

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*



---

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

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit [www.cengage.com/highered](http://www.cengage.com/highered) to search by ISBN#, author, title, or keyword for materials in your areas of interest.

Important Notice: Media content referenced within the product description or the product text may not be available in the eBook version.

**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

© 2019, Cengage Learning EMEA

WCN: 02-300

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced or distributed in any form or by any means, except as permitted by U.S. copyright law, without the prior written permission of the copyright owner.

For product information and technology assistance, contact us at  
**[emea.info@cengage.com](mailto:emea.info@cengage.com)**

For permission to use material from this text or product and for  
permission queries, email **[emea.permissions@cengage.com](mailto: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](http://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](http://www.cengage.co.uk)**

Purchase any of our products at your local college store or at our preferred online store **[www.cengagebrain.com](http://www.cengagebrain.com)**.

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

	<b>1</b>	Overview of Multivariate Methods	1
<b>SECTION I</b>		<b>Preparing for Multivariate Analysis</b>	43
	<b>2</b>	Examining Your Data	45
<b>SECTION II</b>		<b>Interdependence Techniques</b>	119
	<b>3</b>	Exploratory Factor Analysis	121
	<b>4</b>	Cluster Analysis	189
<b>SECTION III</b>		<b>Dependence Techniques – Metric Outcomes</b>	257
	<b>5</b>	Multiple Regression Analysis	259
	<b>6</b>	MANOVA: Extending ANOVA	371
<b>SECTION IV</b>		<b>Dependence Techniques – Non-metric Outcomes</b>	469
	<b>7</b>	Multiple Discriminant Analysis	471
	<b>8</b>	Logistic Regression: Regression with a Binary Dependent Variable	548
<b>SECTION V</b>		<b>Moving Beyond The Basics</b>	601
	<b>9</b>	Structural Equation Modeling: An Introduction	603
	<b>10</b>	SEM: Confirmatory Factor Analysis	658
	<b>11</b>	Testing Structural Equation Models	699
	<b>12</b>	Advanced SEM Topics	726
	<b>13</b>	Partial Least Squares Structural Equation Modeling (PLS-SEM)	759

# Contents

Preface xiv  
Acknowledgments xvii

## 1 Overview of Multivariate Methods 1

### What Is Multivariate Analysis? 3

#### Three Converging Trends 4

Topic 1: Rise of Big Data 4  
Topic 2: Statistical Versus Data Mining Models 7  
Topic 3: Causal Inference 9  
Summary 9

#### Multivariate Analysis in Statistical Terms 9

#### Some Basic Concepts of Multivariate Analysis 10

The Variate 10  
Measurement Scales 11  
Measurement Error and Multivariate Measurement 13

#### Managing the Multivariate Model 14

Managing the Variate 14  
Managing the Dependence Model 17  
Statistical Significance Versus Statistical Power 18  
Review 20

#### A Classification of Multivariate Techniques 21

Dependence Techniques 21  
Interdependence Techniques 25  
**Types of Multivariate Techniques 25**  
Exploratory Factor Analysis: Principal Components and Common Factor Analysis 25  
Cluster Analysis 26  
Multiple Regression 26  
Multivariate Analysis of Variance and Covariance 26  
Multiple Discriminant Analysis 26  
Logistic Regression 27  
Structural Equation Modeling and Confirmatory Factor Analysis 27  
Partial Least Squares Structural Equation Modeling 28  
Canonical Correlation 28  
Conjoint Analysis 28  
Perceptual Mapping 29  
Correspondence Analysis 29

#### Guidelines for Multivariate Analyses and Interpretation 29

Establish Practical Significance as Well as Statistical Significance 30

Recognize That Sample Size Affects All Results 30  
Know Your Data 30  
Strive for Model Parsimony 31  
Look at Your Errors 31  
Simplify Your Models By Separation 31  
Validate Your Results 32

### A Structured Approach to Multivariate Model Building 32

Stage 1: Define the Research Problem, Objectives, and Multivariate Technique to Be Used 33  
Stage 2: Develop the Analysis Plan 33  
Stage 3: Evaluate the Assumptions Underlying the Multivariate Technique 33  
Stage 4: Estimate the Multivariate Model and Assess Overall Model Fit 34  
Stage 5: Interpret the Variate(s) 34  
Stage 6: Validate the Multivariate Model 34  
A Decision Flowchart 34

#### Databases 34

Primary Database 35  
Other Databases 37

#### Organization of the Remaining Chapters 37

Section I: Preparing for a Multivariate Analysis 37  
Section II: Interdependence Techniques 38  
Sections III and IV: Dependence Techniques 38  
Section V: Moving Beyond the Basics 38  
Online Resources: Additional Chapters 38

#### Summary 39

#### Questions 41

#### Suggested Readings and Online Resources 41

#### References 41

## SECTION I Preparing for Multivariate Analysis 43

## 2 Examining Your Data 45

### Introduction 49

### The Challenge of Big Data Research Efforts 49

Data Management	50
Data Quality	50
Summary	51
<b>Preliminary Examination of the Data</b>	51
Univariate Profiling: Examining the Shape of the Distribution	51
Bivariate Profiling: Examining the Relationship Between Variables	52
Bivariate Profiling: Examining Group Differences	53
Multivariate Profiles	54
New Measures of Association	55
Summary	55
<b>Missing Data</b>	56
The Impact of Missing Data	56
Recent Developments in Missing Data Analysis	57
A Simple Example of a Missing Data Analysis	57
A Four-Step Process for Identifying Missing Data and Applying Remedies	58
An Illustration of Missing Data Diagnosis with the Four-Step Process	72
<b>Outliers</b>	85
Two Different Contexts for Defining Outliers	85
Impacts of Outliers	86
Classifying Outliers	87
Detecting and Handling Outliers	88
An Illustrative Example of Analyzing Outliers	91
<b>Testing the Assumptions of Multivariate Analysis</b>	93
Assessing Individual Variables Versus the Variate	93
Four Important Statistical Assumptions	94
<b>Data Transformations</b>	100
Transformations Related to Statistical Properties	101
Transformations Related to Interpretation	101
Transformations Related to Specific Relationship Types	102
Transformations Related to Simplification	103
General Guidelines for Transformations	104
<b>An Illustration of Testing the Assumptions Underlying Multivariate Analysis</b>	105
Normality	105
Homoscedasticity	108
Linearity	108
Summary	112
<b>Incorporating Nonmetric Data with Dummy Variables</b>	112
Concept of Dummy Variables	112
Dummy Variable Coding	113
Using Dummy Variables	113
<b>Summary</b>	114
<b>Questions</b>	115

<b>Suggested Readings and Online Resources</b>	116
<b>References</b>	116

## SECTION II

### Interdependence Techniques

119

#### 3 Exploratory Factor Analysis

121

##### What Is Exploratory Factor Analysis?

124

##### A Hypothetical Example of Exploratory Factor Analysis

126

##### Factor Analysis Decision Process

127

##### Stage 1: Objectives of Factor Analysis

127

Specifying the Unit of Analysis

Achieving Data Summarization Versus Data Reduction

129

Variable Selection

Using Factor Analysis with Other Multivariate Techniques

##### Stage 2: Designing an Exploratory Factor Analysis

132

Variable Selection and Measurement Issues

Sample Size

Correlations among Variables or Respondents

##### Stage 3: Assumptions in Exploratory Factor Analysis

135

Conceptual Issues

Statistical Issues

Summary

##### Stage 4: Deriving Factors and Assessing Overall Fit

136

Selecting the Factor Extraction Method

Stopping Rules: Criteria for the Number of Factors to Extract

Alternatives to Principal Components and Common Factor Analysis

##### Stage 5: Interpreting the Factors

146

The Three Processes of Factor Interpretation

Factor Extraction

Rotation of Factors

Judging the Significance of Factor Loadings

Interpreting a Factor Matrix

##### Stage 6: Validation of Exploratory Factor Analysis

158

Use of Replication or a Confirmatory Perspective

Assessing Factor Structure Stability

Detecting Influential Observations

**Stage 7: Data Reduction—Additional Uses of Exploratory Factor Analysis Results** 159

Selecting Surrogate Variables for Subsequent Analysis 160

Creating Summated Scales 160

Computing Factor Scores 163

Selecting among the Three Methods 164

**An Illustrative Example** 165

Stage 1: Objectives of Factor Analysis 165

Stage 2: Designing a Factor Analysis 165

Stage 3: Assumptions in Factor Analysis 165

Principal Component Factor Analysis: Stages 4–7 168

Common Factor Analysis: Stages 4 and 5 181

A Managerial Overview of the Results 183

**Summary** 184

**Questions** 187

**Suggested Readings and Online Resources** 187

**References** 187

**4 Cluster Analysis** 189

**What Is Cluster Analysis?** 192

Cluster Analysis as a Multivariate Technique 192

Conceptual Development with Cluster Analysis 192

Necessity of Conceptual Support in Cluster Analysis 193

**How Does Cluster Analysis Work?** 193

A Simple Example 194

Objective Versus Subjective Considerations 199

**Cluster Analysis Decision Process** 199

Stage 1: Objectives of Cluster Analysis 199

Stage 2: Research Design in Cluster Analysis 202

Stage 3: Assumptions in Cluster Analysis 211

Stage 4: Deriving Clusters and Assessing Overall Fit 212

Stage 5: Interpretation of the Clusters 227

Stage 6: Validation and Profiling of the Clusters 228

**Implication of Big Data Analytics** 230

Challenges 230

**An Illustrative Example** 230

Stage 1: Objectives of the Cluster Analysis 231

Stage 2: Research Design of the Cluster Analysis 232

Stage 3: Assumptions in Cluster Analysis 235

Stages 4–6: Employing Hierarchical and Nonhierarchical Methods 235

Part 1: Hierarchical Cluster Analysis (Stage 4) 235

Part 2: Nonhierarchical Cluster Analysis

(Stages 4–6) 245

Examining an Alternative Cluster Solution:

Stages 4–6 251

A Managerial Overview of the Clustering Process 252

**Summary** 253

**Questions** 254

**Suggested Readings and Online Resources** 255

**References** 255

## SECTION III

### Dependence Techniques – Metric Outcomes

257

#### 5 Multiple Regression Analysis 259

**What Is Multiple Regression Analysis?** 265

**Multiple Regression in the Era of Big Data** 265

**An Example of Simple and Multiple Regression** 266

Prediction Using a Single Independent Variable:  
Simple Regression 267

Prediction Using Several Independent Variables:  
Multiple Regression 269

Summary 271

**A Decision Process for Multiple Regression Analysis** 272

**Stage 1: Objectives of Multiple Regression** 273

Research Problems Appropriate for Multiple Regression 273

Specifying a Statistical Relationship 274

Selection of Dependent and Independent Variables 275

**Stage 2: Research Design of a Multiple Regression Analysis** 278

Sample Size 278

Creating Additional Variables 281

Overview 286

**Stage 3: Assumptions in Multiple Regression Analysis** 287

Assessing Individual Variables Versus the Variate 287

Methods of Diagnosis 288

Linearity of the Phenomenon 288

Constant Variance of the Error Term 290

Normality of the Error Term Distribution 291

Independence of the Error Terms 291

Summary 292

**Stage 4: Estimating the Regression Model and Assessing Overall Model Fit** 292

Managing the Variate 292

Variable Specification 294	Multivariate Procedures for Assessing Group Differences 377
Variable Selection 295	
Testing the Regression Variate for Meeting the Regression Assumptions 298	<b>A Hypothetical Illustration of MANOVA</b> 381
Examining the Statistical Significance of Our Model 299	Analysis Design 381
Understanding Influential Observations 302	Differences from Discriminant Analysis 381
<b>Stage 5: Interpreting the Regression Variate</b> 308	Forming the Variate and Assessing Differences 382
Using the Regression Coefficients 308	
Assessing Multicollinearity 311	<b>A Decision Process for MANOVA</b> 383
Relative Importance of Independent Variables 317	<b>Stage 1: Objectives of MANOVA</b> 385
Summary 320	When Should We Use MANOVA? 385
<b>Stage 6: Validation of the Results</b> 321	Types of Multivariate Questions Suitable for MANOVA 385
Additional or Split Samples 321	Selecting the Dependent Measures 386
Calculating the PRESS Statistic 321	<b>Stage 2: Issues in the Research Design of MANOVA</b> 387
Comparing Regression Models 322	Types of Research Approaches 387
Forecasting with the Model 322	Types of Variables in Experimental Research 389
<b>Extending Multiple Regression</b> 322	Sample Size Requirements—Overall and by Group 391
Multilevel Models 323	Factorial Designs—Two or More Treatments 391
Panel Models 328	Using Covariates—ANCOVA and MANCOVA 394
<b>Illustration of a Regression Analysis</b> 331	Modeling Other Relationships Between Treatment and Outcome 396
Stage 1: Objectives of Multiple Regression 331	MANOVA Counterparts of Other ANOVA Designs 397
Stage 2: Research Design of a Multiple Regression Analysis 331	A Special Case of MANOVA: Repeated Measures 397
Stage 3: Assumptions in Multiple Regression Analysis 332	<b>Stage 3: Assumptions of ANOVA and MANOVA</b> 398
Stage 4: Estimating the Regression Model and Assessing Overall Model Fit 332	Independence 399
Stage 5: Interpreting the Regression Variate 348	Equality of Variance–Covariance Matrices 399
Stage 6: Validating the Results 353	Normality 400
<b>Evaluating Alternative Regression Models</b> 355	Linearity and Multicollinearity among the Dependent Variables 401
Confirmatory Regression Model 355	Sensitivity to Outliers 401
Use of Summated Scales as Remedies for Multicollinearity 357	<b>Stage 4: Estimation of the MANOVA Model and Assessing Overall Fit</b> 401
Including a Nonmetric Independent Variable 361	Estimation with the General Linear Model 403
A Managerial Overview of the Results 361	Measures for Significance Testing 403
<b>Summary</b> 363	Statistical Power of the Multivariate Tests 403
<b>Questions</b> 366	Estimating Additional Relationships: Mediation and Moderation 407
<b>Suggested Readings and Online Resources</b> 367	<b>Stage 5: Interpretation of the MANOVA Results</b> 410
<b>References</b> 367	Evaluating Covariates 410
<b>6 MANOVA: Extending ANOVA</b> 371	Assessing Effects on the Dependent Variate 411
<b>Re-Emergence of Experimentation</b> 376	Identifying Differences Between Individual Groups 415
<b>Experimental Approaches Versus Other Multivariate Methods</b> 376	Assessing Significance for Individual Outcome Variables 417
<b>MANOVA: Extending Univariate Methods for Assessing Group Differences</b> 377	Interpreting Mediation and Moderation 419
	<b>Stage 6: Validation of the Results</b> 421
	<b>Advanced Issues: Causal Inference in Nonrandomized Situations</b> 421

Causality in the Social and Behavioral Sciences 422  
 The Potential Outcomes Approach 423  
 Counterfactuals in Non-experimental Research  
   Designs 423  
 Propensity Score Models 424  
 Overview 428  
**Summary** 430

**Illustration of a MANOVA Analysis** 430  
 Research Setting 430

**Example 1: Difference Between Two Independent Groups** 432

Stage 1: Objectives of the Analysis 432  
 Stage 2: Research Design of the MANOVA 433  
 Stage 3: Assumptions in MANOVA 433  
 Stage 4: Estimation of the MANOVA Model and Assessing Overall Fit 434  
 Stage 5: Interpretation of the Results 437  
 Summary 438

**Example 2: Difference Between K Independent Groups** 438

Stage 1: Objectives of the MANOVA 438  
 Stage 2: Research Design of MANOVA 439  
 Stage 3: Assumptions IN MANOVA 439  
 Stage 4: Estimation of the MANOVA Model and Assessing Overall Fit 440  
 Stage 5: Interpretation of the Results 443  
 Summary 444

**Example 3: A Factorial Design for MANOVA with Two Independent Variables** 444

Stage 1: Objectives of the MANOVA 445  
 Stage 2: Research Design of the MANOVA 445  
 Stage 3: Assumptions in MANOVA 447  
 Stage 4: Estimation of the MANOVA Model and Assessing Overall Fit 448  
 Stage 5: Interpretation of the Results 451  
 Summary 452

**Example 4: Moderation and Mediation** 452

Moderation of Distribution System ( $X_5$ ) by Firm Size ( $X_3$ ) 453  
 Summary 456  
 Mediation of Distribution System ( $X_5$ ) By Purchase Level ( $X_{22}$ ) 457  
 Summary 459

**A Managerial Overview of the Results** 459

**Summary** 460

**Questions** 463

**Suggested Readings and Online Resources** 464

**References** 464

## SECTION IV

### Dependence Techniques – Non-metric Outcomes

469

#### **7 Multiple Discriminant Analysis** 471

##### **What Is Discriminant Analysis?** 474

The Variate 474  
 Testing Hypotheses 475

##### **Similarities to Other Multivariate Techniques** 476

**Hypothetical Example of Discriminant Analysis** 476  
 A Two-Group Discriminant Analysis: Purchasers Versus Non-purchasers 476  
 A Three-Group Example of Discriminant Analysis: Switching Intentions 481

##### **The Decision Process for Discriminant Analysis** 484

**Stage 1: Objectives of Discriminant Analysis** 484  
 Descriptive Profile Analysis 485  
 Classification Purposes 485

##### **Stage 2: Research Design for Discriminant Analysis** 485

Selecting Dependent and Independent Variables 485  
 Sample Size 487  
 Division of the Sample 488

##### **Stage 3: Assumptions of Discriminant Analysis** 488

Impacts on Estimation and Classification 489  
 Impacts on Interpretation 489

##### **Stage 4: Estimation of the Discriminant Model and Assessing Overall Fit** 490

Selecting an Estimation Method 491  
 Statistical Significance 492  
 Assessing Overall Model Fit 493  
 Casewise Diagnostics 501

##### **Stage 5: Interpretation of the Results** 503

Discriminant Weights 503  
 Discriminant Loadings 503  
 Partial F Values 504  
 Interpretation of Two or More Functions 504  
 Which Interpretive Method to Use? 506

##### **Stage 6: Validation of the Results** 506

Validation Procedures 506  
 Profiling Group Differences 507

##### **A Two-Group Illustrative Example** 508

Stage 1: Objectives of Discriminant Analysis 508  
 Stage 2: Research Design for Discriminant Analysis 508  
 Stage 3: Assumptions of Discriminant Analysis 509  
 Stage 4: Estimation of the Discriminant Model and Assessing Overall Fit 509

Stage 5: Interpretation of the Results	520
Stage 6: Validation of the Results	522
A Managerial Overview	523
<b>A Three-Group Illustrative Example</b>	523
Stage 1: Objectives of Discriminant Analysis	524
Stage 2: Research Design for Discriminant Analysis	524
Stage 3: Assumptions of Discriminant Analysis	524
Stage 4: Estimation of the Discriminant Model and Assessing Overall Fit	525
Stage 5: Interpretation of Three-Group Discriminant Analysis Results	537
Stage 6: Validation of the Discriminant Results	542
A Managerial Overview	543
<b>Summary</b>	544
<b>Questions</b>	546
<b>Suggested Readings and Online Resources</b>	547
<b>References</b>	547

## 8 Logistic Regression: Regression with a Binary Dependent Variable

548

<b>What Is Logistic Regression?</b>	551
<b>The Decision Process for Logistic Regression</b>	552
<b>Stage 1: Objectives of Logistic Regression</b>	552
Explanation	552
Classification	553
<b>Stage 2: Research Design for Logistic Regression</b>	553
Representation of the Binary Dependent Variable	553
Sample Size	555
Use of Aggregated Data	556
<b>Stage 3: Assumptions of Logistic Regression</b>	556
<b>Stage 4: Estimation of the Logistic Regression Model and Assessing Overall Fit</b>	557
Estimating the Logistic Regression Model	558
Assessing the Goodness-of-Fit of the Estimated Model	563
Overview of Assessing Model Fit	571
Casewise Diagnostics	571
Summary	572
<b>Stage 5: Interpretation of the Results</b>	572
Testing for Significance of the Coefficients	573
Interpreting the Coefficients	574
Calculating Probabilities for a Specific Value of the Independent Variable	578
Overview of Interpreting Coefficients	579

<b>Stage 6: Validation of the Results</b>	579
<b>An Illustrative Example of Logistic Regression</b>	580
Stage 1: Objectives of Logistic Regression	580
Stage 2: Research Design for Logistic Regression	580
Stage 3: Assumptions of Logistic Regression	581
Stage 4: Estimation of the Logistic Regression Model and Assessing Overall Fit	581
Stage 5: Interpretation of Results	592
Stage 6: Validation of the Results	596
A Managerial Overview	596
<b>Summary</b>	596
<b>Questions</b>	598
<b>Suggested Readings and Online Resources</b>	598
<b>References</b>	598

## SECTION V Moving Beyond The Basics

601

### 9 Structural Equation Modeling: An Introduction

603

<b>What Is Structural Equation Modeling?</b>	607
Estimation of Multiple Interrelated Dependence Relationships	607
Incorporating Latent Variables Not Measured Directly	608
Defining a Model	610
<b>SEM and Other Multivariate Techniques</b>	613
Similarity to Dependence Techniques	613
Similarity to Interdependence Techniques	613
The Emergence of SEM	614
<b>The Role of Theory in Structural Equation Modeling</b>	614
Specifying Relationships	614
Establishing Causation	615
Developing a Modeling Strategy	618
<b>A Simple Example of SEM</b>	619
Theory	619
Setting Up the Structural Equation Model for Path Analysis	620
The Basics of SEM Estimation and Assessment	621
<b>Six Stages in Structural Equation Modeling</b>	625
<b>Stage 1: Defining Individual Constructs</b>	627
Operationalizing the Construct	627
Pretesting	627

## **Stage 2: Developing and Specifying the Measurement Model** 627

SEM Notation 628  
Creating the Measurement Model 629

## **Stage 3: Designing a Study to Produce Empirical Results** 629

Issues in Research Design 629  
Issues in Model Estimation 633

## **Stage 4: Assessing Measurement Model Validity** 635

The Basics of Goodness-of-Fit 635  
Absolute Fit Indices 636  
Incremental Fit Indices 638  
Parsimony Fit Indices 639  
Problems Associated with Using Fit Indices 639  
Unacceptable Model Specification to Achieve Fit 641  
Guidelines for Establishing Acceptable and Unacceptable Fit 641

## **Stage 5: Specifying the Structural Model** 643

## **Stage 6: Assessing the Structural Model Validity** 644

Competitive Fit 645  
Testing Structural Relationships 647

### **Summary** 648

### **Questions** 649

### **Suggested Readings and Online Resources** 649

### **Appendix 9A: Estimating Relationships Using Path Analysis** 650

### **Appendix 9B: SEM Abbreviations** 653

### **Appendix 9C: Detail on Selected GOF Indices** 654

### **References** 656

## **10 SEM: Confirmatory Factor Analysis** 658

### **What Is Confirmatory Factor Analysis?** 660

CFA and Exploratory Factor Analysis 660  
Measurement Theory and Psychometrics 661  
A Simple Example of CFA and SEM 661  
A Visual Diagram 661

### **SEM Stages for Testing Measurement Theory Validation with CFA** 663

#### **Stage 1: Defining Individual Constructs** 663

#### **Stage 2: Developing the Overall Measurement Model** 663

Unidimensionality 664  
Congeneric Measurement Model 665  
Items per Construct 665  
Reflective Versus Formative Measurement 668

## **Stage 3: Designing a Study to Produce Empirical Results** 670

Measurement Scales in CFA 670  
SEM and Sampling 670  
Specifying the Model 670  
Issues in Identification 671  
Problems in Estimation 673

## **Stage 4: Assessing Measurement Model Validity** 673

Assessing Fit 674  
Path Estimates 674  
Construct Validity 675  
Model Diagnostics 677  
Summary Example 681

### **CFA Illustration** 681

Stage 1: Defining Individual Constructs 682  
Stage 2: Developing the Overall Measurement Model 682  
Stage 3: Designing a Study to Produce Empirical Results 684  
Stage 4: Assessing Measurement Model Validity 685  
HBAT CFA Summary 692  
CFA Results Detect Problems 693  
**Summary** 696  
**Questions** 697  
**Suggested Readings and Online Resources** 697  
**References** 697

## **11 Testing Structural Equation Models** 699

### **What Is a Structural Model?** 700

### **A Simple Example of a Structural Model** 701

### **An Overview of Theory Testing with SEM** 702

### **Stages in Testing Structural Theory** 703

One-Step Versus Two-Step Approaches 703

### **Stage 5: Specifying the Structural Model** 703

Unit of Analysis 704  
Model Specification Using a Path Diagram 704  
Designing the Study 708

### **Stage 6: Assessing the Structural Model Validity** 710

Understanding Structural Model Fit from CFA Fit 710  
Examine the Model Diagnostics 712

### **SEM Illustration** 713

Stage 5: Specifying the Structural Model 713  
Stage 6: Assessing the Structural Model Validity 715

### **Summary** 722

### **Questions** 723

### **Suggested Readings and Online Resources** 723

**Appendix 11A** 724

**References** 725

## 12 Advanced SEM Topics 726

### Reflective Versus Formative Scales 728

Reflective Versus Formative Measurement Theory 728  
Operationalizing a Formative Measure 729  
Differences Between Reflective and Formative Measures 730  
Which to Use—Reflective or Formative? 732

### Higher-Order Factor Models 732

Empirical Concerns 733  
Theoretical Concerns 734  
Using Second-Order Measurement Theories 735  
When to Use Higher-Order Factor Analysis 736

### Multiple Groups Analysis 736

Measurement Model Comparisons 737  
Structural Model Comparisons 741

### Measurement Type Bias 742

Model Specification 742  
Model Interpretation 744

### Relationship Types: Mediation and Moderation 744

Mediation 745  
Moderation 748

### Developments in Advanced SEM Approaches 752

Longitudinal Data 752  
Latent Growth Models 752  
Bayesian SEM 753

### Summary 755

### Questions 756

### Suggested Readings and Online Resources 757

### References 757

## 13 Partial Least Squares Structural Equation Modeling (PLS-SEM) 759

### What is PLS-SEM? 764

Structural Model 764  
Measurement Model 764  
Theory and Path Models in PLS-SEM 765  
The Emergence of SEM 765  
Role of PLS-SEM Versus CB-SEM 766

### Estimation of Path Models with PLS-SEM 766

Measurement Model Estimation 766  
Structural Model Estimation 767  
Estimating the Path Model Using the PLS-SEM Algorithm 767

### PLS-SEM Decision Process 768

### Stage 1: Defining Research Objectives and Selecting Constructs 768

### Stage 2: Designing a Study to Produce Empirical Results 769

Metric Versus Nonmetric Data and Multivariate Normality 769  
Missing Data 770  
Statistical Power 770  
Model Complexity and Sample Size 770

### Stage 3: Specifying the Measurement and Structural Models 771

Measurement Theory and Models 773  
Structural Theory and Path Models 774

### Stage 4: Assessing Measurement Model Validity 774

Assessing Reflective Measurement Models 775  
Assessing Formative Measurement Models 776  
Summary 779

### Stage 5: Assessing the Structural Model 779

Collinearity among Predictor Constructs 779  
Examining the Coefficient of Determination 780  
Effect Size 780  
Blindfolding 780  
Size and Significance of Path Coefficients 780  
Summary 781

### Stage 6: Advanced Analyses Using PLS-SEM 782

Multi-Group Analysis of Observed Heterogeneity 782  
Detecting Unobserved Heterogeneity 782  
Confirmatory Tetrad Analysis 782  
Mediation Effects 782  
Moderation 783  
Higher-Order Measurement Models 783  
Summary 783

### PLS-SEM Illustration 783

Theoretical PLS-SEM Path Model 784

### Stage 4: Assessing Measurement Model Reliability and Validity 785

Path Coefficients 785  
Construct Reliability 786  
Construct Validity 787  
HBAT CCA Summary 790

### Stage 5: Assessing the Structural Model 790

HBAT PLS-SEM Summary 791

### Summary 792

### Questions 793

### Suggested Readings and Online Resources 793

### References 793

### Index 800

# 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 its 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](http://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 8<sup>th</sup> 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 these 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](http://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](http://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](http://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.

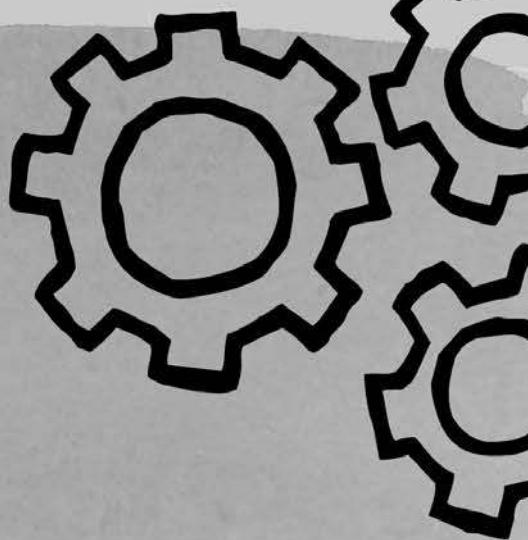
# Teaching & Learning Support Resources

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](http://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](http://cengagebrain.co.uk).



## BE UNSTOPPABLE

Learn more at [cengage.co.uk/education](http://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 ( $\alpha$ )** See *Type I error*.

**Beta ( $\beta$ )** 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 (*summarized 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* ( $\alpha$ ), (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* ( $\alpha$ ). 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* ( $\beta$ ), 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.

Many multivariate techniques are extensions of univariate analysis (analysis of single-variable distributions) and bivariate analysis (cross-classification, correlation, analysis of variance, and simple regression used to analyze two variables). For example, simple regression (with one predictor variable) is extended in the multivariate case to include several predictor variables. Likewise, the single dependent variable found in analysis of variance is extended to include multiple dependent variables in multivariate analysis of variance. Some multivariate techniques (e.g., multiple regression and multivariate analysis of variance) provide a means of performing in a single analysis what once took multiple univariate analyses to accomplish. Other multivariate techniques, however, are uniquely designed to deal with multivariate issues, such as factor analysis, which identifies the structure underlying a set of variables, or discriminant analysis, which differentiates among groups based on a set of variables.

Confusion sometimes arises about what multivariate analysis is because the term is not used consistently in the literature. Some researchers use *multivariate* simply to mean examining relationships between or among more than two variables. Others use the term only for problems in which all the multiple variables are assumed to have a multivariate normal distribution. To be considered truly multivariate, however, all the variables must be random and interrelated in such ways that their different effects cannot meaningfully be interpreted separately. Some authors state that the purpose of multivariate analysis is to measure, explain, and predict the degree of relationship among variates (weighted combinations of variables). Thus, the multivariate character lies in the multiple variates (multiple combinations of variables), and not only in the number of variables or observations. For the purposes of this book, we do not insist on a rigid definition of multivariate analysis. Instead, multivariate analysis will include both multi-variable techniques and truly multivariate techniques, because we believe that knowledge of multivariable techniques is an essential first step in understanding multivariate analysis.

## Some Basic Concepts of Multivariate Analysis

Although the roots of multivariate analysis lie in univariate and bivariate statistics, the extension to the multivariate domain introduces additional concepts and issues of particular relevance. These concepts range from the need for a conceptual understanding of the basic building block of multivariate analysis—the variate—to specific issues dealing with the types of measurement scales used and the statistical issues of significance testing and confidence levels. Each concept plays a significant role in the successful application of any multivariate technique.

### THE VARIATE

As previously mentioned, the building block of multivariate analysis is the **variate**, a linear combination of variables with empirically determined weights. The variables are specified by the researcher whereas the weights are determined by the multivariate technique to meet a specific objective. A variate of  $n$  weighted variables ( $X_1$  to  $X_n$ ) can be stated mathematically as:

$$\text{Variate value} = w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n$$

where  $X_n$  is the observed variable and  $w_n$  is the weight determined by the multivariate technique.

The result is a single value representing a combination of the *entire set* of variables that best achieves the objective of the specific multivariate analysis. In multiple regression, the variate is determined in a manner that maximizes the correlation between the multiple independent variables and the single dependent variable. In discriminant analysis, the variate is formed so as to create scores for each observation that maximally differentiates between groups of observations. In exploratory factor analysis, variates are formed to best represent the underlying structure or patterns of the variables as represented by their intercorrelations.

In each instance, the variate captures the multivariate character of the analysis. Thus, in our discussion of each technique, the variate is the focal point of the analysis in many respects. We must understand not only its

collective impact in meeting the technique's objective but also each separate variable's contribution to the overall variate effect.

## MEASUREMENT SCALES

Data analysis involves the identification and measurement of variation in a set of variables, either among themselves or between a dependent variable and one or more independent variables. The key word here is *measurement* because the researcher cannot identify variation unless it can be measured. Measurement is important in accurately representing the concept of interest and is instrumental in the selection of the appropriate multivariate method of analysis. Data can be classified into one of two categories—nonmetric (qualitative) and metric (quantitative)—based on the type of attributes or characteristics they represent.

The researcher must define the measurement type—nonmetric or metric—for each variable. To the computer, the values are only numbers. As we will see in the following section, defining data as either metric or nonmetric has substantial impact on what the data can represent and how it can be analyzed.

**Nonmetric Measurement Scales** Nonmetric data describe differences in type or kind by indicating the presence or absence of a characteristic or property. These properties are discrete in that by having a particular feature, all other features are excluded; for example, if a person is male, he cannot be female. An “amount” of gender is not possible, just the state of being male or female. Nonmetric measurements can be made with either a nominal or an ordinal scale.

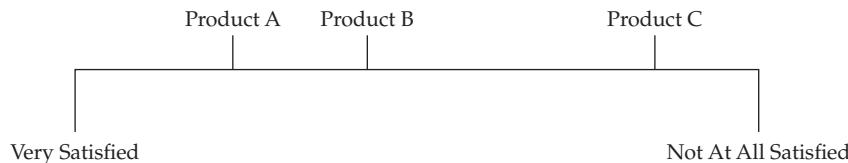
**NOMINAL SCALES** A nominal scale assigns numbers as a way to label or identify subjects or objects. The numbers assigned to the objects have no quantitative meaning beyond indicating the presence or absence of the attribute or characteristic under investigation. Therefore, nominal scales, also known as categorical scales, can only provide the number of occurrences in each class or category of the variable being studied.

For example, in representing gender (male or female) the researcher might assign numbers to each category (e.g., 2 for females and 1 for males). With these values, however, we can only tabulate the number of males and females; it is nonsensical to calculate an average value of gender.

Nominal data only represent categories or classes and do not imply amounts of an attribute or characteristic. Commonly used examples of nominally scaled data include many demographic attributes (e.g., individual's sex, religion, occupation, or political party affiliation), many forms of behavior (e.g., voting behavior or purchase activity), or any other action that is discrete (happens or not).

**ORDINAL SCALES** Ordinal scales are the next “higher” level of measurement precision. In the case of ordinal scales, variables can be ordered or ranked in relation to the amount of the attribute possessed. Every subject or object can be compared with another in terms of a “greater than” or “less than” relationship. The numbers utilized in ordinal scales, however, are really non-quantitative because they indicate only relative positions in an ordered series. Ordinal scales provide no measure of the actual amount or magnitude in absolute terms, only the order of the values. The researcher knows the order, but not the amount of difference between the values.

For example, different levels of an individual consumer's satisfaction with several new products can be illustrated, first using an ordinal scale. The following scale shows a respondent's view of three products.



When we measure this variable with an ordinal scale, we “rank order” the products based on satisfaction level. We want a measure that reflects that the respondent is more satisfied with Product A than Product B and more

satisfied with Product B than Product C, based solely on their position on the scale. We could assign “rank order” values (1 = most satisfied, 2 = next most satisfied, etc.) of 1 for Product A (most satisfaction), 2 for Product B, and 3 for Product C.

When viewed as ordinal data, we know that Product A has the most satisfaction, followed by Product B and then Product C. However, we cannot make any statements on the amount of the differences between products (e.g., we cannot answer the question whether the difference between Products A and B is greater than the difference between Products B and C). We have to use an interval scale (see next section) to assess what is the magnitude of differences between products.

In many instances a researcher may find it attractive to use ordinal measures, but the implications for the types of analyses that can be performed are substantial. The analyst cannot perform any arithmetic operations (no sums, averages, multiplication or division, etc.), thus nonmetric data are quite limited in their use in estimating model coefficients. For this reason, many multivariate techniques are devised solely to deal with nonmetric data (e.g., correspondence analysis) or to use nonmetric data as an independent variable (e.g., discriminant analysis with a nonmetric dependent variable or multivariate analysis of variance with nonmetric independent variables). Thus, the analyst must identify all nonmetric data to ensure that they are used appropriately in the multivariate techniques.

**Metric Measurement Scales** In contrast to nonmetric data, **metric data** are used when subjects differ in amount or degree on a particular attribute. Metrically measured variables reflect relative quantity or degree and are appropriate for attributes involving amount or magnitude, such as the level of satisfaction or commitment to a job. The two different metric measurement scales are interval and ratio scales.

**INTERVAL SCALES** Interval scales and ratio scales (both metric) provide the highest level of measurement precision, permitting nearly any mathematical operation to be performed. These two scales have constant units of measurement, so differences between any two adjacent points on any part of the scale are equal.

In the preceding example in measuring satisfaction, metric data could be obtained by measuring the distance from one end of the scale to each product’s position. Assume that Product A was 2.5 units from the left end, Product B was 6.0 units, and Product C was 12 units. Using these values as a measure of satisfaction, we could not only make the same statements as we made with the ordinal data (e.g., the rank order of the products), but we could also see that the difference between Products A and B was much smaller ( $6.0 - 2.5 = 3.5$ ) than was the difference between Products B and C ( $12.0 - 6.0 = 6.0$ ).

The only real difference between interval and ratio scales is that interval scales use an arbitrary zero point, whereas ratio scales include an absolute zero point. The most familiar interval scales are the Fahrenheit and Celsius temperature scales. Each uses a different arbitrary zero point, and neither indicates a zero amount or lack of temperature, because we can register temperatures below the zero point on each scale. Therefore, it is not possible to say that any value on an interval scale is a multiple of some other point on the scale.

For example, an  $80^{\circ}\text{F}$  day cannot correctly be said to be twice as hot as a  $40^{\circ}\text{F}$  day, because we know that  $80^{\circ}\text{F}$ , on a different scale, such as Celsius, is  $26.7^{\circ}\text{C}$ . Similarly,  $40^{\circ}\text{F}$  on a Celsius scale is  $4.4^{\circ}\text{C}$ . Although  $80^{\circ}\text{F}$  is indeed twice  $40^{\circ}\text{F}$ , one cannot state that the heat of  $80^{\circ}\text{F}$  is twice the heat of  $40^{\circ}\text{F}$  because, using different scales, the heat is not twice as great; that is,  $4.4^{\circ}\text{C} \times 2 \neq 26.7^{\circ}\text{C}$ .

**RATIO SCALES** Ratio scales represent the highest form of measurement precision because they possess the advantages of all lower scales plus an absolute zero point. All mathematical operations are permissible with ratio-scale measurements. The bathroom scale or other common weighing machines are examples of these scales, because they have an absolute zero point and can be spoken of in terms of multiples when relating one point on the scale to another; for example, 100 pounds is twice as heavy as 50 pounds.

**The Impact of Choice of Measurement Scale** Understanding the different types of measurement scales is important for two reasons:

- 1 The researcher must identify the measurement scale of each variable used, so that nonmetric data are not incorrectly used as metric data, and vice versa (as in our earlier example of representing gender as 1 for male and 2 for female). If the researcher incorrectly defines this measure as metric, then it may be used inappropriately (e.g., finding the mean value of gender).
- 2 The measurement scale is also critical in determining which multivariate techniques are the most applicable to the data, with considerations made for both independent and dependent variables. In the discussion of the techniques and their classification in later sections of this chapter, the metric or nonmetric properties of independent and dependent variables are the determining factors in selecting the appropriate technique.

## MEASUREMENT ERROR AND MULTIVARIATE MEASUREMENT

The use of multiple variables and the reliance on their combination (the variate) in multivariate techniques also focuses attention on a complementary issue—measurement error. **Measurement error** is the degree to which the observed values are not representative of the “true” values. Measurement error has many sources, ranging from data entry errors to the imprecision of the measurement (e.g., imposing 7-point rating scales for attitude measurement when the researcher knows the respondents can accurately respond only to a 3-point rating) to the inability of respondents to accurately provide information (e.g., responses as to household income may be reasonably accurate but rarely totally precise). *Thus, all variables used in multivariate techniques must be assumed to have some degree of measurement error.* The measurement error adds “noise” to the observed or measured variables. Thus, the observed value obtained represents both the “true” level and the “noise.” When used to compute correlations or means, the “true” effect is partially masked by the measurement error, causing the correlations to weaken and the means to be less precise. The specific impact of measurement error and its accommodation in dependence relationships is covered in more detail in Chapter 9.

**Validity and Reliability** The researcher’s goal of reducing measurement error can follow several paths. In assessing the degree of measurement error present in any measure, the researcher must address two important characteristics of a measure:

**VALIDITY** **Validity** is the degree to which a measure accurately represents what it is supposed to. For example, if we want to measure discretionary income, we should not ask about total household income. Ensuring validity starts with a thorough understanding of what is to be measured and then making the measurement as “correct” and accurate as possible. However, accuracy does not ensure validity. In our income example, the researcher could precisely define total household income, but it would still be “wrong” (i.e., an invalid measure) in measuring discretionary income because the “correct” question was not being asked.

**RELIABILITY** If validity is assured, the researcher must still consider the reliability of the measurements. **Reliability** is the degree to which the observed variable measures the “true” value and is “error free”; thus, it is the opposite of measurement error. If the same measure is asked repeatedly, for example, more reliable measures will show greater consistency than less reliable measures. The researcher should always assess the variables being used and, if valid alternative measures are available, choose the variable with the higher reliability.

**Employing Multivariate Measurement** In addition to reducing measurement error by improving individual variables, the researcher may also choose to develop **multivariate measurements**, also known as **summed scales**, for which several variables are joined in a **composite measure** to represent a concept (e.g., multiple-item personality

scales or summed ratings of product satisfaction). The objective is to avoid the use of only a single variable to represent a concept and instead to use several variables as **indicators**, all representing differing facets of the concept to obtain a more well-rounded perspective. The use of multiple indicators enables the researcher to more precisely specify the desired responses. It does not place total reliance on a single response, but instead on the “average” or typical response to a set of related responses.

For example, in measuring satisfaction, one could ask a single question, “How satisfied are you?” and base the analysis on the single response. Or a summated scale could be developed that combined several responses of satisfaction (e.g., finding the average score among three measures—overall satisfaction, the likelihood to recommend, and the probability of purchasing again). The different measures may be in different response formats or in differing areas of interest assumed to comprise overall satisfaction.

The guiding premise is that multiple responses reflect the “true” response more accurately than does a single response. The researcher should assess reliability and incorporate scales into the analysis. For a more detailed introduction to multiple measurement models and scale construction, see further discussion in Chapter 3 (Exploratory Factor Analysis) and Chapter 9 (Structural Equations Modeling Overview) or additional resources [48]. In addition, compilations of scales that can provide the researcher a “ready-to-go” scale with demonstrated reliability have been published in recent years [2, 9].

**The Impact of Measurement Error** The impact of measurement error and poor reliability cannot be directly seen because they are embedded in the observed variables. The researcher must therefore always work to increase reliability and validity, which in turn will result in a more accurate portrayal of the variables of interest. Poor results are not always due to measurement error, but the presence of measurement error is guaranteed to distort the observed relationships and make multivariate techniques less powerful. Reducing measurement error, although it takes effort, time, and additional resources, may improve weak or marginal results and strengthen proven results as well.

## Managing the Multivariate Model

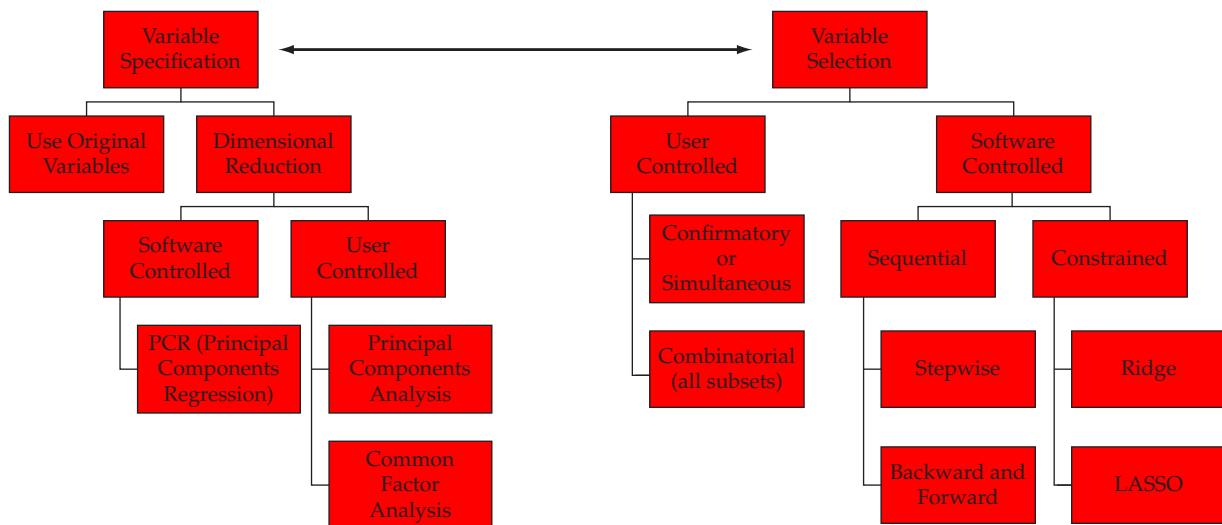
When applying any of the multivariate techniques, the researcher must make a number of important decisions regarding how key elements of the multivariate model itself are handled. In this section we address three such issues that impact “How” multivariate models are used: managing the variate; managing the dependence model and managing statistical significance versus statistical power. Each of these topics focuses on basic principles applicable to all of the statistical models we will discuss in this text.

### MANAGING THE VARIATE

Perhaps the most critical element of a multivariate analysis is the variate, which as described earlier is that set of variables from which multivariate analysis gets its name. The primary benefit of multivariate analysis is to incorporate a set of variables into the analysis for estimating their impacts simultaneously, versus a single variable at a time in univariate analyses. As a result, the researcher has the ability to utilize any number of variables for purposes of explanation and/or prediction.

Any time a variate has two or more variables there is the potential for **multicollinearity**—the degree of correlation among the variables in the variate that may result in a confounding effect in the interpretation of the individual variables of the variate. The explanatory impact of each variable can be divided into (a) the portion of variance the variable shares with other variables in the variate, and (b) the portion which is unique to the variable (see Chapter 5 for a more detailed discussion of shared versus unique effects). Multicollinearity is the measure of the shared variance with other variables in the variate. This distinction becomes critical depending on whether the objective of the analysis is prediction versus explanation. When focused solely on predictions with no need for explanation of the individual variables’ impact, multicollinearity has no real impact. But when the objective shifts to explanation, as it often does with multiple regression, we want to assess the impact for any individual independent variable, and yet the model can only represent the unique effect of each independent variable on the outcome. Any shared explanatory variance with

**Figure 1.2**  
Managing The Variate



other independent variables cannot be attributed to any specific variable. As a result, as multicollinearity increases, the impact of individual variables is under-estimated by the parameters as their explanatory variance shifts from unique explanatory variance to shared explanatory variance. While we do want to understand the unique effects of each variable, multicollinearity tends to diminish the effects attributed to variables based on their correlation with other predictor variables. This results in the principal tradeoff always facing the analyst—including more variables as predictors to increase overall predictive power versus the multicollinearity introduced by more variables which makes it more difficult to attribute the explanatory effect to specific variables.

The decisions by the researcher in managing the variate fall in two primary areas: specifying the independent variables to be included in the analysis and then variable selection during model estimation. Figure 1.2 depicts these two decisions and the options generally available to researchers in each.

**Specifying the Variate Variables** The primary decision here is whether to use the individual variables or to perform some form of dimensional reduction, such as exploratory factor analysis. Using the original variables may seem like the obvious choice as it preserves the characteristics of the variables and may make the results more interpretable and credible. But there are also pitfalls to this approach. First and foremost is the effect of including hundreds and even thousands of variables and then trying to interpret the impact of each variable, thereby identifying the most impactful variables from a large set. Moreover, as the number of variables increases so does the opportunity for multicollinearity that makes distinguishing the impact of individual variables even more difficult.

The alternative is to perform some form of **dimensional reduction**—finding combinations of the individual variables that captures the multicollinearity among a set of variables and allows for a single composite value representing the set of variables. This is the objective of exploratory factor analysis discussed in Chapter 3 which finds these groups of multicollinear variables, forms composites and the researcher then uses the composites in further analyses rather than the original variables. It may seem like dimensional reduction is the option to choose, and many times it is. But the researcher must also recognize that now the “variables” in the analysis are composites, and the impacts for a composite represent the shared effect of those variables, not the individual variables themselves.

As shown in Figure 1.2 dimensional reduction can be performed (a) under the control of the researcher through principal components or common factor analysis, or (b) defined by the software program in techniques like principal components regression (see Chapter 5). We cannot stress the importance of considering the impact of multicollinearity. As you will see in our discussion of each technique, multicollinearity can have a profound effect and markedly shape the interpretation of the results, if not the results themselves. Many researchers are tempted to ignore

this aspect of managing the variate and instead rely solely on the various variable selection techniques described next. But this has its own pitfalls and in most instances the researcher is ceding control of this critical element of the analysis to the software.

**Variable Selection** The second decision to make regarding the variate is if the researcher wants to control the specific variables to be included in the analysis or let the software determine the “best” set of variables to constitute the variate. As with the prior decision, it fundamentally revolves around the level of researcher control. With simultaneous (all variables entered simultaneously) or confirmatory (only a set or sequential sets of variables tested), the researcher can control the exact variables in the model. The combinatorial approach is a variant of the confirmatory approach where all possible combinations of the set of independent variables are estimated and then compared on various model fit criteria. With software control the software employs an algorithm to decide which variables are to be concluded. The most widely used method is the sequential approach, where variables are entered (typically most impactful first) until no other impactful variables can be found. The constrained approach identifies the most impactful variables and constrains all the lesser variables to smaller or even zero estimated parameters. More details on the options in this area are discussed in Chapter 5.

**Reconciling the Two Decisions** These two decisions in managing the variate give the researcher a range of control—complete researcher control over the variables and the model specification, versus total software control in determining the variables input into the analysis and the final set of variables included in the model. While the researcher may be tempted to let the software have complete control based on the notion that “the software knows best,” allowing total software control has potential pitfalls that can seriously impact the results. We therefore encourage researchers to carefully review Chapter 3 (Exploratory factor analysis) where dimensional reduction techniques are discussed, and Chapter 5 (Multiple regression) where the options for variable selection are extensively discussed. There is not a “right” approach for all analyses. Instead, often several principles can be identified:

**SPECIFICATION OF THE VARIATE IS CRITICAL** Many times researchers only focus on how various methods operate in terms of estimating the model of interest. And while technique selection is an important issue, many times the “success or failure” of a research project is dictated by the approach the researcher takes to specification of the variate. Including hundreds or even thousands of variables in an attempt to completely cover all the possible impacts may actually hinder the ability of the model to recover more generalized effects and thus the efficacy of the results. Thus, the more control the researcher retains on which variables are inputs to the model allows for more specificity in how the model answers the specific research question.

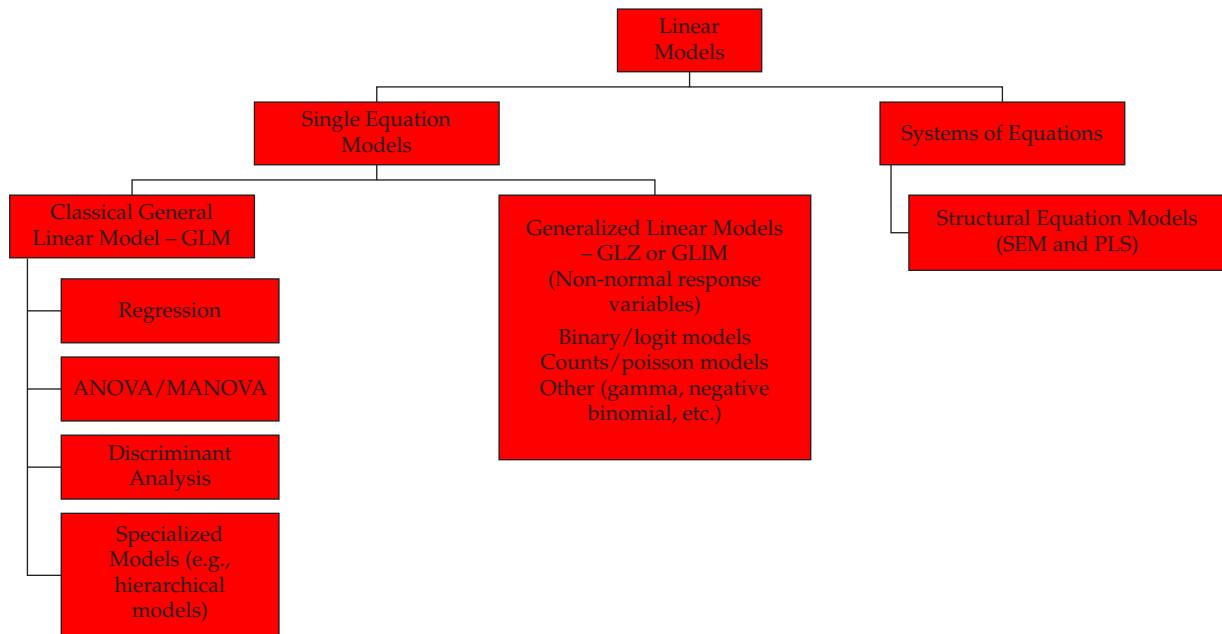
**VARIABLE SELECTION IS NECESSARY** Even with researcher control in specifying the variate, empirical models provide flexibility in testing alternative or competing models which vary in the number of variables included. Any estimated model can have a number of “competing or alternative” models that provide equal or even greater predictive accuracy by including a different set of variables. This is disconcerting to beginning researchers when they find there are many models that work just as well as their selected model, but have different variables and effects. Moreover, the possibility of several alternative models increases as the number of variables grows larger. So researchers should try a number of alternative models in specifying their research, perhaps formulating several alternative models that vary based on whether the researcher controls the process or the software. Chapter 5 explores this consideration by estimating a range of model forms.

**RESEARCHER CONTROL IS PREFERRED** A common theme in both of these decision areas is that while options exist for software control, researcher control provides for the analysis to test many different model specifications (both in variables used and variables selected in the model estimation) and provides a better outcome. Just letting the software make both of these decisions to some extent makes the researcher irrelevant and lets the software dictate the final model. Hopefully the researcher can bring more domain knowledge, coupled with the knowledge of how to “control” the methods to obtain the best outcome, not just let it be dictated by some software algorithm.

## MANAGING THE DEPENDENCE MODEL

While we discuss both interdependence and dependence models in this text and provide a means for selecting between the alternative methods (see Figure 1.6), we also feel that we should focus in more detail on the range of dependence models that are available to researchers. In Figure 1.3 we distinguish between dependence models based on two factors: single equation versus multi-equation models, and among the single equation models, do we use the general linear model or the generalized linear model? We discuss the single versus multiple equation distinction first and then examine the two basic estimation methods for single equation models.

**Figure 1.3**  
Managing Dependence Models



**Single Versus Multiple Equation** The most common applications of multivariate models are the single equation forms—multiple regression, discriminant analysis, ANOVA/MANOVA and logistic regression. All of these model forms provide an approach for specifying a single variate's relationship with an outcome variable. But there are also multiple equation models that enable the researcher to relate different equations to one another, even where they are interrelated (i.e., the outcome measure on one equation becomes the predictor variable in another equation). Most often typified by the structural equation “family” of models (covariance-based structural equation modeling and variance-based partial-least squares modeling), researchers are able to look at sequential models (e.g.,  $X \rightarrow Y$  and then  $Y \rightarrow Z$ ) in a single analysis. Chapters 9 through 13 discuss this family of structural equation modeling approaches in much more detail.

**General Linear Model Versus Generalized Linear Model** The foundation for almost all of the single equation techniques discussed in this book is the **general linear model (GLM)**, which can estimate canonical correlation, multiple regression, ANOVA, MANOVA, and discriminant analysis, as well as all of the univariate group comparisons – t test and such [40, 49, 12]. There is perhaps no single model form more fundamental to inferential statistics than the general linear model. But one limiting characteristic of the GLM is its assumption of an error distribution following the normal distribution. As such, many times we must transform the dependent variable when we know it does not follow a normal distribution (e.g., counts, binary variables, proportions or probabilities).

A second class of linear models is also available that accommodates non-normal outcome variables, thus eliminating the need to transform to dependent variable. Logistic regression discussed in Chapter 8 is one such model form where the dependent variable is binary and the logit transformation “links” the dependent and independent variables. Known as the **generalized linear model (GLZ or GLIM)**, this model provides the researcher with an alternative to the general linear model that is based on a dependent variable exhibiting a normal distribution. While the GLM requires a transformation of a non-normal dependent variable as discussed above, the GLZ can model them directly without transformation. The GLZ model uses maximum likelihood estimation and thus has a different set of model fit measures, including Wald and Likelihood ratio tests and deviance. These fit measures will be discussed in Chapter 8 as a means of assessing a logistic regression model. The GLZ is sometimes referred to as a GLM causing confusion with the general linear model. We make the distinction for purposes of clarification.

While the general linear model has been a staple of inferential statistics, the generalized linear model extends the linear model to a wider range of outcome variables. Outside of the different measures of model fit, both model types are estimated and evaluated in the same manner. Researchers encountering situations in which the dependent variables have a non-normal distribution are encouraged to consider using GLZ models as an alternative to transforming the dependent measures to achieve normality. A more thorough discussion of the GLZ procedure and its many variations are available in several texts [1, 23, 29, 35, 41].

**Summary** The dependence model is the key element of analytics, specifying and then testing the relationships between a variate of independent variables and one or more outcome measures. It has many forms based on the measurement qualities of the variables involved that we discuss in a later chapters. But the analyst must also consider the multiple types of dependence models available to best address the research question. Many times a single equation form is the appropriate model, but there are situations in which the research question actually involves several different relationships that interact together. Estimating each relationship separately will not identify the interrelationships between relationships that may be key to understanding. So in these situations a multi-equation approach is best suited and some form of structural equation modeling (CB-SEM or PLS-SEM) can be employed.

Within the single-equation form there are two types of models, differentiated on the distribution of the dependent measure. The general linear model (GLM) is the model underlying most of the widely used statistical techniques, but it is limited to dependent variables with a normal distribution. For non-normal dependent variables, we can either transform them to hopefully confirm to the normal distribution or use the generalized linear model (GLZ or GLIM) that explicitly allows the researcher to specify the error term distribution and thus avoid transformation of the dependent measure. While the use of maximum likelihood estimation requires a different set of model fit measures than the GLM, they are directly comparable and easily used in the same manner as their GLM counterparts (see Chapter 8 for an example). We encourage researchers to consider the GLZ model when faced with research questions that are not directly estimable by the GLM model.

## STATISTICAL SIGNIFICANCE VERSUS STATISTICAL POWER

All the multivariate techniques, except for cluster analysis and perceptual mapping, are based on the statistical inference of a population’s values or relationships among variables from a randomly drawn sample of that population. A census of the entire population makes statistical inference unnecessary, because any difference or relationship, however small, is true and does exist. Researchers very seldom use a census. Therefore, researchers are often interested in drawing inferences from a sample.

**Types of Statistical Error and Statistical Power** Interpreting statistical inferences requires the researcher to specify the acceptable levels of statistical error that result from using a sample (known as *sampling error*). The most common approach is to specify the level of **Type I error**, also known as **alpha ( $\alpha$ )**. Type I error is the probability of rejecting the null hypothesis when it is actually true—generally referred to as a *false positive*. By specifying an alpha level,

the researcher sets the acceptable limits for error and indicates the probability of concluding that significance exists when it really does not.

When specifying the level of Type I error, the researcher also determines an associated error, termed **Type II error, or beta ( $\beta$ )**. The Type II error is the probability of not rejecting the null hypothesis when it is actually false. An extension of Type II error is  $1 - \beta$ , referred to as the **power** of the statistical inference test. Power is the probability of correctly rejecting the null hypothesis when it should be rejected. Thus, power is the probability that statistical significance will be indicated if it is present. The relationship of the different error probabilities in testing for the difference in two means is shown in Figure 1.4.

Statistical Decision	Reality		Type II error
	No Difference	Difference	
$H_0$ : No Difference	$1 - \alpha$	$\beta$	
$H_a$ : Difference	Type I error	$1 - \beta$	Power

**Figure 1.4**  
Relationship of Error Probabilities in Statistical Inference

Although specifying alpha establishes the level of acceptable statistical significance, it is the level of power that dictates the probability of success in finding the differences if they actually exist. Why not set both alpha and beta at acceptable levels? Because the Type I and Type II errors are inversely related. Thus, Type I error becomes more restrictive (moves closer to zero) as the probability of a Type II error increases. That is, reducing Type I errors reduces the power of the statistical test. Thus, researchers must strike a balance between the level of alpha and the resulting power.

**Impacts on Statistical Power** But why can't high levels of power always be achieved? Power is not solely a function of alpha. Power is determined by three factors: effect size, significance level ( $\alpha$ ) and sample size.

**EFFECT SIZE** The probability of achieving statistical significance is based not only on statistical considerations, but also on the actual size of the effect. Thus, the **effect size** helps researchers determine whether the observed relationship (difference or correlation) is meaningful. For example, the effect size could be a difference in the means between two groups or the correlation between variables. If a weight loss firm claims its program leads to an average weight loss of 25 pounds, the 25 pounds is the effect size. Similarly, if a university claims its MBA graduates get a starting salary that is 50 percent higher than the average, the percentage is the effect size attributed to earning the degree. When examining effect sizes, a larger effect is more likely to be found than a smaller effect and is thus more likely to impact the power of the statistical test.

To assess the power of any statistical test, the researcher must first understand the effect being examined. Effect sizes are defined in standardized terms for ease of comparison. Mean differences are stated in terms of standard deviations, thus an effect size of .5 indicates that the mean difference is one-half of a standard deviation. For correlations, the effect size is based on the actual correlation between the variables.

**SIGNIFICANCE LEVEL ( $\alpha$ )** As alpha ( $\alpha$ ) becomes more restrictive (e.g., moving from .10 to .05 to .01), power decreases. Therefore, as the researcher reduces the chance of incorrectly saying an effect is significant when it is not, the probability of correctly finding an effect decreases. Conventional guidelines suggest alpha levels of .05 or .01. Researchers should consider the impact of a particular alpha level on the power before selecting the alpha level. The relationship of these two probabilities is illustrated in later discussions.

**SAMPLE SIZE** At any given alpha level, increased sample sizes always produce greater power for the statistical test. As sample sizes increase, researchers must decide if the power is too high. By "too high" we mean that by increasing sample size, smaller and smaller effects (e.g., correlations) will be found to be statistically significant, until at very large sample sizes almost any effect is significant. The researcher must always be aware that sample size can

affect the statistical test either by making it insensitive (at small sample sizes) or overly sensitive (at very large sample sizes).

**Managing the Three Elements of Statistical Power** The relationships among alpha, sample size, effect size, and power are complicated, but a number of sources are available for consideration. Cohen [15] examines power for most statistical inference tests and provides guidelines for acceptable levels of power, suggesting that studies be designed to achieve alpha levels of at least .05 with power levels of 80 percent. To achieve such power levels, all three factors—alpha, sample size, and effect size—must be considered simultaneously. These interrelationships can be illustrated by a simple example.

The example involves testing for the difference between the mean scores of two groups. Assume that the effect size is thought to range between small (.2) and moderate (.5). The researcher must now determine the necessary alpha level and sample size of each group. Figure 1.5 illustrates the impact of both sample size and alpha level on power. Note that with a moderate effect size power reaches acceptable levels at sample sizes of 100 or more for alpha levels of both .05 and .01. But when the effect size is small, statistical tests have little power, even with more flexible alpha levels or samples sizes of 200 or more. For example, if the effect size is small a sample of 200 with an alpha of .05 still has only a 50 percent chance of significant differences being found. This suggests that if the researcher expects that the effect sizes will be small the study must have much larger sample sizes and/or less restrictive alpha levels (e.g., .10).

**Figure 1.5**

Power Levels for the Comparison of Two Means: Variations by Sample Size, Significance Level, and Effect Size

Sample Size	alpha ( $\alpha$ ) = .05		alpha ( $\alpha$ ) = .01	
	Effect Size (ES)	Small (.2)	Effect Size (ES)	Moderate (.5)
20	.095	.338	.025	.144
40	.143	.598	.045	.349
60	.192	.775	.067	.549
80	.242	.882	.092	.709
100	.290	.940	.120	.823
150	.411	.990	.201	.959
200	.516	.998	.284	.992

Source: SOLO Power Analysis, BMDP Statistical Software, Inc. [4]

**Using Power with Multivariate Techniques** Researchers can use power analysis either in the study design or after data is collected. In designing research studies, the sample size and alpha level are selected to achieve the desired power. Power also is examined after analysis is completed to determine the actual power achieved so the results can be interpreted. Are the results due to effect sizes, sample sizes, or significance levels? Each of these factors is assessed to determine their impact on the significance or non-significance of the results. Researchers can refer to published studies for specifics on power determination [15] or access websites that assist in planning studies to achieve the desired power or calculate the power of actual results [4, 8]. Specific guidelines for multiple regression and multivariate analysis of variance—the most common applications of power analysis—are discussed in more detail in Chapters 5 and 6.

## REVIEW

Having addressed the issues in extending multivariate techniques from their univariate and bivariate origins, we present a classification scheme to assist in the selection of the appropriate technique by specifying the research objectives (independence or dependence relationship) and the data type (metric or nonmetric). We then briefly introduce each multivariate method discussed in the text.

## Managing the Variate

Dimensional reduction through exploratory factor analysis is a researcher-controlled method for addressing multicollinearity.

Software-controlled variable selection methods should be used with caution.

## Managing Dependence Models

Generalized linear models (GLZ or GLIM) provide an alternative for dependent variables which do not follow the normal distribution.

## Statistical Power Analysis

Researchers should design studies to achieve a power level of .80 at the desired significance level.

More stringent significance levels (e.g., .01 instead of .05) require larger samples to achieve the desired power level.

Conversely, power can be increased by choosing a less stringent alpha level (e.g., .10 instead of .05).

Smaller effect sizes require larger sample sizes to achieve the desired power.

An increase in power is most likely achieved by increasing the sample size.

# A Classification of Multivariate Techniques

To assist you in becoming familiar with the specific multivariate techniques, we present a classification of multivariate methods in Figure 1.6. This classification is based on three judgments the researcher must make about the research objective and nature of the data:

- 1 Can the variables be divided into independent and dependent classifications based on some theory?
- 2 If they can, how many variables are treated as dependent in a single analysis?
- 3 How are the variables, both dependent and independent, measured?

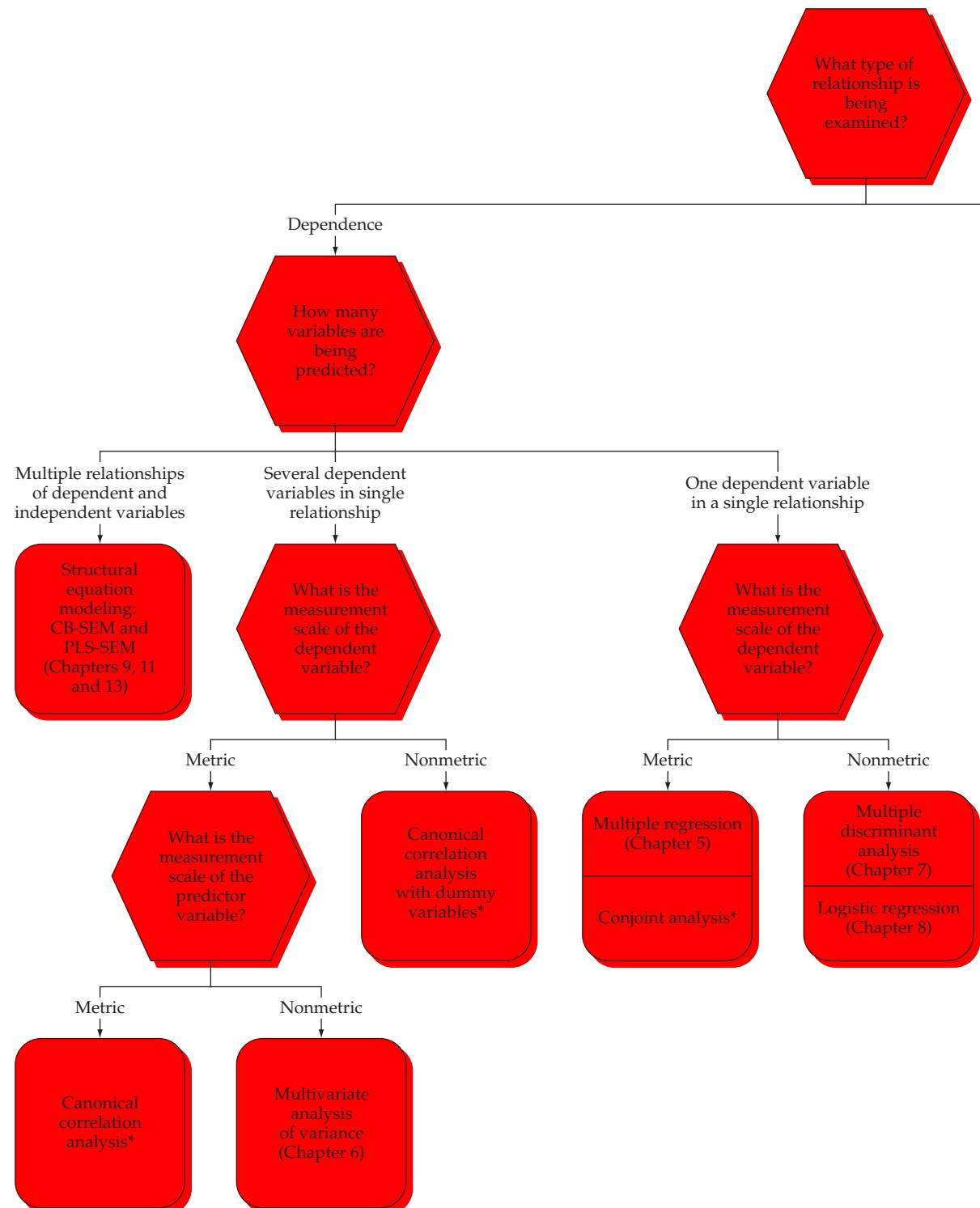
Selection of the appropriate multivariate technique depends on the answers to these three questions.

When considering the application of multivariate statistical techniques, the answer to the first question indicates whether a dependence or interdependence technique should be utilized. Note that in Figure 1.6, the dependence techniques are on the left side and the interdependence techniques are on the right. A **dependence technique** may be defined as one in which a variable or set of variables is identified as the **dependent variable** to be predicted or explained by other variables known as **independent variables**. An example of a dependence technique is multiple regression analysis. In contrast, an **interdependence technique** is one in which no single variable or group of variables is defined as being independent or dependent. Rather, the procedure involves the simultaneous analysis of all variables in the set. Exploratory factor analysis is an example of an interdependence technique. Let us focus on dependence techniques first and use the classification in Figure 1.6 to select the appropriate multivariate method.

## DEPENDENCE TECHNIQUES

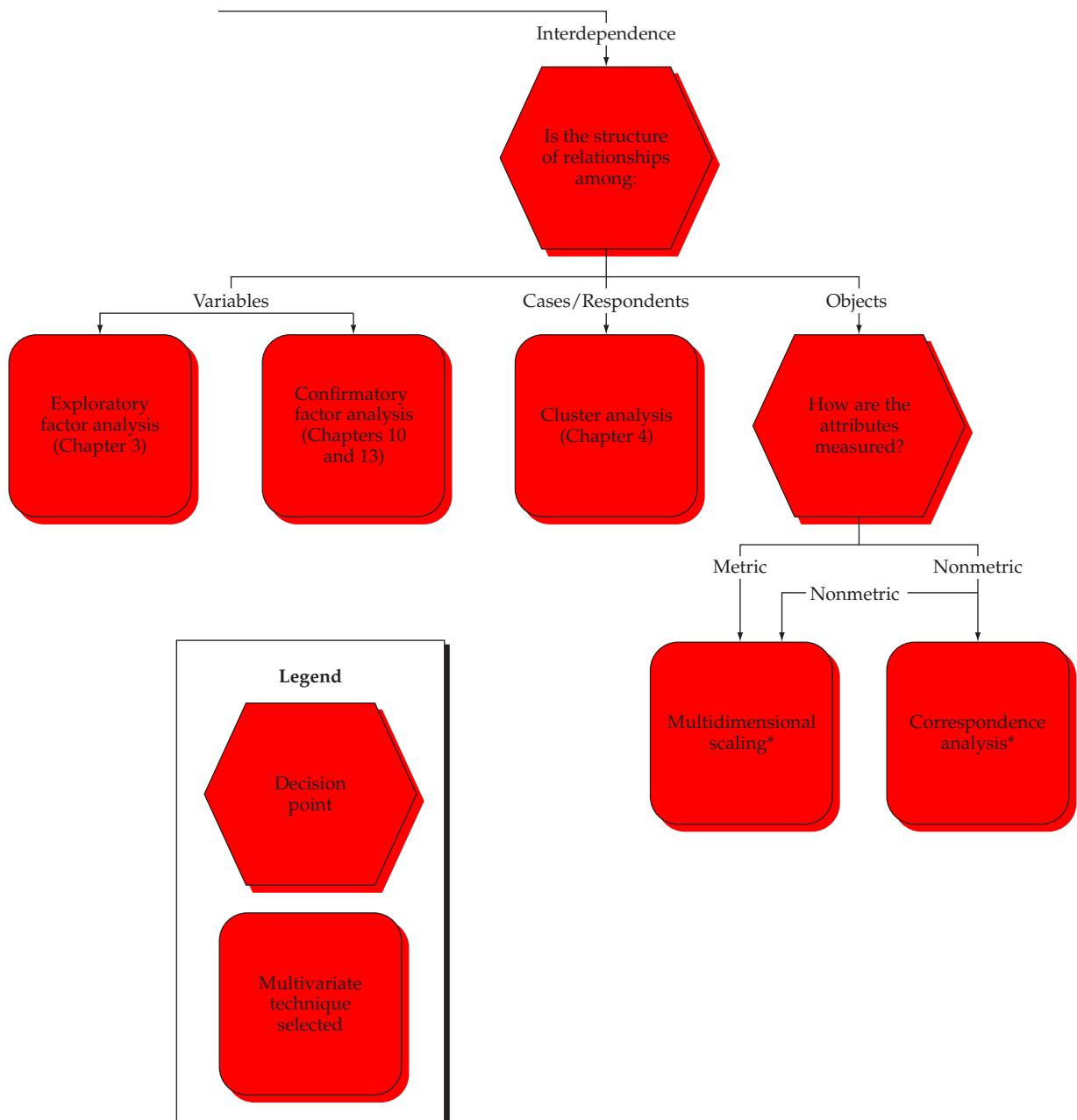
The different dependence techniques can be categorized by two characteristics: (1) the number of dependent variables and (2) the type of measurement scale employed by the variables. First, regarding the number of dependent variables, dependence techniques can be classified as those having a single dependent variable, several dependent variables, or even several dependent/independent relationships. Second, dependence techniques can be further classified as those with either metric (quantitative/numerical) or nonmetric (qualitative/categorical) dependent variables. If the analysis involves a single dependent variable that is metric, the appropriate technique is either multiple regression analysis or conjoint analysis. Conjoint analysis is a special case. It involves a dependence procedure that may treat the dependent

**Figure 1.6**  
Selecting a Multivariate Technique



\*Additional materials on this subject are available in the online resources at the text's websites.

(continued)



variable as either nonmetric or metric, depending on the type of data collected. In contrast, if the single dependent variable is nonmetric (categorical), then the appropriate techniques are multiple discriminant analysis and logistic regression.

When the research problem involves several dependent variables, four other techniques of analysis are appropriate. If the several dependent variables are metric, we must then look to the independent variables. If the independent variables are nonmetric, the technique of multivariate analysis of variance (MANOVA) should be selected. If the independent variables are metric, canonical correlation is appropriate. If the several dependent variables are nonmetric, then they can be transformed through **dummy variable** coding (0–1) and canonical analysis can again be used.<sup>1</sup> Finally, if a set of dependent/independent variable relationships is postulated, then structural equation modeling is appropriate.

A close relationship exists between the various dependence procedures, which can be viewed as a family of techniques. Figure 1.7 defines the various multivariate dependence techniques in terms of the nature and number of dependent and independent variables. As we can see, canonical correlation can be considered to be the general model upon which many other multivariate techniques are based, because it places the least restrictions on the type and number of variables in both the dependent and independent variates. As restrictions are placed on the variates, more precise conclusions can be reached based on the specific scale of data measurement employed. Thus, multivariate techniques range from the general method of canonical analysis to the specialized technique of structural equation modeling.

### FIGURE 1.7

#### The Relationship Between Multivariate Dependence Methods

<b>Canonical Correlation</b>	
$Y_1 + Y_2 + Y_3 + \dots + Y_n$ (metric, nonmetric)	$= X_1 + X_2 + X_3 + \dots + X_n$ (metric, nonmetric)
<b>Multivariate Analysis of Variance</b>	
$Y_1 + Y_2 + Y_3 + \dots + Y_n$ (metric)	$= X_1 + X_2 + X_3 + \dots + X_n$ (nonmetric)
<b>Analysis of Variance</b>	
$Y_1$ (metric)	$= X_1 + X_2 + X_3 + \dots + X_n$ (nonmetric)
<b>Multiple Discriminant Analysis</b>	
$Y_1$ (nonmetric)	$= X_1 + X_2 + X_3 + \dots + X_n$ (metric)
<b>Multiple Regression Analysis</b>	
$Y_1$ (metric)	$= X_1 + X_2 + X_3 + \dots + X_n$ (metric, nonmetric)
<b>Logistic Regression Analysis</b>	
$Y_1$ (binary nonmetric)	$= X_1 + X_2 + X_3 + \dots + X_n$ (metric, nonmetric)
<b>Conjoint Analysis</b>	
$Y_1$ (nonmetric, metric)	$= X_1 + X_2 + X_3 + \dots + X_n$ (nonmetric)
<b>Structural Equation Modeling/PLS</b>	
$Y_1 = X_{11} + X_{12} + X_{13} + \dots + X_{1n}$	
$Y_2 = X_{21} + X_{22} + X_{23} + \dots + X_{2n}$	
$Y_m = X_{m1} + X_{m2} + X_{m3} + \dots + X_{mn}$	

<sup>1</sup> Dummy variables are discussed in greater detail later. Briefly, dummy variable coding is a means of transforming nonmetric data into metric data. It involves the creation of so-called dummy variables, in which 1s and 0s are assigned to subjects, depending on whether they possess a characteristic in question. For example, if a subject is male, assign him a 0, if the subject is female, assign her a 1, or the reverse.

## INTERDEPENDENCE TECHNIQUES

Interdependence techniques are shown on the right side of Figure 1.7. Readers will recall that with interdependence techniques the variables cannot be classified as either dependent or independent. Instead, all the variables are analyzed simultaneously in an effort to find an underlying structure to the entire set of variables or subjects. If the structure of variables is to be analyzed, then exploratory factor analysis or confirmatory factor analysis is the appropriate technique. If cases or respondents are to be grouped to represent structure, then cluster analysis is selected. Finally, if the interest is in the structure of objects, the techniques of perceptual mapping should be applied. As with dependence techniques, the measurement properties of the techniques should be considered. Generally, exploratory factor analysis and cluster analysis are considered to be metric interdependence techniques. However, nonmetric data may be transformed through dummy variable coding for use with special forms of exploratory factor analysis and cluster analysis. Both metric and nonmetric approaches to perceptual mapping have been developed. If the interdependencies of objects measured by nonmetric data are to be analyzed, correspondence analysis is also an appropriate technique.

## Types of Multivariate Techniques

Multivariate analysis is an ever-expanding set of techniques for data analysis that encompasses a wide range of possible research situations as evidenced by the classification scheme just discussed. The more established as well as emerging techniques include the following:

### Interdependence Techniques

- Exploratory Factor Analysis: Principal components and common factor analysis
- Cluster analysis

### Dependence Techniques

- Multiple regression and multiple correlation
- Multivariate analysis of variance and covariance
- Multiple discriminant analysis
- Logistic regression
- Structural equation modeling and confirmatory factor analysis
- Partial least squares structural equation modeling and confirmatory composite analysis
- Canonical correlation analysis
- Conjoint analysis
- Perceptual mapping, also known as multidimensional scaling
- Correspondence analysis

Here we introduce each of the multivariate techniques and briefly define the technique and the objective for its application.

## EXPLORATORY FACTOR ANALYSIS: PRINCIPAL COMPONENTS AND COMMON FACTOR ANALYSIS

Exploratory factor analysis, including both principal component analysis and common factor analysis, is a statistical approach that can be used to analyze interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (factors). The objective is to find a way of condensing the information contained in a number of original variables into a smaller set of variates (factors) with a minimal loss of information. By providing an empirical estimate of the structure of the variables considered, exploratory factor analysis becomes an objective basis for creating summated scales.

A researcher can use factor analysis, for example, to better understand the relationships between customers' ratings of a fast-food restaurant. Assume you ask customers to rate the restaurant on the following six variables: food taste, food temperature, freshness, waiting time, cleanliness, and friendliness of employees. The analyst would like to combine these six variables into a smaller number. By analyzing the customer responses, the analyst might find that the

variables food taste, temperature, and freshness combine together to form a single factor of food quality, whereas the variables waiting time, cleanliness, and friendliness of employees combine to form another single factor, service quality.

## CLUSTER ANALYSIS

Cluster analysis is an analytical technique for developing meaningful subgroups of individuals or objects. Specifically, the objective is to classify a sample of entities (individuals or objects) into a small number of mutually exclusive groups based on the similarities among the entities. In cluster analysis, unlike discriminant analysis, the groups are not predefined. Instead, the technique is used to identify the groups.

Cluster analysis usually involves at least three steps. The first is the measurement of some form of similarity or association among the entities to determine how many groups really exist in the sample. The second step is the actual clustering process, whereby entities are partitioned into groups (clusters). The final step is to profile the persons or variables to determine their composition. Many times this profiling may be accomplished by applying discriminant analysis to the groups identified by the cluster technique.

As an example of cluster analysis, let's assume a restaurant owner wants to know whether customers are patronizing the restaurant for different reasons. Data could be collected on perceptions of pricing, food quality, and so forth. Cluster analysis could be used to determine whether some subgroups (clusters) are highly motivated by low prices versus those who are much less motivated to come to the restaurant based on price considerations.

## MULTIPLE REGRESSION

Multiple regression is the appropriate method of analysis when the research problem involves a single metric dependent variable presumed to be related to two or more metric independent variables. The objective of multiple regression analysis is to predict the changes in the dependent variable in response to changes in the independent variables. This objective is most often achieved through the statistical rule of least squares.

Whenever the researcher is interested in predicting the amount or size of the dependent variable, multiple regression is useful. For example, monthly expenditures on dining out (dependent variable) might be predicted from information regarding a family's income, its size, and the age of the head of household (independent variables). Similarly, the researcher might attempt to predict a company's sales from information on its expenditures for advertising, the number of salespeople, and the number of stores carrying its products.

## MULTIVARIATE ANALYSIS OF VARIANCE AND COVARIANCE

Multivariate analysis of variance (MANOVA) is a statistical technique that can be used to simultaneously explore the relationship between several categorical independent variables (usually referred to as **treatments**) and two or more metric dependent variables. As such, it represents an extension of **univariate analysis of variance (ANOVA)**. Multivariate analysis of covariance (MANCOVA) can be used in conjunction with MANOVA to remove (after the experiment) the effect of any uncontrolled metric independent variables (known as covariates) on the dependent variables. The procedure is similar to that involved in **bivariate partial correlation**, in which the effect of a third variable is removed from the correlation. MANOVA is useful when the researcher designs an experimental situation (manipulation of several nonmetric treatment variables) to test hypotheses concerning the variance in group responses on two or more metric dependent variables.

Assume a company wants to know if a humorous ad will be more effective with its customers than a non-humorous ad. It could ask its ad agency to develop two ads—one humorous and one non-humorous—and then show a group of customers the two ads. After seeing the ads, the customers would be asked to rate the company and its products on several dimensions, such as modern versus traditional or high quality versus low quality. MANOVA would be the technique to use to determine the extent of any statistical differences between the perceptions of customers who saw the humorous ad versus those who saw the non-humorous one.

## MULTIPLE DISCRIMINANT ANALYSIS

Multiple discriminant analysis (MDA) is the appropriate multivariate technique if the single dependent variable is dichotomous (e.g., male–female) or multichotomous (e.g., high–medium–low) and therefore nonmetric. As with

multiple regression, the independent variables are assumed to be metric. Discriminant analysis is applicable in situations in which the total sample can be divided into groups based on a nonmetric dependent variable characterizing several known classes. The primary objectives of multiple discriminant analysis are to understand group differences and to predict the likelihood that an entity (individual or object) will belong to a particular class or group based on several metric independent variables.

Discriminant analysis might be used to distinguish innovators from non-innovators according to their demographic and psychographic profiles. Other applications include distinguishing heavy product users from light users, males from females, national-brand buyers from private-label buyers, and good credit risks from poor credit risks. Even the Internal Revenue Service uses discriminant analysis to compare selected federal tax returns with a composite, hypothetical, normal taxpayer's return (at different income levels) to identify the most promising returns and areas for audit.

## LOGISTIC REGRESSION

Logistic regression models, often referred to as *logit analysis*, are a combination of multiple regression and multiple discriminant analysis. This technique is similar to multiple regression analysis in that one or more independent variables are used to predict a single dependent variable. What distinguishes a logistic regression model from multiple regression is that the dependent variable is nonmetric, as in discriminant analysis. The nonmetric scale of the dependent variable requires differences in the estimation method and assumptions about the type of underlying distribution, yet in most other facets it is quite similar to multiple regression. Thus, once the dependent variable is correctly specified and the appropriate estimation technique is employed, the basic factors considered in multiple regression are used here as well. Logistic regression models are distinguished from discriminant analysis primarily in that they only apply to binary dependent variables, accommodate all types of independent variables (metric and nonmetric) and do not require the assumption of multivariate normality. However, in many instances, particularly with more than two levels of the dependent variable, discriminant analysis is the more appropriate technique.

Assume financial advisors were trying to develop a means of selecting emerging firms for start-up investment. To assist in this task, they reviewed past records and placed firms into one of two classes: successful over a five-year period, and unsuccessful after five years. For each firm, they also had a wealth of financial and managerial data. They could then use a logistic regression model to identify those financial and managerial data that best differentiated between the successful and unsuccessful firms in order to select the best candidates for investment in the future.

## STRUCTURAL EQUATION MODELING AND CONFIRMATORY FACTOR ANALYSIS

Structural equation modeling (SEM) is a technique that allows separate relationships for each of a set of dependent variables. This method of SEM is based on an analysis of only common variance and begins with calculating the covariance matrix, and is often referred to as covariance-based SEM. In its simplest sense, structural equation modeling provides the appropriate and most efficient estimation technique for a series of separate multiple regression equations estimated simultaneously. It is characterized by two basic components: (1) the structural model and (2) the measurement model. The structural model is the *path* model, which relates independent to dependent variables. In such situations, theory, prior experience, or other guidelines enable the researcher to distinguish which independent variables predict each dependent variable. Models discussed previously that accommodate multiple dependent variables—multivariate analysis of variance and canonical correlation—are not applicable in this situation because they allow only a single relationship between dependent and independent variables.

The measurement model enables the researcher to use several variables (indicators) for a single independent or dependent variable. For example, the dependent variable might be a concept represented by a summated scale, such as self-esteem. In a confirmatory factor analysis (CFA) the researcher can assess the contribution of each scale item as well as incorporate how well the scale measures the concept (reliability). The scales are then integrated into the estimation of the relationships between dependent and independent variables in the structural model. This procedure is similar to performing an exploratory factor analysis (discussed in a later section) of the scale items and using the factor scores in the regression.

A study by management consultants identified several factors that affect worker satisfaction: supervisor support, work environment, and job performance. In addition to this relationship, they noted a separate relationship wherein supervisor support and work environment were unique predictors of job performance. Hence, they had two separate, but interrelated relationships. Supervisor support and the work environment not only affected worker satisfaction directly, but had possible indirect effects through the relationship with job performance, which was also a predictor of worker satisfaction. In attempting to assess these relationships, the consultants also developed multi-item scales for each construct (supervisor support, work environment, job performance, and worker satisfaction). SEM provides a means of not only assessing each of the relationships simultaneously rather than in separate analyses, but also incorporating the multi-item scales in the analysis to account for measurement error associated with each of the scales.

## PARTIAL LEAST SQUARES STRUCTURAL EQUATION MODELING

An alternative approach to structural equation modeling is partial least squares structural equation modeling (PLS-SEM), often referred to as variance-based SEM. This method of SEM is based on an analysis of total variance and also includes both a measurement model and a structural model. Theory and prior knowledge or other guidelines enable the researcher to distinguish which independent variables predict each dependent variable. The initial step in applying this method examines the measurement model and is referred to as confirmatory composite analysis. As with CFA, in this step the researcher also identifies the contribution of each measured variable to its construct as well as evaluating the reliability and validity of the measurement models. After the measurement models are determined to be valid and reliable, the analyst examines the structural model. The focus of variance-based SEM is primarily on prediction and explanation of the relationships, whereas with covariance-based SEM the focus is on confirmation of well-established theory.

## CANONICAL CORRELATION

Canonical correlation analysis can be viewed as a logical extension of multiple regression analysis. Recall that multiple regression analysis involves a single metric dependent variable and several metric independent variables. With canonical analysis the objective is to correlate simultaneously several metric dependent variables and several metric independent variables. Whereas multiple regression involves a single dependent variable, canonical correlation involves multiple dependent variables. The underlying principle is to develop a linear combination of each set of variables (both independent and dependent) in a manner that maximizes the correlation between the two sets. Stated in a different manner, the procedure involves obtaining a set of weights for the dependent and independent variables that provides the maximum simple correlation between the set of dependent variables and the set of independent variables.

Assume a company conducts a study that collects information on its service quality based on answers to 50 metrically measured questions. The study uses questions from published service quality research and includes benchmarking information on perceptions of the service quality of “world-class companies” as well as the company for which the research is being conducted. Canonical correlation could be used to compare the perceptions of the world-class companies on the 50 questions with the perceptions of the company. The research could then conclude whether the perceptions of the company are correlated with those of world-class companies. The technique would provide information on the overall correlation of perceptions as well as the correlation between each of the 50 questions.

## CONJOINT ANALYSIS

Conjoint analysis is a dependence technique that brings new sophistication to the evaluation of objects, such as new products, services, or ideas. The most direct application is in new product or service development, allowing for the evaluation of complex products while maintaining a realistic decision context for the respondent. The market researcher is able to assess the importance of attributes as well as the levels of each attribute while consumers evaluate only a few product profiles, which are combinations of product levels.

Assume a product concept has three attributes (price, quality, and color), each at three possible levels (e.g., red, yellow, and blue as the three levels of color). Instead of having to evaluate all 27 ( $3 \times 3 \times 3$ ) possible combinations, a subset (9 or more) can be evaluated for their attractiveness to consumers, and the researcher knows not only how important each attribute is but also the importance of each level (e.g., the attractiveness of red versus yellow versus blue). Moreover, when the consumer evaluations are completed, the results of conjoint analysis can also be used in product design simulators, which show customer acceptance for any number of product formulations and aid in the design of the optimal product.

### PERCEPTUAL MAPPING

In perceptual mapping (also known as *multidimensional scaling*), the objective is to transform consumer judgments of similarity or preference (e.g., preference for stores or brands) into distances represented in multidimensional space. If objects A and B are judged by respondents as being the most similar compared with all other possible pairs of objects, perceptual mapping techniques will position objects A and B in such a way that the distance between them in multidimensional space is smaller than the distance between any other pairs of objects. The resulting perceptual maps show the relative positioning of all objects, but additional analyses are needed to describe or assess which attributes predict the position of each object.

As an example of perceptual mapping, let's assume an owner of a Burger King franchise wants to know whether the strongest competitor is McDonald's or Wendy's. A sample of customers is given a survey and asked to rate the pairs of restaurants from most similar to least similar. The results show that the Burger King is most similar to Wendy's, so the owners know that the strongest competitor is the Wendy's restaurant because it is thought to be the most similar. Follow-up analysis can identify what attributes influence perceptions of similarity or dissimilarity.

### CORRESPONDENCE ANALYSIS

Correspondence analysis is a recently developed interdependence technique that facilitates the perceptual mapping of objects (e.g., products, persons) on a set of nonmetric attributes. Researchers are constantly faced with the need to "quantify the qualitative data" found in nominal variables. Correspondence analysis differs from the interdependence techniques discussed earlier in its ability to accommodate both nonmetric data and nonlinear relationships.

In its most basic form, correspondence analysis employs a contingency table, which is the cross-tabulation of two categorical variables. It then transforms the nonmetric data to a metric level and performs dimensional reduction (similar to exploratory factor analysis) and perceptual mapping. Correspondence analysis provides a multivariate representation of interdependence for nonmetric data that is not possible with other methods.

As an example, respondents' brand preferences can be cross-tabulated on demographic variables (e.g., gender, income categories, occupation) by indicating how many people preferring each brand fall into each category of the demographic variables. Through correspondence analysis, the association, or "correspondence," of brands and the distinguishing characteristics of those preferring each brand are then shown in a two- or three-dimensional map of both brands and respondent characteristics. Brands perceived as similar are located close to one another. Likewise, the most distinguishing characteristics of respondents preferring each brand are also determined by the proximity of the demographic variable categories to the brand's position.

## Guidelines for Multivariate Analyses and Interpretation

As demonstrated throughout this chapter, multivariate analyses' diverse character leads to quite powerful analytical and predictive capabilities. This power becomes especially tempting when the researcher is unsure of the most appropriate analysis design and relies instead on the multivariate technique as a substitute for the necessary conceptual development. And even when applied correctly, the strengths of accommodating multiple variables and relationships create substantial complexity in the results and their interpretation.

Faced with this complexity, we caution the researcher to proceed only when the requisite conceptual foundation to support the selected technique has been developed. We have already discussed several issues particularly applicable to multivariate analyses, and although no single “answer” exists, we find that analysis and interpretation of any multivariate problem can be aided by following a set of general guidelines. By no means an exhaustive list of considerations, these guidelines represent more of a “philosophy of multivariate analysis” that has served us well. The following sections discuss these points in no particular order and with equal emphasis on all.

### ESTABLISH PRACTICAL SIGNIFICANCE AS WELL AS STATISTICAL SIGNIFICANCE

The strength of multivariate analysis is its seemingly magical ability of sorting through a myriad number of possible alternatives and finding those with statistical significance. However, with this power must come caution. Many researchers become myopic in focusing solely on the achieved significance of the results without understanding their interpretations, good or bad. A researcher must instead look not only at the statistical significance of the results but also at their **practical significance**. Practical significance asks the question, “So what?” For any managerial application, the results must offer a demonstrable effect that justifies action. In academic settings, research is becoming more focused not only on the statistically significant results but also on their substantive and theoretical implications, which are many times drawn from their practical significance.

For example, a regression analysis is undertaken to predict repurchase intentions, measured as the probability between 0 and 100 that the customer will shop again with the firm. The study is conducted and the results come back significant at the .05 significance level. Executives rush to embrace the results and modify firm strategy accordingly. What goes unnoticed, however, is that even though the relationship was significant, the predictive ability was poor—so poor that the estimate of repurchase probability could vary by as much as  $\pm 20$  percent at the .05 significance level. The “statistically significant” relationship could thus have a range of error of 40 percentage points! A customer predicted to have a 50 percent chance of return could really have probabilities from 30 percent to 70 percent, representing unacceptable levels upon which to take action. Had researchers and managers probed the practical or managerial significance of the results, they would have concluded that the relationship still needed refinement before it could be relied upon to guide strategy in any substantive sense.

### RECOGNIZE THAT SAMPLE SIZE AFFECTS ALL RESULTS

The discussion of statistical power demonstrated the substantial impact sample size plays in achieving statistical significance, both in small and large sample sizes. For smaller samples, the sophistication and complexity of the multivariate technique may easily result in either (1) too little statistical power for the test to realistically identify significant results or (2) too easily “overfitting” the data such that the results are artificially good because they fit the sample yet provide no generalizability.

A similar impact also occurs for large sample sizes, which as discussed earlier, can make the statistical tests overly sensitive. Any time sample sizes exceed 400 respondents, the researcher should examine all significant results to ensure they have practical significance due to the increased statistical power from the sample size.

Sample sizes also affect the results when the analyses involve groups of respondents, such as discriminant analysis or MANOVA. Unequal sample sizes among groups influence the results and require additional interpretation or analysis. Thus, a researcher or user of multivariate techniques should always assess the results in light of the sample used in the analysis.

### KNOW YOUR DATA

Multivariate techniques, by their very nature, identify complex relationships that are difficult to represent simply. As a result, the tendency is to accept the results without the typical examination one undertakes in univariate and bivariate analyses (e.g., scatterplots of correlations and boxplots of mean comparisons). Such shortcuts can be a

prelude to disaster, however. Multivariate analyses require an even more rigorous examination of the data because the influence of outliers, violations of assumptions, and missing data can be compounded across several variables to create substantial effects.

A wide-ranging set of diagnostic techniques enables discovery of these multivariate relationships in ways quite similar to the univariate and bivariate methods. The multivariate researcher must take the time to utilize these diagnostic measures for a greater understanding of the data and the basic relationships that exist. With this understanding, the researcher grasps not only “the big picture,” but also knows where to look for alternative formulations of the original model that can aid in model fit, such as nonlinear and interactive relationships.

### STRIVE FOR MODEL PARSIMONY

Multivariate techniques are designed to accommodate multiple variables in the analysis. This feature, however, should not substitute for conceptual model development before the multivariate techniques are applied. Although it is always more important to avoid omitting a critical predictor variable, termed **specification error**, the researcher must also avoid inserting variables indiscriminately and letting the multivariate technique “sort out” the relevant variables for two fundamental reasons:

- 1 Irrelevant variables usually increase a technique’s ability to fit the sample data, but at the expense of overfitting the sample data and making the results less generalizable to the population. We address this issue in more detail when the concept of degrees of freedom is discussed in Chapter 5.
- 2 Even though irrelevant variables typically do not bias the estimates of the relevant variables, they can mask the true effects due to an increase in multicollinearity. Multicollinearity represents the degree to which any variable’s effect can be predicted or accounted for by the other variables in the analysis. As multicollinearity rises, the ability to define any variable’s effect is diminished. The addition of irrelevant or marginally significant variables can only increase the degree of multicollinearity, which makes interpretation of all variables more difficult.

Thus, including variables that are conceptually not relevant can lead to several potentially harmful effects, even if the additional variables do not directly bias the model results.

### LOOK AT YOUR ERRORS

Even with the statistical prowess of multivariate techniques, rarely do we achieve the best prediction in the first analysis. The researcher is then faced with the question, “Where does one go from here?” The best answer is to look at the errors in prediction, whether they are the residuals from regression analysis, the misclassification of observations in discriminant analysis, or outliers in cluster analysis. In each case, the researcher should use the errors in prediction not as a measure of failure or merely something to eliminate, but as a starting point for diagnosing the validity of the obtained results and an indication of the remaining unexplained relationships.

### SIMPLIFY YOUR MODELS BY SEPARATION

As you will see in many of the chapters, researchers have at their disposal a wide array of variable transformations (e.g., polynomial terms for non-linear effects) and variable combinations (e.g., interaction terms for moderation effects) that can portray just about any characteristic of a variable. Moreover, as research questions become more complex, more factors must be considered. Yet as we integrate more of these effects into a single model, the interactions among the effects may have unintended impacts on the results or the interpretations of the model. Thus, we encourage researchers to always try and simplify the model and reduce these interactions among effects. One common situation is when moderation effects are expected. A moderator is a variable, such as gender, which is expected to have different model parameters based on the moderator’s value (e.g., different regression coefficients for males versus females). This situation can be estimated in a single model through interaction terms as we will

discuss in later chapters, but doing that in a model with even a few independent variables becomes quite complicated and interpretation of the resulting parameter estimates even more difficult. A more direct approach would be to estimate separate models for males and then females so that each model can be easily interpreted. Another common situation is where the research question can be divided into a series of sub-questions, each representing its own dependence relationship. In these situations you may find that moving to a series of equations, such as structural equation modeling, helps separate the effects so that they are more easily estimated and understood. In all instances researchers should avoid building overly complex models with many variable transformations and other effects, and always attempt to separate the models in some fashion to simplify their basic model form and interpretation.

## VALIDATE YOUR RESULTS

The purpose of validation is to avoid the misleading results that can be obtained from **overfitting** – the estimation of model parameters that over-represent the characteristics of the sample at the expense of generalizability to the population at large. The model fit of any model is an optimistic assessment of how well the model will fit another sample of the population. We most often encounter overfitting when (a) the size of the data set is small, or (b) when the number of parameters in the model is large.

**Split-Sample Validation** The simplest form of validation is the split-sample approach, where the sample is divided into two sub-samples: one used for estimation (**estimation sample**) and the other (**holdout** or **validation sample**) for validation by “holding it out” of the estimation process and then seeing how well the estimated model fits this sample. Since the holdout sample was not used to estimate the model, it provides an independent means of validating the model. We demonstrate this approach in exploratory factor analysis (Chapter 3).

**Cross-validation** While it may seem simple to just divide the sample into two sub-samples, many times either limited sample size or other considerations make this impossible. For these situations cross-validation approaches have been developed. The basic principle of **cross-validation** is that the original sample is divided into a number of smaller sub-samples and the validation fit is the “average” fit across all of the sub-samples. Three of the more popular cross-validation approaches are  $k$ -fold, repeated random/resampling or leave-one-out/jackknife. The  $K$ -fold cross-validation randomly divides the original sample into  $k$  sub-samples ( $k$  folds) and a single sub-sample is held-out as the validation sample while the other  $k - 1$  sub-samples are used for estimation. The process is repeated  $K$  times, each time a different sub-sample being the validation sample. An advantage is a smaller validation sample (e.g., 10%) is possible, thus making it useful for smaller samples. The repeated random/resampling validation approach randomly draws a number of samples to act as validation samples [39]. Its advantage is that the validation sample can be of any size and is not dependent on the number of sub-samples as in the  $k$ -fold approach. Finally, the leave-one-out (or jackknife) approach is an extreme version of the  $k$ -fold approach in that each fold has a single observation (i.e., one observation at a time is left out). This is repeated until all observations have been used once as a validation sample. This approach is demonstrated in discriminant analysis (Chapter 7).

Whenever a multivariate technique is employed, the researcher must strive not only to estimate a significant model but to ensure that it is representative of the population as a whole. Remember, the objective is not to find the best “fit” just to the sample data but instead to develop a model that best describes the population as a whole.

## A Structured Approach to Multivariate Model Building

As we discuss the numerous multivariate techniques available to the researcher and the myriad set of issues involved in their application, it becomes apparent that the successful completion of a multivariate analysis involves more than just the selection of the correct method. Issues ranging from problem definition to a critical diagnosis of the results must be addressed. To aid the researcher or user in applying multivariate methods, a six-step approach to multivariate

analysis is presented. The intent is not to provide a rigid set of procedures to follow but, instead, to provide a series of guidelines that emphasize a model-building approach. This model-building approach focuses the analysis on a well-defined research plan, starting with a conceptual model detailing the relationships to be examined. Once defined in conceptual terms, the empirical issues can be addressed, including the selection of the specific multivariate technique and the implementation issues. After obtaining significant results, we focus on their interpretation, with special attention directed toward the variate. Finally, the diagnostic measures ensure that the model is valid not only for the sample data but that it is as generalizable as possible. The following discussion briefly describes each step in this approach.

This six-step model-building process provides a framework for developing, interpreting, and validating any multivariate analysis. Each researcher must develop criteria for “success” or “failure” at each stage, but the discussions of each technique provide guidelines whenever available. Emphasis on a model-building approach here, rather than just the specifics of each technique, should provide a broader base of model development, estimation, and interpretation that will improve the multivariate analyses of practitioner and academic alike.

## **STAGE 1: DEFINE THE RESEARCH PROBLEM, OBJECTIVES, AND MULTIVARIATE TECHNIQUE TO BE USED**

The starting point for any multivariate analysis is to define the research problem and analysis objectives in conceptual terms before specifying any variables or measures. The role of conceptual model development, or theory, cannot be overstated. No matter whether in academic or applied research, the researcher must first view the problem in conceptual terms by defining the concepts and identifying the fundamental relationships to be investigated.

A conceptual model need not be complex and detailed; instead, it can be just a simple representation of the relationships to be studied. If a dependence relationship is proposed as the research objective, the researcher needs to specify the dependent and independent concepts. For an application of an interdependence technique, the dimensions of structure or similarity should be specified. Note that a concept (an idea or topic), rather than a specific variable, is defined in both dependence and interdependence situations. This sequence minimizes the chance that relevant concepts will be omitted in the effort to develop measures and to define the specifics of the research design. Readers interested in conceptual model development should see Chapter 9.

With the objective and conceptual model specified, the researcher has only to choose the appropriate multivariate technique based on the measurement characteristics of the dependent and independent variables. Variables for each concept are specified prior to the study in its design, but may be respecified or even stated in a different form (e.g., transformations or creating dummy variables) after the data have been collected.

## **STAGE 2: DEVELOP THE ANALYSIS PLAN**

With the conceptual model established and the multivariate technique selected, attention turns to the implementation issues. The issues include general considerations such as minimum or desired sample sizes and allowable or required types of variables (metric versus nonmetric) and estimation methods.

## **STAGE 3: EVALUATE THE ASSUMPTIONS UNDERLYING THE MULTIVARIATE TECHNIQUE**

With data collected, the first task is not to estimate the multivariate model but to evaluate its underlying assumptions, both statistical and conceptual, that substantially affect their ability to represent multivariate relationships. For the techniques based on statistical inference, the assumptions of multivariate normality, linearity, independence of the error terms, and equality of variances must all be met. Assessing these assumptions is discussed in more detail in Chapter 2. Each technique also involves a series of conceptual assumptions dealing with such issues as model formulation and the types of relationships represented. Before any model estimation is attempted, the researcher must ensure that both statistical and conceptual assumptions are met.

## STAGE 4: ESTIMATE THE MULTIVARIATE MODEL AND ASSESS OVERALL MODEL FIT

With the assumptions satisfied, the analysis proceeds to the actual estimation of the multivariate model and an assessment of overall model fit. In the estimation process, the researcher may choose among options to meet specific characteristics of the data (e.g., use of covariates in MANOVA) or to maximize the fit to the data (e.g., rotation of factors or discriminant functions). After the model is estimated, the overall model fit is evaluated to ascertain whether it achieves acceptable levels on statistical criteria (e.g., level of significance), identifies the proposed relationships, and achieves practical significance. Many times, the model will be respecified in an attempt to achieve better levels of overall fit and/or explanation. In all cases, however, an acceptable model must be obtained before proceeding.

No matter what level of overall model fit is found, the researcher must also determine whether the results are unduly affected by any single or small set of observations that indicate the results may be unstable or not generalizable. Ill-fitting observations may be identified as outliers, influential observations, or other disparate results (e.g., single-member clusters or seriously misclassified cases in discriminant analysis).

## STAGE 5: INTERPRET THE VARIATE(S)

With an acceptable level of model fit, interpreting the variate(s) reveals the nature of the multivariate relationship. The interpretation of effects for individual variables is made by examining the estimated coefficients (weights) for each variable in the variate. Moreover, some techniques also estimate multiple variates that represent underlying dimensions of comparison or association. The interpretation may lead to additional respecifications of the variables and/or model formulation, wherein the model is re-estimated and then interpreted again. The objective is to identify empirical evidence of multivariate relationships in the sample data that can be generalized to the total population.

## STAGE 6: VALIDATE THE MULTIVARIATE MODEL

Before accepting the results, the researcher must subject them to one final set of diagnostic analyses that assess the degree of generalizability of the results by the available validation methods. The attempts to validate the model are directed toward demonstrating the generalizability of the results to the total population. These diagnostic analyses add little to the interpretation of the results but can be viewed as “insurance” that the results are the most descriptive of the data, yet generalizable to the population.

## A DECISION FLOWCHART

For each multivariate technique, the six-step approach to multivariate model building will be portrayed in a decision flowchart partitioned into two sections. The first section (Stages 1–3) deals with the issues addressed while preparing for actual model estimation (i.e., research objectives, research design considerations, and testing for assumptions). The second section of the decision flowchart (Stages 4–6) deals with the issues pertaining to model estimation, interpretation, and validation. The decision flowchart provides the researcher with a simplified but systematic method of applying the structural approach to multivariate model building to any application of the multivariate technique.

## Databases

To explain and illustrate each of the multivariate techniques more fully, we use hypothetical data sets throughout the book. The data sets are for HBAT Industries (HBAT), a manufacturer of paper products. Each data set is assumed to be based on surveys of HBAT customers completed on a secure website managed by an established marketing research company. The research company contacts purchasing managers and encourages them to participate. To do so, managers log onto the website and complete the survey. The data sets are supplemented by other information compiled and stored in HBAT’s data warehouse and accessible through its decision support system.

## PRIMARY DATABASE

The primary database, consisting of 100 observations on 18 separate variables, is based on a market segmentation study of HBAT customers. HBAT sells paper products to two market segments: the newsprint industry and the magazine industry. Also, paper products are sold to these market segments either directly to the customer or indirectly through a broker. Two types of information were collected in the surveys. The first type of information was perceptions of HBAT's performance on 13 attributes. These attributes, developed through focus groups, a pretest, and use in previous studies, are considered to be the most influential in the selection of suppliers in the paper industry. Respondents included purchasing managers of firms buying from HBAT, and they rated HBAT on each of the 13 attributes using a 0–10 scale, with 10 being "Excellent" and 0 being "Poor." The second type of information relates to purchase outcomes and business relationships (e.g., satisfaction with HBAT and whether the firm would consider a strategic alliance/partnership with HBAT). A third type of information is available from HBAT's data warehouse and includes information such as size of customer and length of purchase relationship.

By analyzing the data, HBAT can develop a better understanding of both the characteristics of its customers and the relationships between their perceptions of HBAT, and their actions toward HBAT (e.g., satisfaction and likelihood to recommend). From this understanding of its customers, HBAT will be in a good position to develop its marketing plan for next year. Brief descriptions of the database variables are provided in Figure 1.8, in which the variables are classified as either independent or dependent, and either metric or nonmetric. Also, a complete listing and electronic copy of the database are available in the online resources at the text's websites. A definition of each variable and an explanation of its coding are provided in the following sections.

**Figure 1.8**  
Description of Database Variables

Variable Description	Variable Type
<b>Data Warehouse Classification Variables</b>	
X <sub>1</sub> Customer Type	Nonmetric
X <sub>2</sub> Industry Type	Nonmetric
X <sub>3</sub> Firm Size	Nonmetric
X <sub>4</sub> Region	Nonmetric
X <sub>5</sub> Distribution System	Nonmetric
<b>Performance Perceptions Variables</b>	
X <sub>6</sub> Product Quality	Metric
X <sub>7</sub> E-Commerce Activities/Website	Metric
X <sub>8</sub> Technical Support	Metric
X <sub>9</sub> Complaint Resolution	Metric
X <sub>10</sub> Advertising	Metric
X <sub>11</sub> Product Line	Metric
X <sub>12</sub> Salesforce Image	Metric
X <sub>13</sub> Competitive Pricing	Metric
X <sub>14</sub> Warranty and Claims	Metric
X <sub>15</sub> New Products	Metric
X <sub>16</sub> Ordering and Billing	Metric
X <sub>17</sub> Price Flexibility	Metric
X <sub>18</sub> Delivery Speed	Metric
<b>Outcome/Relationship Measures</b>	
X <sub>19</sub> Satisfaction	Metric
X <sub>20</sub> Likelihood of Recommendation	Metric
X <sub>21</sub> Likelihood of Future Purchase	Metric
X <sub>22</sub> Current Purchase/Usage Level	Metric
X <sub>23</sub> Consider Strategic Alliance/Partnership in Future	Nonmetric

**Data Warehouse Classification Variables** As respondents were selected for the sample to be used by the marketing research firm, five variables also were extracted from HBAT's data warehouse to reflect the basic firm characteristics and their business relationship with HBAT. The five variables are as follows:

$X_1$	Customer Type	Length of time a particular customer has been buying from HBAT:
		1 = less than 1 year
		2 = between 1 and 5 years
		3 = longer than 5 years
$X_2$	Industry Type	Type of industry that purchases HBAT's paper products:
		0 = magazine industry
		1 = newsprint industry
$X_3$	Firm Size	Employee size:
		0 = small firm, fewer than 500 employees
		1 = large firm, 500 or more employees
$X_4$	Region	Customer location:
		0 = USA/North America
		1 = outside North America
$X_5$	Distribution System	How paper products are sold to customers:
		0 = sold indirectly through a broker
		1 = sold directly

**Perceptions of HBAT** Each respondent's perceptions of HBAT on a set of business functions were measured on a graphic rating scale, where a 10-centimeter line was drawn between the endpoints, labeled "Poor" and "Excellent," shown here.

As part of the survey, respondents indicated their perceptions by making a mark anywhere on the line. The location of the mark was electronically observed and the distance from 0 (in centimeters) was recorded in the database for that particular survey. The result was a scale ranging from 0 to 10, rounded to a single decimal place. The 13 HBAT attributes rated by each respondent were as follows:

$X_6$	Product Quality	Perceived level of quality of HBAT's paper products
$X_7$	E-Commerce Activities/ Website	Overall image of HBAT's website, especially user-friendliness
$X_8$	Technical Support	Extent to which technical support is offered to help solve product/service issues
$X_9$	Complaint Resolution	Extent to which any complaints are resolved in a timely and complete manner
$X_{10}$	Advertising	Perceptions of HBAT's advertising campaigns in all types of media
$X_{11}$	Product Line	Depth and breadth of HBAT's product line to meet customer needs
$X_{12}$	Salesforce Image	Overall image of HBAT's salesforce
$X_{13}$	Competitive Pricing	Extent to which HBAT offers competitive prices
$X_{14}$	Warranty and Claims	Extent to which HBAT stands behind its product/service warranties and claims

$X_{15}$	New Products	Extent to which HBAT develops and sells new products
$X_{16}$	Ordering and Billing	Perception that ordering and billing is handled efficiently and correctly
$X_{17}$	Price Flexibility	Perceived willingness of HBAT sales reps to negotiate price on purchases of paper products
$X_{18}$	Delivery Speed	Amount of time it takes to deliver the paper products once an order has been confirmed

**Purchase Outcomes** Five specific measures were obtained that reflected the outcomes of the respondent's purchase relationships with HBAT. These measures include the following:

$X_{19}$	Customer Satisfaction	Customer satisfaction with past purchases from HBAT, measured on a 10-point graphic rating scale
$X_{20}$	Likelihood of Recommending HBAT	Likelihood of recommending HBAT to other firms as a supplier of paper products, measured on a 10-point graphic rating scale
$X_{21}$	Likelihood of Future Purchases from HBAT	Likelihood of purchasing paper products from HBAT in the future, measured on a 10-point graphic rating scale
$X_{22}$	Percentage of Purchases from HBAT	Percentage of the responding firm's paper needs purchased from HBAT, measured on a 100-point percentage scale
$X_{23}$	Perception of Future Relationship with HBAT	Extent to which the customer/respondent perceives his or her firm would engage in strategic alliance/partnership with HBAT: 0 = Would not consider 1 = Yes, would consider strategic alliance or partnership

## OTHER DATABASES

Several other specialized databases are used in the text. First, Chapter 6 uses an expanded version of the HBAT database containing 200 respondents (HBAT200) that provides sufficient sample sizes for more complex MANOVA analyses. Chapter 2 uses a smaller database (HATMISS) to illustrate the handling of missing data. The SEM chapters (9, 10, 11, 12, 13) use different databases that meet the unique data requirements for those techniques. In each instance, the database is described more fully in those chapters. All of the databases used in the text are available in the online resources at the text's websites.

## Organization of the Remaining Chapters

The remaining chapters of the text are organized into five sections, each addressing a separate stage in performing a multivariate analysis.

### SECTION I: PREPARING FOR A MULTIVARIATE ANALYSIS

The initial section addresses issues that must be resolved before a multivariate analysis can be performed. Chapter 2 provides a number of methods for addressing a wide range of issues encountered in the dataset, including accommodating missing data, assurance of meeting the underlying statistical assumptions, identifying

outliers that might disproportionately affect the results and a set of variable transformations that can be used to represent unique variable characteristics. These methods enable the researcher to best prepare the data for further analysis.

## SECTION II: INTERDEPENDENCE TECHNIQUES

The two chapters in this section provide the researcher with the ability to “search for structure” in the data and provide some simplification of either the variables or the cases being analyzed. Chapter 3 covers exploratory factor analysis, a technique particularly suited to examining the relationships among variables and the opportunities for creating summated scales or other composite measures. These composite measures become especially useful as the number of variables increases since they provide a means of dimensional reduction. Chapter 4 covers cluster analysis, a technique that identifies groups of similar observations (i.e., clusters) based on a defined set of variables. In this way the large number of observations can be represented by a smaller set of homogeneous groups of observations. In both instances, identifying a small set of representative groups of observations or providing dimensional reduction for a large set of variables can simplify the models so the primary effects become more easily identified.

## SECTIONS III AND IV: DEPENDENCE TECHNIQUES

These sections cover four dependence techniques—multiple regression (Chapter 5), multivariate analysis of variance (Chapter 6), discriminant analysis (Chapter 7) and logistic regression (Chapter 8). Dependence techniques, as noted earlier, enable the researcher to assess the degree of relationship between dependent and independent variables. The dependence techniques vary in the type and character of the relationship as reflected in the measurement properties of the dependent and independent variables. Section III contains Chapter 5 (Multiple regression) and Chapter 6 (MANOVA), both techniques with metric dependent variables. Section IV contains Chapter 7 (Discriminant analysis) and Chapter 8 (Logistic regression) that examine techniques where the dependent variable is non-metric. Each technique is examined for its unique perspective on assessing a dependence relationship and its ability to address a particular type of research objective.

## SECTION V: MOVING BEYOND THE BASICS

This section introduces the researcher to a widely used advanced multivariate technique, structural equation modeling (SEM). Chapters 9 through 12 provide an overview of covariance-based structural equation modeling, focusing on the application of a decision process to SEM analyses, and then extends the SEM discussion to two of the most widely used applications: confirmatory factor analysis (CFA) and structural modeling. In Chapter 13 we provide an overview of variance-based structural equation modeling and the appropriate research applications for this form of SEM analysis when compared to covariance-based SEM.

## ONLINE RESOURCES: ADDITIONAL CHAPTERS

Readers familiar with past editions of the text may note that several techniques are not included in this edition – canonical correlation, conjoint analysis, perceptual mapping and correspondence analysis. Due to the addition of Chapter 13 covering PLS-SEM and the new material in other chapters, these chapters are now available as online resources. The complete chapters, including datasets, are available in the online resources at the text’s websites.

Multivariate data analysis is a powerful tool for researchers. Proper application of these techniques reveals relationships that otherwise would not be identified. This chapter introduces you to the major concepts and helps you to do the following:

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

Some confusion may arise about what multivariate analysis is because the term is not used consistently in the literature. Some researchers use *multivariate* simply to mean examining relationships between or among more than two variables. Others use the term only for problems in which all the multiple variables are assumed to have a multivariate normal distribution. In this book, we do not insist on a rigid definition of multivariate analysis. Instead, multivariate analysis includes both multivariable techniques and truly multivariate techniques, because we believe that knowledge of multivariable techniques is an essential first step in understanding multivariate analysis.

**Discuss the implications of Big Data, the emergence of algorithmic models and causal inference on multivariate analysis** Three emerging factors in the domain of analytics have created a very different and constantly changing environment for researchers over the past decade and will continue in the future. First, the emergence of Big Data has fundamentally influenced at least three aspects of analytics. These aspects include the abundance of data (both variables and observations), changes in the fundamental characteristics of data now available, and the increased desire for data-driven decisions within all types of organizations. As a result, researchers are confronted with a new landscape demanding some basic changes in how they research and conduct data analysis. Along with the emergence of Big Data there is more widespread use of algorithmic models where the emphasis is on prediction rather than explanation. While these models many times provide little insight into the “causes” of the outcomes, they nevertheless can address complicated problems not readily addressed with the more conventional data models approach. Finally, the abundance of data and the addition of entirely new sets of techniques has compelled researchers to strive for more causal inference in their analyses in order to avoid capturing spurious correlations that can result in invalid conclusions.

**Discuss the nature of measurement scales and their relationship to multivariate techniques.** Data analysis involves the identification and measurement of variation in a set of variables, either among themselves or between a dependent variable and one or more independent variables. The key word here is *measurement* because the researcher cannot identify variation unless it can be measured. Measurement is important in accurately representing the research concepts being studied and is instrumental in the selection of the appropriate multivariate method of analysis. Data can be classified into one of two categories—nonmetric (qualitative) and metric (quantitative)—based on the type of attributes or characteristics they represent. The researcher must define the measurement type for each variable. To the computer, the values are only numbers. Whether data are metric or nonmetric substantially affects what the data can represent, how it can be analyzed, and the appropriate multivariate techniques to use.

**Understand the nature of measurement error and its impact on multivariate analysis.** Use of multiple variables and reliance on their combination (the variate) in multivariate methods focuses attention on a complementary issue: measurement error. Measurement error is the degree to which the observed values are not representative of the “true” values. Measurement error has many sources, ranging from data entry errors to the imprecision of the measurement and the inability of respondents to accurately provide information. Thus, all variables used in multivariate techniques must be assumed to have some degree of measurement error. When variables with measurement error are used to compute correlations or means, the “true” effect is partially masked by the measurement error, causing the correlations to weaken and the means to be less precise.

**Examine the researcher options for managing the variate and dependence models** The analyst has two decisions in managing the variate: variable specification and variable selection. Variable specification involves a simple question: Are the variables used in their original form or is some form of dimensional reduction undertaken for simplification purposes and as a remedy for multicollinearity. If dimensional reduction is chosen the researcher can often control the process through exploratory factor analysis or use software options to perform the estimation procedure. For variable selection, the researcher again has the option to control the variables included in the analysis or allow algorithms to

find the “best” variables for model fit. In both instances we recommend that researcher judgment be combined with software control to explore options in the variate being tested. The researcher also has, in addition to the general linear model which is the foundation for most of the statistical techniques, an additional model form—the generalized linear model. This model formulation is comparable with the general linear model, but allows for the analysis of outcome variables that do not exhibit a normal distribution. With this model form the researcher can analyze outcomes such as percentages, binary outcomes and counts by transforming them to approximate the normal distribution.

**Understand the concept of statistical power and the options available to the researcher** Any time a statistical significance test is performed, the researcher should understand the concept of statistical power, which is the probability of finding an effect (i.e., non-zero parameter) when it is present in the data. While much focus is on the alpha level, it is only one of three factors that impact statistical power. The other two factors are effect size (the size of the effect being examined) and the sample size. Sample size plays a critically important role since increasing or decreasing the sample size impacts the significance level of an estimated parameter. Almost any parameter can be found to be significant with a large enough sample size and likewise an insufficient sample size may overlook substantive effects. Through the examination of statistical power the researcher ensures that the estimated significance level had an adequate probability of recovering a significant effect if it is present in the data.

**Determine which multivariate technique is appropriate for a specific research problem.** The multivariate techniques can be classified based on three judgments the researcher must make about the research objective and nature of the data: (1) Can the variables be divided into independent and dependent classifications based on some theory? (2) If they can, how many variables are treated as dependent in a single analysis? and (3) How are the variables, both dependent and independent, measured? Selection of the appropriate multivariate technique depends on the answers to these three questions.

**Define the specific techniques included in multivariate analysis.** Multivariate analysis is an ever-expanding set of techniques for data analysis that encompasses a wide range of possible research situations. Among the more established and emerging techniques are principal components and common factor analysis; multiple regression and multiple correlation; multiple discriminant analysis and logistic regression; canonical correlation analysis; multivariate analysis of variance and covariance; conjoint analysis; cluster analysis; perceptual mapping, also known as multidimensional scaling; correspondence analysis; and structural equation modeling (SEM), which includes confirmatory factor analysis.

**Discuss the guidelines for application and interpretation of multivariate analyses.** Multivariate analyses have powerful analytical and predictive capabilities. The strengths of accommodating multiple variables and relationships create substantial complexity in the results and their interpretation. Faced with this complexity, the researcher is cautioned to use multivariate methods only when the requisite conceptual foundation to support the selected technique has been developed. The following guidelines represent a “philosophy of multivariate analysis” that should be followed in their application:

Establish practical significance as well as statistical significance.

Recognize that sample size affects all results.

Know your data.

Strive for model parsimony.

Look at your errors.

Simplify your models by separation.

Validate your results.

**Understand the six-step approach to multivariate model building.** The six-step model-building process provides a framework for developing, interpreting, and validating any multivariate analysis.

Define the research problem, objectives, and multivariate technique to be used.

Develop the analysis plan.

Evaluate the assumptions.

Estimate the multivariate model and evaluate fit.

Interpret the variates.

Validate the multivariate model.

This chapter introduced the exciting, challenging topic of multivariate data analysis. The following chapters discuss each of the techniques in sufficient detail to enable the novice researcher to understand what a particular technique can achieve, when and how it should be applied, and how the results of its application are to be interpreted.

In your own words, define *multivariate analysis*.

Name the most important factors contributing to the increased application of techniques for multivariate data analysis in the last decade.

What implications does the emergence of Big Data have for researchers?

What are the differences between data models and algorithmic models? How do they impact the type of analysis performed?

What is meant by causal inference? How does it supplement the randomized controlled experiment?

What are the two facets of managing the variate? What role does each play in a typical multivariate analysis?

What types of research questions can the generalized linear model (GLZ) address more directly than the more traditional general linear model (GLM)?

Why is validation so important?

How does cross-validation work? What are some of the more popular types?

List and describe the multivariate data analysis techniques described in this chapter. Cite examples for which each technique is appropriate.

Explain why and how the various multivariate methods can be viewed as a family of techniques.

Why is knowledge of measurement scales important to an understanding of multivariate data analysis?

What are the differences between statistical and practical significance? Is one a prerequisite for the other?

What are the implications of low statistical power? How can the power be improved if it is deemed too low?

Detail the model-building approach to multivariate analysis, focusing on the major issues at each step.

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 Agresti, A. 2015. *Foundations of Linear and Generalized Linear Models*. New York: Wiley.
- 2 Bearden, William O., and Richard G. Netemeyer. 1999. *Handbook of Marketing Scales, Multi-Item Measures for Marketing and Consumer Behavior*, 2nd edn. Thousand Oaks, CA: Sage.
- 3 Bender, E. 2015. Big Data in Biomedicine. *Nature* 527(7576): S1–S22.
- 4 BMDP Statistical Software, Inc. 1991. *SOLO Power Analysis*. Los Angeles.
- 5 Bollen, K. A., and J. Pearl. 2013. Eight Myths About Causality and Structural Equation Models. In *Handbook of Causal Analysis for Social Research*, pp. 301–28. Dordrecht: Springer.
- 6 Boyd, D., and K. Crawford. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication and Society*, 15: 662–79.
- 7 Breiman, L. 2001. Statistical Modeling: The Two Cultures (With Comments and a Rejoinder by the Author). *Statistical Science* 16: 199–231.
- 8 Brent, Edward E., Edward J. Mirielli, and Alan Thompson. 1993. *Ex-Sample™ An Expert System to Assist in Determining Sample Size, Version 3.0*. Columbia, MO: Idea Works.
- 9 Brunner, Gordon C., Karen E. James, and Paul J. Hensel. 2001. *Marketing Scales Handbook, Vol. 3, A Compilation of Multi-Item Measures*. Chicago: American Marketing Association.
- 10 Bullock, Heather E., Lisa L. Harlow and Stanley A. Mulaik. 1994. Causation Issues in Structural Equation Modeling Research. *Structural Equation Modeling: A Multidisciplinary Journal* 1: 253–67.
- 11 Calude, C. S., and G. Longo. 2017. The Deluge of Spurious Correlations in Big Data. *Foundations of Science* 22: 595–612.
- 12 Campbell, K. T., and D. L. Taylor. 1996. Canonical Correlation Analysis as a General Linear Model: A Heuristic Lesson for Teachers and Students. *Journal of Experimental Education* 64: 157–71.
- 13 Cave, Andrew. 2017. What Will We Do When The World's Data Hits 163 Zettabytes In 2025? *Forbes* (13 April).

- 14 Chen, H., R. H. Chiang, and V. C. Storey. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36: 1165–88.
- 15 Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Publishing.
- 16 Davenport, Thomas H., and D. J. Patil. 2012. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review* 90(October): 70–76.
- 17 Einav, L., and J. Levin. 2014. Economics in the Age of Big Data. *Science* 346(6210): 1243089.
- 18 Fan, J., F. Han, and H. Liu. 2014. Challenges of Big Data Analysis. *National Science Review* 1: 293–314.
- 19 Fang B., and P. Zhang. 2016. Big Data in Finance. In S. Yu, and S. Guo (eds.), *Big Data Concepts, Theories, and Applications*. Berlin: Springer.
- 20 Franke, Beate, Jean François Plante, Ribana Roscher, Enshiun Annie Lee, Cathal Smyth, Armin Hatifi, Fuqi Chen et al. 2016. Statistical Inference, Learning and Models in Big Data. *International Statistical Review* 84: 371–89.
- 21 Gandomi, A., and M. Haider. 2015. Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management* 35: 137–44.
- 22 George, G., M. R. Haas, and A. Pentland. 2014. Big Data and Management: From the Editors. *Academy of Management Journal* 57: 321–26.
- 23 Gill, J. 2000. *Generalized Linear Models: A Unified Approach*. Sage University Papers Series on Quantitative Applications in the Social Sciences, No. 134. Thousand Oaks, CA: Sage.
- 24 Glymour, C. 2004. The Automation of Discovery. *Daedalus* 133: 69–77.
- 25 Gow, I. D., D. F. Larcker, and P. C. Reiss. 2016. Causal Inference in Accounting Research. *Journal of Accounting Research* 54: 477–523.
- 26 Grimmer, J. 2015. We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics* 48: 80–3.
- 27 Günther, W. A., M. H. R. Mehrizi, M. Huysman, and F. Feldberg. 2017. Debating Big Data: A Literature Review on Realizing Value From Big Data. *Journal of Strategic Information Systems* 20: 191–209.
- 28 Harford, T. 2014. Big Data: Are We Making a Big Mistake? *Financial Times Magazine*. 28 March.
- 29 Hutcheson, G., and N. Sofroniou. 1999. *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*. Thousand Oaks, CA: Sage.
- 30 Keele, L. 2015. The Statistics of Causal Inference: A View From Political Methodology. *Political Analysis* 23: 313–35.
- 31 Landhuis, E. 2017. Neuroscience: Big Brain, Big Data. *Nature* 541(7638): 559–61.
- 32 Lazer, David, and Jason Radford. 2017. Data ex Machina: Introduction to Big Data. *Annual Review of Sociology* 43: 19–39.
- 33 Marx, V. 2013. Biology: The Big Challenges of Big Data. *Nature* 498(7453): 255–60.
- 34 McAfee, A., E. Brynjolfsson, and T. H. Davenport. 2012. Big Data: The Management Revolution. *Harvard Business Review* 90: 60–8.
- 35 McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd edn. New York: Chapman and Hall.
- 36 Mercer, A. W., F. Kreuter, S. Keeter, and E. A. Stuart. 2017. Theory and Practice in Nonprobability Surveys: Parallels Between Causal Inference and Survey Inference. *Public Opinion Quarterly* 81(S1): 250–71.
- 37 Mooij, J. M., J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. 2016. Distinguishing Cause From Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research* 17: 1103–204.
- 38 Mooij, J. M., J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. 2016. Distinguishing Cause From Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research* 17: 1103–204.
- 39 Mooney, Christopher Z., and Robert D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Thousand Oaks, CA: Sage.
- 40 Muller, K. E. 1982. Understanding Canonical Correlation Through the General Linear Model and Principal Components. *American Statistician* 36: 342–54.
- 41 Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society A* 135: 370–84.
- 42 Pearl, J. 2000. *Causality*. New York: Cambridge University Press.
- 43 Pearl, J. 2009. Causal Inference in Statistics: An Overview. *Statistics Surveys* 3: 96–146.
- 44 Pearl, J. 2014. Interpretation and Identification of Causal Mediation. *Psychological Methods* 19: 459–81.
- 45 Peters, Tom. 1988. *Thriving on Chaos*. New York: Harper and Row.
- 46 Sivarajah, Uthayasankar, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. 2017. Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Business Research* 70: 263–86.
- 47 Stuart, E. A., and S. Naeger. 2017. Introduction to Causal Inference Approaches. *Methods in Health Services Research* 1–13.
- 48 Sullivan, John L., and Stanley Feldman. 1979. *Multiple Indicators: An Introduction*. Thousand Oaks, CA: Sage.
- 49 Thompson, B. 2015. The Case for Using the General Linear Model as a Unifying Conceptual Framework for Teaching Statistics and Psychometric Theory. *Journal of Methods and Measurement in the Social Sciences* 6: 30–41.
- 50 Viktor, Mayer-Schönberger, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.
- 51 Wedel, M., and P. K. Kannan. 2016. Marketing Analytics for Data-Rich Environments. *Journal of Marketing* 80:97–121.
- 52 Wiedermann, W., and A. von Eye. 2015. Direction of Effects in Mediation Analysis. *Psychological Methods* 20: 221–44.

# Preparing for multivariate analysis

## Examining Your Data

## SECTION I

### OVERVIEW

**S**ection I provides a set of tools and analyses that help to prepare the researcher for the increased complexity of a multivariate analysis. The prudent researcher appreciates the need for a higher level of understanding of the data, both in statistical and conceptual terms. Although the multivariate techniques discussed in this text present the researcher with a powerful set of analytical tools, they also pose the risk of further separating the researcher from a solid understanding of the data and leading to the misplaced notions that the analyses present a “quick and easy” means of identifying relationships. As the researcher relies more heavily on these techniques to find the answer and less on a conceptual basis and understanding of the fundamental properties of the data, the risk increases for serious problems in the misapplication of techniques, violation of statistical properties, or the inappropriate inference and interpretation of the results. These risks can never be totally eliminated, but the tools and analyses discussed in this section will improve the researcher’s ability to recognize many of these problems as they occur and to apply the appropriate remedy.

## CHAPTER IN SECTION I

This section consists of Chapter 2, Examining Your Data, which covers the topics of accommodating missing data, meeting the underlying statistical assumptions, identifying outliers that might disproportionately affect the results and useful data transformations. These analyses provide simple empirical assessments that detail the critical statistical properties of the data. The objective of these data examination tasks is as much to reveal what is not apparent as it is to portray the actual data, because the “hidden” effects are easily overlooked. For example, the biases introduced by nonrandom missing data will never be known unless explicitly identified and remedied by the methods discussed in a later section of this chapter. As a result, many times the tasks in this chapter should be looked upon as an “investment” by the researcher that provides some assurance that the “hidden” issues are discovered and not allowed to unduly influence the results.

# 2 Examining Your Data

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

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

## Chapter Preview

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

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

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

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

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

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

*Missing data* are a nuisance to researchers and primarily result from errors in data collection/data entry or from the omission of answers by respondents. Classifying missing data and the reasons underlying their presence are addressed through a series of steps that not only identify the impacts of the missing data, but that also provide remedies for dealing with it in the analysis. *Outliers*, or extreme responses, may unduly influence the outcome of any multivariate analysis. For this reason, methods to assess their impact are discussed. Finally, the *statistical assumptions* underlying most multivariate analyses are reviewed. Before applying any multivariate technique, the researcher must assess the fit of the sample data with the statistical assumptions underlying that multivariate technique. For example, researchers wishing to apply regression analysis (Chapter 5) would be particularly interested in assessing the assumptions of normality, homoscedasticity, independence of error, and linearity. Each of these issues should be addressed to some extent for each application of a multivariate technique.

In addition, this chapter introduces the researcher to several methods of data transformations. These range from incorporating nonmetric variables in applications that require metric variables through the creation of a special type of metric variable known as *dummy* variables to such techniques as binning and logarithmic transformations to represent specific types of relationships. While many forms of transformations are associated with meeting specific statistical properties, they also represent unique ways to modify the character of the data to enhance interpretation or applicability to a specific research question.

## Key Terms

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

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

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

**Boxplot** Method of representing the distribution of a variable. A box represents the major portion of the distribution, and the extensions—called whiskers—reach to the extreme points of the distribution. This method is useful in making comparisons of one or more metric variables across groups formed by a nonmetric variable.

**Cardinality** The number of distinct data values for a variable.

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

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

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

**Comparison group** See *reference category*.

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

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

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

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

**Data transformations** A variable may have an undesirable characteristic, such as non-normality, that detracts from its use in a multivariate technique. A transformation, such as taking the logarithm or square root of the variable, creates a transformed variable that is more suited to portraying the relationship. Transformations may be applied to either the dependent or independent variables, or both. The need and specific type of transformation may be based on theoretical reasons (e.g., transforming a known nonlinear relationship), empirical reasons (e.g., problems identified through graphical or statistical means) or for interpretation purposes (e.g., standardization).

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

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

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

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

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

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

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

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

**Heteroscedasticity** See *homoscedasticity*.

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

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

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

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

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

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

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

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

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

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

**Listwise deletion** See *complete case approach*.

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

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

**Missing at random (MAR)** Classification of *missing data* applicable when missing values of  $Y$  depend on  $X$ , but not on  $Y$ . When missing data are MAR, observed data for  $Y$  are a random sample of the  $Y$  values, but the missing values of  $Y$  are related to some other observed variable ( $X$ ) in the sample. For example, assume two groups based on gender have different levels of missing data between male and female. The data is MAR if the data is missing at random within each group, but the levels of missing data depend on the gender.

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

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

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

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

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

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

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

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

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

**Outlier** An observation that is substantially different from the other observations (i.e., has an extreme value) on one or more characteristics (variables). At issue is its representativeness of the population.

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

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

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

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

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

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

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

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

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

## Introduction

The tasks involved in examining your data may seem mundane and inconsequential, but they are an essential part of any multivariate analysis. Multivariate techniques place tremendous analytical power in the researcher's hands. But they also place a greater burden on the researcher to ensure that the statistical and theoretical underpinnings on which they are based also are supported. By examining the data before the application of any multivariate technique, the researcher gains several critical insights into the characteristics of the data:

- First and foremost, the researcher attains a *basic understanding of the data and relationships between variables*. Multivariate techniques place greater demands on the researcher to understand, interpret, and articulate results based on relationships that are more complex than encountered before. A thorough knowledge of the variable interrelationships can aid immeasurably in the specification and refinement of the multivariate model as well as provide a reasoned perspective for interpretation of the results.
- Second, the researcher ensures that the *data underlying the analysis meet all of the requirements for a multivariate analysis*. Multivariate techniques demand much more from the data in terms of larger datasets and more complex assumptions than encountered with univariate analyses. Missing data, outliers, and the statistical characteristics of the data are all much more difficult to assess in a multivariate context. Thus, the analytical sophistication needed to ensure that these requirements are met forces the researcher to use a series of data examination techniques that are as complex as the multivariate techniques themselves.

Both novice and experienced researchers may be tempted to skim or even skip this chapter to spend more time in gaining knowledge of a multivariate technique(s). The time, effort, and resources devoted to the data examination process may seem almost wasted because many times no corrective action is warranted. The researcher should instead view these techniques as "*investments in multivariate insurance*" that ensure the results obtained from the multivariate analysis are truly valid and accurate. Without such an "investment" it is quite easy, for example, for several unidentified outliers to skew the results, for missing data to introduce a bias in the correlations between variables, or for non-normal variables to invalidate the results. And yet the most troubling aspect of these problems is that they are "hidden," because in most instances the multivariate techniques will go ahead and provide results. Only if the researcher has made the "investment" will the potential for catastrophic problems be recognized and corrected *before* the analyses are performed. These problems can be avoided by following these analyses each and every time a multivariate technique is applied. These efforts will more than pay for themselves in the long run; the occurrence of one serious and possibly fatal problem will make a convert of any researcher. We encourage you to embrace these techniques before problems that arise during analysis force you to do so.

## The Challenge of Big Data Research Efforts

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

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

## DATA MANAGEMENT

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

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

## DATA QUALITY

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

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

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

## SUMMARY

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

## Preliminary Examination of the Data

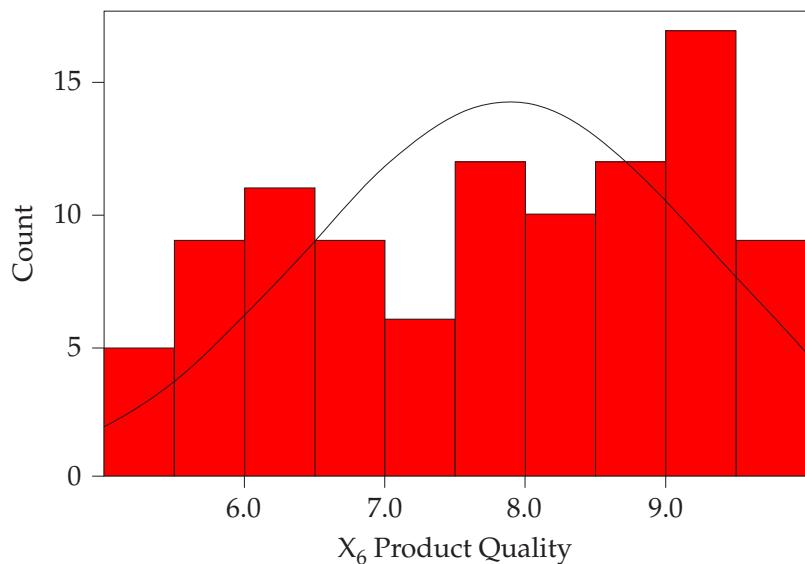
As discussed earlier, the use of multivariate techniques places an increased burden on the researcher to understand, evaluate, and interpret complex results. This complexity requires a thorough understanding of the basic characteristics of the underlying data and relationships. When univariate analyses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

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

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

## UNIVARIATE PROFILING: EXAMINING THE SHAPE OF THE DISTRIBUTION

The starting point for understanding the nature of any variable is to characterize the shape of its distribution. A number of statistical measures are discussed in a later section on normality, but many times the researcher can gain an adequate perspective of the variable through a **histogram**. A histogram is a graphical representation of a single



**Figure 2.1**  
Graphical Representation of Univariate Distribution

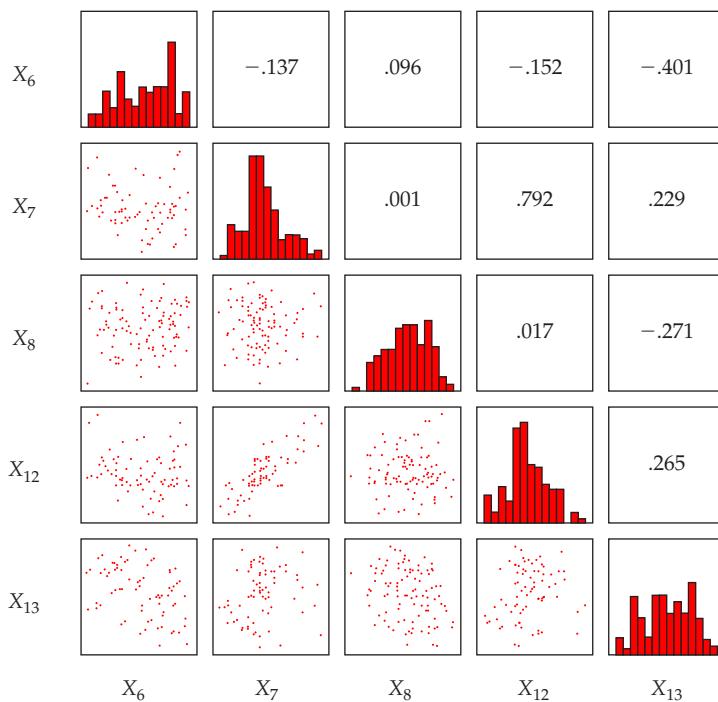
variable that represents the frequency of occurrences (data values) within data categories. The frequencies are plotted to examine the shape of the distribution of values. If the integer values ranged from 1 to 10, the researcher could construct a histogram by counting the number of responses for each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

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

## BIVARIATE PROFILING: EXAMINING THE RELATIONSHIP BETWEEN VARIABLES

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

Of the many types of scatterplots, one format particularly suited to multivariate techniques is the scatterplot matrix, in which the scatterplots are represented for all combinations of variables in the lower portion of the matrix. The diagonal contains histograms of the variables. Scatterplot matrices and individual scatterplots are now available



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

in all popular statistical programs. A variant of the scatterplot is discussed in the following section on outlier detection, where an ellipse representing a specified confidence interval for the bivariate normal distribution is superimposed to allow for outlier identification.

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

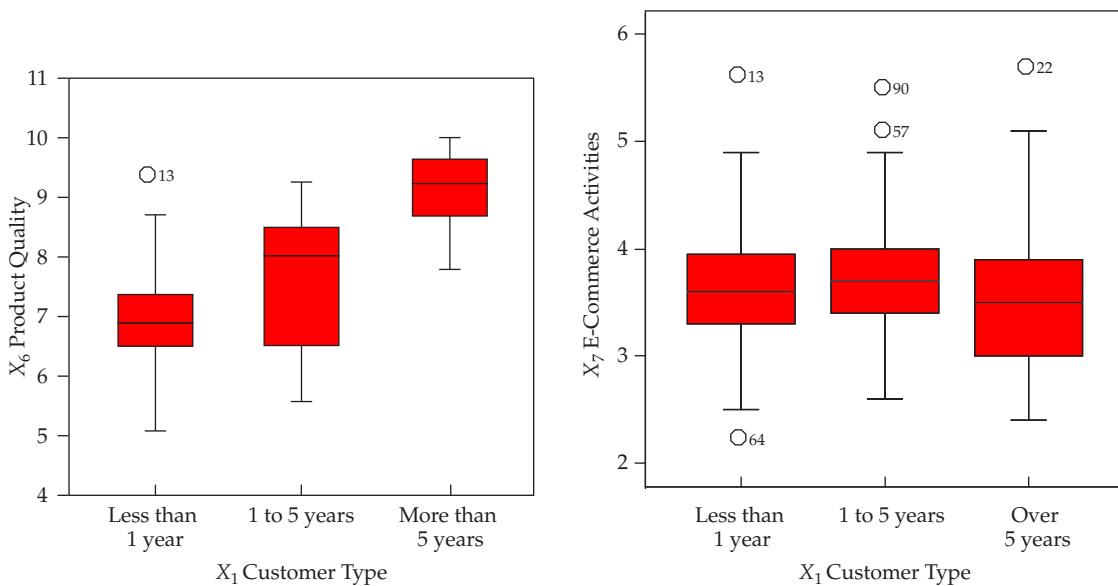
The scatterplot matrix provides a quick and simple method of not only assessing the strength and magnitude of any bivariate relationship, but also a means of identifying any nonlinear patterns that might be hidden if only the bivariate correlations, which are based on a linear relationship, are examined.

## BIVARIATE PROFILING: EXAMINING GROUP DIFFERENCES

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

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

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



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

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

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

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

## MULTIVARIATE PROFILES

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

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

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

Facial representations provide a potent graphical format but also give rise to a number of considerations that affect the assignment of variables to facial features, unintended perceptions, and the quantity of information that can actually be accommodated. Discussion of these issues is beyond the scope of this text, and interested readers are encouraged to review them before attempting to use these methods [87, 88].

## NEW MEASURES OF ASSOCIATION

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

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

## SUMMARY

The researcher can employ any of these methods when examining multivariate data to provide a format that is many times more insightful than just a review of the actual data values. Moreover, the multivariate methods enable the researcher to use a single graphical portrayal to represent a large number of variables, instead of using a large number of the univariate or bivariate methods to portray the same number of variables. And an expanded set of measures of association may assist in identifying previously undiscovered relationships, especially as datasets increase in size.

## Missing Data

**Missing data**, where valid values on one or more variables are not available for analysis, are a fact of life in multivariate analysis. In fact, rarely does the researcher avoid some form of missing data problem. The researcher's challenge is to address the issues raised by missing data that affect the generalizability of the results. To do so, the researcher's *primary concern is to identify the patterns and relationships underlying the missing data in order to maintain as close as possible the original distribution of values when any remedy is applied*. The extent of missing data is a secondary issue in most instances, affecting the type of remedy applied. These patterns and relationships are a result of a **missing data process**, which is any systematic event external to the respondent (such as data entry errors or data collection problems) or any action on the part of the respondent (such as refusal to answer) that leads to missing values. The need to focus on the reasons for missing data comes from the fact that the researcher must understand the processes leading to the missing data in order to select the appropriate course of action.

### THE IMPACT OF MISSING DATA

The effects of some missing data processes are known and directly accommodated in the research plan, as will be discussed later in this section. More often, the missing data processes, particularly those based on actions by the respondent (e.g., non-response to a question or set of questions), are rarely known beforehand. To identify any patterns in the missing data that would characterize the missing data process, the researcher asks such questions as (1) Are the missing data scattered randomly throughout the observations or are distinct patterns identifiable? and (2) How prevalent are the missing data? If distinct patterns are found and the extent of missing data is sufficient to warrant action, then it is assumed that some missing data process is in operation.

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

**Practical Impact** The *practical impact* of missing data is the reduction of the sample size available for analysis. For example, if remedies for missing data are not applied, any observation with missing data on any of the variables will be excluded from the analysis. In many multivariate analyses, particularly survey research applications, missing data may eliminate so many observations that what was an adequate sample is reduced to an inadequate sample. For example, it has been shown that if 10 percent of the data is randomly missing in a set of five variables, on average almost 60 percent of the cases will have at least one missing value [53]. Thus, when complete data are required, the sample is reduced to 40 percent of the original size. In such situations, the researcher must either gather additional observations or find a remedy for the missing data in the original sample.

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

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

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

## RECENT DEVELOPMENTS IN MISSING DATA ANALYSIS

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

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

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

## A SIMPLE EXAMPLE OF A MISSING DATA ANALYSIS

To illustrate the substantive and practical impacts of missing data, Figure 2.4 contains a simple example of missing data among 20 cases. As is typical of many datasets, particularly in survey research, the number of missing data varies widely among both cases and variables.

In this example, we can see that all of the variables ( $V_1$  to  $V_5$ ) have some missing data, with  $V_3$  missing more than one-half (55%) of all values. Three cases (3, 13, and 15) have more than 50 percent missing data and only five cases have complete data. Overall, 23 percent of the data values are missing.

**Practical Impact** From a *practical standpoint*, the missing data in this example can become quite problematic in terms of reducing the sample size. For example, if a multivariate analysis was performed that required complete data on all five variables, the sample would be reduced to only the five cases with no missing data (cases 1, 7, 8, 12, and 20). This sample size is too few for any type of analysis. Among the remedies for missing data that will be discussed in detail in later sections, an obvious option is the elimination of variables and/or cases. In our example, assuming that the conceptual foundations of the research are not altered substantially by the deletion of a variable, eliminating  $V_3$  is one approach to reducing the number of missing data. By just eliminating  $V_3$ , seven additional cases, for a total of 12, now have complete information. If the three cases (3, 13, 15) with exceptionally high numbers of missing data are also eliminated, the total number of missing data is now reduced to only five instances, or 7.4 percent of all values.

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

**Figure 2.4**  
Hypothetical Example of Missing Data

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

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

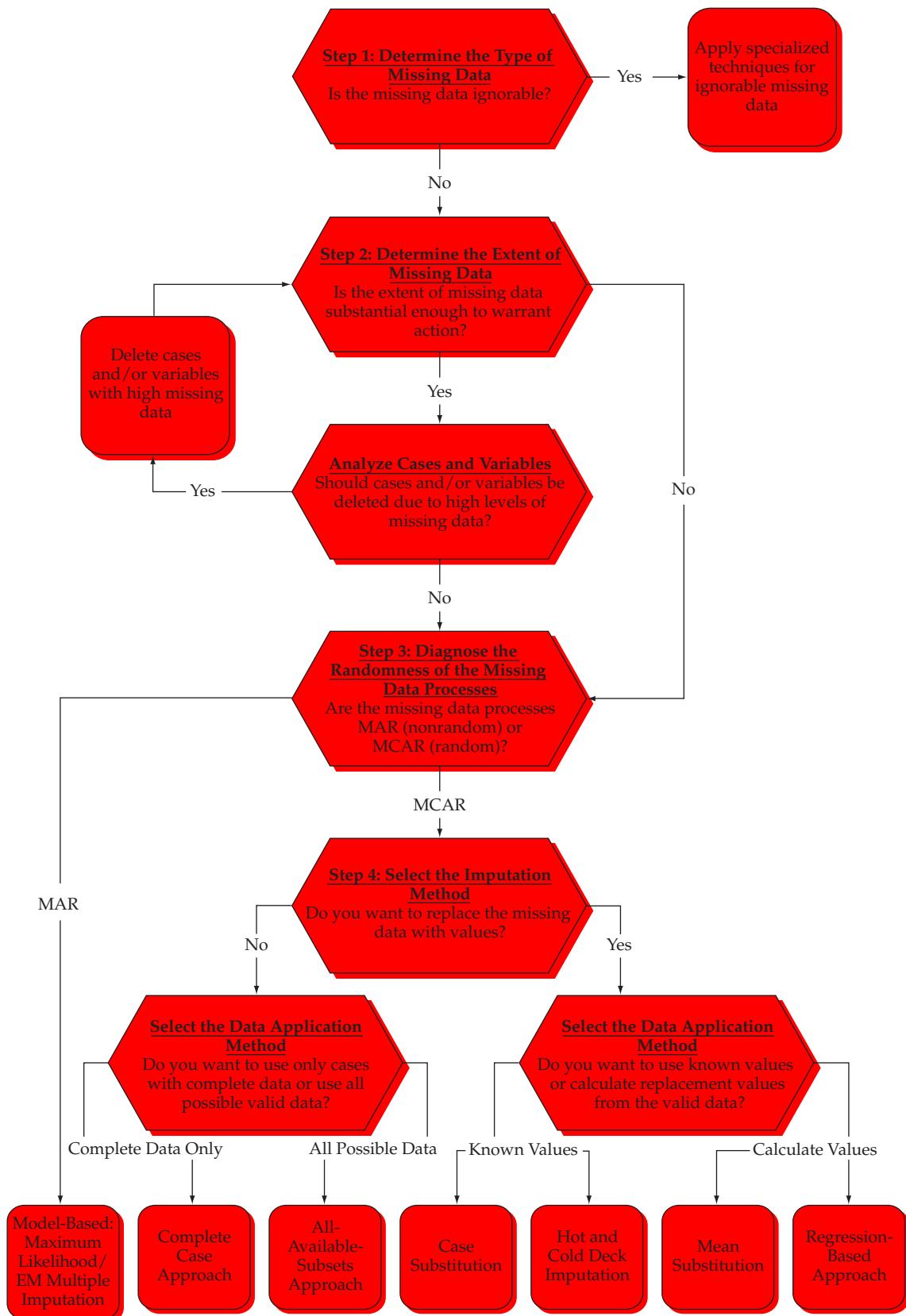
**Overall Impact** As we have seen in the example, finding a remedy for missing data (e.g., deleting cases or variables) can be a practical solution for missing data. Yet the researcher must guard against applying such remedies without diagnosis of the missing data processes. Avoiding the diagnosis may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

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

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

**Figure 2.5**

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



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

**IGNORABLE MISSING DATA** Many times missing data are expected and part of the research design. In these instances, the missing data are termed **ignorable missing data**, meaning that specific remedies for missing data are not needed because the allowances for missing data are inherent in the technique used [62, 83]. The justification for designating missing data as ignorable is that the missing data process is operating at random (i.e., the observed values are a random sample of the total set of values, observed and missing) or explicitly accommodated in the technique used. There are three instances in which a researcher most often encounters ignorable missing data.

**A Sample as Missing Data** The first example encountered in almost all surveys and most other datasets is the ignorable missing data process resulting from taking a sample of the population rather than gathering data from the entire population. In these instances, the missing data are those observations in a population that are not included when taking a sample. The purpose of multivariate techniques is to generalize from the sample observations to the entire population, which is really an attempt to overcome the missing data of observations not in the sample. The researcher makes these missing data ignorable by using probability sampling to select respondents. Probability sampling enables the researcher to specify that the missing data process leading to the omitted observations is random and that the missing data can be accounted for as sampling error in the statistical procedures. Thus, the missing data of the non-sampled observations are ignorable.

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

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

**Censored Data** A third type of ignorable missing data occurs when the data are censored. **Censored data** are observations not complete because of their stage in the missing data process. A typical example is an analysis of the causes of death. Respondents who are still living cannot provide complete information (i.e., cause or time of death) and are thus censored. Another interesting example of censored data is found in the attempt to estimate the heights of the U.S. general population based on the heights of armed services recruits (as cited in [62]). The data are censored because in certain years the armed services had height restrictions that varied in level and enforcement. Thus, the researchers face the task of estimating the heights of the entire population when it is known that certain individuals (i.e., all those below the height restrictions) are not included in the sample. In both instances the researcher’s knowledge of the missing data process allows for the use of specialized methods, such as event history analysis, to accommodate censored data [62].

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

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

**Known Processes** Many missing data processes are *known* to the researcher in that they can be identified due to procedural factors, such as errors in data entry that create invalid codes, disclosure restrictions (e.g., small counts in US Census data), failure to complete the entire questionnaire, or even the morbidity of the respondent. In these

situations, the researcher has little control over the missing data processes, but some remedies may be applicable if the missing data are found to be random.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## How Much Missing Data Is Too Much?

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

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

## Deletions Based on Missing Data

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

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

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

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

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

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

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

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

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

For example, assume that we know the gender of respondents (the  $X$  variable) and are asking about household income (the  $Y$  variable). We find that the missing data are random for both males and females but occur at a much

**Figure 2.6**  
Missing Data Processes: MCAR, MAR and MNAR

Complete Data		Missing Data Process for Y		
X	Y	MCAR:	MAR:	MNAR:
3	9	9	Missing	9
3	5	5	Missing	5
4	1	Missing	Missing	Missing
4	3	3	Missing	Missing
5	2	Missing	2	Missing
6	6	Missing	6	6
7	7	7	7	7
7	4	4	4	Missing
8	5	5	5	5
9	9	Missing	9	9

Characteristics of the Missing Data Process				
Pattern of missing values of Y	Random: Across all values of Y	Random: Across all values of Y	Nonrandom: Only lowest values of Y	
Relationship of X to missingness of Y	No Across all values of X	Yes Lowest values of X	No Across all values of X	

Adapted from [31].

higher frequency for males than females. Even though the missing data process is operating in a random manner within the gender variable, any remedy applied to the missing data will still reflect the missing data process because gender affects the ultimate distribution of the household income values.

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

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

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

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

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

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

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

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

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

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

**Little's MCAR Test** A second approach is an overall test of randomness that determines whether the missing data can be classified as MCAR [58]. This test analyzes the pattern of missing data on all variables and compares it with the pattern expected for a random missing data process. If no significant differences are found, the missing data can be classified as MCAR. If significant differences are found, however, the researcher must use the approaches described previously to identify the specific missing data processes that are nonrandom.

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

**Step 4: Select the Imputation Method** At this step of the process, the researcher must select the approach used for accommodating missing data in the analysis. This decision is based primarily on whether the missing data are MAR or MCAR, but in either case the researcher has several options for imputation [42, 62, 73, 78, 31, 37, 21]. **Imputation** is the process of estimating the missing value based on valid values of other variables and/or cases in the sample. The objective is to employ known relationships that can be identified in the valid values of the sample to assist in estimating the missing values. However, the researcher should carefully consider the use of imputation in each instance because of its potential impact on the analysis [27]:

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

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

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

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

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

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

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

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

Even though the all-available method maximizes the data utilized and overcomes the problem of missing data on a single variable eliminating a case from the entire analysis, several problems can arise. First, correlations may be calculated that are “out of range” and inconsistent with the other correlations in the correlation matrix [65]. Any correlation between  $X$  and  $Y$  is constrained by their correlation to a third variable  $Z$ , as shown in the following formula:

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

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

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

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

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

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

**Case Substitution** In this method, entire observations with missing data are replaced by choosing another non-sampled observation. A common example is to replace a sampled household that cannot be contacted or that has extensive missing data with another household not in the sample, preferably similar to the original observation. This method is most widely used to replace observations with complete missing data, although it can be used to replace observations with lesser amounts of missing data as well. At issue is the ability to obtain these additional observations not included in the original sample.

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

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

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

Although it has the appeal of using relationships already existing in the sample as the basis of prediction, this method also has several disadvantages. First, it reinforces the relationships already in the data. As the use of this method increases, the resulting data become more characteristic of the sample and less generalizable. Second, unless stochastic terms are added to the estimated values, the variance of the distribution is understated. Third, this method assumes that the variable with missing data has substantial correlations with the other variables. If these correlations are not sufficient to produce a meaningful estimate, then other methods, such as mean substitution, are preferable. Fourth, the sample must be large enough to allow for a sufficient number of observations to be used in making each prediction. Finally, the regression procedure is not constrained in the estimates it makes. Thus, the predicted values may not fall in the valid ranges for variables (e.g., a value of 11 may be predicted for a 10-point scale) and require some form of additional adjustment.

Even with all of these potential problems, the regression method of imputation holds promise in those instances for which moderate levels of widely scattered missing data are present and for which the relationships between variables are sufficiently established so that the researcher is confident that using this method will not affect the generalizability of the results.

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

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

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

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

A variation of this method employs maximum likelihood as well, but in an iterative process. The **EM** method [39, 31, 74] is a two-stage method (the E and M stages) in which the E stage makes the best possible estimates of the missing data and the M stage then makes estimates of the parameters (means, standard deviations, or correlations) assuming the missing data were replaced. The process continues going through the two stages until the change in the estimated values is negligible and they replace the missing data. One notable feature is that this method can produce an imputed dataset, although it has been shown to underestimate the standard errors in estimated models [31, 37].

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

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

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

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

- 2 *Estimate the model.* This second step is estimation of the desired analysis. One of the primary advantages of multiple imputation is that it can be used with almost all of the most widely used linear models – t tests, the general linear model (e.g., multiple regression and ANOVA/MANOVA) and the generalized linear model (e.g., logistic regression). Thus, the researcher can perform any of these methods as they would a single dataset with no missing data. Each imputed dataset is analyzed separately, comparable to what is performed in IBM SPSS (SPLIT FILES) or SAS (BY command) with an imputation variable defining the separate imputed datasets.
- 3 *Combining results from multiple imputed datasets.* The final step is to combine the results from the individual imputed datasets into a final set of “combined” results. Procedures are available (e.g., PROC MIANALYZE in SAS) for combining the parameter estimates so that the researcher now has the unbiased estimate of the parameters and the standard errors.

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

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

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

**Figure 2.7**

Comparison of Imputation Techniques for Missing Data

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

## Imputation of Missing Data Based On Extent of Missing Data

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

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

Given the many imputation methods available, the researcher should also strongly consider following a multiple imputation strategy if the MCAR methods are used, whereby a combination of several methods is used. In this approach, two or more methods of imputation are used to derive a composite estimate—usually the mean of the various estimates—for the missing value. The rationale is that the use of multiple approaches minimizes the specific concerns with any single method and the composite will be the best possible estimate. The choice of this approach is primarily based on the trade-off between the researcher's perception of the potential benefits versus the substantially higher effort required to make and combine the multiple estimates. The model-based methods, however, provide the best approaches to avoid biased estimates due to any underlying MAR missing data processes.

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

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

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

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

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

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

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

**Table 2.1** Summary Statistics of Missing Data for Original Sample

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

NA = Not applicable to nonmetric variables

### Summary of Cases

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

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

**Table 2.2 Patterns of Missing Data by Case**

Case	# Missing	% Missing	Missing Data Patterns													
			V <sub>1</sub>	V <sub>2</sub>	V <sub>3</sub>	V <sub>4</sub>	V <sub>5</sub>	V <sub>6</sub>	V <sub>7</sub>	V <sub>8</sub>	V <sub>9</sub>	V <sub>10</sub>	V <sub>11</sub>	V <sub>12</sub>	V <sub>13</sub>	V <sub>14</sub>
205	1	7.1		S												
202	2	14.3	S		S											
250	2	14.3	S		S											
255	2	14.3	S		S											
269	2	14.3	S		S											
238	1	7.1	S													
240	1	7.1	S													
253	1	7.1	S													
256	1	7.1	S													
259	1	7.1	S													
260	1	7.1	S													
228	2	14.3	S			S										
246	1	7.1				S										
225	2	14.3			S	S										
267	2	14.3			S	S										
222	2	14.3			S		S									
241	2	14.3			S		S									
229	1	7.1				S										
216	2	14.3	S			S										
218	2	14.3	S			S										
232	2	14.3	S	S												
248	2	14.3	S	S												
237	1	7.1	S													
249	1	7.1	S													
220	1	7.1	S													
213	2	14.3	S	S												
257	2	14.3	S	S												
203	2	14.3	S					S								
231	1	7.1						S								
219	2	14.3						S	S							
244	1	7.1						S								
227	2	14.3	S						S							
224	3	21.4	S	S					S							
268	1	7.1								S						
235	2	14.3						S			S					
204	3	21.4	S	S							S					
207	3	21.4	S	S							S					
221	3	21.4	S	S					S							
245	7	50.0	S	S		S		S	S	S		S		S	S	
233	7	50.0	S	S		S	S	S	S	S		S		S	S	
261	7	50.0	S	S				S	S	S	S		S			
210	7	50.0		S	S	S	S	S	S	S	S					
263	7	50.0	S		S	S	S	S	S	S	S					
214	7	50.0	S		S		S	S	S	S	S		S		S	

Note: Only cases with missing data are shown.

S = missing data.

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

**Table 2.3** Missing Data Patterns

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

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

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

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

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

**Step 3: Diagnosing the Randomness of the Missing Data Process** The next step is an empirical examination of the patterns of missing data to determine whether the missing data are distributed randomly across the cases and the variables. Hopefully the missing data will be judged MCAR, thus allowing a wider range of remedies in the imputation process. We will first employ a test of comparison between groups of missing and non-missing cases and then conduct an overall test for randomness.

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

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

#### Summary of Cases

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

The first test for assessing randomness is to compare the observations with and without missing data for each variable on the other variables. For example, the observations with missing data on  $V_2$  are placed in one group and those observations with valid responses for  $V_2$  are placed in another group. Then, these two groups are compared to identify any differences on the remaining metric variables ( $V_3$  through  $V_9$ ). Once comparisons have been made on all of the variables, new groups are formed based on the missing data for the next variable ( $V_3$ ) and the comparisons are performed again on the remaining variables. This process continues until each variable ( $V_2$  through  $V_9$ ; remember  $V_1$  has been excluded) has been examined for any differences. The objective is to identify any systematic missing data process that would be reflected in patterns of significant differences.

Table 2.5 contains the results for this analysis of the 64 remaining observations. The only noticeable pattern of significant  $t$  values occurs for  $V_2$ , for which three of the eight comparisons ( $V_4$ ,  $V_5$ , and  $V_6$ ) found significant differences between the two groups. Moreover, only one other instance (groups formed on  $V_4$  and compared on  $V_2$ ) showed a significant difference. This analysis indicates that although significant differences can be found due to the missing data on one variable ( $V_2$ ), the effects are limited to only this variable, making it of marginal concern. If later tests of randomness indicate a nonrandom pattern of missing data, these results would then provide a starting point for possible remedies.

The final test is an overall test of the missing data for being missing completely at random (MCAR). The test makes a comparison of the actual pattern of missing data with what would be expected if the missing data were totally randomly distributed. The MCAR missing data process is indicated by a *nonsignificant* statistical level (e.g., greater than .05), showing that the observed pattern *does not* differ from a random pattern. This test is performed in the Missing Value Analysis module of SPSS as well as several other software packages dealing with missing value analysis.

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

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

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

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

**Table 2.5** Assessing the Randomness of Missing Data Through Group Comparisons of Observations with Missing Versus Valid Data

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

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

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

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

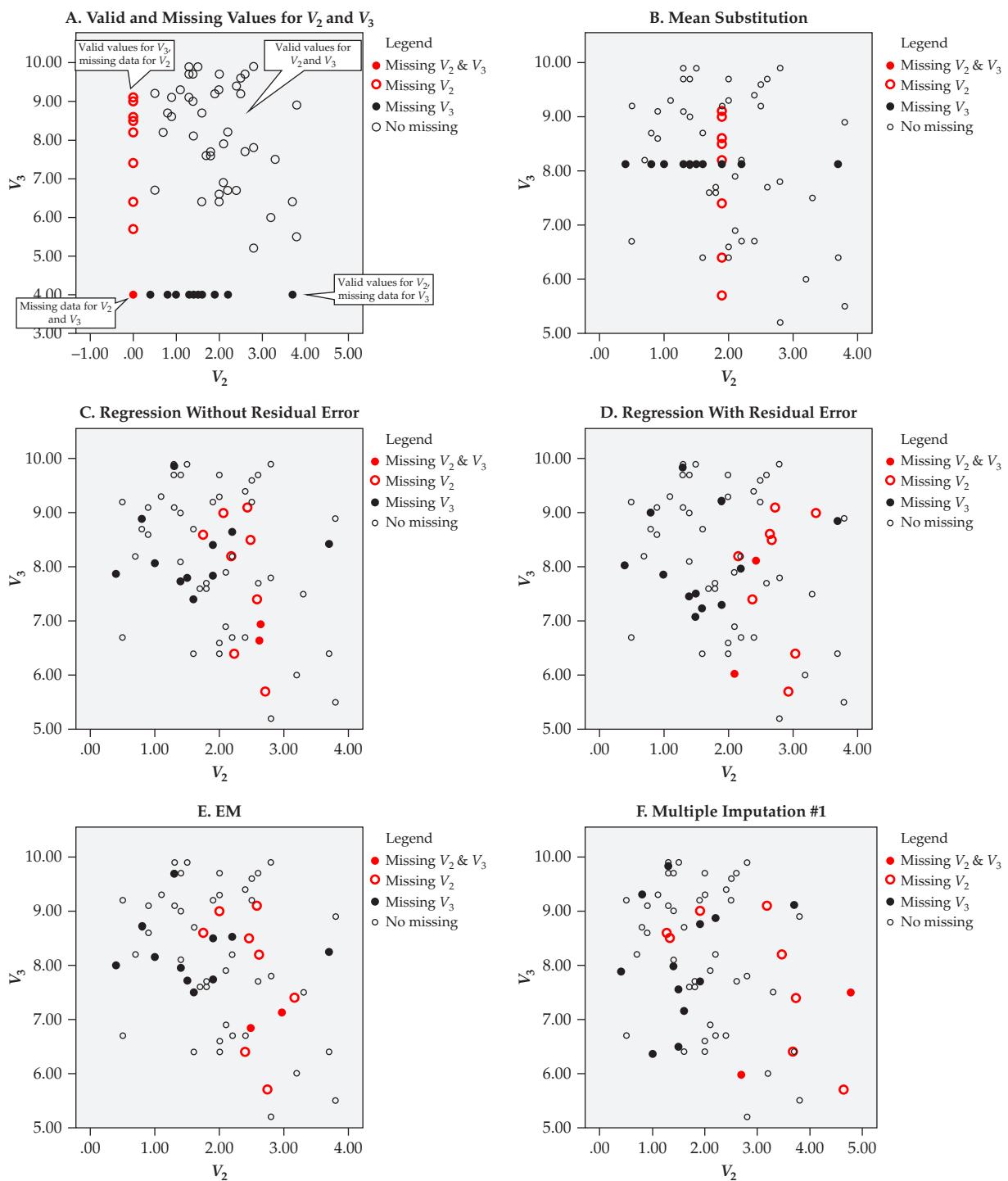
As we can see from the results, the mean substitution method imputed missing data for  $V_3$  as a value of 8.1 and a value of 1.9 for missing data on  $V_2$ . As we will see when we examine these results graphically, some distinct patterns are formed. Also, while the various methods all used different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

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

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

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

Note: Multiple imputation values estimated using auxiliary values.

**Figure 2.8**Comparing Imputed Values for Missing Data on  $V_2$  and  $V_3$ 

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

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

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

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

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

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

<sup>a</sup> Combined results of the five imputations.

NC: Not computed by IBM SPSS.

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

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

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

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

<sup>a</sup> Combined results of the five imputations.

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

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

**Table 2.9** Regression Results for Complete Case and Six Imputation Methods

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

Note: Dependent measure:  $V_9$ .

Bolded items: significant at .05.

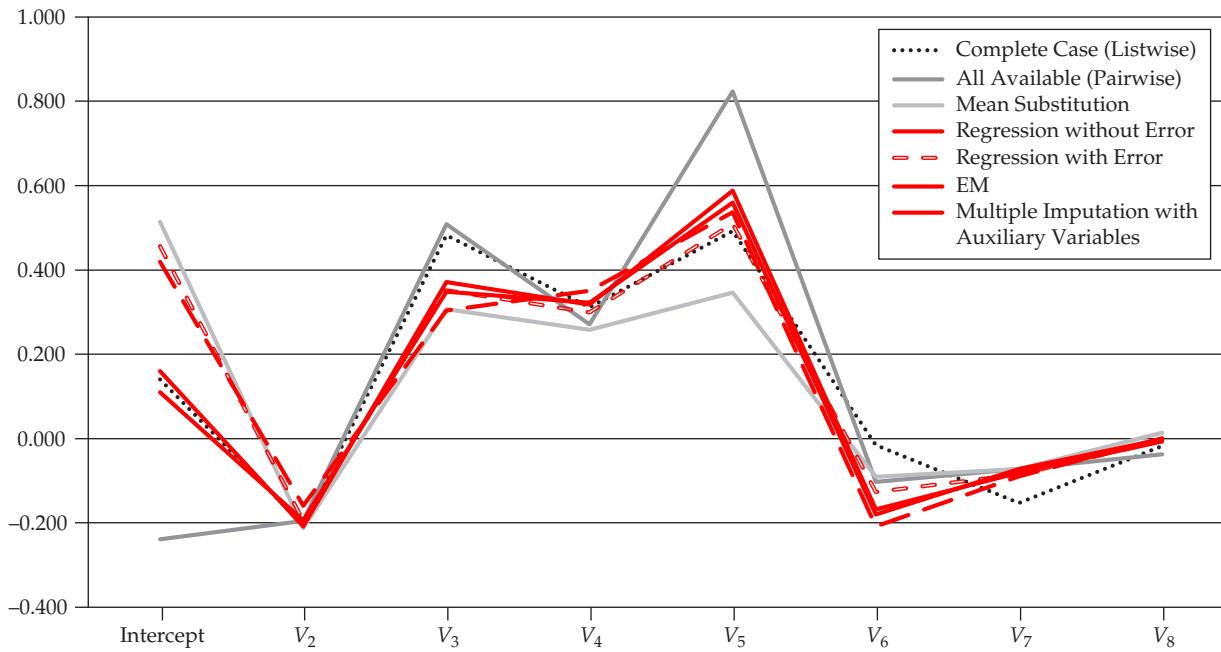
**SUMMARY** The task for the researcher is to coalesce the missing data patterns with the strengths and weaknesses of each approach and then select the most appropriate method. In the instance of differing estimates, the more conservative approach of combining the estimates into a single estimate (the multiple imputation approach) may be the most appropriate choice. Whichever approach is used, the dataset with replacement values should be saved for further analysis. The model-based approaches, especially multiple imputation, provides a means to assess a wider range of imputed values and then make combined estimates of the results across these ranges of variables.

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

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

**Figure 2.9**

## Regression Parameter Estimates for Complete Case and Six Imputation Methods



Dependent Variable: V<sub>9</sub> Independent Variables: V<sub>2</sub>, V<sub>3</sub>, V<sub>4</sub>, V<sub>5</sub>, V<sub>6</sub>, V<sub>7</sub>, V<sub>8</sub>

data process primarily MCAR (missing completely at random). Such a finding provides two advantages to the researcher. First, it should not involve any hidden impact on the results that need to be considered when interpreting the results except in one instance. Second, any of the imputation methods can be applied as remedies for the missing data with no marked biases expected. And the application of the model-based methods of EM and multiple imputation are appropriate for both MCAR and MAR. Thus, the selection need not be based on their ability to handle nonrandom processes, but instead on the applicability of the process and its impact on the results.

- 2 Imputation is the most logical course of action.** Even given the benefit of deleting cases and variables, the researcher is precluded from the simple solution of using the complete case method, because it results in an inadequate sample size. Some form of imputation is therefore needed to maintain an adequate sample size for any multivariate analysis.
- 3 Imputed correlations differ across techniques.** When estimating correlations among the variables in the presence of missing data, the researcher can choose from any number of imputation techniques. In this situation, however, there are differences in the results among these methods. First, there are general consistencies among the all-available information, mean substitution, regression with and without error, EM and multiple imputation method without auxiliary variables methods. There are differences, however, with the complete case approach and the multiple imputation with auxiliary variable methods. Even though the complete information approach would seem the most "safe" and conservative, in this case it is not recommended due to the small sample used (only 26 observations) and its marked differences from the other methods. The differences with the multiple imputation method with auxiliary variables can be attributed to the impact of the auxiliary variables, since multiple imputation without the auxiliary variables is consistent with the other methods. Interestingly, all of these approaches result in comparable regression coefficients except in a few instances for the complete case, all-available and mean substitution methods. Only in the instance of the complete case were the differences substantive enough to warrant concern.

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

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

## Outliers

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

### TWO DIFFERENT CONTEXTS FOR DEFINING OUTLIERS

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

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

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

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

## IMPACTS OF OUTLIERS

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

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

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

**Are Outliers Good or Bad?** Outliers cannot be categorically characterized as either beneficial or problematic, but instead must be viewed within the context of the analysis and should be evaluated by the types of information they may provide. When beneficial, outliers—although different from the majority of the sample—may be indicative of characteristics of the population that would not be discovered in the normal course of analysis. In contrast, problematic outliers are not representative of the population, are counter to the objectives of the analysis, and can seriously distort statistical tests. As such, they do not meet the framework of the analysis being conducted. While they may provide feedback for necessary adjustments to the analysis, they also provide the researcher a means of focusing the analysis and results on the intended population rather than being impacted by observations not even intended for inclusion. In these discussions, outliers are placed in a framework particularly suited for assessing the influence of individual observations and determining whether this influence is helpful or harmful.

## CLASSIFYING OUTLIERS

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

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

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

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

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

- **Unique combinations.** The fourth and final class of outlier contains observations that fall within the ordinary range of values on each of the variables. These observations are not particularly high or low on the variables, but are unique in their combination of values across the variables. In these situations, the researcher should retain the observation unless specific evidence is available that discounts the outlier as a valid member of the population.

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

## DETECTING AND HANDLING OUTLIERS

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

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

**UNIVARIATE DETECTION** The univariate identification of outliers examines the distribution of observations for each variable in the analysis and selects as outliers those cases falling at the outer ranges (high or low) of the distribution. The primary issue is establishing the threshold for designation of an outlier. The typical approach first converts the data values to standard scores, which have a mean of 0 and a standard deviation of 1. Because the values are expressed in a standardized format, comparisons across variables can be made easily. An outlier designation then occurs when an observation falls well to the outer boundaries of the distribution of values, many times identified as cases with standardized values of  $\pm 3$ , which makes them quite unique in terms of that characteristic.

In either case, the researcher must recognize that a certain number of observations may occur normally in these outer ranges of the distribution. The researcher should strive to identify only those truly distinctive observations and designate them as outliers.

**BIVARIATE DETECTION** In addition to the univariate assessment, pairs of variables can be assessed jointly through a scatterplot. Cases that fall markedly outside the range of the other observations will be seen as isolated points in the scatterplot. To assist in determining the expected range of observations in this two-dimensional portrayal, an ellipse representing a bivariate normal distribution’s confidence interval (typically set at the 90% or 95% level) is superimposed over the scatterplot. This ellipse provides a graphical portrayal of the confidence limits and facilitates identification of the outliers. A variant of the scatterplot is termed the influence plot, with each point varying in size in relation to its influence on the relationship.

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

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

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

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

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

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

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

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

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

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

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

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

**Outlier Description and Profiling** Once the potential outliers are identified, the researcher should generate profiles of each outlier observation and identify the variable(s) responsible for its being an outlier. In addition to this visual examination, the researcher can also employ multivariate techniques such as discriminant analysis (Chapter 7) or multiple regression (Chapter 5) to identify the differences between outliers and the other observations. If possible the

## Outlier Detection

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

For small samples (80 or fewer observations), outliers typically are defined as cases with standard scores of 2.5 or greater.

For larger sample sizes, increase the threshold value of standard scores up to 4.

If standard scores are not used, identify cases falling outside the ranges of 2.5 versus 4 standard deviations, depending on the sample size.

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

Use scatterplots with confidence intervals at a specified alpha level.

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

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

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

**Retention or Deletion of the Outlier** After the outliers are identified, profiled and categorized, the researcher must decide on the retention or deletion of each one. Many philosophies among researchers offer guidance as to how to deal with outliers. Our belief is that they should be retained unless demonstrable proof indicates that they are truly aberrant and not representative of any observations in the population. If they do portray a representative element or segment of the population, they should be retained to ensure generalizability to the entire population. As outliers are deleted, the researcher runs the risk of improving the multivariate analysis but limiting its generalizability.

If outliers are problematic in a particular technique, many times they can be accommodated in the analysis in a manner in which they do not seriously distort the analysis. Techniques such as robust regression allow for the retention of outliers, but reduce their impact on the model results. Moreover, research has shown that employing more nonparametric tests can also reduce the influence of outliers [7].

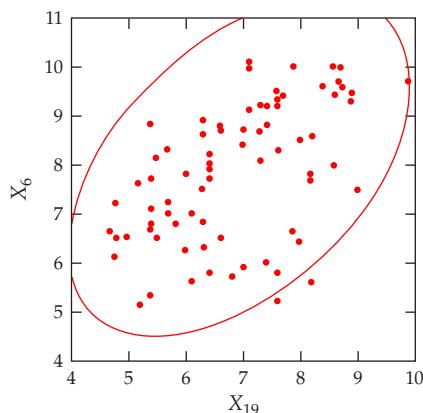
## AN ILLUSTRATIVE EXAMPLE OF ANALYZING OUTLIERS

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

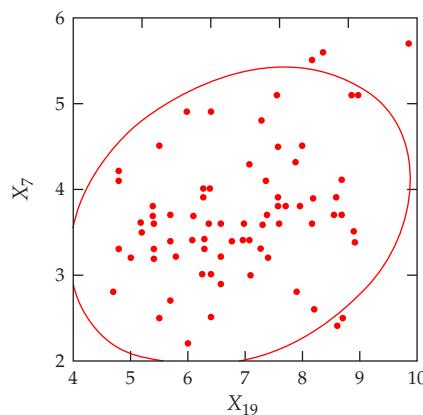
**Outlier Detection** The first step is examination of all the variables from a univariate perspective. Bivariate methods will then be employed to examine the relationships between the dependent variable ( $X_{19}$ ) and each of the independent variables. From each of these scatterplots, observations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

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

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



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



**Table 2.10** Univariate, Bivariate, and Multivariate Outlier Detection Results

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

<sup>a</sup>Mahalanobis  $D^2$  value based on the 13 HBAT perceptions ( $X_6$  to  $X_{18}$ ).

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

**MULTIVARIATE DETECTION** The final diagnostic method is to assess multivariate outliers with the Mahalanobis  $D^2$  measure (see Table 2.10). This analysis evaluates the position of each observation compared with the center of all observations on a set of variables. In this case, all the metric independent variables were used. The calculation of the  $D^2/df$  value ( $df = 13$ ) allows for identification of outliers through an approximate test of statistical significance. Because the sample has only 100 observations, a threshold value of 2.5 will be used rather than the value of 3.5 or 4.0 used in large samples. With this threshold, two observations (98 and 36) are identified as significantly different. It is interesting that these observations were not seen in earlier univariate and bivariate analyses but appear only in the multivariate tests. This result indicates they are not unique on any single variable but instead are unique in combination.

**Retention or Deletion of the Outliers** As a result of these diagnostic tests, no observations demonstrate the characteristics of outliers that should be eliminated. Each variable has some observations that are extreme, and they should be considered if that variable is used in an analysis. No observations are extreme on a sufficient number of variables to be considered unrepresentative of the population. In all instances, the observations designated as outliers, even with the multivariate tests, seem similar enough to the remaining observations to be retained in the multivariate analyses. However, the researcher should always examine the results of each specific multivariate technique to identify observations that may become outliers in that particular application. In the case of regression analysis, Chapter 5 will provide additional methods to assess the relative influence of each observation and provide more insight into the possible deletion of an observation as an outlier.

## Testing the Assumptions of Multivariate Analysis

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

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

### ASSESSING INDIVIDUAL VARIABLES VERSUS THE VARIATE

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

procedures. Discussions in each chapter address the methods used to assess the assumptions underlying the variate for each multivariate technique.

## FOUR IMPORTANT STATISTICAL ASSUMPTIONS

Multivariate techniques and their univariate counterparts are all based on a fundamental set of assumptions representing the requirements of the underlying statistical theory. Although many assumptions or requirements come into play in one or more of the multivariate techniques we discuss in the text, four of them potentially affect every univariate and multivariate statistical technique.

**Normality** The most fundamental assumption in multivariate analysis is **normality**, referring to the shape of the data distribution for an individual metric variable and its correspondence to the **normal distribution**, the benchmark for statistical methods. *If the variation from the normal distribution is sufficiently large, all resulting statistical tests are invalid, because normality is required to use the F and t statistics.* Both the univariate and the multivariate statistical methods discussed in this text are based on the assumption of univariate normality, with the multivariate methods also assuming multivariate normality.

**UNIVARIATE VERSUS MULTIVARIATE NORMALITY** Univariate normality for a single variable is easily tested, and a number of corrective measures are possible, as shown later. In a simple sense, multivariate normality (the combination of two or more variables) means that the individual variables are normal in a univariate sense and that their combinations are also normal. Thus, *if a variable is multivariate normal, it is also univariate normal. However, the reverse is not necessarily true (two or more univariate normal variables are not necessarily multivariate normal).* Thus, a situation in which all variables exhibit univariate normality will help gain, although not guarantee, multivariate normality. Multivariate normality is more difficult to test [36, 83], but specialized tests are available in the techniques most affected by departures from multivariate normality. In most cases assessing and achieving univariate normality for all variables is sufficient, and we will address multivariate normality only when it is especially critical. Even though large sample sizes tend to diminish the detrimental effects of nonnormality, the researcher should always assess the normality for all metric variables included in the analysis.

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

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

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

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

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

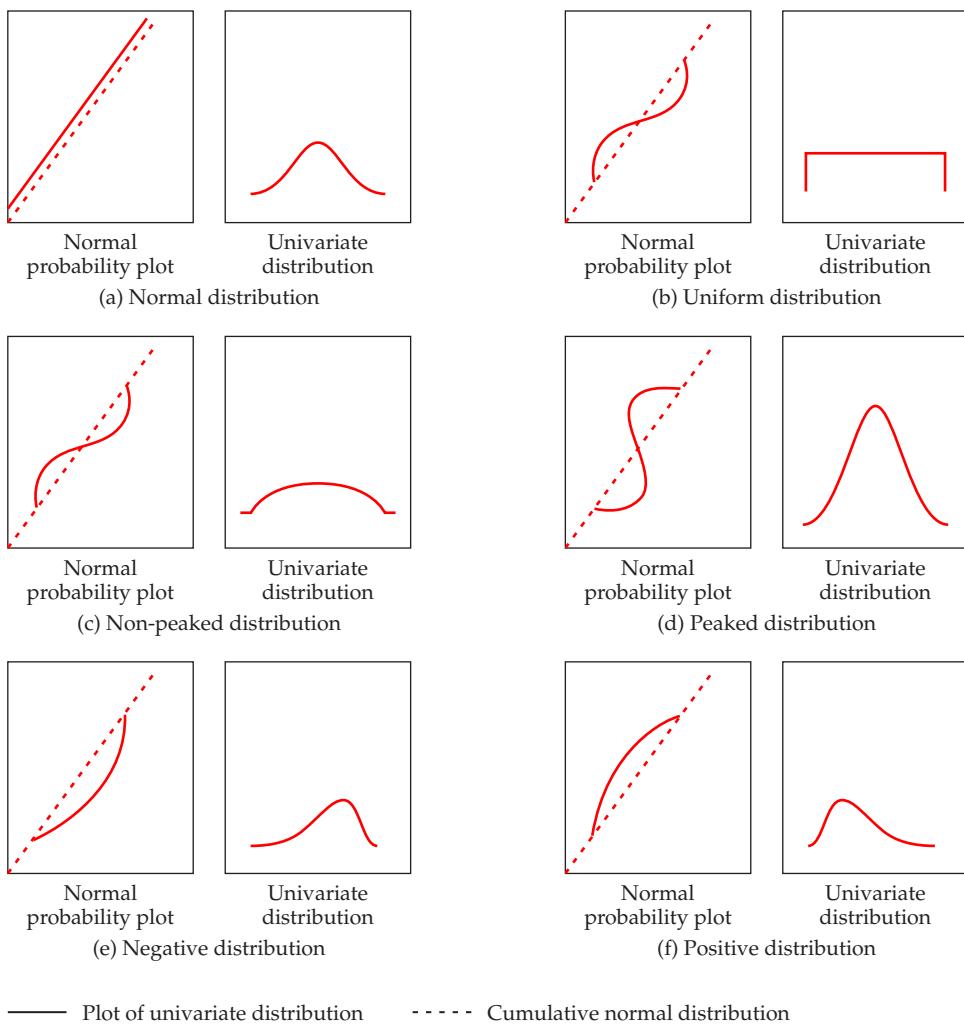
**TESTS OF NORMALITY** Researchers have a number of different approaches to assess normality, but they primarily can be classified as either graphical or statistical. Graphical methods were developed to enable normality assessment without the need for complex computations. They provide the researcher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

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

Figure 2.11 shows several departures from normality and their representation in the normal probability in terms of kurtosis and skewness. First, departures from the normal distribution in terms of kurtosis are easily seen in the normal probability plots. When the line falls below the diagonal, the distribution is flatter than expected. When it goes above the diagonal, the distribution is more peaked than the normal curve. For example, in the normal probability plot of a peaked distribution (Figure 2.11d), we see a distinct S-shaped curve. Initially the distribution is flatter, and the plotted line falls below the diagonal. Then the peaked part of the distribution rapidly moves the plotted line above the diagonal, and eventually the line shifts to below the diagonal again as the distribution flattens. A non-peaked distribution has the opposite pattern (Figure 2.11c). Skewness is also easily seen, most often represented by a simple arc, either above or below the diagonal. A negative skewness (Figure 2.11e) is indicated by an arc below the diagonal, whereas an arc above the diagonal represents a positively skewed distribution (Figure 2.11f). An excellent source for interpreting normal probability plots, showing the various patterns and interpretations, is Daniel and Wood [26]. These specific patterns not only identify non-normality but also tell us the form of the original distribution and the appropriate remedy to apply.

**Statistical Tests** In addition to examining the normal probability plot, one can also use statistical tests to assess normality. A simple test is a rule of thumb based on the skewness and kurtosis values (available as part of the basic descriptive statistics for a variable computed by all statistical programs). The statistic value ( $z$ ) for the skewness value is calculated as:

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

**Figure 2.11** Normal Probability Plots and Corresponding Univariate Distributions

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

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

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

Specific statistical tests for normality are also available in all the statistical programs. The two most common are the Shapiro-Wilks test and a modification of the Kolmogorov-Smirnov test. Each calculates the level of significance for the differences from a normal distribution. The researcher should always remember that tests of significance are less useful in small samples (fewer than 30) and quite sensitive in large samples (exceeding 1,000 observations).

Thus, the researcher should always use both the graphical plots and any statistical tests to assess the actual degree of departure from normality.

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

**Homoscedasticity** The next assumption is related primarily to dependence relationships between variables. **Homoscedasticity** refers to the assumption that dependent variable(s) exhibit equal levels of variance across the range of predictor variable(s). Homoscedasticity is desirable because *the variance of the dependent variable being explained in the dependence relationship should not be concentrated in only a limited range of the independent values*. In most situations, we have many different values of the dependent variable at each value of the independent variable. For this relationship to be fully captured, the dispersion (variance) of the dependent variable values must be relatively equal at each value of the predictor variable. If this dispersion is unequal across values of the independent variable, the relationship is said to be **heteroscedastic**.

**TYPE OF INDEPENDENT VARIABLE** Although the dependent variables must be metric, this concept of an equal spread of variance across independent variables can be applied when the independent variables are either metric or nonmetric. The type of independent variable dictates how homoscedasticity is assessed.

**Metric Independent Variables** The concept of homoscedasticity is based on the spread of dependent variable variance across the range of independent variable values, which is encountered in techniques such as multiple regression. The dispersion of values for the dependent variable should be as large for small values of the independent values as it is for moderate and large values. In a scatterplot, it is seen as an elliptical distribution of points.

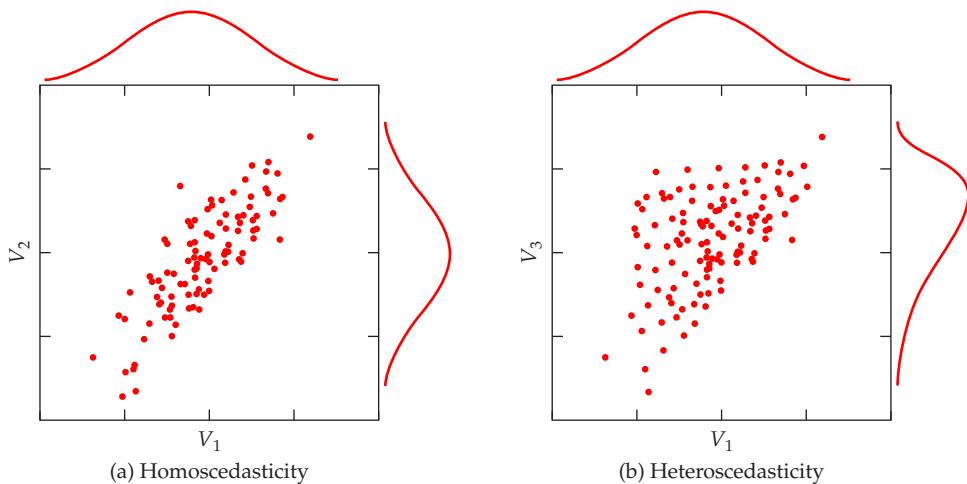
**Nonmetric Independent Variables** In these analyses (e.g., ANOVA and MANOVA) the focus now becomes the equality of the variance (single dependent variable) or the variance/covariance matrices (multiple dependent variables) across the groups formed by the nonmetric independent variables. The equality of variance/covariance matrices is also seen in discriminant analysis, but in this technique the emphasis is on the spread of the independent variables across the groups formed by the nonmetric dependent measure.

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

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

**Skewed Distribution of One or Both Variables** In Figure 2.12a, the scatterplots of data points for two variables ( $V_1$  and  $V_2$ ) with normal distributions exhibit equal dispersion across all data values (i.e., homoscedasticity). However, in Figure 2.12b we see unequal dispersion (heteroscedasticity) caused by skewness of one of the variables ( $V_3$ ). For the different values of  $V_3$ , there are different patterns of dispersion for  $V_1$ .

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

**Figure 2.12** Scatterplots of Homoscedastic and Heteroscedastic Relationships

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

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

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

Boxplots work well to represent the degree of variation between groups formed by a categorical variable. The length of the box and the whiskers each portray the variation of data within that group. Thus, heteroscedasticity would be portrayed by substantial differences in the length of the boxes and whiskers between groups representing the dispersion of observations in each group.

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

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

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

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

**IDENTIFYING NONLINEAR RELATIONSHIPS** The most common way to assess linearity is to examine scatterplots of the variables and to identify any nonlinear patterns in the data. Many scatterplot programs can show the straight line depicting the linear relationship, enabling the researcher to better identify any nonlinear characteristics. An alternative approach is to run a simple regression analysis (the specifics of this technique are covered in Chapter 5) and to examine the **residuals**. The residuals reflect the unexplained portion of the dependent variable; thus, any nonlinear portion of the relationship will show up in the residuals. A third approach is to explicitly model a nonlinear relationship by the testing of alternative model specifications (also known as curve fitting) that reflect the nonlinear elements. A discussion of this approach and residual analysis is found in Chapter 5.

**REMEDIES FOR NONLINEARITY** If a nonlinear relationship is detected, the most direct approach is to transform one or both variables to achieve linearity. A number of available transformations are discussed later in this chapter. An alternative to data transformation is the creation of new variables to represent the nonlinear portion of the relationship. The process of creating and interpreting these additional variables, which can be used in all linear relationships, is discussed in Chapter 5.

**Absence of Correlated Errors** Predictions in any of the dependence techniques are not perfect, and we will rarely find a situation in which they are. However, we do attempt to ensure that any prediction errors are uncorrelated with each other. For example, if we found a pattern that suggests every other error is positive while the alternative error terms are negative, we would know that some unexplained systematic relationship exists in the dependent variable. If such a situation exists, we cannot be confident that our prediction errors are independent of the levels at which we are trying to predict. Some other factor is affecting the results, but is not included in the analysis.

**IDENTIFYING CORRELATED ERRORS** One of the most common violations of the assumption that errors are uncorrelated is due to the data collection process. Similar factors that affect one group may not affect the other. If the groups are analyzed separately, the effects are constant within each group and do not impact the estimation of the relationship. But if the observations from both groups are combined, then the final estimated relationship must be a compromise between the two actual relationships. This combined effect leads to biased results because an unspecified cause is affecting the estimation of the relationship. The common example used is the collection of data within classes or other groups of respondents, where some form of group dynamic may impact each group differently. This situation is addressed in our discussion of multilevel models in Chapter 5.

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

## Testing Statistical Assumptions

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

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

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

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

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

**REMEDIES FOR CORRELATED ERRORS** Correlated errors must be corrected by including the omitted causal factor into the multivariate analysis. In our earlier example, the researcher would add a variable indicating in which class the respondents belonged. The most common remedy is the addition of a variable(s) to the analysis that represents the omitted factor. The key task facing the researcher is not the actual remedy, but rather the identification of the unspecified effect and a means of representing it in the analysis. But beyond just including the variable, the intercorrelation among observations is best handled in some form of multi-level or panel model. Chapter 5 presents a discussion of these extensions of multiple regression which can accommodate grouped/clustered observations or time series data in a structured framework.

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

## Data Transformations

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

of the data) or *data derived* (where the transformations are suggested strictly by an examination of the data). Yet in either case the researcher must proceed many times by trial and error, monitoring the improvement versus the need for additional transformations.

All the transformations described here are easily carried out by simple commands in the popular statistical packages. We focus on transformations that can be computed in this manner, although more sophisticated and complicated methods of data transformation are available (e.g., see Box and Cox [12]).

## TRANSFORMATIONS RELATED TO STATISTICAL PROPERTIES

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

**Achieving Normality and Homoscedasticity** Data transformations provide the principal means of correcting non-normality and heteroscedasticity. In both instances, patterns of the variables suggest specific transformations. For non-normal distributions, the two most common patterns are flat distributions and skewed distributions. For the flat distribution, the most common transformation is the inverse (e.g.,  $1/Y$  or  $1/X$ ). Skewed distributions can be transformed by taking the square root, logarithms, squared, or cubed ( $X^2$  or  $X^3$ ) terms or even the inverse of the variable. Usually negatively skewed distributions are best transformed by employing a squared or cubed transformation, whereas the logarithm or square root typically works best on positive skewness. In many instances, the researcher may apply all of the possible transformations and then select the most appropriate transformed variable.

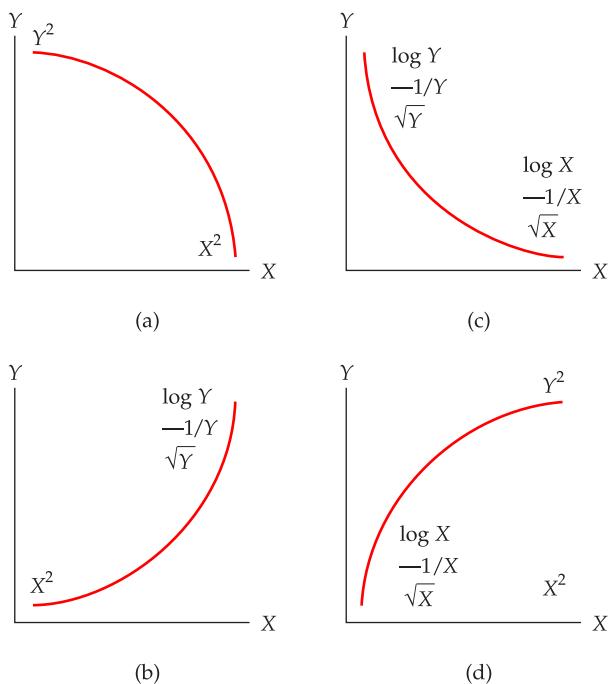
Heteroscedasticity is an associated problem, and in many instances “curing” this problem will deal with normality problems as well. Heteroscedasticity is also a problem with the distribution of the variable(s). When examining the scatterplot, the most common pattern is the cone-shaped distribution. If the cone opens to the right, take the inverse; if the cone opens to the left, take the square root. Some transformations can be associated with certain types of data. For example, frequency counts suggest a square root transformation; proportions are best transformed by the arcsin transformation ( $X_{\text{new}} = 2 \arcsin \sqrt{x_{\text{old}}}$ ) and proportional change is best handled by taking the logarithm of the variable. In all instances, once the transformations have been performed, the transformed data should be tested to see whether the desired remedy was achieved.

**Achieving Linearity** Numerous procedures are available for achieving linearity between two variables, but most simple nonlinear relationships can be placed in one of four categories (see Figure 2.13). In each quadrant, the potential transformations for both dependent and independent variables are shown. For example, if the relationship looks like Figure 2.13a, then either variable can be squared to achieve linearity. When multiple transformation possibilities are shown, start with the top method in each quadrant and move downward until linearity is achieved. An alternative approach is to use additional variables, termed *polynomials*, to represent the nonlinear components. This method is discussed in more detail in Chapter 5.

## TRANSFORMATIONS RELATED TO INTERPRETATION

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

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

**Figure 2.13**

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

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

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

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

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

## TRANSFORMATIONS RELATED TO SPECIFIC RELATIONSHIP TYPES

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

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

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

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

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

## TRANSFORMATIONS RELATED TO SIMPLIFICATION

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

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

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

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

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

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

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

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

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

## GENERAL GUIDELINES FOR TRANSFORMATIONS

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

One caveat for transformations:

---

When explanation is important, beware of transformations!

---

## Transforming Data

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

Noticeable effects should occur when the ratio is less than 4.

When the transformation can be performed on either of two variables, select the variable with the smallest ratio.

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

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

Transformations may change the interpretation of the variables; for example, transforming variables by taking their logarithm translates the relationship into a measure of proportional change (elasticity); always be sure to explore thoroughly the possible interpretations of the transformed variables.

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

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

## An Illustration of Testing the Assumptions Underlying Multivariate Analysis

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

### NORMALITY

The assessment of normality of the metric variables involves both empirical measures of a distribution's shape characteristics (skewness and kurtosis) and the normal probability plots. The empirical measures provide a guide as to the variables with significant deviations from normality, and the normal probability plots provide a visual portrayal of the shape of the distribution. The two portrayals complement each other when selecting the appropriate transformations.

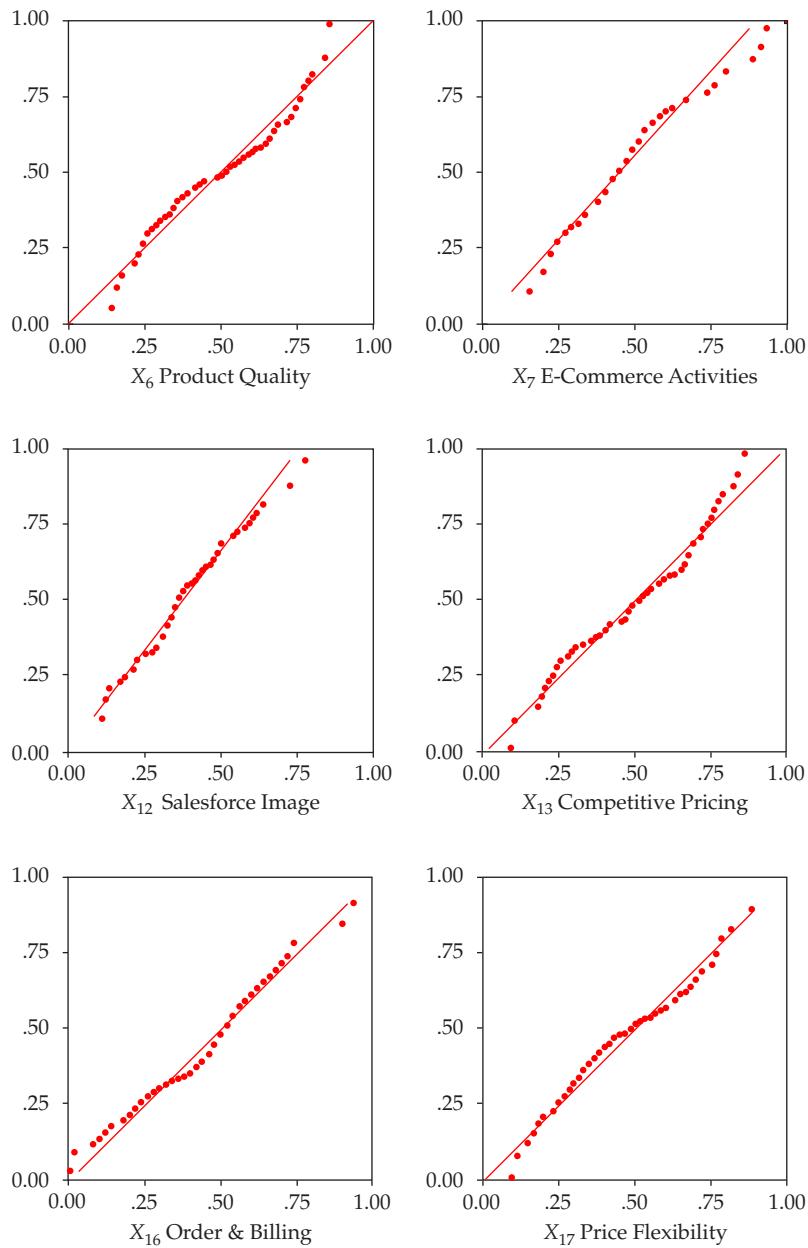
Table 2.11 and Figure 2.14 contain the empirical measures and normal probability plots for the metric variables in our data set. Our first review concerns the empirical measures reflecting the shape of the distribution (skewness and kurtosis) as well as a statistical test for normality (the modified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

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

Firm Characteristics	SHAPE DESCRIPTORS				Tests of Normality				Applicable Remedies	
	Skewness		Kurtosis		Statistic	z value	Significance	Description of the Distribution	Transformation	Significance After Remedy
	Variable	Statistic	z value	Statistic						
$X_6$	-.245	-1.01	-1.132	<b>-2.37</b>	.109	.005	Almost uniform distribution	Squared term	.015	
$X_7$	.660	<b>2.74</b>	.735	<b>1.54</b>	.122	<b>.001</b>	Peaked with positive skew	Logarithm	.037	
$X_8$	-.203	-.84	-.548	-.115	.060	.200 <sup>a</sup>	Normal distribution			
$X_9$	-.136	-.56	-.586	-.123	.051	.200 <sup>a</sup>	Normal distribution			
$X_{10}$	.044	.18	-.888	-.186	.065	.200 <sup>a</sup>	Normal distribution			
$X_{11}$	-.092	-.38	-.522	-.109	.060	.200 <sup>a</sup>	Normal distribution			
$X_{12}$	.377	1.56	.410	.86	.111	<b>.004</b>	Slight positive skew and peakedness			
$X_{13}$	-.240	-1.00	-.903	-.189	.106	<b>.007</b>	Peaked	Clipped term	-	
$X_{14}$	.008	.03	-.445	-.93	.064	.200 <sup>a</sup>	Normal distribution			
$X_{15}$	.299	1.24	.016	.03	.074	.200 <sup>a</sup>	Normal distribution			
$X_{16}$	-.334	-1.39	.244	.51	.129	<b>.000</b>	Negative skewness	Squared term	.066	
$X_{17}$	.323	1.34	-.816	-.171	.101	<b>.013</b>	Peaked, positive skewness	Inverse	.187	
$X_{18}$	-.463	-1.92	.218	.46	.084	.082	Normal distribution			
<b>Performance Measures</b>										
$X_{19}$	.078	.32	-.791	-.165	.078	.137	Normal distribution			
$X_{20}$	.044	.18	-.089	-.19	.077	.147	Normal distribution			
$X_{21}$	-.093	-.39	-.090	-.19	.073	.200 <sup>a</sup>	Normal distribution			
$X_{22}$	-.132	-.55	-.684	-.143	.075	.180	Normal distribution			

<sup>a</sup>Lower bound of true significance.

Note: The z values are derived by dividing the statistics by the appropriate standard errors of .241 (skewness) and .478 (kurtosis). The equations for calculating the standard errors are given in the text.

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

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

Table 2.11 also suggests the appropriate remedy for each of the variables. Two variables ( $X_6$  and  $X_{16}$ ) were transformed by taking the square root.  $X_7$  was transformed by logarithm, whereas  $X_{17}$  was squared and  $X_{13}$  was cubed. Only  $X_{12}$  could not be transformed to improve on its distributional characteristics. For the other five variables, their tests of normality were now either nonsignificant ( $X_{16}$  and  $X_{17}$ ) or markedly improved to more acceptable levels ( $X_6$ ,  $X_7$ , and  $X_{13}$ ). Figure 2.15 demonstrates the effect of the transformation on  $X_{17}$  in achieving normality. The transformed  $X_{17}$  appears markedly more normal in the graphical portrayals, and the statistical descriptors are also improved. The researcher should always examine the transformed variables as rigorously as the original variables in terms of their normality and distribution shape.

In the case of the remaining variable ( $X_{12}$ ), none of the transformations could improve the normality. This variable will have to be used in its original form. In situations where the normality of the variables is critical, the transformed variables can be used with the assurance that they meet the assumptions of normality. But the departures from normality are not so extreme in any of the original variables that they should never be used in any analysis in their original form. If the technique has a robustness to departures from normality, then the original variables may be preferred for the comparability in the interpretation phase.

## HOMOSCEDASTICITY

All statistical packages have tests to assess homoscedasticity on a univariate basis (e.g., the Levene test in SPSS) where the variance of a metric variable is compared across levels of a nonmetric variable. For our purposes, we examine each of the metric variables across the five nonmetric variables in the dataset. These analyses are appropriate in preparation for analysis of variance or multivariate analysis of variance, in which the nonmetric variables are the independent variables, or for discriminant analysis, in which the nonmetric variables are the dependent measures.

Table 2.12 contains the results of the Levene test for each of the nonmetric variables. Among the performance factors, only  $X_4$  (Region) has notable problems with heteroscedasticity. For the 13 firm characteristic variables, only  $X_6$  and  $X_{17}$  show patterns of heteroscedasticity on more than one of the nonmetric variables. Moreover, in no instance do any of the nonmetric variables have more than two problematic metric variables. The actual implications of these instances of heteroscedasticity must be examined whenever group differences are examined using these nonmetric variables as independent variables and these metric variables as dependent variables. The relative lack of either numerous problems or any consistent patterns across one of the nonmetric variables suggests that heteroscedasticity problems will be minimal. If the assumption violations are found, variable transformations are available to help remedy the variance dispersion.

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

The tests for homoscedasticity of two metric variables, encountered in methods such as multiple regression, are best accomplished through graphical analysis, particularly an analysis of the residuals. The interested reader is referred to Chapter 5 for a complete discussion of residual analysis and the patterns of residuals indicative of heteroscedasticity.

## LINEARITY

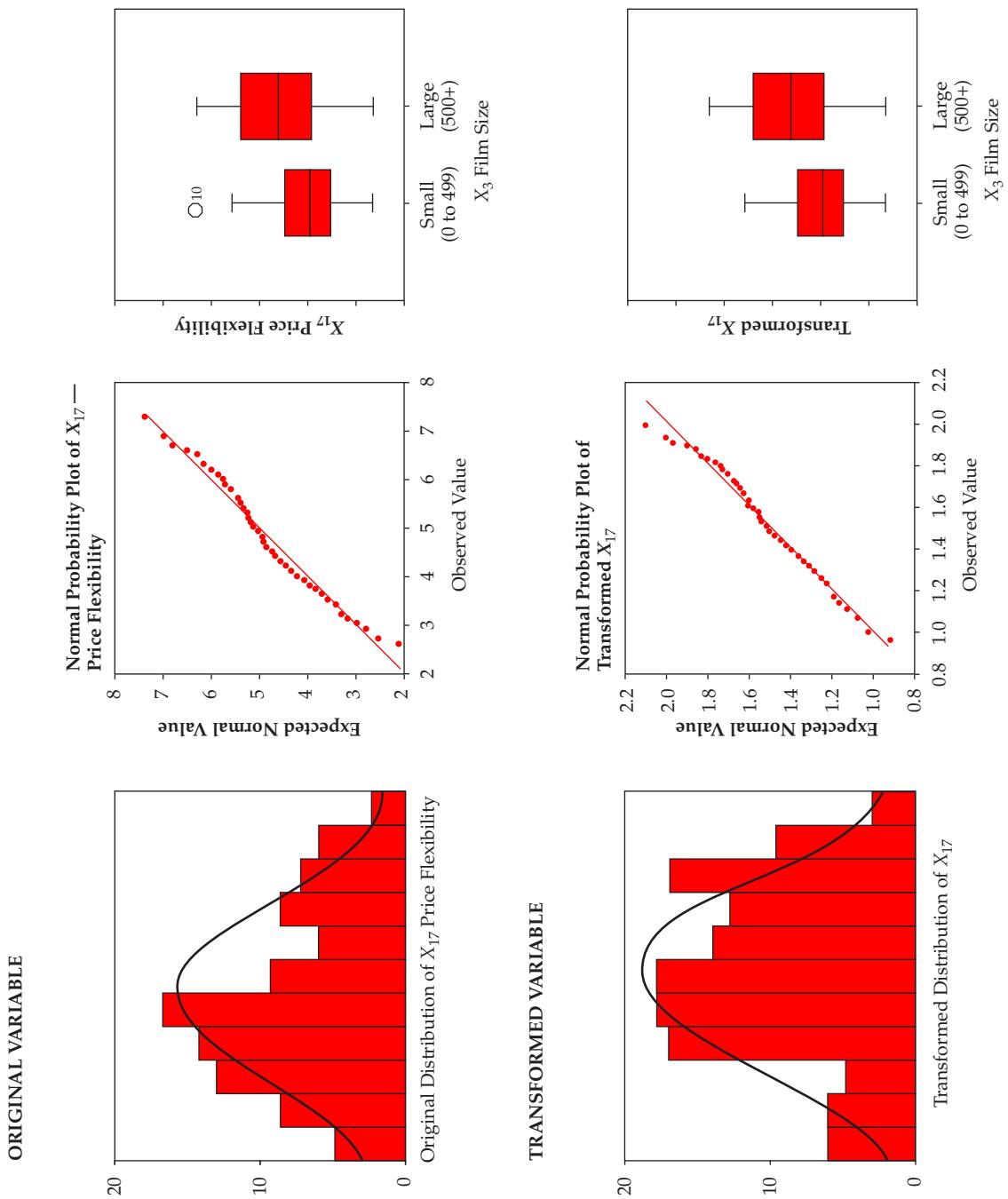
The final assumption to be examined is the linearity of the relationships. In the case of individual variables, this linearity relates to the patterns of association between each pair of variables and the ability of the correlation coefficient to adequately represent the relationship. If nonlinear relationships are indicated, then the researcher can either transform one or both of the variables to achieve linearity or create additional variables to represent the nonlinear components. For our purposes, we rely on the visual inspection of the relationships to determine whether nonlinear

**Table 2.12** Testing for Homoscedasticity

NONMETRIC/CATEGORICAL VARIABLE															
Firm Characteristics	$X_1$ Customer Type			$X_2$ Industry Type			$X_3$ Firm Size			$X_4$ Region			$X_5$ Distribution System		
	Metric	Levene Statistic	Sig.	Levene Statistic	Sig.	Levene Statistic	Sig.	Levene Statistic	Sig.	Levene Statistic	Sig.	Levene Statistic	Sig.	Levene Statistic	Sig.
	$X_6$	<b>17.47</b>	.00	.01	.94	.02	.89	<b>17.86</b>	.00	.48	.49				
$X_7$	.58	.56	.09	.76	.09	.76	.05	.83	.287	.09					
$X_8$	.37	.69	.48	.49	1.40	.24	.72	.40	.11	.74					
$X_9$	.43	.65	.02	.88	.17	.68	.58	.45	1.20	.28					
$X_{10}$	.74	.48	.00	.99	.74	.39	1.19	.28	.69	.41					
$X_{11}$	.06	.95	.15	.70	.09	.76	3.44	.07	1.72	.19					
$X_{12}$	2.46	.09	.36	.55	.06	.80	1.55	.22	1.55	.22					
$X_{13}$	.84	.43	<b>4.43</b>	<b>04</b>	1.71	.19	.24	.63	2.09	.15					
$X_{14}$	2.39	.10	2.53	.11	<b>4.55</b>	<b>.04</b>	.25	.62	.16	.69					
$X_{15}$	1.13	.33	.47	.49	1.05	.31	.01	.94	.59	.45					
$X_{16}$	1.65	.20	.83	.37	.31	.58	2.49	.12	<b>4.60</b>	<b>.03</b>					
$X_{17}$	<b>5.56</b>	<b>.01</b>	2.84	.10	<b>4.19</b>	<b>.04</b>	<b>16.21</b>	<b>.00</b>	.62	.43					
$X_{18}$	.87	.43	.30	.59	.18	.67	2.25	.14	<b>4.27</b>	<b>.04</b>					
Performance Measures															
$X_{19}$	<b>3.40</b>	<b>.04</b>	.00	.96	.73	.39	<b>8.57</b>	.00	.18	.67					
$X_{20}$	1.64	.20	.03	.86	.03	.86	<b>7.07</b>	<b>.01</b>	.46	.50					
$X_{21}$	1.06	.35	.11	.74	.68	.41	<b>11.54</b>	<b>.00</b>	2.67	.10					
$X_{22}$	.15	.86	.30	.59	.74	.39	.00	.99	1.47	.23					

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

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



(Continued)

**Figure 2.15** Continued

Distribution Characteristics Before and After Transformation					
SHAPE DESCRIPTORS					
Variable Form	Statistic	z value <sup>a</sup>	Kurtosis Statistic	Test of Normality	
				Statistic	Significance
Original $X_{17}$	.323	1.34	-.816	-.101	.013
Transformed $X_{17}^b$	-.121	.50	-.803	-1.68	.080
<sup>a</sup> The z values are derived by dividing the statistics by the appropriate standard errors of .241 (skewness) and .478 (kurtosis). The equations for calculating the standard errors are given in the text.					
<sup>b</sup> Logarithmic transformation.					
Levene Test Statistic					
Variable Form	$X_1$ Customer Type	$X_2$ Industry Type	$X_3$ Firm Size	$X_4$ Region	$X_5$ Distribution System
Original $X_{17}$	5.56**	2.84	4.19*	16.21**	.62
$X_{17}$	2.76	2.23	1.20	3.11**	.01

\* Significant at .05 significance level.

\*\* Significant at .01 significance level.

relationships are present. The reader can also refer to Figure 2.2, the scatterplot matrix containing the scatterplot for selected metric variables in the dataset. Examination of the scatterplots does not reveal any apparent nonlinear relationships. Review of the scatterplots not shown in Figure 2.2 also did not reveal any apparent nonlinear patterns. Thus, transformations are not deemed necessary. The assumption of linearity will also be checked for the entire multivariate model, as is done in the examination of residuals in multiple regression.

## SUMMARY

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

## Incorporating Nonmetric Data With Dummy Variables

A critical factor in choosing and applying the correct multivariate technique is the measurement properties of the dependent and independent variables (see Chapter 1 for a more detailed discussion of selecting multivariate techniques). Some of the techniques, such as discriminant analysis or multivariate analysis of variance, specifically require nonmetric data as dependent or independent variables. In many instances, metric variables must be used as independent variables, such as in regression analysis, discriminant analysis, and canonical correlation. Moreover, the interdependence techniques of factor and cluster analysis generally require metric variables. To this point, all discussions assumed metric measurement for variables. What can we do when the variables are nonmetric, with two or more categories? Are nonmetric variables, such as gender, marital status, or occupation, precluded from use in many multivariate techniques? The answer is no, and we now discuss how to incorporate nonmetric variables into many of these situations that require metric variables.

### CONCEPT OF DUMMY VARIABLES

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

First, assume we wish to include gender, which has two categories, female and male. We also have measured household income level by three categories (see Figure 2.16). To represent the nonmetric variable gender, we would create two new dummy variables ( $X_1$  and  $X_2$ ), as shown in Figure 2.16.  $X_1$  would represent those individuals who are

**Figure 2.16** Representing Nonmetric Variables with Dummy Variables

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

female with a value of 1 and would give all males a value of 0. Likewise,  $X_2$  would represent all males with a value of 1 and give females a value of 0. Both variables ( $X_1$  and  $X_2$ ) are not necessary, however, because when  $X_1 = 0$ , gender must be female by definition. Thus, we need include only one of the variables ( $X_1$  or  $X_2$ ) to test the effect of gender.

Correspondingly, if we had also measured household income with three levels, as shown in Figure 2.16, we would first define three dummy variables ( $X_3$ ,  $X_4$ , and  $X_5$ ). In the case of gender, we would not need the entire set of dummy variables, and instead use  $k - 1$  dummy variables, where  $k$  is the number of categories. Thus, we would use two of the dummy variables to represent the effects of household income.

## DUMMY VARIABLE CODING

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

**Indicator Coding** The first approach, known as **indicator coding**, uses three ways to represent the household income levels with two dummy variables, as shown in Figure 2.17. An important consideration is the *reference category or comparison group, the category that received all zeros for the dummy variables*. For example, in regression analysis, the regression coefficients for the dummy variables represent *deviations from the comparison group on the dependent variable*. The deviations represent the differences between the dependent variable mean score for each group of respondents (represented by a separate dummy variable) and the comparison group. This form is most appropriate in a logical comparison group, such as in an experiment. In an experiment with a control group acting as the comparison group, the coefficients are the mean differences on the dependent variable for each treatment group from the control group. Any time dummy variable coding is used, we must be aware of the comparison group and remember the impacts it has in our interpretation of the remaining variables.

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

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

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

## USING DUMMY VARIABLES

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

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

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

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

Determine the type of missing data and whether or not it can be ignored.

Determine the extent of missing data and decide whether respondents or variables should be deleted.

Diagnose the randomness of the missing data.

Select the imputation method for estimating missing data.

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

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

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

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

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

**Determine the best method of data transformation given a specific problem.** When the statistical assumptions are not met, it is not necessarily a “fatal” problem that prevents further analysis. Instead, the researcher may be able to apply any number of transformations to the data in question that will solve the problem and enable the assumptions to be met. Data transformations provide a means of modifying variables for one of two reasons: (1) to correct violations of the statistical assumptions underlying the multivariate techniques, or (2) to improve the relationship (correlation) between variables. Most of the transformations involve modifying one or more variables (e.g., compute the square root, logarithm, or inverse) and then using the transformed value in the analysis. It should be noted that the underlying data are still intact, just their distributional character is changed so as to meet the necessary statistical assumptions.

**Understand how to incorporate nonmetric variables as metric variables.** An important consideration in choosing and applying the correct multivariate technique is the measurement properties of the dependent and independent variables. Some of the techniques, such as discriminant analysis or multivariate analysis of variance, specifically require nonmetric data as dependent or independent variables. In many instances, the multivariate methods require that metric variables be used. Yet nonmetric variables are often of considerable interest to the researcher in a particular analysis. A method is available to represent a nonmetric variable with a set of dichotomous variables, known as dummy variables, so that it may be included in many of the analyses requiring only metric variables. A dummy variable is a dichotomous variable that has been converted to a metric distribution and represents one category of a nonmetric independent variable.

Considerable time and effort can be expended in these activities, but the prudent researcher wisely invests the necessary resources to thoroughly examine the data to ensure that the multivariate methods are applied in appropriate situations and to assist in a more thorough and insightful interpretation of the results.

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

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

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

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

Describe the conditions under which a researcher would delete a case with missing data versus the conditions under which a researcher would use an imputation method.

Evaluate the following statement: In order to run most multivariate analyses, it is not necessary to meet all the assumptions of normality, linearity, homoscedasticity, and independence.

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

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 Aguinis, H., R. K. Gottfredson, and H. Joo. 2013. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods* 16: 270–301.
- 2 Allison, P. D. 2002. *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences No. 07-136. Thousand Oaks, CA: Sage.
- 3 Allison, P. D. 2012. Handling Missing Data by Maximum Likelihood. In SAS Global Forum, Vol. 23. Haverford, PA: Statistical Horizons.
- 4 Allison, Paul D. 2009. Missing Data. Chapter 4 in R. E. Millsap and A. Maydeu-Olivares, *The SAGE Handbook of Quantitative Methods in Psychology*. London: Sage.
- 5 Anderson, Edgar. 1969. A Semigraphical Method for the Analysis of Complex Problems. *Technometrics* 2: 387–91.
- 6 Arbuckle, J. 1996. Full Information Estimation in the Presence of Incomplete Data. In *Advanced Structural Equation Modeling: Issues and Techniques*, G. A. Marcoulides and R. E. Schumacher (eds.). Mahwah, NJ: Lawrence Erlbaum Associates.
- 7 Bakker, M., and J. M. Wicherts. 2014. Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods* 19: 409–27.
- 8 Barton, R. R. 2013. Response Surface Methodology. In *Encyclopedia of Operations Research and Management Science*. New York: Springer, pp. 1307–13.
- 9 Batista, G. E., and M. C. Monard. 2003. An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence* 17: 519–33.
- 10 Bezerra, M. A., R. E. Santelli, E. P. Oliveira, L. S. Villar, and L. A. Escalera. 2008. Response Surface Methodology (RSM) as a Tool for Optimization in Analytical Chemistry. *Talanta* 76: 965–77.
- 11 Bishara, A. J., and J. B. Hittner. 2012. Testing the Significance of a Correlation with Nonnormal Data: Comparison of Pearson, Spearman, Transformation, and Resampling Approaches. *Psychological Methods* 17: 399–417.
- 12 Box, G. E. P., and D. R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society B* 26: 211–43.
- 13 Brown, R. L. 1994. Efficacy of the Indirect Approach for Estimating Structural Equation Models with Missing Data: A Comparison of Five Methods. *Structural Equation Modeling* 1: 287–316.
- 14 Castanedo, F. 2013. A Review of Data Fusion Techniques. *Scientific World Journal* 704504.
- 15 Cheema, J. R. 2014. A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research* 84: 487–508.
- 16 Chen, H., R. H. Chiang, and V. C. Storey. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36: 1165–88.
- 17 Chernoff, Herman. 1978. Graphical Representation as a Discipline. In *Graphical Representation of Multivariate Data*, Peter C. C. Wang (ed.). New York: Academic Press, pp. 1–11.
- 18 Cheung, M. W. L. 2006. Recovering Preipsative Information from Additive Ipsatized Data: A Factor Score Approach. *Educational and Psychological Measurement* 66: 565–88.
- 19 Cody, R. (2017). *Cody's Data Cleaning Techniques Using SAS*. Cary, NC: SAS Institute.
- 20 Cohen, Jacob, Stephen G. West, Leona Aiken, and Patricia Cohen. 2002. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 21 Collins, L. M., J. L. Schafer, and C. M. Kam. 2001. A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods* 6: 330–51.
- 22 Cox, B. E., K. McIntosh, R. D. Reason, and P. T. Terenzini. 2014. Working with Missing Data in Higher Education Research: A Primer and Real-World Example. *Review of Higher Education* 37: 377–402.
- 23 Cunningham, W. H., I. C. Cunningham, and R. T. Green. 1977. The Ipsative Process to Reduce Response Set Bias. *Public Opinion Quarterly* 41: 379–84.
- 24 Dahl, A., V. Iotchkova, A. Baud, Å. Johansson, U. Gyllensten, N. Soranzo, and J. Marchini. 2016. A Multiple Phenotype Imputation Method for Genetic Studies. *Nature Genetics* 48: 466–72.
- 25 Dalal, D. K., and M. J. Zickar. 2012. Some Common Myths About Centering Predictor Variables in Moderated Multiple Regression and Polynomial Regression. *Organizational Research Methods* 15: 339–62.
- 26 Daniel, C., and F. S. Wood. 1999. *Fitting Equations to Data*, 2nd edn. New York: Wiley-Interscience.
- 27 Dempster, A. P., and D. B. Rubin. 1983. Overview. In Madow, Olkin, and Rubin (eds.), *Incomplete Data in Sample Surveys: Theory and Annotated Bibliography*, Vol. 2, New York: Academic Press.

- 28 Duncan, T. E., R. Omen, and S. C. Duncan. 1994. Modeling Incomplete Data in Exercise Behavior Using Structural Equation Methodology. *Journal of Sport and Exercise Psychology* 16: 187–205.
- 29 Dunning, T., and E. Friedman. 2014. *Practical Machine Learning: A New Look at Anomaly Detection*. Sebastopol, CA: O'Reilly Media.
- 30 Edwards, J. R. 2001. Ten Difference Score Myths. *Organizational Research Methods* 4: 265–87.
- 31 Enders, C. K. 2010. *Applied Missing Data Analysis*. New York: Guilford Press.
- 32 Enders, C. K. 2017. Multiple Imputation as a Flexible Tool for Missing Data Handling in Clinical Research. *Behaviour Research and Therapy* 98: 4–18.
- 33 Enders, C. K., and D. Tofighi. 2007. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods* 12: 121–38.
- 34 Feinberg, Stephen. 1979. Graphical Methods in Statistics. *American Statistician* 33: 165–78.
- 35 Fichman, M., and J. N. Cummings. 2003. Multiple Imputation for Missing Data: Making the Most of What You Know. *Organizational Research Methods* 6: 282–308.
- 36 Gnanadesikan, R. 1977. *Methods for Statistical Analysis of Multivariate Distributions*. New York: Wiley.
- 37 Graham, J. W. 2009. Missing Data Analysis: Making it Work in the Real World. *Annual Review of Psychology* 60: 549–76.
- 38 Graham, J. W., A. E. Olchowski, and T. D. Gilreath. 2007. How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science* 8: 206–13.
- 39 Graham, J. W., and S. W. Donaldson. 1993. Evaluating Interventions with Differential Attrition: The Importance of Nonresponse Mechanisms and Use of Follow-up Data. *Journal of Applied Psychology* 78: 119–28.
- 40 Graham, J. W., S. M. Hofer, and D. P. MacKinnon. 1996. Maximizing the Usefulness of Data Obtained with Planned Missing Value Patterns: An Application of Maximum Likelihood Procedures. *Multivariate Behavioral Research* 31: 197–218.
- 41 Hayes, T., and J. J. McArdle. 2017. Evaluating the Performance of CART-Based Missing Data Methods Under a Missing Not at Random Mechanism. *Multivariate Behavioral Research* 52: 113–4.
- 42 Heitjan, D. F. 1997. Annotation: What Can Be Done About Missing Data? Approaches to Imputation. *American Journal of Public Health* 87: 548–50.
- 43 Hertel, B. R. 1976. Minimizing Error Variance Introduced by Missing Data Routines in Survey Analysis. *Sociological Methods and Research* 4: 459–74.
- 44 Hill, R. C., W. E. Griffiths, and G. C. Lim. 2018. *Principles of Econometrics*, 5th edn. Hoboken, NJ: Wiley.
- 45 Hoeffding, W. 1948. A Non-parametric Test of Independence. *Annals of Mathematical Statistics* 19: 546–57.
- 46 Howard, W. J., M. Rhemtulla, and T. D. Little. 2015. Using Principal Components as Auxiliary Variables in Missing Data Estimation. *Multivariate Behavioral Research* 50: 285–99.
- 47 Huff, D. L., and R. R. Batsell. 1977. Delimiting the Areal Extent of a Market Area. *Journal of Marketing Research* 14: 581–5.
- 48 Iacobucci, D., S. S. Posavac F. R., Kardes, M. Schneider, and D. Popovich. 2015. Toward a more nuanced understanding of the statistical properties of a median split. *Journal of Consumer Psychology* 25: 652–65.
- 49 Jayawardene, V., S. Sadiq, and M. Indulska. 2013a. *An Analysis of Data Quality Dimensions*. ITEE Technical Report, School of Information Technology and Electrical Engineering, the University of Queensland.
- 50 Johnson, R. A., and D. W. Wichern. 2002. *Applied Multivariate Statistical Analysis*, 5th edn. Upper Saddle River, NJ: Prentice Hall.
- 51 Keller, S., G. Korkmaz, M. Orr, A. Schroeder, and S. Shipp. 2017. The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches. *Annual Review of Statistics and Its Application* 4: 85–108.
- 52 Khuri, A. I., and S. Mukhopadhyay. 2010. Response Surface Methodology. *Wiley Interdisciplinary Reviews: Computational Statistics* 2: 128–49.
- 53 Kim, J. O., and J. Curry. 1977. The Treatment of Missing Data in Multivariate Analysis. *Sociological Methods and Research* 6: 215–41.
- 54 Kinney, J. B., and G. S. Atwal. 2014. Equitability, Mutual Information, and the Maximal Information Coefficient. *Proceedings of the National Academy of Sciences* 111: 3354–9.
- 55 Kromrey, J. D., and L. Foster-Johnson. 1998. Mean Centering in Moderated Multiple Regression: Much Ado About Nothing. *Educational and Psychological Measurement* 58: 42–67.
- 56 Lemieux, J., and L. McAlister. 2005. Handling Missing Values in Marketing Data: A Comparison of Techniques. *MSI Reports* 2: 41–60.
- 57 Li, P., E. A. Stuart, and D. B. Allison. 2015. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA* 314: 1966–67.
- 58 Little, R. J. A. 1988. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* 83: 1198–1202.
- 59 Little, R. J., and D. B. Rubin. 2014. *Statistical Analysis with Missing Data*. New York: Wiley.
- 60 Little, Todd D., Terrence D. Jorgensen, Kyle M. Lang, E. Whitney, and G. Moore. 2014. On the Joys of Missing Data. *Journal of Pediatric Psychology* 39: 151–62.
- 61 MacCallum, R. C., S. Zhang, K. J. Preacher, and D. D. Rucker. 2002. On the Practice of Dichotomization of Quantitative Variables. *Psychological Methods* 7: 19–40.
- 62 Malhotra, N. K. 1987. Analyzing Marketing Research Data with Incomplete Information on the Dependent Variables. *Journal of Marketing Research* 24: 74–84.
- 63 McCallum, Q. 2012. *Bad Data Handbook*. Sebastopol, CA: O'Reilly Media.
- 64 Meade, A. W., and S. B. Craig. 2012. Identifying carelessness responses in survey data. *Psychological Methods* 17: 437–55.

- 65 Myers, T. A. 2011. Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data. *Communication Methods and Measures* 5: 297–310.
- 66 Newman, D. A. 2014. Missing Data: Five Practical Guidelines. *Organizational Research Methods* 17: 372–411.
- 67 Newman, D. A., and J. M. Cottrell. 2014. Missing Data Bias: Exactly How Bad Is Pairwise Deletion? In C. E. Lance and R. J. Vandenberg (eds.), *More Statistical and Methodological Myths and Urban Legends*. New York: Routledge.
- 68 Preacher, K. J., D. D. Rucker, R. C. MacCallum, and W. A. Nicewander. 2005. Use of the Extreme Groups Approach: A Critical Reexamination and New Recommendations. *Psychological Methods* 10: 178–92.
- 69 Rässler, S. 2016. Data Fusion: Identification Problems, Validity, and Multiple Imputation. *Austrian Journal of Statistics* 33: 153–71.
- 70 Raymonds, M. R., and D. M. Roberts. 1987. A Comparison of Methods for Treating Incomplete Data in Selection Research. *Educational and Psychological Measurement* 47: 13–26.
- 71 Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. 2011. Detecting Novel Associations in Large Data Sets. *Science* 334: 1518–24.
- 72 Robinson, C., and R. E. Schumacker. 2009. Interaction Effects: Centering, Variance Inflation Factor, and Interpretation Issues. *Multiple Linear Regression Viewpoints* 35: 6–11.
- 73 Roth, P. L. 1994. Missing Data: A Conceptual Review for Applied Psychologists. *Personnel Psychology* 47: 537–60.
- 74 Rubin, Donald B. 1976. Inference and Missing Data. *Biometrika* 63: 581–92.
- 75 Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- 76 Rucker, D. D., B. B. McShane, and K. J. Preacher. 2015. A Researcher's Guide to Regression, Discretization, and Median Splits of Continuous Variables. *Journal of Consumer Psychology* 25: 666–78.
- 77 Saha, B., and D. Srivastava. 2014. Data Quality: The Other Face of Big Data. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference*, New York: IEEE, pp. 1294–7.
- 78 Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
- 79 Schafer, J. L., and J. W. Graham. 2002. Missing Data: Our View of the State of the Art. *Psychological Methods* 7: 147–77.
- 80 Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
- 81 Shanock, L. R., B. E. Baran, W. A. Gentry, S. C. Pattison, and E. D. Heggstad. 2010. Polynomial Regression with Response Surface Analysis: A Powerful Approach for Examining Moderation and Overcoming Limitations of Difference Scores. *Journal of Business and Psychology* 25: 543–54.
- 82 Silva, L. O., and L. E. Zárate. 2014. A Brief Review of the Main Approaches for Treatment of Missing Data. *Intelligent Data Analysis* 18: 1177–98.
- 83 Stevens, J. 2001. *Applied Multivariate Statistics for the Social Sciences*, 4th edn. Hillsdale, NJ: Lawrence Erlbaum Publishing.
- 84 Székely, G. J., M. L. Rizzo, and N. K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *Annals of Statistics* 35: 2769–94.
- 85 Thoemmes, F., and N. Rose. 2014. A Cautious Note on Auxiliary Variables That Can Increase Bias in Missing Data Problems. *Multivariate Behavioral Research* 49: 443–59.
- 86 Twala, B., and M. Cartwright. 2010. Ensemble Missing Data Techniques for Software Effort Prediction. *Intelligent Data Analysis* 14: 299–331.
- 87 Wang, Peter C. C. (ed.). 1978. *Graphical Representation of Multivariate Data*. New York: Academic Press.
- 88 Weisberg, S. 1985. *Applied Linear Regression*. New York: Wiley.
- 89 Wessling, K. S., J. Huber, and O. Netzer. 2017. MTurk Character Misrepresentation: Assessment and Solutions. *Journal of Consumer Research* 44: 211–30.
- 90 White, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica: Journal of the Econometric Society* 48: 817–38.
- 91 Wilkinson, L. 1982. An Experimental Evaluation of Multivariate Graphical Point Representations. In *Human Factors in Computer Systems: Proceedings*. New York: ACM Press, pp. 202–9.
- 92 Xiao, Y., R. Song, M. Chen, and H. I. Hall. 2012. Direct and Unbiased Multiple Imputation Methods for Missing Values of Categorical Variables. *Journal of Data Science* 10: 465–81.
- 93 Zhou, X. H., C. Zhou, D. Lui, and X. Ding. 2014. *Applied Missing Data Analysis in the Health Sciences*. New York: Wiley.
- 94 Zijlstra, W. P., L. A. Van Der Ark, and K. Sijtsma. 2007. Outlier Detection in Test and Questionnaire Data. *Multivariate Behavioral Research* 42: 531–55.

# Interdependence techniques

Exploratory Factor Analysis

Cluster Analysis

## SECTION II

### OVERVIEW

The dependence methods described in Sections 3 and 4 provide the researcher with several methods for assessing relationships between one or more dependent variables and a set of independent variables. Many methods are discussed that accommodate all types (metric and nonmetric) and potentially large numbers of both dependent and independent variables that could be applied to sets of observations. Yet what if the variables or observations are related in ways not captured by the dependence relationships? What if the assessment of interdependence (i.e., structure) is missing? One of the most basic abilities of human beings is to classify and categorize objects and information into simpler schema, such that we can characterize the objects within the groups in total rather than having to deal with each individual object. The objective of the methods in this section is to identify the structure among a defined set of variables, observations, or objects. The identification of structure offers not only simplicity, but also a means of description and even discovery.

Interdependence techniques, however, are focused solely on the definition of structure, assessing interdependence without any associated dependence relationships. None of the interdependence techniques will define structure to optimize or maximize a dependence relationship. It is the researcher's task to first utilize

these methods in identifying structure and then to employ it where appropriate. The objectives of dependence relationships are not “mixed” in these interdependence methods—they assess structure for its own sake and no other.

## CHAPTERS IN SECTION II

Section 2 is comprised of two chapters, which cover techniques for assessing structure among variables and cases/objects. The first interdependence technique, Exploratory Factor Analysis (Chapter 3) provides us with a tool for understanding the relationships among variables, a knowledge fundamental to all of our multivariate analyses. The issues of multicollinearity and model parsimony are reflective of the underlying structure of the variables, and factor analysis provides an objective means of assessing the groupings of variables and the ability to incorporate composite variables reflecting these variable groupings into other multivariate techniques.

It is not only variables that have structure, however. Although we assume independence among the observations and variables in our estimation of relationships, we also know that most populations have subgroups sharing general characteristics. Marketers look for target markets of differentiated groups of homogeneous consumers, strategy researchers look for groups of similar firms to identify common strategic elements, and financial modelers look for stocks with similar fundamentals to create stock portfolios. These and many other situations require techniques that find these groups of similar objects based on a set of characteristics.

This goal is met by cluster analysis, the topic of Chapter 4. Cluster analysis is ideally suited for defining groups of objects with maximal similarity within the groups while also having maximum heterogeneity between the groups—determining the most similar groups that are also most different from each other. As we show, cluster analysis has a rich tradition of application in almost every area of inquiry. Yet, its ability to define groups of similar objects is countered by its rather subjective nature and the instrumental role played by the researcher’s judgment in several key decisions. This subjective aspect does not reduce the usefulness of the technique, but it does place a greater burden on the researcher to fully understand the technique and the impact of certain decisions on the ultimate cluster solution.

These techniques provide the researcher with methods that bring order to the data in the form of structure among the observations or variables. In this way, the researcher can better understand the basic structures of the data, not only facilitating the description of the data, but also providing a foundation for a more refined analysis of the dependence relationships.

# 3 Exploratory Factor Analysis

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

- Differentiate exploratory factor analysis techniques from other multivariate techniques.
- Distinguish between exploratory and confirmatory uses of factor analytic techniques.
- Understand the seven stages of applying exploratory factor analysis.
- Distinguish between *R* and *Q* factor analysis.
- Identify the differences between principal component analysis and common factor analysis models.
- Describe how to determine the number of factors to extract.
- Explain the concept of rotation of factors.
- Describe how to name a factor.
- Explain the additional uses of exploratory factor analysis.
- State the major limitations of exploratory factor analytic techniques.

## Chapter Preview

Use of the multivariate statistical technique of exploratory factor analysis increased during the past decade in all fields of social sciences research, particularly business. Yet the technique gained even greater use in the era of “Big Data” outside of academia as analysts in every sector of the economy are faced with an explosion in the number of variables being analyzed (see more detailed discussion in Chapter 1). This increase in the number of variables to be considered increases the need for improved knowledge of the structure and interrelationships of variables. This chapter describes exploratory factor analysis, a technique particularly suitable for analyzing the patterns of complex, multidimensional relationships encountered by researchers. It defines and explains in broad, conceptual terms the fundamental aspects of factor-analytic techniques. Exploratory factor analysis can be utilized to examine the underlying patterns or relationships for a large number of variables and to determine whether the information can be condensed or summarized in a smaller set of factors or components. To further clarify the methodological concepts, basic guidelines for presenting and interpreting the results of these techniques are also included.

## Key Terms

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

**1 – R<sup>2</sup>ratio** Diagnostic measure employed in *variable clustering* to assess whether variables are singularly represented by a *cluster component* or have a substantive *cross-loading*.

**Anti-image correlation matrix** Matrix of the partial correlations among variables after factor analysis, representing the degree to which the factors explain each other in the results. The diagonal contains the *measures of sampling adequacy* for each variable, and the off-diagonal values are partial correlations among variables.

**A priori criterion** A stopping rule for determining the number of factors. This rule is determined solely on the researcher's judgment and experience. The researcher may know the desired structure, be testing a specific structure or other conceptually-based considerations so that the number of factors can be predetermined.

**Bartlett test of sphericity** Statistical test for the overall significance of all correlations within a correlation matrix.

**Cluster analysis** Multivariate technique with the objective of grouping respondents or cases with similar profiles on a defined set of characteristics. Similar to *Q factor analysis*.

**Cluster component** A principal component extracted in *variable clustering* which contains only a subset of the complete variable set.

**Common factor analysis** Factor model in which the factors are based on a reduced correlation matrix. That is, *communalities* are inserted in the diagonal of the *correlation matrix*, and the extracted factors are based only on the *common variance*, with *specific* and *error variance* excluded.

**Common variance** Variance shared with other variables in the factor analysis (i.e., shared variance represented by the squared correlation).

**Commonality** See *communality*.

**Communality** Total amount of variance an original variable shares with all other variables included in the factor analysis. Calculated as the sum of the squared loadings for a variable across the factors.

**Component** See *factor*.

**Component analysis** See *principal component analysis*.

**Composite measure** See *summated scales*.

**Conceptual definition** Specification of the theoretical basis for a concept that is represented by a factor.

**Confirmatory approach** An approach to factor analysis, typically associated with structural equation modeling, that assesses the extent to which a pre-defined structure fits the data. This approach contrasts with an *exploratory approach*, which is data driven and the analysis reveals the structure.

**Convergent validity** The degree to which two measures (scales) of the same concept are correlated. One aspect of construct validity.

**Construct validity** Broad approach to ensure the validity of a set of items as representative of a *conceptual definition*. Includes specific sub-elements of *convergent validity*, *discriminant validity* and *nomological validity*.

**Content validity** Assessment of the degree of correspondence between the items selected to constitute a *summated scale* and its *conceptual definition*.

**Correlation matrix** Table showing the intercorrelations among all variables.

**Cronbach's alpha** Measure of *reliability* that ranges from 0 to 1, with values of .60 to .70 deemed the lower limit of acceptability.

**Cross-loading** A variable has two or more *factor loadings* exceeding the threshold value deemed necessary for significance in the factor interpretation process.

**Discriminant validity** One element of *construct validity* focusing on the degree to which two concepts are distinct. Every scale in the analysis must be shown to have discriminant validity from all other scales.

**Dummy variable** Binary metric variable used to represent a single category of a nonmetric variable.

**Eigenvalue** Represents the amount of variance accounted for by a factor. Calculated as the column sum of squared loadings for a factor; also referred to as the *latent root*.

**EQUIMAX** One of the *orthogonal factor rotation* methods that is a "compromise" between the VARIMAX and QUARTIMAX approaches, but is not widely used.

**Error variance** Variance of a variable due to *measurement error* (e.g., errors in data collection or measurement).

**Exploratory approach** An approach to factor analysis in which the objective is to define the structure within a set of variables, with no pre-specification of number of factors or which variables are part of a factor. Contrasted to a *confirmatory approach* where the structure is pre-defined.

**Face validity** See *content validity*.

**Factor** Linear combination (variate) of the original variables. Factors also represent the underlying dimensions (constructs) that summarize or account for the original set of observed variables.

**Factor indeterminacy** Characteristic of *common factor analysis* such that several different *factor scores* can be calculated for a respondent, each fitting the estimated factor model. It means the factor scores are not unique for each individual.

**Factor loadings** Correlation between the original variables and the factors, and the key to understanding the nature of a particular factor. Squared factor loadings indicate what percentage of the variance in an original variable is explained by a factor.

**Factor matrix** Table displaying the *factor loadings* of all variables on each factor.

**Factor pattern matrix** One of two factor matrices found in an *oblique rotation* that is most comparable to the *factor matrix* in an *orthogonal rotation*.

**Factor rotation** Process of manipulation or adjusting the factor axes to achieve a simpler and pragmatically more meaningful factor solution.

**Factor score** Composite measure created for each observation on each factor extracted in the factor analysis. The factor weights are used in conjunction with the original variable values to calculate each observation's score. The factor score then can be used to represent the factor(s) in subsequent analyses. Factor scores are standardized to have a mean of 0 and a standard deviation of 1. Similar in nature to a *summed scale*.

**Factor structure matrix** A *factor matrix* found in an *oblique rotation* that represents the simple correlations between variables and factors, incorporating the unique variance and the correlations between factors. Most researchers prefer to use the *factor pattern matrix* when interpreting an oblique solution.

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

**Item** See *indicator*.

**Kaiser rule** See *latent root criterion*.

**Latent root** See *eigenvalue*.

**Latent root criterion** One of the stopping rules to determine how many factors to retain in the analysis. In this rule, all factors with *eigenvalues* (*latent roots*) greater than 1.0 are retained.

**Measure of sampling adequacy (MSA)** Measure calculated both for the entire correlation matrix and each individual variable. MSA values above .50 for either the entire matrix or an individual variable indicate appropriateness for performing factor analysis on the overall set of variables or specific variables respectively.

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

**Multicollinearity** Extent to which a variable can be explained by the other variables in the analysis.

**Nomological validity** An element of *construct validity* focusing on the extent to which the scale makes accurate predictions of other concepts in a theoretically-based model.

**Oblique factor rotation** *Factor rotation* computed so that the extracted factors are correlated. Rather than arbitrarily constraining the factor rotation to an *orthogonal* solution, the oblique rotation identifies the extent to which each of the factors is correlated.

**Optimal scaling** Process of transforming nonmetric data (i.e., nominal and ordinal) to a form suitable for use in *principal component analysis*.

**Orthogonal** Mathematical independence (no correlation) of factor axes to each other (i.e., at right angles, or 90 degrees).

**Orthogonal factor rotation** *Factor rotation* in which the factors are extracted so that their axes are maintained at 90 degrees. Each factor is independent of, or *orthogonal* to, all other factors (i.e., correlation between the factors is constrained to be zero).

**Parallel analysis** A stopping rule based on comparing the factor *eigenvalues* to a set of eigenvalues generated from random data. The basic premise is to retain factors that have eigenvalues exceeding those which would be generated by random data.

**Percentage of variance criterion** A stopping rule for the number of factors to retain which is based on the amount of total variance accounted for in a set of factors, or the *communality* of each of the variables. The threshold value is specified by the researcher based on the objectives of the research and judgments about the quality of the data being analyzed.

**Principal component analysis** Factor model in which the factors are based on the total variance. With principal component analysis, unities (1s) are used in the diagonal of the *correlation matrix*; this procedure computationally implies that all the variance is common or shared.

**Q factor analysis** Forms groups of respondents or cases based on their similarity on a set of characteristics (also see the discussion of *cluster analysis* in Chapter 4).

**QUARTIMAX** A type of *orthogonal factor rotation* method focusing on simplifying the columns of a factor matrix. Generally considered less effective than the VARIMAX rotation.

**R factor analysis** Analyzes relationships among variables to identify groups of variables forming latent dimensions (factors).

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

**Reverse scoring** Process of reversing the scores of a variable, while retaining the distributional characteristics, to change the relationships (correlations) between two variables. Used in *summed scale* construction to avoid a canceling out between variables with positive and negative *factor loadings* on the same factor.

**Scale development** A specific process, usually involving both *exploratory* and *confirmatory factor analyses*, that attempts to define a set of variables which represent a concept that cannot be adequately measured by a single variable. For more details, see Chapters 9 and 10.

**Scoring procedure** Saves the scoring coefficients from the factor matrix and then allows them to be applied to new datasets to generate factor scores as a form of replication of the original results.

**Scree test** A stopping rule based on the pattern of eigenvalues of the extracted factors. A plot of the eigenvalues is examined to find an “elbow” in the pattern denoting subsequent factors that are not distinctive.

**Specific variance** Variance of each variable unique to that variable and not explained or associated (correlations) with other variables in the factor analysis. One portion of *unique variance*.

**Stopping rule** A criterion for determining the number of factors to retain in the final results, including the *latent root criterion*, *a priori criterion*, *percentage of variance criterion*, *scree test* and *parallel analysis*.

**Summed 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. In most instances, the separate variables are summed and then their total or average score is used in the analysis.

**Surrogate variable** Selection of a single proxy variable with the highest *factor loading* to represent a factor in the data reduction stage instead of using a *summed scale* or *factor score*.

**Trace** Represents the total amount of variance on which the factor solution is based. The trace is equal to the number of variables, based on the assumption that the variance in each variable is equal to 1.

**Unidimensional** A characteristic of the set of variables forming a *summed scale* where these variables are only correlated with the hypothesized *factor* (i.e., have a high *factor loading* only on this factor).

**Unique variance** Portion of a variable’s total variance that is not *shared variance* (i.e., not correlated with any other variables in the analysis). Has two portions—*specific variance* relating to the variance of the variable not related to any other variables and *error variance* attributable to the measurement errors in the variable’s value.

**Validity** Extent to which a single variable or set of variables (*construct validity*) 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 variable(s), whereas *reliability* relates to the consistency of the variables(s).

**Variable clustering** A variant of principal component analysis which estimates components with only subsets of the original variable set (*cluster components*). These cluster components are typically estimated in a hierarchical fashion by “splitting” a cluster component when two principal components can be extracted. This splitting process continues until some threshold is achieved.

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

**VARIMAX** The most popular *orthogonal factor rotation* method focusing on simplifying the columns in a *factor matrix*. Generally considered superior to other orthogonal factor rotation methods in achieving a simplified factor structure.

## What Is Exploratory Factor Analysis?

Exploratory factor analysis, often referred to as EFA, is an interdependence technique, as defined in Chapter 1, whose *primary purpose is to define the underlying structure among the variables in the analysis*. Obviously, variables play a key role in any multivariate analysis. Whether we are making a sales forecast with regression, predicting success or failure of a new firm with discriminant analysis, grouping observations with cluster analysis or using any of the other multivariate techniques discussed in Chapter 1, we must have a set of variables upon which to form relationships (e.g., What variables best predict sales or success/failure? How do we compare observations for grouping purposes?). As such, variables are the building blocks of relationships.

As we employ multivariate techniques, by their very nature the number of variables increases and the ability to incorporate multiple variables is one of their fundamental benefits. Univariate techniques are limited to a single variable, but multivariate techniques can have tens, hundreds, thousands, or even millions of variables. But how do we describe and represent all of these variables? Certainly, if we have only a few variables, they may all be distinct and different. As we add more and more variables, more and more overlap (i.e., correlation) is likely among the variables. In some instances, such as when we are using multiple measures to overcome measurement error by multi-variable measurement (see Chapter 1 for a more detailed discussion), the researcher even strives for correlations

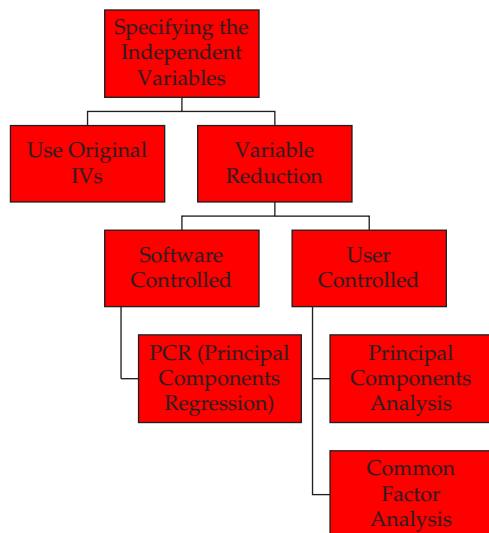
among the variables. As the variables become correlated, the researcher now needs ways in which to manage these variables—grouping highly correlated variables together, labeling or naming the groups, and perhaps even creating a new composite measure that can represent each group of variables.

These issues all impact the task of “Managing the Variate” as introduced in Chapter 1. In this regard, the researcher employs techniques to (a) reduce the number of variables to be analyzed, and (b) select the appropriate variables to be included in the multivariate analyses. Exploratory factor analysis plays a critical role in this first task—reducing the number of variables—by providing the analyst direct control over this process (see Figure 3.1). Various techniques (e.g., principal components regression) provide comparable approaches to data reduction, but EFA provides the only “user-controlled” method of achieving this objective.

We introduce exploratory factor analysis as our first multivariate technique because it can play a unique role in the application of other multivariate techniques. Broadly speaking, factor analysis provides the tools for analyzing the structure of the interrelationships (correlations) among a large number of variables (e.g., test scores or items, questionnaire responses, social media usage patterns, digital tracking) by defining sets of variables that are highly interrelated, known as **factors or components**. These groups of variables (factors or components), which are by definition highly intercorrelated, are assumed to represent dimensions within the data. If we are only concerned with reducing the number of variables, then the dimensions can guide in creating new composite measures. However, if we have a conceptual basis for understanding the relationships between variables, then the dimensions may actually have meaning for what they collectively represent. In the latter case, these dimensions may correspond to concepts that cannot be adequately described by a single measure (e.g., store atmosphere is defined by many sensory elements that must be measured separately but are all interrelated). Exploratory factor analysis presents several ways of representing these groups of variables for use in other multivariate techniques.

We should note at this point that factor-analytic techniques can achieve their purposes from either an exploratory or confirmatory perspective. A continuing debate concerns its appropriate role, where many researchers consider it only an **exploratory approach**, useful in searching for structure among a set of variables or as a data reduction method. In this perspective, exploratory factor analytic techniques “take what the data give you” and do not set any *a priori* constraints on the estimation of components or the number of components to be extracted. For many—if not most—applications, this use of factor analysis is appropriate. In other situations, however, the researcher has preconceived thoughts on the actual structure of the data, based on theoretical support or prior research. For example, the researcher may wish to test hypotheses involving issues such as which variables should be grouped together on a factor (i.e., multiple items were specifically collected to represent store image) or the

**Figure 3.1**  
Variable Reduction Methods in Managing the Variate



precise number of factors. In these instances, the researcher requires that factor analysis take a **confirmatory approach**—that is, assess the degree to which the data meet the expected theoretical structure. The methods we discuss in this chapter do not directly provide the necessary structure for formalized hypothesis testing. We explicitly address the confirmatory perspective of factor analysis in Chapter 9 and 10, and confirmatory composite analysis in Chapter 13. In this chapter, however, we view factor analytic techniques principally from an exploratory or non-confirmatory viewpoint.

## A Hypothetical Example of Exploratory Factor Analysis

Assume that through qualitative research a retail firm identified 80 different characteristics of retail stores and their service that consumers mentioned as influencing their patronage choice among stores. The retailer wants to understand how consumers make decisions but feels that it cannot evaluate 80 separate characteristics or develop action plans for this many variables, because they are too specific. At the same time, the problem is particularly complex since the retailer must understand both in-store and online issues. As a result, the retailer would like to know whether consumers think in more general evaluative dimensions rather than in just the specific items. For example, consumers may consider salespersons to be a more general evaluative dimension that is composed of many more specific characteristics, such as knowledge, courtesy, likeability, sensitivity, friendliness, helpfulness, and so on. At the same time, in-store dimensions are likely to be distinct from online dimensions.

To identify these broader dimensions, the retailer could conduct a survey asking for consumer evaluations on each of the 80 specific items. Exploratory factor analysis would then be used to identify the broader underlying evaluative dimensions. Specific items that correlate highly are assumed to be a member of that broader dimension. These dimensions form the basis for creating composites of specific variables, which in turn allow the dimensions to be interpreted and described. In our example, the exploratory factor analysis might identify such dimensions as product assortment, product quality, prices, store personnel, service, website ease of use, and store atmosphere as the broader evaluative dimensions used by the respondents. Each of these dimensions contains specific items that are a facet of the broader evaluative dimension. From these findings, the retailer may then use the dimensions (factors) to define broad areas for planning and action.

An illustrative example of an application of factor analysis is shown in Figure 3.2, which represents the correlation matrix for nine store image elements. Included in this set are measures of the product offering, store personnel, price levels, and in-store service and experiences. The question a researcher may wish to address is: Are all of these elements separate in their evaluative properties or do they group into some more general areas of evaluation? For example, do all of the product elements group together? Where does price level fit, or is it separate? Do the in-store features (e.g., store personnel, service, and atmosphere) relate to one another? Visual inspection of the original correlation matrix (Figure 3.2, Part 1) does not easily reveal any specific pattern. Among scattered high correlations, variable groupings are not apparent. The application of factor analysis results in the grouping of variables as reflected in Part 2 of Figure 3.2. Here some interesting patterns emerge. First, four variables all relating to the in-store experience of shoppers are grouped together. Then, three variables describing the product assortment and availability are grouped together. Finally, product quality and price levels are grouped. Each group represents a set of highly interrelated variables that may reflect a more general evaluative dimension. In this case, we might label the three variable groupings by the labels in-store experience, product offerings, and value.

This example of exploratory factor analysis demonstrates its basic objective of grouping highly intercorrelated variables into distinct sets (factors). In many situations, these factors can provide a wealth of information about the interrelationships of the variables. In this example, exploratory factor analysis identified for store management a smaller set of concepts to consider in any strategic or tactical marketing plans, while still providing insight into what constitutes each general area (i.e., the individual variables defining each factor).

**Figure 3.2**

**Illustrative Example of the Use of Exploratory Factor Analysis to Identify Structure Within a Group of Variables**

**PART 1: ORIGINAL CORRELATION MATRIX**

	<b>V<sub>1</sub></b>	<b>V<sub>2</sub></b>	<b>V<sub>3</sub></b>	<b>V<sub>4</sub></b>	<b>V<sub>5</sub></b>	<b>V<sub>6</sub></b>	<b>V<sub>7</sub></b>	<b>V<sub>8</sub></b>	<b>V<sub>9</sub></b>
<i>V<sub>1</sub></i> Price Level	1.000								
<i>V<sub>2</sub></i> Store Personnel	.427	1.000							
<i>V<sub>3</sub></i> Return Policy	.302	.771	1.000						
<i>V<sub>4</sub></i> Product Availability	.470	.497	.427	1.000					
<i>V<sub>5</sub></i> Product Quality	.765	.406	.307	.427	1.000				
<i>V<sub>6</sub></i> Assortment Depth	.281	.445	.423	.713	.325	1.000			
<i>V<sub>7</sub></i> Assortment Width	.345	.490	.471	.719	.378	.724	1.000		
<i>V<sub>8</sub></i> In-store Service	.242	.719	.733	.428	.240	.311	.435	1.000	
<i>V<sub>9</sub></i> Store Atmosphere	.372	.737	.774	.479	.326	.429	.466	.710	1.000

**PART 2: CORRELATION MATRIX OF VARIABLES AFTER GROUPING ACCORDING TO FACTOR ANALYSIS**

	<b>V<sub>3</sub></b>	<b>V<sub>8</sub></b>	<b>V<sub>9</sub></b>	<b>V<sub>2</sub></b>	<b>V<sub>6</sub></b>	<b>V<sub>7</sub></b>	<b>V<sub>4</sub></b>	<b>V<sub>1</sub></b>	<b>V<sub>5</sub></b>
<i>V<sub>3</sub></i> Return Policy	1.000								
<i>V<sub>8</sub></i> In-store Service	.773	1.000							
<i>V<sub>9</sub></i> Store Atmosphere	.771	.710	1.000						
<i>V<sub>2</sub></i> Store Personnel	.771	.719	.737	1.000					
<i>V<sub>6</sub></i> Assortment Depth	.423	.311	.429	.445	1.000				
<i>V<sub>7</sub></i> Assortment Width	.471	.435	.466	.490	.724	1.000			
<i>V<sub>4</sub></i> Product Availability	.427	.428	.479	.497	.713	.719	1.000		
<i>V<sub>1</sub></i> Price Level	.302	.242	.372	.427	.281	.354	.470	1.000	
<i>V<sub>5</sub></i> Product Quality	.307	.240	.326	.406	.325	.378	.427	.765	1.000

## Factor Analysis Decision Process

We center the discussion of exploratory factor analysis on the six-stage model-building paradigm introduced in Chapter 1. Figure 3.3 shows the first three stages of the structured approach to multivariate model building, and Figure 3.5 details the final three stages, plus an additional stage (stage 7) beyond the estimation, interpretation, and validation of the factor models, which aids in selecting surrogate variables, computing factor scores, or creating summated scales for use in other multivariate techniques. A discussion of each stage follows.

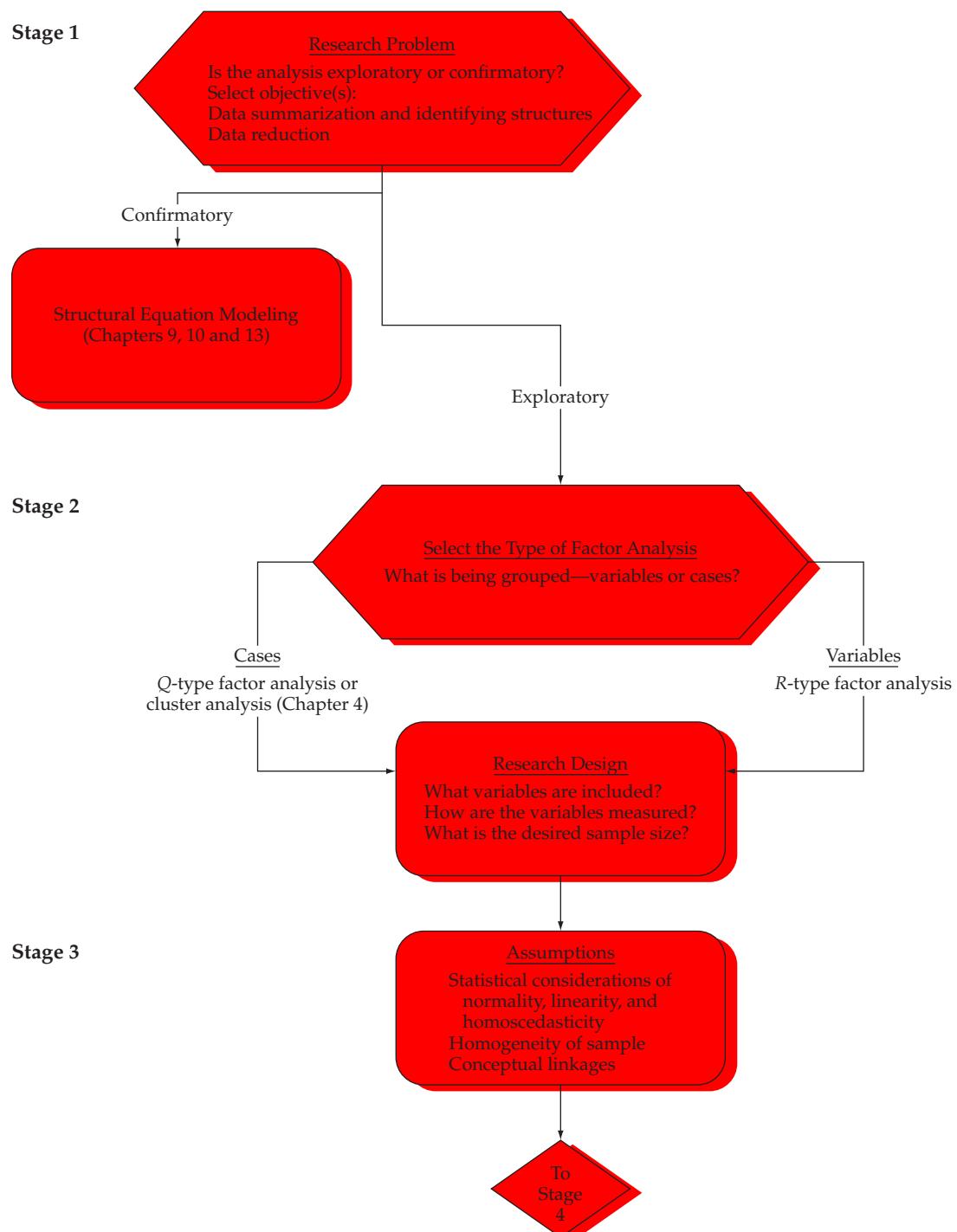
### Stage 1: Objectives of Factor Analysis

The starting point in factor analysis, as with other statistical techniques, is the research problem. The general purpose of exploratory factor analytic techniques is to find a way to condense (summarize) the information contained in a number of original variables into a smaller set of new, composite dimensions or variates (factors) with a minimum loss of information—that is, to search for and define the fundamental constructs or dimensions assumed to underlie the original variables [25, 49]. In meeting its objectives, exploratory factor analysis focuses on four issues: specifying the unit of analysis, achieving data summarization and/or data reduction, variable selection, and using exploratory factor analysis results with other multivariate techniques.

#### SPECIFYING THE UNIT OF ANALYSIS

Up to now, we have defined exploratory factor analysis solely in terms of identifying structure among a set of variables. Exploratory factor analysis is actually a more general model in that it can identify the structure of relationships

**Figure 3.3**  
Stages 1–3 in the Exploratory Factor Analysis Decision Diagram



among either variables or cases (e.g., respondents) by examining either the correlations between the variables or the correlations between the cases.

- If the objective of the research is to summarize the characteristics (i.e., variables), factor analysis would be applied to a **correlation matrix** of the variables. This most common type of factor analysis, referred to as **R factor analysis**, analyzes a set of variables to identify the dimensions for the variables that are latent (not easily observed).
- Exploratory factor analysis also can be applied to a correlation matrix of the individual cases based on their characteristics. Referred to as **Q factor analysis**, this method combines or condenses large numbers of cases into distinctly different groups within a larger population. The Q factor analysis approach is not utilized frequently because of computational difficulties. Instead, most researchers utilize some type of **cluster analysis** (see Chapter 4) to group individual cases. Also see Stewart [55] for other possible combinations of groups and variable types.

Thus, the researcher must first select the unit of analysis for factor analysis: variables or cases. Even though we will focus primarily on structuring variables, the option of employing factor analysis among cases as an alternative to cluster analysis is also available. The implications in terms of identifying similar variables or cases will be discussed in stage 2 when the correlation matrix is defined.

## ACHIEVING DATA SUMMARIZATION VERSUS DATA REDUCTION

Exploratory factor analysis provides the researcher with two distinct, but complementary, outcomes: data summarization and data reduction. In summarizing the data, exploratory factor analysis derives underlying dimensions that describe the data in a much smaller number of concepts than the original individual variables. The researcher may extend this process by modifying the original dimensions (factors) to obtain more interpretable and better understood factors. Data reduction extends this process by deriving an empirical value (factor or summated scale score) for each dimension (factor) and then substituting this value for the original values.

**Data Summarization** The fundamental concept involved in data summarization is the definition of *structure*. Through structure, the researcher can view the set of variables at various levels of generalization, ranging from the most detailed level (individual variables themselves) to the more generalized level, where individual variables are grouped and then viewed not for what they represent individually, but for what they represent collectively in expressing a concept. As we will discuss below, this stage can occur either with a focus on interpretability or without any attempt at interpretation. But before discussing these two approaches, let us consider exactly what exploratory factor analysis attempts to accomplish.

Exploratory factor analysis, as an interdependence technique, differs from the dependence techniques discussed in the next sections (i.e., multiple regression, discriminant analysis, multivariate analysis of variance, or logistic regression) where one or more variables are explicitly considered the criterion or dependent variable(s) and all others are the predictor or independent variables. In exploratory factor analysis, all variables are simultaneously considered with no distinction as to dependent or independent variables. Factor analysis still employs the concept of the **variate**, the linear composite of variables, but in exploratory factor analysis, the variates (factors) are formed to maximize their explanation of the entire variable set, not to predict a dependent variable(s). The goal of data summarization is achieved by defining a small number of factors that adequately represent the original set of variables.

If we were to draw an analogy to dependence techniques, it would be that each of the observed (original) variables is a dependent variable that is a function of some underlying and latent set of factors (dimensions) that are themselves made up of all other variables. Thus, each variable is predicted by all of the factors, and indirectly by all the other variables. Conversely, one can look at each factor (variate) as a dependent variable that is a function of the entire set of observed variables. Either analogy illustrates the differences in purpose between dependence (prediction) and interdependence (identification of structure) techniques. Structure is defined by the interrelatedness among variables allowing for the specification of a smaller number of dimensions (factors) representing the original set of variables.

**DATA SUMMARIZATION WITHOUT INTERPRETATION** In the most basic form, exploratory factor analysis which are based solely on the intercorrelations among variables, with no regard as to whether they represent interpretable dimensions or not. Such methods as principal components regression employ this procedure strictly to reduce the number of variables in the analysis with no specific regard as to what these dimensions represent. This is not to detract from this valuable benefit which can ultimately represent a large number of variables in a much smaller set of dimensions. In doing so, exploratory factor analysis provides a much more parsimonious set of variables for the analysis which can aid in model development and estimation.

**DATA SUMMARIZATION WITH INTERPRETATION** Yet in many other instances, researchers may wish to interpret and label the dimensions for managerial purposes or even utilize the procedure to assist in scale development. **Scale development** is a specific process focused on the identification of a set of items that represent a construct (e.g., store image) in a quantifiable and objective manner. As will be discussed extensively in later sections, researchers can examine each dimension and “fine tune” the variables included on that dimension to be more interpretable and useful in a managerial setting. This provides a straightforward method of achieving the benefits of data summarization while also creating factors for data reduction (see next section) which have substantive managerial impact.

Exploratory factor analysis is also crucial in the process of scale development, where a set of individual variables are selected for their relationship to a more general concept. For example, individual variables might be: “I shop for specials,” “I usually look for the lowest possible prices,” “I shop for bargains,” “National brands are worth more than store brands.” Collectively, these variables might be used to identify consumers who are “price conscious” or “bargain hunters.” The distinguishing feature in this use of exploratory factor analysis is that there is a prespecified structure that is hopefully revealed in the analysis. We will discuss such issues as conceptual definitions, reliability and validity in a later section as they are all issues by which an exploratory factor analysis is evaluated in the scale development process.

The choice of data summarization with or without interpretation will have an impact on the factor extraction method applied—principal components analysis or common factor analysis. In brief, principal components analysis is based on extracting as much as possible of the total variance in the set of variables and is thus ideally suited for use with data summarization with less emphasis on interpretation. When the objective involves interpretation however, the researcher has a choice of methods. Common factor analysis is most applicable to interpretability since it analyzes only the shared variation among the variables, focusing on what is in common among the variables. Yet principal components analysis is widely used in this manner as well and we will discuss the pros and cons of each approach in a later section.

**Data Reduction** Once exploratory factor analysis has been evaluated in the data summarization process, researchers generally also use data reduction techniques to (1) identify representative variables from the much larger set of variables represented by each factor for use in subsequent multivariate analyses, or (2) create an entirely new set of variables, representing composites of the variables represented by each factor, to replace the original set of variables. In both instances, the purpose is to retain the nature and character of the original variables, but reduce the number of actual values included in the analysis (i.e., one per factor) to simplify the subsequent multivariate analysis. Even though the multivariate techniques were developed to accommodate multiple variables, the researcher is always looking for the most parsimonious set of variables to include in the analysis. As discussed in Chapter 1, both conceptual and empirical issues support the creation of composite measures while “Managing the Variate” (see Figure 3.1). Exploratory factor analysis provides the empirical basis for assessing the structure of variables and the potential for creating these composite measures or selecting a subset of representative variables for further analysis.

Data summarization makes the identification of the underlying dimensions or factors as the ends in themselves. Thus, estimates of the factors and the contributions of each variable to the factors (termed *loadings*) provide the basis for finalizing the content of each factor. Then, any of the data reduction approaches can be used with the results of the data summarization process. The combination of data summarization and data reduction can create many options. For example, it might involve simplified data summarization without interpretation, and the creation of factor scores by the software. But it also could involve a more complex scale development process whereby data summarization

plays a critical role in determining the final variables to be retained in each factor, and then some type of score (e.g., summated scales) is created to represent each factor. We will discuss in more detail the issues underlying both data summarization and data reduction in later sections, but have summarized these basic options to acquaint the reader with the analysis alternatives provided by exploratory factor analysis.

## VARIABLE SELECTION

Whether exploratory factor analysis is used for data reduction and/or summarization, the researcher should always consider the conceptual underpinnings of the variables and use judgment as to the appropriateness of the variables for factor analysis.

**Variable Specification** In both uses of exploratory factor analysis, the researcher implicitly specifies the potential dimensions that can be identified through the character and nature of the variables submitted to factor analysis. For example, in assessing the dimensions of store image, if no questions on store personnel were included, factor analysis would not be able to identify this dimension.

**Factors Are Always Produced** The researcher also must remember that exploratory factor analysis will always produce factors. Thus, exploratory factor analysis is always a potential candidate for the “garbage in, garbage out” phenomenon. If the researcher indiscriminately includes a large number of variables and hopes that factor analysis will “figure it out,” then the possibility of poor results is high. The quality and meaning of the derived factors reflect the conceptual underpinnings of the variables included in the analysis. Thus, researchers should use judgment on which set of variables to analyze together and understand that the researcher dictates the structure being examined. For example, a set of variables representing attitudes and opinions would in most instances not be analyzed with another set of variables representing actual behaviors as these are distinctly different sets of variables. This is particularly true when one set may be independent variables and another set will be dependent variables. Exploratory factor analysis cannot distinguish between these sets as can be done in confirmatory factor analysis, so the researcher is best suited to analyze each set separately for their unique structure rather than in a combined analysis.

**Factors Require Multiple Variables** The researcher has ultimate control on which variables are subjected to factor analysis. In some cases, there may only be single variable that measures a concept that is available and thus no need to place that variable in a factor analysis (e.g., probably only a single measure of employee turnover—Yes or No). So these individual variables should not be put into the exploratory factor analysis, but instead “mixed” with the composite scores from data reduction as needed.

Obviously, the use of exploratory factor analysis as a data summarization technique is based on some conceptual basis for the variables analyzed. But even if used primarily for data reduction with no interpretation in the data summarization process, factor analysis is most efficient when conceptually defined dimensions can be represented by the derived factors.

## USING FACTOR ANALYSIS WITH OTHER MULTIVARIATE TECHNIQUES

Exploratory factor analysis, by providing insight into the interrelationships among variables and the underlying structure of the data, is an excellent starting point for many other multivariate techniques. From the data summarization perspective, exploratory factor analysis provides the researcher with a clear understanding of which variables may act in concert and how many variables may actually be expected to have an impact in the analysis.

- Variables determined to be highly correlated and belonging to the same factor would be expected to have similar profiles of differences across groups in multivariate analysis of variance or in discriminant analysis.
- Highly correlated variables, such as those within a single factor, affect the stepwise procedures of multiple regression and discriminant analysis that sequentially enter variables based on their incremental predictive

power over variables already in the model. As one variable from a factor is entered, it becomes less likely that additional variables from that same factor would also be included due to their high correlations with variable(s) already in the model, meaning that they have little incremental predictive power. It does not mean that the other variables of the factor are less important or have less impact, but instead their effect is already represented by the included variable from the factor. Thus, knowledge of the structure of the variables by itself would give the researcher a better understanding of the reasoning behind the entry of variables in this technique. High correlations among independent variables in multiple regression or discriminant analysis are thus an excellent reason for performing data reduction with exploratory factor analysis.

The insight provided by data summarization can be directly incorporated into other multivariate techniques through any of the data reduction techniques. Factor analysis provides the basis for creating a new set of variables that incorporate the character and nature of the original variables in a much smaller number of new variables, whether using representative variables, factor scores, or summated scales. In this manner, problems associated with large numbers of variables or high intercorrelations among variables can be substantially reduced by substitution of the new variables. The researcher can benefit from both the empirical estimation of relationships and the insight into the conceptual foundation and interpretation of the results.

## Stage 2: Designing an Exploratory Factor Analysis

The design of an exploratory factor analysis involves three basic decisions: (1) design of the study in terms of the number of variables, measurement properties of variables, and the types of allowable variables; (2) the sample size necessary, both in absolute terms and as a function of the number of variables in the analysis; and (3) calculation of the input data (a correlation matrix) to meet the specified objectives of grouping variables or respondents.

### VARIABLE SELECTION AND MEASUREMENT ISSUES

Two specific questions must be answered at this point: (1) What types of variables can be used in factor analysis? and (2) How many variables should be included? In terms of the types of variables included, the primary requirement is that a correlation value can be calculated among all variables. Metric variables are easily measured by several types of correlations. Nonmetric variables, however, are more problematic because they cannot use the same types of correlation measures used by metric variables. Although some specialized methods calculate correlations among nonmetric variables, the most prudent approach is to avoid nonmetric variables. If a nonmetric variable must be included, one approach is to develop **dummy variables** (coded 0–1) to represent categories of nonmetric variables. One drawback of this approach, however, is that there is no way for the program to ensure that all of the dummy variables created for a single multi-category variable are represented in a single factor. If all the variables are dummy variables, then specialized forms of factor analysis, such as Boolean factor analysis, are more appropriate [5].

The researcher should also consider the number of variables to include, but still maintain a reasonable number of variables per factor. If a study is being designed to assess a proposed structure (i.e., scale development), the researcher should be sure to include several variables (five or more) that may represent each proposed factor. The strength of exploratory factor analysis lies in finding patterns among groups of variables, and it is of little use in identifying factors composed of only a single variable. Finally, when designing a study to be factor analyzed, the researcher should, if possible, identify several key variables (sometimes referred to as key indicants or marker variables) that closely reflect the hypothesized underlying factors. This identification will aid in validating the derived factors and assessing whether the results have practical significance.

### SAMPLE SIZE

Regarding the sample size question, there are guidelines based on (1) the absolute size of the dataset, (2) the ratio of cases to variables and (3) the “strength” of the factor analysis results. In terms of absolute size, researchers generally would not factor analyze a sample of fewer than 50 observations, and preferably the sample size should be 100 or

larger. Researchers have suggested much larger samples (200 and larger) as the number of variables and expected number of factors increases. In terms of the ratio of observations to variables, the general rule is to have a minimum of five times as many observations as the number of variables to be analyzed, and a more acceptable sample size would have a 10:1 ratio. Some researchers even propose a minimum of 20 cases for each variable. Finally, if the objective of the exploratory factor analysis is to assess preliminary structural patterns, such as in a pilot test for a questionnaire, these ratios can be adjusted downward.

The final guidelines on sample size are based on the “strength” of the exploratory factor analysis results. One measure of how well a variable is accounted for in the retained factors is **communality**—the amount of a variable’s variance explained by its loadings on the factors. The communality is calculated as the sum of the squared loadings across the factors. We will discuss communality later as a measure for retaining variables, but communality also helps identify the “strength” of the factors in explaining each variable. Communalities of .70, for example, can only occur with at least one fairly high loading (e.g., the square root of .70 is .83), so high communalities generally denote factor results that exhibit a series of high loadings for each variable. Fabrigar and Wegener propose the following sample size guidelines: (a) a sample size of 100 is sufficient if all the communalities are .70 or above and there are at least three variables with high loadings on each factor; (b) as the communalities fall to the range of .40 to .70 then the sample size should be at least 200; and (c) if the communalities are below .40 and there are few high loadings per factor, sample sizes of up to 400 are appropriate [23].

In each of these guidelines we can see that the necessary sample size increases as the complexity of the factor analysis increases [46, 27]. For example, 30 variables requires computing 435 correlations in the factor analysis. At a .05 significance level, perhaps even 20 of those correlations would be deemed significant and appear in the factor analysis just by chance. Thus, the researcher should always try to obtain the highest cases-per-variable ratio to minimize the chances of overfitting the data (i.e., deriving factors that are sample-specific with little generalizability). In order to do so, the researcher may employ the most parsimonious set of variables, guided by conceptual and practical considerations, and then obtain an adequate sample size for the number of variables examined. When dealing with smaller sample sizes and/or a lower cases-to-variables ratio, the researcher should always interpret any findings cautiously. The issue of sample size will also be addressed in a later section on interpreting factor loadings.

## CORRELATIONS AMONG VARIABLES OR RESPONDENTS

The final decision in the design of an exploratory factor analysis focuses on calculating the input data for the analysis. We discussed earlier the two forms of factor analysis: *R*-type versus *Q*-type factor analysis. Both types of factor analysis utilize a correlation matrix as the basic data input. With *R*-type factor analysis, the researcher would use a traditional correlation matrix (correlations among variables) as input, but an alternative correlation matrix with *Q*-type.

The researcher could also elect to derive the correlation matrix from the correlations between the individual respondents. In this *Q*-type factor analysis, the results would be a factor matrix that would identify similar individuals. For example, if the individual respondents are identified by number, the resulting factor pattern might tell us that individuals 1, 5, 6, and 7 are similar. Similarly, respondents 2, 3, 4, and 8 would perhaps load together on another factor, and we would label these individuals as similar. From the results of a *Q* factor analysis, we could identify groups or clusters of individuals that exhibit a similar pattern on the variables included in the analysis.

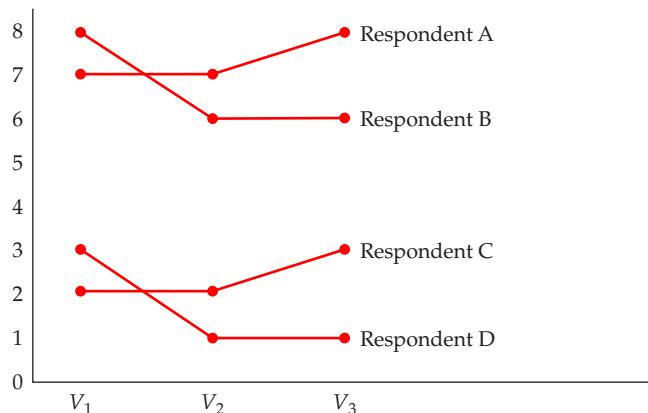
A logical question at this point would be: How does *Q*-type factor analysis differ from cluster analysis, because both approaches compare the pattern of responses across a number of variables and place the respondents in groups? The answer is that *Q*-type factor analysis is based on the intercorrelations between the respondents, whereas cluster analysis forms groupings based on a distance-based similarity measure between the respondents’ scores on the variables being analyzed.

To illustrate this difference, consider Figure 3.4, which contains the scores of four respondents over three different variables. A *Q*-type factor analysis of these four respondents would yield two groups with similar covariance structures, consisting of respondents A and C versus B and D. In contrast, the clustering approach would be sensitive to the actual distances between the respondents’ scores and would lead to a grouping of the closest pairs. Thus, with a cluster analysis approach, respondents A and B would be placed in one group and C and D in the other group.

**Figure 3.4**

Comparisons of Score Profiles for Q-Type Factor Analysis and Cluster Analysis

Respondent	Variables		
	$V_1$	$V_2$	$V_3$
A	7	7	8
B	8	6	6
C	2	2	3
D	3	1	1



If the researcher decides to employ Q-type factor analysis, these distinct differences from traditional cluster analysis techniques should be noted. With the availability of other grouping techniques and the widespread use of factor analysis for data reduction and summarization, the remaining discussion in this chapter focuses on R-type factor analysis, the grouping of variables rather than respondents.

**SUMMARY** Issues in the design of factor analysis are of equally critical importance whether an exploratory or confirmatory perspective is taken. In either perspective, the researcher is relying on the technique to provide insights into the structure of the data, but the structure revealed in the analysis is dependent on decisions by the researcher in such areas as variables included in the analysis, sample size, theoretical foundation, and so on. As such, several key considerations are listed in Rules of Thumb 3-1.

## Exploratory Factor Analysis Design

Exploratory factor analysis is performed most often only on metric variables, although specialized methods exist for the use of dummy variables; a limited number of “dummy variables” can be included in a set of metric variables that are factor analyzed as long as they represent binary attributes.

If a study is being designed to reveal factor structure, strive to have at least five variables for each proposed factor.

For sample size:

The sample must have more observations than variables.

The minimum absolute sample size should be 50 observations, with 100 observations the preferred minimum.

Increase the sample as the complexity of the factor analysis increases (i.e., number of variables and/or factors retained).

Strive to maximize the number of observations per variable, with a desired ratio of at least 5 observations per variable.

Higher communalities among the variables provides support for smaller samples sizes, all other things equal.

## Stage 3: Assumptions in Exploratory Factor Analysis

The critical assumptions underlying exploratory factor analysis are more conceptual than statistical. The researcher is always concerned with meeting the statistical requirement for any multivariate technique, but in exploratory factor analysis the overriding concerns center as much on the character and composition of the variables included in the analysis as on their statistical qualities.

### CONCEPTUAL ISSUES

The conceptual assumptions underlying factor analysis relate to the set of variables selected and the sample chosen. A basic assumption of factor analysis is that some *underlying structure does exist* in the set of selected variables. The presence of correlated variables and the subsequent definition of factors do not guarantee relevance, even if they meet the statistical requirements. It is the responsibility of the researcher to ensure that the observed patterns are conceptually valid and appropriate to study with exploratory factor analysis, because the technique has no means of determining appropriateness other than the correlations among variables. For example, mixing dependent and independent variables in a single factor analysis and then using the derived factors to support dependence relationships is not appropriate.

The researcher must also ensure that the sample is homogeneous with respect to the underlying factor structure. For example, it is inappropriate to apply exploratory factor analysis to a sample of males and females for a set of items known to differ because of gender. When the two subsamples (males and females) are combined, the resulting correlations and factor structure will be a poor representation (and likely incorrect) of the unique structure of each group. Thus, whenever groups are included in the sample that are expected to have different items measuring the same concepts, separate factor analyses should be performed, and the results should be compared to identify differences not reflected in the results of the combined sample.

### STATISTICAL ISSUES

From a statistical standpoint, departures from normality, homoscedasticity, and linearity apply only to the extent that they diminish the observed correlations. Only normality is necessary if a statistical test is applied to the significance of the factors, but these tests are rarely used. And some degree of **multicollinearity** is desirable, because the objective is to identify interrelated sets of variables.

Assuming the researcher has met the conceptual requirements for the variables included in the analysis, the next step is to ensure that the variables are sufficiently intercorrelated to produce representative factors. As we will see, we can assess this degree of interrelatedness from the perspectives of both the overall correlation matrix and individual variables. The following are several empirical measures to aid in diagnosing the factorability of the correlation matrix.

**Overall Measures of Intercorrelation** In addition to the statistical bases for the correlations of the data matrix, the researcher must also ensure that the data matrix has sufficient correlations to justify the application of exploratory factor analysis. If it is found that all of the correlations are low, or that all of the correlations are equal (denoting that no structure exists to group variables), then the researcher should question the application of exploratory factor analysis. To this end, several approaches are available:

**VISUAL INSPECTION** If visual inspection reveals a small number of correlations among the variables greater than .30, then exploratory factor analysis is probably inappropriate. The correlations among variables can also be analyzed by computing the partial correlations among variables. A partial correlation is the correlation that is unexplained when the effects of other variables are taken into account. If “true” factors exist in the data, the partial correlation should be small, because the variable can be explained by the variables loading on the factors. If the partial correlations are high, indicating no underlying factors, then exploratory factor analysis is not appropriate.

The one exception regarding high correlations as indicative of a correlation matrix that is not appropriate occurs when two variables are highly correlated and have substantially higher loadings than other variables on that factor.

Then, their partial correlation may be high because they are not explained to any great extent by the other variables, but do explain each other. This exception is also to be expected when a factor has only two variables loading highly.

A high partial correlation is one with practical and statistical significance, and a rule of thumb would be to consider partial correlations above .7 as high. Major software programs (e.g., IBM SPSS, SAS, Stata, and R) provide the **anti-image correlation matrix**, which is just the negative value of the partial correlation. In each case, larger partial or anti-image correlations (shown on the off-diagonal portion of the matrix) are indicative of variables not suited to this factor analysis.

**BARTLETT TEST** Another method of determining the appropriateness of exploratory factor analysis examines the entire correlation matrix. The **Bartlett test of sphericity** is a statistical test for the presence of correlations among the variables. It provides the statistical significance indicating the correlation matrix has significant correlations among at least some of the variables. The researcher should note, however, that increasing the sample size causes the Bartlett test to become more sensitive in detecting correlations among the variables.

**MEASURE OF SAMPLING ADEQUACY** A third measure to quantify the degree of intercorrelations among the variables and the appropriateness of exploratory factor analysis is the **measure of sampling adequacy (MSA)**. This index ranges from 0 to 1, reaching 1 when each variable is perfectly predicted without error by the other variables. The measure can be interpreted with the following guidelines: .80 or above, meritorious; .70 or above, middling; .60 or above, mediocre; .50 or above, miserable; and below .50, unacceptable [34, 35]. The MSA increases as (1) the sample size increases, (2) the average correlations increase, (3) the number of variables increases, or (4) the number of factors decreases [35]. The researcher should always have an overall MSA value of above .50 before proceeding with the factor analysis. If the MSA value falls below .50, then the variable-specific MSA values (see the following discussion) can identify variables for deletion to achieve an overall value of .50.

**Variable-Specific Measures of Intercorrelation** In addition to a visual examination of a variable's correlations with the other variables in the analysis, the MSA guidelines should be extended to individual variables. The researcher should examine the MSA values for each variable and exclude those falling in the unacceptable range ( $< 0.50$ , shown on the diagonal of the matrix). In deleting variables, the researcher should first delete the variable with the lowest MSA and then recalculate the factor analysis. Continue this process of deleting the variable with the lowest MSA value under .50 until all variables have an acceptable MSA value. Once the individual variables achieve an acceptable level, then the overall MSA can be evaluated and a decision made on continuance of the factor analysis.

The purpose of eliminating individual variables with MSA values under .5 is that these variables typically are poorly represented by the extracted factors. Most often variables with a low MSA end up as single variables on a factor, the result of their lack of association with any of the other variables. We should note that they are not "bad" in some manner, just that they are not correlated highly enough with other variables in the analysis to be suitable for exploratory factor analysis.

## SUMMARY

Exploratory factor analysis is in many ways more impacted by not meeting its underlying conceptual assumptions than by the statistical assumptions. The researcher must be sure to thoroughly understand the implications of ensuring that the data meet the statistical requirements as well as having the conceptual foundation to support the results. In doing so, the researcher should consider several key guidelines as listed in Rules of Thumb 3-2.

## Stage 4: Deriving Factors and Assessing Overall Fit

Once the variables are specified and the correlation matrix is prepared, the researcher is ready to apply exploratory factor analysis to identify the underlying structure of relationships (see Figure 3.5). In doing so, decisions must be made concerning (1) the method of extracting the factors (common factor analysis versus principal components analysis) and (2) the number of factors selected to represent the underlying structure in the data.

## Testing Assumptions of Exploratory Factor Analysis

A strong conceptual foundation needs to support the assumption that a structure does exist before the exploratory factor analysis is performed.

A statistically significant Bartlett's test of sphericity ( $\text{sig.} < 0.50$ ) indicates that sufficient correlations exist among the variables to proceed.

Measure of sampling adequacy (MSA) values must exceed .50 for both the overall test and each individual variable; variables with values less than .50 should be omitted from the factor analysis one at a time, with the smallest one being omitted each time.

## SELECTING THE FACTOR EXTRACTION METHOD

The researcher can choose from two similar, yet unique, methods for defining (extracting) the factors to represent the structure of the variables in the analysis. This decision must combine the objectives of the exploratory factor analysis with knowledge about some basic characteristics of the relationships between variables. Before discussing the two methods available for extracting factors, a brief introduction to partitioning a variable's variance is presented.

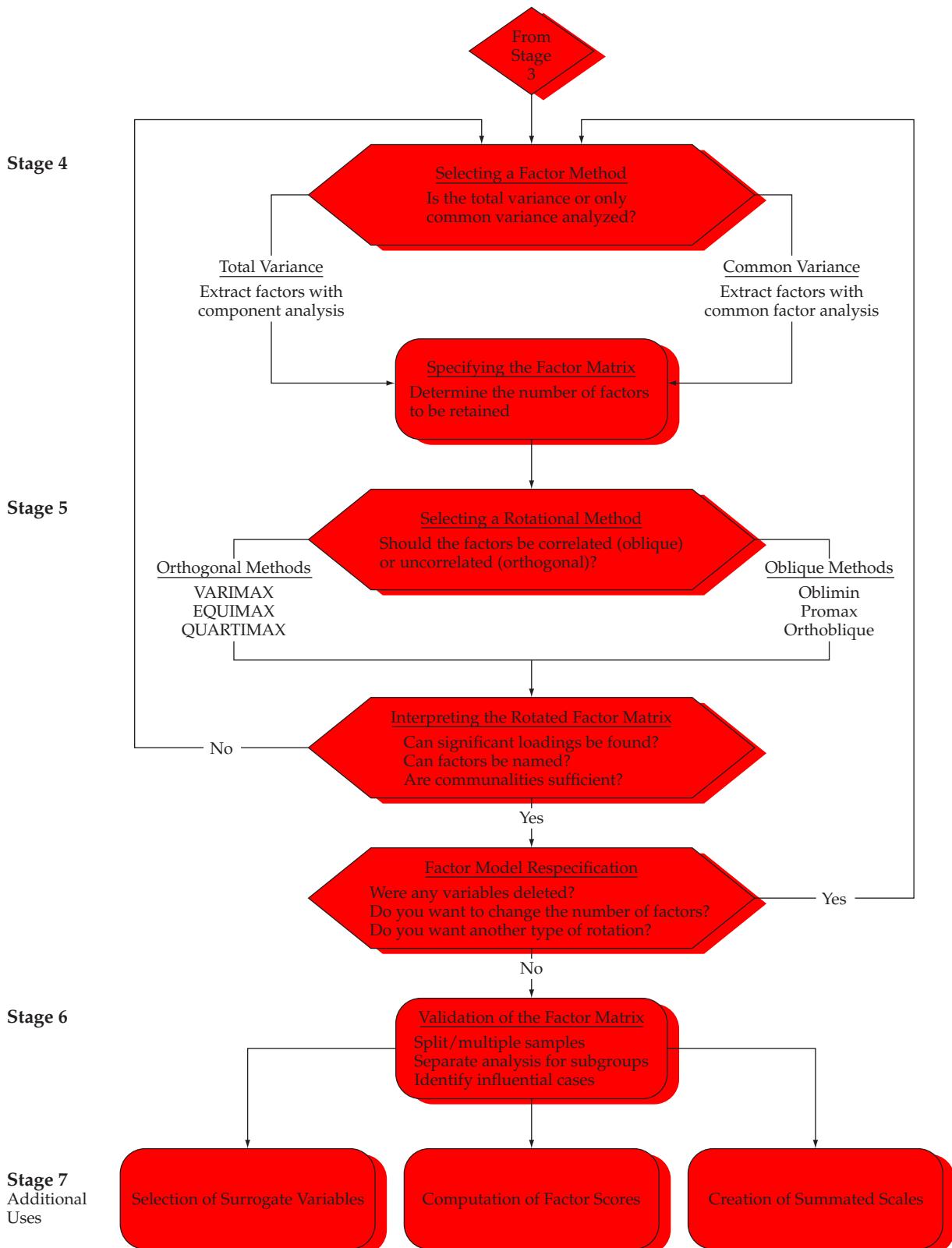
**Partitioning the Variance of a Variable** In order to select between the two methods of factor extraction, the researcher must first have some understanding of the variance for a variable and how it is divided or partitioned. First, remember that variance is a value (i.e., the square of the standard deviation) that represents the total amount of dispersion of values for a single variable about its mean. When a variable is correlated with another variable, we often say it *shares* variance with the other variable, and the amount of sharing between just two variables is simply the squared correlation. For example, if two variables have a correlation of .50, each variable shares 25 percent (.50<sup>2</sup>) of its variance with the other variable.

In exploratory factor analysis, we group variables by their correlations, such that variables in a group (factor) have high correlations with each other. Thus, for the purposes of exploratory factor analysis, it is important to understand how much of a variable's variance is shared with other variables in that factor versus what cannot be shared (e.g., unexplained). The total variance of any variable can be divided (partitioned) into three types of variance:

**COMMON VERSUS UNIQUE VARIANCE** The first partitioning is between common and unique variance. **Common variance** is that variance in a variable that is shared with all other variables in the analysis. This variance is accounted for (shared) based on the variable's correlations with all other variables in the analysis. A variable's communality is the estimate of its shared, or common, variance among the variables as represented by the derived factors. **Unique variance** is that variance associated with only a specific variable and is not represented in the correlations among variables. Variables with high common variance are more amenable to exploratory factor analysis since they correlate more with the other variables in the analysis. The amount of common variance is what is measured in the MSA values and provides an objective manner in which to assess the degree of common variance.

**UNIQUE VARIANCE COMPOSED OF SPECIFIC AND ERROR VARIANCE** While unique variance is not the primary basis for extracting factors, it is useful to try and understand its two potential sources—specific and error—so as to understand ways in which the common variance may be increased or the error variance mitigated in some fashion. **Specific variance** cannot be explained by the correlations to the other variables but is still associated uniquely with a single variable. It reflects the unique characteristics of that variable apart from the other variables in the analysis. **Error variance** is also variance that cannot be explained by correlations with other variables, but it is due to unreliability in the data-gathering process, measurement error, or a random component in the measured phenomenon. While making a precise estimate of the breakdown of unique variance into specific and error variance is not required for the analysis, understanding the sources of unique variance is important for assessing a variable in the factor results, particularly in the scale development process.

**Figure 3.5**  
Stages 4–7 in the Exploratory Factor Analysis Decision Diagram



Thus, the total variance of any variable has two basic sources—common and unique, with unique variance having two sub-parts—specific and error. As a variable is more highly correlated with one or more variables, the common variance (communality) increases. However, if unreliable measures or other sources of extraneous error variance are introduced, then the amount of possible common variance and the ability to relate the variable to any other variable are reduced.

**Common Factor Analysis Versus Principal Component Analysis** With a basic understanding of how variance can be partitioned, the researcher is ready to address the differences between the two methods, known as **common factor analysis** and **principal component analysis**. The selection of one method over the other is based on two criteria: (1) the objectives of the factor analysis and (2) the amount of prior knowledge about the variance in the variables. Principal component analysis is used when the objective is to summarize most of the original information (variance) in a minimum number of factors for prediction purposes. In contrast, common factor analysis is used primarily to identify underlying factors or dimensions that reflect what the variables share in common. The most direct comparison between the two methods is by their use of the explained versus unexplained variance:

**PRINCIPAL COMPONENTS ANALYSIS** Also known as **components analysis**, it considers the total variance and derives factors that contain small proportions of unique variance and, in some instances, error variance. However, the first few factors do not contain enough unique or error variance to distort the overall factor structure. Specifically, with principal component analysis, unities (values of 1.0) are inserted in the diagonal of the correlation matrix, so that the full variance is brought into the factor matrix. Figure 3.6 portrays the use of the total variance in principal component analysis and the differences when compared to common factor analysis.

**COMMON FACTOR ANALYSIS** In contrast, it considers only the common or shared variance, assuming that both the unique and error variance are not of interest in defining the structure of the variables. To employ only common variance in the estimation of the factors, communalities (instead of unities) are inserted in the diagonal. Thus, factors resulting from common factor analysis are based only on the common variance. As shown in Figure 3.6, common factor analysis excludes a portion of the variance included in a principal component analysis.

How is the researcher to choose between the two methods? First, the common factor and principal component analysis models are both widely used. As a practical matter, the principal components model is the typical default method of most statistical programs when performing factor analysis. Beyond the program defaults, distinct instances indicate which of the two methods is most appropriate.

Principal component analysis is most appropriate when:

- data reduction is a primary concern, focusing on the minimum number of factors needed to account for the maximum portion of the total variance represented in the original set of variables,

**Figure 3.6**  
Types of Variance Included in the Factor Matrix

Exploratory Factor Analysis Technique	Diagonal Value of Correlation Matrix	Variance Included in the Analysis		
Principal Components Analysis	Unity (1.0)	Common Variance	Unique Variance	
			Specific Variance	Error Variance
Common Factor Analysis	Commonality	Common Variance	Unique Variance	
			Specific Variance	Error Variance
Variance extracted				
Variance excluded				

- prior knowledge suggests that specific and error variance represent a *relatively small proportion* of the total variance, or
- the principal component results are used as a preliminary step in the scale development process.

Common factor analysis is most appropriate when:

- *the primary objective is to identify the latent dimensions or constructs represented in the common variance* of the original variables, as typified in the scale development process, and
- the researcher has *little knowledge about the amount of specific and error variance* and therefore wishes to eliminate this variance.

Common factor analysis, with its more restrictive assumptions and use of only the latent dimensions (shared common variance), is often viewed as more theoretically-based and generally associated with scale development. Although theoretically sound, however, common factor analysis has several problems. First, common factor analysis suffers from **factor indeterminacy**, which means that for any individual respondent, several different factor scores can be calculated from a single factor model result [40]. No single unique solution is found, as in principal component analysis, but in most instances the differences are not substantial. The second issue involves the calculation of the estimated communalities used to represent the shared variance. Sometimes the communalities are not estimable or may be invalid (e.g., values greater than 1 or less than 0), requiring the deletion of the variable from the analysis.

Does the choice of one model or the other really affect the results? The complications of common factor analysis have contributed to the widespread use of principal component analysis. But the base of proponents for the common factor model is strong. Cliff [16] characterizes the dispute between the two sides as follows:

Some authorities insist that component analysis is the only suitable approach, and that the common factor methods just superimpose a lot of extraneous mumbo jumbo, dealing with fundamentally unmeasurable things, the common factors. Feelings are, if anything, even stronger on the other side. Militant common-factorists insist that components analysis is at best a common factor analysis with some error added and at worst an unrecognizable hodgepodge of things from which nothing can be determined. Some even insist that the term “factor analysis” must not be used when a components analysis is performed.

Although debate remains over which factor model is the more appropriate [6, 26, 39, 54], empirical research demonstrates similar results in many instances [58, 27, 17, 28, 42]. In most applications, both principal component analysis and common factor analysis arrive at essentially identical results if the number of variables exceeds 30 [25] or the communalities exceed .60 for most variables. If the researcher is concerned with the assumptions of principal components analysis, then common factor analysis should also be applied to assess its representation of structure.

When a decision has been made on the factor extraction method, the researcher is ready to extract the initial unrotated factors. By examining the unrotated factor matrix, the researcher can explore the potential for data reduction and obtain a preliminary estimate of the number of factors to extract. Final determination of the number of factors must wait, however, until the results are rotated and the factors are interpreted.

### STOPPING RULES: CRITERIA FOR THE NUMBER OF FACTORS TO EXTRACT

How do we decide on the number of factors to extract? Both factor analysis methods are interested in the best linear combination of variables—best in the sense that the particular combination of original variables accounts for more of the variance in the data as a whole than any other linear combination of variables. Therefore, the first factor may be viewed as the single best summary of linear relationships exhibited in the data. The second factor is defined as the second-best linear combination of the variables, subject to the constraint that it is orthogonal to the first factor.

To be **orthogonal** to the first factor, the second factor must be derived only from the variance remaining after the first factor has been extracted. Thus, the second factor may be defined as the linear combination of variables that accounts for the most variance that is still unexplained after the effect of the first factor has been removed from the data. The process continues extracting factors accounting for smaller and smaller amounts of variance until all of the variance is explained. For example, the components method actually extracts  $n$  factors, where  $n$  is the number of variables in the analysis. Thus, if 30 variables are in the analysis, 30 factors are extracted.

So, what is gained by exploratory factor analysis? In our example of 30 store image variables where 30 factors are extracted, the first factors will hopefully account for a substantial enough portion of the variance so that the researcher can retain only a small number of factors to adequately represent the variance of the entire set of variables. The key question is: *How many factors to extract or retain?*

In deciding when to stop factoring (i.e., how many factors to extract), the researcher must combine a conceptual foundation (How many factors should be in the structure?) with some empirical evidence (How many factors can be reasonably supported?). The researcher generally begins with some predetermined criteria, such as the general number of factors plus some general thresholds of practical relevance (e.g., required percentage of variance explained). These criteria are combined with empirical measures of the factor structure. A definitive quantitative basis for deciding the number of factors to extract has not been developed. However, the following **stopping rules** for the number of factors to extract are currently being utilized.

**A Priori Criterion** The **a priori criterion** is a simple yet reasonable criterion under certain circumstances. When applying it, the researcher already knows how many factors to extract before undertaking the factor analysis. The researcher simply instructs the computer to stop the analysis when the desired number of factors has been extracted. This approach is useful when testing a theory or hypothesis about the number of factors to be extracted. It also can be justified in attempting to replicate another researcher's work and extract the same number of factors that was previously found.

**Latent Root Criterion** The most commonly used technique is the **latent root criterion**, also known as the **Kaiser rule**. This technique is simple to apply, the rationale being that any individual factor should account for the variance of at least a single variable if it is to be retained for interpretation. With principal component analysis each variable by itself contributes a value of 1 (i.e., the value on the diagonal of the correlation matrix) to the total eigenvalue for all variables. The eigenvalue of a single factor is simply the sum of the squared loadings of variables on that factor. The simple rule is: Don't retain any factors which account for less variance than a single variable. Thus, only the factors having **latent roots** or **eigenvalues** greater than 1 are considered significant; all factors with latent roots less than 1 are considered insignificant and are disregarded. This rule is most applicable to principal components analysis where the diagonal value representing the amount of variance for each variable is 1.0. In common factor analysis the diagonal value is replaced with the communality (amount of variance explained) of the variable. In most instances this is a value less than 1.0, so using the latent root criterion on this form of the correlation matrix would be less appropriate. So in common factor analysis many times the latent root criterion is applied to the factor results before the diagonal value is replaced by the communality. Another approach is to extract factors with an eigenvalue less than one, with the level being chosen as approximately the average of the communalities of the items.

This stopping rule has been criticized as being too simplistic and certainly has its drawbacks, working less accurately with a small number of variables or lower communalities. The latent root criterion is most reliable when the number of variables is between 20 and 50 and communalities above .40. If the number of variables is less than 20, the tendency is for this method to extract a conservative number of factors (too few). In contrast, if more than 50 variables are involved, it is not uncommon for too many factors to be extracted [7, 37]. In most instances, the latent root criterion is applied as a first step, and then other criteria are considered in combination with this initial criterion.

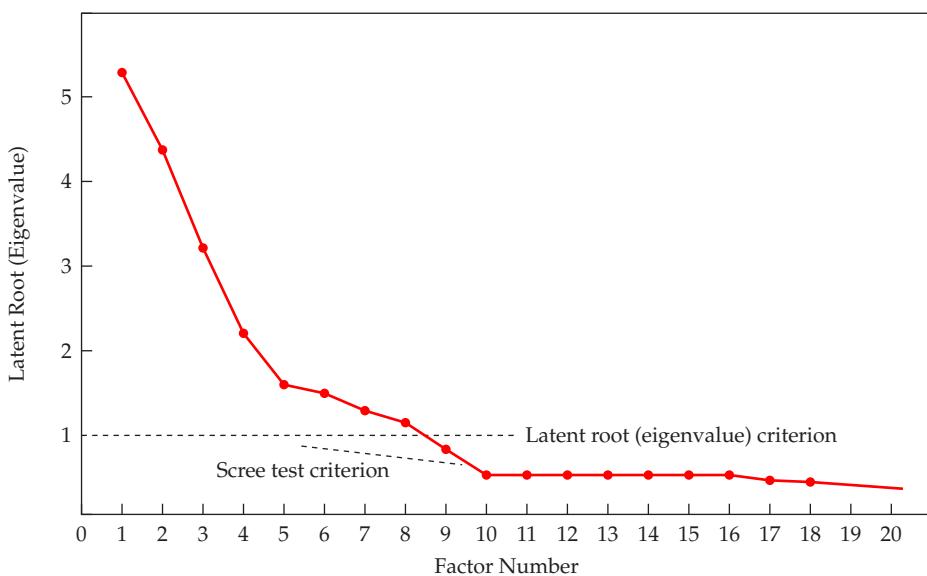
**Percentage of Variance Criterion** The **percentage of variance criterion** is an approach based on achieving a specified cumulative percentage of total variance extracted by successive factors. The purpose is to ensure practical significance for the derived factors by ensuring that they explain at least a specified amount of variance. No absolute threshold has been adopted for all applications. However, in the natural sciences the factoring procedure usually should not be stopped until the extracted factors account for at least 95 percent of the variance, or until the last factor accounts for only a small portion (less than 5%). In contrast, in the social sciences, where information is often less precise, it is not uncommon to consider a solution that accounts for 60 percent of the total variance as satisfactory, and in some instances even less.

A variant of this criterion involves selecting enough factors to achieve a prespecified communality for each of the variables. If theoretical or practical reasons require a certain communality for each variable, then the researcher will include as many factors as necessary to adequately represent each of the original variables. This approach differs from focusing on just the total amount of variance explained, which may diminish the degree of explanation for the individual variables.

**Scree Test Criterion** Recall that with the component analysis factor model the later factors extracted contain both common and unique variance. Although all factors contain at least some unique variance, the proportion of unique variance is substantially higher in later factors. The scree test is used to identify the optimum number of factors that can be extracted before the amount of unique variance begins to dominate the common variance structure [12].

The scree test is derived by plotting the latent roots against the number of factors in their order of extraction, and the shape of the resulting curve is used to evaluate the cut-off point. Figure 3.7 plots the first 18 factors extracted in a study. Starting with the first factor, the plot slopes steeply downward initially and then slowly becomes an approximately horizontal line—an inflection point termed by many as the “elbow.” This point at which the curve first begins to straighten out is considered to represent those factors containing more unique rather than common variance and thus are less suitable for retention. Most researchers do not include the elbow, but rather retain all of the preceding factors. So in this instance if the elbow was at 10 factors, then 9 factors would be retained. Some researchers advocate including the elbow (inflection point). In either case, identifying the elbow is many times difficult and thus the scree test is considered more subjective than other methods. Note that in using the latent root criterion only eight factors would have been considered. In contrast, using the scree test results in one or two more factors being retained.

**Figure 3.7**  
Eigenvalue Plot for Scree Test Criterion



As a general rule, the scree test results in at least one and sometimes two or three more factors being considered for inclusion than does the latent root criterion [12].

**Parallel Analysis** All of the previous stopping rules have been ad hoc criteria that are generalized across all situations. **Parallel analysis** was developed [31] to form a stopping rule based on the specific characteristics (i.e., number of variables and sample size) of the dataset being analyzed. This procedure generates a large number (e.g., 500 or 1000) simulated datasets with random values for the same number of variables and sample size. Each of these simulated datasets is then factor analyzed, either with principal components or common factor methods, and the eigenvalues are averaged for each factor across all the datasets. The result is the average eigenvalue for the first factor, the second factor and so on across the set of simulated datasets. These values are then compared to the eigenvalues extracted for the original data and all factors with eigenvalues above those of the simulated datasets are retained. Variants of this rule suggest taking the upper bound (95th percentile) of the simulated eigenvalues as a more conservative threshold. When used with principal components analysis, parallel analysis may prove to be a “stricter” threshold than the latent root criterion as the number of factors increases (i.e., indicate fewer factors than the latent root criterion). With common factor analysis, the opposite has been observed and it may support later factors that other stopping rules eliminate [9].

While there is still emerging research on the efficacy of parallel analysis, recent efforts have found it quite accurate across a wide range of situations [23, 50, 43, 57]. An extension, the comparative data approach, has recently been posed which incorporates more of the distributional characteristics of the original dataset [50]. While parallel analysis has not been incorporated into the major statistical packages, macros are available to perform this procedure in each package [43] as well as links on our website: <http://www.mvstats.com>.

**Heterogeneity of the Respondents** An assumption of all factor models that impacts the number of factors retained is that the shared variance among variables extends across the entire sample. If the sample is heterogeneous with regard to at least one subset of the variables, then the first factors will represent those variables that are more homogeneous across the entire sample. Variables that are better discriminators between the subgroups of the sample will load on later factors, many times those not selected by the criteria discussed previously [22]. When the objective is to identify factors that discriminate among the subgroups of a sample, the researcher should extract additional factors beyond those indicated by the methods just discussed and examine the additional factors’ ability to discriminate among the groups. If they prove less beneficial in discrimination, the solution can be run again and these later factors eliminated.

**Summary of Factor Selection Criteria** In practice, most researchers seldom use a single criterion in determining how many factors to extract. Instead, they initially use a criterion such as the latent root as a guideline for the first attempt at interpretation. Recent research has shown that most research efforts have employed only a single stopping rule rather than looking for a convergence across a number of stopping rules [30]. Yet no matter what set of stopping rules are used, the researcher must also strive for interpretability if data summarization includes any form of interpretation. After the factors are interpreted, as discussed in the following sections, the practicality of the factors is assessed. Factors identified by other criteria are also interpreted. Selecting the number of factors is interrelated with an assessment of structure, which is revealed in the interpretation phase. Thus, several factor solutions with differing numbers of factors are examined before the structure is well defined. In making the final decision on the factor solution to represent the structure of the variables, the researcher should remember the considerations listed in Rules of Thumb 3-3.

One word of caution in selecting the final set of factors: Negative consequences arise from selecting either too many or too few factors to represent the data. If too few factors are used, then the correct structure is not revealed, and important dimensions may be omitted. If too many factors are retained, then the interpretation becomes more difficult when the results are rotated (as discussed in the next section). Although the factors are independent, you can just as easily have too many factors as too few [62]. By analogy, choosing the number of factors is something

## Choosing Factor Models and Number of Factors

Although both component and common factor analysis models yield similar results in common research settings (30 or more variables or communalities of .60 for most variables):

The component analysis model is most appropriate when data reduction is paramount.

The common factor model is best in well-specified theoretical applications.

Any decision on the number of factors to be retained should be based on several considerations:

Use of several stopping criteria to determine the initial number of factors to retain:

Factors with eigenvalues greater than 1.0.

A predetermined number of factors based on research objectives and/or prior research.

Enough factors to meet a specified percentage of variance explained, usually 60 percent or higher.

Factors shown by the scree test to have substantial amounts of common variance (i.e., factors before inflection point).

Factors above the threshold established by parallel analysis.

More factors when heterogeneity is present among sample subgroups.

Consideration of several alternative solutions (one more and one less factor than the initial solution) to ensure the best structure is identified.

like focusing a microscope. Too high or too low an adjustment will obscure a structure that is obvious when the adjustment is just right. Therefore, by examining a number of different factor structures derived from several trial solutions, the researcher can compare and contrast to arrive at the best representation of the data.

As with other aspects of multivariate models, parsimony is important. The notable exception is when exploratory factor analysis is used strictly for data summarization without interpretation with a specified level of variance to be extracted (i.e., no researcher judgment involved). In all other situations the researcher must balance representativeness and interpretability against parsimony in determining the final set of factors.

## ALTERNATIVES TO PRINCIPAL COMPONENTS AND COMMON FACTOR ANALYSIS

Our discussion to this point has focused on the two approaches most commonly associated with exploratory factor analysis. But as with all multivariate methods, there are always alternative approaches being developed for specialized situations or as an alternative to the traditional methods. In this regard, we will discuss briefly two additional exploratory factor methods that are in use: factor analysis for ordinal/categorical data, which eliminates the requirement for metric data; and variable clustering which uses an extraction approach somewhat similar to cluster analysis.

**Factor Analysis of Categorical Data** As we have discussed throughout this chapter, the fundamental element of exploratory factor analysis is the correlation among variables. We use these interrelationships to identify factors and assess which variables are most representative of each factor. But what are we to do when the data we wish to analyze are not conducive to calculating correlations—Can we perform these same analyses on nonmetric data? This question becomes increasingly important in this era of “Big Data” where data is obtained from many sources and more often is nonmetric in nature rather than the orderly responses we are many times accustomed to in survey data. We discussed earlier representing categorical variables as a series of binary variables, but that is problematic as each category is considered separately and categories from the same variable may load on different factors. Moreover, even academicians face questions of appropriateness of correlations among the items they use, such as the continued debate on whether Likert-type questions (i.e., Agree–Disagree) are metric or nonmetric [32, 11].

**Optimal scaling** is a process to derive interval measurement properties for variables which were originally nominal or ordinal measures. Each original variable is “scaled” so as to achieve maximal fit with the other variables in the analysis. This results in two characteristics to always remember: (a) optimal scaling works within the variable set specified (e.g., maximizes the correlations among the variables in the analysis), but this scaling may change in other analyses, and (b) the researcher sets a series of allowable transformations (e.g., retain ordinality or not) and constraints that impact this process. So while in most situations we feel that metric measures are fairly consistent (even with transformations), optimal scaling may result in more variation in the relationships based on the variables in the analysis.

Originally conceptualized by Fisher [24] and extended by a number of researchers [20, 36, 61, 60, 59], this process is now available in all major statistical packages (e.g., CATPCA, Categorical Principal Components Analysis in SPSS; and PROC PRINQUAL, Principal Components of Qualitative Data in SAS). Once the optimal scaling process is specified and the data transformed, the process proceeds in a manner quite similar to principal components analysis.

**Variable Clustering** As the number of variables increases dramatically, principal components analyses become more difficult in a number of ways. A variant of PCA, termed **variable clustering**, has emerged to address several of these issues and is available in the PROC VARCLUS of SAS [51]. This alternative method of factor extraction using principal components has several distinct benefits:

- It is not impacted by the number of variables included like PCA.
- It is based on a hierarchical process of component formation similar to that seen in cluster analysis.
- It introduces additional diagnostics for selecting the most representative variable(s) for each component.

The procedure employs principal components analysis as the basic factor model in the initial step, but then “splits” each component sequentially, much like divisive hierarchical cluster analysis, to generate “clusters” of variables within non-orthogonal principal components (called **cluster components**). The cluster components provide a much “cleaner” interpretation as they consist of only a subset of the original variables, i.e., the variables in each cluster. To better understand this process, let us examine the basic steps in performing variable clustering:

- 1 Use principal components analysis to generate a two-component solution and perform an oblique rotation.
- 2 Assign each variable to one of the rotated components based on the highest factor loading, forming what are termed cluster components (i.e., only those variables assigned to that component). These cluster components now contain only those assigned variables.
- 3 Then, for each cluster component, apply principal components again and assess if the cluster component can be “split” into two more components. If so, assign the variables to their respective cluster components.
- 4 Continue the process by examining each cluster component to see if it can be split into two components. Stop the process when all cluster components can no longer be split.

This process results in a disjoint hierarchical solution where each variable is contained in a single cluster component. To understand how well each cluster component represents each of its variables, the loading of each variable on its cluster component is compared to the next highest loading on another component. The **1 – R<sup>2</sup> ratio** is calculated for each variable, with lower values indicating a “cleaner” solution:

$$1 - R^2 \text{ ratio} = \frac{1 - R^2 \text{ own cluster}}{1 - R^2 \text{ next closest cluster}}$$

There are several options the researcher can employ to change the procedure (e.g., the threshold determining if a cluster component will split, methods of assigning variables to clusters and even the maximum number of clusters). Variable clustering does, however, have some limitations, particularly in light of its recent development:

- It is considered more subjective than PCA, with fewer objective standards used in estimation of the final solution.
- You must use PROC SCORE to generate cluster/component scores for each observation.

Variable clustering provides an alternative to principal components analysis that researchers are exploring given the dramatic increase in the number of variables being examined for data reduction. It provides an efficient process of generating solutions which many times are found to be “cleaner” than principal components analysis, and the hierarchical nature of the process sometimes provides insights in the interpretation process.

**Summary** Categorical principal components analysis and variable clustering are just two of the increasing number of methods being developed for variable reduction, particularly with large datasets. The ever-increasing number of variables available for analysis in all types of research questions, academic and commercial, are pushing the development of these and other methods. Researchers are encouraged to be attuned to developments in techniques for “Managing The Variate,” both in variable reduction and variable selection (see more discussion in Chapters 1 and 5).

## Stage 5: Interpreting the Factors

As discussed earlier, most applications of exploratory factor analysis involve data summarization with interpretation. Although no unequivocal processes or guidelines determine the interpretation of factors, the researcher with a strong conceptual foundation for the anticipated structure and its rationale has the greatest chance of success. We cannot state strongly enough the importance of a strong conceptual foundation, whether it comes from prior research, theoretical paradigms, qualitative research, or commonly accepted principles. As you will see, the researcher must repeatedly make subjective judgments in decisions such as to the number of factors to extract, the sufficient number of relationships to warrant grouping variables, and how to identify the groupings. As the experienced researcher can attest, almost anything can be uncovered if one tries long and hard enough (e.g., using different factor models, extracting different numbers of factors, using various forms of rotation). It is therefore up to the researcher to be the final arbitrator as to the form and appropriateness of a factor solution, and such decisions are best guided by conceptual rather than empirical bases.

To assist in the process of interpreting a factor structure and selecting a final factor solution, three fundamental processes are described. Within each process, several substantive issues (factor rotation, factor-loading significance, and factor interpretation) are encountered. Thus, after each process is briefly described, each of these processes will be discussed in more detail.

### THE THREE PROCESSES OF FACTOR INTERPRETATION

Factor interpretation is circular in nature. The researcher first evaluates the initial results, then makes a number of judgments in viewing and refining these results, with the distinct possibility that the analysis is respecified, and therefore requires a return to the initial step. Thus, the researcher should not be surprised that several iterations are necessary before a final solution is achieved.

**1. Estimate the Factor Matrix** First, the initial unrotated **factor matrix** is computed, containing the factor loadings for each variable on each factor. **Factor loadings** are the correlation of each variable and the factor. Loadings indicate the degree of correspondence between the variable and the factor, with higher loadings making the variable representative of the factor. In exploratory factor analysis, each variable has loadings on all factors. Factor loadings are the means of interpreting the role each variable plays in defining each factor.

In those situations in which exploratory factor analysis is used strictly for data reduction, there is no need for any interpretation processes. The researcher extracts the factors and determines the number of factors to retain and then proceeds to the data reduction process of creating factors scores, etc. This is illustrated in SAS that has PROC PRINCOMP, a procedure devoted only to principal components and data reduction options. It does not have any of the interpretation methods we will discuss next (i.e., rotation methods), just data summarization without interpretation and data reduction.

**2. Factor Rotation** If interpretation is desired, the unrotated factor solutions achieve the objective of data reduction, but the researcher must ask whether the unrotated factor solution (which fulfills desirable mathematical requirements) will provide information that offers the most adequate interpretation of the variables under examination. In most instances the answer to this question is no, because factor rotation (a more detailed discussion follows in the next section) should simplify the factor structure (i.e., have each variable load highly on only one factor). Therefore, the researcher next employs a rotational method to achieve simpler and theoretically more meaningful factor solutions. In most cases rotation of the factors improves the interpretation by reducing some of the ambiguities that often accompany initial unrotated factor solutions.

**3. Factor Interpretation and Respecification** As a final process, the researcher evaluates the (rotated) factor loadings for each variable in order to determine that variable's role and contribution in determining the factor structure. In the course of this evaluative process, the need may arise to respecify the factor model owing to (1) the deletion of a variable(s) from the analysis, (2) the desire to employ a different rotational method for interpretation, (3) the need to extract a different number of factors, or (4) the desire to change from one extraction method to another. Respecification of a factor model involves returning to the extraction stage (stage 4), extracting factors, and then beginning the process of interpretation once again.

## FACTOR EXTRACTION

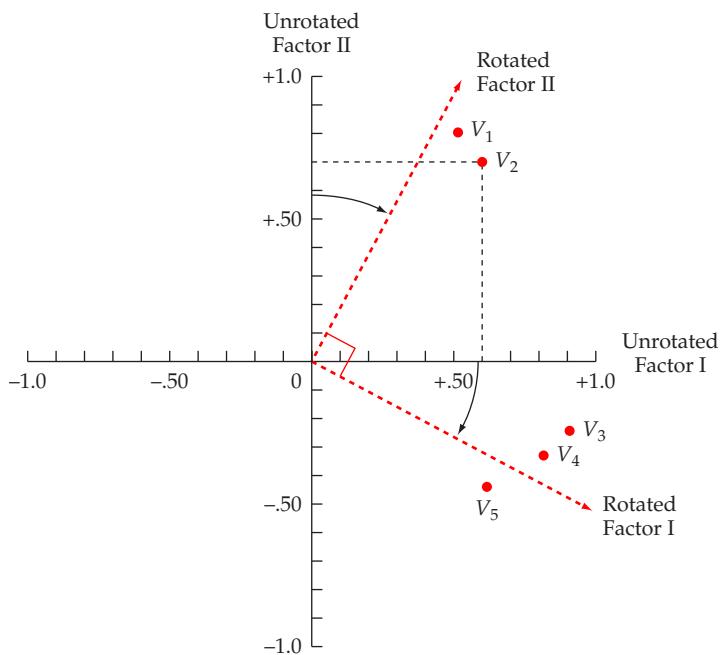
The process of actually estimating the factors and loadings for each variable primarily involve the selection of the principal components versus common factor approach (see earlier discussion). While principal components analysis is most often the default approach in most software, there are a number of options for forms of common factor analysis. For example, both IBM SPSS and SAS have principal axis factor/iterated principal factor analysis (most often used for common factor) along with alpha factoring and image factoring. The researcher is encouraged to explore the different common factor options available as each has a slightly different approach in estimating communalities and extracting factors.

One issue in extraction of the common factor model is the estimation of communalities of each variable. Many times the process identifies variables in which the communality estimates exceed 1.0 and these variables must be eliminated. The researcher is encouraged to utilize the diagnostic information provided by the MSA values as well as identifying any variables with very high bivariate correlations as these many times are problematic at this stage.

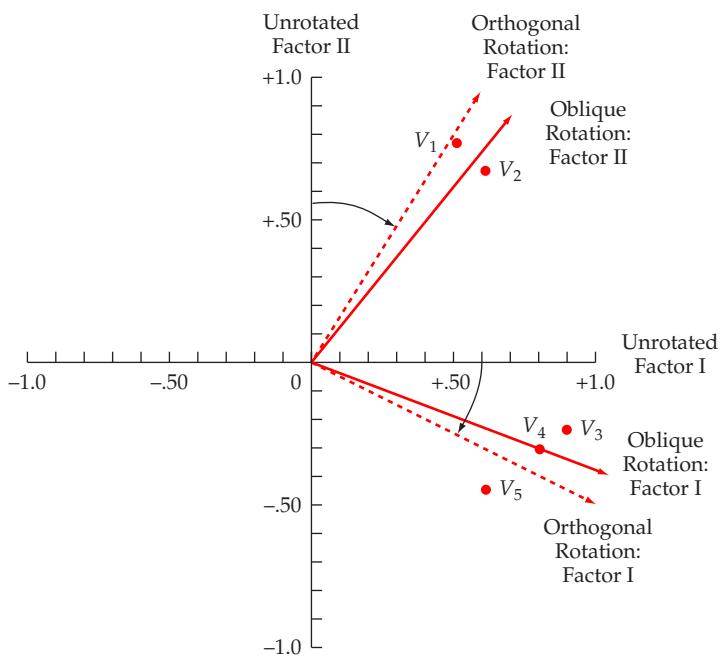
## ROTATION OF FACTORS

Perhaps the most important tool in interpreting factors is **factor rotation**. The term *rotation* means exactly what it implies. Specifically, the reference axes of the factors are turned about the origin until some other position has been reached. As indicated earlier, unrotated factor solutions extract factors in the order of their variance extracted. The first factor tends to be a general factor with almost every variable loading significantly, and it accounts for the largest amount of variance. The second and subsequent factors are then based on the residual amount of variance. Each accounts for successively smaller portions of variance. We should note that the first factor is not necessarily more "important" than the subsequent factors for any other purposes (e.g., prediction), but just that it represents the most variance accounted for in the set of variables. The ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simpler, theoretically more meaningful factor pattern.

**Orthogonal Versus Oblique Rotation** The simplest case of rotation is an **orthogonal factor rotation**, in which the axes are maintained at 90 degrees. It is also possible to rotate the axes and not retain the 90-degree angle between the reference axes. When not constrained to being orthogonal, the rotational procedure is called an **oblique factor rotation**. Orthogonal and oblique factor rotations are demonstrated in Figures 3.8 and 3.9, respectively.



**Figure 3.8**  
Orthogonal Factor Rotation



**Figure 3.9**  
Oblique Factor Rotation

Figure 3.8, in which five variables are depicted in a two-dimensional factor diagram, illustrates factor rotation. The vertical axis represents unrotated factor II, and the horizontal axis represents unrotated factor I. The axes are labeled with 0 at the origin and extend outward to +1.0 or -1.0. The numbers on the axes represent the factor loadings. The five variables are labeled  $V_1$ ,  $V_2$ ,  $V_3$ ,  $V_4$  and  $V_5$ . The factor loading for variable 2 ( $V_2$ ) on unrotated factor II is

determined by drawing a dashed line horizontally from the data point to the vertical axis for factor II. Similarly, a vertical line is drawn from variable 2 to the horizontal axis of unrotated factor I to determine the loading of variable 2 on factor I. A similar procedure followed for the remaining variables determines the factor loadings for the unrotated and rotated solutions, as displayed in Table 3.1 for comparison purposes. On the unrotated first factor, all the variables load fairly high. On the unrotated second factor, variables 1 and 2 are very high in the positive direction. Variable 5 is moderately high in the negative direction, and variables 3 and 4 have considerably lower loadings in the negative direction.

From visual inspection of Figure 3.8, two clusters of variables are obvious. Variables 1 and 2 go together, as do variables 3, 4, and 5. However, such patterning of variables is not so obvious from the unrotated factor loadings, particularly variable 5 which loads somewhat equally on the two unrotated axes. By rotating the original axes clockwise, as indicated in Figure 3.8, the axes are maintained at 90 degrees. This procedure signifies that the factors are mathematically independent and that the rotation has been orthogonal. The rotated loadings portray a much more simplified factor-loading pattern with variables loading highly only on a single factor. Variables 3, 4, and 5 now load high on factor I, and variables 1 and 2 load high on factor II. Thus, the clustering or patterning of these variables into two groups is more obvious after the rotation than before, even though the relative position or configuration of the variables remains unchanged.

The same general principles of orthogonal rotations pertain to oblique rotations. The oblique rotational method is more flexible, however, because the factor axes need not be orthogonal. It is also more realistic because the theoretically important underlying dimensions are not assumed to be uncorrelated with each other. In Figure 3.9 the two rotational methods are compared. Note that the oblique factor rotation represents the clustering of variables even more accurately. This accuracy is a result of the fact that each rotated factor axis is now closer to the respective group of variables. Also, the oblique solution provides information about the extent to which the factors are actually correlated with each other.

Most researchers agree that unrotated solutions are not sufficient. That is, in most cases some form of rotation will improve the interpretation by reducing some of the ambiguities that often accompany the preliminary analysis. The ultimate goal of any rotation is to obtain some theoretically meaningful factors and, if possible, the simplest factor structure. Orthogonal rotational approaches are more widely used because all computer packages with factor analysis contain orthogonal rotation options and they retain orthogonality among the factors. This is particularly useful if data reduction methods are used to create scores for each factor, as these scores are also orthogonal (i.e., no multicollinearity) which may be beneficial in other multivariate techniques. Orthogonal rotations are also utilized more frequently because the analytical procedures for performing oblique rotations are not as well developed and are still subject to some controversy. As a result, the oblique methods are not nearly as widely applied. Several different approaches are available for performing either orthogonal or oblique rotations. However, only a limited number of oblique rotational procedures are available in most statistical packages. Thus, the researcher will probably have to accept the one that is provided.

**Table 3.1** Comparison Between Rotated and Unrotated Factor Loadings

Variables	Unrotated Factor Loadings		Rotated Factor Loadings	
	I	II	I	II
$V_1$	.50	.80	.03	.94
$V_2$	.60	.70	.16	.90
$V_3$	.90	-.25	.95	.24
$V_4$	.80	-.30	.84	.15
$V_5$	.60	-.50	.76	-.13

**Orthogonal Rotation Methods** In practice, the objective of all methods of rotation is to simplify the rows and columns of the factor matrix to facilitate interpretation. In a factor matrix, columns represent factors, with each row corresponding to a variable's loading across the factors. By simplifying the rows, we mean making as many values in each row as close to zero as possible (i.e., maximizing a variable's loading on a single factor). By simplifying the columns, we mean making as many values in each column as close to zero as possible (i.e., making the number of high loadings as few as possible). Three major orthogonal approaches have been developed:

**VARIMAX** The VARIMAX criterion centers on simplifying the columns of the factor matrix. With the VARIMAX rotational approach, the maximum possible simplification is reached if there are only 1s and 0s in a column. That is, the VARIMAX method maximizes the sum of variances of required loadings of the factor matrix. The result is some high loadings (i.e., close to  $\pm 1$  or  $\pm 0$ ) along with some loadings near 0 in each column of the matrix. The logic is that interpretation is easiest when the variable-factor correlations are (1) close to either +1 or -1 thus indicating a clear positive or negative association between the variable and the factor; or (2) close to 0, indicating a clear lack of association. This structure is fundamentally simple. While the QUARTIMAX solution (discussed in the next section) focuses on simplifying rows and is analytically simpler than the VARIMAX solution, VARIMAX seems to give a clearer separation of the factors. In general, Kaiser's experiment [34, 35] indicates that the factor pattern obtained by VARIMAX rotation tends to be more invariant than that obtained by the QUARTIMAX method when different subsets of variables are analyzed. The VARIMAX method has proved successful as an analytic approach to obtaining an orthogonal rotation of factors, and is the most widely used orthogonal rotation method.

**QUARTIMAX** In contrast to VARIMAX, the goal of a QUARTIMAX rotation is to simplify the rows of a factor matrix; that is, QUARTIMAX focuses on rotating the initial factor so that a variable loads high on one factor and as low as possible on all other factors. In these rotations, many variables can load high or near high on the same factor because the technique centers on simplifying the rows. The QUARTIMAX method has not proved especially successful in producing simpler structures. The difficulty is that the method tends to produce a general factor as the first factor on which most, if not all, of the variables have high loadings. Regardless of one's concept of a simpler structure, inevitably it involves dealing with clusters of variables. A method that tends to create a large general factor (i.e., QUARTIMAX) is not consistent with the goals of rotation.

**EQUIMAX** The EQUIMAX approach is a compromise between the QUARTIMAX and VARIMAX approaches. Rather than concentrating either on simplification of the rows or on simplification of the columns, it tries to accomplish some of each. EQUIMAX has not gained widespread acceptance and is used infrequently.

**Oblique Rotation Methods** Oblique rotations are similar to orthogonal rotations, except that oblique rotations allow correlated factors instead of maintaining independence between the rotated factors. Where several choices are available among orthogonal approaches, however, most statistical packages typically provide only limited choices for oblique rotations. For example, IBM SPSS provides OBLIMIN and SAS has PROMAX and ORTHOBLIQUE. The objectives of simplification are comparable to the orthogonal methods, with the added feature of correlated factors. With the possibility of correlated factors, the factor researcher must take additional care to validate obliquely rotated factors, because they have an additional way (non-orthogonality) of becoming specific to the sample and not generalizable, particularly with small samples or a low cases-to-variable ratio.

**Selecting Among Rotational Methods** No specific rules have been developed to guide the researcher in selecting a particular orthogonal or oblique rotational technique and recent research demonstrated the variation in solutions based on the rotation method [52]. In most instances, the researcher simply utilizes the rotational technique provided by the computer program. Most programs have the default rotation of VARIMAX, but all the major rotational methods are widely available. However, no compelling analytical reason suggests favoring one

rotational method over another. The choice of an orthogonal or oblique rotation should be made on the basis of the particular needs of a given research problem. Many times, however, the researcher performs both types of rotations to better assess the underlying structure under the assumption of orthogonality versus relaxing this assumption. To this end, several considerations (in Rules of Thumb 3-4) should guide in selecting the rotational method.

## JUDGING THE SIGNIFICANCE OF FACTOR LOADINGS

In interpreting factors, a decision must be made regarding the factor loadings worth consideration and attention. The following discussion details issues regarding practical and statistical significance, as well as the number of variables, that affect the interpretation of factor loadings.

**Ensuring Practical Significance** The first guideline is not based on any mathematical proposition but relates more to practical significance by making a preliminary examination of the factor matrix in terms of the factor loadings. Because a factor loading is the correlation of the variable and the factor, the squared loading is the amount of the variable's total variance accounted for by the factor. Thus, a .30 loading translates to approximately 10 percent explanation, and a .50 loading denotes that 25 percent of the variance is accounted for by the factor. The loading must exceed .70 for the factor to account for 50 percent of the variance of a variable. Thus, the larger the absolute size of the factor loading, the more important the loading in interpreting the factor matrix. Using practical significance as the criteria, we can assess the loadings as follows:

- Factor loadings less than  $\pm .10$  can be considered equivalent to zero for purposes of assessing simple structure.
- Factor loadings in the range of  $\pm .30$  to  $\pm .40$  are considered to meet the minimal level for interpretation of structure.
- Loadings  $\pm .50$  or greater are considered practically significant.
- Loadings exceeding  $\pm .70$  are considered indicative of well-defined structure and are the goal of any factor analysis.

The researcher should realize that extremely high loadings (.90 and above) are not typical and that the practical significance of the loadings is an important criterion. These guidelines are applicable when the sample size is 100 or larger and where the emphasis is on practical, not statistical, significance.

**Assessing Statistical Significance** As previously noted, a factor loading represents the correlation between an original variable and its factor. In determining a significance level for the interpretation of loadings, an approach similar to determining the statistical significance of correlation coefficients could be used. However, research [18]

## Choosing Factor Rotation Methods

Orthogonal rotation methods:

- Are the most widely used rotational methods.
- Are the preferred method when the research goal is data reduction to either a smaller number of variables or a set of uncorrelated measures for subsequent use in other multivariate techniques.

OblIQUE rotation methods:

- Are best suited to the goal of obtaining several theoretically meaningful factors or constructs, because, realistically, few constructs in the real world are uncorrelated.

If interpretation is important, performing both rotational methods provides useful information on the underlying structure of the variables and the impact of orthogonality on the interpretation of the factors.

has demonstrated that factor loadings have substantially larger standard errors than typical correlations. Thus, factor loadings should be evaluated at considerably stricter levels. The researcher can employ the concept of statistical power discussed in Chapter 1 to specify factor loadings considered significant for differing sample sizes. With the stated objective of obtaining a power level of 80 percent, the use of a .05 significance level, and the proposed inflation of the standard errors of factor loadings, Table 3.2 contains the sample sizes necessary for each factor loading value to be considered significant.

For example, in a sample of 100 respondents, factor loadings of .55 and above are significant. However, in a sample of 50, a factor loading of .75 is required for significance. In comparison with the prior rule of thumb, which denoted all loadings of .30 as having practical significance, this approach would consider loadings of .30 significant only for sample sizes of 350 or greater.

These guidelines are quite conservative when compared with the guidelines of the previous section or even the statistical levels associated with conventional correlation coefficients. Thus, these guidelines should be used as a starting point in factor-loading interpretation, with lower loadings considered significant and added to the interpretation based on other considerations. The next section details the interpretation process and the role that other considerations can play.

**Adjustments Based on the Number of Variables** A disadvantage of both of the prior approaches is that the number of variables being analyzed and the specific factor being examined are not considered. It has been shown that as the researcher moves from the first factor to later factors, the acceptable level for a loading to be judged significant should increase. The fact that unique variance and error variance begin to appear in later factors means that some upward adjustment in the level of significance should be included [34]. The number of variables being analyzed is also important in deciding which loadings are significant. As the number of variables being analyzed increases, the acceptable level for considering a loading significantly decreases. Adjustment for the number of variables is increasingly important as one moves from the first factor extracted to later factors.

Rules of Thumb 3-5 summarizes the criteria for the practical or statistical significance of factor loadings.

**Table 3.2 Guidelines for Identifying Significant Factor Loadings Based on Sample Size**

Factor Loading	Sample Size Needed for Significance <sup>a</sup>
.30	350
.35	250
.40	200
.45	150
.50	120
.55	100
.60	85
.65	70
.70	60
.75	50

<sup>a</sup> Significance is based on a .05 significance level ( $\alpha$ ), a power level of 80 percent, and standard errors assumed to be twice those of conventional correlation coefficients.

Source: Computations made with SOLO Power Analysis, BMDP Statistical Software, Inc., 1993.

## Assessing Factor Loadings

Although factor loadings of  $\pm .30$  to  $\pm .40$  are minimally acceptable, values greater than  $\pm$  are generally considered necessary for practical significance.

To be considered significant:

A smaller loading is needed given either a larger sample size or a larger number of variables being analyzed.

A larger loading is needed given a factor solution with a larger number of factors, especially in evaluating the loadings on later factors.

Statistical tests of significance for factor loadings are generally conservative and should be considered only as starting points needed for including a variable for further consideration.

## INTERPRETING A FACTOR MATRIX

The task of interpreting a factor-loading matrix to identify the structure among the variables can at first seem overwhelming. The researcher must sort through all the factor loadings (remember, each variable has a loading on each factor) to identify those most indicative of the underlying structure. Even a fairly simple analysis of 15 variables on four factors necessitates evaluating and interpreting 60 factor loadings. Using the criteria for interpreting loadings described in the previous section, the researcher finds those distinctive variables for each factor and looks for a correspondence to the conceptual foundation or the managerial expectations for the research to assess practical significance.

In previous discussions we have used the term simple structure as an easily interpretable factor solution. There is no single best way to define simple structure, but Thurstone [56] laid out some general guidelines which if achieved provide an easily interpreted factor solution:

- 1 **Each variable:** should have at least one very low loading (under  $\pm .10$ ).
- 2 **Each factor:** should have at least as many very low loadings as there are factors.
- 3 **Each pair of factors:**
  - a. Some variables have a significant loading (greater than  $.3$  or  $.4$ ) on one factor and very low on the other.
  - b. A substantial percentage of very low loadings on each factor when there are four or more factors.
  - c. Relatively few cross-loadings (i.e., variables with significant loadings on each factor).

In short, the basic idea of simple structure is:

- each variable has a high/significant loading on one factor only, and
- each factor has high/significant loadings for only a subset of items.

Achieving simple structure in many cases is not totally achieved, but is always a desired objective to improve interpretability. In all instances it requires a combination of applying objective criteria with managerial judgment. By following the five-step procedure outlined next, the process can be simplified considerably. After the process is discussed, a brief example will be used to illustrate the process.

**Step 1: Examine the Factor Matrix of Loadings** The factor-loading matrix contains the factor loading of each variable on each factor. They may be either rotated or unrotated loadings, but as discussed earlier, rotated loadings are usually used in factor interpretation unless data reduction is the sole objective. Typically, the factors are arranged as columns; thus, each column of numbers represents the loadings of a single factor.

If an oblique rotation has been used, two matrices of factor loadings are provided. The first is the **factor pattern matrix**, which has loadings that represent the unique contribution of each variable to the factor. The second is

the **factor structure matrix**, which has simple correlations between variables and factors, but these loadings contain both the unique variance between variables and factors and the correlation among factors. As the correlation among factors becomes greater, it becomes more difficult to distinguish which variables load uniquely on each factor in the factor structure matrix. Thus, most researchers report the results of the factor pattern matrix.

**Step 2: Identify the Significant Loading(s) for Each Variable** The search for simple structure should start with the first variable on the first factor and move horizontally from left to right, looking for the highest loading for that variable on any factor. When the highest loading (largest absolute factor loading) is identified, it should be underlined (highlighted) if significant as determined by the criteria discussed earlier. Attention then focuses on the second variable and, again moving from left to right horizontally, looking for the highest loading for that variable on any factor and underlining it. This procedure should continue for each variable until all variables have been reviewed for their highest loading on a factor.

Most factor solutions, however, do not result in a simple structure solution (a single high loading for each variable on only one factor). Thus, the researcher will, after identifying the highest loading for a variable, continue to evaluate the factor matrix by underlining all significant loadings for a variable on all the factors. The process of interpretation would be greatly simplified if each variable had only one significant variable. In practice, however, the researcher may find that one or more variables each has moderate-size loadings on several factors, all of which are significant, and the job of interpreting the factors is much more difficult. When a variable is found to have more than one significant loading, it is termed a **cross-loading**.

The difficulty arises because a variable with several significant loadings (a cross-loading) must be used in labeling all the factors on which it has a significant loading. Yet how can the factors be distinct and potentially represent separate concepts when they “share” variables? This is particularly problematic when engaged in scale development where variables are considered members of only one scale (represented by a factor). Ultimately, the objective is to minimize the number of significant loadings on each row of the factor matrix (i.e., make each variable associate with only one factor). The researcher may find that different rotation methods eliminate any cross-loadings and thus define a simple structure. If simple structure is the objective (i.e., variable interpretation is important), and a variable persists in having cross-loadings, it becomes a candidate for deletion.

So when does a cross-loading represent enough of a problem that an item needs to be considered for removal? While there is mention of the cross-loading problem in almost every discussion on factor interpretation, no frameworks have been developed for identifying and then quantifying their impact. We propose a set of guidelines that help identify which cross-loadings substantively detract from a simple structure objective. The guidelines are based on two basic principles:

**COMPARE VARIANCES, NOT LOADINGS** The fundamental objective of factor extraction is to account for as much of a variable’s variance as possible and is represented as the squared loading. Variances also have the benefit of being a constant measure for comparison (i.e., a difference in variance of .10 is the same across any comparison), but this is not found by directly comparing loadings. For example, the variance represented by a difference in loadings of .1 varies based on the size of the loadings. Comparing loadings of .7 and .6 (a difference of .10) represents a difference in variance of .13 ( $.7^2 - .6^2 = .13$ ). Yet when we compare loadings of .4 and .3, the difference is only a variance of .07 ( $.4^2 - .3^2 = .07$ ). So when we want to truly understand the impact of one loading compared to another, we should compare the differences in variance rather than just the difference in loadings.

**COMPARE AS A RATIO OF VARIANCES** The most direct measure is the ratio of two variances (larger variance / smaller variance) which “scales” the variance difference to the size of the two variances. Just as we noted that a difference in loadings is not equal to the differences in variance as loadings change, the same holds true for the *relative magnitude* of the variance difference. As an example, assume a variance difference of 10 percent. This variance difference seems much more problematic when the two variances are 50 percent and 40 percent (i.e., loadings of .71 and .63) versus when the two variances are 20 percent and 10 percent (i.e., loadings of .44 and .31). So what would illustrate this difference more precisely—a ratio of the larger variance to the smaller variance. In the first case, the ratio would be

1.25, while in the second it would be 2.0. Thus, as the ratio increases, the distinctiveness of the relative magnitude increases.

Building upon these two principles, a simple three-step process emerges for each potential cross-loading:

- 1** Both loadings for a variable must be above the threshold for significance (e.g., .30 or .40)
- 2** Square each of the loadings and then compute the ratio of the larger loading to the smaller loading.
- 3** Designate the pair of loadings as follows based on the ratio:
  - a.** Between 1.0 and 1.5—*problematic cross-loading* and the variable with smaller loading a strong candidate for elimination to achieve simple structure.
  - b.** Between 1.5 and 2.0—*potential cross-loading*, with deletion of a variable based on interpretability of resulting factors.
  - c.** Greater than 2.0—*ignorable cross-loading*, where smaller loading, while significant, can be ignored for purposes of interpretation.

Figure 3.10 provides a simple example of applying these guidelines to a set of factor results for a selected set of loadings from a larger factor matrix. Of the four potential cross-loadings, two were deemed *ignorable* (Var 1, factors 2 and 3; Var 3, factors 1 and 3), one was deemed *potential* (Var 2, factors 1 and 2) and one found to be *problematic* (Var 4, factors 2 and 3). If simple structure and interpretability was desired, then Var 4 would likely be eliminated from the analysis, while Var 2 would be considered a candidate for deletion based on the larger set of variables. Variables 1 and 3 do not have any cross-loadings of note and are strong representative variables of their respective factors.

There are several observations to note about this process. First, in all cases, as the loadings become larger, any problematic cross-loadings become much more unlikely since there is little unaccounted for variance across the remaining variables. For example, a loading of .80 leaves only 34 percent variance ( $1.0 - .8^2 = 34\%$ ) for all of the variables other loadings. In a similar manner, as the maximum loading for a variable approaches the lower bounds of significance, there is the potential for many other loadings to be problematic since there is substantial remaining variance for the other factors. The authors encourage researchers to always keep the objective of simplicity of structure and interpretability as foremost considerations when applying this process. But it does represent a systematic framework for identifying problematic cross-loadings.

**Figure 3.10**  
Example: Identifying Cross-loading Problems

	Factor Loadings Matrix			Squared Loadings			Ratio <sup>1</sup>	Classification	Cross-loading
	Factor 1		Factor 2	Factor 1		Factor 3			
	Factor 1	Factor 2	Factor 3	Factor 2	Factor 1	Factor 3			
Var 1	0.30	0.60	0.40	0.09	0.36	0.16	2.25	Ignorable	
Var 2	0.68	0.54	0.35	0.46	0.29	0.12	1.59	Potential	
Var 3	0.75	0.37	0.47	0.56	0.14	0.22	2.55	Ignorable	
Var 4	0.37	0.53	0.63	0.14	0.28	0.40	1.41	Problematic	

<sup>1</sup> Ratio of larger to smaller squared loadings, designated by shaded values in table

**Step 3: Assess the Communalities of the Variables** Once all the significant loadings and cross-loading issues have been identified, the researcher should look for any variables that are not adequately accounted for by the factor solution. One simple approach is to identify any variable(s) lacking at least one significant loading. Another approach is to examine each variable's communality, representing the amount of variance accounted for by the factor solution for each variable. The researcher should view the communalities to assess whether the variables meet acceptable levels of explanation. For example, a researcher may specify that at least one-half of the variance of each variable must be taken into account. Using this guideline, the researcher would identify all variables with communalities less than .50 as not having sufficient explanation.

**Step 4: Respecify the Factor Model if Needed** When all the significant loadings and cross-loading issues have been identified, and the communalities examined, the researcher may find any one of several problems: (a) a variable has no significant loadings; (b) even with a significant loading, a variable's communality is deemed too low; or (c) a variable has a cross-loading. In this situation, the researcher can take any combination of the following remedies, listed from least to most extreme:

- *Ignore those problematic variables* and interpret the solution as is, which is appropriate if the objective is solely data reduction, but the researcher must still note that the variables in question are poorly represented in the factor solution.
- Evaluate each of those variables for *possible deletion*, depending on the variable's overall contribution to the research as well as its communality index. If the variable is of minor importance to the study's objective or has an unacceptable communality value, it may be eliminated and then the factor model respecified by deriving a new factor solution with those variables eliminated.
- Employ an *alternative rotation method*, particularly an oblique method if only orthogonal methods had been used.
- *Decrease/increase the number of factors retained* to see whether a smaller/larger factor structure will represent those problematic variables.
- *Modify the type of factor model used* (principal component versus common factor) to assess whether varying the type of variance considered affects the factor structure.

No matter which of these options are chosen by the researcher, the ultimate objective should always be to obtain a factor structure with both empirical and conceptual support. As we have seen, many "tricks" can be used to improve upon the structure, but the ultimate responsibility rests with the researcher and the conceptual foundation underlying the analysis.

**Step 5: Label the Factors** When an acceptable factor solution has been obtained in which all variables have a significant loading on a factor, the researcher attempts to assign some meaning to the pattern of factor loadings. Variables with higher loadings are considered more important and have greater influence on the name or label selected to represent a factor. Thus, the researcher will examine all the significant variables for a particular factor and, placing greater emphasis on those variables with higher loadings, will attempt to assign a name or label to a factor that accurately reflects the variables loading on that factor. The signs are interpreted just as with any other correlation coefficients. For each factor, like signs mean the variables are positively related, and opposite signs mean the variables are negatively related. In orthogonal solutions the factors are independent of one another. Therefore, the signs for factor loading relate only to the factor on which they appear, not to other factors in the solution.

This label is not derived or assigned by the factor analysis computer program. Rather, the label is intuitively developed by the researcher based on its appropriateness for representing the underlying dimensions of a particular factor. This procedure is followed for each extracted factor. The final result will be a name or label that is meaningful and represents each of the derived factors as accurately as possible.

As discussed earlier, the selection of a specific number of factors and the rotation method are interrelated. Several additional trial rotations may be undertaken, and by comparing the factor interpretations for several different trial rotations the researcher can select the number of factors to extract. In short, the ability to assign some meaning to the factors, or to interpret the nature of the variables, becomes an extremely important consideration in determining the number of factors to extract.

**An Example of Factor Interpretation** To serve as an illustration of factor interpretation, nine measures were obtained in a pilot test based on a sample of 202 respondents. After estimation of the initial results, further analysis indicated that a three-factor solution was appropriate. Thus, the researcher now has the task of interpreting the factor loadings of the nine variables.

Table 3.3 contains a series of factor-loading matrices. The first to be considered is the unrotated factor matrix (Part a). We will examine the unrotated and rotated factor-loading matrices through the five-step process described earlier.

**STEPS 1 AND 2: EXAMINE THE FACTOR-LOADING MATRIX AND IDENTIFY SIGNIFICANT LOADINGS** Given the sample size of 202, factor loadings of .40 and higher will be considered significant for interpretative purposes. Using this threshold for the factor loadings, we can see that the unrotated matrix does little to identify any form of simple structure. Five of the nine variables appear to have problematic cross-loadings, and for many of the other variables the significant loadings are fairly low. In this situation, rotation may improve our understanding of the relationship among variables.

As shown in Table 3.3b, the VARIMAX rotation improves the structure considerably in two noticeable ways. First, the loadings are improved for almost every variable, with the loadings more closely aligned to the objective of having a high loading on only a single factor. Second, now only one variable ( $V_1$ ) meets the first condition for a cross-loading with two loadings greater than .40. When we compare the squared loadings ( $.505^2 = .255$  and  $.462^2 = .213$ ), the difference is quite small, the first indicator of perhaps a problematic cross-loading. When the ratio of squared loadings is calculated (.255 / .213), the value of 1.19 definitely indicates a problematic cross-loading.

**STEP 3: ASSESS COMMUNALITIES** Only  $V_3$  has a communality that is low (.299). For our purposes  $V_3$  will be retained, but a researcher may consider the deletion of such variables in other research contexts. This illustrates an instance in which a variable has a significant loading, but may still be poorly accounted for by the factor solution.

**Table 3.3** Interpretation of a Hypothetical Factor-Loading Matrix

(a) Unrotated Factor-Loading Matrix				(b) VARIMAX Rotated Factor-Loading Matrix			
	Factor				Factor		
	1	2	3		1	2	3
$V_1$	.611	.250	-.204	$V_1$	.462	.099	.505
$V_2$	.614	-.446	.264	$V_2$	.101	.778	.173
$V_3$	.295	-.447	.107	$V_3$	-.134	.517	.114
$V_4$	.561	-.176	-.550	$V_4$	-.005	.184	.784
$V_5$	.589	-.467	.314	$V_5$	.087	.801	.119
$V_6$	.630	-.102	-.285	$V_6$	.180	.302	.605
$V_7$	.498	.611	.160	$V_7$	.795	-.032	.120
$V_8$	.310	.300	.649	$V_8$	.623	.293	-.366
$V_9$	.492	.597	-.094	$V_9$	.694	-.147	.323

(c) Simplified Rotated Factor-Loading Matrix <sup>1</sup>				(d) Rotated Factor-Loading Matrix with $V_1$ Deleted <sup>2</sup>			
	Factor				Factor		
	1	2	3		1	2	3
$V_7$	.795			$V_2$	.807		
$V_9$	.694			$V_5$	.803		
$V_8$	.623			$V_3$	.524		
$V_5$		.801		$V_7$		.802	
$V_2$		.778		$V_9$		.686	
$V_3$		.517		$V_8$		.655	
$V_4$			.784	$V_4$			.851
$V_6$			.605	$V_6$			.717
$V_1$	.462		.505				

<sup>1</sup> Loadings less than .40 are not shown and variables are sorted by highest loading.

<sup>2</sup>  $V_1$  deleted from the analysis, loadings less than .40 are not shown, and variables are sorted by highest loading.

**STEP 4: RESPECIFY THE FACTOR MODEL IF NEEDED** If we set a threshold value of .40 for loading significance and rearrange the variables according to loadings, the pattern shown in Table 3.3c emerges. Variables  $V_7$ ,  $V_9$ , and  $V_8$  all load highly on factor 1; factor 2 is characterized by variables  $V_5$ ,  $V_2$ , and  $V_3$ ; and factor 3 has two distinctive characteristics ( $V_4$  and  $V_6$ ). Only  $V_1$  has a problematic cross-loading, with significant loadings on both factors 1 and 3. Given that at least two variables are given on both of these factors,  $V_1$  is deleted from the analysis and the loadings recalculated.

**STEP 5: LABEL THE FACTORS** As shown in Table 3.3d, the factor structure for the remaining eight variables is now very well defined, representing three distinct groups of variables that the researcher may now utilize in further research.

As the preceding example shows, the process of factor interpretation involves both objective and subjective judgments. The researcher must consider a wide range of issues, all the time never losing sight of the ultimate goal of defining the best structure of the set of variables. Although many details are involved, some of the general principles are found in Rules of Thumb 3-6.

## Stage 6: Validation of Exploratory Factor Analysis

The sixth stage involves assessing the degree of generalizability of the results to the population and the potential influence of individual cases or respondents on the overall results. The issue of generalizability is critical for each of the multivariate methods, but it is especially relevant for the interdependence methods of exploratory factor analysis and cluster analysis because they describe a data structure that should be representative of the population as a whole. In the validation process, the researcher must address a number of issues in the area of research design and data characteristics as discussed next.

### USE OF REPLICATION OR A CONFIRMATORY PERSPECTIVE

The most direct methods of validating the results are (a) to assess the replicability/generalizability of the results, either with a split sample in the original dataset or with a separate sample, or (b) pursue a confirmatory analysis. If replication is performed, the comparison of two or more factor model results has always been problematic. Several methods have been proposed, ranging from a simple matching index [13] to programs (FMATCH) designed

### Interpreting the Factors

An optimal factor structure, many times termed simple structure, exists when all variables have high loadings only on a single factor and very low loadings on all other factors.

Cross-loadings of a variable (loadings on two factors) can be evaluated by the ratio of their squared loadings and classified as problematic (ratio between 1.0 and 1.5), potential (ratio between 1.5 and 2.0) or ignorable (ratio greater than 2.0). Problematic and perhaps even potential cross-loadings are deleted unless theoretically justified or the objective is strictly data reduction.

Variables should generally have communalities of greater than .50 to be retained in the analysis.

Respecification of a exploratory factor analysis results can include such options as the following:

- Deleting a variable(s)
- Changing rotation methods
- Increasing or decreasing the number of factors

specifically to assess the correspondence between factor matrices [53]. These methods have had sporadic use, due in part to (1) their perceived lack of sophistication, and (2) the unavailability of software or analytical programs to automate the comparisons. Even given these issues, these methods provide some objective basis for comparison when a confirmatory approach is not pursued. The emergence of confirmatory factor analysis (CFA) through structural equation modeling has provided a second approach where direct comparisons are possible. This approach is more complicated, requires specialized software and is discussed in greater detail in Chapters 9 and 10.

### ASSESSING FACTOR STRUCTURE STABILITY

Another aspect of generalizability is the stability of the factor model results. Factor stability is primarily dependent on the sample size and on the number of cases per variable. The researcher is always encouraged to obtain the largest sample possible and develop parsimonious models to increase the cases-to-variables ratio. If sample size permits, the researcher may wish to randomly split the sample into two subsets and estimate factor models for each subset (see discussion in prior section). Comparison of the two resulting factor matrices will provide an assessment of the robustness of the solution across the sample.

### DETECTING INFLUENTIAL OBSERVATIONS

In addition to generalizability, another issue of importance to the validation of exploratory factor analysis is the detection of influential observations. Discussions in Chapter 2 on the identification of outliers and in Chapter 5 on the influential observations in regression both have applicability in factor analysis. The researcher is encouraged to estimate the model with and without observations identified as outliers to assess their impact on the results. If omission of the outliers is justified, the results should have greater generalizability. Also, as discussed in Chapter 5, several measures of influence that reflect one observation's position relative to all others (e.g., covariance ratio) are applicable to factor analysis as well. Finally, the complexity of methods proposed for identifying influential observations specific to factor analysis [14] limits the application of these methods.

## Stage 7: Data Reduction—Additional Uses of Exploratory Factor Analysis Results

Up to this point in the chapter we have focused on the data summarization process, involving the selection of the factor model to be used, the number of factors to retain, and potentially the interpretation process. There may be instances where data summarization will suffice by providing an empirical basis for judging the structure of the variables and the impact of this structure when interpreting the results from other multivariate techniques. One such use is as initial examination of the data preceding a confirmatory factor analysis. But in most other situations, the researcher will engage in data summarization and then proceed to data reduction. Here the purpose is to generally extend the factor results by creating appropriate “replacement” variables representing each factor for subsequent application to other statistical techniques. The two options include the following:

- *Selecting the variable with the highest factor loading* as a surrogate representative for a particular factor dimension.
- *Replacing the original set of variables* with an entirely new, smaller set of variables created either from *summed scales* or *factor scores*.

Either option will provide new variables for use, for example, as the independent variables in a regression or discriminant analysis, as dependent variables in multivariate analysis of variance, or even as the clustering variables in cluster analysis. We discuss each of these options for data reduction in the following sections.

## SELECTING SURROGATE VARIABLES FOR SUBSEQUENT ANALYSIS

If the researcher's objective is simply to identify appropriate variables for subsequent application with other statistical techniques, the researcher has the option of examining the factor matrix and selecting the variable with the highest factor loading on each factor to act as a **surrogate variable** that is representative of that factor. This approach is simple and direct only when one variable has a factor loading that is substantially higher than all other factor loadings. In many instances, however, the selection process is more difficult because two or more variables have loadings that are significant and fairly close to each other, yet only one is chosen as representative of a particular dimension. This decision should be based on the researcher's *a priori* knowledge of theory that may suggest that one variable more than the others would logically be representative of the dimension. Also, the researcher may have knowledge suggesting that a variable with a loading slightly lower is in fact more reliable than the highest-loading variable. In such cases, the researcher may choose the variable that is loading slightly lower as the best variable to represent a particular factor.

The approach of selecting a single surrogate variable as representative of the factor—although simple and maintaining the original variable—has several potential disadvantages.

- It does *not address the issue of measurement error* encountered when using single measures (see the following section for a more detailed discussion).
- It also runs the *risk of potentially misleading results by selecting only a single variable to represent a perhaps more complex result*. For example, assume that variables representing price competitiveness, product quality, and value were all found to load highly on a single factor. The selection of any one of these separate variables would create substantially different interpretations in any subsequent analysis, yet all three may be so closely related as to make any definitive distinction impossible.

In instances where several high loadings complicate the selection of a single variable, the researcher may have no choice but to employ factor analysis as the basis for calculating a summed scale or factor scores instead of the surrogate variable. The objective, just as in the case of selecting a single variable, is to best represent the basic nature of the factor or component.

## CREATING SUMMATED SCALES

Chapter 1 introduced the concept of a **summated scale**, which is formed by combining several individual variables into a single **composite measure**. In simple terms, all of the variables loading highly on a factor are combined, and the total—or more commonly the average score of the variables—is used as a replacement variable. A summated scale provides two specific benefits.

**Reducing Measurement Error** First, it provides a *means of overcoming to some extent the measurement error* inherent in all measured variables. **Measurement error** is the degree to which the observed values are not representative of the actual values due to any number of reasons, ranging from actual errors (e.g., data entry errors) to the inability of individuals to accurately provide information. The impact of measurement error is to partially mask any relationships (e.g., correlations or comparison of group means) and make the estimation of multivariate models more difficult. The summated scale reduces measurement error by using multiple **indicators** (variables) to reduce the reliance on a single response. By using the average or typical response to a set of related variables, the measurement error that might occur in a single question will be reduced.

**Represent Multiple Aspects of a Concept in a Single Measure** Many times we employ more variables in our multivariate models in an attempt to represent the many facets of a concept that we know is quite complex. But in doing so, we complicate the interpretation of the results because of the redundancy in the items (i.e., multicollinearity) associated with the concept. Thus, we would like not only to accommodate the richer descriptions of concepts by using multiple variables, but also to maintain parsimony in the number of variables in our multivariate models. The summated scale, when properly constructed, does combine the multiple indicators into a single measure representing what is held in common across the set of measures.

The process of **scale development** has theoretical and empirical foundations in a number of disciplines, including psychometric theory, sociology, and marketing. Although a complete treatment of the techniques and issues involved are beyond the scope of this chapter, a number of excellent sources are available for further reading on this subject [1, 15, 29, 45, 48] and there is further discussion in Chapters 9 and 10. Additionally, a series of compilations of existing scales may be applied in a number of situations [3, 8, 47]. We discuss here, however, four issues basic to the construction of any summated scale: conceptual definition, dimensionality, reliability, and construct validity.

**CONCEPTUAL DEFINITION** The starting point for creating any summated scale is its **conceptual definition**. The conceptual definition specifies the theoretical basis for the summated scale by defining the concept being represented in terms applicable to the research context. In academic research, theoretical definitions are based on prior research that defines the character and nature of a concept. In a managerial setting, specific concepts may be defined that relate to proposed objectives, such as image, value, or satisfaction. In either instance, creating a summated scale is always guided by the conceptual definition specifying the type and character of the items that are candidates for inclusion in the scale.

**Content validity** is the assessment of the correspondence of the variables to be included in a summated scale and its conceptual definition. Also known as **face validity**, this form of construct validity (see discussion later in this section) subjectively assesses the correspondence between the individual items and the concept through ratings by expert judges, pretests with multiple subpopulations, or other means. The objective is to ensure that the selection of scale items extends past just empirical issues to also match the conceptual definition and include theoretical and practical considerations [15, 48].

**DIMENSIONALITY** An underlying assumption and essential requirement for creating a summated scale is that the items are **unidimensional**, meaning that they are strongly associated with each other and represent a single concept [29, 38]. Factor analysis plays a pivotal role in making an empirical assessment of the dimensionality of a set of items by determining the number of factors and the loadings of each variable on the factor(s). The test of unidimensionality is that each summated scale should consist of items loading highly on a single factor [2, 29, 38, 41]. If a summated scale is proposed to have multiple dimensions, each dimension should be reflected by a separate factor. The researcher can assess unidimensionality with either exploratory factor analysis, as discussed in this chapter, or confirmatory factor analysis, as described in Chapters 9 and 10.

**RELIABILITY** **Reliability** is an assessment of the degree of consistency between multiple measurements of a variable. One form of reliability is *test-retest*. For this method, the consistency is measured between the responses for an individual at two points in time. The objective is to ensure that responses are not too varied across time periods so that a measurement taken at any point in time is reliable. A second and more commonly used measure of reliability is *internal consistency*, which applies to the consistency among the variables in a summated scale. The rationale for internal consistency is that the individual items or indicators of the scale should all be measuring the same construct and thus be highly intercorrelated [15, 41].

Because no single item is a perfect measure of a concept, we must rely on a series of diagnostic measures to assess internal consistency.

**Single Items** The first measures we consider relate to *each separate item*, including the item-to-total correlation (the correlation of the item to the summated scale score) and the inter-item correlation (the correlation among items). Rules of thumb suggest that the item-to-total correlations exceed .50 and that the inter item correlations exceed .30 [48].

**Cronbach's Alpha** The second type of diagnostic measure is the *reliability coefficient*, which assesses the consistency of the entire scale, with **Cronbach's alpha** [19, 41, 44] being the most widely used measure. The generally agreed-upon lower limit for Cronbach's alpha is .70 [48, 47], although it may decrease to .60 in exploratory research [48]. One issue in assessing Cronbach's alpha is its positive relationship to the number of items in the scale. Because increasing the number of items, even with the same degree of intercorrelation, will increase the reliability value, researchers must place more stringent requirements for scales with large numbers of items.

**CFA Measures** Also available are *reliability measures derived from confirmatory factor analysis*. Included in these measures are the composite reliability and the average variance extracted, both discussed in greater detail in Chapter 10.

Each of the major statistical programs now has reliability assessment modules or programs, such that the researcher is provided with a complete analysis of both item-specific and overall reliability measures. Any summated scale should be analyzed for reliability to ensure its appropriateness before proceeding to an assessment of its validity.

**CONSTRUCT VALIDITY** Having ensured that a scale (1) conforms to its conceptual definition, (2) is unidimensional, and (3) meets the necessary levels of reliability, the researcher must make one final assessment: construct validity.

**Construct validity** is the extent to which a scale or set of measures accurately represents the concept of interest. We already described one form of construct validity—content or face validity—in the discussion of conceptual definitions. Other forms of validity are measured empirically by the correlation between theoretically defined sets of variables. The three most widely accepted forms of validity are convergent, discriminant, and nomological validity [10, 45].

**Convergent Validity** Assess the *degree to which two measures of the same concept are correlated*. Here the researcher may look for alternative measures of a concept and then correlate them with the summated scale. High correlations here indicate that the scale is measuring its intended concept.

**Discriminant Validity** The *degree to which two conceptually similar concepts are distinct*. The empirical test is again the correlation among measures, but this time the summated scale is correlated with a similar, but conceptually distinct, measure. Now the correlation should be low, demonstrating that the summated scale is sufficiently different from the other similar concept.

**Nomological Validity** Refers to the *degree that the summated scale makes accurate predictions of other concepts in a theoretically-based model*. The researcher must identify theoretically supported relationships from prior research or accepted principles and then assess whether the scale has corresponding relationships. In summary, convergent validity confirms that the scale is correlated with other known measures of the concept; discriminant validity ensures that the scale is sufficiently different from other similar concepts to be distinct; and nomological validity determines whether the scale demonstrates the relationships shown to exist based on theory or prior research.

A number of differing methods are available for assessing validity, ranging from the multitrait, multimethod (MTMM) matrices to structural equation-based approaches. Although beyond the scope of this text, numerous available sources address both the range of methods available and the issues involved in the specific techniques [10, 33, 45] as well as SEM analyses (see Chapters 9–13).

**Calculating Summated Scales** Calculating a summated scale is a straightforward process whereby the items comprising the summated scale (i.e., the items with high loadings from the factor analysis) are summed or averaged. The most common approach is to take the average of the items in the scale, which provides the researcher with complete control over the calculation and facilitates ease of use in subsequent analyses.

Whenever variables have both positive and negative loadings within the same factor, either the variables with the positive or the negative loadings must have their data values reversed. Typically, the variables with the negative loadings are reverse scored so that the correlations, and the loadings, are now all positive within the factor. **Reverse scoring** is the process by which the data values for a variable are reversed so that its correlations with other variables are reversed (i.e., go from negative to positive). For example, on our scale of 0 to 10, we would reverse score a variable by subtracting the original value from 10 (i.e., reverse score = 10 – original value). In this way, original scores of 10 and 0 now have the reversed scores of 0 and 10. All distributional characteristics are retained; only the distribution is reversed.

The purpose of reverse scoring is to prevent a canceling out of variables with positive and negative loadings. Let us use as an example of two variables with a negative correlation.

We are interested in combining  $V_1$  and  $V_2$ , with  $V_1$  having a positive loading and  $V_2$  a negative loading. If 10 is the top score on  $V_1$ , the top score on  $V_2$  would be 0. Now assume two cases. In case 1,  $V_1$  has a value of 10 and  $V_2$  has a value of 0 (the best case). In the second case,  $V_1$  has a value of 0 and  $V_2$  has a value of 10 (the worst case). If  $V_2$  is not reverse scored, then the scale score calculated by adding the two variables for both cases 1 and 2 is 10, showing no difference, whereas we know that case 1 is the best and case 2 is the worst. If we reverse

score  $V_2$ , however, the situation changes. Now case 1 has values of 10 and 10 on  $V_1$  and  $V_2$ , respectively, and case 2 has values of 0 and 0. The summated scale scores are now 20 for case 1 and 0 for case 2, which distinguishes them as the best and worst situations.

**Summary** Summated scales, one of the recent developments in academic research, has experienced increased application in applied and managerial research as well. The ability of the summated scale to portray complex concepts in a single measure while reducing measurement error makes it a valuable addition in any multivariate analysis. Exploratory factor analysis provides the researcher with an empirical assessment of the interrelationships among variables, essential in forming the conceptual and empirical foundation of a summated scale through assessment of content validity and scale dimensionality (see Rules of Thumb 3-7).

## COMPUTING FACTOR SCORES

The third option for creating a smaller set of variables to replace the original set is the computation of factor scores. **Factor scores** are also composite measures of each factor computed for each subject. Conceptually the factor score represents the degree to which each individual scores high on the group of items with high loadings on a factor. Thus, higher values on the variables with high loadings on a factor will result in a higher factor score. The one key characteristic that differentiates a factor score from a summated scale is that the *factor score is computed-based on the factor loadings of all variables on the factor, whereas the summated scale is calculated by combining only selected variables*. Therefore, although the researcher is able to characterize a factor by the variables with the highest loadings, consideration must also be given to the loadings of other variables, albeit lower, and their influence on the factor score.

Most statistical programs can easily compute factor scores for each respondent. By selecting the factor score option, these scores are saved for use in subsequent analyses. The one disadvantage of factor scores has long been that they are *not easily replicated across studies because they are based on the factor matrix, which is derived separately in each study*. This issue, however, has been substantially reduced by the emergence of scoring procedures in the major software packages. A **scoring procedure** is a process which saves the scoring coefficients from the factor matrix and then allows them to be applied to new datasets. In this manner, exploratory factor analysis can be replicated to generate factor scores across any number of datasets, all based on the original analysis. Facilitated by procedures such as PROC SCORE in SAS and similar procedures in SPSS and other packages, replication is now widely available with minimum effort.

### Summated Scales

A summated scale is only as good as the items used to represent the construct; even though it may pass all empirical tests, it is useless without theoretical justification.

Never create a summated scale without first assessing its unidimensionality with exploratory or confirmatory factor analysis.

Once a scale is deemed unidimensional, its reliability score, as measured by Cronbach's alpha:

should exceed a threshold of .70, although a .60 level can be used in exploratory research;  
the threshold should be raised as the number of items increases, especially as the number of items approaches 10 or more.

With reliability established, validity should be assessed in terms of:

convergent validity scale correlates with other like scales;  
discriminant validity scale is sufficiently different from other related scales;  
nomological validity scale "predicts" as theoretically suggested.

## SELECTING AMONG THE THREE METHODS

To select among the three data reduction options, the researcher must make a series of decisions, weighing the advantages and disadvantages of each approach with the research objectives. The guidelines in Rules of Thumb 3-8 address the fundamental trade-offs associated with each approach.

The decision rule, therefore, would be as follows:

- If data are used only in the original sample, interpretation is less important or orthogonality must be maintained, factor scores are suitable.
- If generalizability or transferability is desired, then summated scales or surrogate variables are more appropriate. If the summated scale is a well-constructed, valid, and reliable instrument, then it is probably the best alternative.
- If the summated scale is untested and exploratory, with little or no evidence of reliability or validity, surrogate variables should be considered if additional analysis is not possible to improve the summated scale.

## Data Reduction—Representing Factor Analysis in Other Analyses

### The Single Surrogate Variable

Advantages:

simple to administer and interpret.

Disadvantages:

does not represent all “facets” of a factor;  
prone to measurement error.

### Factor Scores

Advantages:

represent all variables loading on the factor;  
best method for complete data reduction;  
are by default orthogonal and can avoid complications caused by multicollinearity.

Disadvantages:

interpretation more difficult because all variables contribute through loadings;  
requires additional procedures to replicate across studies.

### Summated Scales

Advantages:

compromise between the surrogate variable and factor score options;  
reduce measurement error;  
represent multiple facets of a concept;  
easily replicated across studies.

Disadvantages:

include only the variables that load highly on the factor and excludes those having little or marginal impact;  
not necessarily orthogonal;  
require extensive analysis of reliability and validity issues.

## An Illustrative Example

In the preceding sections, the major questions concerning the application of exploratory factor analysis were discussed within the model-building framework introduced in Chapter 1. To clarify these topics further, we use an illustrative example of the application of exploratory factor analysis based on data from the database presented in Chapter 1. Our discussion of the empirical example also follows the six-stage model-building process. The first three stages, common to either principal component or common factor analysis, are discussed first. Then, stages 4–6 for principal component analysis will be discussed, along with examples of the additional use of factor results. We conclude with an examination of the differences for common factor analysis in stages 4 and 5.

### STAGE 1: OBJECTIVES OF FACTOR ANALYSIS

Exploratory factor analysis can identify the structure of a set of variables as well as provide a process for data reduction. In our example, the perceptions of HBAT on 13 attributes ( $X_6$  to  $X_{18}$ ) are examined for the following reasons:

- *Data Summarization with Interpretation—Understand whether these perceptions can be “grouped.”* Even the relatively small number of perceptions examined here presents a complex picture of 78 separate correlations. By grouping the perceptions and then engaging in the steps of factor interpretation, HBAT will be able to see the big picture in terms of understanding its customers and what dimensions the customers think about HBAT.
- *Data Reduction—Reduce the 13 variables to a smaller number of composite factors.* If the 13 variables can be represented in a smaller number of composite variables, then the other multivariate techniques can be made more parsimonious. Of course, this approach assumes that a certain degree of underlying order exists in the data being analyzed.

Either or both objectives are typical research questions, making factor analysis applicable to a wide range of research situations. Moreover, as the basis for summated scale development, it has gained much wider use in recent years.

### STAGE 2: DESIGNING A FACTOR ANALYSIS

Understanding the structure of the perceptions of variables requires *R*-type factor analysis and a correlation matrix between variables, not respondents. All the variables are metric and constitute a homogeneous set of perceptions appropriate for exploratory factor analysis.

The sample size in this example is an 8:1 ratio of observations to variables, which falls within acceptable limits. Also, the sample size of 100 provides an adequate basis for the calculation of the correlations between variables.

### STAGE 3: ASSUMPTIONS IN FACTOR ANALYSIS

The underlying statistical assumptions influence exploratory factor analysis to the extent that they affect the derived correlations. Departures from normality, homoscedasticity, and linearity can diminish correlations between variables. These assumptions are examined in Chapter 2, and the reader is encouraged to review the findings. The researcher must also assess the factorability of the correlation matrix.

**Visual Examination of the Correlations** Table 3.4 shows the correlation matrix for the 13 perceptions of HBAT. Inspection of the correlation matrix reveals that 29 of the 78 correlations (37%) are significant at the .01 level, which provides an adequate basis for proceeding to an empirical examination of adequacy for factor analysis on both an overall basis and for each variable. Tabulating the number of significant correlations per variable finds a range from 0 ( $X_{15}$ ) to 9 ( $X_{17}$ ). Although no limits are placed on what is too high or low, variables that have no significant correlations may not be part of any factor, and if a variable has a large number of correlations, it may be part of several factors. We can note these patterns and see how they are reflected as the analysis proceeds.

**Table 3.4** Assessing the Appropriateness of Exploratory Factor Analysis: Correlations, Measures of Sampling Adequacy, and Partial Correlations Among Variables

Correlations Among Variables									Correlations Significant at .01 Level				
	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$	$X_{16}$	$X_{17}$	$X_{18}$
$X_6$ Product Quality	1.000	-.137	.096	.106	-.053	<b>.477</b>	-.152	<b>-.401</b>	.088	.027	.104	<b>-.493</b>	.028
$X_7$ E-Commerce	1.000	.001	.140	<b>.430</b>	-.053	<b>.792</b>	.229	.052	-.027	.156	<b>.271</b>	.192	.3
$X_8$ Technical Support	1.000	.097	-.063	.193	.017	-.271	<b>.797</b>	-.074	.080	-.186	.025	.2	
$X_9$ Complaint Resolution	1.000	.197	<b>.561</b>	<b>.230</b>	-.128	.140	.059	<b>.757</b>	<b>.395</b>	<b>.865</b>	.5		
$X_{10}$ Advertising	1.000	-.012	<b>.542</b>	-.134	.011	.084	.184	<b>.334</b>	<b>.276</b>	.4			
$X_{11}$ Product Line	1.000	-.061	<b>-.495</b>	<b>.273</b>	.046	<b>.424</b>	<b>-.378</b>	<b>.602</b>	.7				
$X_{12}$ Salesforce Image	1.000	<b>.265</b>	.107	.032	.195	<b>.352</b>	<b>.272</b>	.6					
$X_{13}$ Competitive Pricing	1.000	<b>-.245</b>	.023	-.115	<b>.471</b>	-.073	.6						
$X_{14}$ Warranty & Claims	1.000	.035	.197	-.170	.109	.3							
$X_{15}$ Packaging					1.000	.069	.094	.106	0				
$X_{16}$ Order & Billing						1.000	<b>.407</b>	<b>.751</b>	.4				
$X_{17}$ Price Flexibility							1.000	<b>.497</b>	.9				
$X_{18}$ Delivery Speed								1.000	.6				

Note: Bolded values indicate correlations significant at the .01 significance level.

Overall Measure of Sampling Adequacy: .609

Bartlett Test of Sphericity: 948.9

Significance: .000

Measures of Sampling Adequacy and Partial Correlations		$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$	$X_{16}$	$X_{17}$	$X_{18}$
$X_6$ Product Quality		.873												
$X_7$ E-Commerce		.038	.620											
$X_8$ Technical Support		-.049	-.060	.527										
$X_9$ Complaint Resolution		-.082	.117	-.150	.890									
$X_{10}$ Advertising		-.122	.002	.049	.092	.807								
$X_{11}$ Product Line		-.023	-.157	.067	-.152	-.101	.448							
$X_{12}$ Salesforce Image		-.006	-.729	.077	-.154	-.333	.273	.586						
$X_{13}$ Competitive Pricing		.054	-.018	.125	.049	.090	-.088	-.138	.879					
$X_{14}$ Warranty & Claims		.124	.091	-.792	.123	-.020	-.103	-.172	-.019	.529				
$X_{15}$ Packaging		-.076	.091	.143	.061	-.026	-.118	-.054	.015	-.138	.314			
$X_{16}$ Order & Billing		-.189	-.105	.160	-.312	.044	.044	.100	.106	-.250	.031	.859		
$X_{17}$ Price Flexibility		.135	-.134	.031	-.143	-.151	.953	.241	-.212	-.029	-.137	-.037	.442	
$X_{18}$ Delivery Speed		.013	.136	-.028	-.081	.064	-.941	-.254	.126	.070	.090	-.109	-.922	.533

Note: Measures of sampling adequacy (MSA) are on the diagonal, partial correlations in the off-diagonal.

**Bartlett's Test and MSA Values** The researcher can assess the overall significance of the correlation matrix with the Bartlett test and the factorability of the overall set of variables and individual variables using the measure of sampling adequacy (MSA). Because exploratory factor analysis will always derive factors, the objective is to ensure a base level of statistical correlation within the set of variables, such that the resulting factor structure has some objective basis.

In this example, the Bartlett's test finds that the correlations, when taken collectively, are significant at the .0001 level (see Table 3.4). But this test only indicates the presence of non-zero correlations, not the pattern of these correlations. More specific measures related to the patterns of variables and even specific variables are required.

The measure of sampling adequacy (MSA) looks not only at the correlations, but also at patterns between variables. In this situation the overall MSA value falls in the acceptable range (above .50) with a value of .609. Examination of the values for each variable, however, identifies three variables ( $X_{11}$ ,  $X_{15}$ , and  $X_{17}$ ) with MSA values under .50. Because  $X_{15}$  has the lowest MSA value, it will be omitted in the attempt to obtain a set of variables that can exceed the minimum acceptable MSA levels. Recalculating the MSA values after excluding  $X_{15}$  finds that  $X_{17}$  still has an individual MSA value below .50, so it is also deleted from the analysis. We should note at this point that  $X_{15}$  and  $X_{17}$  were the two variables with the lowest and highest number of significant correlations, respectively.

Table 3.5 contains the correlation matrix for the revised set of variables ( $X_{15}$  and  $X_{17}$  deleted) along with the measures of sampling adequacy and the Bartlett test value. In the reduced correlation matrix, 20 of the 55 correlations are statistically significant. As with the full set of variables, the Bartlett test shows that non-zero correlations exist at the significance level of .0001. The reduced set of variables collectively meets the necessary threshold of sampling adequacy with an MSA value of .653. Each of the variables also exceeds the threshold value, indicating that the reduced set of variables meets the fundamental requirements for factor analysis. Finally, examining the partial correlations shows only five with values greater than .50 ( $X_6-X_{11}$ ,  $X_7-X_{12}$ ,  $X_8-X_{14}$ ,  $X_9-X_{18}$ , and  $X_{11}-X_{18}$ ), which is another indicator of the strength of the interrelationships among the variables in the reduced set. It is of note that both  $X_{11}$  and  $X_{18}$  are involved in two of the high partial correlations. Collectively, these measures all indicate that the reduced set of variables is appropriate for factor analysis, and the analysis can proceed to the next stages.

## PRINCIPAL COMPONENT FACTOR ANALYSIS: STAGES 4–7

As noted earlier, factor analysis procedures are based on the initial computation of a complete table of intercorrelations among the variables (correlation matrix). The correlation matrix is then transformed through estimation of a factor model to obtain a factor matrix that contains factor loadings for each variable on each derived factor. The loadings of each variable on the factors are then interpreted to identify the underlying structure of the variables, in this example the perceptions of HBAT. These steps of factor analysis, contained in stages 4–7, are examined first for principal component analysis. Then, a common factor analysis is performed and comparisons made between the two factor models.

**Stage 4: Deriving Factors and Assessing Overall Fit** Given that the principal components method of extraction will be used first, the next decision is to select the number of components to be retained for further analysis. As discussed earlier, the researcher should employ a number of different criteria in determining the number of factors to be retained for interpretation, ranging from the more subjective (e.g., selecting a number of factors *a priori* or specifying the percentage of variance extracted) to the more objective (latent root criterion, scree test or parallel analysis) criteria.

**STOPPING RULES** Table 3.6 contains the information regarding the 11 possible factors and their relative explanatory power as expressed by their eigenvalues. In addition to assessing the importance of each component, we can also use the eigenvalues to assist in selecting the number of factors.

- *A priori criterion.* The researcher is not bound by preconceptions as to the number of factors that should be retained, but practical reasons of desiring multiple measures per factor (at least 2 and preferably 3) dictate that between three and five factors would be best given the 11 variables to be analyzed.

**Table 3.5** Assessing the Appropriateness of Factor Analysis for the Revised Set of Variables ( $X_{15}$  and  $X_{17}$  Deleted): Correlations, Measures of Sampling Adequacy, and Partial Correlations Among Variables

Correlations Among Variables		Correlations Significant at .01 Level													
		$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$	$X_{16}$	$X_{17}$	$X_{18}$	
$X_6$ Product Quality	1.000	-.137	.096	.106	-.053	<b>.477</b>	-.152	<b>-.401</b>	.088	.104	.028	.2			
$X_7$ E-Commerce	1.000	.001	.140	<b>.430</b>	-.053	<b>.792</b>	.229	.052	.156	.192	.2				
$X_8$ Technical Support	1.000	.097	-.063	.193	.017	<b>-.271</b>	<b>.797</b>	.080	.025						
$X_9$ Complaint Resolution	1.000	.197	<b>.561</b>	<b>.230</b>	-.128	.140	<b>.757</b>	<b>.865</b>	.4						
$X_{10}$ Advertising		1.000	-.012	<b>.542</b>	.134	.011	.184	<b>.276</b>	.3						
$X_{11}$ Product Line			1.000	-.061	<b>-.495</b>	<b>.273</b>	<b>.424</b>	<b>.602</b>	.6						
$X_{12}$ Salesforce Image				1.000	<b>.265</b>	.107	.195	<b>.272</b>	.5						
$X_{13}$ Competitive Pricing					1.000	<b>-.245</b>	-.115	-.073	.5						
$X_{14}$ Warranty & Claims						1.000	.197	.109	.3						
$X_{15}$ Order & Billing							1.000	<b>.751</b>	.3						
$X_{16}$ Delivery Speed									1.000					5	

Note: Bolded values indicate correlations significant at the .01 significance level.

Overall Measure of Sampling Adequacy: .653

Bartlett's Test of Sphericity: 619.3

Significance: .000

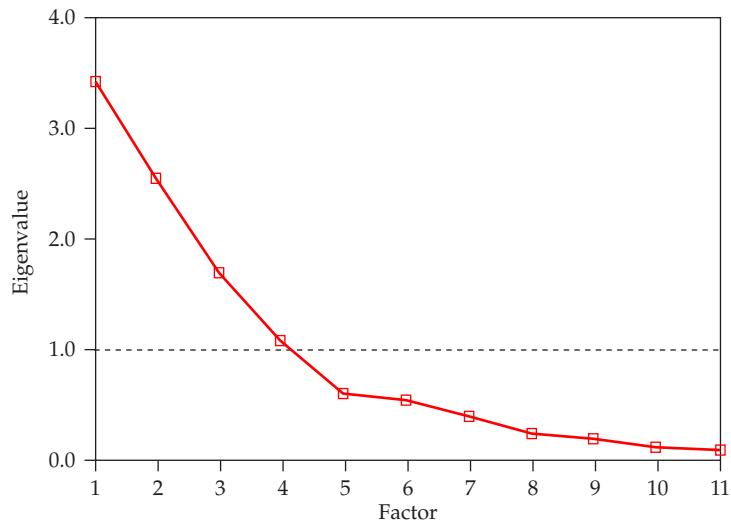
#### Measures of Sampling Adequacy and Partial Correlations

	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{16}$	$X_{18}$
$X_6$ Product Quality	.509										
$X_7$ E-Commerce	.061	.626									
$X_8$ Technical Support	-.045	-.068	.519								
$X_9$ Complaint Resolution	-.062	.097	-.156	.787							
$X_{10}$ Advertising	-.107	-.015	.062	.074	.779						
$X_{11}$ Product Line	-.503	-.101	.117	-.054	.143	.622					
$X_{12}$ Salesforce Image	-.042	-.725	.076	-.124	-.311	.148	.622				
$X_{13}$ Competitive Pricing	.085	-.047	.139	.020	.060	.386	-.092	.753			
$X_{14}$ Warranty & Claims	.122	.100	-.787	.127	-.032	-.246	-.175	-.028	.511		
$X_{16}$ Order & Billing	-.184	-.113	.160	-.322	.040	.261	.113	.101	-.250	.760	
$X_{18}$ Delivery Speed	.355	.040	.017	-.555	-.202	-.529	-.087	-.184	.100	-.369	.666

Note: Measures of sampling adequacy (MSA) are on the diagonal, partial correlations in the off-diagonal.

**Table 3.6** Results for the Extraction of Component Factors

Component	Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.43	31.2	31.2
2	2.55	23.2	54.3
3	1.69	15.4	69.7
4	1.09	9.9	79.6
5	.61	5.5	85.1
6	.55	5.0	90.2
7	.40	3.7	93.8
8	.25	2.2	96.0
9	.20	1.9	97.9
10	.13	1.2	99.1
11	.10	.9	100.0

**Figure 3.11**  
Scree Test for Component Analysis

- *Latent root criterion.* If we retain factors with eigenvalues greater than 1.0, four factors will be retained.
- *Percentage of variance criterion.* The four factors retained represent 79.6 percent of the variance of the 11 variables, deemed sufficient in terms of total variance explained.
- *Scree test.* As shown in Figure 3.11, the scree test indicates that four or perhaps five factors may be appropriate when considering the changes in eigenvalues (i.e., identifying the “elbow” in the eigenvalues at the fifth factor). In viewing the eigenvalue for the fifth factor, its low value (.61) relative to the latent root criterion value of 1.0 precluded its inclusion. If the eigenvalue had been quite close to 1, then it might be considered for inclusion as well.
- *Parallel analysis.* Table 3.7 contains the parallel analysis for the full set of variables as well as the final set of reduced variables. For the parallel analysis, the mean of the eigenvalues of the random datasets is given, along with the 95th percentile which may be used as an even more conservative threshold. For our purposes, we will use the mean values for comparison.

For the full set of 13 variables, we would select five factors based on the latent root criterion, although noting that the fifth factor had an eigenvalue (1.01)—barely exceeding 1. The parallel analysis would indicate four factors, as the mean of the eigenvalues in parallel analysis (1.14) exceeds the actual eigenvalue (1.01). Examining the reduced

**Table 3.7** Parallel Analysis as a Stopping Rule for Principal Components Analysis

PCA Results		Parallel Analysis	
Component	Eigenvalue	Mean	95th Percentile
1	3.57	1.64	1.79
2	3.00	1.47	1.57
3	1.74	1.35	1.44
4	1.29	1.24	1.30
5	1.01	1.14	1.22
6	0.62	1.05	1.11
7	0.55	0.96	1.02
8	0.45	0.88	0.93
9	0.28	0.81	0.88
10	0.20	0.74	0.80
11	0.17	0.66	0.74
12	0.13	0.57	0.63
13	0.01	0.49	0.55

PCA Results		Parallel Analysis	
Component	Eigenvalue	Mean	95th Percentile
1	3.43	1.57	1.72
2	2.55	1.40	1.51
3	1.69	1.27	1.36
4	1.09	1.17	1.24
5	0.61	1.06	1.12
6	0.55	0.97	1.04
7	0.40	0.88	0.95
8	0.25	0.80	0.87
9	0.20	0.72	0.79
10	0.13	0.63	0.70
11	0.10	0.53	0.60

set of 11 variables we would retain four components per the latent root criterion, but parallel analysis retains only three, with the mean eigenvalues greater than the fourth eigenvalue (1.17 versus 1.09). So in both cases parallel analysis is more conservative in retaining components, in each instance providing evidence that the final factor considered for retention, while passing the latent root criterion, may not be suitable. This is as expected when the final factors have eigenvalues very close to 1.

Combining all these criteria together is essential given that there is no single best method for determining the number of factors. In this case it leads to the conclusion to retain four factors for further analysis. When questions arise as to the appropriate number of factors, researchers are encouraged to evaluate the alternative solutions. As will be shown in a later section, the three- and five-factor solutions are somewhat less interpretable or useful, giving additional support to a four factor solution. More importantly, these results illustrate the need for multiple decision criteria in deciding the number of components to be retained.

**Stage 5: Interpreting the Factors** With four factors to be analyzed, the researcher now turns to interpreting the factors. Once the factor matrix of loadings has been calculated, the interpretation process proceeds by examining the unrotated and then rotated factor matrices for significant factor loadings and adequate communalities.

If deficiencies are found (i.e., cross-loadings or factors with only a single variable), respecification of the factors is considered. Once the factors are finalized, they can be described based on the significant factor loadings characterizing each factor.

**STEP 1: EXAMINE THE FACTOR MATRIX OF LOADINGS FOR THE UNROTATED FACTOR MATRIX** Factor loadings, in either the unrotated or rotated factor matrices, represent the degree of association (correlation) of each variable with each factor. The loadings take on a key role in interpretation of the factors, particularly if they are used in ways that require characterization as to the substantive meaning of the factors (e.g., as predictor variables in a dependence relationship). The objective of factor analysis in these instances is to maximize the association of each variable with a single factor, many times through rotation of the factor matrix. The researcher must make a judgment as to the adequacy of the solution in this stage and its representation of the structure of the variables and ability to meet the goals of the research. We will first examine the unrotated factor solution and determine whether the use of the rotated solution is necessary.

Table 3.8 presents the unrotated principal component analysis factor matrix. To begin the analysis, let us explain the numbers included in the table. Five columns of numbers are shown. The first four are the results for the four factors that are extracted (i.e., factor loadings of each variable on each of the factors). The fifth column provides summary statistics detailing how well each variable is explained by the four components (communality), which are discussed in the next section. The first row of numbers at the bottom of each column is the column sum of squared factor loadings (*eigenvalues*) and indicates the relative importance of each factor in accounting for the variance associated with the set of variables. Note that the sums of squares for the four factors are 3.427, 2.551, 1.691, and 1.087, respectively. As expected, the factor solution extracts the factors in the order of their importance, with factor 1 accounting for the most variance, factor 2 slightly less, and so on through all 11 factors. At the far right-hand side of the row is the number 8.756, which represents the total of the four eigenvalues ( $3.427 + 2.551 + 1.691 + 1.087$ ). The total of eigenvalues represents the total amount of variance extracted by the factor solution.

The total amount of variance explained by either a single factor or the overall factor solution can be compared to the total variation in the set of variables as represented by the **trace** of the factor matrix. The trace is the total variance to be explained and is equal to the sum of the eigenvalues of the variable set. In principal components analysis, the trace is equal to the number of variables because each variable has a possible eigenvalue of 1.0. By adding the percentages of trace for each of the factors (or dividing the total eigenvalues of the factors by the trace), we obtain the total percentage of trace extracted for the factor solution. This total is used as an index to determine how well a

**Table 3.8** Unrotated Component Analysis Factor Matrix

<b>Variables</b>	<b>Factor</b>				<b>Communality</b>
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	
$X_6$ Product Quality	.248	-.501	-.081	.670	.768
$X_7$ E-Commerce	.307	.713	.306	.284	.777
$X_8$ Technical Support	.292	-.369	.794	-.202	.893
$X_9$ Complaint Resolution	.871	.031	-.274	-.215	.881
$X_{10}$ Advertising	.340	.581	.115	.331	.576
$X_{11}$ Product Line	.716	-.455	-.151	.212	.787
$X_{12}$ Salesforce Image	.377	.752	.314	.232	.859
$X_{13}$ Competitive Pricing	-.281	.660	-.069	-.348	.641
$X_{14}$ Warranty & Claims	.394	-.306	.778	-.193	.892
$X_{15}$ Order & Billing	.809	.042	-.220	-.247	.766
$X_{16}$ Delivery Speed	.876	.117	-.302	-.206	.914
					Total
Sum of Squares (eigenvalue)	3.427	2.551	1.691	1.087	8.756
Percentage of trace <sup>a</sup>	31.15	23.19	15.37	9.88	79.59

<sup>a</sup>Trace = 11.0 (sum of eigenvalues)

particular factor solution accounts for what all the variables together represent. If the variables are all very different from one another, this index will be low. If the variables fall into one or more highly redundant or related groups, and if the extracted factors account for all the groups, the index will approach 100 percent.

The percentages of trace explained by each of the four factors (31.15%, 23.19%, 15.37%, and 9.88%, respectively) are shown as the last row of values of Table 3.7. The percentage of trace is obtained by dividing each factor's sum of squares (eigenvalues) by the trace for the set of variables being analyzed. For example, dividing the sum of squares of 3.427 for factor 1 by the trace of 11.0 results in the percentage of trace of 31.154 percent for factor 1. The index for the overall solution shows that 79.59 percent of the total variance ( $8.756 \div 11.0$ ) is represented by the information contained in the factor matrix of the four-factor solution. Therefore, the index for this solution is high, and the variables are in fact highly related to one another.

**STEP 2: IDENTIFY THE SIGNIFICANT LOADINGS IN THE UNROTATED FACTOR MATRIX** Having defined the various elements of the unrotated factor matrix, let us examine the factor-loading patterns. As discussed earlier, the factor loadings allow for the description of each factor and the structure in the set of variables.

As anticipated, the first factor accounts for the largest amount of variance in Table 3.8. The second factor is somewhat of a general factor, with half of the variables having a high loading (high loading is defined as greater than .40). The third factor has two high loadings, whereas the fourth factor only has one high loading. Based on this factor-loading pattern with a relatively large number of high loadings on factor 2 and only one high loading on factor 4, interpretation would be difficult and theoretically less meaningful. Therefore, the researcher should proceed to rotate the factor matrix to redistribute the variance from the earlier factors to the later factors. Rotation should result in a simpler and theoretically more meaningful factor pattern. However, before proceeding with the rotation process, we must examine the communalities to see whether any variables have communalities so low that they should be eliminated.

**STEP 3: ASSESS THE COMMUNALITIES OF THE VARIABLES IN THE UNROTATED FACTOR MATRIX** The row sum of squared factor loadings are referred to as *communalities*. The communalities show the amount of variance in a variable that is accounted for by all of the retained factors taken together. The size of the communality is a useful index for assessing how much variance in a particular variable is accounted for by the factor solution. Higher communality values indicate that a large amount of the variance in a variable has been extracted by the factor solution. Small communalities show that a substantial portion of the variable's variance is not accounted for by the factors. Although no statistical guidelines indicate exactly what is "large" or "small," practical considerations are consistent with a lower level of .50 for communalities in this analysis.

The communalities in Table 3.8 are shown at the far right side of the table. For instance, the communality value of .576 for variable  $X_{10}$  indicates that it has less in common with the other variables included in the analysis than does variable  $X_8$ , which has a communality of .893. Both variables, however, still share more than one-half of their variance with the four factors. All of the communalities are sufficiently high to proceed with the rotation of the factor matrix.

**STEPS 2 AND 3: ASSESS THE SIGNIFICANT FACTOR LOADING(S) AND COMMUNALITIES OF THE ROTATED FACTOR MATRIX** Given that the unrotated factor matrix did not have a completely clean set of factor loadings (i.e., had substantial cross-loadings or did not maximize the loadings of each variable on one factor), a rotation technique can be applied to hopefully improve the interpretation. In this case, the VARIMAX rotation is used and its impact on the overall factor solution and the factor loadings are described next.

**Applying the Orthogonal (VARIMAX) Rotation** The VARIMAX-rotated principal component analysis factor matrix is shown in Table 3.9. Note that the total amount of variance extracted is the same in the rotated solution as it was in the unrotated solution, 79.6 percent. Also, the communalities for each variable do not change when a rotation technique is applied. Still, two differences do emerge. First, the variance is redistributed so that the factor-loading pattern and the percentage of variance for each of the factors are slightly different. Specifically, in the VARIMAX-rotated factor solution, the first factor accounts for 26.3 percent of the variance, compared to 31.2 percent in the unrotated solution. Likewise, the other factors also change, the largest change being the fourth factor, increasing from 9.9 percent in the unrotated solution to 16.1 percent in the rotated solution. Thus, the explanatory power shifted slightly to a more

even distribution because of the rotation. Second, the interpretation of the factor matrix is simplified. As will be discussed in the next section, the factor loadings for each variable are maximized for each variable on one factor, except in any instances of cross-loadings.

With the rotation complete, the researcher now examines the *rotated factor matrix* for the patterns of significant factor loadings hoping to find a simplified structure. If any problems remain (i.e., nonsignificant loadings for one or more variables, cross-loadings, or unacceptable communalities), the researcher must consider respecification of the factor analysis through the set of options discussed earlier.

Our first examination is to see if “simple structure” is found in the rotated factor solution. As described earlier, the solution is evaluated from the perspectives of each variable, each factor and each pair of factors. Examination

**Table 3.9** VARIMAX-Rotated Component Analysis Factor Matrices: Full and Reduced Sets of Variables

<b>Full Set of Variables</b>	<b>VARIMAX-ROTATED LOADINGS<sup>a</sup></b>				
	<b>Factor</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
X <sub>18</sub> Delivery Speed	<b>.938</b>	.177	−.005	.052	.914
X <sub>9</sub> Complaint Resolution	<b>.926</b>	.116	.048	.091	.881
X <sub>16</sub> Order & Billing	<b>.864</b>	.107	.084	−.039	.766
X <sub>12</sub> Salesforce Image	.133	<b>.900</b>	.076	−.159	.859
X <sub>7</sub> E-Commerce	.057	<b>.871</b>	.047	−.117	.777
X <sub>10</sub> Advertising	.139	<b>.742</b>	−.082	.015	.576
X <sub>8</sub> Technical Support	.018	−.024	<b>.939</b>	.101	.893
X <sub>14</sub> Warranty & Claims	.110	.055	<b>.931</b>	.102	.892
X <sub>6</sub> Product Quality	.002	−.013	−.033	<b>.876</b>	.768
X <sub>13</sub> Competitive Pricing	−.085	.226	−.246	−.723	.641
X <sub>11</sub> Product Line	<b>.591</b>	−.064	.146	<b>.642</b>	.787
					Total
Sum of Squares (eigenvalue)	2.893	2.234	1.855	1.774	8.756
Percentage of trace	26.30	20.31	16.87	16.12	79.59

<sup>a</sup>Factor loadings greater than .40 are in bold and variables have been sorted by loadings on each factor.

<b>Reduced Set of Variables (X<sub>11</sub> deleted)</b>	<b>VARIMAX-ROTATED LOADINGS<sup>a</sup></b>				
	<b>Factor</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
X <sub>9</sub> Complaint Resolution	.933				.890
X <sub>18</sub> Delivery Speed	.931				.894
X <sub>16</sub> Order & Billing	.886				.806
X <sub>12</sub> Salesforce Image		.898			.860
X <sub>7</sub> E-Commerce		.868			.780
X <sub>10</sub> Advertising		.743			.585
X <sub>8</sub> Technical Support			.940		.894
X <sub>14</sub> Warranty & Claims			.933		.891
X <sub>6</sub> Product Quality				.892	.798
X <sub>13</sub> Competitive Pricing				−.730	.661
					Total
Sum of Squares (eigenvalue)	2.589	2.216	1.846	1.406	8.057
Percentage of trace	25.89	22.16	18.46	14.06	80.57

<sup>a</sup>Factor loadings less than .40 have not been printed and variables have been sorted by loadings.

of the variables shows that each variable has at least one very low loading (under  $\pm .10$ ), with several having two and even one variable with three low loadings. At the factor level, given a four-factor solution, each factor should have at least four very low loadings and this criterion is also met. Finally, for each pair of factors there should be several variables with significant loadings on one of the factors and a very low loading on the other factor, along with relatively few cross-loadings. Examining each pair of factors finds this criterion also met, with only one possible cross-loading. By meeting all of these criteria, the rotated factor solution can be termed a “simple structure.”

In the rotated factor solution (Table 3.9) each of the variables has a significant loading (defined as a loading above .40) on only one factor, except for  $X_{11}$ , which cross-loads on two factors (factors 1 and 4). Moreover, all of the loadings are above .70, meaning that more than one-half of the variance is accounted for by the loading on a single factor. With all of the communalities of sufficient size to warrant inclusion, the only remaining decision is to determine the action to be taken for  $X_{11}$ .

**STEP 4: RESPECIFY THE FACTOR MODEL IF NEEDED** Even though the rotated factor matrix improved upon the simplicity of the factor loadings, the cross-loading of  $X_{11}$  on factors 1 and 4 requires action. The possible actions include ignoring the cross-loading, deleting  $X_{11}$  to eliminate the cross-loading, using another rotation technique, or decreasing the number of factors. The following discussion addresses these options and the course of action chosen.

Examining the correlation matrix in Table 3.5 shows that  $X_{11}$  has high correlations with  $X_6$  (part of factor 4),  $X_9$  (part of factor 1), and  $X_{12}$  (part of factor 2). Thus, it is not surprising that it may have several high loadings. With the loadings of .642 (factor 4) and .591 (factor 1) almost identical, the cross-loading is so substantial as to not be ignorable. Applying the ratio of variances guidelines described earlier, we see that the ratio is  $1.18 (.642^2 \div .591^2)$  or  $.412 \div .349$  is well below the threshold of 1.5, indicating a significant cross-loading. As for employing another rotation technique, additional analysis showed that the other orthogonal methods (QUARTIMAX and EQUIMAX) still had this fundamental problem. Also, the number of factors should not be decreased due to the relatively large explained variance (16.1%) for the fourth factor.

Thus, the course of action taken is to delete  $X_{11}$  from the analysis, leaving 10 variables in the analysis. The rotated factor matrix and other information for the reduced set of 10 variables are also shown in Table 3.9. As we see, the factor loadings for the 10 variables remain almost identical, exhibiting both the same pattern and almost the same values for the loadings. The amount of explained variance increases slightly to 80.6 percent. With the simplified pattern of loadings (all at significant levels), all communalities above 50 percent (and most much higher), and the overall level of explained variance high enough, the 10-variable/four-factor solution is accepted, with the final step being to describe the factors.

**STEP 5: NAMING THE FACTORS** When a satisfactory factor solution has been derived, the researcher next attempts to assign some meaning to the factors. The process involves substantive interpretation of the pattern of factor loadings for the variables, including their signs, in an effort to name each of the factors. Before interpretation, a minimum acceptable level of significance for factor loadings must be selected. Then, all significant factor loadings typically are used in the interpretation process. Variables with higher loadings influence to a greater extent the name or label selected to represent a factor.

Let us look at the results in Table 3.9 to illustrate this procedure. The factor solution was derived from principal component analysis with a VARIMAX rotation of 10 perceptions of HBAT. The cut-off point for interpretation purposes in this example is all loadings  $\pm .40$  or above (see Table 3.2). The cut-off point was set somewhat low to illustrate the factor interpretation process with as many significant loadings as possible. In this example, however, all the loadings are substantially above or below this threshold, making interpretation quite straightforward.

Substantive interpretation is based on the significant loadings. In Table 3.9, loadings above .40 have been noted and the variables are sorted by their loadings on each factor. A marked pattern of variables with high loadings for each factor is evident. Factors 1 and 2 have three variables with significant loadings and factors 3 and 4 have two. Each factor can be named based on the variables with significant loadings:

- *Factor 1 Postsale Customer Service:*  $X_9$  complaint resolution;  $X_{18}$ , delivery speed; and  $X_{16}$ , order and billing;
- *Factor 2 Marketing:*  $X_{12}$  salesforce image;  $X_7$ , e-commerce presence; and  $X_{10}$ , advertising;

- *Factor 3 Technical Support:*  $X_8$  technical support; and  $X_{14}$ , warranty and claims;
- *Factor 4 Product Value:*  $X_6$  product quality; and  $X_{13}$ , competitive pricing.

One particular issue should be noted: In factor 4, competitive pricing ( $X_{13}$ ) and product quality ( $X_6$ ) have opposite signs. This means that product quality and competitive pricing vary together, but move in directions opposite to each other. Perceptions are more positive whether product quality increases or price decreases. This fundamental trade-off leads to naming the factor product value. When variables have differing signs, the researcher needs to be careful in understanding the relationships between variables before naming the factors and must also make special actions if calculating summated scales (see earlier discussion on reverse scoring).

Three variables ( $X_{11}$ ,  $X_{15}$ , and  $X_{17}$ ) were not included in the final factor analysis. When the factor-loading interpretations are presented, it must be noted that these variables were not included. If the results are used in other multivariate analyses, these three could be included as separate variables, although they would not be assured to be orthogonal to the factor scores.

The process of naming factors is based primarily on the subjective opinion of the researcher. Different researchers in many instances will no doubt assign different names to the same results because of differences in their backgrounds and training. For this reason, the process of labeling factors is subject to considerable criticism. If a logical name can be assigned that represents the underlying nature of the factors, it usually facilitates the presentation and understanding of the factor solution and therefore is a justifiable procedure.

**APPLYING AN OBLIQUE ROTATION** The VARIMAX rotation is orthogonal, meaning that the factors remain uncorrelated throughout the rotation process. In many situations, however, the factors need not be uncorrelated and may even be conceptually linked, which requires correlation between the factors. The researcher should always consider applying a nonorthogonal rotation method and assess its comparability to the orthogonal results.

In our example, it is quite reasonable to expect that perceptual dimensions would be correlated. Therefore, the application of the non-orthogonal oblique rotation is justified. Table 3.10 contains the pattern and structure matrices with the factor loadings for each variable on each factor. As discussed earlier, the pattern matrix is typically used for interpretation purposes, especially if the factors have a substantial correlation between them. In this case, the highest correlation between the factors is only  $-.241$  (factors 1 and 2), so that the pattern and structure matrices have quite comparable loadings. By examining the variables loading highly on each factor, we note that the interpretation is exactly the same as found with the VARIMAX rotation. The only difference is that all three loadings on factor 2 are negative, so that if the variables are reverse coded the correlations between factors will reverse signs as well.

**SUMMARY OF FACTOR INTERPRETATION** Both the orthogonal and oblique rotations resulted in acceptable solutions that met the “simple structure” guidelines and were readily interpretable. As an additional assessment of the four factor solution, three- and five-factor solutions (with VARIMAX rotation) were also performed on the 11 variable set (see Table 3.11). The three-factor solution does not meet all of the simple structure guidelines and still has a significant cross-loading. The five-factor solution is also less suitable since it still has a cross-loading and the fifth factor has only one significant loading. Examining these two alternative factor solutions provides additional support for the four-factor solution.

**Stage 6: Validation of Principal Components Analysis** Validation of any exploratory factor analysis result is essential, particularly when attempting to define underlying structure among the variables. Optimally, we would always follow our use of exploratory factor analysis with some form of confirmatory factor analysis, such as structural equation modeling (see Chapter 10), but this type of follow-up is often not feasible. We must look to other means, such as split sample analysis or application to entirely new samples.

In this example, we split the sample into two equal samples of 50 respondents and re-estimate the factor models to test for comparability. Table 3.12 contains the VARIMAX rotations for the split-sample results, along with the communalities. As can be seen, the two VARIMAX rotations are quite comparable in terms of both loadings and communalities for all six perceptions. The only notable occurrence is the presence of a slight cross-loading for  $X_{13}$ .

**Table 3.10** Oblique Rotation of Components Analysis Factor Matrix

PATTERN MATRIX		OBIQUE ROTATED LOADINGS <sup>a</sup>				
		Factor				
		1	2	3	4	Communality <sup>b</sup>
X <sub>9</sub> Complaint Resolution		.943				.890
X <sub>18</sub> Delivery Speed		.942				.894
X <sub>16</sub> Order & Billing		.895				.806
X <sub>12</sub> Salesforce Image			-.897			.860
X <sub>7</sub> E-Commerce			-.880			.780
X <sub>10</sub> Advertising			-.756			.585
X <sub>8</sub> Technical Support				.946		.894
X <sub>14</sub> Warranty & Claims				.936		.891
X <sub>6</sub> Product Quality					.921	.798
X <sub>13</sub> Competitive Pricing					-.702	.661

STRUCTURE MATRIX		OBIQUE ROTATED LOADINGS <sup>a</sup>				
		Factor				
		1	2	3	4	
X <sub>9</sub> Complaint Resolution		.943				
X <sub>18</sub> Delivery Speed		.942				
X <sub>16</sub> Order & Billing		.897				
X <sub>12</sub> Salesforce Image			-.919			
X <sub>7</sub> E-Commerce			-.878			
X <sub>10</sub> Advertising			-.750			
X <sub>8</sub> Technical Support				.944		
X <sub>14</sub> Warranty & Claims				.940		
X <sub>6</sub> Product Quality					.884	
X <sub>13</sub> Competitive Pricing					-.773	

FACTOR CORRELATION MATRIX						
Factor		1	2	3	4	
	1	1.000				
	2	-.241	1.000			
	3	.118	.021	1.000		
	4	.121	.190	.165	1.000	

<sup>a</sup> Factor loadings less than .40 have not been printed and variables have been sorted by loadings on each factor.

<sup>b</sup> Communality values are not equal to the sum of the squared loadings due to the correlation of the factors.

in subsample 1, although the large difference in loadings (.445 versus  $-.709$ ) and the recommended cross-loadings guidelines (i.e., variance ratio of 2.53) indicate assignment of X<sub>13</sub> only to factor 4 appropriate.

With these results we can be reasonably assured that the results are stable within our sample. If possible, we would always like to perform additional work by gathering additional respondents and ensuring that the results generalize across the population or generating new subsamples for analysis and assessment of comparability.

**Table 3.11** Comparison of Three- and Five-Factor VARIMAX Rotated Solutions

Variable	Three-Factor Solution			Five-Factor Solution				
	1	2	3	1	2	3	4	5
$X_{18}$ - Delivery Speed	<b>0.910</b>	0.206	-0.029	<b>0.941</b>	0.115	-0.002	0.047	0.122
$X_9$ - Complaint Resolution	<b>0.903</b>	0.136	0.028	<b>0.928</b>	0.084	0.048	0.089	0.049
$X_{16}$ - Order & Billing	<b>0.825</b>	0.149	0.050	<b>0.866</b>	0.082	0.083	0.038	0.039
$X_{11}$ - Product Line	<b>0.758</b>	-0.306	0.272	<b>0.594</b>	-0.063	0.150	<b>0.643</b>	-0.060
$X_{12}$ - Salesforce Image	0.160	<b>0.878</b>	0.101	0.153	<b>0.877</b>	0.057	-0.130	0.267
$X_7$ - E-Commerce Activities	0.102	<b>0.824</b>	0.087	0.080	<b>0.929</b>	0.019	-0.071	0.108
$X_{10}$ - Advertising	0.217	<b>0.647</b>	-0.017	0.139	0.355	-0.043	-0.038	<b>0.906</b>
$X_{13}$ - Competitive Pricing	-0.293	<b>0.520</b>	<b>-0.404</b>	-0.079	0.326	-0.271	<b>-0.693</b>	-0.110
$X_6$ - Product Quality	0.305	<b>-0.430</b>	0.202	0.008	0.019	-0.031	<b>0.889</b>	-0.091
$X_8$ - Technical Support	0.002	-0.013	<b>0.923</b>	0.018	-0.002	<b>0.940</b>	0.089	-0.041
$X_{14}$ - Warranty & Claims	0.096	0.062	<b>0.917</b>	0.111	0.062	<b>0.932</b>	0.091	0.004

Note: Variable loadings above .40 are in bold

**Table 3.12** Validation of Component Factor Analysis by Split Sample Estimation with VARIMAX Rotation

Split-Sample 1	VARIMAX-ROTATED LOADINGS				Communality
	1	2	3	4	
$X_9$ Complaint Resolution	.924				.901
$X_{18}$ Delivery Speed	.907				.878
$X_{16}$ Order & Billing	.901				.841
$X_{12}$ Salesforce Image		.885			.834
$X_7$ E-Commerce		.834			.733
$X_{10}$ Advertising		.812			.668
$X_8$ Technical Support			.927		.871
$X_{14}$ Warranty & Claims			.876		.851
$X_6$ Product Quality				.884	.813
$X_{13}$ Competitive Pricing		.445		-.709	.709

Split-Sample 2	VARIMAX-ROTATED LOADINGS				Communality
	1	2	3	4	
$X_9$ Complaint Resolution	.943				.918
$X_{18}$ Delivery Speed	.935				.884
$X_{16}$ Order & Billing	.876				.807
$X_{12}$ Salesforce Image		.902			.886
$X_7$ E-Commerce		.890			.841
$X_{10}$ Advertising		.711			.584
$X_8$ Technical Support			.958		.932
$X_{14}$ Warranty & Claims			.951		.916
$X_6$ Product Quality				.889	.804
$X_{13}$ Competitive Pricing				-.720	.699

**Stage 7: Additional Uses of the Exploratory Factor Analysis Results** The researcher has the option of using exploratory factor analysis not only as a data summarization tool, as seen in the prior discussion, but also as a data-reduction tool. In this context, exploratory factor analysis would assist in reducing the number of variables, either through selection of a set of surrogate variables, one per factor, or by creating new composite variables for each factor. The following sections detail the issues in data reduction for this example.

**SELECTING SURROGATE VARIABLES FOR SUBSEQUENT ANALYSIS** Let us first clarify the procedure for selecting surrogate variables. In selecting a single variable to represent an entire factor, it is preferable to use an orthogonal rotation so as to ensure that, to the extent possible, the selected variables be uncorrelated with each other. Thus, on this analysis the orthogonal solution (Table 3.10) will be used instead of the oblique results.

Assuming we want to select only a single variable for further use, attention is on the magnitude of the factor loadings (Table 3.10), irrespective of the sign (positive or negative). Focusing on the factor loadings for factors 1 and 3, we see that the first and second highest loadings are essentially identical (.933 for  $X_9$  and .931 for  $X_{18}$  on factor 1, .940 for  $X_8$  and .933 for  $X_{14}$  on factor 3). If we have no a priori evidence to suggest that the reliability or validity for one of the variables is better than for the other, and if none would be theoretically more meaningful for the factor interpretation, we would select the variable with the highest loading ( $X_9$  and  $X_8$  for factors 1 and 3, respectively). However, researchers must be cautious to not let these single measures provide the sole interpretation for the factor, because each factor is a much more complex dimension than could be represented in any single variable. The difference between the first and second highest loadings for factors 2 and 4 are much greater, making selection of variables  $X_{12}$  (factor 2) and  $X_6$  (factor 4) easier and more direct. For all four factors, however, no single variable represents the component best; thus factor scores or a summated scale would be more appropriate if possible.

**CREATING SUMMATED SCALES** A summated scale is a composite value for a set of variables calculated by such simple procedures as taking the average of the variables in the scale. It is much like the variates in other multivariate techniques, except that the weights for each variable are assumed to be equal in the averaging procedure. In this way, each respondent would have four new variables (summated scales for factors 1, 2, 3, and 4) that could be substituted for the original 13 variables in other multivariate techniques. Exploratory factor analysis assists in the construction of the summated scale by identifying the dimensionality of the variables (defining the factors), which then form the basis for the composite values if they meet certain conceptual and empirical criteria. After the actual construction of the summated scales, which includes reverse scoring of opposite-signed variables (see earlier discussion), the scales should also be evaluated for reliability and validity if possible.

In this example, the four-factor solution suggests that four summated scales should be constructed. The four factors, discussed earlier, correspond to dimensions that can be named and related to concepts with adequate content validity. The dimensionality of each scale is supported by the clean interpretation of each factor, with high factor loadings of each variable on only one factor. The reliability of the summated scales is best measured by Cronbach's alpha, which in this case is .90 for scale 1, .78 for scale 2, .80 for scale 3, and .57 for scale 4. Only scale 4, representing the Product Value factor, has a reliability below the recommended level of .70. A primary reason for the low reliability value is that the scale only has two variables. Future research should strive to find additional items that measure this concept. It will be retained for further use with the caveat of a somewhat lower reliability and the need for future development of additional measures to represent this concept. Also remember that because  $X_{13}$  has a negative relationship (loading) it should be reverse-scored before creating the summated scale.

Although no direct test is available to assess the validity of the summated scales in the HBAT database, one approach is to compare the summated scales with the surrogate variables to see whether consistent patterns emerge. Table 3.13 (Part A) illustrates the use of summated scales as replacements for the original variables by comparing the differences in the surrogate variables across the two regions (USA/North America versus outside North America) of  $X_4$  to those differences of the corresponding summated scales and factor scores.

When viewing the two groups of and factor scores  $X_4$ , we can see that the pattern of differences is consistent.  $X_{12}$  and  $X_6$  (the surrogate variables for factors 2 and 4) and scales 2 and 4 (the summated scales for factors 2 and 4) all have significant differences between the two regions, whereas the measures for the first and third factors ( $X_9$  and  $X_8$ , scales 1 and 3, and factor scores 1 and 3) all show no difference. The summated scales and the surrogate variables all show the same patterns of differences between the two regions, as seen for the factor scores, demonstrating a level of convergent validity between these three measures.

**USE OF FACTOR SCORES** Instead of calculating summated scales, we could calculate factor scores for each of the four factors in our principal component analysis. The factor scores differ from the summated scales in that factor scores are based directly on the factor loadings, meaning that every variable contributes to the factor score based on the size of its loading (rather than calculating the summated scale score as the mean of selected variables with high loadings).

**Table 3.13** Evaluating the Replacement of the Original Variables by Factor Scores or Summated Scales

Statistical Test	Mean Scores		t-test	
	Group 1: USA/North America	Group 2: Outside North America	t value	Significance
<b>Measure</b>				
<b>Representative Variables from Each Factor</b>				
$X_9$ , Complaint Resolution	5.456	5.433	.095	.925
$X_{12}$ , Salesforce Image	4.587	5.466	-4.341	.000
$X_8$ , Technical Support	5.697	5.152	1.755	.082
$X_6$ , Product Quality	8.705	7.238	5.951	.000
<b>Factor Scores</b>				
Factor 1, Customer Service	-.031	.019	-.248	.805
Factor 2, Marketing	-.308	.197	-2.528	.013
Factor 3, Technical Support	.154	-.098	1.234	.220
Factor 4, Product Value	.741	-.474	7.343	.000
<b>Summated Scales</b>				
Scale 1, Customer Service	4.520	4.545	-.140	.889
Scale 2, Marketing	3.945	4.475	-3.293	.001
Scale 3, Technical Support	5.946	5.549	1.747	.084
Scale 4, Product Value	6.391	4.796	8.134	.000
<b>Part B: Correlations Between Summated Scales</b>				
	Scale 1	Scale 2	Scale 3	Scale 4
Scale 1	1.000			
Scale 2	.260**	1.000		
Scale 3	.113	.010	1.000	
Scale 4	.126	-.225*	.228*	1.000
<b>Part C: Correlations Between Factor Scores and Summated Scales</b>				
	Factor 1	Factor 2	Factor 3	Factor 4
Scale 1	.987**	.127	.057	.060
Scale 2	.147	.976**	.008	-.093
Scale 3	.049	.003	.984**	.096
Scale 4	.082	-.150	.148	.964**

\*Significant at .05 level.

\*\*Significant at .01 level.

The first test of comparability of factor scores is similar to that performed with summated scales in assessing the pattern of differences found on  $X_4$  for the surrogate variables and now the factor scores. Just as seen with the summated scales, the patterns of differences were identical, with differences on factor scores 2 and 4 corresponding to the differences in the surrogate variables for factors 2 and 4, with no differences for the others.

The consistency between factor scores and summated scales is also seen in the correlations in Table 3.13. We know that the factor scores, since rotated with a VARIMAX technique, are orthogonal (uncorrelated). In Part B we see that the scales are relatively uncorrelated among themselves (the highest correlation is .260), which matches fairly closely to an orthogonal solution. This pattern also closely matches the oblique solution shown in Table 3.10 (note that the second factor in the oblique solution had all negative loadings, thus the difference between positive and negative correlations among the factors). But how closely do the scales correspond to the factor scores? Part C of Table 3.13 shows the correlations between the summated scales and factor scores, with correlations ranging from .964 to .987. These results further support the use of the summated scales as valid substitutes for factor scores if desired.

**SELECTING BETWEEN DATA REDUCTION METHODS** If the original variables are to be replaced by surrogate variables, factor scores, or summated scales, a decision must be made on which to use. This decision is based on the need for simplicity

(which favors surrogate variables) versus replication in other studies (which favors use of summated scales) versus the desire for orthogonality of the measures (which favors factor scores). Although it may be tempting to employ surrogate variables, the preference among researchers today is the use of summated scales or, to a lesser degree, factor scores. From an empirical perspective, the two composite measures are essentially identical. The correlations in Table 3.13 demonstrate the high correspondence of factor scores to summated scales and the low correlations among summated scales, approximating the orthogonality of the factor scores. The final decision, however, rests with the researcher and the need for orthogonality versus replicability in selecting factor scores versus summated scales.

## COMMON FACTOR ANALYSIS: STAGES 4 AND 5

Common factor analysis is the second major factor analytic model that we discuss. The primary distinction between principal component analysis and common factor analysis is that the latter considers only the common variance associated with a set of variables. This aim is accomplished by factoring a “reduced” correlation matrix with estimated initial communalities in the diagonal instead of unities. The differences between principal component analysis and common factor analysis occur only at the factor estimation and interpretation stages (stages 4 and 5). Once the communalities are substituted on the diagonal, the common factor model extracts factors in a manner similar to principal component analysis. The researcher uses the same criteria for factor selection and interpretation. To illustrate the differences that can occur between common factor and principal component analysis, the following sections detail the extraction and interpretation of a common factor analysis of the 13 HBAT perceptions used in the principal component analysis.

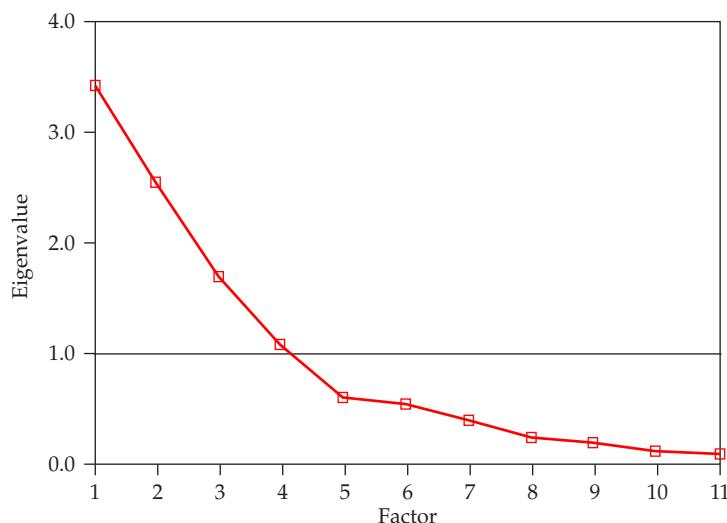
**Stage 4: Deriving Factors and Assessing Overall Fit** The reduced correlation matrix with communalities on the diagonal was used in the common factor analysis. Recalling the procedures employed in the principal component analysis, the original 13 variables were reduced to 11 due to low MSA values for  $X_{15}$  and  $X_{17}$ . We will proceed from this set of 11 variables for the common factor analysis.

The first step is to determine the number of factors to retain for examination and possible rotation. Table 3.14 shows the extraction statistics. Using the original eigenvalues, comparable to principal components analysis, the latent root criterion with a cut-off value of 1.0 for the eigenvalue would retain four factors. Likewise, the scree analysis indicates that four or five factors be retained (see Figure 3.12). In combining these two criteria, we will retain four factors for further analysis because of the low eigenvalue for the fifth factor and to maintain comparability with the principal component analysis. As with the principal component analysis examined earlier, the researcher should employ a combination of criteria in determining the number of factors to retain and may even wish to examine the three-factor solution as an alternative.

Because the final common factor model differs from the initial extraction estimates (e.g., see discussion of Table 3.14 that follows), the researcher should be sure to evaluate the extraction statistics for the final common factor model. Remember that in common factor analysis, only the “common” or shared variance is used. Thus, the trace (sum of all eigenvalues) and the eigenvalues for all factors will be lower when only the common variance is considered.

**Table 3.14** Results for the Extraction of Common Factors: Extraction Method—Principal Axis Factoring

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.427	31.154	31.154	3.215	29.231	29.231
2	2.551	23.190	54.344	2.225	20.227	49.458
3	1.691	15.373	69.717	1.499	13.630	63.088
4	1.087	9.878	79.595	.678	6.167	69.255
5	.609	5.540	85.135			
6	.552	5.017	90.152			
7	.402	3.650	93.802			
8	.247	2.245	96.047			
9	.204	1.850	97.898			
10	.133	1.208	99.105			
11	.098	.895	100.000			



**Figure 3.12**  
Scree Test for Common Factor Analysis

As such, a researcher may wish to be more liberal in making judgments on such issues as variance extracted or the latent root criterion threshold. If the researcher was dissatisfied with the total variance explained, for example, the remedies discussed earlier are still available (such as extracting one or more additional factors to increase explained variance). Also, communalities should also be examined to ensure an adequate level is maintained after extraction.

As also shown in Table 3.14, the eigenvalues for extracted factors can be restated in terms of the common factor extraction process. As shown in Table 3.14, the values for the extracted factors still support four factors because the percentage of total variance explained is still 70 percent. The only substantive difference is for the eigenvalue of factor 4, which falls below the 1.0 threshold. Parallel analysis can be applied to common factor models as well as principal components models and Table 3.15 presents the results using the eigenvalues for the extracted factors. A four-factor model is supported using both the mean eigenvalue as well as the more conservative 95th percentile value. When evaluating the fifth factor, we see that the support for retention is provided when using the mean value, but not the 95th percentile. This casts doubt on including the fifth factor since parallel analysis is prone to including more factors than other common factor stopping rules. So for our purposes the fourth factor is retained, but the fifth factor is not included in subsequent analysis. This also maintains comparability with the principal component analysis results.

The unrotated factor matrix (Table 3.16) shows that the communalities of each variable are comparable to those found in principal component analysis. Because several variables fell below a communality of .50, a five-factor model could be made in an attempt to increase the communalities, as well as the overall variance explained. For our purposes here, however, we interpret the four-factor solution.

**Stage 5: Interpreting the Factors** With the factors extracted and the number of factors determined, we now proceed to interpretation of the factors.

**Table 3.15** Parallel Analysis as a Stopping Rule for Common Factor Analysis

Analysis of Reduced Variable Set (11 variables)			
Common Factor Results		Parallel Analysis	
Factor	Eigenvalue <sup>A</sup>	Mean	95th Percentile
1	3.215	0.680	0.863
2	2.225	0.517	0.644
3	1.499	0.367	0.484
4	0.678	0.256	0.350
5	0.201 <sup>B</sup>	0.156	0.232

<sup>A</sup> Eigenvalues based on communalities (reduced correlation matrix).

<sup>B</sup> Eigenvalue when five factors extracted rather than four.

**Table 3.16** Unrotated Common Factor-Loadings Matrix

	Factor <sup>a</sup>				Communality
	1	2	3	4	
X <sub>18</sub> Delivery Speed	.895				.942
X <sub>9</sub> Complaint Resolution	.862				.843
X <sub>16</sub> Order & Billing	.747				.622
X <sub>11</sub> Product Line	.689	-.454			.800
X <sub>12</sub> Salesforce Image		.805			.990
X <sub>7</sub> E-Commerce Presence		.657			.632
X <sub>13</sub> Competitive Pricing		.553			.443
X <sub>10</sub> Advertising		.457			.313
X <sub>8</sub> Technical Support			.739		.796
X <sub>14</sub> Warranty & Claims			.735		.812
X <sub>6</sub> Product Quality		-.408		.463	.424

<sup>a</sup>Factor loadings less than .40 have not been printed and variables have been sorted by loadings on each factor.

By examining the unrotated loadings (see Table 3.16), we note the need for a factor matrix rotation just as we found in the principal component analysis. Factor loadings were generally not as high as desired, and two variables (X<sub>6</sub> and X<sub>11</sub>) exhibited cross-loadings. Turning then to the VARIMAX-rotated common factor analysis factor matrix (Table 3.17), the information provided is the same as in the principal component analysis solution (e.g., sums of squares, percentages of variance, communalities, total sums of squares, and total variances extracted).

Comparison of the information provided in the rotated common factor analysis factor matrix and the rotated principal component analysis factor matrix shows remarkable similarity. X<sub>11</sub> has substantial cross-loadings on both factors 1 and 4 in both analyses (Tables 3.9 and 3.17). When we apply the variance ratio test for cross-loadings, we see that the ratio is 1.84 ( $.713^2 \div .525^2 = .508 \div .276 = 1.84$ ) which characterizes this as a potential cross-loading. When X<sub>11</sub> is deleted from the analysis, the four-factor solution is almost identical to the principal component analysis. The primary differences between the principal component analysis and common factor analysis are the generally lower loadings in the common factor analysis, owing primarily to the lower communalities of the variables used in common factor analysis. Even with these slight differences in the patterns of loadings, though, the basic interpretations are identical between the principal component analysis and the common factor analysis.

## A MANAGERIAL OVERVIEW OF THE RESULTS

Both the principal component and common factor analyses provide the researcher with several key insights into the structure of the variables from data summarization and options for data reduction. First, concerning the structure of the variables, clearly four separate and distinct dimensions of evaluation are used by the HBAT customers. These dimensions encompass a wide range of elements in the customer experience, from the tangible product attributes (Product Value) to the relationship with the firm (Customer Service and Technical Support) to even the outreach efforts (Marketing) by HBAT. Business planners within HBAT can now discuss plans revolving around these four areas instead of having to deal with all of the separate variables.

It should be noted that the two variables eliminated due to low MSA values may represent potential additional areas within the customer experience that could be represented by additional items to enable factors to be estimated in those areas. It is important to remember that variables excluded from exploratory factor analysis are not necessarily “bad” or deficient, it is just that they are not well correlated with other items included in the analysis and may be single-variable measures of other dimensions not represented by the remaining variables, or they may suggest other concepts that need to be assessed by new multi-item dimensions. When exploratory factor analysis is employed as a precursor to confirmatory factor analysis in scale development, it does provide a means for identifying variables that are not amenable to the confirmatory model (e.g., significant cross-loadings) and provides an efficient screening tool in that analysis. See Chapters 9 and 10 for a much more detailed discussion of confirmatory factor analyses.

**Table 3.17** VARIMAX-Rotated Common Factor Matrix: Full and Reduced Sets of Variables

<b>Full Set of 11 Variables</b>	<b>Factor<sup>a</sup></b>				<b>Communality</b>
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	
X <sub>18</sub> Delivery Speed	.949				.942
X <sub>9</sub> Complaint Resolution	.897				.843
X <sub>16</sub> Order & Billing	.768				.622
X <sub>12</sub> Salesforce Image		.977			.990
X <sub>7</sub> E-Commerce		.784			.632
X <sub>10</sub> Advertising		.529			.313
X <sub>14</sub> Warranty & Claims			.884		.812
X <sub>8</sub> Technical Support			.884		.796
X <sub>11</sub> Product Line	.525			.712	.800
X <sub>6</sub> Product Quality				.647	.424
X <sub>13</sub> Competitive Pricing				-.590	.443
					Total
Sum of Squared Loadings (eigenvalue)	2.635	1.971	1.641	1.371	7.618
Percentage of Trace	23.95	17.92	14.92	12.47	69.25
<b>Reduced Set of 10 Variables</b>	<b>Factor<sup>a</sup></b>				<b>Communality</b>
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	
X <sub>18</sub> Delivery Speed	.925				.885
X <sub>9</sub> Complaint Resolution	.913				.860
X <sub>16</sub> Order & Billing	.793				.660
X <sub>12</sub> Salesforce Image		.979			.993
X <sub>7</sub> E-Commerce		.782			.631
X <sub>10</sub> Advertising		.531			.316
X <sub>8</sub> Technical Support			.905		.830
X <sub>14</sub> Warranty & Claims			.870		.778
X <sub>6</sub> Product Quality				.788	.627
X <sub>13</sub> Competitive Pricing				-.480	.353
					Total
Sum of Squared Loadings (eigenvalue)	2.392	1.970	1.650	.919	6.932
Percentage of Trace	23.92	19.70	16.50	9.19	69.32

<sup>a</sup>Factor loadings less than .40 have not been printed and variables have been sorted by loadings on each factor.

Exploratory factor analysis also provides the basis for data reduction through either summated scales or factor scores. The researcher now has a method for combining the variables within each factor into a single score that can replace the original set of variables with four new composite variables. When looking for differences, such as between regions, these new composite variables can be used so that only differences for composite scores, rather than the individual variables, are analyzed.

The multivariate statistical technique of exploratory factor analysis has been presented in broad conceptual terms. Basic guidelines for interpreting the results were included to clarify further the methodological concepts. An example of the application of exploratory factor analysis was presented based on the HBAT database. This chapter helps you to do the following:

**Differentiate exploratory factor analysis techniques from other multivariate techniques.** Exploratory factor analysis (EFA) can be a highly useful and powerful multivariate statistical technique for effectively extracting information

from large bodies of interrelated data. When variables are correlated, the researcher needs ways to manage these variables: grouping highly correlated variables together, labeling or naming the groups, and perhaps even creating a new composite measure that can represent each group of variables. The primary purpose of exploratory factor analysis is to define the underlying structure among the variables in the analysis. As an interdependence technique, factor analysis attempts to identify structure (groupings among variables) based on relationships represented in a correlation matrix. It is a powerful tool to better understand the structure of the data, and also can be used to simplify analyses of a large set of variables by replacing them with composite variables. When it works well, it points to interesting relationships that might not have been obvious from examination of the raw data alone, or even the correlation matrix.

**Distinguish between exploratory and confirmatory uses of factor analysis techniques.** Factor analysis as discussed in this chapter is primarily an exploratory technique because the researcher has little control over the specification of the structure (e.g., number of factors, loadings of each variable, etc.). Although the methods discussed in this chapter provide insights into the data, any attempt at confirmation will most likely require the use of specific methods discussed in the chapters on structural equation modeling (see Chapters 9, 10 and 13).

**Understand the seven stages of applying factor analysis.** The seven stages of applying exploratory factor analysis include the following:

- Clarifying the objectives of factor analysis
- Designing a factor analysis, including selection of variables and sample size
- Assumptions of exploratory factor analysis
- Deriving factors and assessing overall fit, including which factor model to use and the number of factors
- Rotating and interpreting the factors
- Validation of exploratory factor analysis solutions
- Additional uses of exploratory factor analysis results, such as selecting surrogate variables, creating summated scales, or computing factor scores

**Distinguish between R and Q factor analysis.** The principal use of exploratory factor analysis is to develop a structure among variables, referred to as *R* factor analysis. Factor analysis also can be used to group cases and is then referred to as *Q* factor analysis. *Q* factor analysis is similar to cluster analysis. The primary difference is that *Q* factor analysis uses correlation as the measure of similarity whereas cluster analysis is based on a distance measure.

**Identify the differences between principal component analysis and common factor analysis models.** Three types of variance are considered when applying exploratory factor analysis: common variance and unique variance, which can be further divided into specific and error variance. When you add the three types of variance together, you get total variance. Each of the two methods of developing a factor solution uses different types of variance. Principal component analysis, also known as *components analysis*, considers the total variance, deriving factors that focus on the common variance but also contain small proportions of specific variance and, in some instances, error variance. Principal component analysis is preferred when data reduction is a primary goal and when the researcher feels confident that unique variance is less prevalent so as to not impact the resulting factors. Common factor analysis is based only on common (shared) variance and assumes that both the unique and error variance are not of interest in defining the structure of the variables. It is more useful in the scale development process (i.e., identifying latent constructs) and when the researcher has little knowledge about the unique variance. The two methods achieve essentially the same results in many research situations.

**Describe how to determine the number of factors to extract.** A critical decision in exploratory factor analysis is the number of factors to retain for interpretation and further use. In deciding when to stop factoring (i.e., how many factors to extract), the researcher must combine a conceptual foundation (How many factors should be in the structure?) with some empirical evidence (How many factors can be reasonably supported?). The researcher generally begins with some predetermined criteria, such as the general number of factors, plus some general thresholds of practical relevance (e.g., required percentage of variance explained). These criteria are combined with empirical measures of the factor structure. An exact quantitative basis for deciding the number of factors to extract has not been developed. A number of empirical stopping criteria for the number of factors to extract are available, including the latent root criterion, percentage of variance, the scree test and parallel analysis. These criteria must be balanced against any theoretical bases for establishing the number of factors, such as the a priori criterion.

**Explain the concept of rotation of factors.** Perhaps the most important tool in interpreting factors is factor rotation. The term *rotation* means that the reference axes of the factors are turned about the origin until some other position has been reached. Two types of rotation are orthogonal and oblique. Unrotated factor solutions extract factors in the order of their importance, with the first factor being a general factor with almost every variable loading significantly and accounting for the largest amount of variance. The second and subsequent factors are based on the residual amount of variance, with each accounting for successively smaller portions of variance. The ultimate effect of rotating the factor matrix is to redistribute the variance from earlier factors to later ones to achieve a simpler, theoretically more meaningful factor pattern. Factor rotation assists in the interpretation of the factors by simplifying the structure through maximizing the significant loadings of a variable on a single factor. In this manner, the variables most useful in defining the character of each factor can be easily identified.

**Describe how to name a factor.** Factors represent a composite of many variables. When an acceptable factor solution has been obtained in which all variables have a significant loading on a factor, the researcher attempts to assign some meaning to the pattern of factor loadings. Variables with higher loadings are considered more important and have greater influence on the name or label selected to represent a factor. The significant variables for a particular factor are examined and, placing greater emphasis on those variables with higher loadings, a name or label is assigned to a factor that accurately reflects the variables loading on that factor. The presence of cross-loadings (i.e., variables with significant loadings on more than one factor) may indicate deletion of that variable from the analysis since it does not represent simple structure and complicates the naming process. The researcher identifies the variables with the greatest contribution to a factor and assigns a “name” to represent the factor’s conceptual meaning.

**Explain the additional uses of exploratory factor analysis.** Depending on the objectives for applying exploratory factor analysis, the researcher may stop with data summarization (with or without interpretation) or further engage in one of the methods for data reduction. If the objective is simply to identify logical combinations of variables and better understand the interrelationships among variables, then data summarization with interpretation will suffice. If the objective, however, is to identify appropriate variables for subsequent application to other statistical techniques, then some form of data reduction will be employed. One of the data reduction options of exploratory factor analysis is to select a single (surrogate) variable with the highest factor loading. In doing so, the researcher identifies a single variable as the best representative for all variables in the factor. A second option for data reduction is to calculate a summated scale, where variables with the highest factor loadings are summed. A single summated score represents the factor, but only selected variables contribute to the composite score. A third option for data reduction is to calculate factor scores for each factor, where each variable contributes to the score based on its factor loading. This single measure is a composite variable that reflects the relative contributions of all variables to the factor. If the summated scale is valid and reliable, it is probably the best of these three data reduction alternatives.

**State the major limitations of exploratory factor analytic techniques.** Three of the most frequently cited limitations are as follows:

Because many techniques for performing exploratory factor analyses are available, controversy exists over which technique is the best.

The subjective aspects of exploratory factor analysis (i.e., deciding how many factors to extract, which technique should be used to rotate the factor axes, which factor loadings are significant) are all subject to many differences in opinion.

The problem of reliability is real.

Like any other statistical procedure, an exploratory factor analysis starts with a set of imperfect data. When the data vary because of changes in the sample, the data-gathering process, or the numerous kinds of measurement errors, the results of the analysis also may change. The results of any single analysis are therefore less than perfectly dependable.

The potential applications of exploratory factor analysis to problem solving and decision making in business research are numerous. Exploratory factor analysis is a much more complex and involved subject than might be indicated here. This problem is especially critical because the results of a single-factor analytic solution frequently look plausible. It is important to emphasize that plausibility is no guarantee of validity or stability.

What are the differences between the objectives of data summarization and data reduction?

How can exploratory factor analysis help the researcher improve the results of other multivariate techniques?

What guidelines can you use to determine the number of factors to extract? Explain each briefly.

How do you use the factor-loading matrix to interpret the meaning of factors?

When would the researcher use an oblique rotation instead of an orthogonal rotation? What are the basic differences between them?

What are the criteria used in determining the variables that will represent a factor (e.g., be used in naming that factor)?

How and when should factor scores be used in conjunction with other multivariate statistical techniques?

What are the differences between factor scores and summated scales? When is each most appropriate?

What are the differences between exploratory factor analysis and confirmatory factor analysis? When are each most applicable?

What is the difference between Q-type factor analysis and cluster analysis?

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 American Psychological Association. 1985. *Standards for Educational and Psychological Tests*. Washington, DC: APA.
- 2 Anderson, J. C., D. W. Gerbing, and J. E. Hunter. 1987. On the Assessment of Unidimensional Measurement: Internal and External Consistency and Overall Consistency Criteria. *Journal of Marketing Research* 24: 432–7.
- 3 Bearden, W. O., and R. G. Netemeyer. 1999. *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior*, 2nd edn. Newbury Park, CA: Sage.
- 4 Bentler, Peter M. 1995. *EQS Structural Equations Program Manual*. Los Angeles: BMDP Statistical Software.
- 5 BMDP Statistical Software, Inc. 1992. *BMDP Statistical Software Manual, Release 7, Vols. 1 and 2*. Los Angeles: BMDP Statistical Software.
- 6 Borgatta, E. F., K. Kercher, and D. E. Stull. 1986. A Cautionary Note on the Use of Principal Components Analysis. *Sociological Methods and Research* 15: 160–8.
- 7 Browne, M. W. 1968. A Comparison of Factor Analytic Techniques. *Psychometrika* 33: 267–334.
- 8 Bruner, G. C., Karen E. James, and P. J. Hensel. 2001. *Marketing Scales Handbook, Vol. 3, A Compilation of Multi-Item Measures*. Chicago: American Marketing Association.
- 9 Buja, A., and N. Eyuboglu. 1992. Remarks on Parallel Analysis. *Multivariate Behavioral Research* 27: 509–40.
- 10 Campbell, D. T., and D. W. Fiske. 1959. Convergent and Discriminant Validity by the Multitrait-Multimethod Matrix. *Psychological Bulletin* 56: 81–105.
- 11 Carifio, J. and R. Perla. 2007. Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences* 2: 106–16.
- 12 Cattell, R. B. 1966. The Scree Test for the Number of Factors. *Multivariate Behavioral Research* 1: 245–76.
- 13 Cattell, R. B., K. R. Balcar, J. L. Horn, and J. R. Nesselroade. 1969. Factor Matching Procedures: An Improvement of the s Index; with Tables. *Educational and Psychological Measurement* 29: 781–92.
- 14 Chatterjee, S., L. Jamieson, and F. Wiseman. 1991. Identifying Most Influential Observations in Factor Analysis. *Marketing Science* 10: 145–60.
- 15 Churchill, G. A. 1979. A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research* 16: 64–73.
- 16 Cliff, N. 1987. *Analyzing Multivariate Data*. San Diego: Harcourt Brace Jovanovich.
- 17 Cliff, N. 1988. The Eigenvalues-Greater-Than-One Rule and the Reliability of Components. *Psychological Bulletin* 103: 276–9.
- 18 Cliff, N., and C. D. Hamburger. 1967. The Study of Sampling Errors in Factor Analysis by Means of Artificial Experiments. *Psychological Bulletin* 68: 430–45.
- 19 Cronbach, L. J. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 31: 93–96.
- 20 De Leeuw, J. 2006. Nonlinear Principal Components Analysis. In M. Greenacre and J. Blasius (eds.), *Multiple Correspondence Analysis and Related Methods*, pp. 107–33. Boca Raton, FL: CRC Press.
- 21 Dillon, W. R., and M. Goldstein. 1984. *Multivariate Analysis: Methods and Applications*. New York: Wiley.
- 22 Dillon, W. R., N. Mulani, and D. G. Frederick. 1989. On the Use of Component Scores in the Presence of Group Structure. *Journal of Consumer Research* 16: 106–112.

- 23 Fabrigar, L. R. and D. T. Wegener. 2012. *Factor Analysis*. New York: Oxford University Press.
- 24 Fisher, R. A. 1938. *Statistical Methods for Research Workers*. 10th edn. Edinburgh: Oliver and Boyd.
- 25 Gorsuch, R. L. 1983. *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 26 Gorsuch, R. L. 1990. Common Factor Analysis Versus Component Analysis: Some Well and Little Known Facts. *Multivariate Behavioral Research* 25: 33–39.
- 27 Guadagnoli, E. and W. Velicer. 1988. Relation of Sample Size to the Stability of Component Patterns. *Psychological Bulletin* 103: 265–75.
- 28 Harman, H. H. 1967. *Modern Factor Analysis*, 2nd edn. Chicago: University of Chicago Press.
- 29 Hattie, J. 1985. Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement* 9: 139–64.
- 30 Henson, R. K., and J. K. Roberts. 2006. Use of Exploratory Factor Analysis in Published Research. *Educational and Psychological Measurement* 66: 393–416.
- 31 Horn, J. L. 1965. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika* 30: 179–85.
- 32 Jamieson, S. 2004. Likert Scales: How to (Ab)use Them. *Medical Education* 38: 1212–18.
- 33 Jöreskog, K.G. & Sörbom, D. 2015. *LISREL 9.20 for Windows* [Computer software]. Skokie, IL: Scientific Software International, Inc.
- 34 Kaiser, H. F. 1970. A Second-Generation Little Jiffy. *Psychometrika* 35: 401–15.
- 35 Kaiser, H. F. 1974. Little Jiffy, Mark IV. *Educational and Psychological Measurement* 34: 111–17.
- 36 Kruskal, J. B., and R. N. Shepard. 1974. A Nonmetric Variety of Linear Factor Analysis. *Psychometrika* 38: 123–57.
- 37 Linn, R. L. 1968. A Monte Carlo Approach to the Number of Factors Problem. *Psychometrika* 33: 37–71.
- 38 McDonald, R. P. 1981. The Dimensionality of Tests and Items. *British Journal of Mathematical and Social Psychology* 34: 100–17.
- 39 Mulaik, S. A. 1990. Blurring the Distinction Between Component Analysis and Common Factor Analysis. *Multivariate Behavioral Research* 25: 53–9.
- 40 Mulaik, S. A., and R. P. McDonald. 1978. The Effect of Additional Variables on Factor Indeterminacy in Models with a Single Common Factor. *Psychometrika* 43: 177–92.
- 41 Nunnally, J. L. 1978. *Psychometric Theory*, 2nd edn. New York: McGraw-Hill.
- 42 Nunnally, J. 1978. *Psychometric Theory*. New York: McGraw-Hill.
- 43 O'Connor, B. P. 2000. SPSS and SAS Programs for Determining the Number of Components Using Parallel Analysis and Velicer's MAP Test. *Behavior Research Methods, Instruments, & Computers* 32: 396–402.
- 44 Peter, J. P. 1979. Reliability: A Review of Psychometric Basics and Recent Marketing Practices. *Journal of Marketing Research* 16: 6–17.
- 45 Peter, J. P. 1981. Construct Validity: A Review of Basic Issues and Marketing Practices. *Journal of Marketing Research* 18: 133–45.
- 46 Preacher, K. J. and R. C. MacCallum. 2003. Repairing Tom Swift's Electric Factor Analysis Machine. *Understanding Statistics* 2: 13–43.
- 47 Robinson, J. P., P. R. Shaver, and L. S. Wrightman. 1991. *Measures of Personality and Social Psychological Attitudes*. San Diego: Academic Press.
- 48 Robinson, J. P., P. R. Shaver, and L. S. Wrightsman. 1991. Criteria for Scale Selection and Evaluation. In *Measures of Personality and Social Psychological Attitudes*, J. P. Robinson, P. R. Shaver, and L. S. Wrightsman (eds.). San Diego: Academic Press.
- 49 Rummel, R. J. 1970. *Applied Factor Analysis*. Evanston, IL: Northwestern University Press.
- 50 Ruscio, J and B. Roche. 2012. Determining the Number of Factors to Retain in an Exploratory Factor Analysis using Comparison Data of Known Factorial Structure. *Psychological Assessment* 24: 282–92.
- 51 SAS Institute Inc. 2008. *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.
- 52 Sass, Daniel A. and Thomas A. Schmitt. 2010. A Comparative Investigation of Rotation Criteria Within Exploratory Factor Analysis. *Multivariate Behavioral Research* 45: 73–103.
- 53 Smith, Scott M. 1989. *PC-MDS: A Multidimensional Statistics Package*. Provo, UT: Brigham Young University Press.
- 54 Snook, S. C., and R. L. Gorsuch. 1989. Principal Component Analysis Versus Common Factor Analysis: A Monte Carlo Study. *Psychological Bulletin* 106: 148–54.
- 55 Stewart, D. W. 1981. The Application and Misapplication of Factor Analysis in Marketing Research. *Journal of Marketing Research* 18: 51–62.
- 56 Thurstone, L. L. 1947. *Multiple-Factor Analysis*. Chicago: University of Chicago Press.
- 57 Turner, N. E. 1998. The Effect of Common Variance and Structure on Random Data Eigenvalues: Implications for the Accuracy of Parallel Analysis. *Educational & Psychological Measurement* 58: 541–68.
- 58 Velicer, W. F., and D. N. Jackson. 1990. Component Analysis Versus Common Factor Analysis: Some Issues in Selecting an Appropriate Procedure. *Multivariate Behavioral Research* 25: 1–28.
- 59 Winsberg, S. and J. O. Ramsay. 1983. Monotone Spline Transformations for Dimension Reduction. *Psychometrika* 48: 575–95.
- 60 Young, F. W. 1981. Quantitative Analysis of Qualitative Data. *Psychometrika* 46: 357–88.
- 61 Young, F. W., Y. Takane, and J. de Leeuw. 1978. The Principal Components of Mixed Measurement Level Multivariate Data: An Alternating Least Squares Method with Optimal Scaling Features. *Psychometrika* 43: 279–81.
- 62 Zwick, W. R. and W. F. Velicer. 1986. Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin* 99: 432–42.

# 4 Cluster Analysis

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

- Define cluster analysis, its roles, and its limitations.
- Identify the types of research questions addressed by cluster analysis.
- Understand how interobject similarity is measured.
- Understand why different distance measures are sometimes used.
- Understand the differences between hierarchical and nonhierarchical clustering techniques.
- Know how to interpret results from cluster analysis.
- Follow the guidelines for cluster validation.

## Chapter Preview

Researchers often encounter situations best resolved by defining groups of homogeneous objects, whether they are individuals, firms, or even behaviors. Strategy options based on identifying groups within the population, such as segmentation and target marketing, would not be possible without an objective methodology. This same need is encountered in other areas, ranging from the physical sciences (e.g., creating a biological taxonomy for the classification of various animal groups—*insects* versus *mammals* versus *reptiles*) to the social sciences (e.g., analyzing various psychiatric profiles). In all instances, the researcher is searching for a “natural” structure among the observations based on a multivariate profile.

The most commonly used technique for this purpose is cluster analysis. Cluster analysis groups individuals or objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters. The attempt is to maximize the homogeneity of objects within the clusters while also maximizing the heterogeneity between the clusters. This chapter explains the nature and purpose of cluster analysis and provides the researcher with an approach for obtaining and using cluster results.

## Key Terms

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

**Absolute Euclidean distance** See *squared Euclidean distance*.

**Agglomerative methods** *Hierarchical procedure* that begins with each *object* or observation in a separate cluster. In each subsequent step, the two clusters that are most similar are combined to build a new aggregate cluster. The process is repeated until all objects are finally combined into a single cluster. This process is the opposite of the *divisive method*.

**Average linkage** *Hierarchical clustering algorithm* that represents *similarity* as the average distance from all objects in one cluster to all objects in another. This approach tends to combine clusters with small variances.

**Boundary point** An object in the *density-based approach* which does not have the required *density* to constitute a cluster by itself, but does fall within the *neighborhood* of an existing cluster.

**Centering** A form of standardization where the mean of a variable is subtracted from the values of all objects. This equates the means of all centered variables at zero. Similar to z scores, except that the variances are not equivalent as well (i.e., centered values are not divided by the standard deviation of the variable).

**Centroid method** *Hierarchical clustering algorithm* in which *similarity* between clusters is measured as the distance between *cluster centroids*. When two clusters are combined, a new centroid is computed. Thus, cluster centroids migrate, or move, as the clusters are combined.

**City-block distance** Method of calculating distances based on the sum of the absolute differences of the coordinates for the *objects*. This method assumes that the variables in the *cluster variate* are uncorrelated and that unit scales are compatible.

**Cluster centroid** Average value of the objects contained in the cluster on all the variables in the *cluster variate*.

**Cluster seed** Initial value or starting point for a cluster. These values are selected to initiate *nonhierarchical procedures*, in which clusters are built around these prespecified points.

**Cluster solution** A specific number of clusters selected as representative of the data structure of the sample of *objects*.

**Cluster variate** Set of variables or characteristics representing the *objects* to be clustered and used to calculate the *similarity* between objects.

**Clustering algorithm** Set of rules or procedures; similar to an equation.

**Clustering variables** Set of variables selected/specify by the researcher to form the *cluster variate* and be the basis for calculating *interobject similarity* between *objects*.

**Complete-linkage method** *Hierarchical clustering algorithm* in which *interobject similarity* is based on the maximum distance between *objects* in two clusters (the distance between the most dissimilar members of each cluster). At each stage of the *agglomeration*, the two clusters with the smallest maximum distance (most similar) are combined.

**Component** A probability distribution of a subsample of objects which uniquely defines a cluster based on a high degree of *similarity* between the objects in the *model-based approach*.

**Core point** An object in the *density-based approach* which has a *density* above the researcher-specified level necessary to constitute a cluster.

**Cubic clustering criterion (CCC)** A direct measure of *heterogeneity* in which the highest CCC values indicate the final *cluster solution*.

**Curse of dimensionality** Related to the increase in the number of clustering variables, the impact is to diminish the ability of distance-based measures to distinguish similarity among the cases.

**Dendrogram** Graphical representation (tree graph) of the results of a *hierarchical procedure* in which each *object* is arrayed on one axis, and the other axis portrays the steps in the *hierarchical procedure*. Starting with each object represented as a separate cluster, the dendrogram shows graphically how the clusters are combined at each step of the procedure until all are contained in a single cluster.

**Density** The number of points within a *neighborhood* as used in the *density-based approach*.

**Density-based approach** Clustering method based on forming clusters based on objects which are closely spaced together (i.e., high density), using areas with a low density of objects as areas separating clusters.

**Density-reachable point** See *boundary point*.

**Diameter method** See *complete-linkage method*.

**Divisive method** *Hierarchical clustering algorithm* that begins with all *objects* in a single cluster, which is then divided at each step into two additional clusters that contain the most dissimilar objects. The single cluster is divided into two clusters, then one of these two clusters is split for a total of three clusters. This continues until all observations are in single-member clusters. This method is the opposite of the *agglomerative method*.

**Entropy group** Group of *objects* independent of any cluster (i.e., they do not fit into any cluster) that may be considered outliers and possibly eliminated from the cluster analysis.

**Euclidean distance** Most commonly used measure of the *similarity* between two *objects*. Essentially, it is a measure of the length of a straight line drawn between two objects when represented graphically.

**Farthest-neighbor method** See *complete-linkage method*.

**Heterogeneity** A measure of diversity of all observations across all clusters that is used as a general element in *stopping rules*. A large increase in heterogeneity when two clusters are combined indicates that a more natural structure exists when the two clusters are separate.

**Hierarchical procedures** Stepwise clustering procedures involving a combination (or division) of the objects into clusters. The two alternative procedures are the *agglomerative* and *divisive methods*. The result is the construction of a hierarchy, or treelike structure (*dendrogram*), depicting the formation of the clusters. Such a procedure produces  $N - 1$  cluster solutions, where  $N$  is the number of objects. For example, if the agglomerative procedure starts with five objects in separate clusters, it will show how four clusters, then three, then two, and finally one cluster are formed.

**Interobject similarity** The correspondence or association of two *objects* based on the variables of the *cluster variate*. Similarity can be measured in two ways. First is a measure of association, with higher positive correlation coefficients representing greater similarity. Second, proximity, or closeness, between each pair of objects can assess similarity. When measures of distance or difference are used, smaller distances or differences represent greater similarity.

**K-means** A group of *nonhierarchical clustering algorithms* that work by partitioning observations into a user-specified number of clusters and then iteratively reassigning observations until some numeric goal related to cluster distinctiveness is met.

**Mahalanobis distance ( $D^2$ )** Standardized form of *Euclidean distance*. Scaling responses in terms of standard deviations standardizes the data with adjustments made for correlations between the variables.

**Manhattan distance** See *city-block distance*.

**Mixture model** The statistical model underlying the model-based approach where the sample of points are assumed to be composed of a mixture of probability distributions (known as *components*), each representing a different cluster.

**Model-based approach** The only clustering approach with a statistical foundation, it differs primarily in that *similarity* based on distance/proximity is replaced by probability of membership in a specific cluster. Based upon the concept of a *mixture model*, it has become popular since it has measures of fit that allow for direct comparisons of different *cluster solutions*.

**Nearest-neighbor method** See *single-linkage method*.

**Neighborhood** The area around a point used to calculate the *density* of objects associated with that point. The diameter of the neighborhood is specified by the analyst and is used in the *density-based approach* to define *core*, *boundary* or *noise* points.

**Noise point** An object in the *density-based approach* which does not have the *density* necessary to be defined as a *core point* nor does it fall within the neighborhood of an existing cluster (i.e., not a *boundary point*). Also termed an outlier.

**Nonhierarchical procedures** Procedures that produce only a single cluster solution for a set of *cluster seeds* and a given number of clusters. Instead of using the tree-like construction process found in the *hierarchical procedures*, cluster seeds are used to group objects within a prespecified distance of the seeds. *Nonhierarchical procedures* do not produce results for all possible numbers of clusters as is done with a hierarchical procedure.

**Object** Person, product or service, firm, or any other entity that can be evaluated on a number of attributes.

**Optimizing procedure** *Nonhierarchical clustering* procedure that allows for the reassignment of *objects* from the originally assigned cluster to another cluster on the basis of an overall optimizing criterion.

**Parallel coordinates graph** See *profile diagram*.

**Profile diagram** Graphical representation of data that aids in screening for outliers or the interpretation of the final cluster solution. Typically, the variables of the cluster variate or those used for validation are listed along the horizontal axis, and the scale is the vertical axis. Separate lines depict the scores (original or standardized) for individual objects or cluster centroids/objects.

**Response-style effect** Series of systematic responses by a respondent that reflect a bias or consistent pattern. Examples include responding that an object always performs excellently or poorly across all attributes with little or no variation.

**Root mean square standard deviation (RMSSTD)** The square root of the variance of the new cluster formed by joining the two clusters across the *cluster variate*. Large increases indicate that the two clusters separately represent a more natural data structure than when joined.

**Row-centering standardization** See *within-case standardization*.

**Similarity** See *interobject similarity*.

**Single-linkage method** *Hierarchical clustering algorithm* in which *similarity* is defined as the minimum distance between any single *object* in one cluster and any single *object* in another, which simply means the distance between the closest objects in two clusters. This procedure has the potential for creating less compact, or even chain-like, clusters. It differs from the *complete-linkage method*, which uses the maximum distance between objects in the cluster.

**Squared Euclidean distance** Measure of *similarity* that represents the sum of the squared distances without taking the square root (as done to calculate *Euclidean distance*).

**Stopping rule** *Clustering algorithm* for determining the final number of clusters to be formed. With no *stopping rule* inherent in cluster analysis, researchers developed several criteria and guidelines for this determination. Two classes of rules that are applied post hoc and calculated by the researcher are (1) measures of similarity and (2) adapted statistical measures.

**Taxonomy** Empirically derived classification of actual *objects* based on one or more characteristics, as typified by the application of cluster analysis or other grouping procedures. This classification can be contrasted to a *typology*.

**Typology** Conceptually based classification of objects based on one or more characteristics. A typology does not usually attempt to group actual observations, but instead provides the theoretical foundation for the creation of a *taxonomy*, which groups actual observations.

**Ward's method** Hierarchical clustering algorithm in which the similarity used to join clusters is calculated as the sum of squares between the two clusters summed over all variables. This method has the tendency to result in clusters of approximately equal size due to its minimization of within-group variation.

**Within-case standardization** Method of standardization in which a respondent's responses are not compared to the overall sample but instead to the respondent's own responses. In this process, also known as ipsitizing, the respondents' average responses are used to standardize their own responses.

## What Is Cluster Analysis?

Cluster analysis is a group of multivariate techniques whose primary purpose is to group objects based on the characteristics they possess. It has been referred to as Q analysis, typology construction, classification analysis, and numerical taxonomy. This variety of names is due to the usage of clustering methods in such diverse disciplines as psychology, biology, sociology, economics, engineering, and business. Although the names differ across disciplines, the methods all have a common dimension: classification according to relationships among the objects being clustered [3, 4, 6, 25, 45, 52]. This common dimension represents the essence of all clustering approaches—the classification of data as suggested by natural groupings of the data themselves. Cluster analysis is comparable to exploratory factor analysis (see Chapter 3) in its objective of assessing structure. Cluster analysis differs from exploratory factor analysis, however, in that cluster analysis groups objects, whereas exploratory factor analysis is primarily concerned with grouping variables. Additionally, exploratory factor analysis makes the groupings based on patterns of variation (correlation) in the data whereas cluster analysis makes groupings on the basis of distance (proximity).

### CLUSTER ANALYSIS AS A MULTIVARIATE TECHNIQUE

Cluster analysis classifies **objects** (e.g., respondents, products, or other entities), on a set of user selected characteristics (**clustering variables**). The resulting clusters should exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity. Thus, if the classification is successful, the objects within clusters will be close together when plotted geometrically, and different clusters will be far apart.

The concept of the variate is again important in understanding how cluster analysis mathematically produces results. The **cluster variate** represents a mathematical representation of the selected set of clustering variables which compares the objects' similarities.

The variate in cluster analysis is determined quite differently from other multivariate techniques. Cluster analysis is the only multivariate technique that does not estimate the variate empirically but instead uses the variate as specified by the researcher. The focus of cluster analysis is on the comparison of objects based on the variate, not on the estimation of the variate itself. This distinction makes the researcher's definition of the variate a critical step in cluster analysis.

### CONCEPTUAL DEVELOPMENT WITH CLUSTER ANALYSIS

Cluster analysis has been used in every research setting imaginable. Ranging from the derivation of taxonomies in biology for grouping all living organisms, to psychological classifications based on personality and other personal traits, to segmentation analyses of markets, cluster analysis applications have focused largely on grouping individuals. However, cluster analysis can classify objects other than individual people, including the market structure of firms, analyses of the similarities and differences among new products, and performance evaluations of firms to identify groupings based on the firms' strategies or strategic orientations. In many instances, however, the grouping is actually a means to an end in terms of a conceptually defined goal. The more common roles cluster analysis can play in conceptual development include data reduction and hypothesis generation.

**Data Reduction** A researcher may be faced with a large number of observations that are meaningless unless classified into manageable groups. Cluster analysis can perform this data reduction procedure objectively by reducing the information from an entire population or sample to information about specific groups. For example, if we can

understand the attitudes of a population by identifying the major groups within the population, then we have reduced the data for the entire population into profiles of a number of groups. In this fashion, the researcher provides a more concise, understandable description of the observations, with minimal loss of information.

**Hypothesis Generation** Cluster analysis is also useful when a researcher wishes to develop hypotheses concerning the nature of the data or to examine previously stated hypotheses. For example, a researcher may believe that attitudes toward the consumption of diet versus regular soft drinks could be used to separate soft-drink consumers into logical segments or groups. Cluster analysis can classify soft-drink consumers by their attitudes about diet versus regular soft drinks, and the resulting clusters, if any, can be profiled for demographic similarities and differences.

The large number of applications of cluster analysis in almost every area of inquiry creates not only a wealth of knowledge on its use, but also the need for a better understanding of the technique to minimize its misuse.

## NECESSITY OF CONCEPTUAL SUPPORT IN CLUSTER ANALYSIS

Believe it or not, cluster analysis can be criticized for working too well in the sense that statistical results are produced even when a logical basis for clusters is not apparent. Thus, the researcher should have a strong conceptual basis to deal with issues such as why groups exist in the first place and what variables logically explain why objects end up in the groups that they do. Even if cluster analysis is being used in conceptual development as just mentioned, some conceptual rationale is essential. The following are the most common criticisms that must be addressed by conceptual rather than empirical support:

- *Cluster analysis is descriptive, atheoretical, and non-inferential.* Cluster analysis has no statistical basis upon which to draw inferences from a sample to a population, and many contend that it is only an exploratory technique. Nothing guarantees unique solutions, because the cluster membership for any number of solutions is dependent upon many elements of the procedure, and many different solutions can be obtained by varying one or more elements.
- *Cluster analysis will always create clusters, regardless of the actual existence of any structure in the data.* When using cluster analysis, the researcher is making an assumption of some structure among the objects. The researcher should always remember that just because clusters can be found does not validate their existence. Only with strong conceptual support and then validation are the clusters potentially meaningful and relevant.
- *The cluster solution is not generalizable because it is totally dependent upon the variables used as the basis for the similarity measure.* This criticism can be made against any statistical technique, but cluster analysis is generally considered more dependent on the measures used to characterize the objects than other multivariate techniques. With the cluster variate completely specified by the researcher, the addition of spurious variables or the deletion of relevant variables can have a substantial impact on the resulting solution. As a result, the researcher must be especially cognizant of the variables used in the analysis, ensuring that they have strong conceptual support.

Thus, in any use of cluster analysis the researcher must take particular care in ensuring that strong conceptual support predates the application of the technique. Only with this support in place should the researcher then address each of the specific decisions involved in performing a cluster analysis.

## How Does Cluster Analysis Work?

Cluster analysis performs a task innate to all individuals—pattern recognition and grouping. The human ability to process even slight differences in innumerable characteristics is a cognitive process inherent in human beings that is not easily matched even with all of our technological advances. Take for example the task of analyzing and grouping human faces. Even from birth, individuals can quickly identify slight differences in facial expressions and group different faces in homogeneous groups while considering hundreds of facial characteristics. Yet we still struggle with

facial recognition programs to accomplish the same task. The process of identifying natural groupings is one that can become quite complex rather quickly.

To demonstrate how cluster analysis operates, we examine a simple example that illustrates some of the key issues: measuring similarity, forming clusters, and deciding on the number of clusters that best represent structure. We also briefly discuss the balance of objective and subjective considerations that must be addressed by any researcher.

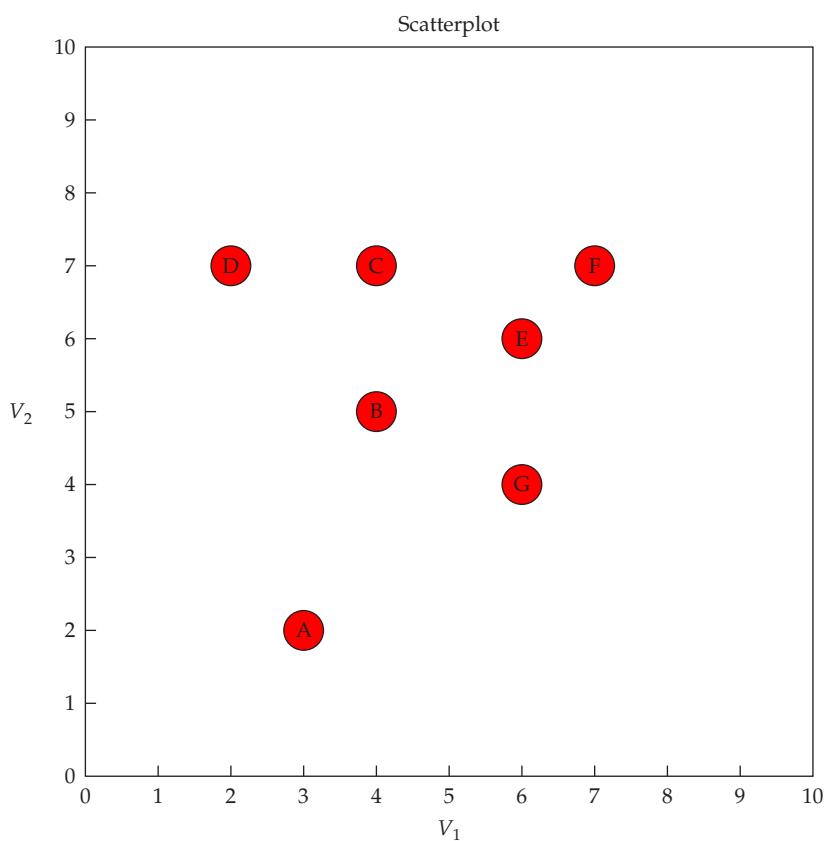
## A SIMPLE EXAMPLE

The nature of cluster analysis and the basic decisions on the part of the researcher will be illustrated by a simple example involving identification of customer segments in a retail setting.

Suppose a marketing researcher wishes to determine market segments in a community based on patterns of loyalty to brands and stores. A small sample of seven respondents is selected as a pilot test of how cluster analysis is applied. Two measures of loyalty— $V_1$  (store loyalty) and  $V_2$  (brand loyalty)—were measured for each respondent on a 0–10 scale. The values for each of the seven respondents are shown in Figure 4.1, along with a scatter diagram depicting each observation on the two variables.

Clustering Variable	Data Values						
	Respondents						
	A	B	C	D	E	F	G
$V_1$	3	4	4	2	6	7	6
$V_2$	2	5	7	7	6	7	4

**Figure 4.1**  
Data Values and Scatterplot of the Seven Observations Based on the Two Clustering Variables ( $V_1$  and  $V_2$ )



The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups. To accomplish this task, we must address three basic questions:

- 1 How do we measure similarity?** We require a method of simultaneously comparing observations on the two clustering variables ( $V_1$  and  $V_2$ ). Several methods are possible, including the correlation between objects or perhaps a measure of their proximity in two-dimensional space such that the distance between observations indicates similarity.
- 2 How do we form clusters?** No matter how similarity is measured, the procedure must group those observations that are most similar into a cluster, thereby determining the cluster group membership of each observation for each set of clusters formed. The number of cluster solutions can range from all objects being in separate clusters to a single cluster containing all objects.
- 3 How many groups do we form?** The final task is to select one cluster solution (i.e., set of clusters) as the final solution. In doing so, the researcher faces a trade-off: fewer clusters and less homogeneity within clusters versus a larger number of clusters and more within-group homogeneity. Simple structure, in striving toward parsimony, is reflected in as few clusters as possible. Yet as the number of clusters decreases, the heterogeneity within the clusters necessarily increases, thus making the differences between clusters less distinct. A balance must be made between defining the most basic structure (fewer clusters) that still achieves an acceptable level of homogeneity within the clusters and heterogeneity between the clusters.

With procedures for addressing each of these issues, we can perform a cluster analysis. We will illustrate the principles underlying each of these issues through our simple example.

**Measuring Similarity** The first task is developing some measure of similarity between each object to be used in the clustering process. **Similarity** represents the degree of correspondence among objects across all of the clustering variables. Many times our similarity measures are really dissimilarity measures in that smaller numbers represent greater similarity and larger numbers represent less similarity. This is the case with most distance measures, where larger numbers represent greater distances between objects and thus more dissimilarity (i.e., less similarity).

In our example, similarity must be determined between each of the seven observations (respondents A–G) so as to compare each observation to each other. We will measure similarity according to the Euclidean (straight-line) distance between each pair of observations (see Table 4.1) based on the two characteristics ( $V_1$  and  $V_2$ ). In this two-dimensional case (where each characteristic forms one axis of the graph) we can view distance as the proximity of each point to the others. In using distance as the measure of proximity, we must remember that smaller distances indicate greater similarity, such that observations E and F are the most similar (1.414), and A and F are the most dissimilar (6.403).

**Forming Clusters** With similarity measures calculated, we now move to forming clusters based on the similarity measure of each observation. Typically we form a number of cluster solutions (a two-cluster solution, a

**Table 4.1** Proximity Matrix of Euclidean Distances Between Observations

Observation	Observation						
	A	B	C	D	E	F	G
A	—						
B	3.162	—					
C	5.099	2.000	—				
D	5.099	2.828	2.000	—			
E	5.000	2.236	2.236	4.123	—		
F	6.403	3.606	3.000	5.000	1.414	—	
G	3.606	2.236	3.606	5.000	2.000	3.162	—

three-cluster solution, etc.). Once clusters are formed, we then select the final cluster solution from the set of possible solutions. First we will discuss how clusters are formed and then examine the process for selecting a final cluster solution.

Having calculated the similarity measure, we must develop a procedure for forming clusters. As shown later in this chapter, many methods have been proposed, but for our purposes here, we use this simple rule:

*Identify the two most similar (closest) observations not already in the same cluster and combine them.*

We apply this rule repeatedly to generate a number of cluster solutions, starting with each observation as its own “cluster” and then combining two clusters at a time until all observations are in a single cluster. This process is termed a **hierarchical procedure** because it moves in a stepwise fashion to form an entire range of cluster solutions. It is also an **agglomerative method** because clusters are formed by combining existing clusters.

Table 4.2 details the steps of the hierarchical agglomerative process, first depicting the initial state with all seven observations in single-member clusters, joining them in an agglomerative process until only one cluster remains. The clustering process results in six cluster solutions, ranging from six clusters to a single cluster:

**Step 1:** Identify the two closest observations (E and F with distance of 1.414) and combine them into a cluster, moving from *seven to six clusters*.

**Step 2:** Find the next closest pairs of observations. In this case, three pairs have the same distance of 2.000 (E–G, C–D, and B–C). For our purposes, choose the observations E–G. G is a single-member cluster, but E was combined in the prior step with F. So, the cluster formed at this stage now has three members: G, E, and F and there are *five clusters*.

**Step 3:** Combine the single-member clusters of C and D so that we now have *four clusters*.

**Step 4:** Combine B with the two-member cluster C–D that was formed in step 3. At this point, we now have *three clusters*: cluster 1 (A), cluster 2 (B, C, and D), and cluster 3 (E, F, and G).

**Step 5:** The next smallest distance is 2.236 for three pairs of observations (E–B, B–G, and C–E). We use only one of these distances, however, as each observation pair contains a member from each of the two existing clusters (B, C, and D versus E, F, and G). Combine the two three-member clusters into a single six-member cluster with observation A in a single member cluster for a *two cluster solution*.

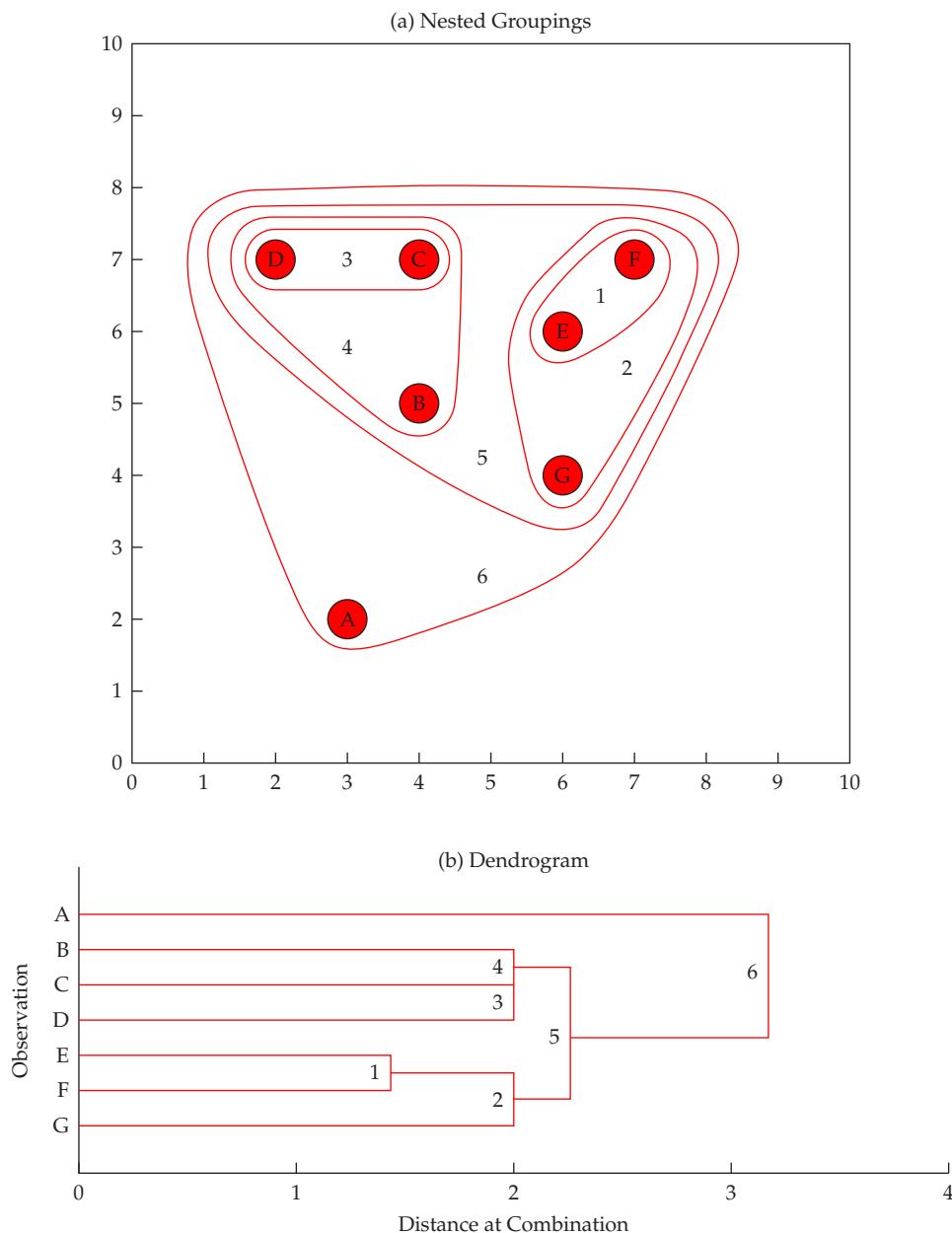
**Step 6:** Combine observation A with the remaining cluster (six observations) into a *single cluster* at a distance of 3.162. You will note that distances smaller or equal to 3.162 are not used because they are between members of the same cluster.

The hierarchical clustering process can be portrayed graphically in several ways. Figure 4.2 illustrates two such methods. First, because the process is hierarchical, the clustering process can be shown as a series of nested groupings

**Table 4.2** Agglomerative Hierarchical Clustering Process

Step	AGGLOMERATION PROCESS			CLUSTER SOLUTION		
	Minimum Distance Between Unclustered Observations <sup>a</sup>		Observation Pair	Cluster Membership (A) (B) (C) (D) (E) (F) (G)	Number of Clusters	Overall Similarity Measure (Average) Within-Cluster Distance
	Initial Solution					
1	1.414	E-F	(A) (B) (C) (D) (E-F) (G)	6	7	0 1.414
2	2.000	E-G	(A) (B) (C) (D) (E-F-G)	5	6	2.192
3	2.000	C-D	(A) (B) (C-D) (E-F-G)	4	5	2.144
4	2.000	B-C	(A) (B-C-D) (E-F-G)	3	4	2.234
5	2.236	B-E	(A) (B-C-D-E-F-G)	2	3	2.896
6	3.162	A-B	(A-B-C-D-E-F-G)	1	2	3.420

<sup>a</sup>Euclidean distance between observations



**Figure 4.2**  
Graphical Portrayals  
of the Hierarchical  
Clustering Process

(see Figure 4.2a). This process, however, can represent the proximity of the observations for only two or three clustering variables in the scatterplot or three-dimensional graph. A more common approach is a **dendrogram**, which represents the clustering process in a tree-like graph. The horizontal axis represents the agglomeration coefficient, in this instance the distance used in joining clusters. This approach is particularly useful in identifying outliers, such as observation A. It also depicts the relative size of varying clusters, although it becomes unwieldy when the number of observations increases.

**Determining the Number of Clusters in the Final Solution** A hierarchical method results in a number of cluster solutions—in this case starting with a seven single member clusters and ending in a one-cluster solution. Which solution do we choose? We know that as we move from single-member clusters in the seven-cluster

solution, heterogeneity increases as we join observations together. So why not stay at seven clusters, the most homogeneous possible? If all observations are treated as their own unique cluster, no data reduction has taken place and no true segments have been found. The goal is identifying segments by combining observations, but at the same time introducing only small amounts of heterogeneity.

**MEASURING HETEROGENEITY** Any measure of **heterogeneity** of a cluster solution should represent the overall diversity among observations in all clusters. In the initial solution of an agglomerative approach where all observations are in separate clusters, no heterogeneity exists. As observations are combined to form clusters, heterogeneity increases. The measure of heterogeneity thus should start with a value of zero and increase to show the level of heterogeneity as clusters are combined.

In this example, we use a simple measure of heterogeneity: the average of all distances between observations within clusters (see Table 4.2). As already described, the measure should increase as clusters are combined:

- In the initial solution with seven clusters, our overall similarity measure is 0—no observation is paired with another.
- Six clusters: The overall similarity is the distance between the two observations (1.414) joined in step 1.
- Five clusters: Step 2 forms a three-member cluster (E, F, and G), so that the overall similarity measure is the mean of the distances between E and F (1.414), E and G (2.000), and F and G (3.162), for an average of 2.192.
- Four clusters: In the next step a new two-member cluster is formed with a distance of 2.000, which causes the overall average to fall slightly to 2.144.
- Three, two, and one clusters: The final three steps form new clusters in this manner until a single-cluster solution is formed (step 6), in which the average of all distances in the distance matrix is 3.420.

**SELECTING A FINAL CLUSTER SOLUTION** Now, how do we use this overall measure of similarity to select a **cluster solution?** Remember that we are trying to get the simplest structure possible that still represents homogeneous groupings. A fundamental tenet of comparing cluster solutions: When monitoring the heterogeneity measure as the number of clusters decreases, large increases in heterogeneity indicate that two rather dissimilar clusters were joined at that stage.

From Table 4.2, we can see that the overall measure of heterogeneity increases as we combine clusters until we reach the final one-cluster solution. Table 4.3 shows the increases in similarity and proportionate increases across the cluster solutions. To select a final cluster solution, we examine the proportionate changes in the homogeneity measure to identify large increases indicative of merging dissimilar clusters:

- When we first join two observations (step 1) we establish the minimum heterogeneity level within clusters, in this case 1.414.
- In Step 2 we see a substantial increase in heterogeneity from Step 1, but in the next two steps (3 and 4), the overall measure does not change substantially, which indicates that we are forming other clusters with essentially the same heterogeneity of the existing clusters.
- When we get to step 5, which combines the two three-member clusters, we see a large increase (.662 or 29.6%). This change indicates that joining these two clusters resulted in a single cluster that was markedly less homogeneous. As a result, we would consider the three-cluster solution of step 4 much better than the two-cluster solution found in step 5.
- We can also see that in step 6 the overall measure again increased markedly, indicating when this single observation was joined at the last step, it substantially changed the cluster homogeneity. Given the rather unique profile of this observation (observation A) compared to the others, it might best be designated as a member of the **entropy group**, those observations that are outliers and independent of the existing clusters.

Thus, when reviewing the range of cluster solutions, the three-cluster solution of step 4 seems the most appropriate for a final cluster solution, with two equally sized clusters and the single outlying observation.

**Table 4.3** Selecting the Final Cluster Solution Based on Percentage Changes in Heterogeneity

Step	Overall Similarity Measure	Difference in Similarity to Next Step	Percentage Increase in Heterogeneity
			to Next Stage
1	1.414	.778	55.0
2	2.192	-.048	NI
3	2.144	.090	4.2
4	2.234	.662	29.6
5	2.896	.524	18.1
6	3.420	—	—

NI: No increase in heterogeneity

## OBJECTIVE VERSUS SUBJECTIVE CONSIDERATIONS

As is probably clear by now, the selection of the final cluster solution requires substantial researcher judgment and is considered by many as too subjective. Even though many diagnostic measures have been developed to assist in evaluating the cluster solutions, it still falls to the researcher to make the final decision as to the number of clusters to accept as the final solution. Moreover, decisions on the characteristics to be used, the methods of combining clusters, and even the interpretation of cluster solutions rely as much on the judgment of the researcher as any empirical test.

Even this rather simple example of only two characteristics and seven observations demonstrates the potential complexity of performing a cluster analysis. Researchers in realistic settings are faced with analyses containing many more characteristics with many more observations. It is thus imperative researchers employ whatever objective support is available and be guided by reasoned judgment, especially in the design and interpretation stages.

## Cluster Analysis Decision Process

Cluster analysis, like the other multivariate techniques discussed earlier, can be viewed from the six-stage model-building approach introduced in Chapter 1 (see Figure 4.3 for stages 1–3 and Figure 4.6 for stages 4–6). Starting with research objectives that can be either exploratory or confirmatory, the design of a cluster analysis deals with the following:

- Partitioning the data set to form clusters and selecting a cluster solution
- Interpreting the clusters to understand the characteristics of each cluster and develop a name or label that appropriately defines its nature
- Validating the results of the final cluster solution (i.e., determining its stability and generalizability), along with describing the characteristics of each cluster to explain how they may differ on relevant dimensions such as demographics.

The following sections detail all these issues through the six stages of the model-building process.

### STAGE 1: OBJECTIVES OF CLUSTER ANALYSIS

The primary goal of cluster analysis is to partition a set of objects into two or more groups based on the similarity of the objects for a set of specified characteristics (clustering variables forming the cluster variate). In fulfilling this basic objective, the researcher must address two key issues: the research questions being addressed in this analysis and the variables used to characterize objects in the clustering process. We will discuss each issue in the following section.

**Figure 4.3**  
Stages 1–3 of the Cluster Analysis Decision Diagram

**Stage 1**

Research Problem  
Select objective(s):  
Taxonomy description  
Data simplification  
Reveal relationships  
Select clustering variables

**Stage 2**

Research Design Issues  
Can outliers be detected?  
Should the data be standardized?

Select a Similarity Measure  
Are the cluster variables metric or nonmetric?

Is the focus on pattern or proximity?

Proximity:

Distance Measures of Similarity  
Euclidean distance  
City-block distance  
Mahalanobis distance

Pattern:

Correlation Measure of Similarity  
Correlation coefficient

Nonmetric Data:

Association of Similarity  
Matching coefficients

Standardization Options

Standardizing variables  
Standardizing by observation

**Stage 3**

Assumptions  
Is the sample representative of the population?  
Is multicollinearity substantial enough to affect results?

To  
Stage  
4

**Research Questions in Cluster Analysis** In forming homogeneous groups, cluster analysis may address any combination of three basic research questions:

- 1 *Taxonomy description.* The most traditional use of cluster analysis has been for exploratory purposes and the formation of a **taxonomy**—an empirically-based classification of objects. As described earlier, cluster analysis has been used in a wide range of applications for its partitioning ability. Cluster analysis can also generate hypotheses related to the structure of the objects. Finally, although viewed principally as an exploratory technique, cluster analysis can be used for confirmatory purposes. In such cases, a proposed **typology** (theoretically-based classification) can be compared to that derived from the cluster analysis.
- 2 *Data simplification.* By defining structure among the observations, cluster analysis also develops a simplified perspective by grouping observations for further analysis. Whereas exploratory factor analysis attempts to provide dimensions or structure to variables (see Chapter 3), cluster analysis performs the same task for observations. Thus, instead of viewing all of the observations as unique, they can be viewed as members of clusters and profiled by their general characteristics.
- 3 *Relationship identification.* With the clusters defined and the underlying structure of the data represented in the clusters, the researcher has a means of revealing relationships among the observations that typically is not possible with the individual observations. Whether analyses such as discriminant analysis are used to empirically identify relationships, or the groups are examined by more qualitative methods, the simplified structure from cluster analysis often identifies relationships or similarities and differences not previously revealed.

**Selection of Clustering Variables** The objectives of cluster analysis cannot be separated from the selection of clustering variables used to characterize the objects being clustered and calculate the similarity between them. Whether the objective is exploratory or confirmatory, the researcher effectively constrains the possible results by the variables selected for use. The derived clusters reflect the inherent structure of the data and similarity among objects is based solely on the clustering variables. Thus, selecting the variables to be included in the cluster variate must be done with regard to theoretical and conceptual considerations.

Any application of cluster analysis must have some rationale upon which variables are selected to represent similarity among objects. Whether the rationale is based on an explicit theory, past research, or supposition, the researcher must realize the importance of including only those variables that (1) characterize the objects being clustered and (2) relate specifically to the objectives of the cluster analysis. The cluster analysis technique has no means of differentiating relevant from irrelevant variables and derives the most consistent, yet distinct, groups of objects across all variables. Thus, one should never include variables indiscriminately. Instead, carefully choose the variables with the research objective as the criterion for selection.

Let's use the HBAT dataset to provide an example of how to select the appropriate variables for a cluster analysis. First, variables  $X_1$  to  $X_5$  are nonmetric data warehouse classification variables. Thus, they are not appropriate for cluster analysis. Next, let us consider variables  $X_6$  to  $X_{18}$ . These 13 variables are appropriate because they all have a common foundation—they relate to customer's perceptions of the performance of HBAT and they are measured metrically. If we used these perceptions variables for a cluster analysis, the objective would be to see if there are groups of HBAT customers that exhibit distinctively different perceptions of the performance of HBAT between the groups, but similar perceptions within each of the groups. Finally, we need to consider variables  $X_{19}$  to  $X_{23}$ . These variables would not be considered part of the perceptions clustering variables because they are distinct from variables  $X_6$  to  $X_{18}$ . We might consider  $X_{19}$  to  $X_{21}$  as validation measures for our derived clusters or even use them as clustering variables because they all relate to the construct of customer commitment or loyalty. But in both instances they would be included in the analysis separately from the perceptions variables.

## Objectives of Cluster Analysis

Cluster analysis is used for:

Taxonomy description—identifying natural groups within the data.

Data simplification—the ability to analyze groups of similar observations instead of all individual observations.

Relationship identification—the simplified structure from cluster analysis portrays relationships not revealed otherwise.

Theoretical and conceptual considerations must be observed when selecting clustering variables for cluster analysis. Similarity among objects can only be calculated based on the clustering variables, so make sure the variables selected best characterize the features of the objects that meet the research objectives.

## STAGE 2: RESEARCH DESIGN IN CLUSTER ANALYSIS

With the objectives defined and variables selected, the researcher must address five questions before starting the partitioning process:

- 1 What types and how many clustering variables can be included?
- 2 Is the sample size adequate?
- 3 Can outliers be detected and, if so, should they be deleted?
- 4 How should object similarity be measured?
- 5 Should the data be standardized?

Many different approaches can be used to answer these questions. However, none of them has been evaluated sufficiently to provide a definitive answer to any of these questions, and unfortunately, many of the approaches provide different results for the same dataset. Thus, cluster analysis, along with exploratory factor analysis, is as much an art as a science. For this reason, our discussion reviews these issues by providing examples of the most commonly used approaches and an assessment of the practical limitations where possible.

The importance of these issues and the decisions made in later stages becomes apparent when we realize that although cluster analysis is seeking structure in the data, it must actually impose a structure through a selected methodology. Cluster analysis cannot evaluate all the possible partitions because even the relatively small problem of partitioning 25 objects into five non-overlapping clusters involves  $2.431 \times 10^{15}$  possible partitions [4]. Instead, based on the decisions of the researcher, the technique identifies a small subset of possible solutions as “correct.” From this viewpoint, the research design issues and the choice of methodologies made by the researcher perhaps have greater impact than with any other multivariate technique.

**Types and Number of Clustering Variables** The practical considerations impacting variable selection have a direct relationship to the conceptual considerations discussed above. They involve three related issues: type of variables to be included; number of clustering variables and impact of irrelevant or inappropriate variables.

**TYPE OF VARIABLES INCLUDED** While cluster analysis can incorporate either metric or nonmetric variables, it is most limited when the clustering variables are mixed (both metric and nonmetric). Metric variables are the most common form of clustering variable and are typically associated with distance-based similarity measures such as we illustrated in our simple example above. And yet there are multiple ways in which similarity can be calculated for nonmetric data. What is less well developed is a set of clustering variables which has both types. While predictive models in

most instances can easily handle both variable types simultaneously, cluster analysis (and even exploratory factor analysis) are best suited for analysis of a single type of variable.

**NUMBER OF CLUSTERING VARIABLES** One substantial benefit of multivariate analysis in general is the ability to analyze multiple variables simultaneously, i.e., the variate. In cluster analysis this allows for a multidimensional characterization of objects and calculation of similarity. For many of the multivariate techniques we discuss in this text the number of variables included in the analysis is not constrained. In most predictive models, if multicollinearity is mitigated, the number of independent variables is not problematic. And exploratory factor analysis is specifically developed for a large number of variables. But cluster analysis can face serious difficulties as the number of variables increases. This issue relates to the calculation of proximity/similarity that becomes untenable as the dimensionality increases since the distance to even the most similar cases becomes less distinguishable from the majority of points. This is commonly termed the **curse of dimensionality** [8]. Research has shown that this effect starts with dimensionality as low as 20 variables [1]. As a result, the researcher certainly wants to include all of the necessary characteristics that address the research problem, but must also realize that there are practical limits in the number of clustering variables to be included. As research questions become more complex it may be useful to incorporate variable reduction approaches, such as exploratory factor analysis, to assist in providing a complete profile of the objects in the least number of variables possible.

**RELEVANCY OF CLUSTERING VARIABLES** Cluster analysis is one of the few techniques that has no method for detecting the relevancy of clustering variables in generating cluster solutions. There is not any form of variable selection (e.g., sequential selection in linear models) that can eliminate those variables that are not contributing to distinctiveness between clusters. The importance of this is that the cluster results can be affected dramatically by the inclusion of only one or two inappropriate or undifferentiated variables [40, 53]. As a result, the researcher is always encouraged to include those variables with only the strongest conceptual support, but then also examine the cluster solutions and eliminate the variables that are not distinctive (i.e., that do not differ significantly) across the derived clusters. The cluster analysis can even be re-run excluding these variables. Variable selection methods are becoming available to assess the “best” set of variables to meet the clustering objectives [17, 46]. This process enables the cluster techniques to maximally define clusters based only on those variables exhibiting differences across the objects.

**Sample Size** The issue of sample size in cluster analysis does not relate to any statistical inference issues (i.e., statistical power), but instead represents a trade-off between representation and overall sample size. First, the sample size must be large enough to provide sufficient representation of small groups within the population and represent the underlying structure. This issue of representation becomes critical in detecting outliers (see next section), with the primary question being: When an outlier is detected, is it a representative of a small but substantive group? Small groups will naturally appear as small numbers of observations, particularly when the sample size is small. For example, when a sample contains only 100 or fewer observations, groups that actually make up 10 percent of the population may be represented by only one or two observations due to the sampling process. In such instances the distinction between outliers and representatives of a small group is much harder to make. Larger samples increase the chance that small groups will be represented by enough cases to make their presence more easily identified.

As a result, the researcher should ensure the sample size is sufficiently large to adequately represent all of the relevant groups of the population. In determining the sample size, the researcher should specify the group sizes necessary for relevance for the research questions being asked. Obviously, if the analysis objectives require identification of small groups within the population, the researcher should strive for larger sample sizes. If the researcher is interested only in larger groups (e.g., major segments for promotional campaigns), however, then the distinction between an outlier and a representative of a small group is less important and they can both be handled in a similar manner.

But increasing the sample size also presents problems in estimation of some cluster methods. In general this issue arises from the case-to-case level of analysis used by cluster analysis where the size of the sample has a direct bearing on the scale of the problem. In a simple sense, cluster analysis requires a  $N$  by  $N$  matrix of similarities, where  $N$  is

the sample size. In other techniques, such as exploratory factor analysis or predictive models, increasing the sample size doesn't expand the scale of the estimation procedure. But in cluster analysis, going from a 100 by 100 matrix of similarities to a 1 million by 1 million matrix is a substantial undertaking. This is especially problematic for hierarchical methods of cluster formation. So as the sample size increases, the clustering methods must change as well and we will discuss some of these approaches in a later section.

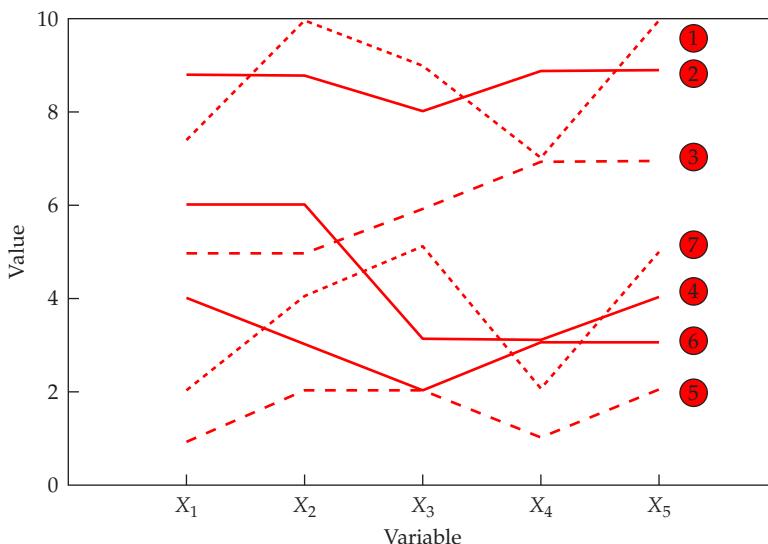
New programs have also been developed for applications using large sample sizes approaching 1,000 observations or more that will be discussed in a later section. For example, IBM SPSS and SAS provide two-step cluster programs that have the ability to quickly determine an appropriate number of groups and then classify them using a nonhierarchical routine. This procedure is relatively new, but it may prove useful in applications with large samples where traditional clustering methods are inefficient.

**Detecting Outliers** In its search for structure, we have already discussed how cluster analysis is sensitive to the inclusion of irrelevant variables. But cluster analysis is also sensitive to outliers (objects different from all others). Outliers can represent either:

- Truly aberrant observations that are not representative of the general population, members of the entropy group.
- Representative observations of small or insignificant segments within the population
- An undersampling of actual group(s) in the population that causes poor representation of the group(s) in the sample

In the first case, the outliers distort the actual structure and make the derived clusters unrepresentative of the actual population structure. In the second case, the outlier is removed so that the resulting clusters more accurately represent the segments in the population with sufficient size. However, in the third case the outliers should be included in the cluster solutions, even if they are underrepresented in the sample, because they represent valid and relevant groups. For this reason, a preliminary screening for outliers is always necessary.

**GRAPHICAL APPROACHES** One of the simplest ways to screen data for outliers is to prepare a graphic **profile diagram** or **parallel coordinates graph**, listing the variables along the horizontal axis and the variable values along the vertical axis. Each point on the graph represents the value of the corresponding variable, and the points are connected to facilitate visual interpretation. Profiles for all objects are then plotted on the graph, a line for each object. Outliers are those respondents that have very different profiles from the more typical respondents. An example of a graphic profile diagram is shown in Figure 4.4. This approach can also be used in Stage 5 for the profiling/interpretation where objects in each cluster are differentiated (e.g., by colors) to provide cluster profiles.



**Figure 4.4**  
Profile Diagram

**EMPIRICAL APPROACHES** Although quite simple, the graphical procedures become cumbersome with a large number of objects and even more difficult as the number of variables increases. Moreover, detecting outliers must extend beyond a univariate approach, because outliers also may be defined in a multivariate sense as having unique profiles across an entire set of variables that distinguish them from all of the other observations. As a result, an empirical measure is needed to facilitate comparisons across objects. For these instances, the procedures for identifying outliers discussed in Chapter 2 can be applied. The combination of bivariate and multivariate approaches provides a comprehensive set of tools for identifying outliers from many perspectives.

Another approach is to identify outliers through the measures of similarity. The most obvious examples of outliers are single observations that are the most dissimilar to the other observations. Before the analysis, the similarities of all observations can be compared to the overall group centroid (typical respondent). Isolated observations showing great dissimilarity can be dropped.

Finally, the actual cluster results can be used in an iterative manner. Here outliers are seen as single member or very small clusters within the cluster solution. These clusters can then be eliminated from the analysis if needed and the analysis performed again. However, as the number of objects increases, multiple iterations are typically needed to identify all of the outliers in this manner. Moreover, some of the clustering approaches are quite sensitive to removing just a few cases [33], so the inclusion of outliers in initial solutions may impact the composition of the other clusters in that solution and when re-run, the retained clusters may differ. Thus, emphasis should be placed on identifying outliers before the analysis begins.

**Measuring Similarity** The concept of similarity is fundamental to cluster analysis. **Interobject similarity** is an empirical measure of correspondence, or resemblance, between objects to be clustered. Comparing the two interdependence techniques (exploratory factor analysis and cluster analysis) will demonstrate how similarity works to define structure in both instances.

- In our discussion of factor analysis, the correlation matrix between all pairs of variables was used to group variables into factors. The correlation coefficient represented the similarity of each variable to another variable when viewed across all observations. Thus, factor analysis grouped together variables that had high correlations among themselves.
- A comparable process occurs in cluster analysis. Here, the similarity measure is calculated for all pairs of objects, with similarity based on the profile of each observation across the clustering variables specified by the researcher. In this way, any object can be compared to any other object through the similarity measure, just as we used correlations between variables in exploratory factor analysis. The cluster analysis procedure then proceeds to group similar objects together into clusters.

Interobject similarity can be measured in a variety of ways, but three methods dominate the applications of cluster analysis: correlational measures, distance measures, and association measures. Both the correlational and distance measures require metric data, whereas the association measures are for nonmetric data.

**CORRELATIONAL MEASURES** The interobject measure of similarity that probably comes to mind first is the correlation coefficient between a pair of objects measured on several variables. In effect, instead of correlating two sets of variables, we invert the data matrix so that the columns represent the objects and the rows represent the variables. Thus, the correlation coefficient between the two columns of numbers is the correlation (or similarity) between the profiles of the two objects. High correlations indicate similarity (the correspondence of patterns across the characteristics) and low correlations denote a lack of it. This procedure is also followed in the application of Q-type exploratory factor analysis (see Chapter 3).

The correlation approach is illustrated by using the example of seven observations shown in Figure 4.4. A correlational measure of similarity does not look at the observed mean value, or magnitude, but instead at the patterns of movement seen as one traces the data for each case over the variables measured; in other words, the similarity in the profiles for each case. In Table 4.4, which contains the correlations among these seven observations, we can see two distinct groups. First, cases 1, 5, and 7 all have similar patterns and corresponding high positive correlations. Likewise, cases 2, 4, and 6 also have high positive correlations among themselves but low or negative correlations

**Table 4.4** Calculating Correlational and Distance Measures of Similarity

Part A: Original Data						
Case	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	X <sub>5</sub>	
1	7	10	9	7	10	
2	9	9	8	9	9	
3	5	5	6	7	7	
4	6	6	3	3	4	
5	1	2	2	1	2	
6	4	3	2	3	3	
7	2	4	5	2	5	

Part B: Similarity Measure: Correlation							
Case	1	2	3	4	5	6	7
1	1.00						
2	-.147	1.00					
3	.000	.000	1.00				
4	.087	.516	-.824	1.00			
5	.963	-.408	.000	-.060	1.00		
6	-.466	.791	-.354	.699	-.645	1.00	
7	.891	-.516	.165	-.239	.963	-.699	1.00

Part C: Similarity Measure: Euclidean Distance							
Case	1	2	3	4	5	6	7
1	nc						
2	3.32	nc					
3	6.86	6.63	nc				
4	10.25	10.20	6.00	nc			
5	15.78	16.19	10.10	7.07	nc		
6	13.11	13.00	7.28	3.87	3.87	nc	
7	11.27	12.16	6.32	5.10	4.90	4.36	nc

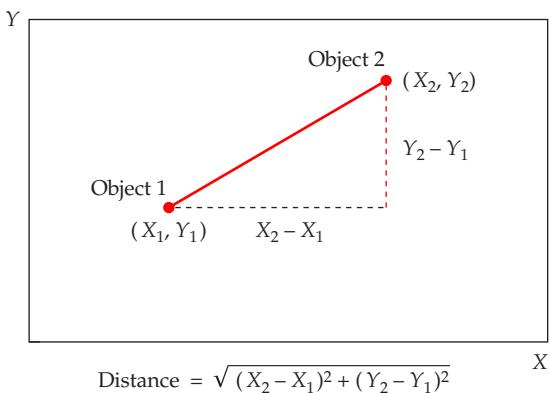
nc = distances not calculated.

with the other observations. Case 3 has low or negative correlations with all other cases, thereby perhaps forming a group by itself.

Correlations represent patterns across the variables rather than the magnitudes, which are comparable to a Q-type exploratory factor analysis (see Chapter 3). Correlational measures are rarely used in most applications of cluster analysis, however, because emphasis is on the magnitudes of the objects, not the patterns of values.

**DISTANCE MEASURES** Even though correlational measures have an intuitive appeal and are used in many other multivariate techniques, they are not the most commonly used measure of similarity in cluster analysis. Instead, the most commonly used measures of similarity are distance measures. These distance measures represent similarity as the proximity of observations to one another across the variables in the cluster variate. Distance measures are actually a measure of dissimilarity, with larger values denoting lesser similarity. Distance is converted into a similarity measure by using an inverse relationship.

A simple illustration of using distance measures was shown in our hypothetical example (see Figure 4.2), in which clusters of observations were defined based on the proximity of observations to one another when each observation's scores on two variables were plotted graphically. Even though proximity may seem to be a simple concept, several distance measures are available, each with specific characteristics.



**Figure 4.5**  
An Example of Euclidean Distance Between Two Objects Measured on Two Variables, X and Y

- **Euclidean distance** is the most commonly recognized measure of distance, many times referred to as straight-line distance. An example of how Euclidean distance is obtained is shown geometrically in Figure 4.5. Suppose that two points in two dimensions have coordinates  $(X_1, Y_1)$  and  $(X_2, Y_2)$ , respectively. The Euclidean distance between the points is the length of the hypotenuse of a right triangle, as calculated by the formula under the figure. This concept is easily generalized to more than two variables.
- **Squared (or absolute) Euclidean distance** is the sum of the squared differences without taking the square root. The squared Euclidean distance has the advantage of not having to take the square root, which speeds computations markedly. It is the recommended distance measure for the centroid and Ward's methods of clustering.
- **City-block (Manhattan) distance** is not based on Euclidean distance. Instead, it uses the sum of the absolute differences of the variables (i.e., the two sides of a right triangle rather than the hypotenuse). This procedure is the simplest to calculate, but may lead to invalid clusters if the clustering variables are highly correlated [51]. It has been found, however, to be more effective when the number of clustering variables becomes larger [1].
- **Mahalanobis distance ( $D^2$ )** is a generalized distance measure that accounts for the correlations among variables in a way that weights each variable equally. It also relies on standardized variables and will be discussed in more detail in a following section. Although desirable in many situations, it is not available as a proximity measure in either SAS or IBM SPSS.

Other distance measures (other forms of differences or the powers applied to the differences) are available in many clustering programs. The researcher is encouraged to explore alternative cluster solutions obtained when using different distance measures in an effort to best represent the underlying data patterns. Although these distance measures are said to represent similarity, in a very real sense they better represent dissimilarity, because higher values typically mean relatively less similarity. Greater distance means observations are less similar. Some software packages actually use the term *dissimilarity* because of this fact.

**COMPARING DISTANCE TO CORRELATIONAL MEASURES** The difference between correlational and distance measures perhaps can be best illustrated by referring again to Figure 4.4. Distance measures focus on the magnitude of the values and portray as similar the objects that are close together, even if they have different patterns across the variables. In contrast, correlation measures focus on the patterns across the variables and do not consider the magnitude of the differences between objects. Let us look at our seven observations to see how these approaches differ.

Table 4.4 contains the values for the seven observations on the five variables ( $X_1$  to  $X_5$ ), along with both distance and correlation measures of similarity. Cluster solutions using either similarity measure seem to indicate three clusters, but the membership in each cluster is quite different.

With smaller distances representing greater similarity, we see that cases 1 and 2 form one group (distance of 3.32), and cases 4, 5, 6, and 7 (distances ranging from 3.87 to 7.07) make up another group. The distinctiveness of these two groups from each other is shown in that the smallest distance between the two clusters is 10.20. These two clusters represent observations with higher versus lower values. A third group, consisting of only case 3, differs from the other two groups because it has values that are both low and high.

Using the correlation as the measure of similarity, three clusters also emerge. First, cases 1, 5, and 7 are all highly correlated (.891 to .963), as are cases 2, 4, and 6 (.516 to .791). Moreover, the correlations between clusters are generally close to zero or even negative. Finally, case 3 is again distinct from the other two clusters and forms a single-member cluster.

A correlational measure focuses on patterns rather than the more traditional distance measure and requires a different interpretation of the results by the researcher. Because of this, the researcher will not focus on the actual group centroids on the clustering variables, as is done when distance measures are used. Interpretation depends much more heavily on patterns that become evident in the results.

**WHICH DISTANCE MEASURE IS BEST?** In attempting to select a particular distance measure, the researcher should remember the following caveats:

- Different distance measures or a change in the scales of the variables may lead to different cluster solutions. Thus, it is advisable to use several measures and compare the results with theoretical or known patterns.
- When the variables are correlated (either positively or negatively), the Mahalanobis distance measure is likely to be the most appropriate because it adjusts for correlations and weights all variables equally. When not available, the researcher may wish to avoid using highly redundant variables as input to cluster analysis.

**ASSOCIATION MEASURES FOR NONMETRIC DATA** Association measures of similarity are used to compare objects whose characteristics are measured only in nonmetric terms (nominal or ordinal measurement). As an example, respondents could answer yes or no on a number of statements. An association measure could assess the degree of agreement or matching between each pair of respondents. The simplest form of association measure would be the percentage of times agreement occurred (both respondents said yes to a question or both said no) across the set of questions.

Extensions of this simple matching coefficient have been developed to accommodate multicategory nominal variables and even ordinal measures. Many computer programs, however, offer only limited support for association measures, and the researcher is forced to first calculate the similarity measures and then input the similarity matrix into the cluster program. Reviews of the various types of association measures can be found in several sources [20, 32, 33, 52, 21].

**SELECTING A SIMILARITY MEASURE** Although three different forms of similarity measures are available, the most frequently used and preferred form is the distance measure for several reasons. First, the distance measure best represents the concept of proximity, which is fundamental to cluster analysis. Correlational measures, although having widespread application in other techniques, represent patterns rather than proximity. Second, cluster analysis is typically associated with characteristics measured by metric variables. In some applications, nonmetric characteristics dominate, but most often the characteristics are represented by metric measures making distance again the preferred measure. As noted in an earlier discussion, most similarity measures are appropriate for either metric or nonmetric data, but are not applicable when both types of variables are in the cluster variate. Thus, in any situation, the researcher is provided measures of similarity that can represent the proximity of objects across a set of metric or nonmetric variables.

**Standardizing the Data** With the similarity measure selected, the researcher must address one more question: Should the data be standardized before similarities are calculated? In answering this question, the researcher must consider that most cluster analyses using distance measures are quite sensitive to differing scales or magnitudes among the variables. In general, variables with larger dispersion (i.e., larger standard deviations) have more impact on the final similarity value.

Clustering variables that are not all of the same scale should be standardized whenever necessary to avoid instances where a variable's influence on the cluster solution is greater than it should be [5]. We will now examine several approaches to standardization available to researchers.

**STANDARDIZING THE VARIABLES** The most common form of standardization is the conversion of each variable to standard scores (also known as Z scores) by subtracting the mean and dividing by the standard deviation for each variable. This option can be found in all computer programs and many times is even directly included in the cluster analysis procedure. The process converts each raw data score into a standardized value with a mean of 0 and a standard deviation of 1, and in turn, eliminates the bias introduced by the differences in the scales of the several attributes or variables used in the analysis.

There are two primary benefits from standardization. First, it is much easier to compare between variables because they are on the same scale (a mean of 0 and standard deviation of 1). Positive values are above the mean, and negative values are below. The magnitude represents the number of standard deviations the original value is from the mean. Second, no difference occurs in the standardized values when only the scale changes. For example, when we standardize a measure of time duration, the standardized values are the same whether measured in minutes or seconds.

Thus, using standardized variables truly eliminates the effects due to scale differences not only across variables, but for the same variable as well. The need for standardization is minimized when all of the variables are measured on the same response scale (e.g., a series of attitudinal questions), but becomes quite important whenever variables using quite different measurement scales are included in the cluster variate.

**USING A STANDARDIZED DISTANCE MEASURE** A The most common approach is to substitute the standardized values for the original clustering variables and then perform the cluster analysis. This option is directly available in many clustering programs or the data can be standardized before application of the cluster procedure.

Another measure of Euclidean distance that directly incorporates a standardization procedure is the Mahalanobis distance ( $D^2$ ). The Mahalanobis approach not only performs a standardization process on the data by scaling in terms of the standard deviations but it also sums the pooled within-group variance–covariance, which adjusts for correlations among the variables. Highly correlated sets of variables in cluster analysis can implicitly overweight one set of variables in the clustering procedures (see discussion of multicollinearity in stage 3). In short, the Mahalanobis generalized distance procedure computes a distance measure between objects comparable to  $R^2$  in regression analysis. Although many situations are appropriate for use of the Mahalanobis distance, not all programs include it as a measure of similarity. In such cases, the researcher usually selects the squared Euclidean distance.

**STANDARDIZING BY OBSERVATION** Up to now we discussed standardizing only variables. Why might we standardize respondents or cases? Let us take a simple example.

Suppose we collected a number of ratings on a 10-point scale of the importance for several attributes used in purchase decisions for a product. We could apply cluster analysis and obtain clusters, but one distinct possibility is that what we would get are clusters of people who said everything was important, some who said everything had little importance, and perhaps some clusters in between. What we are seeing are patterns of responses specific to an individual. These patterns may reflect a specific way of responding to a set of questions, such as yea-sayers (answer favorably to all questions) or naysayers (answer unfavorably to all questions).

These patterns of yea-sayers and naysayers represent what are termed **response-style effects**. If we want to identify groups according to their response style and even control for these patterns, then the typical standardization through calculating Z scores is not appropriate. What is desired in most instances is the relative importance of one variable to another for each individual. In other words, is attribute 1 more important than the other attributes, and can clusters of respondents be found with similar patterns of importance? In this instance, standardizing by respondent would standardize each question not to the sample's average but instead to that respondent's average score. This **within-case** or **row-centering standardization** can be quite effective in removing response-style effects and is especially suited to many forms of attitudinal data [50]. We should note that this approach is similar to a correlational measure in highlighting the pattern across variables, but the proximity of cases still determines the similarity value.

**SHOULD YOU STANDARDIZE?** Standardization provides a remedy to a fundamental issue in similarity measures, particularly distance measures, and many recommend its widespread use [28, 32]. However, the researcher should not apply standardization without consideration for its consequences of removing some natural relationship reflected in the scaling of the variables [54], while others have said it may be appropriate [3]. Some researchers demonstrate that it may not even have noticeable effects [18, 40]. Thus, no single reason tells us to use standardized variables versus unstandardized variables. The decision to standardize should be based on both empirical and conceptual issues reflecting both the research objectives and the empirical qualities of the data. For example, a researcher may wish to consider standardization if clustering variables with substantially different scales or if preliminary analysis shows that the cluster variables display large differences in standard deviations.

## Research Design in Cluster Analysis

Remember, the research design issues and the choice of methodologies made by the researcher perhaps have greater impact on cluster analysis than with any other multivariate technique.

Selection of variables must consider the following issues:

While metric variables are most often used, nonmetric variables can form the basis of similarity measures as well. But few similarity measures are available that allow for using both metric and nonmetric variables in the same cluster variate.

The effectiveness of similarity measures to differentiate among objects becomes diminished as the number of clustering variables increases. While there is no specific upper limit, this effects have been demonstrated with as few as 20 clustering variables.

Theoretical and practical considerations should be used to ensure that only relevant variables are included in the analysis, since even a small number of irrelevant/indistinguishable variables can impact the cluster solutions.

The sample size required is not based on statistical considerations for inference testing, but rather:

Sufficient size is needed to ensure representativeness of the population and its underlying structure, particularly small groups within the population.

Minimum group sizes are based on the relevance of each group to the research question and the confidence needed in characterizing that group.

Similarity measures calculated across the entire set of clustering variables allow for the grouping of observations and their comparison to each other:

Distance measures are most often used as a measure of similarity, with higher values representing greater dissimilarity (distance between cases), not similarity.

Less frequently used are correlational measures, where large values do indicate similarity.

Given the sensitivity of some procedures to the similarity measure used, the researcher should employ several distance measures and compare the results from each with other results or theoretical/known patterns.

Outliers can severely distort the representativeness of the results if they appear as structure (clusters) inconsistent with the research objectives.

They should be removed if the outlier represents:

Aberrant observations not representative of the population,

Observations of small or insignificant segments within the population and of no interest to the research objectives.

They should be retained if an undersampling/poor representation of relevant groups in the population; the sample should be augmented to ensure representation of these groups.

Outliers can be identified based on the similarity measure by:

Finding observations with large distances from all other observations,

Graphic profile diagrams highlighting outlying cases,

Their appearance in cluster solutions as single-member or small clusters.

Clustering variables that have scales using widely differing numbers of scale points or that exhibit large differences in standard deviations should be standardized:

The most common standardization conversion is Z scores.

If groups are to be identified according to an individual's response style, then within-case or row-centering standardization is appropriate.

### STAGE 3: ASSUMPTIONS IN CLUSTER ANALYSIS

Cluster analysis is not a statistical inference technique in which parameters from a sample are assessed as representing a population. Instead, cluster analysis is a method for quantifying the structural characteristics of a set of observations. As such, it has strong mathematical properties but not statistical foundations. The requirements of normality, linearity, and homoscedasticity that were so important in other techniques really have little bearing on cluster analysis. The researcher must focus, however, on three other critical issues: assumption of existing structure; representativeness of the sample and multicollinearity among variables in the cluster variate.

**Structure Exists** A fundamental assumption of all interdependence techniques is that a “natural” structure of objects exists which is to be identified by the technique. Exploratory factor analysis assumed that multicollinearity among variables demonstrated some interpretable relationship among those variables. Likewise, cluster analysis assumes there is some natural grouping of objects in the sample to be analyzed. This is important as the saying “Cluster analysis will generate clusters” is true, whether the resulting cluster solution is an interpretable structure or just a partitioning of the data, even if randomly dispersed, into a specified number of clusters. This assumption highlights the need for validating the cluster solution in terms of the research objectives (see Step 6 for a more detailed discussion of validation).

**Representativeness of the Sample** Rarely does the researcher have a census of the population to use in the cluster analysis. Usually, a sample of cases is obtained and the clusters are derived in the hope that they represent the structure of the population. The researcher must therefore be confident that the obtained sample is truly representative of the population. As mentioned earlier, outliers may really be only an undersampling of divergent groups that, when discarded, introduce bias in the estimation of structure. The researcher must realize that cluster analysis is only as good as the representativeness of the sample. Therefore, all efforts should be made to ensure that the sample is representative and the results are generalizable to the population of interest.

**Impact of Multicollinearity** Multicollinearity is an issue in other multivariate techniques because of the difficulty in discerning the true impact of multicollinear variables. In cluster analysis the effect is different because multicollinearity is actually a form of implicit weighting. Let us start with an example that illustrates the effect of multicollinearity.

Suppose that respondents are being clustered on 10 variables, all attitudinal statements concerning a service. When multicollinearity is examined, we see two sets of variables, the first made up of eight statements and the second consisting of the remaining two statements. If our intent is to really cluster the respondents on the dimensions of the service (in this case represented by the two groups of variables), then using the original 10 variables will be quite misleading. Because each variable is weighted equally in cluster analysis, the first dimension will have four times as many chances (eight items compared to two items) to affect the similarity measure. As a result, similarity will be predominantly affected by the first dimension with eight items rather than the second dimension with two items.

Multicollinearity acts as a weighting process not apparent to the observer but affecting the analysis nonetheless. For this reason, the researcher is encouraged to examine the variables used in cluster analysis for substantial multicollinearity and, if found, either reduce the variables to equal numbers in each set or use a distance measure that takes multicollinearity into account. Another possible solution involves using exploratory factor analysis prior to clustering and either selecting one cluster variable from each factor or using the resulting factor scores as cluster variables. Recall that principal components or varimax rotated factors are uncorrelated. In this way, the research can take a proactive approach to dealing with multicollinearity.

One last issue is whether to use factor scores in cluster analysis. The debate centers on research showing that the variables that truly discriminate among the underlying groups are not well represented in most factor solutions. Thus, when factor scores are used, it is quite possible that a poor representation of the actual structure of the data will be obtained [47]. The researcher must deal with both multicollinearity and discriminability of the variables to arrive at the best representation of structure.

## Assumptions in Cluster Analysis

An implicit assumption in applying cluster analysis is that there are natural groupings in the data. Remember that cluster analysis will identify cluster solutions even with randomly dispersed data, so it is essential that the analyst understand the nature of the data.

Input variables should be examined for substantial multicollinearity and if present:

Reduce the variables to equal numbers in each set of correlated measures, or

Use a distance measure that compensates for the correlation, such as Mahalanobis distance.

Take a proactive approach and include only cluster variables that are not highly correlated.

## STAGE 4: DERIVING CLUSTERS AND ASSESSING OVERALL FIT

With the clustering variables selected and the similarity matrix calculated, the partitioning process begins (see Figure 4.6). The researcher must:

- Select the partitioning procedure used for forming clusters.
- Potentially respecify initial cluster solutions by eliminating outliers or small clusters.
- Make the decision on the number of clusters in the final cluster solution.

These three decisions have substantial implications not only on the results that will be obtained but also on the interpretation that can be derived from the results [34]. First, we examine the available partitioning procedures and select the approach best suited for both the data and the research purpose. As we form our initial cluster solutions, examine the results to identify and potentially eliminate any outliers or otherwise irrelevant clusters. Once we have a suitable set of cluster solutions, we then decide on a final cluster solution(s) by defining the number of clusters and membership for each observation.

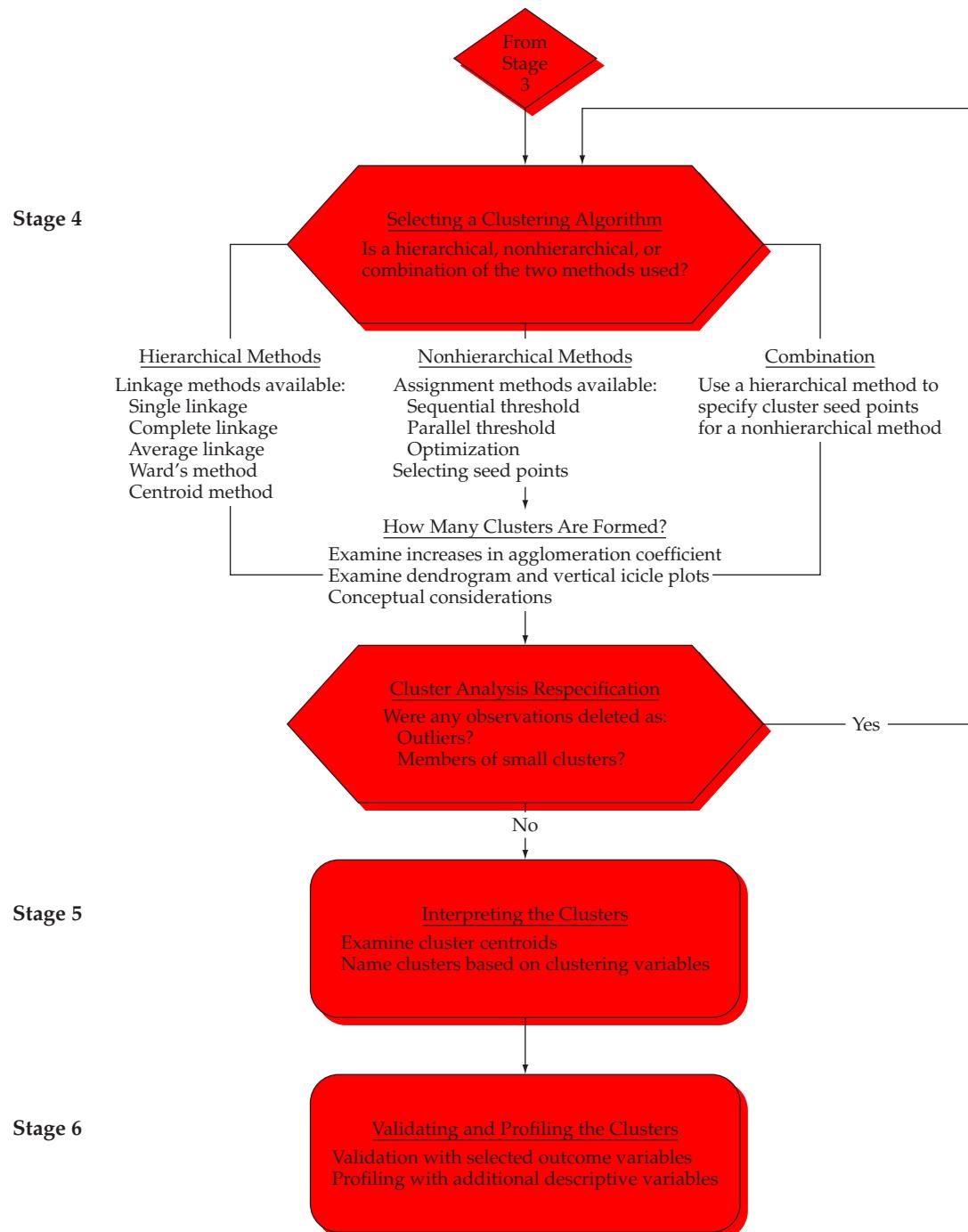
**Partitioning Procedures** All partitioning procedures work on a simple principle. They seek to group objects into clusters which seek to maximize the between-group variation (i.e., distance between groups) while minimizing the within-group variation (i.e., differences of in-group members) as shown in Figure 4.7. The two most common and illustrative partitioning procedures are hierarchical and nonhierarchical procedures. We encountered a **hierarchical** procedure in our earlier example of clustering the seven objects where all objects were initially in separate clusters and then sequentially joined two-clusters at a time until only a single cluster remained. A **nonhierarchical** procedure is quite different in that the number of clusters is specified by the analyst and then the set of objects are formed into that set of groupings. So while we see a specific number of clusters, we must perform a separate analysis for each potential cluster solution.

In the discussion that follows we will first examine the hierarchical and nonhierarchical procedures and then compare them as to their strengths and weaknesses. We will then examine briefly two emerging alternative procedures: density-based procedures which have a fundamentally different way of representing similarity and model-based procedures which employ a statistically-based method of nonhierarchical modeling to identify groupings of objects to form clusters.

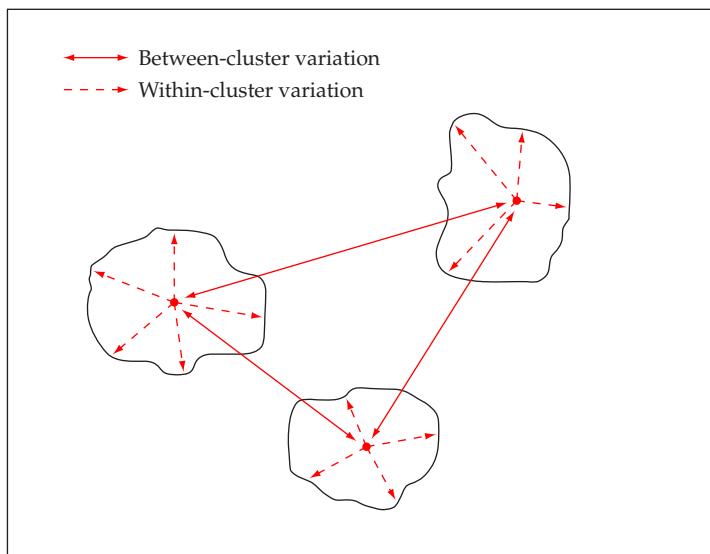
**Hierarchical Cluster Procedures** Hierarchical procedures involve a series of  $n - 1$  clustering decisions (where  $n$  equals the number of observations) that combine observations into a hierarchy or a tree-like structure. The two basic types of hierarchical clustering procedures are agglomerative and divisive. In the **agglomerative methods**, each object or observation starts out as its own cluster and is successively joined, the two most similar clusters at a time until

**Figure 4.6**

Stages 4–6 of the Cluster Analysis Decision Diagram



only a single cluster remains. In **divisive methods** all observations start in a single cluster and are successively divided (first into two clusters, then three, and so forth) until each is a single-member cluster. In Figure 4.8, agglomerative methods move from left to right, and divisive methods move from right to left. Because most commonly used computer packages use agglomerative methods, and divisive methods act almost as agglomerative methods in reverse, we focus here on the agglomerative methods.



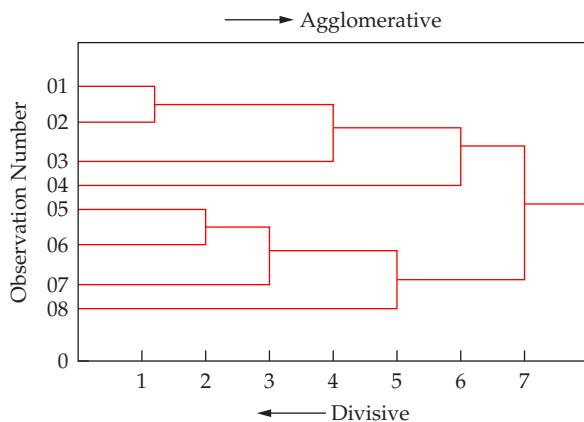
**Figure 4.7**  
Cluster Diagram Showing  
Between- and Within-Cluster  
Variation

To understand how a hierarchical procedure works, we will examine the most common form—the agglomerative method—which follows a simple, repetitive process:

- 1 Start with all observations as their own cluster (i.e., each observation forms a single-member cluster), so that the number of clusters equals the number of observations.
- 2 Using the similarity measure, combine the two most similar clusters (termed a clustering algorithm) into a new cluster (now containing two observations), thus reducing the number of clusters by one.
- 3 Repeat the clustering process again, using the similarity measure and clustering algorithm to combine the two most similar clusters into a new cluster.
- 4 Continue this process, at each step combining the two most similar clusters into a new cluster. Repeat the process a total of  $n - 1$  times until all observations are contained in a single cluster.

Assume that we had 100 observations. We would initially start with 100 separate clusters, each containing a single observation. At the first step, the two most similar clusters would be combined, leaving us with 99 clusters. At the next step, we combine the next two most similar clusters, so that we then have 98 clusters. This process continues until the last step where the final two remaining clusters are combined into a single cluster.

An important characteristic of hierarchical procedures is that the results at an earlier stage are always nested within the results at a later stage, creating a similarity to a tree. For example, an agglomerative six-cluster solution is



**Figure 4.8**  
Dendrogram Illustrating  
Hierarchical Clustering

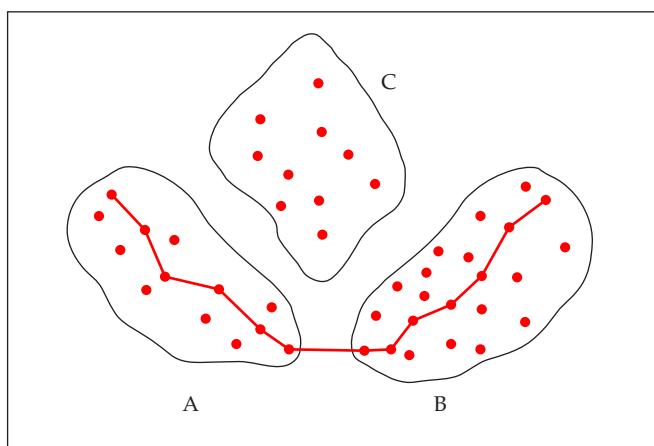
obtained by joining two of the clusters found at the seven-cluster stage. Because clusters are formed only by joining existing clusters, any member of a cluster can trace its membership in an unbroken path to its beginning as a single observation. This process is shown in Figure 4.8; the representation is referred to as a **dendrogram** or tree graph, which can be useful, but becomes unwieldy with large applications. The dendrogram is widely available in most clustering software.

**CLUSTERING ALGORITHMS** The **clustering algorithm** in a hierarchical procedure defines how similarity is specified between multiple-member clusters in the clustering process. When joining two single-member clusters, their similarity is simply the similarity between the single object in each cluster. But how do we measure similarity between clusters when one or both clusters have multiple members? Do we select one member to act as a “typical” member and measure similarity between these members of each cluster, or do we create some composite member to represent the cluster, or even combine the similarities between all members of each cluster? We could employ any of these approaches, or even devise other ways to measure similarity between multiple-member clusters. Among numerous approaches, the five most popular agglomerative algorithms are (1) single-linkage, (2) complete-linkage, (3) average linkage, (4) centroid method, and (5) Ward’s method. In our discussions we will use distance as the similarity measure between observations, but other similarity measures could be used just as easily.

**Single-Linkage** The **single-linkage method** (also called the **nearest-neighbor method**) defines the similarity between clusters as the shortest distance from any object in one cluster to any object in the other. This rule was applied in the example at the beginning of this chapter and enables us to use the original distance matrix between observations without calculating new distance measures. Just find all the distances between observations in the two clusters and select the smallest as the measure of cluster similarity.

This method is probably the most versatile agglomerative algorithm, because it can define a wide range of clustering patterns (e.g., it can represent clusters that are concentric circles, like rings of a bull’s-eye). This flexibility also creates problems, however, when clusters are poorly delineated. In such cases, single-linkage procedures can form long, snake-like chains [34, 43]. Individuals at opposite ends of a chain may be dissimilar, yet still within the same cluster. Many times, the presence of such chains may contrast with the objectives of deriving the most compact clusters. Thus, the researcher must carefully examine the patterns of observations within the clusters to ascertain whether these chains are occurring. It becomes increasingly difficult by simple graphical means as the number of clustering variables increases, and requires that the researcher carefully profile the internal homogeneity among the observations in each cluster.

An example of this arrangement is shown in Figure 4.9. Three clusters (A, B, and C) are to be joined. The single-linkage algorithm, focusing on only the closest points in each cluster, would link clusters A and B because of their short distance at the extreme ends of the clusters. Joining clusters A and B creates a cluster that encircles cluster C. Yet in striving for within-cluster homogeneity, it would be much better to join cluster C with either A or B. This illustrates the principal disadvantage of the single-linkage algorithm.



**Figure 4.9**  
Example of Single Linkage  
Joining Dissimilar Clusters  
A and B

**Complete-Linkage** The **complete-linkage method** (also known as the **farthest-neighbor** or **diameter method**) is comparable to the single-linkage algorithm, except that cluster similarity is based on maximum distance between observations in each cluster. Similarity between clusters is the smallest (minimum diameter) sphere that can enclose all observations in both clusters. This method is called complete-linkage because all objects in a cluster are linked to each other at some maximum distance. Thus, within-group similarity equals group diameter.

This technique eliminates the chaining problem identified with single-linkage and has been found to generate the most compact clustering solutions [5]. Even though it represents only one aspect of the data (i.e., the farthest distance between members), many researchers find it the appropriate for a wide range of clustering applications [31].

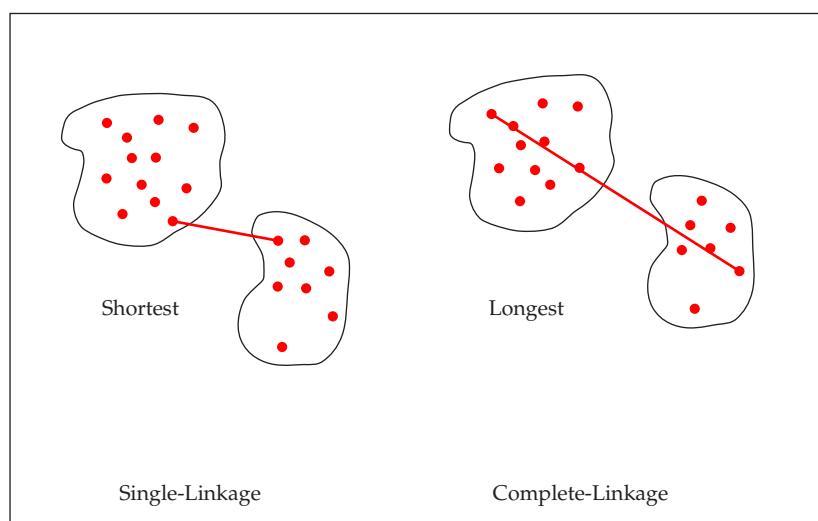
Figure 4.10 compares the shortest (single-linkage) and longest (complete-linkage) distances in representing similarity between clusters. Yet both measures reflect only one aspect of the data. The use of the single-linkage reflects only a closest single pair of objects, and the complete-linkage also reflects a single pair, this time the two most extreme.

**Average Linkage** The **average linkage** procedure differs from the single-linkage or complete-linkage procedures in that the similarity of any two clusters is the average similarity of all individuals in one cluster with all individuals in another. This algorithm does not depend just on extreme values (closest or farthest pairs) as do single-linkage or complete-linkage. Instead, similarity is based on all members of the clusters rather than on a single pair of extreme members and is thus less affected by outliers. Average linkage approaches, as a type of compromise between single- and complete-linkage methods, tend to generate clusters with small within-cluster variation. They also tend toward the production of clusters with approximately equal within-group variance.

**Centroid Method** In the **centroid method** the similarity between two clusters is the distance between the cluster centroids. **Cluster centroids** are the mean values of the observations on the variables in the cluster variate. In this method, every time individuals are grouped, a new centroid is computed. Cluster centroids migrate as cluster mergers take place. In other words, a cluster centroid changes every time a new individual or group of individuals is added to an existing cluster.

These methods are the most popular in the physical and life sciences (e.g., biology) but may produce messy and often confusing results. The confusion occurs because of reversals, that is, instances when the distance between the centroids of one pair may be less than the distance between the centroids of another pair merged at an earlier combination. The advantage of this method, like the average linkage method, is that it is less affected by outliers than are other hierarchical methods.

**Ward's Method** Ward's method differs from the previous methods in that the similarity between two clusters is not a single measure of similarity, but rather the sum of squares within the clusters summed over all variables. It is quite similar to the simple heterogeneity measure used in the example at the beginning of the chapter to assist in



**Figure 4.10**  
Comparison of Distance  
Measures for Single-Linkage  
and Complete-Linkage

determining the number of clusters. In the Ward's procedure, the selection of which two clusters to combine is based on which combination of clusters minimizes the within-cluster sum of squares across the complete set of disjoint or separate clusters. At each step, the two clusters combined are those that minimize the increase in the total sum of squares across all variables in all clusters.

This procedure tends to combine clusters with a small number of observations, because the sum of squares is directly related to the number of observations involved. The use of a sum of squares measure makes this method easily distorted by outliers [40]. Moreover, the Ward's method also tends to produce clusters with approximately the same number of observations. If the researcher expects or desires the clustering patterns to reflect somewhat equally sized clusters, then this method is quite appropriate. However, the use of this method also makes it more difficult to identify clusters representing small proportions of the sample.

**OVERVIEW** Hierarchical clustering procedures are a combination of a repetitive clustering process combined with a clustering algorithm to define the similarity between clusters with multiple members. The process of creating clusters generates a tree-like diagram that represents the combinations/divisions of clusters to form the complete range of cluster solutions. Hierarchical procedures generate a complete set of cluster solutions, ranging from all single-member clusters to the one-cluster solution where all observations are in a single cluster. In doing so, the hierarchical procedure provides an excellent framework with which to compare any set of cluster solutions and help in judging how many clusters should be retained.

**Nonhierarchical Clustering Procedures** In contrast to hierarchical methods, **nonhierarchical procedures** do not involve the tree-like construction process. Instead, they assign objects into clusters once the number of clusters is specified. For example, a six-cluster solution is not just a combination of two clusters from the seven-cluster solution, but is based only on finding the best six-cluster solution. The nonhierarchical cluster software programs usually proceed through two steps:

- 1 Specify cluster seeds.** The first task is to identify starting points, known as **cluster seeds**, for each cluster. A cluster seed may be prespecified by the researcher or observations may be selected, usually in a random process.
- 2 Assignment.** With the cluster seeds defined, the next step is to assign each observation to one of the cluster seeds based on similarity. Many approaches are available for making this assignment (see later discussion in this section), but the basic objective is to assign each observation to the most similar cluster seed. In some approaches, as observations are added to existing clusters and the cluster composition changes, objects may be reassigned/switched to other clusters that are more similar than their original cluster assignment.

We discuss several different approaches for selecting cluster seeds and assigning objects in the next sections.

**SELECTING CLUSTER SEEDS** Even though the nonhierarchical clustering algorithms discussed in the next section differ in the manner in which they assign observations to the clusters, they all face the same initial problem: How do we select the cluster seeds? The different approaches can be classified into two basic categories:

**Researcher Specified** In this approach, the researcher provides the seed points based on external data. The two most common sources of the seed points are prior research or data from another multivariate analysis. Many times the researcher has knowledge of the cluster profiles being researched. For example, prior research may have defined segment profiles and the task of the cluster analysis is to assign individuals to the most appropriate segment cluster. It is also possible that other multivariate techniques may be used to generate the seed points. One common example is the use of a hierarchical clustering algorithm to establish the number of clusters and then generate seed points from these results (a more detailed description of this approach is contained in the following section). The common element is that the researcher, while knowing the number of clusters to be formed, also has information about the basic character of these clusters.

**Sample Generated** The second approach is to generate the cluster seeds from the observations of the sample, either in some systematic fashion or simply through random selection. For example, in the FASTCLUS program in SAS,

the default approach is to specify the first seed as the first observation in the data set with no missing values. The second seed is the next complete observation (no missing data) that is separated from the first seed by a specified minimum distance. The default option is a zero minimum distance to ensure that they are not exactly identical. After all seeds are selected, the program assigns each observation to the cluster with the nearest seed. The researcher can specify that the cluster seeds be revised (updated) by calculating seed cluster means each time an observation is assigned. In contrast, the K-means program in IBM SPSS can select the necessary seed points randomly from among the observations. In any sample-generated method the researcher relies on the selection process to choose seed points that reflect natural clusters as starting points for the clustering algorithms. A potentially serious limitation is that replication of the results is difficult if the initial cluster seeds vary across analyses due to (a) differing random objects are selected as cluster seeds in each analysis or (b) the observations are reordered for each analysis. Varying the initial cluster seeds is a form of validation we will discuss in Step 6, but it is an issue impacting replication.

Whether researcher-specified or sample generated, the researcher must be aware of the impact of the cluster seed selection process on the final results. All of the clustering algorithms, even those of an optimizing nature (see the following discussion), will generate different cluster solutions depending on the initial cluster seeds. The differences among cluster solutions will hopefully be minimal using different seed points, but they underscore the importance of cluster seed selection and its impact on the final cluster solution.

**NONHIERARCHICAL CLUSTERING ALGORITHMS** The **k-means algorithm** [27, 29] is by far the most popular nonhierarchical clustering algorithm in use today. The k-means algorithm works by partitioning the data into a user-specified number of clusters and then iteratively reassigning observations to clusters until some numerical criterion is met. The criterion specifies a goal related to minimizing the distance of observations from one another in a cluster and maximizing the distance between clusters. The name comes from defining each of  $k$  clusters with its centroid (i.e., the mean of all objects in the cluster). The k-means algorithm only works with metric data.

The key element in any nonhierarchical algorithm is the assignment and potential reassignment of objects to clusters. Three basic options exist for this process: sequential, parallel, and optimization [24].

**Sequential Threshold Method** Starts by selecting one cluster seed and includes all objects within a prespecified distance. A second cluster seed is then selected, and all objects within the prespecified distance of that seed are included. A third seed is then selected, and the process continues as before. The primary disadvantage of this approach is that when an observation is assigned to a cluster, it cannot be reassigned to another cluster, even if that cluster seed is more similar.

**Parallel Threshold Method** The parallel threshold method considers all cluster seeds simultaneously and assigns observations within the threshold distance to the nearest seed.

**Optimizing Method** The third method, referred to as the **optimizing procedure**, is similar to the other two nonhierarchical procedures except that it allows for reassignment of observations to a seed other than the one with which it was originally associated. If, in the course of assigning observations, an observation becomes closer to another cluster that is not the cluster to which it is currently assigned, then an optimizing procedure switches the observation to the more similar (closer) cluster.

K-means is so commonly used that the term is many times used to refer to nonhierarchical cluster analysis in general. For example, in IBM SPSS, the nonhierarchical clustering routine is referred to as K-means. Its widespread use comes from its ability to analyze a large number of objects as each object can be processed sequentially where in hierarchical approaches a complete similarity matrix between all objects is needed. As such, it is the favored method for processing large datasets in many different forms. As will be discussed later, it is even used in conjunction with hierarchical methods to “pre-process” a large dataset and create a fairly large number of clusters, for example 100, that then can be analyzed by a hierarchical method.

**Should Hierarchical or Nonhierarchical Methods be Used?** A definitive answer to this question cannot be given for two reasons. First, the research problem at hand may suggest one method or the other. Second, what we learn with continued application to a particular context may suggest one method over the other as more suitable for that context. These reasons are borne out in the widespread applications of both approaches in today's analytical landscape, whether it be in academia or the practitioner domain. We will examine the strengths and weaknesses of each method to provide the analyst with a framework for making their own decisions on the most applicable approach in their specific research setting.

**PROS AND CONS OF HIERARCHICAL METHODS** Hierarchical clustering techniques have long been the more popular clustering method, with Ward's method and average linkage probably being the best available [40]. Besides the fact that hierarchical procedures were the first clustering methods developed, they do offer several advantages that lead to their widespread usage:

- 1 *Simplicity:* Hierarchical techniques, with their development of the tree-like structures depicting the clustering process, do afford the researcher with a simple, yet comprehensive, portrayal of the entire range of clustering solutions. In doing so, the researcher can evaluate any of the possible clustering solutions from one analysis.
- 2 *Measures of similarity:* The widespread use of the hierarchical methods led to an extensive development of similarity measures for almost any type of clustering variables, either metric or nonmetric. As a result, hierarchical techniques can be applied to almost any type of research question.
- 3 *Speed:* Hierarchical methods have the advantage of generating an entire set of clustering solutions (from all separate clusters to one cluster) in an expedient manner. This enables the researcher to examine a wide range of alternative clustering solutions, varying measures of similarities and linkage methods, in an efficient manner.

Even though hierarchical techniques have been widely used, they do have several distinct disadvantages that affect any of their cluster solutions:

- 1 *Permanent combinations:* Hierarchical methods can be misleading because undesirable early combinations of clusters may persist throughout the analysis and lead to artificial results. This "never split" characteristic allows for the creation of the tree-like structure and dendrogram, also magnifies the substantial impact of outliers on hierarchical methods, particularly with the complete-linkage and Ward's methods. It also results in the possible chain-like clusters when using the single-linkage method.
- 2 *Impact of outliers:* To reduce the impact of outliers, the researcher may find it useful to cluster analyze the data several times, each time deleting problematic observations or outliers. The deletion of cases, however, even those not found to be outliers, can many times distort the solution. Moreover, the deletion of outliers in any cluster solution requires a potentially subjective decision by the analyst. Thus, the researcher must employ extreme care in the deletion of observations for any reason.
- 3 *Large samples:* Although computations of the clustering process are relatively fast, hierarchical methods are not amenable to analyzing large samples. As sample size increases, the data storage requirements increase dramatically. For example, a sample of 400 cases requires storage of approximately 80,000 similarities, which increases to almost 125,000 for a sample of 500. Even with today's computing power, these data requirements can limit the application in many instances. The researcher may take a random sample of the original observations to reduce sample size but must now question the representativeness of the sample taken from the original sample.

**EMERGENCE OF NONHIERARCHICAL METHODS** Nonhierarchical methods have gained increased acceptability and usage, but any application depends on the ability of the researcher to select the seed points according to some practical, objective, or theoretical basis. In these instances, nonhierarchical methods offer several advantages over hierarchical techniques.

- 1 The results are less susceptible to:
  - a. outliers in the data,
  - b. the distance measure used, and
  - c. the inclusion of irrelevant or inappropriate variables.

- 2** Nonhierarchical methods can analyze extremely large data sets because they do not require the calculation of similarity matrices among all observations, but instead just the similarity of each observation to the cluster centroids. Even the optimizing algorithms that allow for reassignment of observations between clusters can be readily applied to all sizes of datasets.

Although nonhierarchical methods do have several distinct advantages, several shortcomings can markedly affect their use in many types of applications.

- 1** The benefits of any nonhierarchical method are realized only with the use of nonrandom (i.e., specified) seed points. Thus, the use of nonhierarchical techniques with random seed points is generally considered inferior to hierarchical techniques.
- 2** Even a nonrandom starting solution does not guarantee an optimal clustering of observations. In fact, in many instances the researcher will get a different final solution for each set of specified seed points. How is the researcher to select the optimum answer? Only by analysis and validation can the researcher select what is considered the best representation of structure, realizing that many alternatives may be as acceptable.
- 3** The tendency of the k-means cluster analysis to only produce clusters which are spherical in shape and equally sized [14].
- 4** Nonhierarchical methods are also not as efficient when examining a large number of potential cluster solutions. Each cluster solution is a separate analysis, in contrast to the hierarchical techniques that generate all possible cluster solutions in a single analysis. Thus, nonhierarchical techniques are not as well suited to exploring a wide range of solutions based on varying elements such as similarity measures, observations included, and potential seed points.

**A COMBINATION OF BOTH METHODS** Many researchers recommend a combination approach using both methods. In this way, the advantages of one approach can compensate for the weaknesses of the other [40]. This can be accomplished in two steps:

- 1** First, a hierarchical technique is used to generate a complete set of cluster solutions, establish the applicable cluster solutions (see next section for discussion of this topic), and establish the appropriate number of clusters.
- 2** After outliers are eliminated, the remaining observations can then be clustered by a nonhierarchical method using the number of clusters determined by the hierarchical approach and even determining seed points from the cluster centroids of the hierarchical solution.

In this way, the advantages of the hierarchical methods are complemented by the ability of the nonhierarchical methods to refine the results by allowing the switching of cluster membership. We demonstrate this approach in our empirical example later in the chapter.

A different combination of hierarchical and nonhierarchical methods has been implemented in both SAS (CLUSTER in Enterprise Miner) and IBM SPSS (Two-Stage) with particular emphasis on analyzing large datasets. In SAS, for example, a three step process (nonhierarchical → hierarchical → nonhierarchical) is followed:

- 1** A large number of cluster seeds (default of 50) are chosen and then a nonhierarchical algorithm is applied to obtain a preliminary set of clusters.
- 2** The preliminary set of clusters are then analyzed with a hierarchical algorithm utilizing Ward's method, generating cluster solutions from  $N$  (number of preliminary clusters from Step 1) to a single cluster. The number of clusters is determined using the CCC criterion (to be discussed in the next section).
- 3** The selected hierarchical cluster solution from Step 2 is used to generate cluster centroids which act as cluster seeds. These cluster seeds are then used in a nonhierarchical analysis of the original sample to generate the final cluster solution.

The IBM SPSS approach is similar in nature, using a sequential process for the first step to generate a much smaller number of preliminary clusters which can then be analyzed with a hierarchical algorithm. In the SAS and IBM SPSS

## Selecting a Clustering Approach

Selection of hierarchical or nonhierarchical methods is based on:

Hierarchical clustering solutions are preferred when:

A wide range of alternative clustering solutions is to be examined.

The sample size is moderate (under 300–400, not exceeding 1000) or a sample of the larger data set is acceptable.

Nonhierarchical clustering methods are preferred when:

The number of clusters is known and/or initial seed points can be specified according to some practical, objective, or theoretical basis.

Outliers cause concern, because nonhierarchical methods generally are less susceptible to outliers.

A combination approach using a hierarchical approach followed by a nonhierarchical approach is often advisable:

A hierarchical approach is used to select the number of clusters and profile cluster centers that serve as initial cluster seeds in the nonhierarchical procedure.

A nonhierarchical method then clusters all observations using the seed points to provide more accurate cluster memberships.

methods both methods are used to highlight their advantages—hierarchical methods to generate and evaluate a large range of cluster solutions and nonhierarchical methods to analyze even very large samples and generate an “optimal” clustering given a specified number of clusters.

**Should the Cluster Analysis be Respecified?** Even before identifying an acceptable cluster analysis solution (see next section), the researcher should examine the fundamental structure represented in the potential cluster solutions. Of particular note are widely disparate cluster sizes or clusters of only one or two observations. Generally, one-member or extremely small clusters are not acceptable given the research objectives, and thus should be eliminated. In a hierarchical solution outliers may be detected by their addition in the agglomeration schedule at a later stage, along with small or single member clusters in either hierarchical or nonhierarchical cluster solutions.

Researchers must examine widely varying cluster sizes from a conceptual perspective, comparing the actual results with the expectations formed in the research objectives. From a practical perspective, more troublesome are single-member clusters, which may be outliers not detected in earlier analyses. If a single-member cluster (or one of small size compared with other clusters) appears, the researcher must decide whether it represents a valid structural component in the sample or should be deleted as unrepresentative. If any observations are deleted, especially when hierarchical solutions are employed, the researcher should rerun the cluster analysis and start the process of defining clusters anew. This may result in several iterations of analysis and respecification to identify all of the objects to be eliminated.

**How Many Clusters Should be Formed?** Perhaps the most critical issue for any researcher performing either a hierarchical or nonhierarchical cluster analysis is determining the number of clusters most representative of the sample’s data structure [15]. This decision is critical for hierarchical techniques because even though the process generates the complete set of cluster solutions, the researcher must select the cluster solution(s) to represent the data structure (also known as the **stopping rule**). Hierarchical cluster results do not always provide unambiguous information as to the best number of clusters. Therefore, researchers commonly use a stopping rule that suggests two or more cluster solutions which can be compared before making the final decision. And with nonhierarchical cluster procedures the analyst must provide the number of clusters or they are determined from prior hierarchical analyses.

As discussed earlier in our simple clustering example, any stopping rule is based on a foundational principle: a natural increase in heterogeneity comes from the reduction in number of clusters. **Heterogeneity** refers to how different the observations in a cluster are from each other (i.e., heterogeneity refers to a lack of similarity among group members). As a result, all stopping rules share in some form a common element in evaluating the trend in heterogeneity across cluster solutions to identify marked increases. We then look for substantive increases in this trend which indicates to relatively distinct clusters were joined and that the cluster structure before joining is a potential candidate for the final solution. The stopping rules must be developed to accommodate the natural increase in heterogeneity or if not, in most instances the two-cluster solution would always be chosen because the value of any stopping rule is normally highest when going from two to one cluster.

Unfortunately, no singular objective stopping rule exists [10, 28]. Because no internal statistical criterion is used for inference, such as the statistical significance tests of other multivariate methods, researchers have developed many criteria for approaching the problem. In doing so, two issues have emerged:

- These ad hoc procedures must be computed by the researcher and often involve fairly complex approaches [3, 41]. Many of the procedures are ad hoc, involve fairly complex procedures and have developed in very specific areas of study [3, 41, 2].
- Many of these criteria are specific to a particular software program and are not easily calculated if not provided by the program. So even the set of stopping rules we review in this section are not all contained in a single software package. As such, the researcher must understand the basic principles of stopping rules and how they are represented in whatever analysis situation they are facing.

With their focus on the varying levels of heterogeneity found in differing cluster solutions, stopping rules can be placed into one of two general classes, as described next.

**MEASURES OF HETEROGENEITY CHANGE** One class of stopping rules examines some measure of heterogeneity change between cluster solutions at each successive decrease in the number of clusters. A cluster solution is identified as a candidate for the final cluster solution when the heterogeneity change measure makes a sudden jump. A simple example was used at the beginning of the chapter, which looked for large increases in the average within-cluster distance. When a large increase occurs, the researcher selects the prior cluster solution on the logic that its combination caused a substantial increase in heterogeneity. This type of stopping rule has been shown to provide fairly accurate decisions in empirical studies [41], but it is not uncommon for a number of cluster solutions to be identified by these large increases in heterogeneity. It is then the researcher's task to select a final cluster solution from these selected cluster solutions. Most of the measures of heterogeneity change have been developed within the hierarchical procedures, where the sequential set of cluster solutions makes these measures easily calculated. They can be applied for nonhierarchical methods as well, but are more complicated since each cluster solution is a separate analysis and thus requires aggregating this information across analyses. Stopping rules using this approach range from a simple measure of percentage change to more complex measures of statistical significance of heterogeneity change.

**Percentage Changes in Heterogeneity** Probably the simplest and most widespread rule is a simple percentage change in some measure of heterogeneity. A typical example is using the agglomeration coefficient from hierarchical methods in SPSS, which measures heterogeneity as the distance at which clusters are formed (if a distance measure of similarity is used) or the within-cluster sum of squares if the Ward's method is used. With this measure, the percentage increase in the agglomeration coefficient can be calculated for each cluster solution. Then the researcher selects cluster solutions as a potential final solution when the percentage increase is markedly larger than occurring at other steps.

**Measures of Variance Change** The **root mean square standard deviation (RMSSTD)** is the square root of the variance of a cluster or set of clusters. As such, it is used in several different ways. For example, in the hierarchical procedure in SAS (PROC CLUSTER), the RMSSTD value is for the newly formed cluster. When representing the heterogeneity of two joined clusters, large values suggest the joining of two quite dissimilar clusters, indicating the previous cluster solution (in which the two clusters were separate) was a candidate for selection as the final cluster solution. In other software packages the RMSSTD may be for individual clusters or for the set of clusters overall. Here, larger values

of the RMSSTD of a cluster or cluster solution indicate a set of clusters with greater heterogeneity. In all instances, however, larger values represent more heterogeneous cluster(s).

**Statistical Measures of Heterogeneity Change** A series of test statistics attempts to portray the degree of heterogeneity for each new cluster solution (i.e., separation between the two clusters most recently joined). One of the most widely used is a pseudo  $T^2$  statistic, which compares the goodness-of-fit of  $k$  clusters to  $k - 1$  clusters. Highly significant values indicate that the  $k$  cluster solution is more appropriate than the  $k - 1$  cluster solution (i.e., the  $k - 1$  cluster solution was the result of joining two very different clusters from the  $k$  solution). For example, if we observe a very high pseudo  $T^2$  statistic at the six-cluster stage, then a candidate solution would be the seven-cluster solution, since six clusters were achieved by joining two very different clusters from the seven-cluster solution (see Figure 4.11). The researcher should not consider any significant value, but instead look to those values markedly more significant than for other solutions.

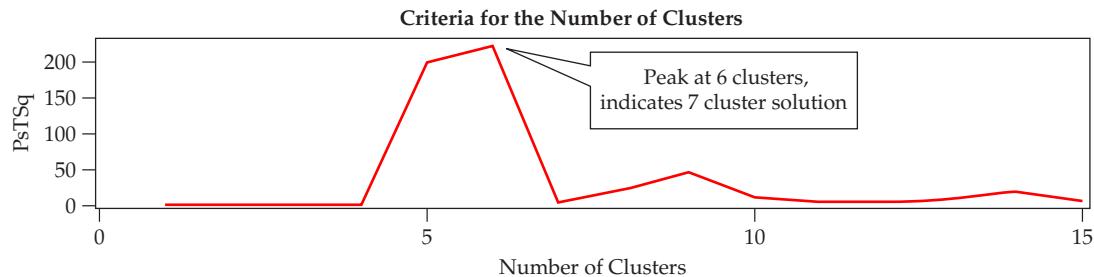
**DIRECT MEASURES OF HETEROGENEITY** A second general class of stopping rules attempts to directly measure heterogeneity of each cluster solution and then allow analyst to evaluate each cluster solution against a criterion measure. There are a large number of these types of measures developed across the differing domains, but at least three general types emerge.

**Comparative Cluster Heterogeneity** The most common measure in this class is the **cubic clustering criterion (CCC)** [41] contained in SAS, a measure of the deviation of the clusters from an expected distribution of points formed by a multivariate uniform distribution. Here the researcher selects the cluster solution with the largest value of CCC (i.e., the cluster solution where CCC peaks as shown in Figure 4.12) [49]. Despite its inclusion in SAS and its advantage of selecting a single-cluster solution, it has been shown to many times generate too many clusters as the final solution [41] and is based on the assumption that the variables are uncorrelated. However, it is a widely used measure and is generally as efficient as any other stopping rule [41].

**Statistical Significance of Cluster Variation** The pseudo  $F$  statistic measures the separation among all the clusters at the current level by the ratio of between-cluster variance (separation of clusters) to within-cluster variance (homogeneity of clusters). Thus, a higher pseudo  $F$  indicates a cluster solution maximizing between-cluster differences while minimizing within-cluster similarity (see Figure 4.12). As with other stopping rules there is no specific value which dictates selection of a cluster solution, but instead the analyst identifies cluster solutions by high relative values of the pseudo  $F$  statistic.

**Internal Validation Index** The final set of measures are many times referred to as internal validity measures as they quantify the characteristics of the cluster solution along two general dimensions: separation and compactness. One such measure is the Dunn index [16] which is the ratio between the minimal within-cluster distance to maximal between-cluster distance. As with other measures, the objective is to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart, as compared to the within cluster variance. For a cluster solution, a higher Dunn index indicates better clustering. Similar measures are the DB index [13] and the silhouette index [48].

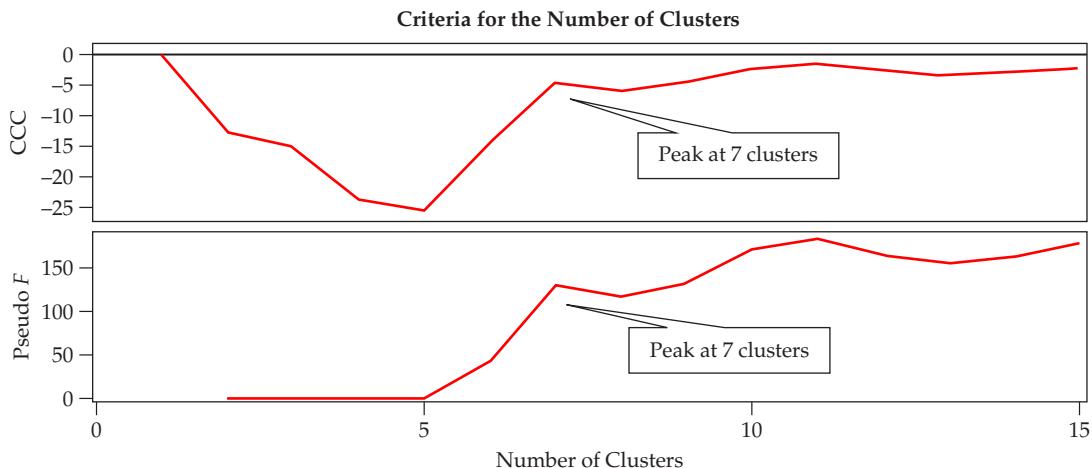
**Figure 4.11**  
Measures of Change in Heterogeneity: Pseudo  $T^2$



Note: Adapted from SAS output

**Figure 4.12**

Direct Measures of Heterogeneity – CCC and Pseudo F



Note: Adapted from SAS output

**SUMMARY** Given the number of stopping rules available and the lack of evidence supporting any single stopping rule, it is suggested that a researcher employ several stopping rules and look for a consensus cluster solution(s). Even with a consensus based on empirical measures, however, the researcher should complement the empirical judgment with any conceptualization of theoretical relationships that may suggest a natural number of clusters. One might start this process by specifying some criteria based on practical considerations, such as saying, “My findings will be more manageable and easier to communicate if I use three to six clusters,” and then solving for this number of clusters and selecting the best alternative after evaluating all of them. In the final analysis, however, it is probably best to compute a number of different cluster solutions (e.g., two, three, four) and then decide among the alternative solutions by using a priori criteria, practical judgment, common sense, or theoretical foundations. The cluster solutions will be improved by restricting the solution according to conceptual aspects of the problem.

## Respecifying the Cluster Solution

Before attempting to identify the final cluster solution(s), the results, whether they be for hierarchical or nonhierarchical methods, should be examined for single-member or extremely small clusters which are of insufficient size for inclusion in the final solution.

These inappropriate clusters should be eliminated from the sample and the analysis run again until all of the inappropriate clusters are eliminated.

## Deriving the Final Cluster Solution

No single objective procedure is available to determine the correct number of clusters; rather the researcher can evaluate alternative cluster solutions on two general types of stopping rules:

Measures of heterogeneity change:

These measures, whether they be percentage changes in heterogeneity, measures of variance change (RMSSTD) or statistical measure of change (pseudo  $T^2$ ), all evaluate the change in heterogeneity when moving from  $k$  to  $k - 1$  clusters.

Candidates for a final cluster solution are those cluster solutions which preceded a large increase in heterogeneity by joining two clusters (i.e., a large change in heterogeneity going from  $k$  to  $k - 1$  clusters would indicate that the  $k$  cluster solution is better).

**Direct measures of heterogeneity:**

These measures directly reflect the compactness and separation of a specific cluster solution. These measures are compared across a range of cluster solutions, with the cluster solution(s) exhibiting more compactness and separation being preferred.

Among the most prevalent measures are the CCC (cubic clustering criterion), a statistical measure of cluster variation (pseudo F statistic) or the internal validation index (Dunn's index).

Cluster solutions ultimately must have theoretical validity assessed through external validation.

**Other Clustering Approaches** While hierarchical and nonhierarchical approaches are the predominant clustering approaches used across both academia and practitioner domains, there have emerged at least two other approaches that have gained use in specific research situations. These two approaches differ in how to represent similarity, whether it be density (density-based approach) or probabilistically (model-based approach). We will provide a brief introduction to each approach in enough detail to enable analysts to determine if either of these approaches is suited to their research situation.

**DENSITY-BASED APPROACH** The **density-based approach** is based on the simple notion that clusters can be identified by “dense” clusters of objects within the sample, separated by regions of lower object density [19]. This corresponds to the “natural” groupings we infer when we look at plots of objects – quickly picking out the higher density areas of objects. The primary advantages of this approach are that it can identify clusters of arbitrary shape (e.g., not just spherical shapes), can easily vary in the number of objects in each cluster and incorporates the identification of outliers (i.e., those objects in very sparse regions) as a fundamental part of the process.

In defining clusters as maximal sets of density connected points, the analyst must make two fundamental decisions:

- $\epsilon$ , the radius around a point that defines a point’s **neighborhood**, and
- the minimum number of objects (minObj) necessary within a neighborhood to define it a cluster.

Notice that the analyst does not have to specify the number of clusters or other basic characteristics of the desired cluster solution. The process, using just these two parameters, analyzes the data and arrives at a configuration of clusters. In determining clusters, all objects are classified into one of three types:

- **Core points**—objects which have a neighborhood which meets the minimum number of objects requirement. These points are central to the definition of any cluster.
- **Boundary or density-reachable points**—these objects don’t meet the minObj requirement for their neighborhood, but are in the neighborhood of a core point. These points can’t be core points, but become a part of cluster around a core point.
- **Outliers or noise points**—any object which is not a core point or a boundary point.

The process for defining clusters is quite similar to nonhierarchical clustering in the assignment of objects, although the number of clusters is not specified:

- 1 Select an unclassified point and compute its neighborhood to see if it is a core point. If yes, start a cluster around this point. If not, move on to another point.
- 2 Once a core point is found, define it as a cluster and identify all boundary points for that core point. Remember that boundary points must be within the neighborhood of the core point. Then, expand the cluster by adding these boundary points to the cluster. Now the neighborhood of the cluster expands to include those added boundary points, which allows for even more boundary points to be added to the cluster. We can see that this is quite similar to the linkage methods in hierarchical methods or the assignment process in nonhierarchical.

- 3** The process continues until all points within the expanded cluster neighborhoods are assigned to that cluster and those not assigned are now classified as outliers. Outliers are points that are neither core points nor are they within the neighborhood of any cluster member. At this point the analyst must decide on the role of outliers and if retained or not.

This procedure, which was popularized by DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [19] is available in most software programs. While it does not generate a range of cluster solutions nor provide the insight into structure of a hierarchical routine, it does have several distinct advantages:

- Ability to identify clusters of any arbitrary shape,
- Ability to process very large samples,
- Requires specification of only two parameters,
- No prior knowledge of number of clusters,
- Explicit designation of outliers as separate from objects assigned to clusters,
- Applicable to a “mixed” set of clustering variables (i.e., both metric and nonmetric).

The density-based approach has been widely used in many application areas, particularly those dealing with large samples and a large and diverse number of clustering variables. It should be noted that while our focus is many times on social science research, this method has wide application in areas such as the physical sciences (astronomy, biology, earth science and geography). As familiarity spreads as to the impact of specifying the parameters of  $\epsilon$  and minObj and variants of the procedure evolve for specific research situations, we expect to see an even wider usage of this approach.

**MODEL-BASED APPROACH** The **model-based approach** is similar in its basic concept to the other clustering methods, but uses differing probability distributions of objects as the basis for forming groups rather than groupings of similarity in distance or high density [22, 23, 7, 39]. It varies from all other clustering approaches in that it is a statistical model whereas the other approaches are varying types of algorithms. The underlying statistical premise is “similarity is probability” versus ‘similarity is distance’. Yet in most other regards the model-based approach is similar to other clustering approaches, perhaps most akin to a statistical version of nonhierarchical models with many extensions.

So what is the source of this probability? The basic model is what is termed a **mixture model** where objects are assumed to be represented by a mixture of probability distributions (known as **components**), each representing a different cluster. These different groups of objects are based on the similar patterns on the clustering variables for each group. Of particular note is that the clustering variables can be a metric, nonmetric or both. Objects which are highly similar to other similar objects thus have a higher probability of being in the same unobserved cluster/component versus other possible clusters formed by other sets of objects. Each subsample of similar objects, ultimately representing a cluster, is modeled separately and the overall population is modeled as a mixture or weighted sum of these subsamples [22, 39].

Forming clusters occurs in two distinct steps:

- 1** Having specified the number of distributions (i.e., desired clusters) to be estimated, estimate the parameters of the assumed probability distributions (i.e., subsamples of objects with differing means and covariances on the clustering variables) through application of the EM algorithm.
- 2** Using the estimated distributions, calculate the posterior probabilities of cluster membership for each object in the sample in each distribution. Cluster membership for each object is through assignment to the cluster/probability distribution with the highest probability.

Given the statistical nature of this approach, overall measures of fit are possible and models with differing numbers of components/clusters can be compared through the Bayesian Information Criterion (BIC) or other measures. This allows the analyst to assess the impact on overall fit due to varying the number of underlying distributions (clusters), with the best fit indicating the appropriate clustering solution and thus the number of clusters/components in a model.

There are several advantages associated with the model-based approach when compared to other clustering approaches:

- Can be applied to any combination of clustering variables (metric and/or nonmetric)
- Statistical tests are available to compare different models and determine best model fit to define best cluster solution
- Missing data can be directly handled
- No scaling issues or transformations of variables needed
- Very flexible with options for simple and complicated distributional forms
- Once the cluster solution is finalized, the model can be expanded to include both antecedent/predictor variables of cluster membership along with outcome/validation variables.

While these advantages do address issues with many of the other clustering approaches, the biggest difference of the model-based approach is its fundamental approach to data analysis, i.e., a statistical model versus an algorithmic model as discussed in Chapter 1. Many analysts may be more or less inclined towards this approach based on their background and training, but it does provide a totally different approach to the clustering process from the other methods discussed in this chapter.

It should be noted that this approach has many names across varying domains, including mixture model clustering [39, 38], model-based clustering [22, 23, 7, 9] and latent class cluster analysis [36]. But all of these methods share the same fundamental approach of employing a statistical model to estimate probability distributions within the sample and then assign objects to clusters based on these probabilities.

**Overview** The multiple approaches to clustering and the options/issues with each approach characterize both the promise and the peril of cluster analysis for the analyst. The ability to identify homogeneous subsamples within the overall sample has many applications and uses, but the analyst is first faced with a myriad set of decisions on which basic approach to select – hierarchical versus nonhierarchical or even some combination of these, not to mention density-based and model-based approaches. Once this choice is made, there are additional decisions within each approach that substantially impact what clusters are ultimately found.

But most challenging is not the complexity of these approaches, but ultimately the relevancy of the results. Many times analysts are dissuaded by the differing solutions obtained due to the approaches used or even just decisions within each approach. And yet the analyst must understand that these differences are not reflective of some deficiencies in one method versus another, but our fundamental understanding of the basic elements in each clustering approach and the impact of specific decisions (e.g., distance measure selected for similarity, linkage method used in hierarchical models, number of clusters in nonhierarchical models or the radius for neighborhoods in density models). The analyst must come to accept that there is no definitive correct cluster solution, but instead a range of results that must be evaluated to best achieve the research objectives.

## STAGE 5: INTERPRETATION OF THE CLUSTERS

The interpretation stage involves examining each cluster in terms of the cluster variate to name or assign a label accurately describing the nature of the clusters. When starting the interpretation process, one measure frequently used is the cluster's centroid. If the clustering procedure was performed on the raw data, it would be a logical description. If the data were standardized or if the cluster analysis was performed using exploratory factor analysis results (component factors), the researcher would have to go back to the raw scores for the original variables. As the number of clustering variables or the number of cluster increases, viewing tabular values for the centroids becomes increasingly difficult. It is in these instances that graphical approaches become useful, particularly a parallel coordinates graph as a profile diagram.

To clarify this process, let us refer to the example of diet versus regular soft drinks. Let us assume that an attitude scale was developed that consisted of statements regarding consumption of soft drinks, such as "diet soft drinks taste harsher," "regular soft drinks have a fuller taste," "diet drinks are healthier," and so forth. Further, let us assume that

demographic and soft-drink consumption data were also collected. We can first examine the average score profiles on the attitude statements for each group and assign a descriptive label to each cluster. Many times discriminant analysis is applied to generate score profiles, but we must remember that statistically significant differences would not indicate an optimal solution because statistical differences are expected, given the objective of cluster analysis. Examination of the profiles allows for a rich description of each cluster. For example, two of the clusters may have favorable attitudes about diet soft drinks and the third cluster negative attitudes. Moreover, of the two favorable clusters, one may exhibit favorable attitudes toward only diet soft drinks, whereas the other may display favorable attitudes toward both diet and regular soft drinks. From this analytical procedure, one would evaluate each cluster's attitudes and develop substantive interpretations to facilitate labeling each. For example, one cluster might be labeled "health- and calorie-conscious," whereas another might be labeled "get a sugar rush."

The profiling and interpretation of the clusters, however, achieve more than just description and are essential elements in selecting between cluster solutions when the stopping rules indicate more than one appropriate cluster solution.

- They provide a means for assessing the correspondence of the derived clusters to those proposed by prior theory or practical experience. If used in a confirmatory mode, the cluster analysis profiles provide a direct means of assessing the correspondence.
- The cluster profiles also provide a route for making assessments of practical significance. The researcher may require that substantial differences exist on a set of clustering variables and the cluster solution be expanded until such differences arise.

In assessing either correspondence or practical significance, the researcher compares the derived clusters to a preconceived typology. This more subjective judgment by the researcher combines with the empirical judgment of the stopping rules to determine the final cluster solution to represent the data structure of the sample.

## STAGE 6: VALIDATION AND PROFILING OF THE CLUSTERS

Given the somewhat subjective nature of cluster analysis about selecting an optimal cluster solution, the researcher should take great care in validating and ensuring practical significance of the final cluster solution. Although no single method exists to ensure validity and practical significance, several approaches have been proposed to provide some basis for the researcher's assessment.

**Validating the Cluster Solution** Validation includes attempts by the researcher to assure that the cluster solution is representative of the general population, and thus is generalizable to other objects and is stable over time.

**CROSS-VALIDATION** The most direct approach in this regard is to cluster analyze separate samples, comparing the cluster solutions and assessing the correspondence of the results [11, 12]. This approach, however, is often impractical because of time or cost constraints or the unavailability of objects (particularly consumers) for multiple cluster analyses. In these instances, a common approach is to split the sample into two groups. Each cluster is analyzed separately, and the results are then compared. Cross-tabulation also can be used for a single sample, because the members of any specific cluster in one solution should stay together in a cluster in another solution. Therefore, the cross-tabulation should display obvious patterns of matching cluster membership. Other approaches include (1) a modified form of split sampling whereby cluster centers obtained from one cluster solution are employed to define clusters from the other observations and the results are compared [37], and (2) a direct form of cross-validation [45].

For any of these methods, stability of the cluster results can be assessed by the number of cases assigned to the same cluster across cluster solutions. Generally, a very stable solution would be produced with less than 10 percent of observations being assigned to a different cluster. A stable solution would result with between 10 and 20 percent assigned to a different group, and a somewhat stable solution when between 20 and 25 percent of the observations are to a different cluster than the initial one.

**ESTABLISHING CRITERION VALIDITY** The researcher may also attempt to establish some form of criterion or predictive validity. To do so, the researcher selects variable(s) not used to form the clusters but known to vary across the clusters. In our example, we may know from past research that attitudes toward diet soft drinks vary by age. Thus, we can

statistically test for the differences in age between those clusters that are favorable to diet soft drinks and those that are not. The variable(s) used to assess predictive validity should have strong theoretical or practical support because they become the benchmark for selecting among the cluster solutions.

**Profiling the Cluster Solution** The profiling stage involves describing the characteristics of each cluster on variables that are not among the clustering variables or validation variables. The variables used at this stage are typically of two types: (1) descriptive variables such as demographic variables, psychographic profiles, consumption patterns or other behavioral measures which help identify the clusters within the general population or (2) predictive variables which are hypothesized as reasons leading to the clustering of objects in clusters. While variables in the descriptive category may be selected just based on practical relevance, attempts to understand the impact of variables on the formation of clusters require theoretical support as well. In both instances, discriminant analysis or other techniques are used to identify which variables do differ across the clusters. While we expected all the clustering variables to vary across clusters since that was the objective of cluster analysis, with these variables we are looking for key distinguishing variables from either the descriptive or predictive variables. Using discriminant analysis, the researcher compares average score profiles for the clusters. The categorical dependent variable is the previously identified clusters, and the independent variables are the demographics, psychographics, and so on.

From this analysis, assuming statistical significance, the researcher could conclude, for example, that the “health-and calorie-conscious” cluster from our previous diet soft drink example consists of better-educated, higher-income professionals who are moderate consumers of soft drinks.

For graphical or simple comparisons between clusters, the researcher may choose to **center** each variable by subtracting the overall mean for that variable from each observation. The result is a set of variables with a mean of zero but retaining their unique variability. This step simply facilitates interpretation when the variables do not have the same means, but is not used in the actual clustering process. Analysts should be cautioned, however, to remember that when using centered variables the point of reference is the individual (i.e., how much a single variable differs from the object’s average response) versus the differences between groups.

In short, the profile analysis focuses on describing not what directly determines the clusters but rather on the characteristics of the clusters after they are identified. Moreover, the emphasis is on the characteristics that differ

## Interpreting, Profiling, and Validating Clusters

All clusters should be significantly different across the set of clustering variables.

The cluster centroid, a mean profile of the cluster on each clustering variable, is particularly useful in the interpretation stage:

Interpretation involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters.

Cluster solutions failing to show substantial variation indicate other cluster solutions should be examined.

The cluster centroid should also be assessed for correspondence with the researcher's prior expectations based on theory or practical experience.

Validation is essential in cluster analysis because the clusters are descriptive of structure and require additional support for their relevance:

Cross-validation empirically validates a cluster solution by creating two subsamples (randomly splitting the sample) and then comparing the two cluster solutions for consistency with respect to number of clusters and the cluster profiles.

Validation is also achieved by examining differences on variables not included in the cluster analysis but for which a theoretical and relevant reason enables the expectation of variation across the clusters.

significantly across the clusters and those that could predict membership in a particular cluster. Profiling often is an important practical step in clustering procedures, because identifying characteristics like demographics enables segments to be identified or located with easily obtained information.

## Implication of Big Data Analytics

As discussed in Chapter 1, Big Data analytics are becoming increasingly important to academic and practitioners alike. And there is perhaps no multivariate technique which has more potential in the domain of Big Data analytics, yet also faces the most challenges to its application, than cluster analysis. In terms of contributions, Big Data is perhaps best known for exactly that—more data observations for analysis than ever before. In this domain the analyst can quickly see the advantages of simplification—reducing the large number of observations into a much smaller number of groupings from which the general nature and character of the entire dataset can be observed. In many ways cluster analysis is comparable to dimensional reduction discussed in Chapter 3, differing only in the clustering of diverse objects (e.g., customers, text, genes, etc.) rather than attributes of those objects.

We have long seen the advantages of such simplification, perhaps no more so that in the areas of consumer analytics, driven primarily by segmentation schemes. But today this advantages is extended to a diverse set of other areas, such as information retrieval, pattern recognition, textual analysis, spatial observations, web applications or even DNA analysis. In each area cluster analysis is playing a critical and influential role [excellent overviews in 2, 26].

### CHALLENGES

Two specific challenges discussed previously are exacerbated in the application of cluster analysis to larger datasets. The first arises from the case level of analysis of methods such as hierarchical procedures where the size of the sample has a direct bearing on the scale of the problem. These methods require a  $N$  by  $N$  matrix of similarities, where  $N$  is the sample size. So as the sample size increases to the sizes found today in Big Data applications, only certain clustering methods can be used or we need to employ two-stage procedures to handle the large sample size.

The second challenge emerges when clustering high-dimensional data (i.e., a large number of object attributes). As discussed in Chapter 1 another characteristic of Big Data is not only access to large samples, but also the explosion in the number and types of attributes available for each object. With this comes two distinct issues: *variable relevancy* and *determining object similarity*. As discussed in Step 2, these issues have a substantial impact on the effectiveness of cluster analysis. Up to this point most research has been based on a reasonable number of clustering variables with substantial conceptual support. But the analysis of Big Data necessarily takes on more of an exploratory perspective where both of these issues can become quite problematic. While notable strides have been made to address these issues [35, 30], the analyst must always be aware of their impact and thus the type of cluster solution that emerges.

We mention these complications not only because they are more frequently encountered in many research settings today, but more importantly because they emphasize the importance of researcher judgment in cluster analysis. As seen in prior discussions, there are many options for each decision, resulting in an almost incalculable number of combinations. It is incumbent on the analyst to understand the implications of each separate decision and their combination. So as “automatic” clustering approaches emerge which make most of the critical clustering decisions, including even selecting the number of clusters, we urge caution before employing such methods so as to truly understand the underlying choices being made in all phases of the clustering process.

## An Illustrative Example

We will use the HBAT database to illustrate the application of cluster analysis techniques. Customer perceptions of HBAT provide a basis for illustrating one of the most common applications of cluster analysis—the formation of customer segments. In our example, we follow the stages of the model-building process, starting with setting

objectives, then addressing research design issues, and finally partitioning respondents into clusters and interpreting and validating the results. The following sections detail these procedures through each of the stages.

## STAGE 1: OBJECTIVES OF THE CLUSTER ANALYSIS

The first stage in applying cluster analysis involves determining the objectives to be achieved. Once the objectives have been agreed upon, the HBAT research team must select the clustering variables to be used in the clustering process.

**Clustering Objectives** Cluster analysis can achieve any combination of three objectives: taxonomy development, data simplification, and relationship identification. In this situation, HBAT is primarily interested in the segmentation of customers (taxonomy development), although additional uses of the derived segments are possible.

The primary objective is to develop a taxonomy that segments objects (HBAT customers) into groups with similar perceptions. Once identified, strategies with different appeals can be formulated for the separate groups—the requisite basis for market segmentation. Cluster analysis, with its objective of forming homogeneous groups that are as distinct between one another as possible, provides a unique methodology for developing taxonomies with maximal managerial relevance.

In addition to forming a taxonomy that can be used for segmentation, cluster analysis also facilitates data simplification and even identification of relationships. In terms of data simplification, segmentation enables categorization of HBAT customers into segments that define the basic character of group members. In an effective segmentation, customers are viewed not as only individuals, but also as members of relatively homogeneous groups portrayed through their common profiles. Segments also provide an avenue to examine relationships previously not studied. A typical example is the estimation of the impact of customer perceptions on sales for each segment, enabling the researcher to understand what uniquely impacts each segment rather than the sample as a whole. For example, do customers with more favorable perceptions of HBAT also purchase more?

**Clustering Variables** The research team has decided to look for clusters based on the variables that indicate how HBAT customers rate the firm's performance on several key attributes. These attributes are measured in the database with variables  $X_6$  to  $X_{18}$ . The research team knows that multicollinearity can be an issue in using cluster analysis. Thus, having already analyzed the data using exploratory factor analysis, they decide to use variables that are not strongly correlated to one another. To do so, they select a single variable to represent each of the four factors (an example of surrogate variable approach, see Chapter 3) plus  $X_{15}$ , which was eliminated from the exploratory factor analysis by the MSA test because it did not share enough variance with the other variables. The variables included in the cluster analysis are ratings of HBAT on various firm attributes:

- $X_6$  product quality (representative of the factor *Product Value*)
- $X_8$  technical support (representative of the factor *Technical Support*)
- $X_{12}$  salesforce image (representative of the factor *Marketing*)
- $X_{15}$  new product development (not included in extracted factors)
- $X_{18}$  delivery speed (representative of the factor *Post-sale Customer Service*)

Table 4.5 displays the descriptive statistics for these variables.

**Table 4.5** Descriptive Statistics for Cluster Variables

	N	Minimum	Maximum	Mean	Std. Deviation
$X_6$ , Product Quality	100	5	10	7.81	1.40
$X_8$ , Technical Support	100	1.3	8.5	5.37	1.53
$X_{12}$ , Salesforce Image	100	2.9	8.2	5.12	1.07
$X_{15}$ , New Products	100	1.7	9.5	5.15	1.49
$X_{18}$ , Delivery Speed	100	1.6	5.5	3.89	0.73

## STAGE 2: RESEARCH DESIGN OF THE CLUSTER ANALYSIS

In preparing for a cluster analysis, the researcher must address four issues in research design: detecting outliers, determining the similarity measure to be used, deciding the sample size, and standardizing the variable and/or objects. Each of these issues plays an essential role in defining the nature and character of the resulting cluster solutions.

**Detecting Outliers** The first issue is to identify any outliers in the sample before partitioning begins. Even though the univariate procedures discussed in Chapter 2 did not identify any potential candidates for designation as outliers, multivariate procedures are used because objects must also be evaluated on the pattern of values across the entire set of clustering variables used in defining similarity. Outliers are observations that are different or dissimilar from all other observations in the sample. In our example, we will refer to the term *dissimilarity*, because larger distances mean less similar observations. We will focus specifically on finding observations that are potentially quite different than the others.

We begin looking for outliers by determining the average dissimilarity for each observation based upon the distance measures for each observation. There is no single best way to identify the most dissimilar objects. One measure available in most statistical software is the a matrix of pairwise proximity measures showing, for example, the Euclidean distance from each observation to every other observation. In IBM SPSS, you can do this from the correlation dropdown menu or as an option under the hierarchical clustering routine. Observations with relatively large pair-wise proximities (large differences between) become outlier candidates. This method has several disadvantages. First, large pair-wise proximity values may just represent observations on the “opposite” sides of the sample and not outliers in regard to other observations. Second, determining which observations have the largest average distances can prove difficult when the number of observations exceeds 20 or so. In the HBAT example, the 100 observations would produce a proximity, or, more precisely, a dissimilarity matrix that is 100 rows by 100 columns. Thus, a smaller summary analysis of how different each observation is from an average respondent makes this process much easier.

A second approach is to use a measure of proximity to the centroid of the sample (i.e., a point representing the average of all of the clustering variables). Recall from earlier in the chapter that the Euclidean distance measure is easily generalized to more than two dimensions and we can use this approach to find the observations that have relatively high dissimilarity. A typical respondent can be thought of as one that responds with the central tendency on each variable (assuming the data follow a conventional distribution). In other words, an average respondent provides responses that match the mean of any variable. Thus, the average distance of each observation from the typical respondent (centroid) can be used as a measure of dissimilarity to the sample as a whole.

One such measure is the Mahalanobis  $D^2$ , which we discussed in Chapter 2 and is also seen in Chapter 5 (Multiple Regression) as a measure to identify outliers. But the Mahalanobis  $D^2$  is many times only available in conjunction with a specific technique (e.g., as an outlier measure in multiple regression). As a result we can easily calculate a variant measure which provides the same basic measure of dissimilarity for each observation from the sample centroid through the following steps:

- Select an observation and compute the difference between the observed value for a clustering variable and the variable mean. The process is repeated for each clustering variable.
- The differences are squared, similar to the calculation of Euclidean distance (see Figure 4.5) and as is typically done in computing measures of variation to avoid the problem of having negative and positive differences cancel each other out.
- The squared differences for each variable are summed to get a total for each observation across all clustering variables. This represents the squared distance from the typical or average respondent.
- Finally, the square root of that sum is taken to create an estimate of the dissimilarity of this observation based on distance from the typical respondent profile.
- The process is repeated for each observation. Observations with the highest dissimilarities have the potential to be outliers.

Researchers looking for outliers do not focus on the absolute value of dissimilarity. Rather, researchers are simply looking for any values that are relatively large compared to the others. Thus, sorting the observations from the most to the least dissimilar can be convenient. A starting point is to select the observations with the highest five or ten percent of the dissimilarity values for more thorough examination.

Table 4.6 illustrates the process for observation 100. We begin by taking each the observation's score on each cluster variable and subtracting the mean for that variable from the observation. In this case, the observation value was 7.9 on  $X_6$ —Product Quality—and we would subtract the variable mean (7.81 for  $X_6$ ) to yield a value of 0.09. The process is repeated for each cluster variable. In this example, the clustering variables are  $X_6$ ,  $X_8$ ,  $X_{12}$ ,  $X_{15}$ , and  $X_{18}$ . Each row in Table 4.7 shows how the difference is taken for each individual variable. These differences are then squared and a sum total calculated (1.63) and a dissimilarity value (the square root of the sum total) of 1.28.

Most standard statistical packages do not include a function that produces this table; program syntax provides an option, but a spreadsheet can be used to do these simple computations and then sort the observations by dissimilarity. The book's websites provide an example spreadsheet that performs this analysis.

Table 4.7 lists the 10 observations (ten percent of the sample) with the highest average dissimilarities. Two observations—87 and 6—display relatively large values. Each has a dissimilarity of over 5 (5.58 and 5.30, respectively). In contrast, the next largest dissimilarity is for observation 90 (4.62), which is only slightly larger than the fourth highest proximity of 4.57 for observation 53. These two observations—87 and 6—stand out over the others as having relatively high average distances. At this point, we will not eliminate any observations but will continue to watch for other potential signs that observations 87 and/or 6 may truly be outliers. Generally, researchers are more comfortable deleting observations as outliers when multiple pieces of evidence are present.

**Defining Similarity** The next issue involves the choice of a similarity measure to be used as input to the hierarchical clustering algorithm. The researcher does not have to actually perform these computations separately, but rather only needs to specify which approach will be used by the cluster program. Correlational measures are not used, because when segments are identified with cluster analysis we should consider the magnitude of the perceptions (favorable versus unfavorable) as well as the pattern. Correlational measures only consider the patterns of the responses, not the absolute values. Cluster analysis objectives are therefore best accomplished with a distance measure of similarity.

Given that all five clustering variables are metric, squared Euclidean distance is chosen as the similarity measure. Either squared Euclidean distance or Euclidean distance is typically the default similarity measure in statistical packages. Multicollinearity has been addressed by selecting variables that are not highly correlated with each other based on the previous exploratory factor analysis (see results in Chapter 3).

**Sample Size** The third issue relates to the adequacy of the sample of 100 observations. This issue is not a statistical (inferential) issue. Instead, it relates to the ability of the sample to identify managerially useful segments. That is, segments with a large enough sample size to be meaningful. In our example, the HBAT research team believes

**Table 4.6 Example of Centering When Calculating Dissimilarity for Observation 100**

	Observed Values for Observation 100	Less	Variable Means (see Table 4.5)	Difference	Squared Difference
$X_6$	7.9	—	7.81	= 0.09	0.0081
$X_8$	4.4	—	5.37	= -0.97	0.9409
$X_{12}$	4.8	—	5.12	= -0.32	0.1024
$X_{15}$	5.8	—	5.15	= 0.65	0.4225
$X_{18}$	3.5	—	3.89	= -0.39	0.1521
Total Differences Squared					1.63
Square Root of Total					1.28

**Table 4.7** Largest Dissimilarity Values for Identifying Potential Outliers

Observation	Differences from Mean for Each Observation:					Squared Differences from Mean					
	$X_6$	$X_8$	$X_{12}$	$X_{15}$	$X_{18}$	$X_6$	$X_8$	$X_{12}$	$X_{15}$	$X_{18}$	Dissimilarity
87	-2.81	-4.07	-0.22	2.45	-0.79	7.90	16.52	0.05	6.00	0.62	5.58
6	-1.31	-2.27	-1.42	4.35	-0.59	1.72	5.13	2.02	18.92	0.34	5.30
90	-2.31	2.34	3.08	-0.25	1.01	5.34	5.45	9.47	0.06	1.03	4.62
53	1.59	-0.57	-0.52	4.05	0.71	2.53	0.32	0.27	16.40	0.51	4.48
44	-2.71	1.23	2.68	0.05	0.61	7.34	1.53	7.17	0.00	0.38	4.05
41	0.49	-2.07	0.08	-3.45	0.01	0.24	4.26	0.01	11.90	0.00	4.05
72	-1.11	-2.37	-0.62	-2.65	-0.79	1.23	5.59	0.39	7.02	0.62	3.85
31	-0.91	3.14	-0.42	-1.85	-0.59	0.83	9.83	0.18	3.42	0.34	3.82
22	1.79	1.43	2.68	1.35	0.41	3.20	2.06	7.17	1.82	0.17	3.80
88	-0.11	2.64	-0.82	2.55	0.41	0.01	6.94	0.68	6.50	0.17	3.78

segments that represent at least 10 percent of the total sample size will be meaningful. Smaller segments are considered too small to justify development of segment-specific marketing programs. Thus, in our example with a sample of 100 observations, we consider segments consisting of 10 or more observations as meaningful. But initial clusters obtained through hierarchical clustering with as few as five observations may be retained because the cluster size will likely change when observations are reassigned in nonhierarchical clustering. This also corresponds to an upper limit of 10 clusters as the maximum number of clusters that the research team deems useful for HBAT and the desired number is between three and seven clusters.

**Standardization** The final issue involves the type of standardization that may be used. It is not useful to apply within-case standardization because the magnitude of the perceptions is an important element of the segmentation objectives. But the issue of standardizing by variable still remains.

All of the clustering variables are measured on the same scale (0 to 10), so there is no need to standardize because of differences in the scale of the variables. There are, however, other considerations, such as marked differences in the standard deviation of the variables. For example, as shown in Table 4.4, the variables generally have similar amounts of dispersion, with the possible exception of  $X_{18}$  (a small standard deviation of 0.73 compared to all others above one). These differences could affect the clustering results, and standardization would eliminate that possibility. But with only one variable exhibiting a difference, we choose not to standardize in the HBAT example.

Another consideration is the means of the variables used in the cluster analysis. For example, the means of the variables in the HBAT example vary to some degree, ranging from less than four for  $X_{18}$  (3.89) to nearly eight for  $X_6$  (7.81). This does not suggest standardization, but it may make it more difficult to interpret each cluster's meaning. In some situations, the mean-centered values will facilitate interpretation of clusters. Using mean centering does not affect the cluster results, but it often makes it easier to compare the mean values on each variable for each cluster.

Table 4.5 displays the means for the variables across all 100 observations. Mean-centered values for each variable can be obtained by subtracting the mean from each observation. The HBAT researcher performs this task by using the software's compute function and entering the following instruction for each variable:

$$X_{6C} = X_6 - \text{Mean}(X_6)$$

Here  $X_{6C}$  is the name given by the researcher to represent the mean centered values for  $X_6$ ,  $X_6$  is the variable itself, and  $\text{Mean}(X_6)$  represents the mean value for that variable (7.81 in this case). This process is repeated for each cluster variable. Mean-centered variables retain the same information as the raw variables because the standard deviations are the same for each clustering variable and its mean-centered counterpart. The mean for each mean-centered variable

is zero, however, meaning that each mean-centered variable has this common reference point. This is simply a way of recoding the variables to have a common mean. The common mean may make it easier to interpret the cluster profiles.

### STAGE 3: ASSUMPTIONS IN CLUSTER ANALYSIS

In meeting the assumptions of cluster analysis, the researcher is not interested in the statistical qualities of the data (e.g., normality, linearity, etc.) but instead is focused primarily on issues of research design. The two basic issues to be addressed are sample representativeness and multicollinearity among the clustering variables.

**Sample Representativeness** A key requirement for using cluster analysis to meet any of the objectives discussed in stage 1 is that the sample is representative of the population of interest. Whether developing a taxonomy, looking for relationships, or simplifying data, cluster analysis results are not generalizable from the sample unless representativeness is established. The researcher must not overlook this key question, because cluster analysis has no way to determine if the research design ensures a representative sample.

The sample of 100 HBAT customers was obtained through a random selection process from among the entire customer base. All issues concerned with data collection were addressed adequately to ensure that the sample is representative of the HBAT customer base. Thus, we can extend the sample findings to the population of HBAT customers.

**Multicollinearity** If there is multicollinearity among the clustering variables, the concern is that the set of clustering variables is assumed to be independent, but may actually be correlated. This may become problematic if several variables in the set of cluster variables are highly correlated and others are relatively uncorrelated. In such a situation, the correlated variables influence the cluster solution much more so than the several uncorrelated variables.

As indicated earlier, the HBAT research team minimized any effects of multicollinearity through the variable selection process. That is, they chose cluster variables based on the findings of the previous exploratory factor analysis, with  $X_6$ ,  $X_8$ ,  $X_{12}$ , and  $X_{18}$  representing the four factors found in Chapter 3 and  $X_{15}$  a variable not included among the variables in the factors.

### STAGES 4–6: EMPLOYING HIERARCHICAL AND NONHIERARCHICAL METHODS

In applying cluster analysis to the sample of 100 HBAT customers, the research team decided to use both hierarchical and nonhierarchical methods in combination. To do so, they followed a two-part process:

**Part 1 Partitioning:** A hierarchical procedure was used to identify a preliminary set of cluster solutions as a basis for determining the appropriate number of clusters.

**Part 2 Fine Tuning:** Use of a nonhierarchical procedures to “fine-tune” the results and then profile and validate the final cluster solution. The hierarchical and nonhierarchical procedures from IBM SPSS and SAS are used in this analysis, and comparable results would be obtained with most other clustering programs.

### PART 1: HIERARCHICAL CLUSTER ANALYSIS (STAGE 4)

In this step, we utilize the hierarchical clustering procedure’s advantage of quickly examining a wide range of cluster solutions to identify a set of preliminary cluster solutions. This range of solutions is then analyzed by nonhierarchical clustering procedures to determine the final cluster solution. Our emphasis in the hierarchical analysis, therefore, is on Stage 4 (the actual clustering process). The profiling and validation stages (Stages 5 and 6) are then undertaken in step 2—the nonhierarchical process. In the course of performing the hierarchical cluster analysis, the researcher must perform a series of tasks:

**Step 1:** Select the clustering algorithm.

**Step 2:** Generate the cluster results, check for single member or other inappropriate clusters and respecify cluster analysis as needed.

**Step 3:** Select the preliminary cluster solution(s) by applying the stopping rule(s).

**Step 4:** Profile the clustering variables to identify the most appropriate cluster solutions.

In doing so, the researcher must address methodological issues as well as consider managerial and clustering objectives to derive the most representative cluster solution for the sample. In the following sections, we will discuss both types of issues as we address the tasks listed above.

**Step 1: Selecting a Clustering Algorithm** Before actually applying the cluster analysis procedure, we must first ask the following question: Which clustering algorithm should we use? Combined with the similarity measure chosen (squared Euclidean distance), the clustering algorithm provides the means of representing the similarity between clusters with multiple members. Ward's method was selected because of its tendency to generate clusters that are homogeneous and relatively equal in size. Hopefully this will also minimize the emergence of small clusters or outliers.

**Step 2: Initial Cluster Results** With the similarity measure and clustering algorithm defined, the HBAT research team can now apply the hierarchical clustering procedure. The results must be reviewed across the range of cluster solutions to enable us to identify any clusters that may need to be deleted due to small size or other reasons (outliers, unrepresentative, etc.). After review, any identified clusters or data are deleted and the cluster analysis is run again with the reduced dataset.

In IBM SPSS, the clustering schedule contains the following information:

- **Stage:** Recall that the stage is the step in the clustering process where the two most similar clusters are combined. For a hierarchical process, there are always  $N - 1$  stages, where  $N$  is the number of observations being clustered. In this HBAT analysis there is a sample size of 100 resulting in 99 stages.
- **Clusters Combined:** Information detailing which two clusters are combined at each stage. Clusters are labeled by the ID of one of the members of the cluster.
- **Agglomeration Coefficient:** Measures the increase in heterogeneity (reduction in within cluster similarity) that occurs when two clusters are combined. For most hierarchical methods, the agglomeration coefficient is the distance between the two closest observations in the clusters being combined.
- **Stage Cluster First Appears:** Identifies the prior stage at which each cluster being combined was involved. Values of zero indicate the observation is still a single member cluster. That is, the observation has never been combined before that stage.
- **Next Stage in Which New Cluster Appears:** Denotes the next stage at which the new cluster is combined with another cluster.

We can use this information to understand the clustering process at each stage and even follow a cluster throughout the process. Table 4.8 shows a portion of the clustering schedule produced by the hierarchical cluster results. For example, in stage 1 observations 3 and 94, each a single member cluster, combine to create the first cluster with more than one observation. The agglomeration coefficient (which is a within-subjects sum of squares when using Ward's algorithm) is only 0.080. The hierarchical process concludes (stage 99) when cluster 1 combines with cluster 6 to form a single cluster (all 100 observations) with an agglomeration coefficient of 812.8.

**EVALUATING CLUSTER SIZES FOR SINGLE MEMBER CLUSTERS OR OUTLIERS** The agglomeration schedule contains information that helps in identifying either single-member or very small clusters that may be included in the final cluster solutions or act as outliers. Single-member clusters (i.e., observations that have not joined a cluster) are uniquely identified by a zero in the columns *Stage Cluster First Appears*. In Table 4.8 we can see that in stage 1, observations 3 and 94 are joined and they are both single-member clusters. In stage 2, two different single-member clusters are joined together. Then, in stage 18, the cluster formed in stage 1 (observations 3 and 94) is joined with a single-member cluster (observation 38).

We can use this clustering information to identify when single-member clusters are joined much later in the process and are thus likely to be outliers. Obviously if single-member clusters are joined within the set of possible

**Table 4.8** Partial Agglomeration Schedule for Initial HBAT Hierarchical Cluster Solution

Agglomeration Schedule				Stage Cluster First Appears		
Stage	Cluster Combined		Coefficients	Cluster 1	Cluster 2	Next Stage
	Cluster 1	Cluster 2				
1	3	94	.080	0	0	18
2	75	96	.180	0	0	62
.	.	.	.	.	.	.
18	3	38	6.065	1	0	67
.	.	.	.	.	.	.
74	2	98	120.542	59	0	92
75	6	87	125.83	0	0	89
76	32	84	131.506	54	0	86
77	3	50	137.566	67	62	83
.	.	.	.	.	.	.
98	6	11	659.781	97	95	99
99	1	6	812.825	96	98	0

Note: Stages 3–17, 19–73, and 78–97 have been omitted from the table.

cluster solutions, these observations would be deleted as outliers or else a cluster solution would contain a cluster with a single member.

But we should also examine a broader set of cluster solutions to identify any “late” single-member clusters which actually represent outliers. In Table 4.8, for example, stage 75 is where we see that observation 6 is combining with observation 87 to form a cluster. Both are single-member-clusters since under the columns labeled *Stage Cluster First Appears* we see that both have values of zero (indicating they had never joined before). Recall in our earlier discussion of outliers in the HBAT data that these two observations were indicated as the most dissimilar observations in the sample and possibly outliers. At this point, with this additional information, the researcher decides observations 6 and 87 should be removed as outliers. Thus, the cluster analysis will be run again (respecified) after observations 6 and 87 are removed. Observation 84 also remains a single-member cluster until late in the process (stage 76), but we decide to keep it because it did not display as high an average distance from the other observations.

One issue in evaluating the agglomeration schedule is that there is no information as to the size of the clusters being joined except for single member clusters which we discussed earlier. An important task is to identify if possible clusters of unacceptable size before we select a final cluster solution since we will just have to delete the cluster(s) of unacceptable size and start the analysis again. There are several ways we can use the agglomeration schedule toward this goal. First, we can anticipate if smaller clusters are being joined by using the *Stage Cluster First Appears* values to estimate cluster size. Most likely clusters formed at earlier stages are smaller, so if the values in these columns is relatively small, it may indicate a smaller cluster. The researcher may also find it useful to examine the column *Next Stage* for earlier cluster solutions that are small to make sure they are joined before the range of final cluster solutions is evaluated. For example, if two single member clusters were joined and they had a value of 95 in *Next Stage*, that would indicate that they would be present in all of the cluster solutions for Stages 94 and higher. In this analysis, the only problematic cluster looks to be that from Stage 75 discussed earlier. When we view its value for *Next Stage*, we see Stage 89, which indicates it joined with another cluster in the 11 cluster solution. So it should not be a problem when considering cluster solutions of 10 or less. A final approach is to save the cluster solution at the maximum number of clusters to be considered and make sure that all the clusters at that point are at acceptable sizes. In this

analysis when we view the 10 cluster solution, all of the clusters at that stage are at or above the minimum cluster size of 5, so we can proceed knowing that an unacceptably small cluster will not exist in the range of potential cluster solutions. None of these quick checks are sufficient by itself, but in combination they are aimed at eliminating outliers and avoiding the selection of a cluster solution which contains clusters of unacceptable size.

Changes in the agglomeration coefficients can be used along with a number of other diagnostic measures as stopping rules to help identify the appropriate number of clusters. We discuss this process in the next section.

**DENDROGRAM** A final perspective on the agglomeration process is with the dendrogram, a tree-like structure which depicts each stage of the clustering process. Typically, the graph is scaled, so that closer distances between combinations indicate greater homogeneity. The dendrogram is not reproduced here as its size makes it unreadable, but it displays the same pattern as shown in the agglomeration schedule.

**Step 3: Respecified Cluster Results** The deletion of two observations requires that the cluster analysis be performed again on the remaining 98 observations. We now discuss the findings of the new cluster analysis, including examining cluster sizes and the clustering criteria.

**EVALUATING CLUSTER SIZES** The process proceeds as before, with the respecified cluster results first examined for inappropriate cluster sizes. Even though two outliers were eliminated for this analysis, the researcher should still examine the results for any additional single member or extremely small clusters that occur and remove them from further analysis. Examination of the new clustering schedule did not reveal any additional outliers. Moreover, when the ten cluster solution is examined, All of the cluster sizes are above the desired minimum of five observations.

Researchers should be cautioned against “getting into a loop” by continually deleting small clusters and then respecifying the cluster analysis. Judgment must be used in accepting a small cluster and retaining it in the analysis at some point. Otherwise, you may find that the process will start to delete small, but representative, segments. The most problematic small clusters to retain are those that do not combine until very late in the process. But small clusters that are merged in the higher ranges of cluster solutions may be retained. Further, when applying a two-step cluster approach, hierarchical followed by nonhierarchical, final cluster sizes are uncertain until the second step. For several reasons, therefore, care should be taken in deleting small clusters even if they contain slightly fewer observations than would be considered managerially useful.

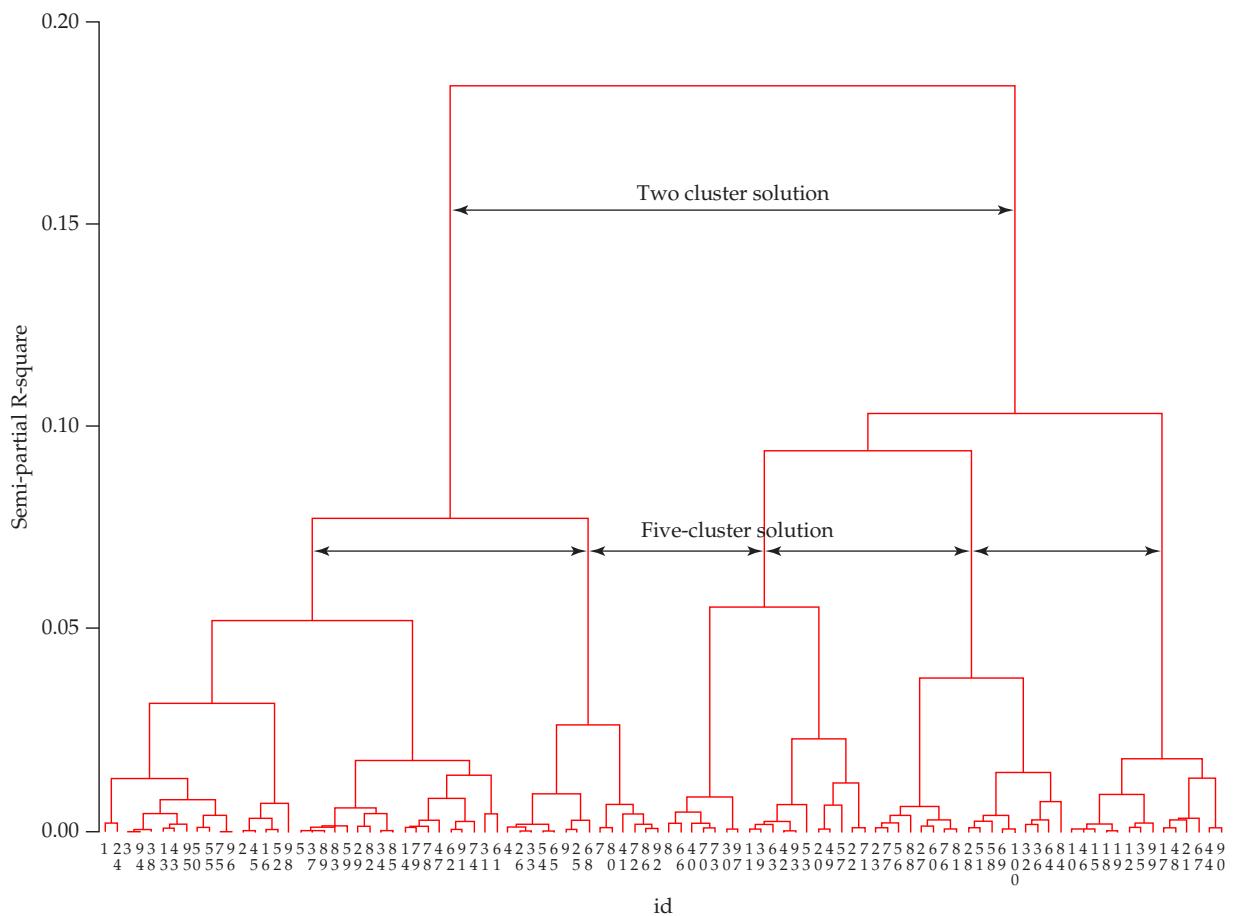
Selection of the linkage method and calculation of similarity also can have a substantive impact in this regard. For example, the HBAT research team chose Ward’s algorithm as the clustering method. This reduces the chance of finding small clusters, because it tends to produce clusters of equal size. Quite different results might be expected if methods such as singe linkage or centroid methods were employed instead.

**USING THE DENDROGRAM.** As noted earlier, the dendrogram is a graphical tree-like depiction of the agglomerative process. While it has limited value as in selecting a specific cluster solution, it does provide a more qualitative assessment of the various cluster solutions in terms of relative size and the changes in heterogeneity across the range of solutions. We can see several characteristics of the clustering process in Figure 4.13, including indications of the two- and five-cluster solutions:

- As we will see in the next section examining the clustering schedule in more detail, the final stage involves joining two equally sized clusters. This is depicted in the dendrogram in the connection of the two clusters at the top of the dendrogram.
- Below this, we see the agglomerative process for each of those two clusters where they are progressively “built up” by the combination of smaller clusters. If we look across the lower portion, we can see emerging clusters, even distinguishing a five-cluster solution of more distinctiveness among the observations
- Finally, we can see that there are no single-member clusters in the later stages of the clustering process since they would be represented by single narrow lines extending upward.

The dendrogram in many instances can provide a useful perspective on the clustering process. Its primary limitation is that it becomes less useful as the sample size increases and is not possible to construct for even

**Figure 4.13**  
Dendrogram For Reduced HBAT Cluster Sample



Note: Adapted from SAS output.

reasonably-sized samples. Here HBAT has only 100 observations and the dendrogram becomes difficult to view in totality. But where possible its use can provide insight into the more quantitative stopping rules discussed below.

**DETERMINING THE PRELIMINARY HIERARCHICAL CLUSTERING SOLUTION(S)** With all of the individual clusters in the ten-cluster solution (the maximum number of clusters considered by HBAT team) meeting the minimum cluster size of five observations, we are assured of the appropriateness of the clustering solution in terms of cluster sizes and the impact of outliers. But we still have not addressed the fundamental question: What is the final cluster solution? We should note that in most situations a single final solution will not be identified in hierarchical analysis. Rather, a set of preliminary cluster solutions is identified. These cluster solutions form the basis for additional profiling and then possible submission to a nonhierarchical analysis from which a final cluster solution is selected. Even though a final cluster solution is not identified at this stage, the researcher must make a critical decision as to how many clusters will be used in the profiling and nonhierarchical analysis. Hopefully, the decision will be relatively clear and the potential cluster solution can be determined. However, researchers routinely decide that a small number of cluster solutions rather than just a single cluster solution emerge from the nonhierarchical procedures. Further analysis of multiple cluster solutions generally provides assurance as to how many clusters are most appropriate.

**ALTERNATIVE STOPPING RULES** The following discussion will focus on the various stopping rules available for hierarchical analyses and then identify a cluster solution(s) for further profiling and use in the nonhierarchical analysis.

**Percentage Changes in Heterogeneity** As discussed earlier, the fundamental principle in all stopping rules is to assess the changes in heterogeneity between cluster solutions. The basic rationale is that when large increases in heterogeneity occur in moving from one stage to the next, the researcher selects the prior cluster solution because the new combination is joining quite different clusters.

The agglomeration coefficient is particularly useful for this stopping rule. Small coefficients indicate that fairly homogeneous clusters are being merged, whereas joining two very different clusters results in a large coefficient. Because each combination of clusters results in increased heterogeneity, we focus on large percentage changes in the coefficient, similar to the scree test in exploratory factor analysis, to identify cluster combination stages that are markedly different. The only caveat is that this approach, although a fairly accurate algorithm, has the tendency to indicate too few clusters.

Table 4.9 contains information from the later stages of the clustering schedule (Stages 90 through 97) for the reduced sample. In this situation we only provided information from the seven-cluster solution forward since this was the upper limit desired by HBAT management. The first four columns contain information similar to that in Table 4.8. However, three additional columns have been added to facilitate an understanding of this solution. The fifth column states the number of clusters that exist upon completing this stage. A sixth column shows the differences in the agglomeration coefficients between a particular stage and the next combination. In other words, it shows how much smaller the coefficient is compared to the next stage. Recall that these coefficients also indicate how much heterogeneity exists in the cluster solution. Thus, the differences indicate how much the heterogeneity increases when you move from one stage to the next and this stopping rule is based on calculating the percentage change in the clustering coefficients for these final stages. Two points worth remembering in interpreting these solutions are:

- By their nature, the size of the agglomeration coefficient gets larger toward the end of the cluster solution.
- Applying a stopping rule is not an exact science.

We now look for relatively large increases in the agglomeration coefficients. The agglomeration coefficient shows rather large increases in going from stages 94 to 95 (465.38 versus 536.24), stages 95 to 96 (536.24 versus 613.79), and stages 96 to 97 (613.79 versus 752.50). The average percentage increase for all stages shown (90 to 97) is 14.2 percent and serves as a rough guide in determining what a large increase is. A graphical depiction of the percentage change is shown in Figure 4.13. Let's examine each stage in more detail to understand how they assist in selecting a final cluster solution.

- Stages 96 and 97 (two to one cluster): Here we condense a two-cluster solution at Stage 96 into a one-cluster solution containing all 98 observations at Stage 97. Combining the two clusters into one yields a percentage increase of 22.6 percent ( $(752.50 - 613.79)/613.79 = .226$ ) Although this is the largest increase, cluster solutions almost always show a large increase at this point indicating a possible two-cluster solution. Yet a two-cluster solution also may represent limited value in meeting many research objectives. Researchers must avoid the temptation to say the two-cluster solution is the best, because it involves the largest change in heterogeneity. A two-cluster solution must be supported by strong theoretical reasoning. For these reasons we will not consider a two-cluster solution for further analysis.
- Stages 95 and 96 (three to two clusters): The percentage increase associated with moving from three to two clusters is 14.5 percent ( $((613.79 - 536.24)/536.24)$ , the second highest if we exclude moving from two to one cluster. Yet it is only slightly above the increase when moving from five to four clusters (see discussion below), so will remain as a possible alternative cluster solution for further consideration.
- Stages 94 and 95 (four to three clusters): The highest percentage increase, excluding moving from two to one clusters, is found when moving from the four-cluster (Stage 94) to the three-cluster solution (Stage 95): 15.2 percent ( $((536.24 - 465.39)/465.39)$ ). As such, the cluster solution associated with four clusters is associated with proportionately less heterogeneity than is the three-cluster solution. The higher value than other possible cluster solutions provides strong support for consideration as a final cluster solution.

**Table 4.9** Agglomeration Schedule and Percentage Change in Heterogeneity for the Reduced HBAT Cluster Sample

Stage	Cluster 1	Combined with Cluster:	Coefficient	Number of Clusters After			Proportionate Increase in Heterogeneity to Next Stage	Stopping Rule	
				Combining	Differences				
90	1	2	297.81	8	28.65	9.6%	HBAT not interested in this many clusters.		
91	22	27	326.46	7	39.11	12.0%	Increase is larger than the previous stage, arguing against combination.		
92	1	5	365.56	6	41.82	11.4%	Increase is relatively small, favoring combination to five clusters.		
93	7	10	407.38	5	58.01	14.2%	Increase is larger than the previous stage, favoring five to four clusters.		
94	1	4	465.39	4	70.86	15.2%	Increase is relatively large, favoring four clusters over three and suggests a possible stopping point.		
95	7	22	536.24	3	77.55	14.5%	Increase is relatively large and favors a three-cluster solution over a two-cluster solution.		
96	7	9	613.79	2	138.71	22.6%	Increase from two to one is relatively large (the increase from two to one is normally large).		
97	1	7	752.50	1	—	—	One-cluster solution not meaningful.		

- Stages 93 and 94 (five to four clusters): Here we see a percentage increase of 14.2 percent, almost equal to that seen for Stages 95/96 and only slightly below the 15.2 percent seen for Stages 94/95. Using more of a scree plot logic, an argument can be made that this would be a stopping point.
- Stages 90, 91 and 92: All of the percentage increases associated with these stages are below the average and less than all the stages following them in the process. They would not be likely candidates for a final cluster solution unless the research objectives necessitated a much larger number of clusters.

Let's review the possibilities for selecting one or more cluster solutions using the percentage change in heterogeneity as our stopping rule.

- While the two-cluster solution is associated with the highest percentage change in heterogeneity, it will not be considered as a final cluster solution given the expected large increase in heterogeneity associated with this stage and its limited usefulness in meeting the research objectives
- The four-cluster solution has the next highest percentage change in heterogeneity and will be considered in further analyses.
- The three-cluster and five-cluster solutions are plausible candidates to compare with a four-cluster solution, particularly if the four-cluster solution proves difficult to interpret or has otherwise undesirable characteristics.

These findings are indicative of the somewhat subjective nature of selecting a cluster solution(s). In this example, several clusters solutions are quite comparable in terms of their associated percentage change in heterogeneity. As a result, we will examine several other stopping rules to see what additional information they can provide in selecting a final cluster solution.

**Statistical Measures of Heterogeneity Change** In addition to the percentage change in heterogeneity, there is also the Pseudo  $T^2$  measure which evaluates the statistical significance of changes in heterogeneity when joining clusters. The Pseudo  $T^2$  values are calculated for the combined clusters, so when we see a large value we “back up” to the previous solution as the cluster solution with less heterogeneity. Table 4.10 and Figure 4.15 show the Pseudo  $T^2$  values for the final stages of the clustering process. We can see that outside of combining all observations into a single cluster (the last stage in the prior clustering schedule), the highest values are for clusters solutions of three and four clusters, indicating the cluster solutions of four and five clusters are most appropriate for consideration, followed by a three-cluster solution. This corresponds to our conclusions drawn from the percentage changes in heterogeneity.

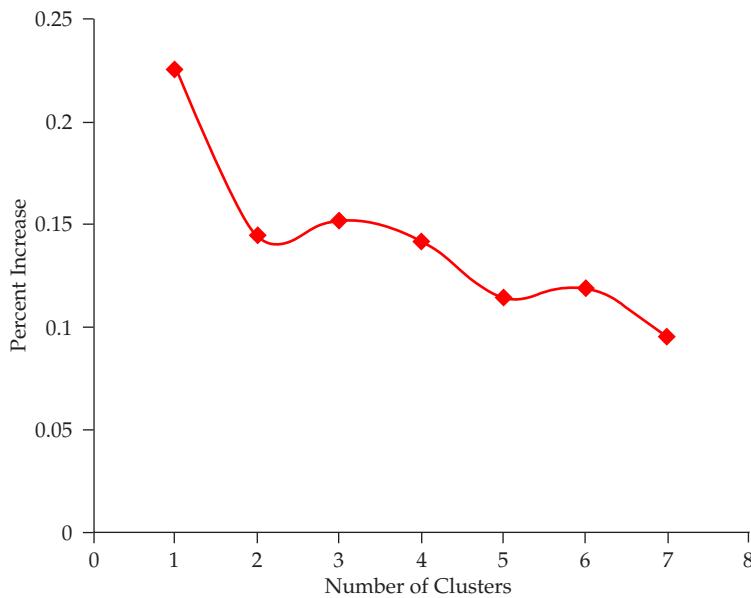
**Direct Measure of Heterogeneity** The CCC value indicates the heterogeneity at each stage of the clustering process, with higher values indicating the more homogeneous cluster solutions. As we view the CCC values in Table 4.10 and Shown in Figure 4.15, we see that none of the cluster solutions are particularly distinctive as indicated by “peaks” in the values. As seen in many situations, the values are all negative, which indicative of more homogenous samples and/or the presence of outliers. Given the elimination of outliers, the CCC supports the results of the other stopping rules which indicates a broader range of solutions, all relatively similar.

**Statistical Significance of Cluster Variation** The final stopping rule measure is the Pseudo  $F$  statistic, which is a measure of the homogeneity of each cluster solution. As with the CCC, we should select cluster solutions with relatively higher values of the Pseudo  $F$  value. As we review the values in Table 4.10 and Figure 4.15, we see little variation across the cluster solutions, a common pattern across all of the stopping rules, with a five- cluster solution having a slightly higher value than the other solutions being considered.

**Selecting a Hierarchical Cluster Solution** As we have seen across all of the stopping rules, the three-, four- and five-cluster solutions are quite similar in terms of their structure. In some instances the four- or five-cluster solution is indicated, but never with a marked difference from these other solutions. As a result, the four-cluster solution will be selected as the cluster solution for consideration in subsequent profiling and as the starting point for the

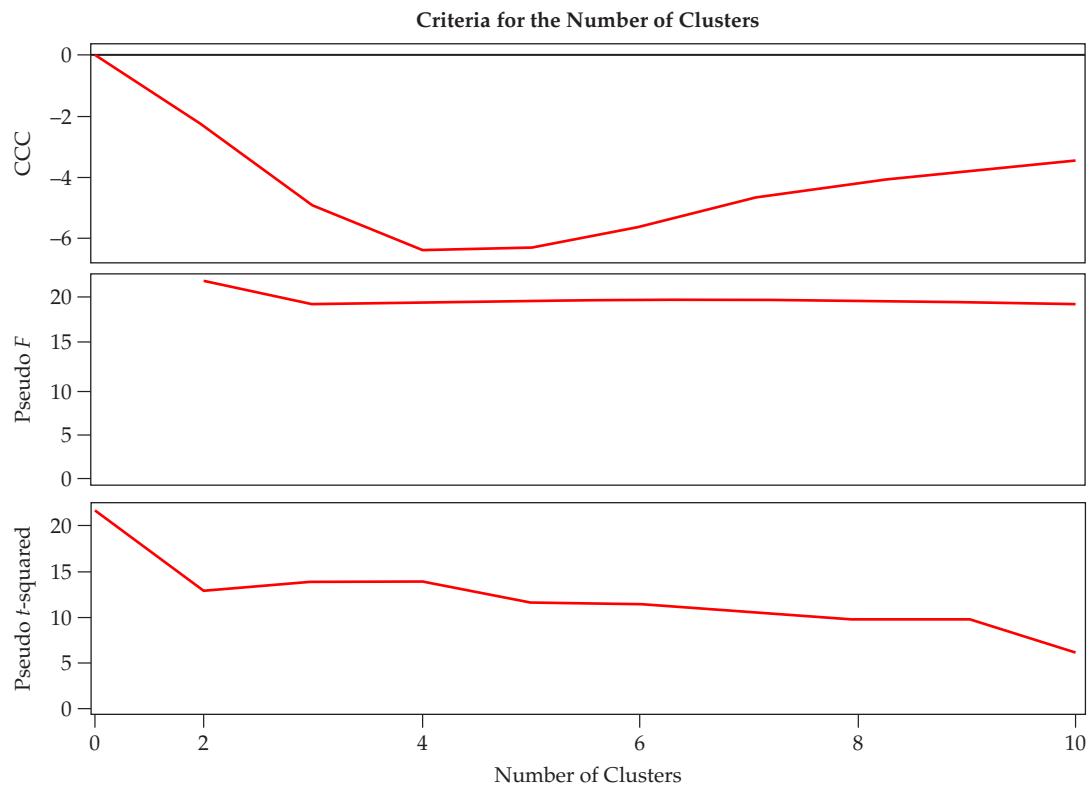
**Table 4.10** Additional Stopping Rule Measures for Reduced HBAT Cluster Sample

Number of Clusters	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared
10	-3.5	19.2	6.2
9	-3.8	19.4	9.9
8	-4.2	19.6	9.8
7	-4.7	19.8	10.6
6	-5.6	19.5	11.4
5	-6.3	19.7	11.6
4	-6.4	19.3	13.9
3	-4.9	19.2	13.9
2	-2.3	21.7	12.9
1	0.00	.	21.7



**Figure 4.14**  
Stopping Rule: Percent Change in Heterogeneity

**Figure 4.15**  
Additional Stopping Rules: Direct Measures and a Statistical Measure of Change in Heterogeneity

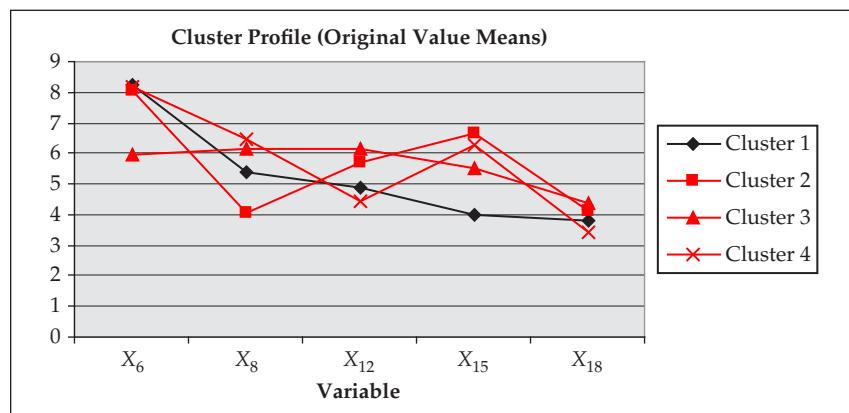


nonhierarchical process. As can be seen from application of all of these stopping rules there many times is no single definitive cluster solution, rather a set of potential cluster solutions for further consideration. This highlights both the need for a rigorous validation process as well as the judgment of researchers as to which cluster solution best meets the research objectives.

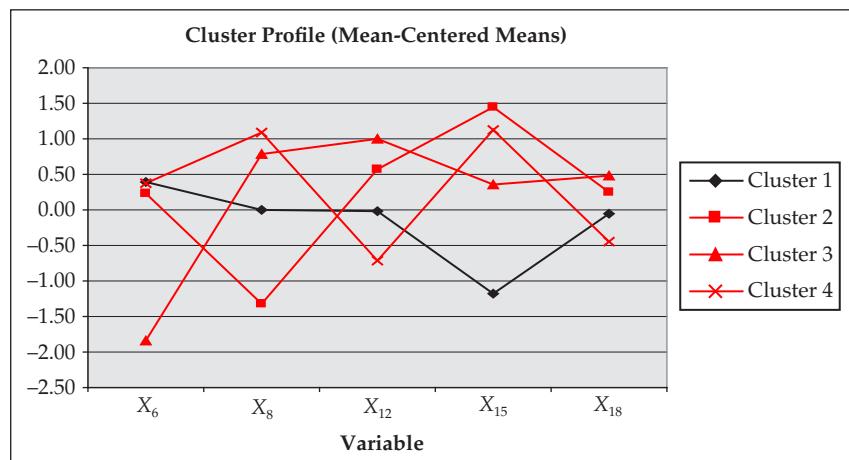
**Step 4: Profiling the Clustering Variables** Before proceeding to the nonhierarchical analysis, we will profile the clustering variables for the four-cluster solution to confirm that the differences between clusters are distinctive and significant in light of the research question and to define the characteristics of the clusters. The profiling information is shown in Figure 4.16.

Let us examine the distinctiveness first. At the far right side of the figure are the  $F$  statistics from one-way ANOVAs that examine whether there are statistically significant differences between the four clusters on each of the five clustering variables. The independent variable is cluster membership (which of the four clusters each of the 98 observations were placed in by the clustering process), and the dependent variables are the five clustering variables. The results show there are significant differences between the clusters on all five variables. The significant  $F$  statistics provide initial evidence that each of the four clusters is distinctive. This is generally the outcome for the clustering variables as this is the primary task of the clustering algorithm. If variables are found to not be significantly different, they are candidates for elimination since they do not provide a means of distinctiveness across the clusters.

Variable	Means from Hierarchical Cluster Analysis									
	Mean Values Cluster Number:				Mean-Centered Values Cluster Number:					
	1	2	3	4	1	2	3	4	F	Sig
$X_6$ Product Quality	8.21	8.04	5.97	8.18	0.40	0.23	-1.84	0.37	14.56	0.000
$X_8$ Technical Support	5.37	4.04	6.16	6.47	0.00	-1.33	0.78	1.09	12.64	0.000
$X_{12}$ Salesforce Image	4.91	5.69	6.12	4.42	-0.02	0.57	1.00	-0.72	11.80	0.005
$X_{15}$ New Products	3.97	6.63	5.51	6.28	-1.18	1.45	0.36	1.13	62.74	0.000
$X_{18}$ Delivery Speed	3.83	4.14	4.37	3.45	-0.06	0.25	0.48	-0.44	5.49	0.002
Cluster sample sizes	49	18	14	17	49	18	14	17		



**Figure 4.16**  
Profile of Four Clusters from  
Hierarchical Cluster Analysis



Now we examine the means of the five-cluster variables to address the profile of each cluster, what clustering variables are high or low and how it corresponds to the research objectives. This stage is particularly important if selecting a relatively small number of clusters. For example, many two-cluster solutions have one cluster as high on all variables and the other cluster as low on all variables. Or a three-cluster solution may generate clusters that are relatively high, moderate and low on all variables. These types of solutions may be useful in some situations, but many times the research is looking for more varied profiles (e.g., high on some clustering variables and low on other variables). It may be necessary to select cluster solutions with larger number of cluster to see these patterns emerge.

This stage in the profiling process is based on interpretation of both the mean values and the mean-centered values. Cluster 1 contains 49 observations and has a relatively lower mean on  $X_{15}$  (New Products) than the other three clusters. The means of the other three clusters are somewhat above average. Cluster 2 contains 18 observations and is best characterized by two variables: a very low mean on  $X_8$  (Technical Support) and the highest score on  $X_{15}$  (New Products). Cluster 3 has 14 observations and is best characterized by a relatively low score on  $X_6$  (product quality). Cluster 4 has 17 observations and is characterized by a relatively low score on  $X_{12}$  (Salesforce Image). These results indicate that each of the four clusters exhibit somewhat distinctive characteristics. Moreover, no clusters contain less than 10 percent of observations. Therefore, the four-cluster solution is sufficiently favorable to indicate moving on to nonhierarchical clustering. The cluster sizes will change in the nonhierarchical analysis and observations will be reassigned. As a result, the final meanings of the four clusters will be determined in the nonhierarchical analysis.

## PART 2: NONHIERARCHICAL CLUSTER ANALYSIS (STAGES 4–6)

The hierarchical clustering method facilitated a comprehensive evaluation of a wide range of cluster solutions. These solutions were impacted, however, by a common characteristic—once observations are joined in a cluster, they are never separated (reassigned) in the clustering process. In the hierarchical clustering process, we selected the algorithm (Ward's) that minimized the impact of this process. But nonhierarchical clustering methods have the advantage of being able to better “optimize” cluster solutions by reassigning observations until maximum homogeneity (similarity) within clusters is achieved.

This second part in the clustering process uses results of the hierarchical process to execute nonhierarchical clustering. Specifically, the number of clusters is determined from the hierarchical results. Nonhierarchical procedures then develop “optimal” cluster solutions. The cluster solutions are then compared in terms of criterion validity as well as applicability to the research question to select a single solution as the final cluster solution.

**Stage 4: Deriving Clusters and Assessing Overall Fit** The primary objective of the second part is using nonhierarchical techniques to adjust, or “fine-tune,” the results from the hierarchical procedures. In performing a nonhierarchical cluster, researchers must make two decisions:

- 1 How will seed points for the clusters be generated?
- 2 What clustering algorithm will be used?

The following discussion addresses both of these points by demonstrating how to use the hierarchical results to improve the nonhierarchical procedure.

**SPECIFYING CLUSTER SEED POINTS** The first task in nonhierarchical cluster analysis is to select the method for specifying cluster seeds. The cluster seeds are the initial starting point for each cluster. From there, the clustering algorithm assigns observations to each seed and forms clusters. There are two methods for selecting cluster seed points: generation from the sample by the cluster software (i.e., random selection of observations to act as seed points) and specification by the researcher. The most common approach is the use of sample-generated seed points, either because there is not a hierarchical analysis to rely upon or generating the seeds points is too cumbersome. Sample-generated methods sometimes produce clusters that are difficult to replicate across samples because different observations are selected as seed points for each analysis, plus they have no direct relationship to the hierarchical results except for the number of clusters.

In contrast, researcher-specified cluster seeds provide some conceptual or empirical basis for selecting the seed points. Researcher-specified methods reduce problems with replicability, but choosing the best seed points can be

difficult, and with some software packages inserting researcher determined cluster seeds is complicated. In most instances researcher-specified cluster seeds are based on the hierarchical solution. This involves either selecting a single observation from each cluster to represent the cluster or, more commonly, to use the cluster centroids as the seed points. Note that deriving the cluster centroids typically requires additional analysis to (a) select the cluster solution(s) to be used in the nonhierarchical analysis and (b) to derive the centroids by profiling each cluster solution. These profiles are not typically generated in the hierarchical analysis, because doing so requires a tremendous effort, as  $N - 1$  cluster solutions (97 solutions in the HBAT example) are generated, and deriving a profile for each would be time consuming and inefficient. In our example, all five clustering variables will be used in the nonhierarchical analysis. Thus, the cluster seed points would require initial values on each variable for each cluster.

To illustrate the most common approach, the research team decides to use the random initial seed points identified by the software. These seed points are affected by the ordering of the observations in the data file. To evaluate cluster solution stability, some researchers reorganize the data (change the order of the observations) and rerun the cluster analysis. If the cluster solutions change substantially, which indicates they are highly unstable, researcher-specified seed points may be need to be used.

**SELECTING A CLUSTERING ALGORITHM** The analyst must now select the clustering algorithm to be used in forming clusters. A primary benefit of nonhierarchical methods is that since each cluster solution is a separate analysis, with cluster membership for one cluster solution (e.g., four clusters) totally separate from another cluster solution (e.g., five clusters). This is in contrast to hierarchical methods, where cluster solutions formed later in the clustering process are directly based on combining two clusters formed earlier in the process. Moreover, nonhierarchical cluster methods also have the ability to develop final cluster assignments that are not based on the order in which observations are processed. This is because observations assigned to a particular cluster early in the clustering process can later be reassigned (moved from one cluster to another) to another cluster formed later in the process if that improves cluster homogeneity. For these reasons, nonhierarchical methods are generally preferred, when possible, for their “fine-tuning” of an existing cluster solution from a hierarchical process.

For the HBAT example, we selected the optimizing algorithm in IBM SPSS that allows for reassignment of observations among clusters until a minimum level of heterogeneity is reached. Using this algorithm, observations are initially grouped to the closest cluster seed. When all observations are assigned, each observation is evaluated to see if it is still in the closest cluster. If it is not, it is reassigned to a closer cluster. The process continues until the homogeneity within clusters cannot be increased by further movement (reassignment) of observations between clusters.

**FORMING CLUSTERS** With the cluster seeds and clustering algorithm specified, the clustering process can begin. To execute the nonhierarchical cluster, we specify the number of clusters as four, based on the results of the hierarchical cluster solution. Using the k-means optimizing algorithm, the process continues to reassign observations until reassignment will not improve within-cluster homogeneity.

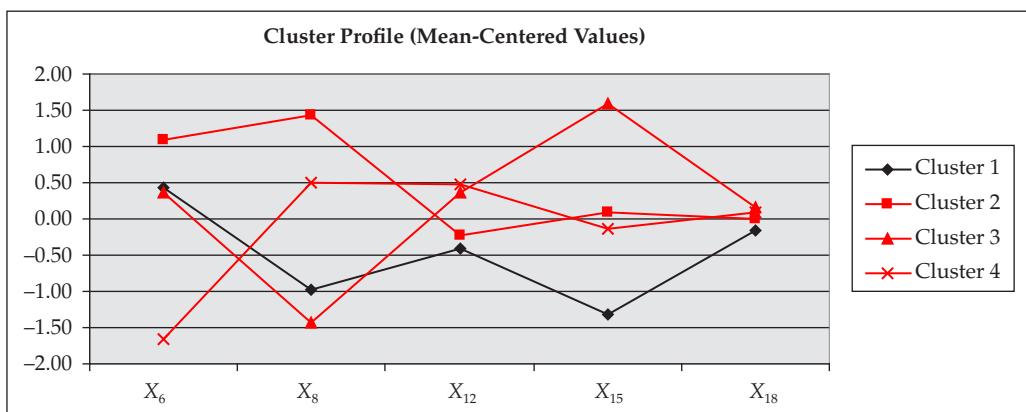
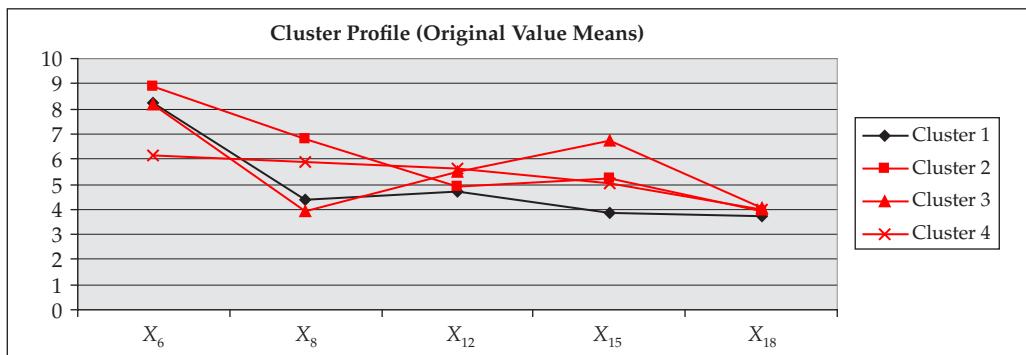
Results from the nonhierarchical four-cluster solution are shown in Figure 4.17. There are two notable differences between the hierarchical and nonhierarchical results:

- **Cluster Sizes.** The nonhierarchical solution, perhaps due to the ability to reassign observations between clusters, has a more even dispersion of observations among the clusters. As an example, nonhierarchical analysis resulted in cluster sizes of 25, 29, 17, and 27, compared to clusters of 49, 18, 14, and 17 in the hierarchical analysis.
- **Significance of Clustering Variable Differences.** Another fundamental difference between the two cluster solutions is the ability of the nonhierarchical process to delineate clusters that are usually more distinctive than the hierarchical cluster solution. Figure 4.17 includes ANOVA results showing the differences in variable means across four clusters. Given that the five clustering variables were used to produce the clusters, the results should be statistically significant. The  $F$ -values indicate that the means of four of the five variables are significantly different. Only the means of  $X_{18}$  (delivery speed) are not significantly different across groups. In fact, three of the five clustering variables have very large  $F$ -values ( $X_6$ ,  $X_8$ , and  $X_{15}$ ). Thus, the nonhierarchical results suggest that the cluster solution is adequately discriminating observations, with the exception of  $X_{18}$ , delivery speed.

**Figure 4.17**

Profile of Four Clusters from K-Means Cluster Solution

Variable	Mean Values Cluster Number:				Mean-Centered Values Cluster Number:					F	Sig
	1	2	3	4	1	2	3	4			
$X_6$ Product Quality	8.25	8.91	8.18	6.14	0.44	1.10	0.37	-1.67	55.06	0.000	
$X_8$ Technical Support	4.40	6.80	3.92	5.86	-0.97	1.43	-1.44	0.50	45.56	0.000	
$X_{12}$ Salesforce Image	4.70	4.89	5.49	5.59	-0.42	-1.33	0.37	0.47	4.56	0.005	
$X_{15}$ New Products	3.83	5.25	6.75	5.01	-1.32	0.10	1.60	-0.14	25.56	0.000	
$X_{18}$ Delivery Speed	3.72	3.89	4.05	3.98	-0.17	0.01	0.17	0.10	0.88	0.002	
Cluster sample sizes	25	29	17	27	25	29	17	27			



The nonhierarchical clustering process produced a four-cluster solution based on the software-generated seed points. Further analysis in terms of profiling the solutions and assessing their criterion validity will provide the elements needed to select a final cluster solution.

**Stage 5: Profiling the Clustering Variables** The HBAT research team first characterizes the clusters by analyzing the pattern of cluster means and mean-centered values shown in Figure 4.17, which are plotted in the profile diagram in the figure. Interpretation begins by looking for extreme values associated with each cluster. In other words, variable means that are the highest or lowest compared to other clusters are useful in distinguishing between the clusters on the clustering variables.

- Cluster 1 has 25 observations and is most distinguished by a relatively low mean for new products ( $X_{15}$ ). The means for the other variables (except product quality) also are relatively low. Thus, this cluster represents a market segment characterized by the belief that HBAT does not perform well in general, particularly in offering new products, and the overall lower means suggest this segment is not a likely target for new product introductions.

- Cluster 2 has 29 observations and is most distinguished by relatively higher means on technical support ( $X_8$ ) and on product quality ( $X_6$ ). Therefore, HBAT interprets this market segment as believing that HBAT provides strong support for its high-quality products. HBAT believes this is a favorable segment for other products and services. This is the largest cluster.
- Cluster 3 has 17 observations and is distinguished by a relatively higher mean for new products ( $X_{15}$ ). In contrast, cluster 3 has a relatively lower mean for technical support ( $X_8$ ). Thus, this segment suggests that HBAT offers new and innovative products, but that its support of these new products is not very good. It should be noted that the new product mean of 6.75, although the highest for any cluster, is still only moderate overall, and HBAT could improve in this area even with this cluster. Cluster 3 is somewhat the opposite of cluster 2.
- Cluster 4 has 27 observations and is distinguished by the lowest mean of all clusters on product quality ( $X_6$ ). All of the means of variable  $X_6$  in the other clusters are much higher. Moreover, the means on the other clustering variables are relatively average for this cluster, with the exception of a somewhat higher mean on salesforce image ( $X_{12}$ ) and technical support ( $X_6$ ). Thus, the segment is characterized as one that indicates that HBAT's products are below average in quality, but the salesforce and technical support are slightly better than average. This is the second largest cluster.

The cluster profiles offer the HBAT research team the ability to evaluate the usefulness of the four-cluster solution in terms of the size of the clusters and the differentiated profiles of each. In this case, distinctive profiles of each cluster provide a basis for differentiated approaches in appealing to each cluster. But the research team must still validate the cluster solution in terms of relevant outcome variables and their unique profiles on an additional set of identifying variables.

**Stage 6: Validation and Profiling the Clusters** In this final stage, the processes of validation and profiling are critical due to the exploratory and often atheoretical basis for the cluster analysis. Researchers should perform tests to confirm the validity of the cluster solution while also ensuring the cluster solution has practical significance. Researchers who minimize or skip this step risk accepting a cluster solution that is specific only to the sample and has limited generalizability, or even little use beyond its mere description of the data on the clustering variables.

**CLUSTER STABILITY** At this point, researchers often assess the stability of the cluster solution. Given that the software chose the initial seed points, factors such as the ordering of the cases in the data can affect cluster membership. To do so, the researcher can sort the observations in a different order and then perform the cluster analysis once again (with the new starting point selected by the software, but with the same number of clusters specified). The cluster solutions can then be compared to see if the same clusters are identified. A cross-classification of cluster membership between solutions should reveal mostly matches between the two solutions. In other words, the observations that cluster together in one analysis should for the most part cluster together in the subsequent cluster solution. This is best illustrated by an example.

The stability of the HBAT four-group nonhierarchical cluster solution is examined by comparing two different solutions using sample-generated seed points. The researcher first sorts the observations into a different order. To do so, the researcher selects a variable from the dataset and uses the IBM SPSS sort function to change the order of the observations. In this case, the observations were sorted by customer type ( $X_1$ ), ranging from those with the least time doing business with HBAT to those with the most time doing business with HBAT. The k-means algorithm is

**Table 4.11** Cross-Classification to Assess Cluster Stability

		Cluster Number from Second K-Means				
Cluster Number from First K-Means		1	2	3	4	Total
1		0	0	1	24	25
2		2	21	0	6	29
3		0	0	17	0	17
4		22	0	5	0	27
Total		24	21	23	30	98

once again used to place observations into one of four clusters. Following the clustering routine, a cross-classification is performed (much like a confusion matrix in discriminant analysis), using the cluster membership variable from the first k-means solution as one variable and the cluster membership variable from the second k-means solution as the other variable. The results are shown in Table 4.11.

Most observations are grouped with the same observations they clustered with in the first nonhierarchical solution. Although cluster identifiers vary between solution, we see a high degree of correspondence between solutions. For example, cluster 1 in the first solution becomes cluster 4 in the second solution (as indicated by the 24 in the first row, fourth column of the cross-classification), all but one of the observations ends up clustering together. The one observation not staying together is now in cluster 3. For cluster 2, 8 of the 29 observations end up not clustering together. All cluster 3 observations stay together. Cluster 4, which is now cluster 1, retains 22 of the 27 original members. Perfect cross-validation would appear if only one cell in each row or column of the cross-classification contained a value. Thus, all but 14 observations have retained the same cluster membership across solutions—a result that supports the validity of a four-cluster solution. In other words, the four-cluster solution appears relatively stable with only 14 percent of the cases switching clusters between solutions. Additional cluster analyses conducted based on sorting the data in a different way could be conducted to further examine the stability of the data.

**ASSESSING CRITERION VALIDITY** To assess predictive validity, we focus on variables that have a theoretically-based relationship to the clustering variables but were not included in the cluster solution. Given this relationship, we should see significant differences in these variables across the clusters. If significant differences do exist on these variables, we can draw the conclusion that the clusters depict groups that have predictive validity.

For this purpose, we consider four outcome measures from the HBAT dataset which have managerial relevance to the research team:

- $X_{19}$  Satisfaction
- $X_{20}$  Likelihood to Recommend
- $X_{21}$  Likelihood to Purchase
- $X_{22}$  Purchase Level

**Table 4.12** Multivariate F Results Assessing Cluster Solution Criterion Validity

Variable	Cluster Number	Cluster Mean	Multivariate F <sup>*</sup>	Univariate F	Sig.
$X_{19}$ Satisfaction	1	6.76	2.23	5.98	0.01
	2	7.44			
	3	7.39			
	4	6.34			
$X_{20}$ Likely to Recommend	1	6.89		3.06	0.032
	2	7.46			
	3	7.14			
	4	6.68			
$X_{21}$ Likely to Purchase	1	7.74		3.53	0.018
	2	8.09			
	3	7.83			
	4	7.33			
$X_{22}$ Purchase Level	1	58.70		6.21	0.001
	2	62.17			
	3	60.92			
	4	53.17			

\*Multivariate F has 12,241 degrees of freedom and univariate Fs each have 3,94 degrees of freedom.

A MANOVA model (see Chapter 6) was estimated using the four criterion validity variables as the dependent variables and cluster membership as an independent variable. MANOVA was selected because the dependent variables are known to correlate with each other. Table 4.12 displays the results. First, the overall MANOVA model is significant ( $F = 2.23, P = .01$ ), providing initial support for the idea that these variables can be predicted by knowing to which segment an HBAT customer belongs. The individual univariate  $F$ -statistics are also significant, further verifying this finding.

The results demonstrate, therefore, that the cluster solution can predict other key outcomes, which provides evidence of criterion validity. For example, cluster 2, which HBAT believed was receptive to more business based on its cluster profile (described above), displays the highest scores on each of these key outcome variables. Thus, HBAT will likely find the cluster solution useful in predicting other key outcomes and forming appropriate strategies.

**PROFILING THE FINAL CLUSTER SOLUTION** The final task is to profile the clusters on a set of additional variables not included in the clustering variate or used to assess predictive validity. The importance of identifying unique profiles on these additional variables is in assessing both the practical significance and the theoretical basis of the identified clusters. In assessing practical significance, researchers often require that the clusters exhibit differences on a set of additional variables that can be used to identify them in the general population or the customer base.

In this example, five characteristics of the HBAT customers are available. These include  $X_1$  (Customer Type),  $X_2$  (Industry Type),  $X_3$  (Firm Size),  $X_4$  (Region), and  $X_5$  (Distribution System). Each of these variables is nonmetric, similar to the variable representing cluster membership for each observation. Thus, cross-classification is used to test the relationships.

Results of the cross-classification are provided in Table 4.13. Significant chi-square values are observed for three of the five profile variables. Several patterns are evident. For example, cluster 4 consists almost entirely of customers from outside of the USA/North America (26 out of 27). In contrast, cluster 2 consists predominantly of customers from the USA/North America. From these variables, distinctive profiles can be developed for each cluster. These profiles support the distinctiveness of the clusters on variables not used in the analysis at any prior point.

**Table 4.13** Results of Cross-Classification of Clusters on  $X_1, X_2, X_3, X_4$ , and  $X_5$

		Number of Cases Per Cluster				
Customer Characteristics		1	2	3	4	Total
$X_1$ Customer Type	Less than 1 year	8	5	5	12	30
	1 to 5 years	8	6	6	15	35
	More than 5 years	9	18	6	0	33
	Total ( $\chi^2 = 24.4, p < .001$ )	25	29	17	27	98
$X_2$ Industry Type	Magazine industry	8	21	10	12	51
	Newsprint industry	17	8	7	15	47
	Total ( $\chi^2 = 10.1, p < .05$ )	25	29	17	27	98
$X_3$ Firm Size	Small (0 to 499)	11	19	7	10	47
	Large (500+)	14	10	10	17	51
	Total ( $\chi^2 = 5.4, p > .1$ )	25	29	17	27	98
$X_4$ Region	USA/North America	14	14	8	1	39
	Outside North America	11	15	9	26	59
	Total ( $\chi^2 = 28.3, p < .001$ )	25	29	17	27	98
$X_5$ Distribution system	Indirect through broker	13	14	8	20	55
	Direct to customer	12	15	9	7	13
	Total ( $\chi^2 = 5.2, p > .1$ )	25	29	17	27	98

A successful segmentation analysis not only requires the identification of homogeneous groups (clusters), but also that the homogeneous groups are identifiable (uniquely described by other variables). When cluster analysis is used to verify a typology or other proposed grouping of objects, associated variables—either antecedents or outcomes—typically are profiled to ensure correspondence of the identified clusters within a larger theoretical model.

### EXAMINING AN ALTERNATIVE CLUSTER SOLUTION: STAGES 4–6

The four-cluster solution was examined first because it had the largest reduction in the agglomeration schedule error coefficient and this was generally supported by the other stopping rules (other than the two-group solution—see Tables 4.9 and 4.10). The HBAT management team then considered looking at both the five-cluster and three-cluster solutions. After reflection, management suggested that a smaller number of clusters would mean fewer market segments to develop separate strategies, and the result would likely be lower costs to execute the strategies. Moreover, the three-cluster solution not only had fewer clusters, but also exhibited the second largest increase in heterogeneity from three clusters to two, indicating that three clusters are substantially more distinct than two. As a result, the research team decided to examine the nonhierarchical three-cluster solution.

The results of the three-cluster solution are shown in Table 4.14. Cluster 1 has 44 customers, whereas clusters 2 and 3 each have 27 customers. Significant differences exist between the three clusters on three variables— $X_6$ ,  $X_{12}$ , and  $X_{15}$ —so the solution is discriminating between the three customer groups.

To interpret the clusters, we examine both the means and the mean-centered values. HBAT is perceived very unfavorably by cluster 1. Three of the variables ( $X_{12}$ ,  $X_{15}$ , and  $X_{18}$ ) are rated very poorly, whereas  $X_8$  is only average (5.3). Only  $X_6$  (Product Quality) is rated favorably (8.4). Thus, HBAT is definitely doing poorly with cluster 1 and needs improvement. Cluster 2 views HBAT more favorably than does cluster 1 with one big exception. HBAT performs slightly above average on four of the five variables ( $X_8$ ,  $X_{12}$ ,  $X_{15}$ , and  $X_{18}$ ) according to cluster 2. The score on  $X_{12}$  (5.7) is clearly the highest among all clusters. However, its rating on  $X_6$  is 6.1. This is by far the lowest rating on this variable across the three clusters. HBAT is therefore overall viewed slightly more favorably by cluster 2 than cluster 1, but has an issue with perceived product quality relative to the other groups. Cluster 3 customers view HBAT relatively favorably. Indeed, HBAT performs quite high on  $X_6$  (Product Quality) and the highest of all customer segments on  $X_{15}$  (New Products). Thus, HBAT may consider maintaining an emphasis on newness and innovativeness among customers in this group.

Criterion validity for the three-cluster solution was examined using the same approach as with the four-cluster solution. Variables  $X_{19}$ ,  $X_{20}$ ,  $X_{21}$ , and  $X_{22}$  were submitted to a MANOVA analysis as dependent variables, and the independent variable was cluster membership. The overall  $F$ -statistic for the MANOVA, as well as the univariate  $F$  statistics, were all significant, thus providing evidence of criterion validity.

The final task is to profile the three clusters so management can determine the characteristics of each cluster and target them with different strategies. Results of the cross-classification are provided in Table 4.15. As was determined earlier, significant chi-square values are observed for three of the five profile variables. Several patterns are evident. For example, cluster 2 consists entirely of customers from outside of the USA/North America (27 out of 27).

**Table 4.14** Means from K-Means Three-Cluster Solution

Variable	Mean Values Cluster Number:			Mean-Centered Values Cluster Number:				<i>F</i>	Sig
	1	2	3	1	2	3			
$X_6$ Product Quality	8.4	6.1	8.7	0.58	-1.70	0.91	79.78	0.000	
$X_8$ Technical Support	5.3	5.5	5.5	-0.03	0.12	0.16	0.16	0.851	
$X_{12}$ Salesforce Image	4.8	5.7	5.1	-0.31	0.61	-0.04	6.90	0.002	
$X_{15}$ New Products	4.0	5.3	6.7	-1.18	0.11	1.57	89.76	0.000	
$X_{18}$ Delivery Speed	3.7	4.1	4.0	-0.14	0.19	0.09	1.98	0.144	
Cluster sample sizes	44	27	27	44	27	27			

**Table 4.15** Cross-classifications from Three-Cluster Solution

Customer Characteristics		Number of Cases Per Customer			
		1	2	3	Total
$X_1$ Customer Type	Less than 1 year	13	10	7	30
	1 to 5 years	12	17	6	35
	More than 5 years	19	0	14	33
	Total ( $\chi^2 = 29.2; p < .001$ )	44	27	27	98
$X_2$ Industry Type	Magazine industry	20	14	17	51
	Newsprint industry	24	13	10	47
	Total ( $\chi^2 = 2.1; p = .36$ )	44	27	27	98
$X_3$ Firm Size	Small (0 to 499)	22	10	15	47
	Large (500+)	22	17	12	51
	Total ( $\chi^2 = 2.0; p = .37$ )	44	27	27	98
$X_4$ Region	USA/North America	25	0	14	39
	Outside North America	19	27	13	59
	Total ( $\chi^2 = 34.2, p < .001$ )	44	27	27	98
$X_5$ Distribution System	Indirect through broker	20	21	14	55
	Direct to customer	24	6	13	43
	Total ( $\chi^2 = 7.8, p < .05$ )	44	27	27	98

In contrast, clusters 1 and 3 are rather evenly split between customers from the USA/North America. Other differences indicate that cluster 2 customers are not among the customers who have been with HBAT the longest. These profiles also support the distinctiveness of the clusters on variables not used in the analysis at any prior point. Moreover, these findings can be used to develop different strategies for each customer cluster.

The question remains: Which cluster solution is best? Each solution, including the five-cluster solution, which was not discussed, has strengths and weaknesses. The four-cluster and five-cluster solutions provide more differentiation between the customers, and each cluster represents a smaller and more homogeneous set of customers. In contrast, the three-cluster solution is more parsimonious and likely easier and less costly for HBAT management to implement. So, ultimately, the question of which is best is not by determined by statistical results alone.

## A MANAGERIAL OVERVIEW OF THE CLUSTERING PROCESS

The cluster analyses (hierarchical and nonhierarchical) were successful in performing a market segmentation of HBAT customers. The process not only created homogeneous groupings of customers based on their perceptions of HBAT, but also found that these clusters met the tests of predictive validity and distinctiveness on additional sets of variables, which are all necessary for achieving practical significance. The segments represent quite different customer perspectives of HBAT, varying in both the types of variables that are viewed most positively as well as the magnitude of the perceptions.

One issue that can always be questioned is the selection of a “final” cluster solution. In this example, both the three-cluster and four-cluster solutions exhibited distinctiveness and strong relationships to the relevant outcomes. Moreover, the solutions provide a basic, but useful, delineation of customers that vary in perceptions, buying behavior, and demographic profile. The selection of the best cluster solution needs to involve all interested parties; in this example, the research team and management both have to provide input so a consensus can be reached.

Cluster analysis can be a very useful data-reduction technique. But its application is more an art than a science, and the technique can easily be abused or misapplied. Different similarity measures and different algorithms can and do affect the results. If the researcher proceeds cautiously, however, cluster analysis can be an invaluable tool in identifying latent patterns by suggesting useful groupings (clusters) of objects that are not discernible through other multivariate techniques. This chapter helps you to do the following:

**Define cluster analysis, its roles, and its limitations.** Cluster analysis is a group of multivariate methods whose primary purpose is to group objects based on the characteristics they possess. Cluster analysis classifies objects (e.g., respondents, products, or other entities) so that each object is very similar to others in the cluster based on a set of selected characteristics known as clustering variables. The resulting clusters of objects should exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity. If the process is successful, the objects within clusters will be close together when plotted geometrically, and different clusters will be far apart. Among the more common roles cluster analysis plays are: (1) data reduction of the type common when a researcher is faced with a large number of observations that can be meaningfully classified into groups or segments and (2) hypothesis generation where cluster analysis is used to develop hypotheses concerning the nature of the data or to examine previously stated hypotheses. The most common criticisms, and therefore limitations, of cluster analysis are: (1) it is descriptive, atheoretical, and non-inferential; (2) it will always create clusters, regardless of the “true” existence of any structure in the data; and (3) cluster solutions are not generalizable, because they are totally dependent upon the variables used as the basis for the similarity measure. Overcoming these criticisms is based as much on conceptual support for the objectives of cluster analysis, the clustering variables selected and the theoretical and practical significance used in selecting a final cluster solution.

**Identify types of research questions addressed by cluster analysis.** In forming homogeneous groups, cluster analysis can address any combination of three basic research questions: (1) taxonomy description (the most traditional use of cluster analysis has been for exploratory purposes and the formation of a taxonomy—an empirically-based classification of objects); (2) data simplification (by defining structure among the observations, cluster analysis also develops a simplified perspective by grouping observations for further analysis); and (3) relationship identification (with the clusters defined and the underlying structure of the data represented in the clusters, the researcher has a means of revealing relationships among the observations that typically is not possible with the individual observations).

**Understand how interobject similarity is measured.** Interobject similarity can be measured in a variety of ways. Three methods dominate applications of cluster analysis: distance measures, correlational measures, and association measures. Each method represents a particular perspective on similarity, dependent on both its objectives and type of data. Both the correlational and distance measures require metric data, whereas the association measures are for nonmetric data. Correlational measures are rarely used, because the emphasis in most applications of cluster analysis is on the objects' magnitudes, not the patterns of values. Distance measures are the most commonly used measures of similarity in cluster analysis. The distance measures represent similarity, because the proximity of observations to one another across the variables in the cluster variate. However, the term *dissimilarity* may be more appropriate for distance measures, because higher values represent more dissimilarity, not more similarity.

**Understand why different similarity measures are sometimes used.** Euclidean distance is the most commonly recognized measure of distance, and is many times referred to as *straight-line* distance. This concept is easily generalized to more than two variables. Squared (or absolute) Euclidean distance is the sum of the squared differences without taking the square root. City-block (Manhattan) distance is not based on Euclidean distance. Instead, it uses the sum of the absolute differences of the variables (i.e., the two sides of a right triangle rather than the hypotenuse). This procedure is the simplest to calculate, but may lead to invalid clusters if the clustering variables are highly correlated. Researchers sometimes run multiple cluster solutions using different distance measures to compare the results. Alternatively, correlation measures can be used to represent similarity when the researcher is more interested in patterns than in profiles based on similarity of cluster members.

**Understand the differences between hierarchical and nonhierarchical clustering techniques.** A wide range of partitioning procedures has been proposed for cluster analysis. The two most widely used procedures are hierarchical versus nonhierarchical. Hierarchical procedures involve a series of  $n - 1$  clustering decisions (where  $n$  equals the number of observations) that combine observations into a hierarchy or tree-like structure. In contrast to hierarchical methods, nonhierarchical procedures do not involve the tree-like construction process. Instead, they assign objects into clusters once the number of clusters to be formed is specified. For example, if a six-cluster solution is specified, then the resulting clusters are not just a combination of two clusters from the seven-cluster solution, but are based only on finding the best six-cluster solution. The process has two steps: (1) identifying starting points, known as cluster seeds, for each cluster, and (2) assigning each observation to one of the cluster seeds based on similarity within the group. Nonhierarchical cluster methods are often referred to as k-means. The emergence of both density-based and model-based approaches provide alternative methods when either the dataset or research question are less amenable to the hierarchical or nonhierarchical approaches.

**Know how to interpret results from cluster analysis.** Results from hierarchical cluster analysis are interpreted differently from nonhierarchical cluster analysis. Perhaps the most crucial issue for any researcher is determining the number of clusters. The researcher must select the cluster solution(s) that will best represent the data by applying a stopping rule. The fundamental principle underlying all of the stopping rules for hierarchical methods is to identify cluster solutions with more homogeneity/less heterogeneity than other possible cluster solutions. Since heterogeneity will always increase as the number of clusters decreases, the stopping rules look to identify “jumps” or a large increase in heterogeneity, which indicates that combining clusters at that stage of the process was accomplished by joining two quite different clusters. This then indicates that the prior solution where the clusters were separate is preferable. In hierarchical cluster analysis, the agglomeration schedule becomes crucial in determining these stopping rules. The researcher also must analyze cluster solutions for distinctiveness and the possibility of outliers, which would be identified by very small cluster sizes or by observations that join clusters late in the agglomeration schedule. Hierarchical cluster results, including the number of clusters and possibly the cluster seed points, can then be inputs to a nonhierarchical approach, where the critical question is how many clusters to form. Since the number of clusters is already established, focus is on interpretation and profiling the clusters on the clustering variables and on other variables, including validation (see next section for more discussion). The profiling stage involves describing the characteristics of each cluster to explain how they may differ on relevant dimensions. The procedure begins after the clusters are identified and typically involves the use of discriminant analysis or ANOVA. Clusters should be distinct and consistent with theory and can be explained.

**Follow the guidelines for cluster validation.** The most direct approach to validation is to cluster analyze separate samples, comparing the cluster solutions and assessing the correspondence of the results. The researcher also can attempt to establish some form of criterion or predictive validity. To do so, variable(s) not used to form the clusters but known to vary across the clusters are selected and compared. The comparisons can usually be performed with either MANOVA/ANOVA or a cross-classification table. The cluster solution should also be examined for stability. Nonhierarchical approaches are particularly susceptible to unstable results because of the different processes for selecting initial seed values as inputs. Thus, the data should be sorted into several different orders and the cluster solution redone with the clusters from the different solutions examined against one another for consistency.

What are the fundamental objectives that can be achieved through cluster analysis?

What are the basic stages in the application of cluster analysis?

What should the researcher consider when selecting a similarity measure to use in cluster analysis?

How does the researcher know whether to use hierarchical or nonhierarchical cluster techniques? Under which conditions would each approach be used?

What are the alternative approaches to the hierarchical and nonhierarchical approaches? When are they best employed?

How does a researcher decide the number of clusters to have in a hierarchical cluster solution?

How can researchers use graphical portrayals of the cluster procedure?

What is the difference between the interpretation stage and the profiling and validation stages?

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 Aggarwal, C. C., A. Hinneburg, and D. A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional spaces. In *ICDT Vol. 1*, pp. 420–34.
- 2 Aggarwal, Charu C., and Chandan K. Reddy. 2014. *Data Clustering: Algorithms and Applications*. Boca Raton, FL: CRC Press.
- 3 Aldenderfer, Mark S., and Roger K. Blashfield. 1984. *Cluster Analysis*. Thousand Oaks, CA: Sage.
- 4 Anderberg, M. 1973. *Cluster Analysis for Applications*. New York: Academic.
- 5 Baeza-Yates, R. A. 1992. Introduction to Data Structures and Algorithms Related to Information Retrieval. In *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates (eds.). Upper Saddle River, NJ: Prentice-Hall, pp. 13–27.
- 6 Bailey, Kenneth D. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques*. Thousand Oaks, CA: Sage.
- 7 Banfield, J. D., and A. E. Raftery. 1993. Model-Based Gaussian and non-Gaussian Clustering. *Biometrics* 49: 803–21.
- 8 Bellman, Richard Ernest (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.
- 9 Bensmail, H., G. Celeux, A. E. Raftery, and C. P. Robert. 1997. Inference in Model-Based Cluster Analysis. *Statistics and Computing* 7: 1–10.
- 10 Bock, H. H. 1985. On Some Significance Tests in Cluster Analysis. *Communication in Statistics* 3: 1–27.
- 11 Breckenridge, J. N. 2000. Validating Cluster Analysis: Consistent Replication and Symmetry. *Multivariate Behavioral Research* 35: 261–85.
- 12 Breckenridge, James N. (2000) Validating Cluster Analysis: Consistent Replication and Symmetry. *Multivariate Behavioral Research* 35: 261–85.
- 13 Davies, David L., and Donald W. Bouldin. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1 2: 224–7.
- 14 de Craen, Saskia, Jacques J. F. Commandeur, Laurence E. Frank, and Willem J. Heiser. 2006. Effects of Group Size and Lack of Sphericity on the Recovery of Clusters in K-means Cluster Analysis. *Multivariate Behavioral Research* 41: 127–45.
- 15 Dubes, R. C. 1987. How Many Clusters Are Best—An Experiment. *Pattern Recognition* 20: 645–63.
- 16 Dunn, J. 1974. Well Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics* 4: 95–104.
- 17 Dy, Jennifer G. and Carla E. Brodley. 2004. Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research* 5: 845–89.
- 18 Edelbrock, C. 1979. Comparing the Accuracy of Hierarchical Clustering Algorithms: The Problem of Classifying Everybody. *Multivariate Behavioral Research* 14: 367–84.
- 19 Ester, M., H. P. Kriegel, J. Sander, and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD'96* 34: 226–31.
- 20 Everitt, B., S. Landau, and M. Leese. 2001. *Cluster Analysis*, 4th ed. New York: Arnold Publishers.
- 21 Finch, H. 2005. Comparison of Distance Measures in Cluster Analysis with Dichotomous Data. *Journal of Data Science* 3: 85–100.
- 22 Fraley, C. and A. E. Raftery. 1998. How Many Clusters? Which Clustering Method? Answer via Model-Based Cluster Analysis. *The Computer Journal* 41: 578–88.
- 23 Fraley, C. and A. E. Raftery. 1999. MCLUST: Software for Model-Based Cluster Analysis. *Journal of Classification* 16: 297–306.
- 24 Green, P. E. 1978. *Analyzing Multivariate Data*. Hinsdale, IL: Holt, Rinehart and Winston.
- 25 Green, P. E., and J. Douglas Carroll. 1978. *Mathematical Tools for Applied Multivariate Analysis*. New York: Academic Press.
- 26 Han, J., and M. Kamber. 2001. *Data Mining Concept and Technology*. London: Academic Press.
- 27 Hartigan, J. A. 1975. *Clustering Algorithms* (Vol. 209). New York: Wiley.

- 28 Hartigan, J. A. 1985. Statistical Theory in Clustering. *Journal of Classification* 2: 63–76.
- 29 Hartigan, J. A., and M. A. Wong. 1979. Algorithm AS 136: A k-means Clustering Algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 28: 100–8.
- 30 Houle, M., H. P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. 2010. Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? In *Scientific and Statistical Database Management*. Berlin: Springer, pp. 482–500.
- 31 Jain, A. K., and R. C. Dubes. 1988. *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall.
- 32 Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. Data Clustering: A Review. *ACM Computing Surveys* 31: 264–323.
- 33 Jardine, N., and R. Sibson. 1975. *Mathematical Taxonomy*. New York: Wiley.
- 34 Ketchen, D. J., and C. L. Shook. 1996. The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal* 17: 441–58.
- 35 Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek. 2009. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. *ACM Transactions on Knowledge Discovery from Data* 3:article 1.
- 36 Magidson, J., and J. Vermunt. 2002. Latent Class Models for Clustering: A Comparison with K-means. *Canadian Journal of Marketing Research* 20: 36–43.
- 37 McIntyre, R. M., and R. K. Blashfield. 1980. A Nearest-Centroid Technique for Evaluating the Minimum-Variance Clustering Procedure. *Multivariate Behavioral Research* 15: 225–38.
- 38 McLachlan, G. J., and D. Peel. 1996. An algorithm for unsupervised learning via normal mixture models. In D. L. Dowe, K. B. Korb, and J. J. Oliver (eds.), *Information, Statistics and Induction in Science*. Singapore: World Scientific, pp. 354–63.
- 39 McLachlan, G. J., and K. E. Basford. 1988. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.
- 40 Milligan, G. 1980. An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika* 45: 325–42.
- 41 Milligan, Glenn W., and Martha C. Cooper. 1985. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* 50(2): 159–79.
- 42 Morrison, D. 1967. Measurement Problems in Cluster Analysis. *Management Science* 13: 775–80.
- 43 Nagy, G. 1968. State of the Art in Pattern Recognition. *Proceedings of the IEEE* 56: 836–62.
- 44 Overall, J. 1964. Note on Multivariate Methods for Profile Analysis. *Psychological Bulletin* 61: 195–8.
- 45 Punj, G., and D. Stewart. 1983. Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research* 20: 134–48.
- 46 Raftery, A. E., and N. Dean. 2006. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association* 101: 168–78.
- 47 Rohlff, F. J. 1970. Adaptive Hierarchical Clustering Schemes. *Systematic Zoology* 19: 58.
- 48 Rousseeuw, Peter J. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53–65.
- 49 Sarle, W. S. 1983. *Cubic Clustering Criterion*, SAS Technical Report A-108. Cary, NC: SAS Institute, Inc.
- 50 Schaninger, C. M., and W. C. Bass. 1986. Removing Response-Style Effects in Attribute-Determinance Ratings to Identify Market Segments. *Journal of Business Research* 14: 237–52.
- 51 Shephard, R. 1966. Metric Structures in Ordinal Data. *Journal of Mathematical Psychology* 3: 287–315.
- 52 Sneath, P. H. A., and R. R. Sokal. 1973. *Numerical Taxonomy*. San Francisco: Freeman Press.
- 53 Steinley, D. and M. J. Brusco. 2008. Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures. *Psychometrika* 73: 125.
- 54 Steinley, Douglas, and Michael J. Brusco. 2008. A New Variable Weighting and Selection Procedure for K-means Cluster Analysis. *Multivariate Behavioral Research* 43: 77–108.

# Dependence techniques – metric outcomes

**Multiple Regression Analysis**

**MANOVA: Extending ANOVA**

## SECTION III

### OVERVIEW

**S**ections 3 and 4 deal with what many would term the essence of multivariate analysis: dependence techniques. As noted in Chapter 1, dependence techniques are based on the use of a set of independent variables to predict and explain one or more dependent variables. The researcher, whether faced with dependent variables of a metric or nonmetric nature, has a variety of dependence methods available to assist in the process of relating independent variables to dependent variables. Given the multivariate nature of these methods, all of the dependence techniques accommodate multiple independent variables while also allowing multiple dependent variables in certain situations. Thus, the researcher has a set of techniques that should allow for the analysis of almost any type of research question involving a dependence relationship. They also provide the opportunity not only for increased prediction capability, but also for enhanced explanation of the dependent variable's relationship to the independent variables. Explanation becomes increasingly important as the research questions begin to address issues concerning how the relationship between independent and dependent variables operates.

For readers familiar with past editions or looking for a discussion of conjoint analysis or canonical correlation, we refer you to the text available in the online resources at the text's websites. Conjoint Analysis presents us with a technique unlike any of the other multivariate methods in that the researcher determines the values of the independent nonmetric variables in a quasi-experimental fashion. Once designed, the respondent provides information regarding only the dependent variable. Although it places more responsibility on the researcher, conjoint analysis provides a powerful tool for understanding complex decision processes. Canonical Correlation is the most generalized form of multivariate analysis, which accommodates multiple dependent and independent variables. In situations in which variates exist for both the dependent and independent variables, canonical correlation provides a flexible method for both prediction and explanation.

## CHAPTERS IN SECTION III

Section 3 covers the two primary dependence techniques which are based on metric outcomes: multiple regression and analysis of variance/multivariate analysis of variance. Multiple Regression Analysis (Chapter 5) focuses on what is perhaps the most fundamental of all multivariate techniques and a building block for our discussion of the other dependence methods. Whether assessing the conformity to underlying statistical assumptions, measuring predictive accuracy, or interpreting the variate of independent variables, the issues discussed in Chapter 5 will be seen as crucial in many of the other techniques as well. In Chapter 6, Multivariate Analysis of Variance, the discussion differs in several ways from the prior techniques; it is suited to the analysis of multiple metric dependent variables and nonmetric independent variables. Although this technique is a direct extension of simple analysis of variance, the multiple metric dependent variables make both prediction and explanation more difficult.

This section provides the researcher with exposure to the dependence techniques available for addressing metric outcomes, each suited to a specific task and relationship. When you complete this section, the issues regarding selecting from these methods should be apparent, and you should feel comfortable in selecting from these techniques and analyzing their results.

# 5

# Multiple Regression Analysis

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

Determine when regression analysis is the appropriate statistical tool in analyzing a problem.

Understand how regression helps us make both predictions and explanations using the least squares concept.

Use dummy variables with an understanding of their interpretation.

Be aware of the assumptions underlying regression analysis and how to assess them.

Understand the implications of managing the variate and its implications on the regression results

Address the implications of user- versus software-controlled variable selection and explain the options available in software controlled variable selection.

Interpret the results of regression and variable importance, especially in light of multicollinearity.

Apply the diagnostic procedures necessary to assess influential observations.

Understand the benefits gained from the extended forms of regression, namely multi-level models and panel models

## Chapter Preview

This chapter describes multiple regression analysis as it is used to solve important research problems in both academia and organizational settings. Regression analysis is by far the most widely used and versatile dependence technique, equally applicable to research questions involving either prediction or explanation. Applications of the method range from the most general problems to the most specific, in each instance relating a factor (or factors) to a specific outcome. From a prediction perspective, regression analysis is the foundation for business forecasting models, ranging from the econometric models that predict the national economy based on certain inputs (income levels, business investment, etc.) to models of a firm's performance in a market if a specific marketing strategy is followed. In an explanatory function, regression models are also used to study how consumers make decisions or form impressions and attitudes. Other applications include evaluating the determinants of effectiveness for a program (e.g., what factors aid in maintaining quality) and determining the feasibility of a new product or the expected return for a new stock issue. Even though these examples illustrate only a small subset of all applications, they demonstrate that regression analysis is a powerful analytical tool designed to explore all types of dependence relationships.

Multiple regression analysis is a statistical technique within the general linear model used to analyze the relationship between a single dependent variable and several independent variables. As noted in Chapter 1, its basic formulation is:

$$\begin{array}{rcl} Y_1 & = & X_1 + X_2 + X_n \\ & \text{(metric)} & \text{(metric)} \end{array}$$

This chapter presents guidelines for judging the appropriateness of multiple regression for various types of problems. Suggestions are provided for interpreting the results of its application from a managerial as well as a statistical viewpoint. Possible transformations of the data to remedy violations of various model assumptions are examined, along with a series of diagnostic procedures that identify observations with particular influence on the results. Particular emphasis is placed on interpretation of the regression results for explanatory purposes with a focus on addressing the complicating factors of multicollinearity. Readers who are already knowledgeable about multiple regression procedures can skim the early portions of the chapter, but for those who are less familiar with the subject, this chapter provides a valuable background for the study of multivariate data analysis. Finally, multiple regression is a fundamental building block for understanding partial least squares structural equation modeling described in Chapter 13.

## Key Terms

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

**Adjusted coefficient of determination (adjusted  $R^2$ )** Modified measure of the *coefficient of determination* that takes into account the number of independent variables included in the regression equation and the sample size. Although the addition of independent variables will always cause the coefficient of determination to rise, the adjusted coefficient of determination may fall if the added independent variables have little explanatory power or if the *degrees of freedom* become too small. This statistic is quite useful for comparison between equations with different numbers of independent variables, differing sample sizes, or both.

**All-possible-subsets regression** Method of selecting the variables for inclusion in the regression model that considers all possible combinations of the independent variables. For example, if the researcher specifies four potential independent variables, this technique would estimate all possible regression models with one, two, three, and four variables. The technique would then identify the model(s) with the best predictive accuracy.

**Atomistic fallacy** Incorrect conclusions about group level behavior based on individual behavior. The reverse of *ecological fallacy*.

**Backward elimination** Method of selecting variables for inclusion in the regression model that starts by including all independent variables in the model and then eliminating those variables not making a significant contribution to prediction.

**Beta coefficient** Standardized regression coefficient (see *standardization*) that allows for a direct comparison between coefficients as to their relative explanatory power of the dependent variable. Whereas *regression coefficients* are expressed in terms of the units of the associated variable, thereby making comparisons inappropriate, beta coefficients use standardized data and can be directly compared.

**Coefficient of determination ( $R^2$ )** Measure of the proportion of the variance of the dependent variable about its mean that is explained by the independent, or predictor, variables. The coefficient can vary between 0 and 1. If the regression model is properly applied and estimated, the researcher can assume that the higher the value of  $R^2$ , the greater the explanatory power of the regression equation, and therefore the better the prediction of the dependent variable.

**Collinearity** Expression of the relationship between two (collinearity) or more (multicollinearity) independent variables. Two independent variables are said to exhibit complete collinearity if their correlation coefficient is 1, and complete lack of collinearity if their correlation coefficient is 0. *Multicollinearity* occurs when any single independent variable is highly correlated with a set of other independent variables. An extreme case of collinearity/multicollinearity is *singularity*, in which an independent variable is perfectly predicted (i.e., correlation of 1.0) by another independent variable (or more than one).

**Condition index** Measure of the relative amount of variance associated with an eigenvalue. A large condition index indicates a high degree of *collinearity*.

**Context** Any external factor outside the unit of analysis that not only impacts the outcome of multiple individuals, but also creates differences between individuals in separate contexts and foster dependencies between the individuals in a single context.

**Cook's distance (D)** Summary measure of the influence of a single case (observation) based on the total changes in all other residuals when the case is deleted from the estimation process. Large values (usually greater than 1) indicate substantial influence by the case in affecting the estimated regression coefficients.

**Correlation coefficient (r)** Coefficient that indicates the strength of the association between any two metric variables. The sign (+ or -) indicates the direction of the relationship. The value can range from +1 to -1, with +1 indicating a perfect positive relationship, 0 indicating no relationship, and -1 indicating a perfect negative or reverse relationship (as one variable grows larger, the other variable grows smaller).

**COVRATIO** Measure of the influence of a single observation on the entire set of estimated regression coefficients. A value close to 1 indicates little influence. If the COVRATIO value minus 1 is greater than  $\pm 3p/n$  (where  $p$  is the number of independent variables + 1, and  $n$  is the sample size), the observation is deemed to be influential based on this measure.

**Criterion variable (Y)** See *dependent variable*.

**Degrees of freedom (df)** Value calculated from the total number of observations minus the number of estimated *parameters*.

These parameter estimates are restrictions on the data because, once made, they define the population from which the data are assumed to have been drawn. For example, in estimating a regression model with a single independent variable, we estimate two parameters, the *intercept* ( $b_0$ ) and a *regression coefficient* for the independent variable ( $b_1$ ). In estimating the random error, defined as the sum of the *prediction errors* (actual minus predicted dependent values) for all cases, we would find  $(n - 2)$  degrees of freedom. Degrees of freedom provide a measure of how restricted the data are to reach a certain level of prediction. If the number of degrees of freedom is small, the resulting prediction may be less generalizable because all but a few observations were incorporated in the prediction. Conversely, a large degrees-of-freedom value indicates the prediction is fairly robust with regard to being representative of the overall sample of respondents.

**Deleted residual** Process of calculating *residuals* in which the influence of each observation is removed when calculating its residual. This is accomplished by omitting the  $i$ th observation from the regression equation used to calculate its predicted value.

**Dependent variable (Y)** Variable being predicted or explained by the set of independent variables.

**DFBETA** Measure of the change in a regression coefficient when an observation is omitted from the regression analysis. The value of DFBETA is in terms of the coefficient itself; a standardized form (SDFBETA) is also available. No threshold limit can be established for DFBETA, although the researcher can look for values substantially different from the remaining observations to assess potential influence. The SDFBETA values are scaled by their standard errors, thus supporting the rationale for cut-offs of 1 or 2, corresponding to confidence levels of .10 or .05, respectively.

**DFFIT** Measure of an observation's impact on the overall model fit, which also has a standardized version (SDFFIT). The best rule of thumb is to classify as influential any standardized values (SDFFIT) that exceed  $2/\sqrt{p/n}$ , where  $p$  is the number of independent variables + 1 and  $n$  is the sample size. There is no threshold value for the DFFIT measure.

**Dummy variable** Independent variable used to account for the effect that different levels of a nonmetric variable have in predicting the dependent variable. To account for  $L$  levels of a nonmetric independent variable,  $L - 1$  dummy variables are needed. For example, gender is measured as male or female and could be represented by two dummy variables,  $X_1$  and  $X_2$ . When the respondent is male,  $X_1 = 1$  and  $X_2 = 0$ . Likewise, when the respondent is female,  $X_1 = 0$  and  $X_2 = 1$ . However, when  $X_1 = 1$ , we know that  $X_2$  must equal 0. Thus, we need only one variable, either  $X_1$  or  $X_2$ , to represent gender. We need not include both variables because one is perfectly predicted by the other (a *singularity*) and the *regression coefficients* cannot be estimated. If a variable has three levels, only two dummy variables are needed. Thus, the number of dummy variables is one less than the number of levels of the nonmetric variable. The two most common methods of determining the values of the dummy values are *indicator coding* and *effects coding*.

**Ecological fallacy** Incorrectly drawing conclusions about individual behavior from relationships based on aggregated data.

**Effects coding** Method for specifying the *reference category* for a set of *dummy variables* in which the reference category receives a value of -1 across the set of dummy variables. In our example of dummy variable coding for gender, we coded the dummy variable as either 1 or 0. But with effects coding, the value of -1 is used instead of 0. With this type of coding, the coefficients for the dummy variables become group deviations on the dependent variable from the mean of the dependent variable across all groups. Effects coding contrasts with *indicator coding*, in which the reference category is given the value of zero across all dummy variables and the coefficients represent group deviations on the dependent variable from the reference group.

**Fixed effect** Default estimation of effects in regression where a point estimate with no variation is made for a parameter estimate (either an intercept or coefficient).

**Forward addition** Method of selecting variables for inclusion in the regression model by starting with no variables in the model and then adding one variable at a time based on its contribution to prediction.

**Functional relationship** Dependence relationship that has no error in prediction. Also see *statistical relationship*.

**Hat matrix** Matrix that contains values for each observation on the diagonal, known as *hat values*, which represent the impact of the observed dependent variable on its predicted value. If all cases have equal influence, each would have a value of  $p/n$ , where  $p$  equals the number of independent variables + 1, and  $n$  is the number of cases. If a case has no influence, its value would be  $-1/n$ , whereas total domination by a single case would result in a value of  $(n - 1)/n$ . Values exceeding  $2p/n$  for larger samples, or  $3p/n$  for smaller samples ( $n \leq 30$ ), are candidates for classification as influential observations.

**Hat value** See *hat matrix*.

**Heteroscedasticity** See *homoscedasticity*.

**Hierarchical data structure** Observations which have a natural nesting effect created by *contexts*.

**Homoscedasticity** Description of data for which the variance of the error terms ( $\epsilon$ ) appears constant over the range of values of an independent variable. The assumption of equal variance of the population error  $\epsilon$  (where  $\epsilon$  is estimated from the sample value  $e$ ) is critical to the proper application of linear regression. When the error terms have increasing or modulating variance, the data are said to be *heteroscedastic*. The discussion of *residuals* in this chapter further illustrates this point.

**Independent variable** Variable(s) selected as predictors and potential explanatory variables of the dependent variable.

**Indicator coding** Method for specifying the *reference category* for a set of *dummy variables* where the reference category receives a value of 0 across the set of dummy variables. The *regression coefficients* represent the group differences in the dependent variable from the reference category. Indicator coding differs from *effects coding*, in which the reference category is given the value of -1 across all dummy variables and the regression coefficients represent group deviations on the dependent variable from the overall mean of the dependent variable.

**Influential observation** An observation that has a disproportionate influence on one or more aspects of the regression estimates. This influence may be based on extreme values of the independent or dependent variables, or both. Influential observations can either be “good,” by reinforcing the pattern of the remaining data, or “bad,” when a single or small set of cases unduly affects the regression estimates. It is not necessary for the observation to be an *outlier*, although many times outliers can be classified as influential observations as well.

**Intercept ( $b_0$ )** Value on the *Y* axis (dependent variable axis) where the line defined by the regression equation  $Y = b_0 + b_1X_1$  crosses the axis. It is described by the constant term  $b_0$  in the regression equation. In addition to its role in prediction, the intercept may have a managerial interpretation. If the complete absence of the independent variable has meaning, then the intercept represents that amount. For example, when estimating sales from past advertising expenditures, the intercept represents the level of sales expected if advertising is eliminated. But in many instances the constant has only predictive value because in no situation are all independent variables absent. An example is predicting product preference based on consumer attitudes. All individuals have some level of attitude, so the intercept has no managerial use, but it still aids in prediction.

**Intraclass correlation (ICC)** Measure of the degree of dependence among individuals within a higher-level grouping.

**Least squares** Estimation procedure used in simple and multiple regression whereby the *regression coefficients* are estimated so as to minimize the total sum of the squared *residuals*.

**Level-1** The lowest level in a *hierarchical data structure*, most often associated with individuals.

**Level-2** The first level of a *hierarchical data structure* where observations from *Level-1* are grouped together because of a *context* present at Level-2. Common examples are students (Level-1) within classrooms (Level-2).

**Leverage points** Type of *influential observation* defined by one aspect of influence termed *leverage*. These observations are substantially different on one or more independent variables, so that they affect the estimation of one or more *regression coefficients*.

**Linearity** Term used to express the concept that the model possesses the properties of additivity and homogeneity. In a simple sense, linear models predict values that fall in a straight line by having a constant unit change (slope) of the dependent variable for a constant unit change of the *independent variable*. In the population model  $Y = b_0 + b_1X_1 + \epsilon$ , the effect of changing  $X_1$  by a value of 1.0 is to add  $b_1$  (a constant) units of  $Y$ .

**Mahalanobis distance ( $D^2$ )** Measure of the uniqueness of a single observation based on differences between the observation's values and the mean values for all other cases across all independent variables. The source of influence on regression results is for the case to be quite different on one or more predictor variables, thus causing a shift of the entire regression equation.

**Measurement error** Degree to which the data values do not truly measure the characteristic being represented by the variable. For example, when asking about total family income, many sources of measurement error (e.g., reluctance to answer full amount, error in estimating total income) make the data values imprecise.

**Mediation** The effect of an independent variable “works through” an intervening or mediating variable.

**Moderator effect** Effect in which a third independent variable (the moderator variable) causes the relationship between a dependent/independent variable pair to change, depending on the value of the moderator variable. It is also known as an interactive effect and is similar to the interaction effect seen in analysis of variance methods.

**Multicollinearity** See *collinearity*.

**Multilevel model (MLM)** Extension of regression analysis that allows for the incorporation of both individual (*Level-1*) and contextual (*Level-2*) effects with the appropriate statistical treatment.

**Multiple regression** Regression model with two or more independent variables.

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

**Null plot** Plot of *residuals* versus the predicted values that exhibits a random pattern. A null plot is indicative of no identifiable violations of the assumptions underlying regression analysis.

**Outlier** In strict terms, an observation that has a substantial difference between the actual value for the dependent variable and the predicted value. Cases that are substantially different with regard to either the dependent or independent variables are often termed *outliers* as well. In all instances, the objective is to identify observations that are inappropriate representations of the population from which the sample is drawn, so that they may be discounted or even eliminated from the analysis as unrepresentative.

**Panel analysis** See *panel models*.

**Panel models** Regression-based analytical technique designed to handle cross-sectional analyses of longitudinal or time-series data.

**Parameter** Quantity (measure) characteristic of the population. For example,  $\mu$  and  $\sigma^2$  are the symbols used for the population parameters mean ( $\mu$ ) and variance ( $\sigma^2$ ). They are typically estimated from sample data in which the arithmetic average of the sample is used as a measure of the population average and the variance of the sample is used to estimate the variance of the population.

**Part correlation** Value that measures the strength of the relationship between a dependent and a single independent variable when the predictive effects of the other independent variables in the regression model are removed. The objective is to portray the unique predictive effect due to a single independent variable among a set of independent variables. Differs from the *partial correlation coefficient*, which is concerned with incremental predictive effect.

**Partial correlation coefficient** Value that measures the strength of the relationship between the criterion or dependent variable and a single independent variable when the effects of the other independent variables in the model are held constant. For example,  $r_{YX_2|X_1}$  measures the variation in  $Y$  associated with  $X_2$  when the effect of  $X_1$  on both  $X_2$  and  $Y$  is held constant. This value is used in sequential variable selection methods of regression model estimation (e.g., *stepwise*, *forward addition*, or *backward elimination*) to identify the independent variable with the greatest incremental predictive power beyond the independent variables already in the regression model.

**Partial F (or t) values** The partial  $F$  test is simply a statistical test for the additional contribution to prediction accuracy of a variable above that of the variables already in the equation. When a variable ( $X_a$ ) is added to a regression equation after other variables are already in the equation, its contribution may be small even though it has a high correlation with the dependent variable. The reason is that  $X_a$  is highly correlated with the variables already in the equation. The partial  $F$  value is calculated for all variables by simply pretending that each, in turn, is the last to enter the equation. It gives the additional contribution of each variable above all others in the equation. A low or insignificant partial  $F$  value for a variable not in the equation indicates its low or insignificant contribution to the model as already specified. A  $t$  value may be calculated instead of  $F$  values in all instances, with the  $t$  value being approximately the square root of the  $F$  value.

**Partial regression plot** Graphical representation of the relationship between the dependent variable and a single independent variable. The scatterplot of points depicts the partial correlation between the two variables, with the effects of other independent variables held constant (see *partial correlation coefficient*). This portrayal is particularly helpful in assessing the form of the relationship (linear versus nonlinear) and the identification of *influential observations*.

**Polynomial** Transformation of an independent variable to represent a curvilinear relationship with the dependent variable. By including a squared term ( $X^2$ ), a single inflection point is estimated. A cubic term estimates a second inflection point. Additional terms of a higher *power* can also be estimated.

**Power** Probability that a significant relationship will be found if it actually exists. Complements the more widely used *significance level alpha* ( $\alpha$ ).

**Prediction error** Difference between the actual and predicted values of the dependent variable for each observation in the sample (see *residual*).

**Predictor variable ( $X_n$ )** See *independent variable*.

**PRESS statistic** Validation measure obtained by eliminating each observation one at a time and predicting this dependent value with the regression model estimated from the remaining observations.

**Psychologicistic fallacy** The failure to acknowledge the impact of *contexts* and group effects on relationships at the individual level.

**Random effect** Makes an estimate of the variability or distribution of a parameters across a set of groups. Method of quantifying the variability of a parameter (intercept or coefficient) within a group by a form of pooling across groups.

**Reference category** The omitted level of a nonmetric variable when a *dummy variable* is formed from the nonmetric variable.

**Regression coefficient ( $b_n$ )** Numerical value of the parameter estimate directly associated with an independent variable; for example, in the model  $Y = b_0 + b_1X_1$  the value  $b_1$  is the regression coefficient for the variable  $X_1$ . The regression coefficient represents the amount of change in the dependent variable for a one-unit change in the independent variable. In the multiple predictor model (e.g.,  $Y = b_0 + b_1X_1 + b_2X_2$ ), the regression coefficients are partial coefficients because each takes into account not only the relationships between  $Y$  and  $X_1$  and between  $Y$  and  $X_2$ , but also between  $X_1$  and  $X_2$ . The coefficient is not limited in range, because it is based on both the degree of association and the scale units of the independent variable. For instance, two variables with the same association to  $Y$  would have different coefficients if one independent variable was measured on a 7-point scale and another was based on a 100-point scale.

**Regression coefficient variance-decomposition matrix** Method of determining the relative contribution of each *eigenvalue* to each estimated coefficient. If two or more coefficients are highly associated with a single eigenvalue (*condition index*), an unacceptable level of *multicollinearity* is indicated.

**Regression variate** Linear combination of weighted independent variables used collectively to predict the dependent variable.

**Residual (e or  $\varepsilon$ )** Error in predicting our sample data. Seldom will our predictions be perfect. We assume that random error will occur, but we assume that this error is an estimate of the true random error in the population ( $\varepsilon$ ), not just the error in prediction for our sample ( $e$ ). We assume that the error in the population we are estimating is distributed with a mean of 0 and a constant (*homoscedastic*) variance.

**Sampling error** The expected variation in any estimated parameter (*intercept* or *regression coefficient*) that is due to the use of a sample rather than the population. Sampling error is reduced as the sample size is increased and is used to statistically test whether the estimated parameter differs from zero.

**SDFBETA** See *DFBETA*.

**SDFFIT** See *DFFIT*.

**Significance level (alpha  $\alpha$ )** Commonly referred to as the level of statistical significance, the significance level represents the probability the researcher is willing to accept that the estimated coefficient is classified as different from zero when it actually is not. This is also known as Type I error. The most widely used level of significance is .05, although researchers use levels ranging from .01 (more demanding) to .10 (less conservative and easier to find significance).

**Simple regression** Regression model with a single independent variable, also known as bivariate regression.

**Singularity** The extreme case of *collinearity* or *multicollinearity* in which an independent variable is perfectly predicted (a correlation of  $\pm 1.0$ ) by one or more independent variables. Regression models cannot be estimated when a singularity exists. The researcher must omit one or more of the independent variables involved to remove the singularity.

**Specification error** Error in predicting the dependent variable caused by excluding one or more relevant independent variables. This omission can bias the estimated coefficients of the included variables as well as decrease the overall predictive power of the regression model.

**Standard error** Expected distribution of an estimated regression coefficient. The standard error is similar to the standard deviation of any set of data values, but instead denotes the expected range of the coefficient across multiple samples of the data. It is useful in statistical tests of significance that test to see whether the coefficient is significantly different from zero (i.e., whether the expected range of the coefficient contains the value of zero at a given level of confidence). The *t* value of a *regression coefficient* is the coefficient divided by its standard error.

**Standard error of the estimate (SE<sub>E</sub>)** Measure of the variation in the predicted values that can be used to develop confidence intervals around any predicted value. It is similar to the standard deviation of a variable around its mean, but instead is the expected distribution of predicted values that would occur if multiple samples of the data were taken.

**Standardization** Process whereby the original variable is transformed into a new variable with a mean of 0 and a standard deviation of 1. The typical procedure is to first subtract the variable mean from each observation's value and then divide by the standard deviation. When all the variables in a *regression variate* are standardized, the  $b_0$  term (the *intercept*) assumes a value of 0 and the *regression coefficients* are known as *beta coefficients*, which enable the researcher to compare directly the relative effect of each independent variable on the dependent variable.

**Standardized residual** Rescaling of the *residual* to a common basis by dividing each residual by the standard deviation of the residuals. Thus, standardized residuals have a mean of 0 and standard deviation of 1. Each standardized residual value can now be viewed in terms of standard errors in middle to large sample sizes. This provides a direct means of identifying outliers as those with values above 1 or 2 for confidence levels of .10 and .05, respectively.

**Statistical relationship** Relationship based on the correlation of one or more independent variables with the dependent variable. Measures of association, typically correlations, represent the degree of relationship because there is more than one value of the dependent variable for each value of the independent variable. Also see *functional relationship*.

**Stepwise estimation** Method of selecting variables for inclusion in the regression model that starts by selecting the best predictor of the dependent variable. Additional independent variables are selected in terms of the incremental explanatory power they can add to the regression model. Independent variables are added as long as their *partial correlation coefficients* are statistically significant. Independent variables may also be dropped if their predictive power drops to a nonsignificant level when another independent variable is added to the model.

**Studentized residual** A commonly used variant of the *standardized residual*. It differs from other methods in how it calculates the standard deviation used in *standardization*. To minimize the effect of any observation on the standardization process, the standard deviation of the residual for observation  $i$  is computed from regression estimates omitting the  $i$ th observation in the calculation of the regression estimates.

**Sum of squared errors (SS<sub>E</sub>)** Sum of the squared prediction errors (*residuals*) across all observations. It is used to denote the variance in the dependent variable not yet accounted for by the regression model. If no independent variables are used for prediction, it becomes the squared errors using the mean as the predicted value and thus equals the *total sum of squares*.

**Sum of squares regression (SS<sub>R</sub>)** Sum of the squared differences between the mean and predicted values of the dependent variable for all observations. It represents the amount of improvement in explanation of the dependent variable attributable to the independent variable(s).

**Suppression effect** The instance in which the expected relationships between independent and dependent variables are hidden or suppressed when viewed in a bivariate relationship. When additional independent variables are entered, the *multicollinearity* removes “unwanted” shared variance and reveals the “true” relationship.

**Tolerance** Commonly used measure of *collinearity* and *multicollinearity*. The tolerance of variable  $i$  ( $\text{TOL}_i$ ) is  $1 - R^2 i$ , where  $R^2 i$  is the coefficient of determination for the prediction of variable  $i$  by the other independent variables in the *regression variate*. As the tolerance value grows smaller, the variable is more highly predicted by the other independent variables (collinearity).

**Total sum of squares (SS<sub>T</sub>)** Total amount of variation that exists to be explained by the independent variables. This baseline value is calculated by summing the squared differences between the mean and actual values for the dependent variable across all observations.

**Transformation** A variable may have an undesirable characteristic, such as non-normality, that detracts from the ability of the *correlation coefficient* to represent the relationship between it and another variable. A transformation, such as taking the logarithm or square root of the variable, creates a new variable and eliminates the undesirable characteristic, allowing for a better measure of the relationship. Transformations may be applied to either the dependent or independent variables, or both. The need and specific type of transformation may be based on theoretical reasons (such as transforming a known nonlinear relationship) or empirical reasons (identified through graphical or statistical means).

**Variance inflation factor (VIF)** Indicator of the effect that the other independent variables have on the standard error of a *regression coefficient*. The variance inflation factor is directly related to the *tolerance* value ( $\text{VIF}_i = 1/\text{TOL}_i$ ). Large VIF values also indicate a high degree of *collinearity* or *multicollinearity* among the independent variables.

## What Is Multiple Regression Analysis?

Multiple regression analysis is a statistical technique that can be used to analyze the relationship between a single **dependent (criterion) variable** and several **independent (predictor) variables**. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the single dependent value selected by the researcher. Each independent variable is weighted by the regression analysis procedure to ensure maximal prediction from the set of independent variables. The weights denote the relative contribution of the independent variables to the overall prediction and facilitate interpretation as to the influence of each variable in making the prediction, although correlation among the independent variables complicates the interpretative process. Thus, it can equally achieve the objectives of prediction or explanation. The set of weighted independent variables forms the **regression variate**, a linear combination of the independent variables that best predicts the dependent variable (Chapter 1 contains a more detailed explanation of the variate). The regression variate, also referred to as the *regression equation* or *regression model*, is the most widely known example of a variate among the multivariate techniques.

As noted in Chapter 1, multiple regression analysis is a dependence technique. Thus, to use it you must be able to divide the variables into dependent and independent variables. Regression analysis is also a statistical tool that should be used only when both the dependent and independent variables are metric. However, under certain circumstances it is possible to include nonmetric data either as independent variables (by transforming either ordinal or nominal data with dummy variable coding) or the dependent variable (by the use of a binary measure in the specialized technique of logistic regression, see Chapter 8). In summary, to apply multiple regression analysis: (1) the data must be metric or appropriately transformed, and (2) before deriving the regression equation, the researcher must decide which variable is to be dependent and which remaining variables will be independent.

## Multiple Regression in the Era of Big Data

Multiple regression has without dispute been the dominant analytical technique for researchers in the era of scientific explanation and the reliance on quantitative methods. Yet one question faces all researchers—Have the requirements for Big Data analytics caused researchers to move on to methods other than multiple regression? While in this chapter we will discuss all of the abilities of multiple regression to accommodate many types of variables and relationships, it is still a model based on linear relationships among metric variables. And as we have discussed in Chapter 1 and

other chapters, the scope and variety of variables now available within Big Data may place limitations on multiple regression as the primary analytical tool. To best address this question, we need to focus on the purpose of the analysis: prediction versus explanation. It is within these two objectives that we see the benefits and pitfalls of multiple regression best addressed.

The emphasis on data-driven decision-making, as discussed in Chapter 1, has placed an emphasis on predictive models in this era of Big Data. With so many decisions now being automated on the online environment and the timeliness required of decision-making, there is a need for highly predictive models, even if they are less helpful in explanation. This is the domain of the data mining/algorithmic model discussed in Chapter 1, with many of the machine learning models such as neural networks, decision trees and support vector machines being prime examples. Their emphasis is on prediction and they truly represent in many instances, such as neural networks, the “black box” approach to modeling. As such, their ability is to develop unique and complicated model forms that make explanation more difficult yet still achieve the purposes of prediction. So in many of these domains multiple regression does have limited applications.

But multiple regression is a primary example of the opposite type of model, the statistical/data model, which is oriented toward confirmation of a proposed model with an emphasis on explanation. And this emphasis on explanation makes multiple regression very well suited for several key areas of analytics. The first is forecasting, where an understanding of the basic “causes” of a process are necessary before projections into the future are made. Multiple regression and its variants of time-series modeling and structural equation modeling provide a framework for the in-depth understanding of the process being investigated. This is critical for forecasting where the model is expected to extrapolate into future conditions that may have not been encountered in the past. A second area of analytics extremely well suited for multiple regression are the academic and managerial areas. In academic research, knowledge creation is the fundamental objective and thus the statistical/data model will always hold prominence. And with managerial areas, organizations need to understand how to “manage” the process. It is very hard to incentivize a manager on increasing customer satisfaction or improving firm performance if there are no objective models that provide the insights into how this is done. This is also the role of multiple regression, to provide an objective means of quantifying the impact of potential factors on a specified outcome.

The era of Big Data presents multiple regression with many challenges, from the multiplicity of types of variables to the broad number of variables (even instances when number of variables exceeds the sample size) to the number of observations being considered [120]. But even with these challenges, multiple regression still provides a foundational statistical/data model well suited for a wide range of research problems focused both in prediction as well as explanation. This is not to discount the role of the emerging set of data mining/algorithmic techniques, but multiple regression still has a primary role to play in analytics today and in the future.

## An Example of Simple and Multiple Regression

The objective of regression analysis is to predict a single dependent variable from the knowledge of one or more independent variables. When the problem involves a single independent variable, the statistical technique is called **simple regression**. When the problem involves two or more independent variables, it is termed **multiple regression**. The following example will demonstrate the application of both simple and multiple regression. Readers interested in a more detailed discussion of the underlying foundations and basic elements of regression analysis are referred to the Basic Stats appendix on the text’s websites.

To illustrate the basic principles involved, results from a small study of eight families regarding their credit card usage are provided. The purpose of the study was to determine which factors affected the number of credit cards used. Three potential factors were identified (family size, family income, and number of automobiles owned), and data were collected from each of the eight families (see Figure 5.1). In the terminology of regression analysis, the dependent variable ( $Y$ ) is the number of credit cards used and the three independent variables ( $V_1$ ,  $V_2$ , and  $V_3$ ) are family size, family income, and number of automobiles owned, respectively.

Family ID	Number of Credit Cards Used ( $Y$ )	Family Size ( $V_1$ )	Family Income (\$000) ( $V_2$ )	Number of Automobiles Owned ( $V_3$ )
1	4	2	14	1
2	6	2	16	2
3	6	4	14	2
4	7	4	17	1
5	8	5	18	3
6	7	5	21	2
7	8	6	17	1
8	10	6	25	2

**Figure 5.1**  
Credit Card Usage Survey Results

## PREDICTION USING A SINGLE INDEPENDENT VARIABLE: SIMPLE REGRESSION

As researchers, a starting point in any regression analysis is identifying the single independent variable that achieves the best prediction of the dependent measure. Based on the concept of minimizing the sum of squared errors of prediction (see the Basic Stats appendix on the text's websites for more detail), we can select the "best" independent variable based on the **correlation coefficient**, because the higher the correlation coefficient, the stronger the relationship and the greater the predictive accuracy. In the regression equation, we represent the **intercept** as  $b_0$ . The amount of change in the dependent variable due to the independent variable is represented by the term  $b_1$ , also known as a **regression coefficient**. Using a mathematical procedure known as **least squares** [67, 85, 123], we can estimate the values of  $b_0$  and  $b_1$  such that the **sum of squared errors** ( $SS_E$ ) of prediction is minimized. The **prediction error**, the difference between the actual and predicted values of the dependent variable, is termed the **residual (e or  $\epsilon$ )**.

Figure 5.2 contains a correlation matrix depicting the association between the dependent ( $Y$ ) variable and independent ( $V_1$ ,  $V_2$ , or  $V_3$ ) variables that can be used in selecting the best independent variable. Looking down the first column, we can see that  $V_1$ , family size, has the highest correlation with the dependent variable and is thus the best candidate for our first simple regression. The correlation matrix also contains the correlations among the independent variables, which we will see is important in multiple regression (two or more independent variables).

We can now estimate our first simple regression model for the sample of eight families and see how well the description fits our data. The regression model can be stated as follows:

$$\text{Predicted number of credit cards used} = \text{Intercept} + \text{Change in number of credit cards used associated with a unit change in family size}$$

or

$$\hat{Y} = b_0 + b_1 V_1$$

For this example, the appropriate values are a constant ( $b_0$ ) of 2.87 and a regression coefficient ( $b_1$ ) of .97 for family size.

Variable	$Y$	$V_1$	$V_2$	$V_3$
Number of Credit Cards Used	1.000			
$V_1$ Family Size	.866	1.000		
$V_2$ Family Income	.829	.673	1.000	
$V_3$ Number of Automobiles	.342	.192	.301	1.000

**Figure 5.2**  
Correlation Matrix for the Credit Card Usage Study

**Interpreting the Simple Regression Model** With the intercept and regression coefficient estimated by the least squares procedure, attention now turns to interpretation of these two values:

**REGRESSION COEFFICIENT** The estimated change in the dependent variable for a unit change of the independent variable. If the regression coefficient is found to be statistically significant (i.e., the coefficient is significantly different from zero), the value of the regression coefficient indicates the extent to which the independent variable is associated with the dependent variable.

**INTERCEPT** Interpretation of the intercept is somewhat different. The intercept has explanatory value only within the range of values for the independent variable(s). Moreover, its interpretation is based on the characteristics of the independent variable:

- In simple terms, the intercept has interpretive value only if zero is a conceptually valid value for the independent variable (i.e., the independent variable can have a value of zero and still maintain its practical relevance). For example, assume that the independent variable is advertising dollars. If it is realistic that, in some situations, no advertising is done, then the intercept will represent the value of the dependent variable when advertising is zero.
- If the independent value represents a measure that never can have a true value of zero (e.g., attitudes or perceptions), the intercept aids in improving the prediction process, but has no explanatory value.

For some special situations where the specific relationship is known to pass through the origin, the intercept term may be suppressed (called *regression through the origin*). In these cases, the interpretation of the residuals and the regression coefficients changes slightly.

**INTERPRETING THE VARIATE** Our regression model predicting credit card holdings indicates that for each additional family member, the credit card holdings are higher on average by .97. The constant 2.87 can be interpreted only within the range of values for the independent variable. In this case, a family size of zero is not possible, so the intercept alone does not have practical meaning. However, this impossibility does not invalidate its use, because it aids in the prediction of credit card usage for each possible family size (in our example from 1 to 5). The simple regression equation and the resulting predictions and residuals for each of the eight families are shown in Figure 5.3.

Because we used only a sample of observations for estimating a regression equation, we can expect that the regression coefficients will vary if we select another sample of observations and estimate another regression equation. We do not want to take repeated samples, so we need an empirical test to see whether the regression coefficient we

**Figure 5.3**  
Simple Regression Results Using Family Size as the Independent Variable

Regression Variate: $Y = b_0 + b_1 V_1$						
Prediction Equation: $Y = 2.87 + .97V_1$						
Family ID	Number of Credit		Simple Regression		Prediction Error	
	Cards Used	Family Size ( $V_1$ )	Prediction	Prediction Error	Squared	
1	4	2	4.81	−.81	.66	
2	6	2	4.81	1.19	1.42	
3	6	4	6.75	−.75	.56	
4	7	4	6.75	.25	.06	
5	8	5	7.72	.28	.08	
6	7	5	7.72	−.72	.52	
7	8	6	8.69	−.69	.48	
8	10	6	8.69	1.31	1.72	
Total					5.50	

estimated has any real value (i.e., is it different from zero?) or could we possibly expect it to equal zero in another sample. To address this issue, regression analysis allows for the statistical testing of the intercept and regression coefficient(s) to determine whether they are significantly different from zero (i.e., they do have an impact that we can expect with a specified probability to be different from zero across any number of samples of observations). Later in the chapter, we will discuss in more detail the concept of significance testing for specific coefficients.

**Assessing Prediction Accuracy** While we may find that our variable in the regression equation is statistically significant, how do we judge the overall model? To this end, we have two measures – one a measure of the percentage of variance in the dependent variable that is accounted for by the variate (i.e., the coefficient of determination) and a second measure that is an absolute measure of the variability of the predicted outcomes (i.e., the standard error of the estimate).

**COEFFICIENT OF DETERMINATION ( $R^2$ )** The most commonly used measure of predictive accuracy for the regression model is the **coefficient of determination ( $R^2$ )**. Calculated as the squared correlation between the actual and predicted values of the dependent variable, it represents the combined effects of the entire variate (one or more independent variables plus the intercept) in predicting the dependent variable. It ranges from 1.0 (perfect prediction) to 0.0 (no prediction). Because it is the squared correlation of actual and predicted values, it also represents the amount of variance in the dependent variable explained by the independent variable(s).

In our example, the simple regression model has a total prediction error of 5.5 (see Figure 5.3), meaning that it accounted for 16.5 ( $22.0 - 5.5 = 16.5$ ) of the total prediction error of 22.0. Because the coefficient of determination is the amount of variation accounted for by the regression model, the simple regression model with one independent variable has an  $R^2$  of 75 percent ( $16.5/22.0 = .75$ ).

**STANDARD ERROR OF THE ESTIMATE** Another measure of predictive accuracy is the expected variation in the predicted values, termed the **standard error of the estimate ( $SE_E$ )**. Defined simply as the standard deviation of the predicted values, it allows the researcher to understand the confidence interval that can be expected for any prediction from the regression model. Obviously smaller confidence intervals denote greater predictive accuracy. This becomes particularly important as a “check” on model fit. Recent research surveyed researchers with results of varying levels of model fit and many times the results were perceived to be more predictable than could be justified by the model [110]. Examination of the  $SE_E$  provided an additional measure that improved assessments of the models. Thus, this measure of practical relevance is important in overall model evaluation.

We can also calculate the standard error of the estimate ( $SE_E$ ) as .957, giving a 95 percent confidence interval of 2.34, which was obtained by multiplying the  $SE_E$  by the  $t$ -value, in this case 2.477 (see the Basic Stats appendix on the text’s websites for a description of these calculations).

**OVERALL ASSESSMENT OF THE MODEL** Both of these measures of predictive accuracy can now be used to assess not only this simple regression model, but also the improvement made when more independent variables are added in a multiple regression model.

The interested reader is referred again to the Basic Stats appendix on the text’s websites where all of these basic concepts are described in more detail and calculations are provided using the credit card example.

## PREDICTION USING SEVERAL INDEPENDENT VARIABLES: MULTIPLE REGRESSION

We previously demonstrated how simple regression can help improve our prediction of a dependent variable (e.g., by using data on family size, we predicted the number of credit cards a family would use much more accurately than we could by simply using the arithmetic average). This result raises the question of whether we could improve our prediction even further by using additional independent variables (e.g., other data obtained from the families). Would our prediction be improved if we used not only data on family size, but data on another variable, perhaps family income or number of automobiles owned by the family?

**The Impact of Multicollinearity** The ability of an additional independent variable to improve the prediction of the dependent variable is related not only to its correlation to the dependent variable, but also to the correlation(s) of the additional independent variable to the independent variable(s) already in the regression equation. **Collinearity** is the association, measured as the correlation, between two independent variables. **Multicollinearity** refers to the correlation among three or more independent variables (evidenced when one is regressed against the others). Although a precise distinction separates these two concepts in statistical terms, it is rather common practice to use the terms interchangeably.

As might be expected, correlation among the independent variables can have a marked impact on the regression model in several different aspects:

**IMPACTS MEASURES OF PREDICTIVE POWER** The impact of multicollinearity is to reduce any single independent variable's unique predictive power by the extent to which it is associated with the other independent variables. As collinearity increases, the unique variance explained by each independent variable decreases and the shared prediction percentage rises. Because this shared prediction can count only once, it has two noticeable effects: (a) the predictive effect attributable to any of the independent variables is based solely on its unique predictive power, thus multicollinearity has the impact of reducing the impacted variable's regression coefficients, and (b) the predictive accuracy is still improved through the increase in shared variance, but highly redundant variables will only incrementally add to the shared variance beyond what one variable would add individually. Thus, the overall prediction increases much more slowly as independent variables with high multicollinearity are added.

**FAVORS VARIABLES WITH LOW MULTICOLLINEARITY** To maximize the prediction from a given number of independent variables, the researcher should look for independent variables that have low multicollinearity with the other independent variables but also have high correlations with the dependent variable.

We revisit the issues of collinearity and multicollinearity in later sections to discuss their implications for both the selection of independent variables and the interpretation of the regression variate.

**The Multiple Regression Equation** As noted earlier, multiple regression is the use of two or more independent variables in the prediction of the dependent variable. *The task for the researcher is to expand upon the simple regression model by adding independent variable(s) that have the greatest additional predictive power.* Even though we can determine any independent variable's association with the dependent variable through the correlation coefficient, the extent of the incremental predictive power for any additional variable is many times as much determined by its multicollinearity with other variables already in the regression equation. We can look to our credit card example to demonstrate these concepts.

**THE NEW VARIATE** To improve further our prediction of credit card holdings, let us use additional data obtained from our eight families. The second independent variable to include in the regression model is family income ( $V_2$ ), which has the next highest correlation with the dependent variable. Although  $V_2$  does have a fair degree of correlation with  $V_1$  already in the equation, it is still the next best variable to enter because  $V_3$  has a much lower correlation with the dependent variable. We simply expand our simple regression model to include two independent variables as follows:

$$\text{Predicted number of credit cards used} = b_0 + b_1 V_1 + b_2 V_2 + e$$

where:

$b_0$  = constant number of credit cards independent of family size and income

$b_1$  = change in credit card usage associated with unit change in family size

$b_2$  = change in credit card usage associated with unit change in family income

$V_1$  = family size

$V_2$  = family income

$e$  = prediction error (residual)

The multiple regression model with two independent variables, when estimated with the least squares procedure, provides a constant of .482 with regression coefficients of .63 and .216 for  $V_1$  and  $V_2$ , respectively.

**PREDICTIVE ACCURACY** We can again find our residuals by predicting  $Y$  and subtracting the prediction from the actual value. We then square the resulting prediction error, as in Figure 5.4. The sum of squared errors for the multiple regression model with family size and family income is 3.04. This result can be compared to the simple regression model value of 5.50 (Figure 5.3), which uses only family size for prediction.

When family income is added to the regression analysis,  $R^2$  also increases to .86. The inclusion of family income in the regression analysis increases the prediction by 11 percent (.86 – .75), all due to the unique incremental predictive power of family income.

**Adding a Third Independent Variable** We have seen an increase in prediction accuracy gained in moving from the simple to multiple regression equation, but we must also note that at some point the addition of independent variables will become less advantageous and even in some instances counterproductive. The addition of more independent variables is based on trade-offs between increased predictive power versus overly complex and even potentially misleading regression models.

The survey of credit card usage provides one more possible addition to the multiple regression equation, the number of automobiles owned ( $V_3$ ). If we now specify the regression equation to include all three independent variables, we can see some improvement in the regression equation, but not nearly of the magnitude seen earlier. The  $R^2$  value increases to .87, only a .01 increase over the previous multiple regression model. Moreover, as we discuss in a later section, the regression coefficient for  $V_3$  is not statistically significant. Therefore, in this instance, the researcher is best served by employing the multiple regression model with two independent variables (family size and income) and not employing the third independent variable (number of automobiles owned) in making predictions.

## SUMMARY

Regression analysis is a simple and straightforward dependence technique that can provide both prediction and explanation to the researcher. The prior example illustrated the basic concepts and procedures underlying regression analysis in an attempt to develop an understanding of the rationale and issues of this procedure in its most basic form. The following sections discuss these issues in much more detail and provide a decision process for applying regression analysis to any appropriate research problem.

**Figure 5.4**

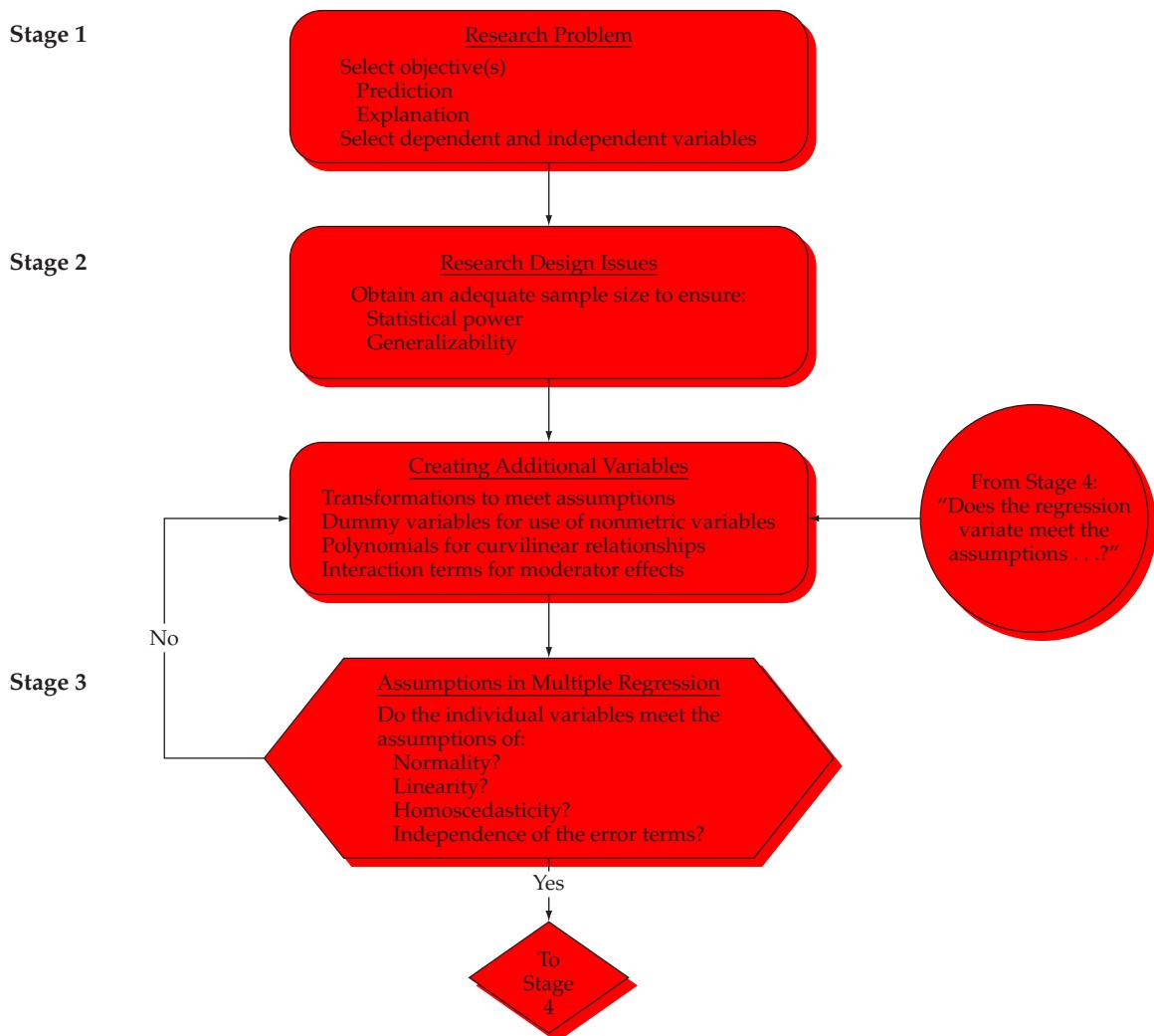
Multiple Regression Results Using Family Size and Family Income as Independent Variables

Multiple Regression Results Using Family Size and Family Income as Independent Variables						
Family ID	Used Credit Cards	Number of Credit Cards		Family Income ( $V_2$ )	Multiple Regression Prediction	Prediction Error
		Family Size ( $V_1$ )	Multiple Prediction			Prediction Error Squared
1	4	2	4.76	14	-.76	.58
2	6	2	5.20	16	.80	.64
3	6	4	6.03	14	-.03	.00
4	7	4	6.68	17	.32	.10
5	8	5	7.53	18	.47	.22
6	7	5	8.18	21	-1.18	1.39
7	8	6	7.95	17	.05	.00
8	10	6	9.67	25	.33	.11
Total						3.04

## A Decision Process for Multiple Regression Analysis

In the previous sections we discussed examples of simple regression and multiple regression. In those discussions, many factors affected our ability to find the best regression model. To this point, however, we examined these issues only in simple terms, with little regard to how they combine in an overall approach to multiple regression analysis. In the following sections, the six-stage model-building process introduced in Chapter 1 will be used as a framework for discussing the factors that affect the creation, estimation, interpretation, and validation of a regression analysis. The process begins (Stage 1) with specifying the objectives of the regression analysis, including the selection of the dependent and independent variables. In Stage 2 the researcher then proceeds to design the regression analysis, considering such factors as sample size and the need for variable transformations. With the regression model formulated, the assumptions underlying regression analysis are first tested for the individual variables (Stage 3). When all assumptions are met, then the model is estimated (Stage 4). Once results are obtained, diagnostic analyses are performed to ensure that the overall model meets the regression assumptions and that no observations have undue influence on the results. The next stage (Stage 5) is the interpretation of the regression variate; it examines the role played by each independent variable in the prediction of the dependent measure. Finally, in Stage 6 the results are validated to ensure generalizability to the population. Figures 5.5 and 5.12 represent Stages 1–3 and 4–6, respectively, in providing a graphical representation of the model-building process for multiple regression, and the following sections discuss each step in detail.

**Figure 5.5**  
Stages 1–3 in the Multiple Regression Decision Diagram



## Stage 1: Objectives of Multiple Regression

Multiple regression analysis, a form of the general linear model, is a multivariate statistical technique used to examine the relationship between a single dependent variable and a set of independent variables. The necessary starting point in multiple regression, as with all multivariate statistical techniques, is the research problem. The flexibility and adaptability of multiple regression allow for its use with almost any dependence relationship. In selecting suitable applications of multiple regression, the researcher must consider three primary issues:

- The appropriateness of the research problem
- Specification of a statistical relationship
- Selection of the dependent and independent variables.

### RESEARCH PROBLEMS APPROPRIATE FOR MULTIPLE REGRESSION

Multiple regression is by far the most widely used multivariate technique of those examined in this text. With its broad applicability, multiple regression has been used for many purposes. The ever-widening applications of multiple regression fall into two broad classes of research problems: prediction and explanation. Prediction involves the extent to which the regression variate (one or more independent variables) can predict the dependent variable. Explanation examines the regression coefficients (their magnitude, sign, and statistical significance) for each independent variable and attempts to develop a substantive or theoretical reason for the effects of the independent variables. These research problems are not mutually exclusive, and an application of multiple regression analysis can address either or both types of research problem.

**Prediction with Multiple Regression** One fundamental purpose of multiple regression is to predict the dependent variable with a set of independent variables. In doing so, multiple regression fulfills one of two objectives:

**MAXIMIZE PREDICTIVE ACCURACY** The first objective is to maximize the overall predictive power of the independent variables as represented in the variate. As shown in our earlier example of predicting credit card usage, the variate is formed by estimating regression coefficients for each independent variable so as to be the optimal predictor of the dependent measure. Predictive accuracy is always crucial to ensuring the validity of the set of independent variables. Measures of predictive accuracy are developed and statistical tests are used to assess the significance of the predictive power. In all instances, whether or not the researcher intends to interpret the coefficients of the variate, the regression analysis must achieve acceptable levels of predictive accuracy to justify its application. The researcher must ensure that both statistical and practical significance are considered (see the discussion of Stage 4).

In certain applications focused solely on prediction, the researcher is primarily interested in achieving maximum prediction, and interpreting the regression coefficients is relatively unimportant. Instead, the researcher employs the many options in both the form and the specification of the independent variables that may modify the variate to increase its predictive power, often maximizing prediction at the expense of interpretation. One specific example is a variant of regression, time series analysis, in which the sole purpose is prediction and the interpretation of results is useful only as a means of increasing predictive accuracy.

**MODEL COMPARISON** Multiple regression can also achieve a second objective of comparing two or more sets of independent variables to ascertain the predictive power of each variate. Illustrative of a confirmatory approach to modeling, this use of multiple regression is concerned with the comparison of results across two or more alternative or competing models. The primary focus of this type of analysis is the relative predictive power among models, although in any situation the prediction of the selected model must demonstrate both statistical and practical significance.

**Explanation with Multiple Regression** Multiple regression also provides a means of objectively assessing the degree and character of the relationship between dependent and independent variables by forming the variate of independent variables and then examining the magnitude, sign, and statistical significance of the regression coefficient for each independent variable. In this manner, the independent variables, in addition to their collective prediction of

the dependent variable, may also be considered for their individual contribution to the variate and its predictions. Interpretation of the variate may rely on any of three perspectives: the importance of the independent variables, the types of relationships found, or the interrelationships among the independent variables.

**RELATIVE IMPORTANCE OF INDEPENDENT VARIABLES** The most direct interpretation of the regression variate is a determination of the relative importance of each independent variable in the prediction of the dependent measure. In all applications, the selection of independent variables should be based on their theoretical relationships to the dependent variable. Regression analysis then provides a means of objectively assessing the magnitude and direction (positive or negative) of each independent variable's relationship. The character of multiple regression that differentiates it from its univariate counterparts is the simultaneous assessment of relationships between each independent variable and the dependent measure. In making this simultaneous assessment, the relative importance of each independent variable is determined.

**NATURE OF RELATIONSHIPS WITH DEPENDENT VARIABLES** In addition to assessing the importance of each variable, multiple regression also affords the researcher a means of assessing the nature of the relationships between the independent variables and the dependent variable. The assumed relationship is a linear association based on the correlations among the independent variables and the dependent measure. Transformations or additional variables are available to assess whether other types of relationships exist, particularly curvilinear relationships. This flexibility ensures that the researcher may examine the true nature of the relationship beyond the assumed linear relationship.

**NATURE OF RELATIONSHIPS AMONG INDEPENDENT VARIABLES** Finally, multiple regression provides insight into the relationships among independent variables in their prediction of the dependent measure. These interrelationships are important for two reasons. First, correlation among the independent variables may make some variables redundant in the predictive effort. As such, they are not needed to produce the optimal prediction given the other independent variable(s) in the regression equation. In such instances, the independent variable will have a strong individual relationship with the dependent variable (substantial bivariate correlations with the dependent variable), but this relationship is markedly diminished in a multivariate context (the partial correlation with the dependent variable is low when considered with other variables in the regression equation). What is the “correct” interpretation in this situation? Should the researcher focus on the strong bivariate correlation to assess importance, or should the diminished relationship in the multivariate context form the basis for assessing the variable's relationship with the dependent variable?

Here the researcher must rely on the theoretical bases for the regression analysis to assess the “true” relationship for the independent variable. *In such situations, the researcher must guard against determining the importance of independent variables based solely on the derived variate, because relationships among the independent variables may mask or confound relationships that are not needed for predictive purposes but represent substantive findings nonetheless.* The interrelationships among variables can extend not only to their predictive power but also to interrelationships among their estimated effects, which is best seen when the effect of one independent variable is contingent on another independent variable.

Multiple regression provides diagnostic analyses that can determine whether such effects exist based on empirical or theoretical rationale. Indications of a high degree of interrelationships (multicollinearity) among the independent variables may suggest the use of summated scales or factor scores, as discussed in Chapter 3.

## SPECIFYING A STATISTICAL RELATIONSHIP

Multiple regression is appropriate when the researcher is interested in a statistical, not a functional, relationship. For example, let us examine the following relationship:

$$\text{Total cost} = \text{Variable cost} + \text{Fixed cost}$$

If the variable cost is \$2 per unit, the fixed cost is \$500, and we produce 100 units, we assume that the total cost will be exactly \$700 and that any deviation from \$700 is caused by our inability to measure cost because the relationship between costs is fixed. It is called a **functional relationship** because we expect no error in our prediction. As such, we always know the impact of each variable in calculating the outcome measure.

But in our earlier example dealing with sample data representing human behavior, we assumed that our description of credit card usage was only approximate and not a perfect prediction. It was defined as a **statistical relationship** because some random component is always present in the relationship being examined. A statistical relationship is characterized by two elements:

- 1** When multiple observations are collected, more than one value of the dependent value will usually be observed for any value of an independent variable.
- 2** Based on the use of a random sample, the error in predicting the dependent variable is also assumed to be random, and for a given independent variable we can only hope to estimate the average value of the dependent variable associated with it.

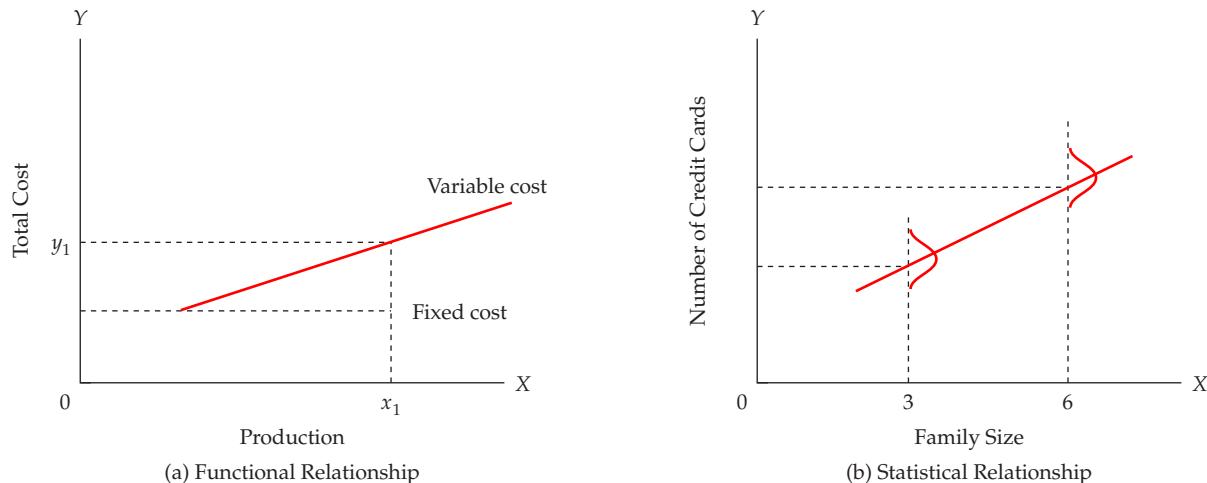
For example, in our simple regression example, we found two families with two members, two with four members, and so on, who had different numbers of credit cards. The two families with four members held an average of 6.5 credit cards, and our prediction was 6.75. Our prediction is not as accurate as we would like, but it is better than just using the average of 7 credit cards. The error is assumed to be the result of random behavior among credit card holders.

In summary, a functional relationship calculates an exact value, whereas a statistical relationship estimates an average value. Both of these relationships are displayed in Figure 5.6. Throughout this book, we are concerned with statistical relationships. Our ability to employ just a sample of observations and then use the estimation methods of the multivariate techniques and assess the significance of the independent variables is based on statistical theory. In doing so, we must be sure to meet the statistical assumptions underlying each multivariate technique, because they are critical to our ability to make unbiased predictions of the dependent variable and valid interpretations of the independent variables.

### SELECTION OF DEPENDENT AND INDEPENDENT VARIABLES

The ultimate success of any multivariate technique, including multiple regression, starts with the selection of the variables to be used in the analysis. Because multiple regression is a dependence technique, the researcher must specify which variable is the dependent variable and which variables are the independent variables. Although many times the options seem apparent, the researcher should always consider three issues that can affect any decision: strong theory, measurement error, and specification error.

**Figure 5.6**  
Comparison of Functional and Statistical Relationships



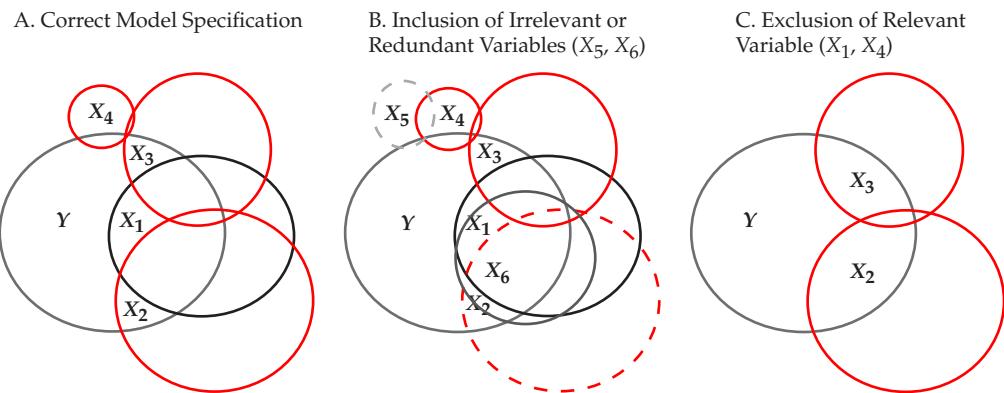
**Strong Theory** The selection of both types of variables should be based principally on conceptual or theoretical grounds, even when the objective is solely for prediction. Chapters 1 and 9 discuss the role of theory in multivariate analysis, and those issues strongly apply to multiple regression. The researcher should always perform the fundamental decisions of variable selection, even though many options and program features are available to assist in model estimation. If the researcher does not exert judgment during variable selection, but instead (1) selects variables indiscriminately or (2) allows for the selection of an independent variable to be based solely on empirical bases, several of the basic tenets of model development will be violated.

**Measurement Error** The selection of variables in the analysis, either the dependent or independent variable, is often dictated by the research problem. In all instances, the researcher must be aware of the **measurement error** in these variables, especially in the dependent variable. Measurement error refers to the degree to which the variable is an accurate and consistent measure of the concept being studied. For example, if the variable used as the dependent measure has substantial measurement error, then even the best independent variables may be unable to achieve acceptable levels of predictive accuracy. Similarly, independent variables with a high degree of measurement error will not accurately represent the relationship with the dependent variable. Although measurement error can come from several sources (see Chapter 1 for a more detailed discussion), multiple regression has no direct means of correcting for known levels of measurement error for the dependent or independent variables. But the researcher has two approaches for addressing the problems associated with measurement error.

**USE OF COMPOSITES** Chapter 3 discussed in detail the use of composite measures, whether they be factor scores or summated scales, as a means of addressing measurement error among a set of variables. Composite measures are most applicable to independent variables since they can join together a set of variables into a composite where the shared variance in the composite has a much smaller degree of measurement error than the individual variables. Composite measures can also be used for dependent measures where perhaps there are a number of alternative measures of the outcome and a composite better reflects the concept than any one of the individual variables. Summated scales can be directly incorporated into multiple regression by replacing either dependent or independent variables with the summated scale values. Many times summated scales are the best “first step” as a remedy for measurement error.

**STRUCTURAL EQUATION MODELING** The technique of structural equation modeling is focused on addressing measurement error through composites as part of the technique rather than separately as described above. Whether it be traditional covariance-based SEM (Chapters 9 to 12) or Partial Least Squares (Chapter 13), both approaches incorporate the estimation of composites and then their use in estimating the dependence relationships into a single technique. Most often structural equation modeling is applied to research questions where data collection is specifically oriented towards generating the variables needed for specifying composites, but partial least squares has more flexibility in accommodating a more diverse set of variables in forming composites. But as might be expected, the formation of composites whether with covariance or variance based SEM can be rather complicated.

**Specification Error** Perhaps the most problematic issue in independent variable selection is **specification error**, which concerns the inclusion of irrelevant variables or the omission of relevant or redundant variables from the set of independent variables. Both types of specification error can have substantial impacts on any regression analysis, although in quite different ways. Figure 5.7 depicts these two situations, and both are described in more detail in the sections below. Part A depicts the correct model specification, three collinear independent variables ( $X_1$ ,  $X_2$ , and  $X_3$ ) along with  $X_4$  that is generally unrelated to the other three variables. The overall fit of the variate (i.e., the  $R^2$ ) is the total overlap of the variate with the dependent variable Y. The regression coefficients are depicted by the area with each independent variable's circle that overlaps with the dependent variable and does not overlap with any other variable. So for example, with  $X_2$ , even though it has a substantial overlap with Y, the regression coefficient will only reflect the small area not overlapped by  $X_1$ . There are similar situations for  $X_1$  and  $X_3$ . For  $X_4$ , even though its overlap is not as large as those of the other independent variables, its regression coefficient will reflect almost all of that overlap since it is not correlated with the other variables to any substantive degree.

**Figure 5.7****Examples of Specification Error: Exclusion and Inclusion of Variables**

**INCLUSION OF IRRELEVANT OR REDUNDANT VARIABLE** Although the *inclusion of irrelevant variables* does not bias the results for the other independent variables, it does impact the regression variate. First, it reduces model parsimony, which may be critical in the interpretation of the results. Second, the additional variables may mask or replace the effects of more useful variables, especially if some sequential form of model estimation is used (see the discussion of Stage 4 for more detail). Finally, the additional variables may make the testing of statistical significance of the independent variables less precise and reduce the statistical and practical significance of the analysis.

Part B of Figure 5.7 depicts some of these situations. First,  $X_5$  is essentially an irrelevant variable, with little overlap with the dependent variable  $Y$  and not even any impact on  $X_4$ . Thus, its impact is minimal. But  $X_6$  has substantial effect. First,  $X_6$  represents a variable that is highly correlated with  $X_1$  and  $X_2$  yet adds very little explanatory power (i.e., very small additional overlap with  $Y$ ). But it does have an effect on  $X_1$  and  $X_2$  by overlapping their unique portions of  $Y$  and thus detracting from the coefficients of  $X_1$  and  $X_2$ . Then  $X_6$  will appear non-significant in the estimated model since it adds no unique effect, but will also diminish the effects of  $X_1$  and  $X_2$ . We can expect similar effects for the other variables if we include redundant variables in the regression model that add little explanatory power yet detract from the other variable's unique effects.

**EXCLUSION OF RELEVANT VARIABLE** Given the problems associated with adding irrelevant variables, should the researcher be concerned with *excluding relevant variables*? The answer is definitely yes, because the exclusion of relevant variables can seriously bias the results and negatively affect any interpretation of them. In the simplest case, the omitted variables are uncorrelated with the included variables, and the only effect is to reduce the overall predictive accuracy of the analysis. But when correlation exists between the included and omitted variables, the effects of the included variables become biased to the extent that they are correlated with the omitted variables. The greater the correlation, the greater the bias. The estimated effects for the included variables now represent not only their actual effects but also the effects that the included variables share with the omitted variables. These effects can lead to serious problems in model interpretation and the assessment of statistical and managerial significance.

Part C of Figure 5.7 represents the impacts of omitted relevant variables. The omission of  $X_4$  reduces the overall model to some extent, but has no impact on the other variables. But the omission of  $X_1$  is quite different. The overall model fit is reduced somewhat (i.e., the non-overlap portion of  $X_1$ ), but  $X_2$  and  $X_3$  still have substantial overlap with  $Y$ . The greater impact is now the estimated effects of  $X_2$  and  $X_3$ , which have increased dramatically with  $X_1$  removed. If prediction is the only objective, then these effects do not matter. But if explanation is important and  $X_1$  should be considered, then omitting  $X_1$  will result in an over-statement of the impacts of  $X_2$  and  $X_3$ .

**WHICH PROBLEM TO AVOID?** The researcher must be careful in the selection of the variables to avoid both types of specification error. Perhaps most troublesome is the omission of relevant variables, because the variables' effect cannot be assessed without their inclusion (see Rules of Thumb 5-1). In most instances these types of considerations are not addressed empirically,

## Meeting Multiple Regression Objectives

Only structural equation modeling (SEM) can directly accommodate measurement error, but using summated scales can mitigate it when using multiple regression.

When in doubt, include potentially irrelevant variables (they can only confuse interpretation) rather than possibly omitting a relevant variable (which can bias all regression estimates).

but rather by the researcher's judgment on the variables to be included in the estimated model. As we will discuss in later sections, these include not only decisions about the set of variables to be included in the analysis, but also the type of variable selection approach the researcher uses for specifying the final model. These potential influences on any results heighten the need for theoretical and practical support for all variables included or excluded in a multiple regression analysis.

## Stage 2: Research Design of a Multiple Regression Analysis

Adaptability and flexibility are two principal reasons for multiple regression's widespread usage across a wide variety of applications. As you will see in the following sections, multiple regression can represent a wide range of dependence relationships. In doing so, the researcher incorporates three features:

- *Sample size.* Multiple regression maintains the necessary levels of statistical power and practical/statistical significance across a broad range of sample sizes.
- *Unique elements of the dependence relationship.* Even though independent variables are assumed to be metric and have a linear relationship with the dependent variable, both assumptions can be relaxed by creating additional variables to represent these special aspects of the relationship.
- *Nature of the relationship of the independent variables.* The use of moderation and mediation provide the researcher with additional effects beyond the direct relationship between independent and dependent variables.

Each of these features plays a key role in the application of multiple regression to many types of research questions while maintaining the necessary levels of statistical and practical significance.

### SAMPLE SIZE

The sample size used in multiple regression is perhaps the single most influential element under the control of the researcher in designing the analysis. The effects of sample size are seen most directly in the statistical power of the significance testing and the generalizability of the result. Both issues are addressed in the following sections.

**Statistical Power and Sample Size** The size of the sample has a direct impact on the appropriateness and the statistical power of multiple regression, as discussed in Chapter 1. Small samples, usually characterized as having fewer than 30 observations, are appropriate for analysis only by simple regression with a single independent variable. Even in these situations, only strong relationships can be detected with any degree of certainty. Likewise, large samples of 1,000 observations or more make the statistical significance tests overly sensitive, often indicating that almost any relationship is statistically significant. With such large samples the researcher must ensure that the criterion of practical significance is met along with statistical significance.

**POWER LEVELS IN VARIOUS REGRESSION MODELS** In multiple regression **power** refers to the probability of detecting as statistically significant a specific level of  $R^2$  or a regression coefficient at a specified significance level for a specific

sample size (see Chapter 1 for a more detailed discussion). Sample size plays a role in not only assessing the power of a current analysis, but also in anticipating the statistical power of a proposed analysis.

Figure 5.8 illustrates the *interplay among the sample size, the significance level ( $\alpha$ ) chosen, and the number of independent variables* in detecting a significant  $R^2$ . The table values are the minimum  $R^2$  that the specified sample size will detect as statistically significant at the specified alpha ( $\alpha$ ) level with a power (probability) of .80.

For example, if the researcher employs five independent variables, specifies a .05 significance level, and is satisfied with detecting the  $R^2$  80 percent of the time it occurs (corresponding to a power of .80), a sample of 50 respondents will detect  $R^2$  values of 23 percent and greater. If the sample is increased to 100 respondents, then  $R^2$  values of 12 percent and above will be detected. But if 50 respondents are all that are available, and the researcher wants a .01 significance level, the analysis will detect  $R^2$  values only in excess of 29 percent.

**SAMPLE SIZE REQUIREMENTS FOR DESIRED POWER** The researcher can also consider the *role of sample size in significance testing before collecting data*. If weaker relationships are expected, the researcher can make informed judgments as to the necessary sample size to reasonably detect the relationships, if they exist.

For example, Figure 5.8 demonstrates that sample sizes of 100 will detect fairly small  $R^2$  values (10% to 15%) with up to 10 independent variables and a significance level of .05. However, if the sample size falls to 50 observations in these situations, the minimum  $R^2$  that can be detected doubles.

The researcher can also determine the sample size needed to detect effects for individual independent variables given the expected effect size (correlation), the  $\alpha$  level, and the power desired. The possible computations are too numerous for presentation in this discussion, and the interested reader is referred to texts dealing with power analyses [29] or to a computer program to calculate sample size or power for a given situation [15].

**SUMMARY** The researcher must always be aware of the anticipated power of any proposed multiple regression analysis. It is critical to understand the elements of the research design that can be changed to meet the requirements for an acceptable analysis, and sample size is a particularly important consideration [80].

**Generalizability and Sample Size** In addition to its role in determining statistical power, sample size also affects the generalizability of the results by the ratio of observations to independent variables. A general rule is that the ratio should never fall below 5:1, meaning that five observations are made for each independent variable in the variate. Although the minimum ratio is 5:1, the desired level is between 15 to 20 observations for each independent variable. When this level is reached, the results should be generalizable if the sample is representative. However, if a stepwise procedure is employed, the recommended level increases to 50:1 because this technique selects only the strongest relationships within the dataset and suffers from a greater tendency to become sample-specific [125]. In cases for which the available sample does not meet these criteria, the researcher should be certain to validate the generalizability of the results.

Sample Size	<i>Significance Level (<math>\alpha</math>) = .01</i>				<i>Significance Level (<math>\alpha</math>) = .05</i>			
	<i>No. of Independent Variables</i>				<i>No. of Independent Variables</i>			
	2	5	10	20	2	5	10	20
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1,000	1	2	2	3	1	1	2	2

**Figure 5.8**  
Minimum  $R^2$  That can be Found Statistically Significant with a Power of .80 for Varying Numbers of Independent Variables and Sample Sizes

Note: Values represent percentage of variance explained.

NA = not applicable.

**DEFINING DEGREES OF FREEDOM** As this ratio falls below 5:1, the researcher encounters the risk of overfitting the variate to the sample, making the results too specific to the sample and thus lacking generalizability. In understanding the concept of overfitting, we need to address the statistical concept of **degrees of freedom**. In any statistical estimation procedure, the researcher is making estimates of parameters from the sample data. In the case of regression, the parameters are the regression coefficients for each independent variable and the constant term. As described earlier, the regression coefficients are the weights used in calculating the regression variate and indicate each independent variable's contribution to the predicted value. What, then, is the relationship between the number of observations and variables? Let us look at a simple view of estimating parameters for some insight into this problem.

Each observation represents a separate and independent unit of information (i.e., one set of values for each independent variable). In a simplistic view, the researcher could dedicate a single variable to perfectly predicting only one observation, a second variable to another observation, and so forth. If the sample is relatively small, then predictive accuracy could be quite high, and many of the observations would be perfectly predicted. As a matter of fact, if the number of estimated **parameters** (regression coefficients and the constant) equals the sample size, perfect prediction will occur even if all the variable values are random numbers. This scenario would be totally unacceptable and considered extreme overfitting because the estimated parameters have no generalizability, but relate only to the sample data. *Moreover, anytime a variable is added to the regression equation, the  $R^2$  value will increase.*

**DEGREES OF FREEDOM AS A MEASURE OF GENERALIZABILITY** What happens to generalizability as the sample size increases? We can perfectly predict one observation with a single variable, but what about all the other observations? Thus, the researcher is searching for the best regression model, one with the highest predictive accuracy for the largest (most generalizable) sample. The degree of generalizability is represented by the degrees of freedom, calculated as:

$$\text{Degrees of freedom } (df) = \text{Sample size} - \text{Number of estimated parameters}$$

or:

$$\text{Degrees of freedom } (df) = N - (\text{Number of independent variables} + 1)$$

The larger the degrees of freedom, the more generalizable are the results. Degrees of freedom increase for a given sample by reducing the number of independent variables. Thus, the objective is to achieve the highest predictive accuracy with the most degrees of freedom. In our preceding example, where the number of estimated parameters equals the sample size, we have perfect prediction, but *zero degrees of freedom!* The researcher must reduce the number of independent variables (or increase the sample size), lowering the predictive accuracy but also increasing the degrees of freedom. No specific guidelines determine how large the degrees of freedom are, just that they are indicative of the generalizability of the results and give an idea of the overfitting for any regression model as shown in Rules of Thumb 5-2.

## Sample Size Considerations

Simple regression can be effective with a sample size of 20, but maintaining power at .80 in multiple regression requires a minimum sample of 50 and preferably 100 observations for most research situations.

The minimum ratio of observations to variables is 5:1, but the preferred ratio is 15:1 or 20:1, which should increase when stepwise estimation is used.

Maximizing the degrees of freedom improves generalizability and addresses both model parsimony and sample size concerns.

## CREATING ADDITIONAL VARIABLES

The basic relationship represented in multiple regression is the *linear* association between metric dependent and independent variables based on the product-moment correlation. One problem often faced by researchers is the desire to incorporate nonmetric data, such as gender or occupation, into a regression equation. Yet, as we already discussed, regression is limited to metric data. Moreover, regression's inability to directly model nonlinear relationships may constrain the researcher when faced with situations in which a nonlinear relationship (e.g., U-shaped) is suggested by theory or detected when examining the data.

**Data Transformations** In these situations, new variables must be created by **transformations**, because multiple regression is totally reliant on creating new variables in the model to incorporate nonmetric variables or represent any effects other than linear relationships. We also will encounter the use of transformations discussed in Chapter 2 as a means of remedying violations of some statistical assumptions, but our purpose here is to provide the researcher with a means of modifying either the dependent or independent variables for one of two reasons:

- 1 Improve or modify the relationship between independent and dependent variables
- 2 Enable the use of nonmetric variables in the regression variate.

Data transformations may be based on reasons that are either *theoretical* (transformations whose appropriateness is based on the nature of the data) or *data derived* (transformations that are suggested strictly by an examination of the data). In either case the researcher must proceed many times by trial and error, constantly assessing the improvement versus the need for additional transformations. We explore these issues with discussions of data transformations that enable the regression analysis to best represent the actual data and a discussion of the creation of variables to supplement the original variables.

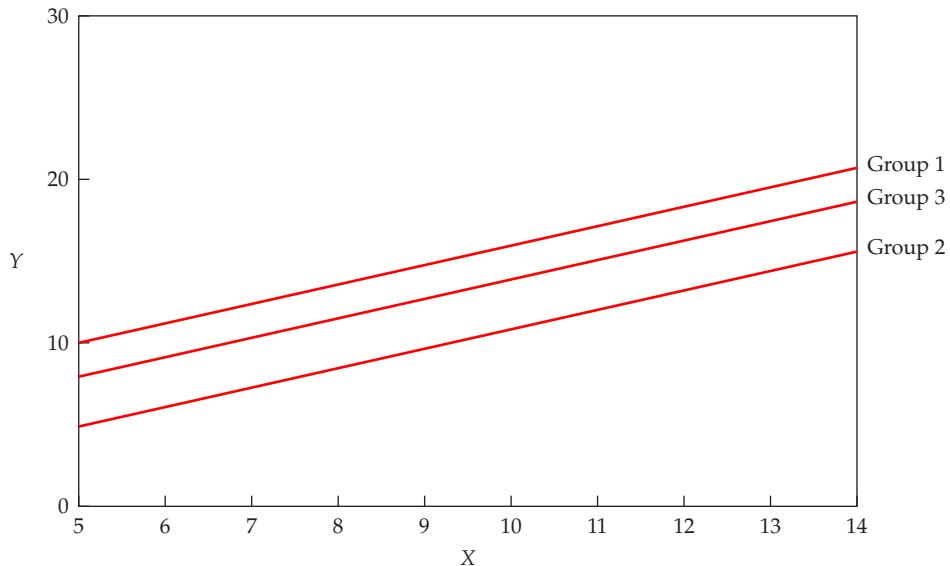
All the transformations we describe are easily carried out by simple commands in all the popular statistical packages. Although we focus on transformations that can be computed in this manner, other more sophisticated and complicated methods of data transformation are available (e.g., see Box and Cox [18]).

**Incorporating Nonmetric Data with Dummy Variables** One common situation faced by researchers is the desire to utilize nonmetric independent variables. Yet, to this point, all our illustrations assumed metric measurement for both independent and dependent variables. When the dependent variable is measured as a dichotomous (0, 1) variable, either discriminant analysis or a specialized form of regression (logistic regression), discussed in later chapters, is appropriate. What can we do when the independent variables are nonmetric and have two or more categories? Chapter 2 introduced the concept of dichotomous variables, known as **dummy variables**, which can act as replacement independent variables. Each dummy variable represents one category of a nonmetric independent variable, and any nonmetric variable with  $k$  categories can be represented as  $k - 1$  dummy variables.

**INDICATOR CODING: THE MOST COMMON FORMAT** Of the two forms of dummy variable coding, the most common is **indicator coding** in which each category of the nonmetric variable is represented by either 1 or 0. *The regression coefficients for the dummy variables represent differences on the dependent variable for each group of respondents from the reference category* (i.e., the omitted group that received all zeros). *These group differences can be assessed directly, because the coefficients are in the same units as the dependent variable.*

This form of dummy-variable coding can be depicted as differing intercepts for the various groups, with the reference category represented in the constant term of the regression model (see Figure 5.9). In this example, a three-category nonmetric variable is represented by two dummy variables ( $D_1$  and  $D_2$ ) representing groups 1 and 2, with group 3 the reference category. The regression coefficients are 2.0 for  $D_1$  and -3.0 for  $D_2$ . These coefficients translate into three parallel lines. The reference group (in this case group 3) is defined by the regression equation with both dummy variables equaling zero. Group 1's line is two units above the line for the reference group. Group 2's line is three units below the line for reference group 3. The parallel lines indicate that dummy variables do not change the nature of the relationship, but only provide for differing intercepts among the groups.

**Figure 5.9**  
Incorporating Nonmetric Variables Through Dummy Variables



Regression Equations with Dummy Variables ( $D_1$ and $D_2$ )	
Specified	$Y = a + b_1X + b_2D_1 + b_3D_2$
Estimated	
Overall	$Y = 2 + 1.2X + 2D_1 - 3D_2$
Group Specific	
Group 1 ( $D_1 = 1, D_2 = 0$ )	$Y = 2 + 1.2X + 2(1)$
Group 2 ( $D_1 = 0, D_2 = 1$ )	$Y = 2 + 1.2X - 3(1)$
Group 3 ( $D_1 = 0, D_2 = 0$ )	$Y = 2 + 1.2X$

This form of coding is most appropriate when a logical reference group is present, such as in an experiment. Any time dummy-variable coding is used, we must be aware of the comparison group and remember that the coefficients represent the differences in group means from this group.

**EFFECTS CODING** An alternative method of dummy-variable coding is termed **effects coding**. It is the same as indicator coding except that the comparison or omitted group (the group that got all zeros) is now given the value of  $-1$  instead of  $0$  for the dummy variables. *Now the coefficients represent differences for any group from the mean of all groups rather than from the omitted group.* Both forms of dummy-variable coding will give the same predictive results, coefficient of determination, and regression coefficients for the continuous variables. The only differences will be in the interpretation of the dummy-variable coefficients. In effects coding the intercept is the unweighted average of the group means, so that unequal group sizes create slight interpretation differences from indicator coding [4].

**Representing Curvilinear Effects with Polynomials** Several types of data transformations are appropriate for linearizing a curvilinear relationship. Direct approaches, discussed in Chapter 2, involve modifying the values

through some arithmetic transformation (e.g., taking the square root or logarithm of the variable). However, such transformations are subject to the following limitations:

- They are applicable only in a simple curvilinear relationship (a relationship with only one turning or inflection point).
- They do not provide any statistical means for assessing whether the curvilinear or linear model is more appropriate.
- They accommodate only univariate relationships and not the interaction between variables when more than one independent variable is involved.

We now discuss a means of creating new variables to explicitly model the curvilinear components of the relationship and address each of the limitations inherent in data transformations.

**SPECIFYING A CURVILINEAR EFFECT** Power transformations of an independent variable that add a nonlinear component for each additional power of the independent variable are known as **polynomials**. The power of 1 ( $X^1$ ) represents the linear component and is the form discussed so far in this chapter. The power of 2, the variable squared ( $X^2$ ), represents the quadratic component. In graphical terms,  $X^2$  represents the first inflection point of a curvilinear relationship. A cubic component, represented by the variable cubed ( $X^3$ ), adds a second inflection point. With these variables, and even higher powers, we can accommodate more complex relationships than are possible with only transformations. For example, in a simple regression model, a curvilinear model with one turning point can be modeled with the equation:

$$Y = b_0 + b_1X_1 + b_2X_1^2$$

where:

$b_0$  = intercept

$b_1X_1$  = linear effect of  $X_1$

$b_2X_1^2$  = curvilinear effect of  $X_1$

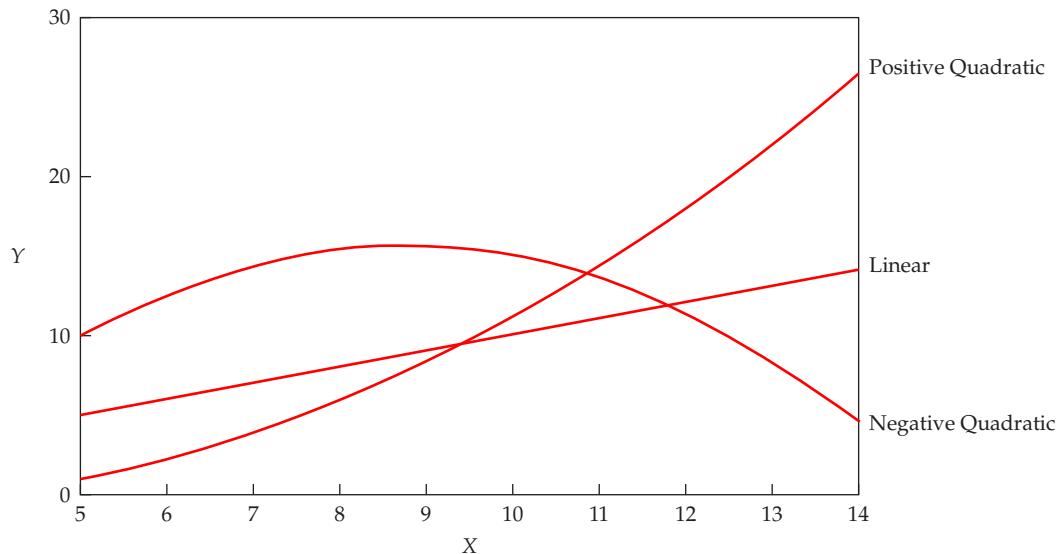
Although any number of nonlinear components may be added, the cubic term is usually the highest power used. Multivariate polynomials are created when the regression equation contains two or more independent variables. We follow the same procedure for creating the polynomial terms as before, but must also create an additional term, the interaction term ( $X_1X_2$ ), which is needed for each variable combination to represent fully the multivariate effects. In graphical terms, a two-variable multivariate polynomial is portrayed by a surface with one peak or valley. For higher-order polynomials, the best form of interpretation is obtained by plotting the surface from the predicted values.

**INTERPRETING A CURVILINEAR EFFECT** As each new variable is entered into the regression equation, we can also perform a direct statistical test of the nonlinear components, which we cannot do with data transformations. However, multicollinearity can create problems in assessing the statistical significance of the individual coefficients to the extent that the researcher should assess incremental effects as a measure of any polynomial terms in a three-step process:

- 1 Estimate the original regression equation.
- 2 Estimate the curvilinear relationship (original equation plus polynomial term).
- 3 Assess the change in  $R^2$ . If it is statistically significant, then a significant curvilinear effect is present. The focus is on the incremental effect, not the significance of individual variables.

Three (two nonlinear and one linear) relationships are shown in Figure 5.10. For interpretation purposes, the positive quadratic term indicates a  $\cup$ -shaped curve, whereas a negative coefficient indicates a  $\cap$ -shaped curve. The use of a cubic term can represent such forms as the S-shaped or growth curve quite easily, but it is generally best to plot the values to interpret the actual shape.

**Figure 5.10**  
Representing Nonlinear Relationships with Polynomials



**HOW MANY TERMS SHOULD BE ADDED?** Common practice is to start with the linear component and then sequentially add higher-order polynomials until non-significance is achieved. The use of polynomials, however, also has potential problems. First, each additional term requires a degree of freedom, which may be particularly restrictive with small sample sizes. This limitation does not occur with data transformations. Also, multicollinearity is introduced by the additional terms and makes statistical significance testing of the polynomial terms inappropriate. Instead, the researcher must compare the  $R^2$  values from the equation model with linear terms to the  $R^2$  for the equation with the polynomial terms. Testing for the statistical significance of the incremental  $R^2$  is the appropriate manner of assessing the impact of the polynomials.

**Representing Interaction or Moderator Effects** The nonlinear relationships just discussed require the creation of an additional variable (e.g., the squared term) to represent the changing slope of the relationship over the range of the independent variable. This variable focuses on the relationship between a single independent variable and the dependent variable. But what if an independent-dependent variable relationship is affected by another independent variable? This situation is termed a **moderator effect**, which occurs when the moderator variable, a second independent variable, changes the *form* of the relationship between another independent variable and the dependent variable. This extension to the effect of an independent variable with the dependent variable has long been a primary topic of interest since it addresses the fundamental question of “When” does this effect occur [1]. It is also known as an *interaction effect* and is similar to the interaction term found in analysis of variance and multivariate analysis of variance (see Chapter 6 for more detail on interaction terms).

**EXAMPLES OF MODERATOR EFFECTS** The most common moderator effect employed in multiple regression is the *quasi* or *bilinear moderator*, in which the slope of the relationship of one independent variable ( $X_1$ ) changes across values of the moderator variable ( $X_2$ ) [63, 106].

In our earlier example of credit card usage, assume that family income ( $X_2$ ) was found to be a positive moderator of the relationship between family size ( $X_1$ ) and credit card usage ( $Y$ ). It would mean that the expected change in credit card usage based on family size ( $b_1$ , the regression coefficient for  $X_1$ ) might be lower for families with low

incomes and higher for families with higher incomes. Without the moderator effect, we assumed that family size had a constant effect on the number of credit cards used, but the interaction term tells us that this relationship changes, depending on family income level. Note that it does not necessarily mean the effects of family size or family income by themselves are unimportant, but instead that the interaction term complements their explanation of credit card usage.

**ADDING THE MODERATOR EFFECT** The moderator effect is represented in multiple regression by a term quite similar to the polynomials described earlier to represent nonlinear effects. The moderator term is a compound variable formed by multiplying  $X_1$  by the moderator  $X_2$ , which is entered into the regression equation. In fact, the nonlinear term can be viewed as a form of interaction, where the independent variable “moderates” itself, thus the squared term ( $X_1 X_2$ ). The moderated relationship is represented as:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$$

where:

$b_0$  = intercept

$b_1 X_1$  = linear effect of  $X_1$

$b_2 X_2$  = linear effect of  $X_2$

$b_3 X_1 X_2$  = moderator effect of  $X_2$  on  $X_1$

Because of the multicollinearity among the old and new variables, an approach similar to testing for the significance of polynomial (nonlinear) effects is employed. To determine whether the moderator effect is significant, the researcher follows a three-step process:

- 1 Estimate the original (unmoderated) equation.
- 2 Estimate the moderated relationship (original equation plus moderator variable).
- 3 Assess the change in  $R^2$ : If it is statistically significant, then a significant moderator effect is present. Only the incremental effect is assessed, not the significance of individual variables.

**INTERPRETING MODERATOR EFFECTS** The interpretation of the regression coefficients changes slightly in moderated relationships. *The  $b_3$  coefficient, the moderator effect, indicates the unit change in the effect of  $X_1$  as  $X_2$  changes. The  $b_1$  and  $b_2$  coefficients now represent the effects of  $X_1$  and  $X_2$ , respectively, when the other independent variable is zero.* In the unmoderated relationship, the  $b_1$  coefficient represents the effect of  $X_1$  across all levels of  $X_2$ , and similarly for  $b_2$ . Thus, in unmoderated regression, the regression coefficients  $b_1$  and  $b_2$  are averaged across levels of the other independent variables, whereas in a moderated relationship they are separate from the other independent variables. To determine the total effect of an independent variable, the separate and moderated effects must be combined. The overall effect of  $X_1$  for any value of  $X_2$  can be found by substituting the  $X_2$  value into the following equation:

$$b_{X_1(\text{overall})} = b_{X_1} + b_3 X_2$$

For example, assume a moderated regression resulted in the following coefficients:  $b_1 = 2.0$  and  $b_3 = .5$ . If the value of  $X_2$  ranges from 1 to 7, the researcher can calculate the total effect of  $X_1$  at any value of  $X_2$ . When  $X_2$  equals 3, the total effect of  $X_1$  is 3.5 [ $2.0 + .5(3)$ ]. When  $X_2$  increases to 7, the total effect of  $X_1$  is now 5.5 [ $2.0 + .5(7)$ ].

We can see the moderator effect at work, making the relationship of  $X_1$  and the dependent variable change, given the level of  $X_2$ . Researchers should always explore the full range of the moderation effects and software supplements for SAS and IBM SPSS have been developed for this purpose [55]. Excellent discussions of moderated relationships in multiple regression are available in a number of sources [29, 63, 106]. However, there are also distinct issues with the interpretation of interactions and their impact on other results [43, 44]. Researchers are cautioned on their widespread use unless distinct effects are hypothesized.

**Summary** The creation of new variables provides the researcher with great flexibility in representing a wide range of relationships within regression models (see Rules of Thumb 5-3). Yet too often the desire for better model fit leads to the inclusion of these special relationships without theoretical support that can lead to misleading results [30]. In these instances, the researcher is running a much greater risk of finding results with little or no generalizability. Instead, in using these additional variables, the researcher must be guided by theory that is supported by empirical analysis. In this manner, both practical and statistical significance can hopefully be met.

**Mediation** The concept of mediation is not a data transformation, but instead a re-specification of the causal relationships in the analysis. **Mediation** occurs when the effect of an independent variable may “work through” an intervening variable (the mediating variable) to predict the dependent variable. In this situation the independent variable may have a direct effect on the dependent measure as well as an indirect effect through the mediating variable to the dependent variable. While most commonly associated with ANOVA and MANOVA models (see Chapter 6 for an extended discussion), its relevance in the discussion of multiple regression relates to the roles of independent variables. A simple example illustrates how mediation may occur.

Assume that consumer behavior research indicates a sequential cognitive process occurs—customer satisfaction leads to future purchase intentions, but also leads to higher loyalty attitudes which are related to higher future purchase intentions. Now if we were to perform a simple multiple regression analysis we would place both customer satisfaction and loyalty as independent variables and future purchase intentions as the dependent measure. We would then determine the relative importance of customer satisfaction and loyalty through their regression coefficients. If both are significant, then we can assume that there is some effect on future purchase intentions from both customer satisfaction and loyalty. While these estimated coefficients are empirically correct, we have understated the effect of customer satisfaction—What about the role of customer satisfaction in predicting loyalty? The full impact of customer satisfaction is actually both its direct effect on future purchase intentions plus some additional effect (an indirect effect) from its relationship to loyalty and then on to future purchase intentions.

Since both customer satisfaction and loyalty are included in the model, the definition of loyalty as a mediator instead of as just another independent variable is conceptual, not empirical. Introducing mediation into the analysis requires conceptual support for the proposed causal path (i.e., the path from the independent variable to the mediator) and empirical evidence that might be available [78]. But as seen in the earlier example, the “true” effect of customer satisfaction on purchase intention would not be seen without the specification of loyalty as a mediating effect. The introduction of mediation into the analysis provides a framework for understanding more than just direct effects, and an integration of mediation and moderation has been developed [39, 54]. As the number of mediation effects move beyond a single independent variable and mediator, the researcher should consider the path models of structural equation analysis (see Chapters 9 and 13 for introductions) where specifying several mediating effects in a single analysis is possible. It is beyond the scope of this chapter to address all of the issues surrounding mediation effects, but it is critical for the researcher to understand the implications of identifying potential mediators and the role they should play in the analysis in order to accurately portray the total effects of the independent variables.

## OVERVIEW

A wide number of issues are addressed in this stage, ranging from the impact of sample size to the nature of the relationships to be included in the analysis. A common theme in all of these issues is the role of researcher judgment, which can play a substantive positive or negative role in the ultimate results. We encourage researchers to employ these elements, whether it be increasing sample size for purposes of statistical power, introducing data transformations to improve the ability to represent the accurate relationships of the variables or even employ moderation and/or mediation effects. But in all instances these actions should be guided with an understanding of their potential impact since indiscriminate use of these options can introduce as many detrimental effects as positive effects.

## Variable Transformations

Nonmetric variables can only be included in a regression analysis by creating dummy variables.

Dummy variables can only be interpreted in relation to their reference category.

Adding an additional polynomial term represents another inflection point in the curvilinear relationship.

Quadratic and cubic polynomials are generally sufficient to represent most curvilinear relationships.

Assessing the significance of a polynomial or interaction term is accomplished by evaluating incremental  $R^2$ , not the significance of individual coefficients, due to high multicollinearity.

## Mediation and Moderation

Selecting a variable to be either a mediator or moderator should be based on conceptual rather than empirical grounds.

## Stage 3: Assumptions in Multiple Regression Analysis

We have shown how improvements in prediction of the dependent variable are possible by adding independent variables and even transforming them to represent aspects of the relationship that are not linear. To do so, however, we must make several assumptions about the relationships between the dependent and independent variables that affect the statistical procedure (least squares) used for multiple regression. In the following sections we discuss testing for the assumptions and corrective actions to take if violations occur.

The basic issue is whether, in the course of calculating the regression coefficients and predicting the dependent variable, the assumptions of regression analysis are met. Are the errors in prediction a result of an actual absence of a relationship among the variables, or are they caused by some characteristics of the data not accommodated by the regression model? The assumptions to be examined are in four areas:

- 1 Linearity of the phenomenon measured
- 2 Constant variance of the error terms
- 3 Normality of the error term distribution
- 4 Independence of the error terms.

### ASSESSING INDIVIDUAL VARIABLES VERSUS THE VARIATE

Before addressing the individual assumptions, we must first understand that the assumptions underlying multiple regression analysis apply both to the individual variables (dependent and independent) and to the relationship as a whole. Chapter 2 examined the available methods for assessing the assumptions for individual variables. In multiple regression, once the variate is derived, it acts collectively in predicting the dependent variable, *which necessitates assessing the assumptions not only for individual variables but also for the variate itself*. This section focuses on examining the variate and its relationship with the dependent variable for meeting the assumptions of multiple regression. These analyses actually must be performed *after* the regression model has been estimated in Stage 4. Thus, testing for assumptions must occur not only in the initial phases of the regression, but also after the model has been estimated.

A common question is posed by many researchers: Why examine the individual variables when we can just examine the variate and avoid the time and effort expended in assessing the individual variables? The answer rests in the insight gained in examining the individual variables in two areas:

- Have assumption violations for individual variable(s) caused their relationships to be misrepresented?
- What are the sources and remedies of any assumptions violations for the variate?

Only with a thorough examination of the individual variables will the researcher be able to address these two key questions. If only the variate is assessed, then the researcher not only has little insight into how to correct any problems, but perhaps more importantly does not know what opportunities were missed for better representations of the individual variables and, ultimately, the variate.

## METHODS OF DIAGNOSIS

The principal measure of prediction error for the variate is the *residual*—the difference between the observed and predicted values for the dependent variable. When examining residuals, some form of standardization is recommended to make the residuals directly comparable. (In their original form, larger predicted values naturally have larger residuals.) The most widely used standardization approach is the **studentized residual**, which differs from other methods in how it calculates the standard deviation. To minimize the effect of any observation on the standardization process, the standard deviation of the residual for observation is computed from regression estimates omitting the *i*th observation in the calculation of the regression estimates. The studentized residual values correspond to t values, making it quite easy to assess the statistical significance of particularly large residuals.

**Graphical Diagnostics** Plotting the residuals versus the independent or predicted variables is a basic method of identifying assumption violations for the overall relationship. Use of the residual plots, however, depends on several key considerations.

**BASIC RESIDUAL PLOT** The most common residual plot involves the residuals ( $r_i$ ) versus the predicted dependent values ( $\hat{Y}_i$ ). For a simple regression model, the residuals may be plotted against either the dependent or independent variables, because they are directly related. In multiple regression, however, only the predicted dependent values represent the total effect of the regression variate. Thus, unless the residual analysis intends to concentrate on only a single variable, the predicted dependent variables are used.

**ASSESSING VIOLATIONS** Violations of each assumption can be identified by specific patterns of the residuals. Figure 5.11 contains a number of residual plots that address the basic assumptions discussed in the following sections. One plot of special interest is the **null plot** (Figure 5.11a), the plot of residuals when all assumptions are met. The null plot shows the residuals falling randomly, with relatively equal dispersion about zero and no strong tendency to be either greater or less than zero. Likewise, no pattern is found for large versus small values of the independent variable. The remaining residual plots will be used to illustrate methods of examining for violations of the assumptions underlying regression analysis. In the following sections, we examine a series of statistical tests that can complement the visual examination of the residual plots.

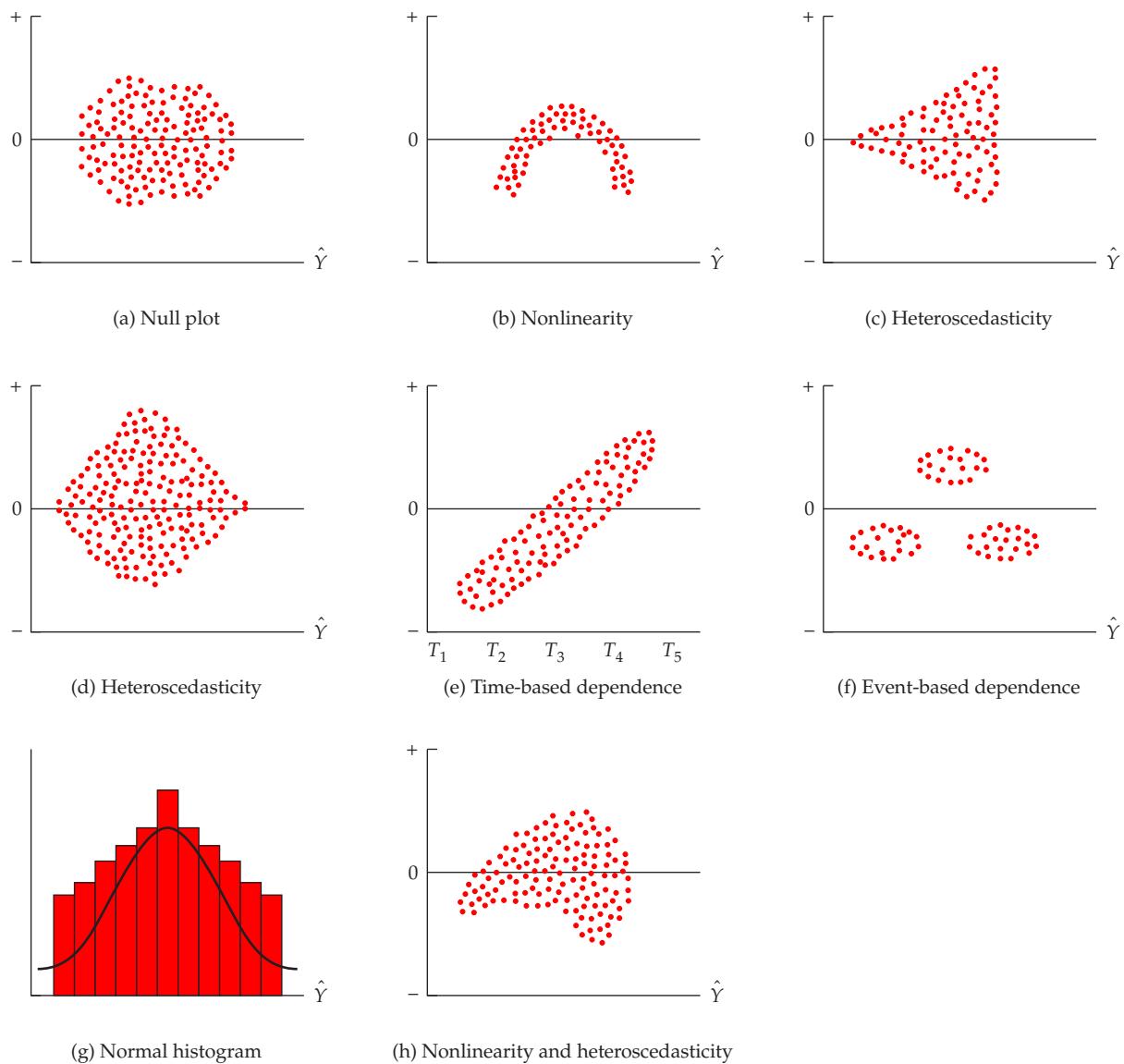
## LINEARITY OF THE PHENOMENON

The **linearity** of the relationship between dependent and independent variables represents the degree to which the change in the dependent variable is associated with the independent variable. The regression coefficient is assumed to be constant across the range of values for the independent variable. The concept of correlation, the measure of association underlying regression analysis, is based on a linear relationship, thus making it a critical issue in representing the “true” relationship between variables in the analysis. Moreover, violations of the linearity assumption are not overcome by increasing the sample size, as is the case with other assumptions (e.g., normality).

Linearity of any bivariate relationship is easily examined through residual plots. Figure 5.11b shows a typical pattern of residuals indicating the existence of a nonlinear relationship not represented in the current model. Any consistent curvilinear pattern in the residuals indicates that corrective action will increase both the predictive accuracy of the model and the validity of the estimated coefficients. Corrective action can take one of three forms:

- Transforming the data values (e.g., logarithm, square root, etc.) of one or more independent variables to achieve linearity is discussed in Chapter 2 [83].

**Figure 5.11**  
Graphical Analysis of Residuals



- Directly including the nonlinear relationships in the regression model, such as through the creation of polynomial terms as discussed in Stage 2.
- Using specialized methods such as nonlinear regression specifically designed to accommodate the curvilinear effects of independent variables or more complex nonlinear relationships.

**Identifying the Independent Variables for Action** How do we determine which independent variable(s) to select for corrective action? In multiple regression with more than one independent variable, an examination of the residuals shows only the combined effects of all independent variables, but we cannot examine any independent variable separately in a residual plot. To do so, we use what are called **partial regression plots**, which show the relationship of a single independent variable to the dependent variable, controlling for the effects of all other independent variables. As such, the partial regression plot portrays the unique relationship between dependent and independent variables.

These plots differ from the residual plots just discussed because the line running through the center of the points, which was horizontal in the earlier plots (refer to Figure 5.11), will now slope up or down depending on whether the regression coefficient for that independent variable is positive or negative.

Examining the observations around this line is done exactly as before, but now the curvilinear pattern indicates a nonlinear relationship between a specific independent variable and the dependent variable. This method is more useful when several independent variables are involved, because we can tell which specific variables violate the assumption of linearity and apply the needed remedies only to them. Also, the identification of outliers or influential observations is facilitated on the basis of one independent variable at a time.

## CONSTANT VARIANCE OF THE ERROR TERM

The presence of unequal variances (**heteroscedasticity**) is one of the most common assumption violations. In these instances, the error terms (residuals) are not constant across the range of the independent variable. This lack of constant variance in the residuals does not bias the estimated coefficients, but it does cause inaccurate estimation of the standard errors of the estimates (most often underestimated). This can cause inflated Type I error rates or decreased statistical power [98].

**Diagnosis** Diagnosis is made with residual plots or simple statistical tests. Plotting the residuals (studentized) against the predicted dependent values and comparing them to the null plot (see Figure 5.11a) shows a consistent pattern if the variance is not constant. Perhaps the most common pattern is triangle-shaped in either direction (Figure 5.11c). A diamond-shaped pattern (Figure 5.11d) can be expected in the case of percentages where more variation is expected in the midrange than at the tails. Many times, a number of violations occur simultaneously, such as in nonlinearity and heteroscedasticity (Figure 5.11h). Remedies for one of the violations often correct problems in other areas as well.

Each statistical computer program has statistical tests for heteroscedasticity. For example, IBM SPSS provides the Levene test for homogeneity of variance, which measures the equality of variances for a single pair of variables. Its use is particularly recommended because it is less affected by departures from normality, another frequently occurring problem in regression.

**Remedies** If heteroscedasticity is present, three remedies are available:

**VARIABLE TRANSFORMATION** The most direct remedy is the transformation of the offending variable(s) with a one of the variance-stabilizing transformations discussed in Chapter 2. After transformation the variables should exhibit homoscedasticity and can be used directly in the regression model. The disadvantage of this approach, however, is that the transformation process many times complicates interpretation of the transformed variable.

**WEIGHTED LEAST SQUARES** A second remedy is to use the procedure of weighted least squares analysis. This procedure “weights” each observation based on its variance and thus mitigates the variations in variance of the residuals seen in heteroscedasticity. The use of this approach, however, requires a series of assumptions concerning the distribution of the residuals and is a more complicated estimation process.

**HETEROSCEDASTICITY-CONSISTENT STANDARD ERRORS** A third approach that has become more widespread in recent years is the use of robust standard errors or heteroscedasticity-consistent standard errors (HCSE) [75, 79, 124]. These estimates of the standard errors are corrected for any heteroscedasticity that may be present and thus are a much simpler and direct option than either variable transformations or weighted least squares [56]. There are a series of estimates of HCSE (HC1, HC2, HC3 and HC4), but HC3 or HC4 are generally regarded as the most appropriate to use [98]. Direct estimates of HCSE are available in the major software programs, either as an option or supplemental analysis [56]. At a minimum the researcher should obtain these “corrected” standard errors and compare them to the original standard errors to assess the degree of impact of heteroscedasticity.

## NORMALITY OF THE ERROR TERM DISTRIBUTION

While technically the assumption of normality applies only to the error terms/residuals, any attempt to remedy non-normality involves assessing the non-normality of the independent or dependent variables or both [105]. The simplest diagnostic for the set of independent variables in the equation is a histogram of residuals, with a visual check for a distribution approximating the normal distribution (see Figure 5.11g). Although attractive because of its simplicity, this method is particularly difficult in smaller samples, where the distribution is often not normal. A better method is the use of **normal probability plots**. They differ from residual plots in that the standardized residuals are compared with the normal distribution. The normal distribution makes a straight diagonal line, and the plotted residuals are compared with the diagonal. If a distribution is normal, the residual line closely follows the diagonal. The same procedure can compare the dependent or independent variables separately to the normal distribution [33]. Chapter 2 provides a more detailed discussion of the interpretation of normal probability plots.

Regression analysis is generally considered robust to violations of normality when the sample size exceeds 200 observations, but researchers are always encouraged to make an assessment of the normality of the residuals to identify problematic issues. In smaller samples variables can be transformed to achieve normality to correct for the assumption violations. When the dependent variables are known to follow non-normal distributions (e.g., counts, proportions or probabilities, binary variables) researchers are encouraged to explore the use of generalized linear models (see Chapter 1 for more discussion) that explicitly incorporate these error term distributions other than the normal distribution. This provides the researcher with a method for avoiding transformations of the dependent measure just to achieve normality.

## INDEPENDENCE OF THE ERROR TERMS

We assume in regression that each predicted value is independent, which means that the predicted value is not related to any other prediction; that is, they are not grouped or sequenced by any variable. We can best identify such an occurrence by plotting the residuals against any possible grouping or sequencing variable. If the residuals are independent, the pattern should appear random and similar to the null plot of residuals. Violations will be identified by a consistent pattern in the residuals. Figure 5.11e displays a residual plot that exhibits an association between the residuals and time, a common sequencing variable. Another frequent pattern is shown in Figure 5.11f. This pattern occurs when basic model conditions change but are not included in the model. For example, swimsuit sales are measured monthly for 12 months, with two winter seasons versus a single summer season, yet no seasonal indicator is estimated. The residual pattern will show negative residuals for the winter months versus positive residuals for the summer months.

The types of grouping or sequencing variables fall into two basic classes: time series data and clustered observations. Time series data represents observations on the same unit (e.g., person or object) over multiple occasions. It is similar to repeated measures in many experimental situations (see Chapter 6 for more discussion). In both instances we assume that the observations for any individual/object are related and thus not independent. We can apply data transformations such as first differencing in time series analysis or centering in repeated measures. Both of these situations can be accommodated by panel analysis, an extension of multiple regression to accommodate both cross-sectional and time series/repeated measures in a single framework. We will discuss panel model later in this chapter.

The second type of grouping/sequencing variable is found when the data is hierarchically distributed (i.e., there are groups of observations that form a nested structure in the data). The classic example is in an educational setting, where individual students can be grouped by class, then classes within schools and so on. These groups can all be interrelated within the group (e.g., the common impact of one teacher versus another) and thus violate the independence assumption. A class of models termed multilevel or hierarchical models have been developed to specifically address this issue and provide a remedy for the dependence among observations. This class of models will also be discussed later in the chapter along with panel models as extensions of regression analysis to address these types of research situations.

## Assessing Statistical Assumptions

Testing assumptions must be done not only for each dependent and independent variable, but for the variate as well.

Graphical analyses (i.e., partial regression plots, residual plots, and normal probability plots) are the most widely used methods of assessing assumptions for the variate.

Most remedies for problems found in the variate must be accomplished by modifying one or more independent variables as described in Chapter 2.

Heteroscedasticity can be remedied by use of heteroscedasticity-consistent standard errors (HCSE).

### SUMMARY

Analysis of residuals, whether with the residual plots or statistical tests, provides a simple yet powerful set of analytical tools for examining the appropriateness of our regression model. Too often, however, these analyses are not made, and the violations of assumptions are left intact. Thus, users of the results are unaware of the potential inaccuracies that may be present, which range from inappropriate tests of the significance of coefficients (either showing significance when it is not present, or vice versa) to the biased and inaccurate predictions of the dependent variable. We strongly recommend that these methods be applied for each set of data and regression model (see Rules of Thumb 5-4). Application of the remedies, particularly transformations of the data, will increase confidence in the interpretations and predictions from multiple regression. When transformations are not desirable, then alternative measures (e.g., HCSE estimates) are available along with alternative model forms (e.g., multilevel or panel models) when data transformations are not sufficient.

## Stage 4: Estimating the Regression Model and Assessing Overall Model Fit

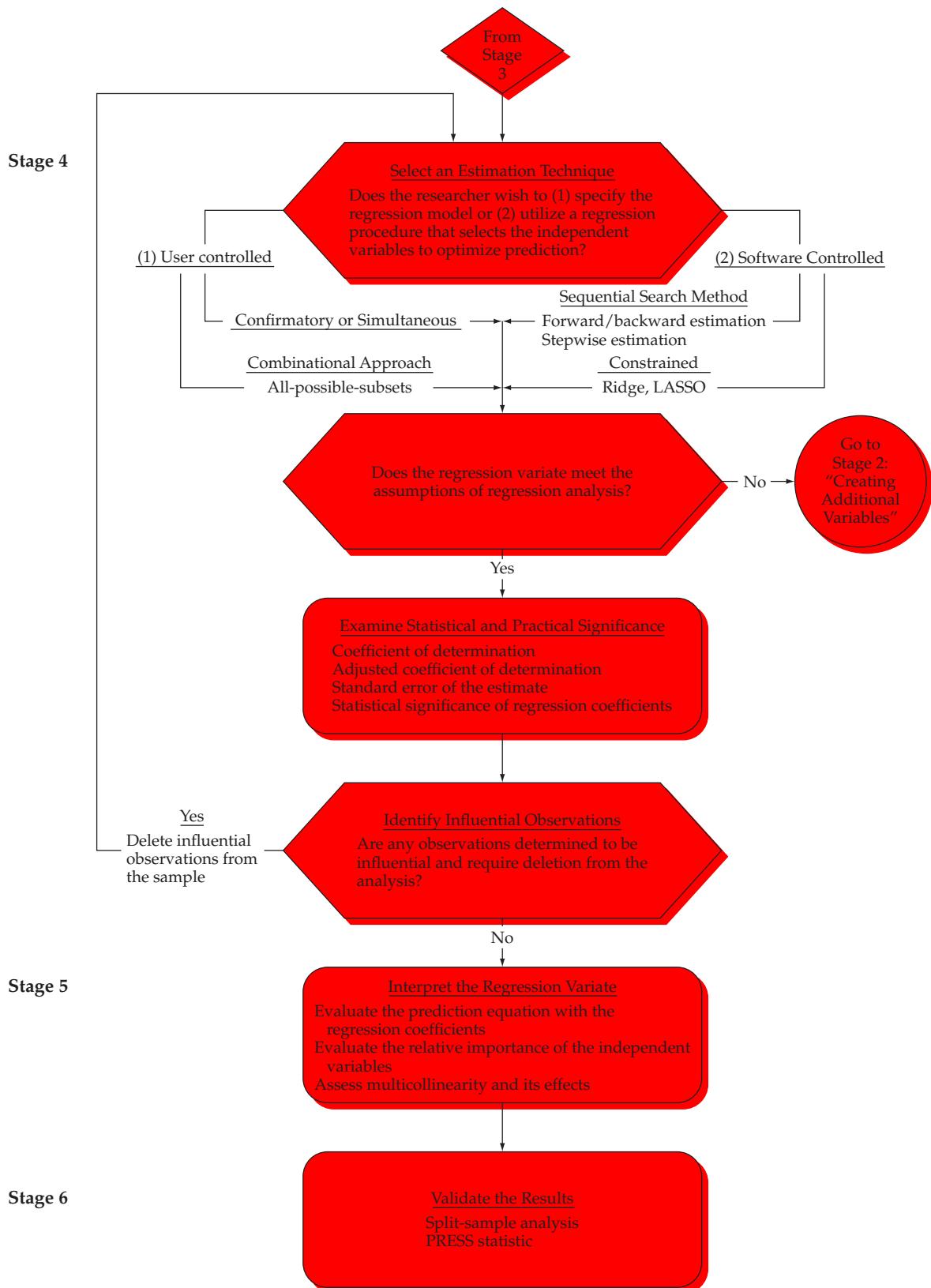
Having specified the objectives of the regression analysis, selected the independent and dependent variables, addressed the issues of research design, and assessed the variables for meeting the assumptions of regression, the researcher is now ready to estimate the regression model and assess the overall predictive accuracy of the independent variables (see Figure 5.12). In this stage, the researcher must accomplish three basic tasks:

- 1 Select a method for specifying the regression model to be estimated.
- 2 Assess the statistical significance of the overall model in predicting the dependent variable.
- 3 Determine whether any of the observations exert an undue influence on the results.

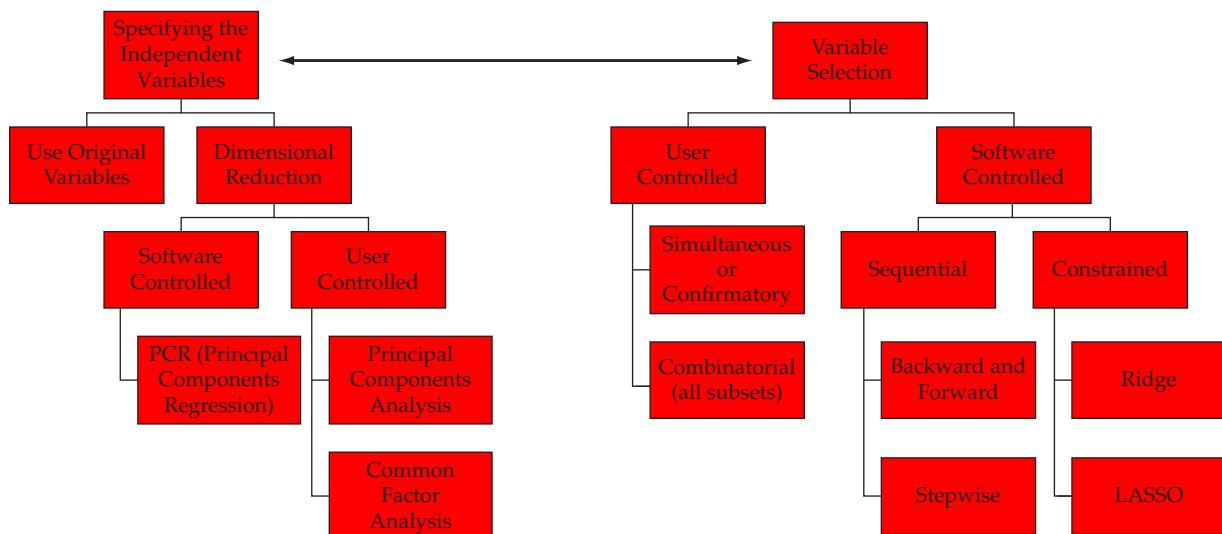
### MANAGING THE VARIATE

Perhaps the most critical task in any regression analysis is the correct specification of the variate in the final model. As we have discussed earlier, issues such as multicollinearity and specification error play a pivotal role in the results from the regression analysis and most of these issues are resolved as much by researcher judgment as empirical remedies. In Chapter 1 we introduced the framework of Managing The Variate (see Figure 5.13) that we will revisit here as it encompasses both the issues of variable specification (pre-analysis decisions) and variable selection (specifying the variate during the estimation process). The following sections will provide more detail on the various options available to the researcher in each of these decision areas.

**Figure 5.12**  
Stages 4–6 in the Multiple Regression Decision Diagram



**Figure 5.13**  
Managing The Variate



## VARIABLE SPECIFICATION

As discussed in Chapter 1, the basic decision facing the researcher at this stage is whether to use the independent variables in their original form or perform some form of dimensional reduction on the set of independent variables. Use of the original variables provides the researcher with direct measures of the variables of interest, which may be particularly important as explanation becomes an important objective. But as the number of variables increases the issue of interpretability emerges as described in Stage 1, especially as the researcher attempts to avoid specification error by not including relevant variables.

If dimensional reduction is performed, the researcher again can choose one of two options: perform the dimensional reduction before the analysis is conducted through some form of exploratory factor analysis (see Chapter 3 for a more extended discussion) or employ a software-controlled approach such as principal components regression where the software performs the dimensional reduction without researcher intervention. The advantages of dimensional reduction in any form have been discussed earlier when addressing the issues associated with multicollinearity and measurement error.

As a remedy for multicollinearity, redundant variables are formed into some form of composite variable that then replaces the individual variables in the analysis. The effect is to concentrate all of the predictive effect into a single measure versus having several measures all “sharing” in that effect. For example, measurement error composites provide greater reliability and validity about the central concept underlying the variables, relying on a “convergence” of the variables to represent a shared perspective on the concept that is much stronger than any single measure. The issue of reducing measurement error through composites is a fundamental element of structural equation modeling and we encourage researchers to acquaint themselves with the concept of a latent construct introduced in Chapters 9 and 10, and further explained in Chapters 11 to 13.

The appropriate choice is obvious in the extreme situations—a very few or a very large number of variables. But most research situations fall between the extremes and the researcher’s judgment has substantial impact. So we recommend researchers explore all alternatives to understand the implications of each approach. In our HBAT example later in the chapter we will examine regression models using all thirteen independent variables as well as substituting the factors developed in Chapter 3 in our exploratory factor analysis. In that manner we can see the conclusions shared between the two approaches as well as the unique conclusions for each approach.

## VARIABLE SELECTION

In most instances of multiple regression, the researcher has a number of possible independent variables from which to choose for inclusion in the regression equation. Sometimes the set of independent variables is exactly specified and the regression model is essentially used in a confirmatory approach. But in most instances, the researcher may choose to specify the variables to be included in the variate (by explicit specification or using the combinatorial approach) or use the estimation technique to pick and choose among the set of independent variables with either sequential search or constrained processes. The objective should always be to find the best regression model, whether through one or more of these approaches. Each of these four basic approaches is discussed next.

**Confirmatory or Simultaneous Specification** The simplest, yet perhaps most demanding, approach for specifying the regression model is to employ a confirmatory (also known as a simultaneous) approach wherein the researcher specifies the exact set of independent variables to be included. As compared with the other approaches to be discussed next, the researcher has total control over the variable selection. Although the confirmatory specification is simple in concept, the researcher is completely responsible for the trade-offs between more independent variables and greater predictive accuracy versus model parsimony, multicollinearity and concise explanation. Particularly problematic are specification errors of either omission or inclusion discussed earlier in Stage 1. Guidelines for model development are discussed in Chapter 1. The researcher must avoid being guided by empirical information and instead rely heavily on theoretical justification for a truly confirmatory approach.

**Combinatorial Approach** The other form of user-controlled variable selection is the combinatorial approach, primarily a generalized search process across all possible combinations of independent variables. The best-known procedure is **all-possible-subsets regression**, which is exactly as the name suggests. All possible combinations of the independent variables are examined, and the best-fitting set of variables is identified. For example, a model with 10 independent variables has 1,024 possible regressions (1 equation with only the constant, 10 equations with a single independent variable, 45 equations with all combinations of two variables, and so on). With computerized estimation procedures, this process can be managed today even for rather large problems, thereby identifying the best overall regression equation for any number of measures of predictive fit.

Usage of this approach has decreased due to criticisms of its (1) atheoretical nature and (2) lack of consideration of such factors as multicollinearity, the identification of outliers and influentials, and the interpretability of the results. When these issues are considered, the “best” equation may involve serious problems that affect its appropriateness, and another model may ultimately be selected. This approach can, however, provide insight into the number of regression models that are roughly equivalent in predictive power, yet possess quite different combinations of independent variables.

**Sequential Search Methods** In a marked contrast to the user-controlled approaches, sequential search methods have in common the general approach of estimating the regression equation by considering a set of variables defined by the researcher, and then a software algorithm selectively adding or deleting among these variables until some overall criterion measure is achieved. This approach provides an objective method for selecting variables that maximizes the prediction while employing the smallest number of variables. Two types of sequential search approaches are (1) stepwise estimation and (2) forward addition and backward elimination. In each approach, variables are individually assessed for their contribution to prediction of the dependent variable and added to or deleted from the regression model based on their relative contribution. The stepwise procedure is discussed and then contrasted with the forward addition and backward elimination procedures.

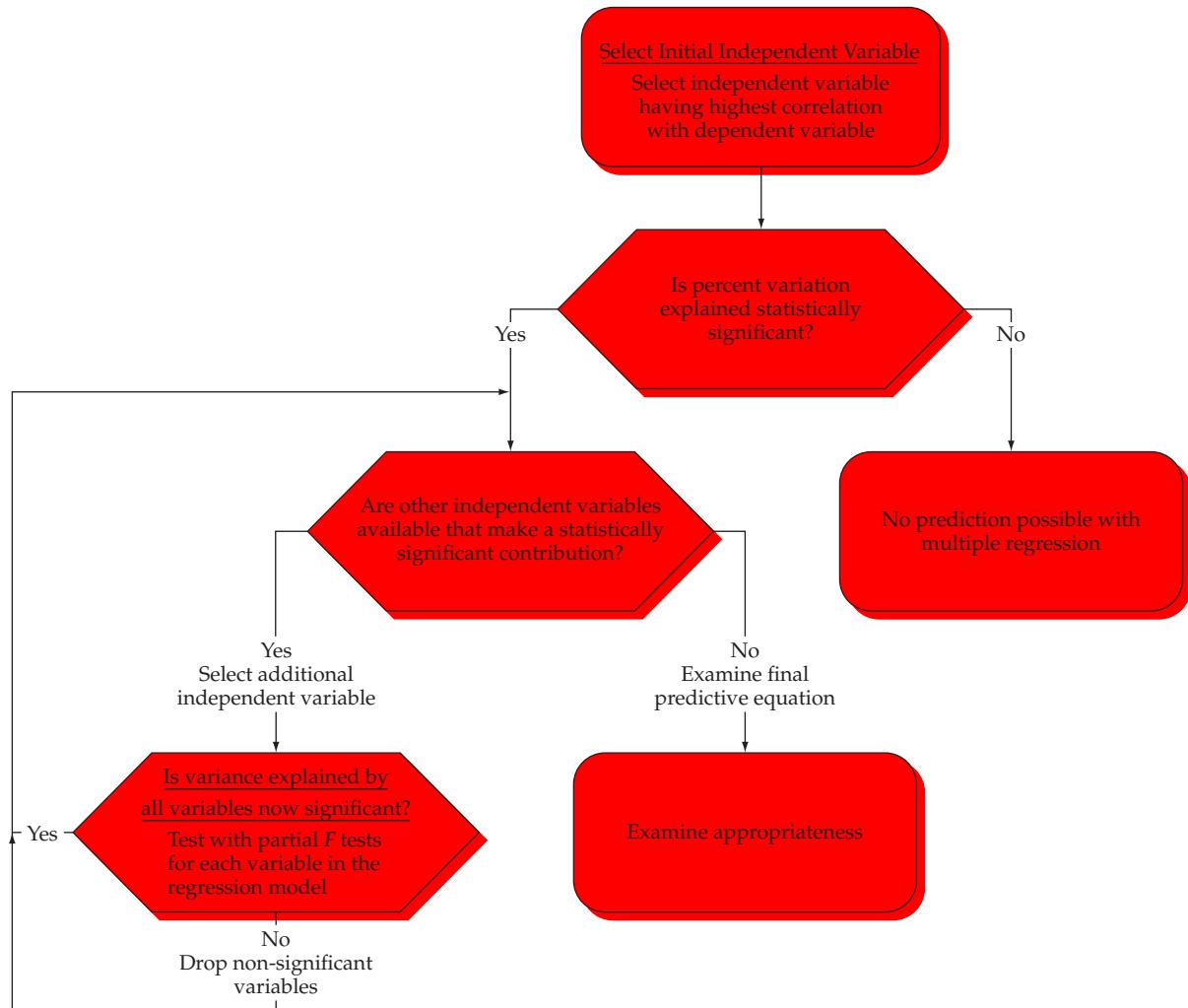
**STEPWISE ESTIMATION** Perhaps the most popular sequential approach to variable selection is **stepwise estimation**. This approach enables the researcher to examine the contribution of each independent variable to the regression model. Each variable is considered for inclusion prior to developing the equation. The independent variable with the greatest contribution is added first. Independent variables are then selected for inclusion based on their incremental

contribution over the variable(s) already in the equation. The stepwise procedure is illustrated in Figure 5.14. The specific issues at each stage are as follows:

- 1 Start with the simple regression model by selecting the one independent variable that is the most highly correlated with the dependent variable. The equation would be  $Y = b_0 + b_1X_1$ .
- 2 Examine the **partial correlation coefficients** to find an additional independent variable that explains the *largest statistically significant* portion of the unexplained (error) variance remaining from the first regression equation.
- 3 Recompute the regression equation using the two independent variables, and examine the **partial F value** for the original variable in the model to see whether it still makes a significant contribution, given the presence of the new independent variable. If it does not, eliminate the variable. This ability to eliminate variables already in the model distinguishes the stepwise model from the forward addition/backward elimination models. If the original variable still makes a significant contribution, the equation would be  $Y = b_0 + b_1X_1 + b_2X_2$ .
- 4 Continue this procedure by examining all independent variables not in the model to determine whether one would make a *statistically significant addition to the current equation* and thus should be included in a revised

**Figure 5.14**

Flowchart of the Stepwise Estimation Method



equation. If a new independent variable is included, examine all independent variables previously in the model to judge whether they should be kept.

- 5 Continue adding independent variables until none of the remaining candidates for inclusion would contribute a statistically significant improvement in predictive accuracy. This point occurs when all of the remaining partial regression coefficients are non-significant.

A potential bias in the stepwise procedure results from considering only one variable for selection at a time. Suppose variables  $X_3$  and  $X_4$  together would explain a significant portion of the variance (each given the presence of the other), but neither is significant by itself. In this situation, neither would be considered for the final model. Also, as will be discussed later, multicollinearity among the independent variables can substantially affect all sequential estimation methods.

**FORWARD ADDITION AND BACKWARD ELIMINATION** The procedures of **forward addition** and **backward elimination** are largely trial-and-error processes for finding the best regression estimates. The forward addition model is similar to the stepwise procedure in that it builds the regression equation starting with a single independent variable, whereas the backward elimination procedure starts with a regression equation including all the independent variables and then deletes independent variables that do not contribute significantly. *The primary distinction of the stepwise approach from the forward addition and backward elimination procedures is its ability to add or delete variables at each stage. Once a variable is added or deleted in the forward addition or backward elimination schemes, the action cannot be reversed at a later stage.* Thus, the ability of the stepwise method to add and delete makes it the preferred method among most researchers.

**CAVEATS TO SEQUENTIAL SEARCH METHODS** To many researchers, the sequential search methods seem the perfect solution to the dilemma faced in the confirmatory approach by achieving the maximum predictive power with only those variables that contribute in a statistically significant amount. Yet in the selection of variables for inclusion in the regression variate, three critical caveats markedly affect the resulting regression equation.

**Impact of Multicollinearity** The multicollinearity among independent variables has substantial impact on the final model specification. Let us examine the situation with two independent variables that have almost equal correlations with the dependent variable and are also highly correlated with each other. The criterion for inclusion or deletion in these approaches is to maximize the incremental predictive power of the additional variable. *If one of these variables enters the regression model, it is highly unlikely that the other variable will also enter because these variables are highly correlated and separately show little unique variance (see the later discussion on multicollinearity).* For this reason, the researcher must assess the effects of multicollinearity in model interpretation by not only examining the final regression equation, but also examining the direct correlations of all potential independent variables. This knowledge will help the researcher to avoid concluding that the independent variables that do not enter the model are inconsequential when in fact they may be highly related to the dependent variable, but also correlated with variables already in the model. Although the sequential search approaches will maximize the predictive ability of the regression model, the researcher must be careful in using these methods in establishing the impact of independent variables without considering multicollinearity among independent variables.

**Loss of Researcher Control** All sequential search methods create a loss of control on the part of the researcher. Even though the researcher does specify the variables to be considered for the regression variate, it is the estimation technique, interpreting the empirical data, that specifies the final regression model. In many situations, complications such as multicollinearity can result in a final regression model that achieves the highest levels of predictive accuracy, but has little managerial relevance in terms of variables included and so on. Yet in such instances, what recourse does the researcher have? The ability to specify the final regression model has been relinquished by the researcher. Use of these estimation techniques must consider trade-offs between the advantages found in these methods versus the lack of control in establishing the final regression model.

**Increased Alpha Level** The third caveat pertains primarily to the stepwise procedure. In this approach, multiple significance tests are performed in the model estimation process. To ensure that the overall error rate across all significance tests is reasonable, the researcher should employ more conservative thresholds (e.g., .01) in adding or deleting variables.

The sequential estimation methods have become widely used because of their efficiency in selecting that subset of independent variables that maximizes the predictive accuracy. With this benefit comes the potential for misleading results in explanation where only one of a set of highly correlated variables is entered into the equation and a loss of control in model specification. These potential issues do not suggest that sequential search methods should be avoided, just that the researcher must realize the issues (pro and con) involved in their use.

**Constrained Methods** The final approach to variable selection is an emerging set of techniques which act upon the actual regression estimates to “shrink” the estimates based upon their variance. These techniques are especially useful when (a) there are high degrees of multicollinearity or (b) the number of variables exceeds the number of observations in the sample.

**RIDGE REGRESSION** The first constrained method employing shrinkage was ridge regression, first proposed in 1970 as a means of combating high levels of multicollinearity [59]. The regression estimates are “shrunk” based on  $\lambda$ , the tuning parameter or ridge estimator. The objective is to reduce the regression estimates that are “inflated” by the effects of multicollinearity closer towards their true values. The tuning parameter controls the degree of shrinkage, so that if it is set to zero then no shrinkage occurs (estimates equal original regression coefficients), but as the value is increased then all of the coefficients are decreased. In this manner the more important coefficients are highlighted as those of lesser value decrease towards zero [129]. One limitation of ridge regression is that it will not set any coefficient to zero, thus all variables are retained in the equation. Ridge regression works particularly well when there are a number of variables with coefficients that are small or even zero. If all of the coefficients are moderate and of relatively equal size, then the shrinkage is relatively constant across the set and little differentiation occurs. While default values of the tuning parameter generally work well, the researcher can vary the parameter to review the impacts on the parameter estimates.

**LASSO** A more recent development is LASSO (least absolute shrinkage and selection operator), which builds upon the ridge regression approach by adding a variable selection component as well [114, 115]. The variable selection component comes from the addition of an additional constraint on the estimated coefficients. The result is small coefficients can be reduced to zero, thus eliminating them from the variate. In this manner, LASSO both constrains the estimates to magnify the larger coefficients and forces the smaller coefficients to zero, thus eliminating them from the model. It works particularly well when the number of variables is very large, and even when the number of variables exceeds the sample size.

**Review of the Model Selection Approaches** Whether the user exerts control over the variable selection process or a software-controlled method is used, the most important criterion is the researcher’s substantive knowledge of the research context and any theoretical foundation that allows for an objective and informed perspective as to the variables to be included as well as the expected signs and magnitude of their coefficients (see Rules of Thumb 5-5). Without this knowledge, the regression results can have high predictive accuracy but little managerial or theoretical relevance. Each estimation method has advantages and disadvantages, such that no single method is always preferred over the other approaches. As such, the researcher should never totally rely on any one of these approaches without understanding how the implications of the estimation method relate to the researcher’s objectives for prediction and explanation and the theoretical foundation for the research. Many times the use of two or more methods in combination may provide a more balanced perspective for the researcher versus using only a single method and trying to address all of the issues affecting the results.

## TESTING THE REGRESSION VARIATE FOR MEETING THE REGRESSION ASSUMPTIONS

With the independent variables selected and the regression coefficients estimated, the researcher must now assess the estimated model for meeting the assumptions underlying multiple regression. As discussed in Stage 3, the individual

## Variable Specification

While the researcher may choose to use the original set of variables in model estimation, issues of multicollinearity in most situations warrants consideration of some form of dimensional reduction.

## Variable Selection

No matter which variable technique is chosen, theory must be a guiding factor in evaluating the final regression model because:

Confirmatory specification, the only method to allow direct testing of a prespecified model, is also the most complex from the perspectives of specification error, model parsimony, and achieving maximum predictive accuracy.

Combinatorial estimation, while considering all possible models, still removes control from the researcher in terms of final model specification, even though the researcher can view the set of roughly equivalent models in terms of predictive accuracy.

Sequential search (e.g., stepwise), although maximizing predictive accuracy, represents a completely “automated” approach to model estimation, leaving the researcher almost no control over the final model specification.

Constrained variable selection has the same caveats as sequential search methods.

No single method is best, and the prudent strategy is to employ a combination of approaches to capitalize on the strengths of each to reflect the theoretical basis of the research question.

variables must meet the assumptions of linearity, constant variance, independence, and normality. In addition to the individual variables, the regression variate must also meet these assumptions. The diagnostic tests discussed in Stage 3 can be applied to assessing the collective effect of the variate through examination of the residuals. If substantial violations are found, the researcher must take corrective actions on one or more of the independent variables and then re-estimate the regression model.

## EXAMINING THE STATISTICAL SIGNIFICANCE OF OUR MODEL

If we were to take repeated random samples of respondents and estimate a regression equation for each sample, we would not expect to get exactly the same values for the regression coefficients each time. Nor would we expect the same overall level of model fit. Instead, some amount of random variation due to sampling error should cause differences among many samples. From a researcher’s perspective, we take only one sample and base our predictive model on it. With only this one sample, we need to test the hypothesis that our regression model can represent the population rather than just our one sample. These statistical tests take two basic forms: a test of the variation explained (coefficient of determination) and a test of each regression coefficient.

**Significance of the Overall Model: Testing the Coefficient of Determination** To test the hypothesis that the amount of variation explained by the regression model is more than the baseline prediction (i.e., that  $R^2$  is significantly greater than zero), the  $F$  ratio is calculated as:

$$F \text{ ratio} = \frac{\frac{\text{SS}_{\text{regression}}}{df_{\text{regression}}}}{\frac{\text{SS}_{\text{residual}}}{df_{\text{residual}}}}$$

where

$df_{\text{regression}}$ =Number of estimated coefficients (including intercept) – 1

$df_{\text{residual}}$ =Sample size – Number of estimated coefficients (including intercept)

Three important features of this ratio should be noted:

**RATIO OF VARIANCES** Dividing each sum of squares by its appropriate degrees of freedom ( $df$ ) results in an estimate of the variance. The top portion of the  $F$  ratio is the variance explained by the regression model, while the bottom portion is the unexplained variance.

**SIZE OF  $R^2$**  Intuitively, if the ratio of the explained variance to the unexplained is high, the regression variate must be of significant value in explaining the dependent variable. Using the  $F$  distribution, we can make a statistical test to determine whether the ratio is different from zero (i.e., statistically significant). In those instances in which it is statistically significant, the researcher can feel confident that the regression model is not specific to just this sample, but would be expected to be significant in multiple samples from this population.

**ASSESS PRACTICAL SIGNIFICANCE** Although larger  $R^2$  values result in higher  $F$  values, the researcher must base any assessment of practical significance separate from statistical significance. Because statistical significance is really an assessment of the impact of sampling error, the researcher must be cautious of always assuming that statistically significant results are also practically significant. This caution is particularly relevant in the case of large samples where even small  $R^2$  values (e.g., 5% or 10%) can be statistically significant, but such levels of explanation would not be acceptable for further action on a practical basis.

In our example of credit card usage, the  $F$  ratio for the simple regression model is  $(16.5 \div 1) / (5.50 \div 6) = 18.0$ . The tabled  $F$  statistic of 1 with 6 degrees of freedom at a significance level of .05 yields 5.99. Because the  $F$  ratio is greater than the table value, we reject the hypothesis that the reduction in error we obtained by using family size to predict credit card usage was a chance occurrence. This outcome means that, considering the sample used for estimation, we can explain 18 times more variation than when using the average, which is not likely to happen by chance (less than 5% of the time). Likewise, the  $F$  ratio for the multiple regression model with two independent variables is  $(18.96 \div 2) / (3.04 \div 5) = 15.59$ . The multiple regression model is also statistically significant, indicating that the additional independent variable was substantial in adding to the regression model's predictive ability.

**Adjusting the Coefficient of Determination** As was discussed earlier in defining degrees of freedom, the addition of a variable will always increase the  $R^2$  value. This increase then creates concern with generalizability because  $R^2$  will increase even if non-significant predictor variables are added. The impact is most noticeable when the sample size is close in size to the number of predictor variables (termed *overfitting*—when the degrees of freedom is small). With this impact minimized when the sample size greatly exceeds the number of independent variables, a number of guidelines have been proposed as discussed earlier (e.g., 10 to 15 observations per independent variable to a minimum of 5 observations per independent variable). Yet what is needed is a more objective measure relating the level of overfitting to the  $R^2$  achieved by the model.

This measure involves an adjustment based on the number of independent variables relative to the sample size. In this way, adding non-significant variables just to increase the  $R^2$  can be discounted in a systematic manner. As part of all regression programs, an **adjusted coefficient of determination (adjusted  $R^2$ )** is given along with the coefficient of determination. Interpreted the same as the unadjusted coefficient of determination, the adjusted  $R^2$  decreases as we have fewer observations per independent variable. The adjusted  $R^2$  value is particularly useful in comparing across regression equations involving different numbers of independent variables or different sample sizes because it makes allowances for the degrees of freedom for each model.

In our example of credit card usage,  $R^2$  for the simple regression model is .751, and the adjusted  $R^2$  is .709. As we add the second independent variable,  $R^2$  increases to .861, but the adjusted  $R^2$  increases to only .806. When we add the third variable,  $R^2$  increases to only .872, and the adjusted  $R^2$  decreases to .776. Thus, while we see the  $R^2$  always increased when adding variables, the decrease in the adjusted  $R^2$  when adding the third variable indicates an overfitting of the data. When we discuss assessing the statistical significance of regression coefficients in the next section, we will see that the third variable was not statistically significant. The adjusted  $R^2$  not only reflects overfitting, but also the addition of variables that do not contribute significantly to predictive accuracy.

**Significance Tests of Regression Coefficients** Statistical significance testing for the estimated coefficients in regression analysis is appropriate and necessary when the analysis is based on a sample of the population rather than a census. When using a sample, the researcher is not just interested in the estimated regression coefficients for that sample, but is also interested in how the coefficients are expected to vary across repeated samples. The interested reader can find more detailed discussion of the calculations underlying significance tests for regression coefficients in the Basic Stats appendix on the text's websites.

**ESTABLISHING A CONFIDENCE INTERVAL** Significance testing of regression coefficients is a statistically-based probability estimate of whether the estimated coefficients across a large number of samples of a certain size will indeed be different from zero. To make this judgment, a confidence interval must be established around the estimated coefficient. If the confidence interval does not include the value of zero, then it can be said that the coefficient's difference from zero is statistically significant. To make this judgment, the researcher relies on three concepts:

**Significance Level** Establishing the **significance level (alpha)** denotes the chance the researcher is willing to take of being wrong about whether the estimated coefficient is different from zero. A value typically used is .05. As the researcher desires a smaller chance of being wrong and sets the significance level smaller (e.g., .01 or .001), the statistical test becomes more demanding. Increasing the significance level to a higher value (e.g., .10) allows for a larger chance of being wrong, but also makes it easier to conclude that the coefficient is different from zero.

**Sampling Error** The reason for variation in the estimated regression coefficients for each sample drawn from a population is sampling error. For small sample sizes, the sampling error is larger and the estimated coefficients will most likely vary widely from sample to sample. As the size of the sample increases, the samples become more representative of the population (i.e., sampling error decreases), and the variation in the estimated coefficients for these large samples become smaller. This relationship holds true until the analysis is estimated using the population. Then the need for significance testing is eliminated because the sample is equal to, and thus perfectly representative of, the population (i.e., no sampling error).

**Standard Error** The expected variation of the estimated coefficients (both the constant and the regression coefficients) due to sampling error is represented by the **standard error**. The standard error acts like the standard deviation of a variable by representing the expected dispersion of the *coefficients* estimated from repeated samples of this size.

With the significance level selected and the standard error calculated, we can establish a confidence interval for a regression coefficient based on the standard error just as we can for a mean based on the standard deviation. For example, setting the significance level at .05 would result in a confidence interval of  $\pm 1.96 \times$  standard error, denoting the outer limits containing 95 percent of the coefficients estimated from repeated samples.

**APPLYING THE CONFIDENCE INTERVAL** With the confidence interval in hand, the researcher now must ask three questions about the statistical significance of any regression coefficient:

- 1 *Was statistical significance established?* The researcher sets the significance level from which the confidence interval is derived (e.g., a significance level of 5 percent for a large sample establishes the confidence interval at  $\pm 1.96 \times$  standard error). A coefficient is deemed statistically significant if the confidence interval does not include zero.
- 2 *How does the sample size come into play?* If the sample size is small, sampling error may cause the standard error to be so large that the confidence interval includes zero. However, if the sample size is larger, the test has greater precision because the variation in the coefficients becomes less (i.e., the standard error is smaller). Larger samples do not guarantee that the coefficients will not equal zero, but instead make the test more precise.
- 3 *Does it provide practical significance in addition to statistical significance?* As we saw in assessing the  $R^2$  value for statistical significance, just because a coefficient is statistically significant does not guarantee that it also is practically significant. Be sure to evaluate the sign and size of any significant coefficient to ensure that it meets the research needs of the analysis.

Significance testing of regression coefficients provides the researcher with an empirical assessment of their “true” impact. Although it is not a test of validity, it does determine whether the impacts represented by the coefficients are generalizable to other samples from this population.

## UNDERSTANDING INFLUENTIAL OBSERVATIONS

Up to now, we focused on identifying general patterns within the entire set of observations. Here we shift our attention to individual observations, with the objective of finding the observations that:

- lie outside the general patterns of the data set
  - or
- strongly influence the regression results.

These observations are not necessarily “bad” in the sense that they must be deleted. In many instances they represent the distinctive elements of the dataset. However, we must first identify them and assess their impact before proceeding [13]. This section introduces the concept of influential observations and their potential impact on the regression results.

**Types of Influential Observations** Influential observations, in the broadest sense, include all observations that have a disproportionate effect on the regression results. The three basic types are based upon the nature of their impact on the regression results:

**OUTLIERS** Observations that have large residual values and can be identified only with respect to a specific regression model are **outliers** [2]. Outliers were traditionally the only form of influential observation considered in regression models, and specialized regression methods (e.g., robust regression) were even developed to deal specifically with outliers’ impact on the regression results [8, 100]. Chapter 2 provides additional procedures for identifying outliers as well as a discussion in a later section of this chapter.

**LEVERAGE POINTS** Observations that are distinct from the remaining observations based on their independent variable values are known as **leverage points**. Their impact is particularly noticeable in the estimated coefficients for one or more of the independent variables.

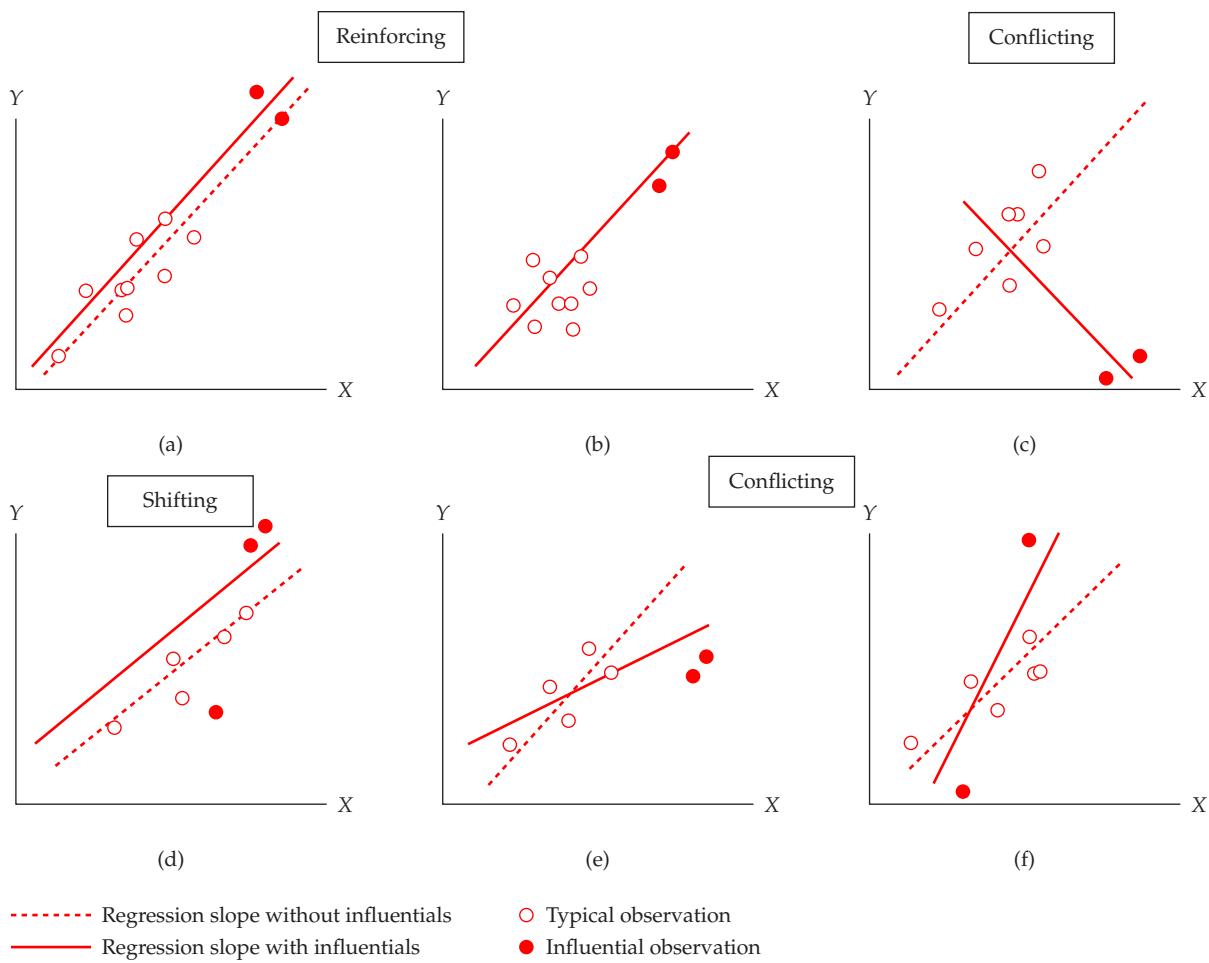
**INFLUENTIAL OBSERVATIONS** This is the broadest category, including all observations that have a disproportionate effect on the regression results. Influential observations potentially include outliers and leverage points but may include other observations as well. Also, not all outliers and leverage points are necessarily influential observations.

**Impacts of Influential Observations** Influential observations many times are difficult to identify through the traditional analysis of residuals when looking for outliers. Their patterns of residuals would go undetected because the residual for the influential points (the perpendicular distance from the point of the estimated regression line) would not be so large as to be classified as an outlier. Thus, focusing only on large residuals would generally ignore these influential observations. Figure 5.15 illustrates several types of impacts of influential observations and their corresponding pattern of residuals:

**REINFORCING** In Figure 5.15a, the influential point is a “good” one, reinforcing the general pattern of the data and lowering the standard error of the prediction and coefficients. It is a leverage point but has a small or zero residual value, because it is predicted well by the regression model. In Figure 5.15b, two influential observations almost totally account for the observed relationship, because without them no real pattern emerges from the other data points. They also would not be identified if only large residuals were considered, because their residual value would be small.

**CONFLICTING** Influential points can have an effect that is *contrary* to the general pattern of the remaining data but still have small residuals. In Figure 5.15c, we see a profound effect in which the influential observations counteract the general pattern of all the remaining data. In this case, the real data would have larger residuals than the bad

**Figure 5.15**  
Patterns of Influential Observations



influential points. Multiple influential points may also work toward the same result. In Figure 5.15e, two influential points have the same relative position, making detection somewhat harder. In Figure 5.15f, influentials have quite different positions but a similar effect on the results.

**SHIFTING** Influential observations may affect all of the results in a similar manner. One example is shown in Figure 5.15d, where the slope remains constant but the intercept is shifted. Thus, the relationship among all observations remains unchanged except for the shift in the regression model. Moreover, even though all residuals would be affected, little in the way of distinguishing features among them would assist in diagnosis.

These examples illustrate that we must develop an expanded toolkit of methods for identifying these influential cases. Procedures for identifying all types of influential observations are becoming quite widespread, yet are still less well known and utilized infrequently in regression analysis. All computer programs provide an analysis of residuals from which those with large values (particularly standardized residuals greater than 2.0) can be easily identified. Moreover, most computer programs now provide at least some of the diagnostic measures for identifying leverage points and other influential observations.

**Identifying Influential Observations** In the following discussion, we discuss a four-step process of identifying outliers, leverage points, and influential observations. As noted before, an observation may fall into one or more of these classes, and the course of action to be taken depends on the judgment of the researcher, based on the best available evidence.

**STEP 1: EXAMINING RESIDUALS AND PARTIAL REGRESSION PLOTS** Residuals are instrumental in detecting violations of model assumptions, and they also play a role in identifying observations that are outliers on the dependent variable once the model has been estimated. We employ two methods of detection: the analysis of residuals and partial regression plots.

**Analysis of Residuals** The residual is the primary means of classifying an observation as an outlier. The residual for the  $i_{th}$  observation is calculated as the actual minus predicted values of the dependent variable, or:

$$\text{Residual}_i = Y_i - \hat{Y}_i$$

The residual can actually take many forms based on the results of two procedures: the cases used for calculating the predicted value, and the use (or nonuse) of some form of standardization. We will review each procedure in the following sections and then discuss how they are “combined” to derive specific types of residuals.

- **Cases used in calculating the residual.** We have already seen how we calculate the residual using all of the observations, but a second form, the deleted residual, differs from the normal residual in that the  $i_{th}$  observation is omitted when estimating the regression equation used to calculate the predicted value for that observation. Thus, each observation has no impact on its own predicted value in the deleted residual. The deleted residual is less commonly used, although it has the benefit of reducing the influence of the observation on its calculation.
- **Standardizing the residual.** The second procedure in defining a residual involves whether to standardize the residuals. Residuals that are not standardized are in the scale of the dependent variable, which is useful in interpretation but gives no insight as to what is too large or small enough not to consider. **Standardized residuals** are the result of a process of creating a common scale by dividing each residual by the standard deviation of residuals. After standardization, the residuals have a mean of zero and a standard deviation of one. With a fairly large sample size (50 or above), standardized residuals approximately follow the  $t$  distribution, such that residuals exceeding a threshold such as 1.96 (the critical  $t$  value at the .05 confidence level) can be deemed statistically significant. A stricter test of significance has also been proposed, which accounts for multiple comparisons being made across various sample sizes [18].

A special form of standardized residual is the **studentized residual**. It is similar in concept to the deleted residual, but in this case the  $i_{th}$  observation is eliminated when deriving the standard deviation used to standardize the  $i_{th}$  residual. The studentized residual eliminates the case’s impact on the standardization process and offers a “less influenced” residual measure. It can be evaluated by the same criteria as the standardized residual.

The five types of residuals typically calculated by combining the options for calculation and standardization are (1) the normal residual, (2) the deleted residual, (3) the standardized residual, (4) the studentized residual, and (5) the studentized deleted residual. Each type of residual offers unique perspectives on both the predictive accuracy of the regression equation by its designation of outliers and the possible influences of the observation on the overall results.

**Partial Regression Plots** To graphically portray the impact of individual cases, the partial regression plot is most effective. Because the slope of the regression line of the partial regression plot is equal to the variable’s coefficient in the regression equation, an outlying case’s impact on the regression slope (and the corresponding regression equation coefficient) can be readily seen. The effects of outlying cases on individual regression coefficients are portrayed visually. Again, most computer packages have the option of plotting the partial regression plot, so the researcher need look only for outlying cases separated from the main body of observations. A visual comparison of the partial regression plots with and without the observation(s) deemed influential can illustrate their impact.

**STEP 2: IDENTIFYING LEVERAGE POINTS** Our next step is finding those observations that are substantially different from the remaining observations on one or more independent variables. These cases are termed **leverage points** in that they may “lever” the relationship in their direction because of their difference from the other observations. There are two measures generally used to identify leverage points: Hat values and Mahalanobis distance.

**Hat Matrix** When only two predictor variables are involved, plotting each variable on an axis of a two-dimensional plot will show those observations substantially different from the others. Yet, when a larger number of predictor variables are included in the regression equation, the task quickly becomes impossible through univariate methods. However, we are able to use a special matrix, the **hat matrix**, which contains values (**hat values**) for each observation that indicate leverage. The hat values represent the combined effects of all independent variables for each case.

Hat values (found on the diagonal of the hat matrix) measure two aspects of influence. First, for each observation, the hat value is a measure of the distance of the observation from the mean center of all other observations on the independent variables (similar to the Mahalanobis distance discussed next). Second, large diagonal values also indicate that the observation carries a disproportionate weight in determining its predicted dependent variable value, thus minimizing its residual. This is an indication of influence, because the regression line must be closer to this observation (i.e., strongly influenced) for the small residual to occur. This is not necessarily “bad,” as illustrated in the text, when the influential observations fall in the general pattern of the remaining observations.

What is a large hat value? The range of possible hat values is between 0 and 1, and the average value is  $p/n$ , where  $p$  is the number of predictors (the number of coefficients plus one for the constant) and  $n$  is the sample size. The rule of thumb for situations in which  $p$  is greater than 10 and the sample size exceeds 50 is to select observations with a leverage value greater than twice the average ( $2p/n$ ). When the number of predictors is less than 10 or the sample size is less than 50, use of three times the average ( $3p/n$ ) is suggested. The more widely used computer programs all have options for calculating and printing the leverage values for each observation. The analyst must then select the appropriate threshold value ( $2p/n$  or  $3p/n$ ) and identify observations with values larger than the threshold.

**Mahalanobis Distance** A measure comparable to the hat value is the **Mahalanobis distance ( $D^2$ )**, which considers only the distance of an observation from the mean values of the independent variables and not the impact on the predicted value. The Mahalanobis distance is another means of identifying outliers. It is limited in this purpose because threshold values depend on a number of factors, and a rule of thumb threshold value is not possible. It is possible, however, to determine statistical significance of the Mahalanobis distance from published tables [8]. Yet even without the published tables, the researcher can look at the values and identify any observations with substantially higher values than the remaining observations. For example, a small set of observations with the highest Mahalanobis values that are two to three times the next highest value would constitute a substantial break in the distribution and another indication of possible leverage.

**STEP 3: SINGLE-CASE DIAGNOSTICS** Up to now we have found outlying points on the predictor and criterion variables but have not formally estimated the influence of a single observation on the results. In this third step, all the methods rely on a common proposition: the most direct measure of influence involves deleting one or more observations and observing the changes in the regression results in terms of the residuals, individual coefficients, or overall model fit. The researcher then needs only to examine the values and select those observations that exceed the specified value. We have already discussed one such measure, the studentized deleted residual, but will now explore several other measures appropriate for diagnosing individual cases.

**Influences on Individual Coefficients** The impact of deleting a single observation on each regression coefficient is shown by the **DFBETA** and its standardized version the **SDFBETA**. Calculated as the change in the coefficient when the observation is deleted, DFBETA is the relative effect of an observation on each coefficient. Guidelines for identifying particularly high values of SDFBETA suggest that a threshold of  $\pm 1.0$  or  $\pm 2.0$  be applied to small sample sizes, whereas  $\pm 2/\sqrt{n}$  should be used for medium and larger data sets.

**Overall Influence Measures** These measures assess the impact on overall model fit. **Cook's distance ( $D_i$ )** is considered the single most representative measure. It captures the impact of an observation from two sources: the size of

changes in the predicted values when the case is omitted (outlying studentized residuals) as well as the observation's distance from the other observations (leverage). A rule of thumb is to identify observations with a Cook's distance of 1.0 or greater, although the threshold of  $4/(n - k - 1)$ , where  $n$  is the sample size and  $k$  is the number of independent variables, is suggested as a more conservative measure in small samples or for use with larger datasets. Even if no observations exceed this threshold, however, additional attention is dictated if a small set of observations has substantially higher values than the remaining observations.

A similar measure is the **COVRATIO**, which estimates the effect of the observation on the efficiency of the estimation process. Specifically, COVRATIO represents the degree to which an observation impacts the standard errors of the regression coefficients. It differs from the DFBETA and SDFBETA in that it considers all coefficients collectively rather than each coefficient individually. A threshold can be established at  $1 \pm 3p/n$ . Values above the threshold of  $1 + 3p/n$  make the estimation process more efficient, whereas those less than  $1 - 3p/n$  detract from the estimation efficiency. This allows the COVRATIO to act as another indicator of observations that have a substantial influence both positively and negatively on the set of coefficients.

A third measure is **SDFFIT**, the degree to which the fitted values change when the case is deleted. A cut-off value  $2\sqrt{(k + 1)/(n - k - 1)}$  has been suggested to detect substantial influence. Even though both Cook's distance and SDFFIT are measures of overall fit, they must be complemented by the measures of steps 1 and 2 to enable us to determine whether influence arises from the residuals, leverage, or both. An unstandardized version (**DFFIT**) is also available.

**STEP 4: SELECTING INFLUENTIAL OBSERVATIONS** The identification of influential observations is more a process of convergence by multiple methods than a reliance on a single measure. Because no single measure totally represents all dimensions of influence, it is a matter of interpretation, although these measures typically identify a small set of observations.

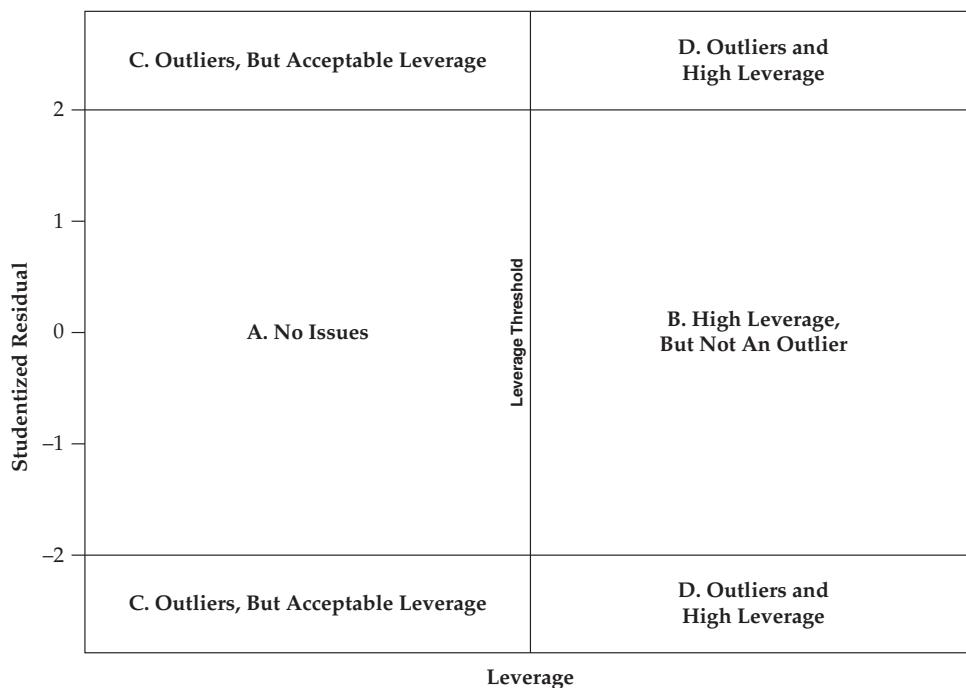
One particularly useful diagnostic tool is a plot of the two fundamental elements of an influential observation – leverage values versus standardized or studentized residuals (see Figure 5.16). By using the threshold values for studentized residuals ( $\pm 1.96$ ) and leverage, observations can be placed in one of four categories:

- A. No issues.** These are the observations that are fit well by the model (acceptable residuals) and also do not exhibit any extreme values on the independent variables (acceptable leverage).
- B. High leverage, but not an outlier.** These observations have high or extreme values on one or more independent variables, but are still predicted within acceptable limits by the regression model. They may act as the reinforcing influential observations described earlier.
- C. Outliers, but acceptable leverage.** These observations are outliers, but we can assume that the reason is not because of some extreme values on the independent variables, but instead some other fundamental problem. Any number of issues could relate to the poor fit, but it most likely is not because of the values of the independent variables.
- D. Outliers and high leverage.** This final set of observations are those that have very different values on the independent variables, yet are still not well predicted by the model. These most likely correspond to those conflicting influential observations that represent distinct impacts on the regression results.

While this graph does not provide a complete perspective of all aspects of outliers, it does provide the researcher with a concise portrayal of these two key elements of influential observations.

In addition to the leverage and residual graph, the researcher can also identify those observations with large values on the diagnostic measures. In doing so, first identify all observations exceeding the threshold values, and then examine the range of values in the dataset being analyzed and look for large gaps between the highest values and the remaining data. Some additional observations will be detected that should be classified as influential.

**Figure 5.16**  
Outlier and Leverage Diagnostics



**Remedies for Influentials** The need for additional study of leverage points and influentials is highlighted when we see the substantial extent to which the generalizability of the results and the substantive conclusions (the importance of variables, level of fit, and so forth) can be changed by only a small number of observations. Whether good (accentuating the results) or bad (substantially changing the results), these observations must be identified to assess their impact. Influentials, outliers, and leverage points are based on one of four conditions, each of which has a specific course of corrective action:

- 1 *An error in observations or data entry.* Remedy by correcting the data or deleting the case.
- 2 *A valid but exceptional observation that is explainable by an extraordinary situation.* Remedy by deletion of the case unless variables reflecting the extraordinary situation are included in the regression equation.
- 3 *An exceptional observation with no likely explanation.* Presents a special problem because it lacks reasons for deleting the case, but its inclusion cannot be justified either, suggesting analyses with and without the observations to make a complete assessment.
- 4 *An ordinary observation in its individual characteristics but exceptional in its combination of characteristics.* Indicates modifications to the conceptual basis of the regression model and should be retained.

In all situations, the researcher is encouraged to delete truly exceptional observations but still guard against deleting observations that, although different, are representative of the population. If, however, deletion cannot be justified, several more “robust” estimation techniques are available, among them robust regression [27, 101, 99]. Remember that the objective is to ensure the most representative model for the sample data so that it will best reflect the population from which it was drawn. This practice extends beyond achieving the highest predictive fit, because some outliers may be valid cases that the model should attempt to predict, even if poorly. The researcher should also be aware of instances in which the results would be changed substantially by deleting just a single observation or a small number of observations.

## Statistical Significance and Influential Observations

Always ensure practical significance when using large sample sizes, because the model results and regression coefficients could be deemed irrelevant even when statistically significant due just to the statistical power arising from large sample sizes.

Use the adjusted  $R^2$  as your measure of overall model predictive accuracy.

Statistical significance is required for a relationship to have validity, but statistical significance without theoretical support does not support validity.

Although outliers may be easily identifiable, the other forms of influential observations requiring more specialized diagnostic methods can be equal to or have even more impact on the results.

Use the case-wise diagnostics to better understand the impact of each observation and the potential sources of its impact—residuals or leverage.

## Stage 5: Interpreting the Regression Variate

The researcher's next task is to interpret the regression variate by evaluating the estimated regression coefficients for their explanation of the dependent variable. The researcher must evaluate not only the regression model that was estimated but also the potential independent variables that were omitted if a sequential search or combinatorial approach was employed. In those approaches, multicollinearity may substantially affect the variables ultimately included in the regression variate. Thus, in addition to assessing the estimated coefficients, the researcher must also evaluate the potential impact of omitted variables to ensure that the managerial significance is evaluated along with statistical significance.

### USING THE REGRESSION COEFFICIENTS

The estimated regression coefficients, termed the  $b$  coefficients, represent both the type of relationship (positive or negative) and the strength of the relationship between independent and dependent variables in the regression variate. The sign of the coefficient denotes whether the relationship is positive or negative, and the value of the coefficient indicates the change in the dependent value each time the independent variable changes by one unit.

For example, in the simple regression model of credit card usage with family size as the only independent variable, the coefficient for family size was .971. This coefficient denotes a positive relationship showing that as a family adds a member, credit card usage is expected to increase by almost one (.971) credit card. Moreover, if the family size decreases by one member, credit card usage would also decrease by almost one credit card (−.971).

The regression coefficients play two key functions in meeting the objectives of prediction and explanation for any regression analysis.

**Prediction** Prediction is an integral element in regression analysis, both in the estimation process as well as in forecasting situations. As described in the first section of the chapter, regression involves the use of a variate (the regression model) to estimate a single value for the dependent variable. This process is used not only to calculate the predicted values in the estimation procedure, but also with additional samples used for validation or forecasting purposes.

**ESTIMATION** First, in the ordinary least squares (OLS) estimation procedure used to derive the regression variate, a prediction of the dependent variable is made for each observation in the dataset. The estimation procedure sets the weights of the regression variate to minimize the residuals (e.g., minimizing the differences between predicted and

actual values of the dependent variable). No matter how many independent variables are included in the regression model, a single predicted value is calculated. As such, the predicted value represents the total of all effects of the regression model and allows the residuals, as discussed earlier, to be used extensively as a diagnostic measure for the overall regression model.

**FORECASTING** Although prediction is an integral element in the estimation procedure, the real benefits of prediction come in forecasting applications. A regression model is used in these instances for prediction with a set of observations not used in estimation. For example, assume that a sales manager developed a forecasting equation to forecast monthly sales of a product line. After validating the model, the sales manager inserts the upcoming month's expected values for the independent variables and calculates an expected sales value.

A simple example of a forecasting application can be shown using the credit card example. Assume that we are using the following regression equation that was developed to estimate the number of credit cards ( $Y$ ) held by a family:

$$Y = .286 + .635V_1 + .200V_2 + .272V_3$$

Now, suppose that we have a family with the following characteristics: Family size ( $V_1$ ) of 2 persons, family income ( $V_2$ ) of 22 (\$22,000), and number of autos ( $V_3$ ) being 3. What would be the expected number of credit cards for this family?

We would substitute the values for  $V_1$ ,  $V_2$ , and  $V_3$  into the regression equation and calculate the predicted value:

$$\begin{aligned} Y &= .286 + .635(2) + .200(22) + .272(3) \\ &= .286 + 1.270 + 4.40 + .819 \\ &= 6.775 \end{aligned}$$

Our regression equation would predict that this family would have 6.775 credit cards.

**Explanation** Many times the researcher is interested in more than just prediction. It is important for a regression model to have accurate predictions to support its validity, but many research questions are more focused on assessing the nature and impact of each independent variable in making the prediction of the dependent variable. In the multiple regression example discussed earlier, an appropriate question is to ask which variable—family size or family income—has the larger effect in predicting the number of credit cards used by a family. Independent variables with larger regression coefficients, all other things equal, would make a greater contribution to the predicted value. Insights into the relationship between independent and dependent variables are gained by examining the relative contributions of each independent variable. In our simple example, a marketer looking to sell additional credit cards and looking for families with higher numbers of cards would know whether to seek out families based on family size or family income.

**INTERPRETATION WITH REGRESSION COEFFICIENTS** Thus, for explanatory purposes, the regression coefficients become indicators of the relative impact and importance of the independent variables in their relationship with the dependent variable. Unfortunately, in many instances the regression coefficients do not give us this information directly, the key issue being “all other things equal.” As we will see, the scale of the independent variables also comes into play. To illustrate, we use a simple example.

Suppose we want to predict the amount a married couple spends at restaurants during a month. After gathering a number of variables, it was found that two variables, the husband's and wife's annual incomes, were the best predictors. The following regression equation was calculated using a least squares procedure:

$$Y = 30 + 4INC_1 + .004INC_2$$

where:

$INC_1$  = Husband's annual income (in \$1,000s)

$INC_2$  = Wife's annual income (in dollars)

If we just knew that  $INC_1$  and  $INC_2$  were annual incomes of the two spouses, then we would probably conclude that the income of the husband was much more important (actually 1,000 times more) than that of the wife. On closer examination, however, we can see that the two incomes are actually equal in importance, the difference being in the way each was measured. The husband's income is in thousands of dollars, such that a \$40,000 income is used in the equation as 40, whereas a wife's \$40,000 income is entered as 40,000. If we predict the restaurant dollars due just to the wife's income, it would be \$160 ( $40,000 \times .004$ ), which would be exactly the same for a husband's income of \$40,000 ( $40 \times 4$ ). Thus, each spouse's income is equally important, but this interpretation would probably not occur through just an examination of the regression coefficients.

In order to use the regression coefficients for explanatory purposes, we must first ensure that all of the independent variables are measured on comparable scales. Yet even then, differences in variability from one variable to another variable can affect the size of the regression coefficients. What is needed is a way to make all independent variables comparable in both scale and variability. We can achieve both these objectives and resolve this problem in explanation by using a modified regression coefficient called the beta coefficient.

**STANDARDIZING THE REGRESSION COEFFICIENTS: BETA COEFFICIENTS** The variation in response scale and variability across variables makes direct interpretation problematic. Yet, what if each of our independent variables had been standardized before we estimated the regression equation? **Standardization** converts variables to a common scale and variability, the most common being a mean of zero (0.0) and standard deviation of one (1.0). In this way, we make sure that all variables are comparable. If we still want the original regression coefficients for predictive purposes, is our only recourse to standardize all the variables and then perform a second regression analysis?

Luckily, multiple regression gives us not only the regression coefficients, but also coefficients resulting from the analysis of standardized data termed **beta ( $\beta$ ) coefficients**. Their advantage is that they eliminate the problem of dealing with different units of measurement (as illustrated previously) and thus reflect the relative impact on the dependent variable of a change in one standard deviation in either variable. Now that we have a common unit of measurement, we can determine which variable has the most impact. We will return to our credit card example to see the differences between the regression ( $b$ ) and beta ( $\beta$ ) coefficients.

In the credit card example, the regression ( $b$ ) and beta ( $\beta$ ) coefficients for the regression equation with three independent variables ( $V_1$ ,  $V_2$ , and  $V_3$ ) are shown in Figure 5.17.

Interpretation using the regression versus the beta coefficients yields substantially different results. The regression coefficients indicate that  $V_1$  is markedly more impact than either  $V_2$  or  $V_3$ , which are roughly comparable. The beta coefficients tell a different story.  $V_1$  is still the most impactful, but  $V_2$  is now almost as impactful, while  $V_3$  now is marginally important at best. These simple results portray the inaccuracies in interpretation that may occur when regression coefficients are used with variables of differing scale and variability.

Although the beta coefficients represent an objective measure of importance that can be directly compared, three cautions must be observed in their use:

- First, they should be used as a *guide to the relative importance of individual independent variables only when collinearity is minimal*. As we will see in the next section, collinearity can distort the contributions of any independent variable even if beta coefficients are used. Moreover, there are other measures of relative importance that better accommodate the impacts of multicollinearity.

Variable	Coefficients	
	Regression ( $b$ )	Beta ( $\beta$ )
$V_1$ Family Size	.635	.566
$V_2$ Family Income	.200	.416
$V_3$ Number of Autos	.272	.108

**Figure 5.17**  
Regression Coefficients versus Standardized Beta Coefficients

- Second, the beta values can be *interpreted only in the context of the other variables in the equation*. For example, a beta value for family size reflects its importance only in relation to family income, not in any absolute sense. If another independent variable were added to the equation, the beta coefficient for family size would probably change, because some relationship between family size and the new independent variable is likely.
- Third, the scale of beta weights after standardization is in terms of standard deviations of the original variable. The beta weight now represents the *change in the dependent measure for a one standard deviation change in the independent variable*. This transformation of scale may impact how a one unit change in standard deviation is perceived versus a one unit change in the original scale.

In summary, beta coefficients should be used only as a guide to the relative importance of the independent variables included in the equation and only for those variables with minimal multicollinearity. We will discuss a number of alternative measures of relative importance in a later section.

## ASSESSING MULTICOLLINEARITY

A key issue in interpreting the regression variate is the correlation among the independent variables. This problem is one of data, not of model specification. The ideal situation for a researcher would be to have a number of independent variables highly correlated with the dependent variable, but with little correlation among themselves. If you refer back to Chapter 3 and our discussion of exploratory factor analysis, the use of factor scores that are orthogonal (uncorrelated) was suggested to achieve such a situation.

Yet in most situations, particularly situations involving consumer response data, some degree of multicollinearity is unavoidable. On some other occasions, such as using dummy variables to represent nonmetric variables or polynomial terms for nonlinear effects, the researcher is creating situations of high multicollinearity. The researcher's task includes the following:

- Understand new measures of correlation which incorporate multicollinearity.
- Assess the degree of multicollinearity.
- Determine its impact on the results.
- Apply the necessary remedies if needed.

In the following sections we discuss in detail some new measures of correlation, useful diagnostic procedures, the effects of multicollinearity, and then possible remedies.

**Measure of Correlation Incorporating Multicollinearity** We have already discussed several times the impact of multicollinearity creating shared variance with another variable(s), but have not addressed how this can be quantified. The Pearson correlation we are accustomed to is a bivariate or **zero-order correlation**. It represents only the association between two variables, not accounting for the variation shared with any other variables and is the type of correlation that appears in the correlation matrix. But we also know that the correlation matrix shows correlations of every variable with all other variables, and since those values are not generally zero, there is always some level of multicollinearity. Moreover, the level of multicollinearity differs for each variable as each variable will have differing correlations with the other variables in the analysis.

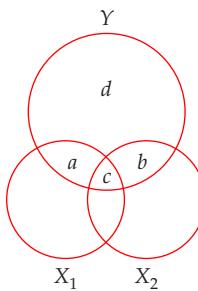
There are two other forms of correlation that reflect the degree of shared variance from multicollinearity (see Figure 5.18). First, the bivariate or zero correlation of  $X_2$  and  $Y$  includes both  $b$  (unique variance) and  $c$  (shared variance) as we have already discussed. This differs from the **semi-partial or part correlation**, which only contains the unique variance ( $b$ ). The **partial correlation** also only has the unique variance ( $b$ ), but differs in that the denominator is not the total variance of  $Y$ , but instead only the portion of  $Y$  not explained by any other variables ( $d + b$ ). These distinctions are important, as you will later see, that the squared semi-partial correlation quantifies the amount that  $R^2$  will decrease when the variable is removed, while the partial correlation is used in stepwise regression for selecting additional variables to add to the model. Thus, each of these correlations provides specific perspectives into the impact of multicollinearity on the bivariate correlation.

- Correlation of  $X_2$  and  $Y$ :  

$$(b + c) / (a + b + c + d)$$
- Semipartial (Part) Correlation of  $X_2$  and  $Y$ :  

$$b / (a + b + c + d)$$
- Partial Correlation of  $X_2$  and  $Y$ :  

$$b / (d + b)$$



**Figure 5.18**  
**Partial and Part Correlation: Reflecting Multicollinearity**

$a$  = variance of  $Y$  uniquely explained by  $X_1$   
 $b$  = variance of  $Y$  uniquely explained by  $X_2$   
 $c$  = variance of  $Y$  explained jointly by  $X_1$  and  $X_2$   
 $d$  = variance of  $Y$  not explained by  $X_1$  or  $X_2$

This example illustrates how the bivariate correlation can be modified to reflect multicollinearity. Note that in our example we only had one other variable, but both correlations can easily be extended to include any number of other variables contributing to the shared variance arising from multicollinearity.

**Identifying Multicollinearity** The simplest and most obvious means of identifying collinearity is an examination of the correlation matrix for the independent variables. We could look for other variables that are highly correlated with a specific independent variable, but that only reflects collinearity. The presence of high correlations (generally .70 and higher) is the first indication of substantial collinearity. Lack of any high correlation values, however, does not ensure a lack of collinearity. Collinearity may be due to the combined effect of two or more other independent variables (termed *multicollinearity*).

To assess multicollinearity, we need a measure expressing the degree to which each independent variable is explained by the set of other independent variables. *In simple terms, each independent variable becomes a dependent variable and is regressed against the remaining independent variables.* Two approaches are available to assess multicollinearity. First, overall measures (tolerance and its inverse, the variance inflation factor) of multicollinearity show the level of multicollinearity for each variable. Second, a decomposition of the multicollinearity among variables can identify specific sets of variables where the multicollinearity may originate.

**OVERALL MEASURES** The first diagnosis of multicollinearity should be with the overall measures of multicollinearity for each variable. While these measures do not provide any information as to the source of the multicollinearity (i.e., which other variables are collinear), they do provide an indication if multicollinearity does exist and if further examination is needed.

**Tolerance** The first overall measure of multicollinearity is **tolerance**, which is defined as the amount of variability of the selected independent variable *not explained by the other independent variables*. Thus, for any regression model with two or more independent variables the tolerance can be simply defined in two steps:

- 1 Take each independent variable, one at a time, and calculate  $R^{2*}$ —the amount of that independent variable that is explained by all of the other independent variables in the regression model. In this process, the selected independent variable is made a dependent variable predicted by all the other remaining independent variables.
- 2 Tolerance is then calculated as  $1 - R^{2*}$ . For example, if the other independent variables explain 25 percent of independent variable  $X_1$  ( $R^{2*} = .25$ ), then the tolerance value of  $X_1$  is .75 ( $1.0 - .25 = .75$ ).

The tolerance value should be high, which means a small degree of multicollinearity (i.e., the other independent variables do not collectively have any substantial amount of shared variance). Determining the appropriate levels of tolerance will be addressed in a following section.

**Variance Inflation Factor** A second measure of multicollinearity is the **variance inflation factor (VIF)**, which is calculated simply as the inverse of the tolerance value. In the preceding example with a tolerance of .75, the VIF would

be 1.33 ( $1.0 \div .75 = 1.33$ ). Thus, instances of higher degrees of multicollinearity are reflected in lower tolerance values and higher VIF values. The VIF gets its name from the fact that the square root of the ( $\sqrt{VIF}$ ) is the degree to which the standard error has been increased due to multicollinearity. Let us examine a couple of examples to illustrate the interrelationship of tolerance, VIF, and the impact on the standard error.

For example, if the VIF equals 1.0 (meaning that tolerance equals 1.0 and thus no multicollinearity), then the  $\sqrt{VIF} = 1$  and the standard error is unaffected. However, assume that the tolerance is .25 (meaning that there is fairly high multicollinearity, because 75 percent of the variable's variance is explained by other independent variables). In this case the VIF is 4.0 ( $1.0 \div .25 = 4$ ) and the standard error has been doubled ( $\sqrt{4} = 2$ ) due to multicollinearity.

The VIF translates the tolerance value, which directly expresses the degree of multicollinearity, into an impact on the estimation process. As the standard error is increased, it makes the confidence intervals around the estimated coefficients larger, thus making it harder to demonstrate that the coefficient is significantly different from zero.

**DECOMPOSITION OF MULTICOLLINEARITY** The decompositional method provides the researcher a means to identify the set(s) of variables which have high multicollinearity [13]. There are two components that depict the overall level of multicollinearity as well as its presence across the independent variables:

- *Condition index.* This represents the collinearity of combinations of variables in the dataset.
- *Regression coefficient variance-decomposition matrix.* This shows the proportion of variance for each regression coefficient (and its associated independent variable) attributable to each condition index.

We combine these two components in a two-step procedure:

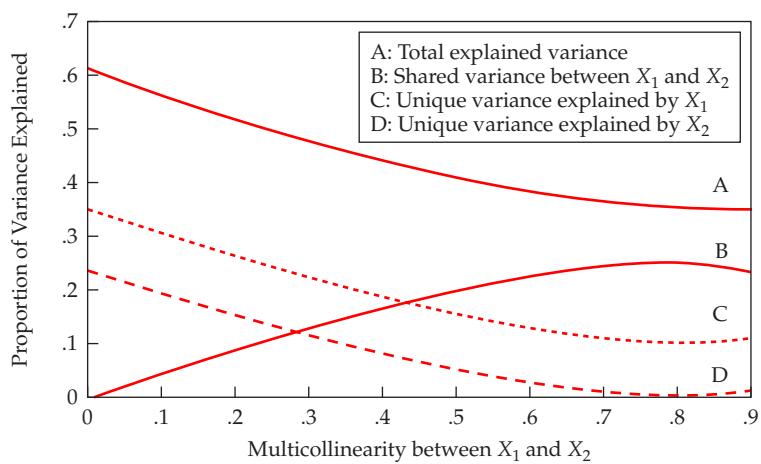
- 1 Identify all condition indices above a threshold value. The threshold value usually is in a range of 15 to 30, with 30 the most commonly used value.
- 2 For all condition indices exceeding the threshold, identify variables associated with that condition index that have variance proportions above 90 percent. A collinearity problem is indicated when a condition index identified in step 1 as above the threshold value accounts for a substantial proportion of variance (.90 or above) for *two or more* coefficients.

The application of the decompositional method is only needed if there are indications of high multicollinearity. But it does provide a means of identifying the sets of variables that should be examined more closely if there is a need to remedy multicollinearity.

**The Effects of Multicollinearity** The effects of multicollinearity can be categorized in terms of estimation or explanation. In either instance, however, the underlying reason is the same: Multicollinearity creates "shared" variance between variables, thus decreasing the ability to predict the dependent measure as well as ascertain the relative roles of each independent variable. Figure 5.19 portrays the proportions of shared and unique variance for two independent variables in varying instances of collinearity. If the collinearity of these variables is zero, then the individual variables predict 36 and 25 percent of the variance in the dependent variable, for an overall prediction ( $R^2$ ) of 61 percent. As multicollinearity increases, the total variance explained decreases (*estimation*). Moreover, the amount of unique variance for the independent variables is reduced to levels that make estimation of their individual effects quite problematic (*explanation*). The following sections address these impacts in more detail.

**IMPACTS ON ESTIMATION** Multicollinearity can have substantive effects not only on the predictive ability of regression model (as just described), but also on the estimation of the regression coefficients and their statistical significance tests.

**Singularity** First, the extreme case of multicollinearity in which two or more variables are perfectly correlated, termed **singularity**, prevents the estimation of any coefficients. Although singularities may occur naturally among the independent variables, many times they are a result of researcher error. A common mistake is to include all of the dummy variables used to represent a nonmetric variable, rather than omitting one as the reference category. Also,

**Figure 5.19**

Proportions of Unique and Shared Variance by Levels of Multicollinearity

Correlation between dependent and independent variables:  
 $X_1$  and dependent (.60),  $X_2$  and dependent (.50)

actions such as including a summated scale along with the individual variables that created it will result in singularities. For whatever reason, however, the singularity must be removed before the estimation of coefficients can proceed.

**INCREASES IN STANDARD ERRORS** As multicollinearity increases, the ability to demonstrate that the estimated regression coefficients are significantly different from zero can become markedly impacted due to increases in the standard error as shown in the VIF value. This issue becomes especially problematic at smaller sample sizes, where the standard errors are generally larger due to sampling error.

**Reversal of Signs of Coefficients** Apart from affecting the statistical tests of the coefficients or the overall model, high degrees of multicollinearity can also result in regression coefficients being incorrectly estimated and even having the wrong signs. Two examples illustrate this point.

Our first example (see Figure 5.20) illustrates the situation of reversing signs due to high negative correlation between two variables. In Example A it is clear in examining the correlation matrix and the simple regressions that the relationship between  $Y$  and  $V_1$  is positive, whereas the relationship between  $Y$  and  $V_2$  is negative. The multiple regression equation, however, does not maintain the relationships from the simple regressions. It would appear to the casual observer examining only the multiple regression coefficients that both relationships ( $Y$  and  $V_1$ ,  $Y$  and  $V_2$ ) are negative, when we know that such is not the case for  $Y$  and  $V_1$ . The sign of  $V_1$ 's regression coefficient is wrong in an intuitive sense, but the strong negative correlation between  $V_1$  and  $V_2$  results in the reversal of signs for  $V_1$ . Even though these effects on the estimation procedure occur primarily at relatively high levels of multicollinearity (above .80), the possibility of counterintuitive and misleading results necessitates a careful scrutiny of each regression variate for possible multicollinearity.

A similar situation can be seen in Example B of Figure 5.20. Here, both  $Z_1$  and  $Z_2$  are positively correlated with the dependent measure (.293 and .631, respectively), but have a higher intercorrelation (.642). In this regression model, even though both bivariate correlations of the independent variables are positive with the dependent variable and the two independent variables are positively intercorrelated, when the regression equation is estimated the coefficient for  $Z_1$  becomes negative (-.343) and the other coefficient is positive (.702). This typifies the case of high multicollinearity reversing the signs of the weaker independent variables (i.e., lower correlations with the dependent variable).

In some instances, this reversal of signs is expected and desirable. Termed a **suppression effect**, it denotes instances when the “true” relationship between the dependent and independent variable(s) has been hidden in the bivariate correlations (e.g., the expected relationships are non-significant or even reversed in sign). By adding additional independent variables and inducing multicollinearity, some unwanted shared variance is accounted for and the remaining unique variance allows for the estimated coefficients to be in the expected direction. More detailed descriptions of all of the potential instances of suppression effects are shown in [29].

**Figure 5.20**

Regression Estimates with Multicollinearity Data

EXAMPLE A				EXAMPLE B			
Data				Data			
ID	Y	V <sub>1</sub>	V <sub>2</sub>	ID	Y	Z <sub>1</sub>	Z <sub>2</sub>
1	5	6	13	1	3.7	3.2	2.9
2	3	8	13	2	3.7	3.3	4.2
3	9	8	11	3	4.2	3.7	4.9
4	9	10	11	4	4.3	3.3	5.1
5	13	10	9	5	5.1	4.1	5.5
6	11	12	9	6	5.2	3.8	6.0
7	17	12	7	7	5.2	2.8	4.9
8	15	14	7	8	5.6	2.6	4.3
				9	5.6	3.6	5.4
				10	6.0	4.1	5.5
Correlation Matrix				Correlation Matrix			
	Y	V <sub>1</sub>	V <sub>2</sub>		Y	Z <sub>1</sub>	Z <sub>2</sub>
Y	1.0			Y	1.0		
V <sub>1</sub>	.823	1.0		Z <sub>1</sub>	.293	1.0	
V <sub>2</sub>	-.977	-.913	1.0	Z <sub>2</sub>	.631	.642	1.0
Regression Estimates				Regression Estimates			
Simple Regression (V <sub>1</sub> ):				Simple Regression (Z <sub>1</sub> ):			
$Y = -4.75 + 1.5V_1$				$Y = 2.996 + .525Z_1$			
Simple Regression (V <sub>2</sub> ):				Simple Regression (Z <sub>2</sub> ):			
$Y = 29.75 - 1.95V_2$				$Y = 1.999 + .587Z_2$			
Multiple Regression (V <sub>1</sub> , V <sub>2</sub> ):				Multiple Regression (Z <sub>1</sub> , Z <sub>2</sub> ):			
$Y = 44.75 - .75V_1 - 2.7V_2$				$Y = 2.659 - .343Z_1 + .702Z_2$			

However, in other instances, the theoretically supported relationships are reversed because of multicollinearity, leaving the researcher to explain why the estimated coefficients are reversed from the expected sign. In these instances, the researcher may need to revert to using the bivariate correlations to describe the relationship rather than the estimated coefficients that are impacted by multicollinearity.

The reversal of signs may be encountered in all of the estimation procedures, but is seen more often in confirmatory estimation processes where a set of variables is entered into the regression model and the likelihood of weaker variables being affected by multicollinearity is increased.

**IMPACTS ON EXPLANATION** The effects on explanation primarily concern the ability of the regression procedure and the researcher to represent and understand the effects of each independent variable in the regression variate. As multicollinearity occurs (even at the relatively low levels of .30 or so) the process for identifying the unique effects of independent variables becomes increasingly difficult. This has an impact on several aspects of explanation.

**Interpretation of Coefficients** Remember that the regression coefficients represent the amount of unique variance explained by each independent variable. As multicollinearity results in larger portions of shared variance and lower levels of unique variance, the effects of the individual independent variables become less distinguishable. It is even possible to find those situations in which multicollinearity is so high that none of the independent regression coefficients are statistically significant, yet the overall regression model has a significant level of predictive accuracy. We will discuss measures of relative importance in a later section that portrays both the unique and shared variance of an independent variable.

**Interpretation of Shared Variance** As multicollinearity increases, the amount of shared variance may increase dramatically, yet there has been little attempt to understand the composition of this shared variance and how it might be interpreted. Recent research [103] has proposed a method to “decompose” the shared variance based on the work of Mood [82]. Somewhat similar to dominance analysis, a technique for assessing a variable’s relative importance, this decompositional approach can provide some perspectives on how to interpret the shared variance portion of  $R^2$ .

More recent research employs a structural equation modeling approach to decompose the shared variance among independent variables into criterion-relevant (related to the dependent variable) versus criterion-irrelevant (not related to the dependent variable) [10]. The criterion-irrelevant effects are similar to suppression effects described earlier and by eliminating these effects from the estimation process more accurate regression weights can be estimated.

**IGNORABLE EFFECTS** There are situations in which a high degree of multicollinearity may be expected and thus ignored. The most common is when interaction terms are included in the analysis (e.g., moderation effects). Since the interaction terms are a product of two other variables in the equation, multicollinearity is to be expected. Similarly, when polynomials are included to represent non-linear effects of variables there can be high multicollinearity. This is why many times significance tests for the interaction or polynomial terms are accomplished through incremental fit tests of significance. Also, there may be multicollinearity among the dummy variables used to represent nonmetric variables. Finally, if there are variables included as control variables, then there is no need for interpretation and they can exhibit high multicollinearity as well.

**How Much Multicollinearity is Too Much?** Because the tolerance value is the amount of a variable unexplained by the other independent variables, small tolerance values (and thus large VIF values because  $VIF = 1 \div \text{tolerance}$ ) denote high collinearity. A common cut-off threshold is a tolerance value of .10, which corresponds to a VIF value of 10. However, particularly when sample sizes are smaller, the researcher may wish to be more restrictive due to the increases in the standard errors due to multicollinearity. With a VIF threshold of 10, this tolerance would correspond to standard errors being “inflated” more than three times ( $\sqrt{10} = 3.16$ ) what they would be with no multicollinearity.

Each researcher must determine the degree of collinearity that is acceptable, because most defaults or recommended thresholds still allow for substantial collinearity. Some suggested guidelines for bivariate and multicollinearity follow.

**BIVARIATE CORRELATIONS** When assessing bivariate correlations, two issues should be considered. First, correlations of even .70 (which represents “shared” variance of 50%) can impact both the explanation and estimation of the regression results. Moreover, even lower correlations can have an impact if the correlation between the two independent variables is greater than either independent variable’s correlation with the dependent measure (e.g., the situation in our earlier example of the reversal of signs). These patterns should be examined for each variable with the highest two or three correlations, particularly if they involve correlations with different signs [128].

**TOLERANCE OR VIF** The suggested cut-off for the tolerance value is .10 (or a corresponding VIF of 10.0), which corresponds to a multiple correlation of .95 with the other independent variables. When values at this level are encountered, multicollinearity problems are quite likely. However, problems are likely at much lower levels as well [88]. For example, a VIF of 5.3 corresponds to a multiple correlation of .9 between one independent variable and all other independent variables. Even a VIF of 3.0 represents a multiple correlation of .82, which would be considered high if between dependent and independent variables.

Therefore, the researcher should always assess the degree and impact of multicollinearity even when the diagnostic measures are substantially below the suggested cut-off (e.g., VIF values of 3 to 5). While multicollinearity is many times the result of a large number of variables, understanding the patterns of correlations among the strongest three or four variables can provide substantial insight into the impact of multicollinearity [128].

We strongly suggest that the researcher always specify the allowable tolerance values in regression programs, because the default values for excluding collinear variables allow for an extremely high degree of collinearity. For example, the default tolerance value in IBM SPSS for excluding a variable is .0001, which means that until more than 99.99 percent of variance is predicted by the other independent variables, the variable could be included in the

regression equation. Estimates of the actual effects of high collinearity on the estimated coefficients are possible but beyond the scope of this text (see Neter et al. [85]).

**Remedies for Multicollinearity** The remedies for multicollinearity range from modification of the regression variate to the use of specialized estimation procedures. Once the degree of collinearity has been determined, the researcher has a number of options:

**DELETE COLLINEAR VARIABLES** Omit one or more highly correlated independent variables and identify other independent variables to help the prediction. The researcher should be careful when following this option, however, to avoid creating specification error when deleting one or more independent variables.

**APPLY DIMENSIONAL REDUCTION** As discussed earlier, dimensional reduction provides an alternative to using the original independent variables by substituting values for composites/factors in place of the original variables. The composites are estimated to “combine” highly collinear variables into a single measure so that their predictive effect can be retained, but without the multicollinearity from the entire set of individual variables.

**SPECIFIC ESTIMATION TECHNIQUES** Use a more sophisticated method of analysis such as Bayesian regression or regression on principal components to obtain a model that more clearly reflects the simple effects of the independent variables. These procedures are discussed in more detail in several texts [13, 85]. Variable selection techniques discussed earlier such as ridge regression and LASSO also help mitigate multicollinearity effects.

**DO NOTHING** Use the model with the highly correlated independent variables for prediction only (i.e., make no attempt to interpret the regression coefficients), while acknowledging the lowered level of overall predictive ability. Use the simple correlations between each independent variable and the dependent variable to understand the independent-dependent variable relationship.

Each of these options requires that the researcher make a judgment on the variables included in the regression variate, which should always be guided by the theoretical background of the study.

## RELATIVE IMPORTANCE OF INDEPENDENT VARIABLES

As we have seen in the past discussions, the presence of multicollinearity complicates the interpretation of the regression coefficients, especially when assessing the impact of independent variables on the dependent measure. Something as simple as the bivariate correlations overstates the impact of individual variables, but using only the estimated regression coefficients can have complications as well. To this end a series of new measures of **relative importance** have been developed to provide an assessment of the overall impact of the independent variables, accounting for both shared and unique variance explained and in measures that are comparable across all the independent variables [69]. These newer measures help clarify which independent variables are contributing most to a regression model by providing supplemental information to the traditional regression results.

The following discussion will first review the available direct measures of variable importance available from the regression results and correlations used in the estimation process. Then a series of relative importance measures will be discussed that extend the basic regression results by either performing a series of regression models or undertake some independent variable transformations.

**Direct Measures of Variable Importance** Researchers have long relied on the regression results and measures of correlation to make assessments of variable importance. We will quickly review those measures and discuss their advantages and limitations.

**BIVARIATE CORRELATIONS** The first and most basic measure of a variable’s impact is the bivariate correlation with the dependent variable. It represents the fundamental relationship of regression analysis and thus should always be considered. In some ways it represents our most fundamental information about the relationship used in regression analysis and thus provides a starting point for understanding a variable’s impact. It is limited, however, as we have seen in past discussions by the presence of multicollinearity. This is perhaps best reflected in the fact that the sum of individual

bivariate correlations (remember that squared correlations equal variance explained) is always greater than the regression model's  $R^2$ . Thus, while representing the fundamental relationships that should be represented in the regression model, the bivariate correlations do have limitations in terms of their actual impact when combined in the variate.

**SQUARED SEMI-PARTIAL CORRELATION** This represents the percentage of the variance in the dependent variables that is unique to the independent variables and not any of the other variate variables. It can also be viewed as the decrease in  $R^2$  that results from removing a predictor from the model. Thus, while representing the unique variance of the independent variable, it does not reflect any of the shared effect with other variables.

**REGRESSION WEIGHTS** The estimated regression coefficients reflect the “unique” relationship, but also the variable’s overall impact is diminished by its multicollinearity with other variables in the model. As a result, it does not reflect any of the shared impact with other variables.

**BETA (STANDARDIZED) WEIGHTS** The regression weights are expressed on a standardized scale and thus are now free of the effects of differing scales (e.g., 1–10 scale for one variable and 1 to a 100 scale for another variable) in impacting the estimated coefficients. They are not, however, adjusted for in any manner for multicollinearity and thus suffer the same limitations as the regression weights in expressing only unique variance and no measure of shared variance.

**Measures of Relative Importance** The measures of relative importance discussed below provide differing perspectives on (a) an independent variable’s unique and shared impacts on the dependent measures while (b) expressing these effects in a manner which makes them directly comparable (e.g., summing to  $R^2$  or 100 percent). While several different approaches are taken by the various measures, they all provide additional insights into the total impact of the different independent variables in the presence of multicollinearity [87, 69]. While we can only provide an overview of the various measures and their relative benefits, we encourage researchers to explore these measures as they develop and new measures emerge. Also, these measures will be illustrated in the HBAT example at the end of this chapter.

**ALL POSSIBLE SUBSETS REGRESSION** This method is not strictly a measure of relative importance, but is discussed briefly as it is the foundation for several of the measures that follow. The concept of all possible subsets regression, as discussed earlier in the variable selection section, is to estimate all of the possible regression equations that can be formed from differing combinations of the independent variables and then compare them on some criterion measure (e.g., adjusted  $R^2$ ). Thus, when rank-ordered by the criterion measure, the researcher can identify the variables in the models with the best predictive fit. As will be seen when examining these results, many models are generally very close in predictive fit and thus it is impractical for the researcher to make any assessments just on these results. One drawback is that analyses quickly become quite complex, as even a model with just 20 independent variables requires over a million regression models to be estimated.

**STRUCTURE COEFFICIENTS** The **structure coefficients** are the bivariate correlations of each independent variable with the predicted value, not the dependent variable as seen in the input correlation matrix [32]. As such, they are a measure of the relative contribution to the predicted value. They do not make any distinction between unique versus shared variance, just as the case with the bivariate correlations with the dependent variable. But their relationship to the predicted value makes them quite useful in comparisons to the regression weights. Variables with small regression weights yet large squared structure coefficients have a high shared predictive effect, but a small unique effect. In a counter situation, large regression weights but small structure coefficients indicate variables that may exhibit suppression effects [87].

**COMMONALITY ANALYSIS** Based upon all possible subsets regression, **commonality analysis** divides each independent variable’s impact into a unique component (the squared semi-partial correlation) and the shared component, with both of these totaling across all variables to the model  $R^2$  [25, 86, 94]. For each variable’s shared component, it identifies each unique combination of variables (e.g., for  $V_1$ ,  $V_2$ , and  $V_3$  the shared variance is divided into  $V_1-V_2$ ,  $V_1-V_3$ ,  $V_2-V_3$  and  $V_1-V_2-V_3$ ). There are also combined effects for each variable (unique and shared). One

aspect of commonality analysis that can be particularly useful is that negative effects (either unique or shared) are indicators of suppression effects and they can be summed across all effects to get a total suppression effect [95].

**DOMINANCE ANALYSIS** The **dominance analysis** method is also based on all possible subsets regression, but provides a different comparison among the independent variables [23, 24]. The basic measure of impact is the average squared semi-partial correlation across all regression models in which that variable is included. Also provided are two measures of dominance between each pair of variables: (a) complete dominance is when one variable always has the greater squared semi-partial correlation, no matter what other variables are in the model, (b) conditional dominance is when one variable exceeds the other in some model specifications, but they reverse in other model specifications. A third measure of general dominance simply focuses on the overall importance weights and not specific combinations. As such, it is the average contribution of an independent variable to the dependent variable, both on its own and when considering all the other independent variables. Also, general dominance weights sum to the overall model  $R^2$ .

**RELATIVE WEIGHTS** The final measure of relative importance is **relative weights**, which is a two-step process involving transformation of the independent variables [64]. A new set of uncorrelated predictors ( $Z_j$ ) that are maximally related to the original set of correlated predictors ( $X_i$ ) are created. The  $Z$  are first used as independent variables in predicting the dependent variable and then the estimated  $Z$  values are predicted by the independent variables. The relative weights are the product of the squared standardized regression weights from two equations [64]. The relative weights sum to  $R^2$ , but do not distinguish between unique and shared variance. This approach has generated substantial application, including an online tool [118]. There has also been considerable research in examining the impacts of measurement error and sampling error [65], statistical significance [119], an extension to logistic regression [116] and even multiple outcome measures [72]. While there have been some criticisms of relative weights [111], they are generally regarded as one of the more useful measures [66, 84].

**WHICH MEASURE TO USE?** No matter which measure of relative importance is used, researchers should consider these measures as a critical supplement to the more traditional measures of variable importance [117]. All of the measures we have discussed, even the direct measures, provide some insights into variable importance. Figure 5.21 provides a comparison of these measures on a range of variable importance characteristics. The reader should note that no measure

**Figure 5.21**  
Comparing Measures of Variable Importance

Measures of Variable Importance	Always Totals $R^2$	Values are		Identifies Multi-collinearity	Total Effect	Direct Effect	Partial Effect	Identifies Suppressor
		Identical When Predictors Uncorrelated	Identifies					
Bivariate correlation		X				X		
Beta weights		X			X			
Structure Coefficients						X		
Commonality Coefficients	X							X
Unique		X			X			
Common			X	X				
Dominance Analysis						X	X	X
Complete					X			
Conditional								X
General	X							
Relative Weights	X			X				

Source: [84]

## Interpreting the Regression Variate

Examine the bivariate correlations with the dependent variable as “starting point” for the nature of the relationships among dependent and independent variables.

Use the results of the regression model to interpret the unique impact of each independent variable relative to the other variables in the model, because model respecification can have a profound effect on the remaining variables:

Use beta weights as a measure of comparing relative importance among independent variables, but remember that they only reflect unique impact.

Regression coefficients describe changes in the dependent variable, but can be difficult in comparing across independent variables if the response formats vary.

Multicollinearity may be considered “good” when it reveals a suppressor effect, but generally it is viewed as harmful because increases in multicollinearity:

Reduce the overall  $R^2$  that can be achieved.

Confound estimation of the regression coefficients.

Negatively affect the statistical significance tests of coefficients.

Generally accepted levels of multicollinearity (tolerance values up to .10, corresponding to a VIF of 10) almost always indicate problems with multicollinearity, but these problems may also be seen at much lower levels of collinearity and multicollinearity:

Bivariate correlations of .70 or higher may result in problems, and even lower correlations may be problematic if they are higher than the correlations between the independent and dependent variables.

Values much lower than the suggested thresholds (VIF values of even 3 to 5) may result in interpretation or estimation problems, particularly when the relationships with the dependent measure are weaker.

Additional measures of variable importance (e.g., commonality analysis, dominance analysis and relative weights) all represent both the unique and shared impact of the independent variables. The shared impact, not reflected in the regression coefficients, is the result of multicollinearity among independent variables.

provides all of the possible characteristics and each has its specific insights. While to some extent the choice of one of these measures may rest on which foundational approach is considered more appropriate (all possible subsets regression or data transformations), measures from each approach (dominance analysis and relative weights) are considered to be the most successful in representing relative importance in the presence of multicollinearity. One challenge to these methods comes from interaction and polynomial terms, but integration of higher-order effects is also possible [73]. Fostering the more widespread use of these measures is the availability of software supplements for SAS and SPSS [69], Excel [69] and R [87].

## SUMMARY

Interpreting the regression weights of the variables in the regression model is a critical step in achieving any objective of regression analysis requiring explanation. As we have seen, the estimated regression weights are critical elements in the regression model, but multicollinearity among the independent variables results in shared variance in the prediction of the dependent variable that is not directly represented in the estimated regression weights. As the level of multicollinearity increases, the proportion of the overall explained variance that is shared also increases. As a result, the research needs to employ measures that capture not only the unique effects of each independent variable, but also some insight into its contribution to the shared explanatory effect of the dependent measure. The impact played by multicollinearity accentuates the importance of “Managing the variate” through the options available in both variable specification and variable selection.

## Stage 6: Validation of the Results

After identifying the best regression model, the final step is to ensure that it represents the general population (generalizability) and is appropriate for the situations in which it will be used (transferability). The best guideline is the extent to which the regression model matches an existing theoretical model or set of previously validated results on the same topic. In many instances, however, prior results or theory are not available. Thus, we also discuss empirical approaches to model validation.

### ADDITIONAL OR SPLIT SAMPLES

The most appropriate empirical validation approach is to test the regression model on a new sample drawn from the general population. A new sample will ensure representativeness and can be used in several ways. First, the original model can predict values in the new sample and predictive fit can be calculated. Second, a separate model can be estimated with the new sample and then compared with the original equation on characteristics such as the significant variables included; sign, size, and relative importance of variables; and predictive accuracy. In both instances, the researcher determines the validity of the original model by comparing it to regression models estimated with the new sample.

Many times the ability to collect new data is limited or precluded by such factors as cost, time pressures, or availability of respondents. Then, the researcher may divide the sample into two parts: an estimation subsample for creating the regression model and the holdout or validation subsample used to test the equation. Many procedures, both random and systematic, are available for splitting the data, each drawing two independent samples from the single dataset. All the popular statistical packages include specific options to allow for estimation and validation on separate subsamples. Chapter 1 provides a discussion of the basic approaches to validation, including cross-validation.

Whether a new sample is drawn or not, it is likely that differences will occur between the original model and other validation efforts. The researcher's role now shifts to being a mediator among the varying results, looking for the best model across all samples. The need for continued validation efforts and model refinements reminds us that no regression model, unless estimated from the entire population, is the final and absolute model.

### CALCULATING THE PRESS STATISTIC

An alternative approach to obtaining additional samples for validation purposes is to employ the original sample in a specialized manner by calculating the **PRESS statistic (Prediction Sum of Squares)**, a measure similar to  $R^2$  used to assess the predictive accuracy of the estimated regression model. It differs from the prior approaches in that not one, but  $n - 1$  regression models are estimated in a procedure similar to the jackknife validation approach. The procedure omits one observation in the estimation of the regression model and then predicts the omitted observation with the estimated model. Thus, the observation cannot affect the coefficients of the model used to calculate its predicted value. The procedure is applied again, omitting another observation, estimating a new model, and making the prediction. The residuals for the "hold-out" observations can then be summed to provide an overall measure of predictive fit as well as compared to the original model to assess the increase in sum of squares of the residuals due to the validation procedure.

A relative measure of prediction is  $P^2$ , the coefficient of prediction, which substitutes PRESS for SSE:

$$P^2 = 1 - \frac{\text{PRESS}}{\left[ \frac{n}{n-1} \right]^2 \text{SST}}$$

This measure is an "out-of-sample" measure of expected predicted accuracy similar to  $R^2$  and acts similar to adjusted  $R^2$  in that a decrease in  $P^2$  indicates the inclusion of an irrelevant variable or deleting a relevant variable.

## COMPARING REGRESSION MODELS

When comparing regression models, the most common standard used is overall predictive fit.  $R^2$  provides us with this information, but it has one drawback: As more variables are added,  $R^2$  will always increase. Thus, by including all independent variables, we will never find another model with a higher  $R^2$ , but we may find that a smaller number of independent variables result in an almost identical value. Therefore, to compare between models with different numbers of independent variables, we use the adjusted  $R^2$ . The adjusted  $R^2$  is also useful in comparing models between different datasets, because it will compensate for the different sample sizes.

The most popular alternatives to  $R^2$  are the Akaike Information Criterion [3] and the Bayesian Information Criterion (or Schwartz Information Criterion) [104]. These information measure statistics are designed to identify the model with the best predictive power. Usually, the model that gives the smallest value of the AIC (or BIC) statistic is the preferred one. The BIC is claimed to be an improvement over the AIC since the AIC is inclined to overfitting the data, but both measures work in a consistent manner [121]. These measures can compare any set of models as long as the dependent measure and sample remain consistent.

## FORECASTING WITH THE MODEL

Forecasts can always be made by applying the estimated model to a new set of independent variable values and calculating the dependent variable values. However, in doing so, we must consider several factors that can have a serious impact on the quality of the new predictions:

- When applying the model to a new sample, we must remember that the predictions now have not only the sampling variations from the original sample, but also those of the newly drawn sample. Thus, we should always calculate the confidence intervals of our predictions in addition to the point estimate to see the expected range of dependent variable values.
- We must make sure that the conditions and relationships measured at the time the original sample was taken have not changed materially. For instance, in our credit card example, if most companies started charging higher fees for their cards, actual credit card holdings might change substantially, yet this information would not be included in the model.
- Finally, do not use the model to estimate beyond the range of independent variables found in the sample. For instance, in our credit card example, if the largest family had 6 members, it might be unwise to predict credit card holdings for families with 10 members. One cannot assume that the relationships are the same for values of the independent variables substantially greater or less than those in the original estimation sample.

## Extending Multiple Regression

As might be expected, the widespread usage of multiple regression across all areas of analytics, both academic and organizational, has spawned a myriad set of variants for addressing a wide range of issues, many of which have already been discussed. Among the most widely used are the following: robust regression to deal with outliers without their deletion, quantile regression where the dependent variable is divided into subgroups and models estimated for each, constrained regression where constraints are placed on the range or even direction of the parameter estimates, regression with censored or truncated outcome data (similar to survival analysis), regression with measurement error corrections and some other multiple equation regression models – seemingly unrelated regression and multivariate regression. It is beyond the scope of this chapter to delve into all of these methods, but the researcher should be aware that there are a wide variety of regression-based models for dealing with a range of different research questions.

In the following discussion we will introduce two additional regression-based models that have become increasingly popular – multilevel/hierarchical models and panel models. Each is designed to analyze unique types of data structures – nested or hierarchical data for multilevel models and cross-sectional longitudinal data for panel models.

## MULTILEVEL MODELS

The rise of multilevel models across a wide range of disciplines has acknowledged that multilevel or hierarchical effects (1) foster theoretical frameworks more appropriate for many of the actual settings in which they are studied, as well as (2) provide a unified framework for addressing many of the statistical issues which occur naturally when hierarchical data structures are present. Moreover, the software and resources for multilevel modeling have developed to the point that any researcher should incorporate multilevel modeling when appropriate. The following discussion focuses on first providing an introduction to multilevel modeling, the importance of contextual effects and the resulting hierarchical data structures, levels of effects in hierarchical models and its widespread use across disciplines and basic resources available. The next section details some of the basic concepts in multilevel modeling, matching measurement properties to level, the intraclass correlation, random versus fixed effects, sample size considerations, while the final section describes a five-stage modeling strategy to develop multilevel models.

**Introduction to MLM** Context matters! All individuals are embedded within contexts that have an influence on their behavior. **Contexts** are any external factor outside the unit of analysis that not only impact the outcome of multiple individuals, but also create differences between individuals in separate contexts and foster dependencies between the individuals in a single context. Failure to acknowledge the impact of contexts and group effects is termed the **psychologic fallacy** [36]. Contexts create data structures that are termed nested, hierarchical or clustered, and have long been acknowledged as a fundamental effect applicable to almost all research settings [19, 12]. Perhaps the most widely used example is student achievement as the outcome measure of interest. We know that student achievement is based to some degree on a number of student attributes, which we define as **Level-1**, the most fundamental unit of analysis. But we also assume that the context for that outcome (e.g., the classroom setting with the teacher, resources, other students, etc. which is termed **Level-2**) has an impact. Each level above Level-1 is an aggregation into groups of observations in the lower level (i.e., students into classrooms). The primary objective of any analysis of **hierarchical data structure** is to accurately assess the effects of each attribute of Level-1 (the student) and characteristic of Level-2 (the classroom) while also determining the extent to which Level-1 and Level-2 collectively impact the outcome. A **multilevel model (MLM)** is an extension of regression analysis that allows for the incorporation of both individual (Level-1) and contextual (Level-2) effects with the appropriate statistical treatment. As we will discuss in more detail later, the effects of context can create dependencies among the individuals within each context (e.g., students within a class) and thus violate the basic statistical assumption of independence. We should note that for simplicity we will discuss hierarchical data at two levels, but the technique can accommodate hierarchical structures with many levels (e.g., students in classrooms in schools in school districts in states).

**A SIMPLE EXAMPLE** Let's take a simplified look at how multilevel models work in a simple two-level model, with individual observations (e.g., single wage earners) at Level-1 grouped together by a Level-2 variable (e.g., regions). In our example, the basic model is that single wage earner income is a function of educational level. A regional predictor is the percentage of the regional budget devoted to education, which is proposed to assist in individuals achieving their educational level.

**Separate Equations by Region** The simplest way to assess the impact of the Level-2 variable on the Level-1 equations is to estimate a separate equation for each Level-2 variable (e.g., region). Intercepts and slopes for educational level can now vary by region. But the researcher faces two additional questions. First, what about the dependencies among individuals within regions? We have to assume that they may be correlated and thus violate the statistical assumption of independence, even if randomly sampled within the region. The dependency comes from knowing that the single wage earners, when grouped by region, are correlated in their outcome measures. Also, even when we do find that there are differences between regions on the regression coefficients, we do not know what about the regions impacts these changes. Since the context/group level characteristics (e.g., percentage of regional budget devoted to education) are constant within each group (e.g., region), analyzing each group separately doesn't allow for these group characteristics to be included since there isn't any variance on this variable within the group.

Before the development of multilevel models researchers were forced into two compromises. First, violations of the assumptions of independence were generally ignored. Second, variables characterizing the regions were entered directly into a single regression equation pooled across regions, even though they were measured at a different level and thus a region's single value was used for all individuals within that region. Both of these actions violated tenets fundamental to multiple regression, but were the only course of action at the time.

**The Multilevel Model Approach** Multilevel models approach the problem in a somewhat more complex, but fundamentally simple fashion—separate equations for each level that are then combined. The Level-1 equations are like our basic regression models before, varying by region:

$$\text{Level-1 equation (by region): } \text{Single wage earner income} = \text{intercept} + \text{Coefficient}_{\text{Education}} * \text{Educational level}$$

But multilevel models add a second type of equation—the Level-2 equation. The purpose of the Level-2 equation, which can have a set of characteristics of the level (e.g., region), is to now predict the intercepts and regression coefficients within each region. Thus, the dependent variables of the Level-2 equations are the slopes and coefficients of the Level-1 model, not the dependent variable of student achievement. For example, the Level-2 equation for the Level-1 coefficients would be:

$$\text{Level-2 equation (by region) Coefficient}_{\text{Education}} = \text{intercept} + \text{Coefficient}_{\text{PctBudgetEducation}} * \text{Pct budget in education}$$

When these two equations are combined, we get a new equation which has an “average” intercept and coefficient of the Level-1 variable that does not vary across regions (i.e., fixed effects), an estimate of the impact of the region characteristics on the Level-1 parameters (i.e., another fixed effect) and estimates of the variation of the Level-1 intercepts and regression coefficients across groups/regions (i.e., random effects). Additional terms allow for the dependencies among observations within groups.

While this is a very simplified view of how the models are estimated, it does highlight one feature of the multilevel model—the effects of each level are directed at the results of the level below—which is theoretically consistent with how we expect nested or hierarchical effects to occur. The higher-level effects “work through” the lower level effects, somewhat akin to mediation effects discussed earlier.

**BENEFITS OF MLM** The benefits of MLM are both methodological and conceptual [58]. From a *methodological perspective*, the most important improvement is the incorporation of the dependencies among observations introduced by the nested data structure. These dependencies create two statistical problems when not corrected by using MLM. First, the standard errors of coefficients are biased downward (i.e., smaller), thus making it easier to find statistical significance than it should be [107]. This is especially problematic for higher-level variables that are many times the key variables of interest [77]. Second, the effective sample size that should be used is smaller than the overall sample because of the nested data structure [112]. This impacts the estimates of statistical power and overstates the level of statistical power that is actually achieved. In addition to addressing these issues, MLM can directly incorporate repeated measure data as well as be applied directly to more complicated stratified sampling plans that have a nesting effect.

From a *conceptual perspective*, MLM allows for inclusion of characteristics from multiple levels of a conceptual model while retaining the hierarchical nature of the effects. Many conceptual models are theorized in a hierarchical structure (e.g., educational outcomes, organizational behavior, social and political groups, etc.). MLM provides a method that “preserves” the level’s effects to relationships within that level (i.e., independent variables in the level’s equation are only variables measured at that level). This also allows for an apportionment of impact to the various levels that was not available before. Finally, as will be discussed in the section on fixed versus random effects, the use of random effects allows for a generalization of the impacts to the population that is not possible with fixed effects.

**BRIEF HISTORY OF MLM** These effects have long been recognized and it was not until early pioneering work in several areas that a more unified approach was developed [16, 22, 46]. From these and other early efforts came research and application of multilevel models come from a wide range of disciplines, among them education [16, 48, 34, 92, 20], health

science [38, 96, 90], public policy [102], sociology [37, 81, 89], organizational behavior [74, 60, 21, 35, 50] and many others. As a result, what is today generally referred to as multilevel models is also known by a number of other names as developed in various disciplines, including hierarchical linear models, mixed models, random-coefficients, random parameter or random-effects models and nested or clustered data models.

**RESOURCES FOR MLM** One of the most positive benefits of the multi-disciplinary development of MLM is the diversity of software available. The first software widely used for MLM was HLM that is still widely used today [93]. A more recent development has been MLwiN [26] and multilevel modules are available in many statistical programs (Stata, MPLUS) or through basic statistical methods (e.g., MIXED models in IBM SPSS and SAS). Besides software availability, there are multiple excellent texts with a more detailed explanation of MLM from any number of perspectives [62, 45, 47, 108, 6] as well as texts oriented towards specific software – IBM SPSS [58], SAS [122], MPLUS [41] Stata [91] and R [42].

**Basic Concepts and Issues** Three fundamental concepts distinguish MLM from more basic regression models: matching measurement properties to level, the intra-class correlation coefficient, fixed versus random effects and sample size at each level. Each of these concepts plays a critical role in understanding and operationalizing MLM.

**MATCHING MEASUREMENT PROPERTIES TO LEVEL** As noted earlier, distinguishing variables by their appropriate level has conceptual and operational benefits. From a conceptual perspective, matching a variable to its appropriate level ensures that it will be included in the appropriate set of effects. But just as important is to match the variable so that it represents a “true” characteristic of that level and avoids misrepresentation of the effect [62, 71]. While it may seem obvious that the measure and level should correspond, there are common instances when they do not match. The most widespread is when the characteristic of a higher-level is derived from characteristics of a lower level, principally through aggregation. In our example this could occur by adding a new variable, percentage male, as a regional characteristic, but then interpreting it as a gender effect at the individual level. This approach results in an **ecological fallacy** since it has been consistently shown that aggregated values have quite different relationships than individual values [97]. Thus, it is incorrect to infer individual behavior based solely on group-level relationships. The reverse has been termed the **atomistic fallacy**, where group level relationships are assumed to equate to individual-level relationships [70]. While MLM can separate effects by level, it is the responsibility of the researcher to ensure that measures are included at their appropriate level.

**INTRACLASS CORRELATION (ICC)** The **intraclass correlation (ICC)** is a measure of the degree of dependence among individuals within a higher-level grouping. Examples of the ICC at Level-2 might be students in the same class, employees under the same supervisor or repeated responses from the same individual. At higher levels the ICC might be classes within a school or supervisors in a firm. In all these instances, it represents the expected correlation in the dependent variable between two randomly drawn units in a level. So in our example, it would be the expected correlation in single wage earner income between two randomly drawn individuals in a region. As this correlation increases, it demonstrates that the individuals in that region are not independent, thus violating this assumption of regression. It can also be thought of as the degree of covariance among error terms within the group [62]. As it increases it represents that the groups are more homogeneous and thus different from one another. In assessing the size of the ICC, researchers should note that ICC values as low as .10 have been found to be impactful when group sizes are large [113]. A simple measure of the ICC is a one-factor ANOVA, where the independent variable is the Level-2 variable. The ICC is then defined as the proportion of variance that exists between groups compared to the total variation.

The usefulness of the ICC comes from its use as both a diagnostic tool and a statistical adjustment for dependency. As a diagnostic tool, if the ICC for a level is non-significant, there is no reason to perform a MLM with that level since there are no significant effects. Moreover, the ICC can be used to estimate the extent to which the standard errors need to be increased to accurately reflect the Type I error [14]. The ICC also provides a means of calculating the effective sample size in clustered sampling designs such that the appropriate standard errors are utilized [68].

Increased ICC values can be offset in some regard by smaller group sizes, but this diminishes the ability to make group-level estimates. We will discuss this issue in more detail in the section below on sample size.

**FIXED VERSUS RANDOM EFFECTS** Perhaps the concept most unfamiliar to most researchers using regression is fixed versus random effects [11, 28]. A common distinction between fixed and random effects is that fixed effects are used for population parameters (i.e., single fixed value across the population) and random effects are used when the variable represents only a sample of possible effects. As an example, gender would be assumed to be a fixed effect since there are a set number of categories that never change. But what about classes chosen randomly in a school? Here we know that there are other classes that are not included. In this situation, a random effect for class would provide information about the variation or range of values across this sample of classes that could be extrapolated to the population of classes.

Technically, a **fixed effect** is the “default” in regression, such that a parameter estimate (either an intercept or coefficient) is made as a point estimate with no variation. We make an estimate of its variation across repeated samples with the standard error, but the parameter itself is considered as a fixed or single effect. So in almost all regression analyses we assume fixed parameters. But in a MLM setting, fixed effects for Level-2 and higher can become problematic. For example, assume that we just want to estimate a single fixed effect for each group at Level-2 that would require a dummy variable (the fixed effect) for each Level-2 group. For situations with a large number of groups, this introduces a large number of parameter estimates for each dummy variable (i.e., number of groups minus 1) into the model. Moreover, since groups at higher levels may have relatively few observations per group, then that estimate can have high variance, large standard error and be impacted by just a single observation. As a result, while fixed effects are unbiased, they do have the potential to have high variability as well as using a large number of parameters. A **random effect** attempts to reduce that variability of effect within a group by a form of pooling across groups and focusing on the best estimate of the variability or distribution of effects across the set of groups. In making the estimate of the variability of effects across all the groups, it has a tendency to minimize outlier groups to achieve a more stable estimate of variation, but at the expense of some degree of bias.

The use of random effects for Level-2 and higher is based on several considerations. First, if we used fixed effects, we would essentially be estimating a separate equation for each group to avoid the effect of dependency among observations. Yet for a Level-2 model, the dependent variable is the Level-1 parameter (either intercept or coefficient). Using a fixed effect here would be impossible since the dependent variable value would be the same for all observations within the group. This would also prevent any group characteristics to act as independent variables since they would all have the same value as well. So we need some method of “pooling” across the groups and still address the issue of dependency among observations. This is where the random effect comes into play. The relationships of the Level-2 equation are now estimated as random effects, such that we can “pool” across groups and estimate relationships between group characteristics we couldn’t do earlier. The trade-off becomes one of bias versus precision. The fixed effects equations generate many more parameters, each unbiased but less precise (few observations per parameter) and limited in types of effects that can be estimated. The random effects introduce some bias, but with much more precision as to the variance of the effects and the ability to introduce a set of Level-2 characteristics into the analysis. This is why MLM employ random effects at Level-2 and higher, thus the name “random coefficients” models.

A second consideration between fixed and random effects is the impact of endogeneity – what we termed earlier as bias due to omitted variables. Fixed effect models of grouped data can eliminate many forms of endogeneity since they in essence estimate model effects separately for each group (i.e., all the potential endogeneity effects are constant for observations within the group, thus having no effect). But since random effects “pool” across groups, endogeneity can have an impact and must be eliminated in some other fashion. This consideration comes less into consideration in MLM, but will be more impactful in panel models to be discussed in a later section.

Finally, is the choice between fixed and random effects just a conceptual issue? While the researcher should consider the implications of fixed versus random effects based on the issues discussed above, there is an empirical test to also provide guidance. The Hausman test [53] compares the results of the fixed effects model to that of the random effects model. A significance level of  $p < 0.05$  supports the hypothesis that there is a difference between

the two models and thus the fixed effects model should be used. If there is not a significant difference, then the random effects model can be used since it is preferred over the fixed effects model. As noted earlier, random effects are generally the default in MLM, but the choice is more critical in panel models.

**SAMPLE SIZE BY LEVEL** The sample size is dependent upon level and is based on the number of units of analysis. At Level-1 all of the sample is used, so the sample size is the number of observations. But as we move to higher levels, the unit of analysis changes to the number of groups per level (e.g., number of classroom, not the number of students). Moreover, the number of observations per group is less important than the number of groups. Group sizes of even five observations or more are acceptable as long as the number of groups is sufficient. So how many groups and observations per group are acceptable? Heck and Thomas [57] propose the 20/30 rule—at least 20 groups and 30 observations per group. Hox [61] varies this slightly with a 30/30 rule—at least 30 groups and 30 observations per group. Yet many research situations will vary in terms of individuals or groups, so which is most important to consider. In a practical sense, it is the number of groups at any level, since the groups are now the units of observation for the analysis. It is much better to have 30 groups of 10 observations than 10 groups of 60 observations since there are more groups for estimating the parameters. So increasing the number of groups at any level is always useful for estimating effects at that level and increases statistical power.

So how small can a group size be? Group sizes of five are encountered in many research situations and provide reasonable estimates. Moreover, they do not impact the power of the tests since that comes from the number of groups. But smaller group sizes will limit the power of the random coefficient tests—the differences between groups on the intercepts or independent variables [109]. So should group sizes be as large as possible? A trade-off should be considered as larger group sizes increase the impact of the ICC, especially as the ICC increases [7]. This is a logical outcome since larger group sizes reinforce the ICC across a larger number of observations in the group.

So what is the best sample size at each level? Obviously, all things equal, more observations are better. But the researcher needs to consider how observations are distributed in the nested structure. The limiting consideration is the smallest number of groups at any level. Following the 30/30 rule may be fairly easy for Level-2/Level-1, but it becomes much more difficult as we move to Level-3/Level-2 data structure. That does not mean that a smaller number of groups than 20 or 30 necessarily means the level effects cannot be estimated, but the research should always remember that the number of groups determines the sample size for that level's effects. If you only had eight groups at a level, would you feel comfortable estimating a regression equation with only eight observations? The types of effects being estimated and the requisite precision are always considerations on which levels to include in an analysis.

**SUMMARY** The issues of appropriate measurement by level, the intraclass correlation, fixed versus random effects and sample size by level are among the most impactful for a MLM design. Our discussion has touched on the most basic considerations in each of these issues and considerable research exists which provides more insight into each issue. Moreover, issues such as missing data and imputation in nested data structures are also critical in many research situations. We encourage the researcher to investigate these issues in more detail before undertaking a MLM analysis.

**Five-Stage Modeling Strategy** While the decisions involved in MLM may seem somewhat arbitrary, consensus on a five-stage modeling strategy has evolved which tests progressively more complex models [58, 62]. As the researcher finds sufficient model improvement at each stage the models can incorporate more effects at different levels. In this discussion we will focus on a two-level model, but this process can be extended to any pair of levels. In our example Level-1 will be individuals and Level-2 groups of individuals (the first contextual factor). We will use the term coefficient to refer to the regression coefficient estimated for an independent variable at a specific level. Also, the criterion for moving to the next level is incremental model fit since these are nested models.

**STEP 1: SUFFICIENT VARIATION AT LEVEL 2** The first step is to make sure that there is enough variation between groups to justify including the level in the analysis. If there are not differences between groups, then there will not be enough difference to estimate group-level effects and the ICC will be so low as to have no effect. The most direct test is a

simple one-way ANOVA with the single independent variable being Level-2 groups. If this does not reveal group differences, then no reason to include the Level-2 effects. In terms of the equations at each level, the Level-1 equation is just the single overall intercept and the Level-2 equation is the group-specific intercepts (means).

**STEP 2: LEVEL-1 MODEL WITH LEVEL-2 EFFECTS** In this model, the basic regression equation for Level-1 is specified—the intercept and independent variables for Level-1 are specified as fixed effects and the Level-2 effects (intercepts) are added. Here we test the basic ability of Level-2 effects to relate to Level-1 effects that we demonstrated in general in Step 1. The Level-2 model is just an intercept to control for group differences.

**STEP 3: INTRODUCE LEVEL-2 INDEPENDENT VARIABLES** At this step the Level-2 characteristics are introduced into the Level-2 equations to establish their relationships with the Level-1 parameters. Remember that the Level-2 equations have as their dependent variable the Level-1 intercept or coefficient for that group. So we now have three fixed effects—Level-1 intercept and coefficient and Level-2 coefficient along with a random effect for group differences.

**STEP 4: TEST FOR RANDOM COEFFICIENTS** The next model is one that is many times most associated with MLM and that is a test for random coefficients of the Level-1 variables. We have the fixed effects of the Level-1 intercept and coefficients, plus any Level-2 coefficients for explanatory variables. The model also provides the random effects for both intercepts and coefficients—Do the intercepts vary across groups? Do the Level-1 coefficients vary across groups? Most often each Level-1 variable is tested separately and then all significant random effects entered into the final model.

**STEP 5: ADD CROSS-LEVEL INTERACTIONS TO EXPLAIN VARIATIONS IN COEFFICIENTS** This final step is to identify which Level-2 characteristics are related to the Level-1 variables that had random coefficients (i.e., varied across groups). Interaction terms are used to test this cross-level effect in a manner similar to moderation discussed earlier. Significant interaction terms demonstrate the relationships between Level-2 variables and the differences between coefficients across groups.

**REVIEW** The five-stage process builds incrementally more complex models to ascertain at each step whether an additional model specification is justified—are the groups sufficiently different, how does the group and its characteristics impact the basic regression model, do the intercepts vary across groups, do the coefficients vary across groups and finally do the group characteristics relate to any variation in coefficients across groups. While the researcher may focus on results from one of the more complex models, we strongly recommend that all steps be performed so the researcher more fully understands the nature of the effects at each step.

**Summary** Multilevel models have gained widespread acceptance across a wide range of disciplines because they allow researchers to address research questions in a more realistic framework—one where the relationship of interest takes is impacted by contextual factors. No researcher wants to fall prey to the psychologicistic fallacy and assume that no force outside the unit of analysis have any impact. In some sense we are looking at higher-level effects as moderators—the individual group behavior is conditional on its higher-level group. In this manner researchers can (a) accommodate these types of effects in a framework which separates the Level-1 effects of interest from those of a higher-level (b) in a statistically appropriate manner and (c) correcting for any dependencies among observations due to these groupings.

The evolution of this technique has eliminated software availability and resources as reasons to not use the technique. Thus we strongly recommend that researchers become familiar with MLM as a “standard” technique when dealing with these types of effects and data structures.

## PANEL MODELS

In an analysis framework somewhat similar to multilevel models, **panel models** or **panel analysis** is a regression-based analytical technique designed to handle cross-sectional analyses of longitudinal or time-series data. In the cross-sectional perspective, it accommodates the typical regression analysis with a series of independent variables being related to an outcome variable. The analysis could be analyzing individuals, firms, brands, schools, countries

or other analysis units, and what factors influence their outcomes. But the unique element of panel models is that they also accommodate longitudinal data for both the dependent and independent variables. So instead of having to conduct separate analyses for each year or pool these data across years in some manner, panel models were designed to address the necessary characteristics of this type of data.

**Similarity to Multilevel Models** For readers who are acquainted with multilevel models or reviewed our earlier discussion, some of these concepts seem familiar. Using the terminology of multilevel models, each analysis unit (e.g., individual, firm, school, etc.) forms a group (i.e., Level-2), with the longitudinal data for that analysis unit the observations within the group (i.e., Level-1). In multilevel models we were concerned about dependencies within the group. With longitudinal data we know that we have some patterns of serial correlation that must be accounted for. Panel models will utilize many of the same concepts, particularly fixed versus random effects, to analyze this specific form of grouped data.

The basic panel model uses a data structure that can be visualized as rows representing an analysis unit (e.g., individual, brand, school, etc.) with both dependent measure and independent variables. This is comparable to a typical regression analysis. But there is one additional variable – a time period indicator. This allows multiple years to be contained in the same dataset so that the complete dataset has all the observations for each analysis unit in each of the time periods. As will be discussed in a later section, the researcher can now select whether to employ a fixed or random effect for each estimated parameter as well as identify if there are any time-related effects.

**Benefits of Panel Models** The ability to integrate cross-sectional and longitudinal analyses into a single framework has several benefits over other methods. First, by using a fixed effects estimate for an effect, the omitted variable problem (endogeneity) is accounted for. In doing so the model can also account for heterogeneity among the analysis units. The combined use of fixed and random effects allows for precise control over variation due to heterogeneity across analysis units and dependencies within groups. The ability to employ a full range of independent variable also allows for the testing of more complex models than alternative methods

**Background** Panel models provide an alternative to any number of other analytical techniques devoted to longitudinal data. A first approach would be repeated cross-sectional models, which suffer from the inability to “link” the analyses across years that we discussed in the multilevel model section. A second approach is the event history model, including such techniques as survival analysis, failure time analysis, and hazards or risks models. This model type is quite applicable to specific processes (e.g., survival) and emerging methods are facilitating the inclusion of more covariates (i.e., independent variables). But these models are still disadvantaged by the limited variates they can accommodate and the resulting focus on prediction rather than explanation, along with some complex data management issues. The third alternative method is time series analysis, which is best characterized as few variables, many time periods. It has unique advantages in the extensive analytical procedures for estimation and predictions and the detailed insights into patterns across time. It is generally limited in terms of the complexity of the relationships being evaluated and requires a fairly large number of time periods (generally at least 30).

Panel models provide an alternative analytical framework that has been a central analytical method in areas such as economics [5, 49, 27], epidemiology [51], sociology [17, 52] political science [9, 126] and spatial analysis [40]. The ability to include a large number of cross-sectional units and time periods while also integrating the analytical capabilities of regression for explanatory purposes makes panel models a perfect match for many types of research questions, particularly those utilizing secondary data.

**Basic Issues** As with our discussion of multilevel models we will only be able to discuss briefly some of the fundamental issues involved in panel models and expose the reader to the types of issues that must be addressed in this method. In the sections below we will cover the types of variables than can be incorporated into panel models, the types of models that can be estimated by combining fixed and random effects, some of the basic questions when choosing fixed versus random effects and the ability to add a time dimension to the results.

**TYPE OF VARIABLES** There are four basic types of variables that can be used in a panel model. First are variables that differ between units of analysis, but don't change over time (e.g., race and gender, type of firm or level of school). The second variable type can change over time, but are the same for all units of analysis at any given time period (e.g., national economic indicators). A third type is a variable that can vary over both time and units (e.g., income, firm expenditures, school student composition). The final variable is one that varies over time in a predictable pattern (e.g., any measure of age or tenure). The inclusion of some variables, such as those that vary over units but not time, are dependent on whether fixed or random effects are used. The others, such as those with a time element will depend on how the time aspect is modeled. But panel analysis has options by which all of these variable types can be included in some fashion.

**TYPES OF MODELS** As we saw with multilevel models, panel models can be used for a range of models, from the simple to those with quite complex relationships. The most basic model type is a simple pooled regression, which disregards the interdependencies among observations within a unit of analysis and estimates a single intercept and slope coefficient that are constant across time and units. This is equivalent to running a basic regression model where the time period variable is disregarded and all observations analyzed together. While this model does have issues with disregarding the dependencies within units, it can form a baseline model for comparison purposes. But the analyst must be cautioned as the estimated coefficients may provide misleading results. The next set of models focused on varying levels of unit-specific results (similar to the random effects in the multilevel models) where intercepts, coefficients or both vary by unit. The unique model in panel models when compared to multilevel models is the time dependent effects model, where the intercept and coefficients may vary over time as well.

**SELECTING BETWEEN FIXED VERSUS RANDOM EFFECTS** The choice between fixed and random effects in panel models is primarily focused on the trade-off between controlling for endogeneity with fixed effects versus the statistical efficiency, yet biased results that result from random effects. As we discussed briefly in our earlier discussion of fixed effects, they provide a simple way to control for variables that have not or cannot be measured. This is done by analyzing each unit separately with the fixed effects. In this way, unobserved variables that do not change over time are constant over the time periods and thus cannot have any effect on the results. By using each unit as its own control, and making the observations relative to each other within the group, most forms of endogeneity are eliminated. It cannot account for unmeasured variables that change over time, but all other forms are addressed. Moreover, variables that do not vary by unit, such as gender, etc., cannot be analyzed since they have no variance over the time period. Thus, fixed effects use within-unit change for estimating all effects and increase consistency at the expense of efficiency.

Random effects, however, present a different set of benefits and limitations. They require fewer estimated parameters (primarily by not estimating an intercept for each unit) and thus increase the degrees of freedom available. Also, time-invariant variables such as gender or race can now be included in the analysis. But all of these benefits must be weighed against the potential bias from endogeneity effects that can impact random effects estimates. Thus, the choice between fixed and random effects is the trade-off of controlling endogeneity versus increased efficiency and expanded variable types.

**ADDING TIME** In addition to the cross-sectional effects we have discussed so far, panel models also provide for estimating time-variant effects just as was possible for unit-specific effects. To estimate the time-variant effects there needs to be at least enough time periods for a basic relationship to be estimated (five or more). As the number of time periods increases, the possible types of time-variant effects also increase.

**Summary** Panel models are an analytical framework designed for a specific data structure – cross-sectional analyses coupled with longitudinal data. Many of the statistical concepts underlying panel models are shared with multilevel models, making each model type more accessible to potential users. We encourage researchers to consider panel models whenever their data structures dictate, and also hope that researchers will now expand their research efforts to incorporate longitudinal aspects with the availability of this technique.

## Illustration of a Regression Analysis

The issues concerning the application and interpretation of regression analysis have been discussed in the preceding sections by following the six-stage model-building framework introduced in Chapter 1 and discussed earlier in this chapter. To provide an illustration of the important questions at each stage, an illustrative example is presented here detailing the application of multiple regression to a research problem specified by HBAT. Chapter 1 introduced a research setting in which HBAT had obtained a number of measures in a survey of customers. To demonstrate the use of multiple regression, we show the procedures used by researchers to attempt to predict customer satisfaction of the individuals in the sample with a set of 13 independent variables.

### STAGE 1: OBJECTIVES OF MULTIPLE REGRESSION

HBAT management has long been interested in more accurately predicting the satisfaction level of its customers. If successful, it would provide a better foundation for their marketing efforts. To this end, researchers at HBAT proposed that multiple regression analysis should be attempted to predict the customer satisfaction based on their perceptions of HBAT's performance. In addition to finding a way to accurately predict satisfaction, the researchers also were interested in identifying the factors that lead to increased satisfaction for use in differentiated marketing campaigns. Thus, explanation was a critical objective since this analysis was intended to provide management with the key elements necessary to improve customer satisfaction.

To apply the regression procedure, researchers selected customer satisfaction ( $X_{19}$ ) as the dependent variable ( $Y$ ) to be predicted by 13 independent variables representing perceptions of HBAT's performance ( $X_6$  to  $X_{18}$ ). Brief descriptions of the independent variables are included in Table 5.1.

The relationship among the 13 independent variables and customer satisfaction was assumed to be statistical, not functional, because it involved perceptions of performance and may include levels of measurement error.

### STAGE 2: RESEARCH DESIGN OF A MULTIPLE REGRESSION ANALYSIS

The HBAT survey obtained 100 respondents from their customer base. All 100 respondents provided complete responses, resulting in 100 observations available for analysis. The first question to be answered concerning sample size is the level of relationship ( $R^2$ ) that can be detected reliably with the proposed regression analysis.

Figure 5.7 indicates that the sample of 100, with 13 potential independent variables, is able to detect relationships with  $R^2$  values of approximately 23 percent at a power of .80 with the significance level set at .01. If the significance level is relaxed to .05, then the analysis will identify relationships explaining about 18 percent of the variance. The sample of 100 observations also meets the guideline for the minimum ratio of observations to independent variables (5:1) with an actual ratio of 7:1 (100 observations with 13 variables).

The proposed regression analysis was deemed sufficient to identify not only statistically significant relationships but also relationships that had managerial significance. Although HBAT researchers can be reasonably assured that they are not in danger of overfitting the sample, they should still validate the results if at all possible to ensure the generalizability of the findings to the entire customer base, particularly when using a stepwise estimation technique.

**Table 5.1** Dependent and Independent Variables Included in Multiple Regression Analysis

$X_6$	Product Quality	$X_{13}$	Competitive Pricing
$X_7$	E-Commerce	$X_{14}$	Warranty & Claims
$X_8$	Technical Support	$X_{15}$	New Products
$X_9$	Complaint Resolution	$X_{16}$	Ordering & Billing
$X_{10}$	Advertising	$X_{17}$	Price Flexibility
$X_{11}$	Product Line	$X_{18}$	Delivery Speed
$X_{12}$	Salesforce Image		

## STAGE 3: ASSUMPTIONS IN MULTIPLE REGRESSION ANALYSIS

Meeting the assumptions of regression analysis is essential to ensure that the results obtained are truly representative of the sample and that we obtain the best results possible. Any serious violations of the assumptions must be detected and corrected if at all possible. Analysis to ensure the research is meeting the basic assumptions of regression analysis involves two steps: (1) testing the individual dependent and independent variables and (2) testing the overall relationship after model estimation. This section addresses the assessment of individual variables. The overall relationship will be examined after the model is estimated.

The three assumptions to be addressed for the individual variables are linearity, constant variance (homoscedasticity), and normality. For purposes of the regression analysis, we summarize the results found in Chapter 2 detailing the examination of the dependent and independent variables.

First, scatterplots of the individual variables did not indicate any nonlinear relationships between the dependent variable and the independent variables. Tests for heteroscedasticity found that only two variables ( $X_6$  and  $X_{17}$ ) had minimal violations of this assumption, with no corrective action needed. Finally, in the tests of normality, six variables ( $X_6$ ,  $X_7$ ,  $X_{12}$ ,  $X_{13}$ ,  $X_{16}$ , and  $X_{17}$ ) were found to violate the statistical tests. For all but one variable ( $X_{12}$ ), transformations were sufficient remedies. In the HBAT example, we will first provide the results using the original variables and then compare these findings with the results obtained using the transformed variables.

Although regression analysis has been shown to be quite robust even when the normality assumption is violated, researchers should estimate the regression analysis with both the original and transformed variables to assess the consequences of non-normality of the independent variables on the interpretation of the results. To this end, the original variables are used first and later results for the transformed variables are shown for comparison.

## STAGE 4: ESTIMATING THE REGRESSION MODEL AND ASSESSING OVERALL MODEL FIT

With the regression analysis specified in terms of dependent and independent variables, the sample deemed adequate for the objectives of the study, and the assumptions assessed for the individual variables, the model-building process now proceeds to estimation of the regression model and assessing the overall model fit. As depicted in Figure 5.13, there are two options for variable selection: user- or software-controlled. For purposes of illustration, the stepwise procedure is employed to select variables for inclusion in the regression variate. Given the relatively small number of variables it was not deemed necessary to use a constrained approach to variable selection. After the regression model has been estimated, the variate will be assessed for meeting the assumptions of regression analysis. Finally, the observations will be examined to determine whether any observation should be deemed influential. Each of these issues is discussed in the following sections.

**Stepwise Estimation: Selecting the First Variable** The stepwise estimation procedure maximizes the incremental explained variance at each step of model building. In the first step, the highest bivariate correlation (also the highest partial correlation, because no other variables are in the equation) will be selected. The process for the HBAT example follows.

Table 5.2 displays all the correlations among the 13 independent variables and their correlations with the dependent variable ( $X_{19}$ , Customer Satisfaction). Examination of the correlation matrix (looking down the first column) reveals that complaint resolution ( $X_9$ ) has the highest bivariate correlation with the dependent variable (.603). The first step is to build a regression equation using just this single independent variable. The results of this first step appear as shown in Table 5.3.

**OVERALL MODEL FIT** From Table 5.3 the researcher can address issues concerning both overall model fit as well as the stepwise estimation of the regression model.

**Multiple R** Multiple  $R$  is the correlation coefficient (at this step) for the simple regression of  $X_9$  and the dependent variable. It has no plus or minus sign because in multiple regression the signs of the individual variables may vary, so this coefficient reflects only the degree of association. In the first step of the stepwise estimation, the Multiple  $R$  is the same as the bivariate correlation (.603) because the equation contains only one variable.

**Table 5.2** Correlation Matrix: HBAT Data

	$X_{19}$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$	$X_{16}$	$X_{17}$	$X_{18}$	
<b>Dependent Variable</b>															
$X_{19}$ Customer Satisfaction															
<b>Independent Variables</b>															
$X_6$ Product Quality		<b>.486</b>	1.000												
$X_7$ E-Commerce		<b>.283</b>	-.137	1.000											
$X_8$ Technical Support		.113	.096	.001	1.000										
$X_9$ Complaint Resolution		<b>.603</b>	.106	.140	.097	1.000									
$X_{10}$ Advertising		<b>.305</b>	-.153	<b>.430</b>	-.063	<b>.197</b>	1.000								
$X_{11}$ Product Line		<b>.551</b>	<b>.477</b>	-.053	<b>.193</b>	<b>.561</b>	-.012	1.000							
$X_{12}$ Salesforce Image		<b>.500</b>	-.152	<b>.792</b>	.017	<b>.230</b>	<b>.542</b>	-.061	1.000						
$X_{13}$ Competitive Pricing		<b>-.208</b>	<b>-.401</b>	<b>.229</b>	<b>-.271</b>	-.128	.134	<b>-.495</b>	<b>.265</b>	1.000					
$X_{14}$ Warranty & Claims		<b>.178</b>	.088	.052	<b>.797</b>	.140	.011	<b>.273</b>	.107	<b>-.245</b>	1.000				
$X_{15}$ New Products		.071	.027	-.027	-.074	.059	.084	.046	.032	.023	.035	1.000			
$X_{16}$ Order & Billing		<b>.522</b>	.104	.156	.080	<b>.757</b>	<b>.184</b>	<b>.424</b>	<b>.195</b>	-.115	<b>.197</b>	.069	1.000		
$X_{17}$ Price Flexibility		.056	<b>-.493</b>	<b>.271</b>	<b>-.186</b>	.395	.334	<b>.378</b>	<b>.352</b>	<b>.471</b>	<b>-.170</b>	.094	<b>.407</b>	1.000	
$X_{18}$ Delivery Speed		<b>.577</b>	.028	<b>.192</b>	.025	<b>.865</b>	<b>.276</b>	<b>.602</b>	<b>.272</b>	-.073	.109	.106	<b>.751</b>	<b>.497</b>	1.000

Note: Items in bold are significant at .05 level.

**R Square**  $R^2$  is the correlation coefficient squared ( $.603^2 = .364$ ), also referred to as the *coefficient of determination*. This value indicates the percentage of total variation of  $Y$  ( $X_{19}$ , Customer Satisfaction) explained by the regression model consisting of  $X_9$ .

**Standard Error of the Estimate** The standard error of the estimate is another measure of the accuracy of our predictions. It is the square root of the sum of the squared errors divided by the degrees of freedom, also represented by the square root of the  $MS_{\text{residual}}$  ( $\sqrt{89.45 \div 98} = .955$ ). It represents an estimate of the standard deviation of the actual dependent values around the regression line; that is, it is a measure of variation around the regression line. The standard error of the estimate also can be viewed as the standard deviation of the prediction errors; thus it becomes a measure to assess the absolute size of the prediction error. It is used also in estimating the size of the confidence interval for the predictions. See Neter *et al.* [85] for details regarding this procedure.

**ANOVA and F Ratio** The ANOVA analysis provides the statistical test for the overall model fit in terms of the  $F$  ratio. The total sum of squares ( $51.178 + 89.450 = 140.628$ ) is the squared error that would occur if we used only the mean of  $Y$  to predict the dependent variable. Using the values of  $X_9$  reduces this error by 36.4 percent ( $51.178 \div 140.628$ ). This reduction is deemed statistically significant with an  $F$  ratio of 56.070 and a significance level of .000.

**VARIABLES IN THE EQUATION (STEP 1)** In step 1, a single independent variable ( $X_9$ ) is used to calculate the regression equation for predicting the dependent variable. For each variable in the equation, several measures need to be defined: the regression coefficient, the standard error of the coefficient, the  $t$  value of variables in the equation, and the collinearity diagnostics (tolerance and VIF).

**Table 5.3** Example Output: Step 1 of HBAT Multiple Regression Example

Step 1-Variable Entered: $X_9$ Complaint Resolution												
Multiple R										.603		
Coefficient of Determination ( $R^2$ )										.364		
Adjusted $R^2$										.357		
Standard error of the estimate										.955		
Analysis of Variance												
	Sum of Squares			df	Mean Square			F	Sig.			
Regression	51.178			1	51.178			56.070	.000			
Residual	89.450			98	.913							
Total	140.628			99								
Variables Entered into the Regression Model												
Variables	Regression Coefficients				Statistical Significance			Correlations				
	Entered	Std. B	Error	Beta	t	Sig.	Zero-order	Partial	Part	Collinearity Statistics		
(Constant)	3.680	.443			8.310	.000				Tolerance		
$X_9$ Complaint Resolution	.595	.079	.603		7.488	.000	.603	.603	.603	VIF		
Variables Not Entered into the Regression Model												
	Statistical Significance				Partial Correlation			Collinearity Statistics				
	Beta In	t	Sig.					Tolerance	VIF			
$X_6$ Product Quality	.427	6.193	.000				.532	.989		1.011		
$X_7$ E-Commerce	.202	2.553	.012				.251	.980		1.020		
$X_8$ Technical Support	.055	.675	.501				.068	.991		1.009		
$X_{10}$ Advertising	.193	2.410	.018				.238	.961		1.040		
$X_{11}$ Product Line	.309	3.338	.001				.321	.685		1.460		
$X_{12}$ Salesforce Image	.382	5.185	.000				.466	.947		1.056		
$X_{13}$ Competitive Pricing	-.133	-1.655	.101				-.166	.984		1.017		
$X_{14}$ Warranty & Claims	.095	1.166	.246				.118	.980		1.020		
$X_{15}$ New Products	.035	.434	.665				.044	.996		1.004		
$X_{16}$ Order & Billing	.153	1.241	.218				.125	.427		2.341		
$X_{17}$ Price Flexibility	-.216	-2.526	.013				-.248	.844		1.184		
$X_{18}$ Delivery Speed	.219	1.371	.173				.138	.252		3.974		

**Regression Coefficients ( $b$  and Beta)** The regression coefficient ( $b$ ) and the standardized coefficient ( $\beta$ ) reflect the change in the dependent measure for each unit change in the independent variable. Comparison between regression coefficients allows for a relative assessment of each variable's importance in the regression model.

The value .595 is the regression coefficient ( $b_9$ ) for the independent variable ( $X_9$ ). The predicted value for each observation is the intercept (3.680) plus the regression coefficient (.595) times its value of the independent variable ( $Y = 3.680 + .595X_9$ ). The standardized regression coefficient, or beta value, of .603 is the value calculated from standardized data. With only one independent variable, the squared beta coefficient equals the coefficient of determination. The beta value enables you to compare the effect of  $X_9$  on  $Y$  to the effect of other independent variables on  $Y$  at each stage, because this value reduces the regression coefficient to a comparable unit, the number of standard deviations. (Note that at this time we have no other variables available for comparison.)

**Standard Error of the Coefficient** The standard error of the regression coefficient is an estimate of how much the regression coefficient will vary between samples of the same size taken from the same population. In a simple sense, it is the standard deviation of the estimates of  $b_9$  across multiple samples. If one were to take multiple samples of the same sample size from the same population and use them to calculate the regression equation, the standard error is an estimate of how much the regression coefficient would vary from sample to sample. A smaller standard error implies more reliable prediction and therefore smaller confidence intervals.

The standard error of  $b_9$  is .079, denoting that the 95 percent confidence interval for  $b_9$  would be  $.595 \pm (1.96 \times .079)$ , or ranging from a low of .44 to a high of .75. The value of  $b_9$  divided by the standard error ( $.595 \div .079 = 7.488$ ) is the calculated  $t$  value for a  $t$  test of the hypothesis  $b_9 = 0$  (see following discussion).

**The  $t$  Value of Variables in the Equation** The  $t$  value of variables in the equation, as just calculated, measures the significance of the partial correlation of the variable reflected in the regression coefficient. As such, it indicates whether the researcher can confidently say, with a stated level of error, that the coefficient is not equal to zero.  $F$  values may be given at this stage rather than  $t$  values. They are directly comparable because the  $t$  value is approximately the square root of the  $F$  value.

The  $t$  value is also particularly useful in the stepwise procedure in helping to determine whether any variable should be dropped from the equation once another independent variable has been added. The calculated level of significance is compared to the threshold level set by the researcher for dropping the variable. In our example, we set a .10 level for dropping variables from the equation. The critical value for a significance level of .10 with 98 degrees of freedom is 1.658. As more variables are added to the regression equation, each variable is checked to see whether it still falls within this threshold. If it falls outside the threshold (significance greater than .10), it is eliminated from the regression equation, and the model is estimated again.

In our example, the  $t$  value (as derived by dividing the regression coefficient by the standard error) is 7.488, which is statistically significant at the .000 level. It gives the researcher a high level of assurance that the coefficient is not equal to zero and can be assessed as a predictor of customer satisfaction.

**Correlations** Three different correlations are given as an aid in evaluating the estimation process. The *zero-order correlation* is the simple bivariate correlation between the independent and dependent variable. The *partial correlation* denotes the incremental predictive effect, controlling for other variables in the regression model on both dependent and independent variables. This measure is used for judging which variable is next added in sequential search methods. Finally, the *part or semi-partial correlation* denotes the unique effect attributable to each independent variable.

For the first step in a stepwise solution, all three correlations are identical (.603) because no other variables are in the equation. As variables are added, these values will differ, each reflecting their perspective on each independent variable's contribution to the regression model.

**Collinearity Diagnostics** Both collinearity measures (tolerance and VIF) are given to provide a perspective on the impact of collinearity on the independent variables in the regression equation. Remember that the tolerance value is the amount of an independent variable's predictive capability that is not predicted by the other independent variables in the equation. Thus, it represents the unique variance remaining for each variable. The VIF is the inverse of the tolerance value.

In the case of a single variable in the regression model, the tolerance is 1.00, indicating that it is totally unaffected by other independent variables (as it should be since it is the only variable in the model). Also, the VIF is 1.00. Both values indicate a complete lack of multicollinearity.

**VARIABLES NOT IN THE EQUATION** With  $X_9$  included in the regression equation, 12 other potential independent variables remain for inclusion to improve the prediction of the dependent variable. For each of these variables, four types of measures are available to assess their potential contribution to the regression model: partial correlations, collinearity measures, standardized coefficients (Beta), and  $t$  values.

**Partial Correlation and Collinearity Measures** The partial correlation is a measure of the variation in  $Y$  that can be accounted for by each of these additional variables, controlling for the variables already in the equation (only  $X_9$  in step 1). As such, the sequential search estimation methods use this value to denote the next candidate for inclusion. If the variable with the highest partial correlation exceeds the threshold of statistical significance required for inclusion, it will be added to the regression model at the next step.

The partial correlation represents the correlation of each variable not in the model with the unexplained portion of the dependent variable. As such, the contribution of the partial correlation (the squared partial correlation) is that percentage of the unexplained variance that is explained by the addition of this independent variable. Assume that the variable(s) in the regression model already account for 60 percent of the dependent measure ( $R^2 = .60$  with unexplained variance = .40). If a partial correlation has a value of .5, then the additional explained variance it accounts for is the square of the partial correlation times the amount of unexplained variance. In this simple example, that is  $.52 \times .40$ , or 10 percent. By adding this variable, we would expect the  $R^2$  value to increase by 10 percent (from .60 to .70).

For our example, the values of partial correlations range from a high of .532 to a low of .044. The  $X_6$ , with the highest value of .532, should be the variable next entered if this partial correlation is found to be statistically significant (see later section). It is interesting to note, however, that  $X_6$  had only the sixth highest bivariate correlation with  $X_{19}$ . Why was it the second variable to enter the stepwise equation, ahead of the variables with higher correlations? The variables with the second, third, and fourth highest correlations with  $X_{19}$  were  $X_{18}$  (.577),  $X_{11}$  (.551), and  $X_{16}$  (.522). Both  $X_{18}$  and  $X_{16}$  had high correlations with  $X_9$ , reflected in their rather low tolerance values of .252 and .427, respectively. It should be noted that this fairly high level of multicollinearity is not unexpected, because these three variables ( $X_9$ ,  $X_{16}$ , and  $X_{18}$ ) constituted the first factor derived in Chapter 3. The  $X_{11}$ , even though it did not join this factor, is highly correlated with  $X_9$  (.561) to the extent that the tolerance is only .685. Finally,  $X_{12}$ , the fifth highest bivariate correlation with  $X_{19}$ , only has a correlation with  $X_9$  of .230, but it was just enough to make the partial correlation slightly lower than that of  $X_6$ . The correlation of  $X_9$  and  $X_6$  of only .106 resulted in a tolerance of .989 and transformed the bivariate correlation of .486 into a partial correlation of .532, which was highest among all the remaining 12 variables.

If  $X_6$  is added, then the  $R^2$  value should increase by the partial correlation squared times the amount of unexplained variance (change in  $R^2 = .5322 \times .636 = .180$ ). Because 36.4 percent was already explained by  $X_9$ ,  $X_6$  can explain only 18.0 percent of the remaining variance.

**Standardized Coefficients** For each variable not in the equation, the standardized coefficient (Beta) that it would have if entered into the equation is calculated. In this manner, the researcher can assess the relative magnitude of this variable if added to those variables already in the equation. Moreover, it allows for an assessment of practical significance in terms of relative predictive power of the added variable.

In Table 5.3, we see that  $X_6$ , the variable with the highest partial correlation, also has the highest Beta coefficient if entered. Even though the magnitude of .427 is substantial, it can also be compared with the beta for the variable now in the model ( $X_9$  with a beta of .603), indicating that  $X_6$  will make a substantive contribution to the explanation of the regression model, as well as to its predictive capability.

**The t Values of Variables Not in the Equation** The  $t$  value measures the significance of the partial correlations for variables not in the equation. They are calculated as a ratio of the additional sum of squares explained by including a particular variable and the sum of squares left after adding that same variable. If this  $t$  value does not exceed a specified significance level (e.g., .05), the variable will not be allowed to enter the equation. The tabled  $t$  value for a significance level of .05 with 97 degrees of freedom is 1.98.

Looking at the column of  $t$  values in Table 5.3, we note that six variables ( $X_6$ ,  $X_7$ ,  $X_{10}$ ,  $X_{11}$ ,  $X_{12}$ , and  $X_{17}$ ) exceed this value and are candidates for inclusion. Although all are significant, the variable added will be that

variable with the highest partial correlation. We should note that establishing the threshold of statistical significance before a variable is added precludes adding variables with no significance even though they increase the overall  $R^2$ .

**LOOKING AHEAD** With the first step of the stepwise procedure completed, the final task is to evaluate the variables not in the equation and determine whether another variable meets the criteria and can be added to the regression model. As noted earlier, the partial correlation must be great enough to be statistically significant at the specified level (generally .05). If two or more variables meet this criterion, then the variable with the highest partial correlation is selected.

As described earlier,  $X_6$  (Product Quality) has the highest partial correlation at this stage, even though four other variables had higher bivariate correlations with the dependent variable. In each instance, multicollinearity with  $X_9$ , entered in the first step, caused the partial correlations to decrease below that of  $X_6$ .

We know that a significant portion of the variance in the dependent variable is explained by  $X_9$ , but the stepwise procedure indicates that if we add  $X_6$  with the highest partial correlation coefficient with the dependent variable and a  $t$  value is significant at the .05 level, we will make a significant increase in the predictive power of the overall regression model. Thus, we can now look at the new model using both  $X_9$  and  $X_6$ .

**Stepwise Estimation: Adding a Second Variable ( $X_6$ )** The next step in a stepwise estimation is to check and delete any of the variables in the equation that now fall below the significance threshold, and once done, add the variable with the highest statistically significant partial correlation. The following section details the newly formed regression model and the issues regarding its overall model fit, the estimated coefficients, the impact of multicollinearity, and identification of a variable to add in the next step.

**OVERALL MODEL FIT** As described in the prior section,  $X_6$  was the next variable to be added to the regression model in the stepwise procedure. The multiple  $R$  and  $R^2$  values have both increased with the addition of  $X_6$  (see Table 5.4). The  $R^2$  increased by 18.0 percent, the amount we derived in examining the partial correlation coefficient from  $X_6$  of .532 by multiplying the 63.6 percent of variation that was not explained after step 1 by the partial correlation squared ( $63.6 \times .532^2 = 18.0$ ). Then, of the 63.3 percent unexplained with  $X_9$ ,  $(.532)^2$  of this variance was explained by adding  $X_6$ , yielding a total variance explained ( $R^2$ ) of .544. The adjusted  $R^2$  also increased to .535 and the standard error of the estimate decreased from .955 to .813. Both of these measures also demonstrate the improvement in the overall model fit.

**ESTIMATED COEFFICIENTS** The regression coefficient for  $X_6$  is .364 and the beta weight is .427. Although not as large as the beta for  $X_9$  (.558),  $X_6$  still has a substantial impact in the overall regression model. The coefficient is statistically significant and multicollinearity is minimal with  $X_9$  (as described in the earlier section). Thus, tolerance is quite acceptable with a value of .989 indicating that only 1.1 percent of either variable is explained by the other.

**IMPACT OF MULTICOLLINEARITY** The lack of multicollinearity results in little change for either the value of  $b_9$  (.550) or the beta of  $X_9$  (.558) in step 1. It further indicates that variables  $X_9$  and  $X_6$  are relatively independent (the simple correlation between the two variables is .106). If the effect of  $X_6$  on  $Y$  were totally independent of the effect of  $X_9$ , the  $b_9$  coefficient would not change at all. The  $t$  values indicate that both  $X_9$  and  $X_6$  are statistically significant predictors of  $Y$ . The  $t$  value for  $X_9$  is now 8.092, whereas it was 7.488 in step 1. The  $t$  value for  $X_6$  relates to the contribution of this variable given that  $X_5$  is already in the equation. Note that the  $t$  value for  $X_6$  (6.193) is the same value shown for  $X_6$  in step 1 under the heading "Variables Not Entered into the Regression Model" (see Table 5.3).

**Table 5.4** Example Output: Step 2 of HBAT Multiple Regression Example

Step 2 – Variable Entered: $X_6$ Product Quality									
Multiple R									.738
Coefficient of Determination ( $R^2$ )									.544
Adjusted $R^2$									.535
Standard error of the estimate									.813
Analysis of Variance									
	Sum of Squares		df	Mean Square		F		Sig.	
Regression	76.527		2	38.263		57.902		.000	
Residual	64.101		97	.661					
Total	140.628		99						
Variables Entered into the Regression Model									
Variables Entered	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics
	B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance VIF
(Constant)	1.077	.564		1.909	.059				
$X_9$ Complaint Resolution	.550	.068	.558	8.092	.000	.603	.635	.555	.989 1.011
$X_6$ Product Quality	.364	.059	.427	6.193	.000	.486	.532	.425	.989 1.011
Variables Not Entered into the Regression Model									
	Statistical Significance				Collinearity Statistics				
	Beta In		t	Sig.	Partial Correlation		Tolerance	VIF	
$X_7$ E-Commerce	.275		4.256	.000	.398		.957	1.045	
$X_8$ Technical Support	.018		.261	.794	.027		.983	1.017	
$X_{10}$ Advertising	.228		3.423	.001	.330		.956	1.046	
$X_{11}$ Product Line	.066		.683	.496	.070		.508	1.967	
$X_{12}$ Salesforce Image	.477		8.992	.000	.676		.916	1.092	
$X_{13}$ Competitive Pricing	.041		.549	.584	.056		.832	1.202	
$X_{14}$ Warranty & Claims	.063		.908	.366	.092		.975	1.026	
$X_{15}$ New Products	-.026		.382	.703	.039		.996	1.004	
$X_{16}$ Order & Billing	.129		1.231	.221	.125		.427	2.344	
$X_{17}$ Price Flexibility	.084		.909	.366	.092		.555	1.803	
$X_{18}$ Delivery Speed	-.334		2.487	.015	.246		.247	4.041	

**IDENTIFYING VARIABLES TO ADD** Because  $X_9$  and  $X_6$  both make significant contributions, neither will be dropped in the stepwise estimation procedure. We can now ask “Are other predictors available?” To address this question, we can look in Table 5.4 under the section “Variables Not Entered into the Regression Model.”

Looking at the partial correlations for the variables not in the equation in Table 5.4, we see that  $X_{12}$  has the highest partial correlation (.676), which is also statistically significant at the .000 level. This variable would explain 45.7 percent of the previously unexplained variance ( $.676^2 = .457$ ), or 20.9 percent of the total variance ( $.676^2 \times .456$ ). This substantial contribution actually slightly surpasses the incremental contribution of  $X_6$ , the second variable entered in the stepwise procedure.

**Stepwise Estimation: A Third Variable ( $X_{12}$ ) is Added** The next step in a stepwise estimation follows the same pattern of (1) first checking and deleting any variables in the equation falling below the significance threshold and then (2) adding the variable with the highest statistically significant partial correlation. The following section describes the newly formed regression model and the issues regarding its overall model fit, the estimated coefficients, the impact of multicollinearity, and identification of a variable to add in the next step (see Table 5.5).

**Table 5.5** Example Output: Step 3 of HBAT Multiple Regression Example

Step 3 – Variable Entered: $X_{12}$ Salesforce Image										
Multiple R										.868
Coefficient of Determination ( $R^2$ )										.753
Adjusted $R^2$										.745
Standard error of the estimate										.602
Analysis of Variance										
	Sum of Squares			df	Mean Square			F	Sig.	
Regression	105.833			3	35.278			97.333	.000	
Residual	34.794			96	.362					
Total	140.628			99						
Variables Entered into the Regression Model										
	Regression Coefficients				Statistical Significance		Correlations			Collinearity Statistics
Variables Entered	B	Std. Error	Beta		t	Sig.	Zero-order	Partial	Part	Tolerance VIF
(Constant)	−1.569	.511			−3.069	.003				
$X_9$ Complaint Resolution	.433	.052	.439		8.329	.000	.603	.648	.423	.927 1.079
$X_6$ Product Quality	.437	.044	.512		9.861	.000	.486	.709	.501	.956 1.046
$X_{12}$ Salesforce Image	.530	.059	.477		8.992	.000	.500	.676	.457	.916 1.092
Variables Not Entered into the Regression Model										
	Statistical Significance					Collinearity Statistics				
	Beta In		t	Sig.		Partial Correlation		Tolerance	VIF	
$X_7$ E-Commerce	−.232		−2.890	.005		−.284		.372	2.692	
$X_8$ Technical Support	.013		.259	.796		.027		.983	1.017	
$X_{10}$ Advertising	−.019		−.307	.760		−.031		.700	1.428	
$X_{11}$ Product Line	.180		2.559	.012		.254		.494	2.026	
$X_{13}$ Competitive Pricing	−.094		−1.643	.104		−.166		.776	1.288	
$X_{14}$ Warranty & Claims	.020		.387	.700		.040		.966	1.035	
$X_{15}$ New Products	.016		.312	.755		.032		.996	1.004	
$X_{16}$ Order & Billing	.101		1.297	.198		.132		.426	2.348	
$X_{17}$ Price Flexibility	−.063		−.892	.374		−.091		.525	1.906	
$X_{18}$ Delivery Speed	.219		2.172	.032		.217		.243	4.110	

**OVERALL MODEL FIT** Entering  $X_{12}$  into the regression equation gives the results shown in Table 5.5. As predicted, the value of  $R^2$  increases by 20.9 percent (.753 − .544 = .209). Moreover, the adjusted  $R^2$  increases to .745 and the standard error of the estimate decreases to .602. Again, as was the case with  $X_6$  in the previous step, the new variable entered ( $X_{12}$ ) makes substantial contribution to overall model fit.

**ESTIMATED COEFFICIENTS** The addition of  $X_{12}$  brought a third statistically significant predictor of customer satisfaction into the equation. The regression weight of .530 is complemented by a beta weight of .477, second highest among the three variables in the model (behind the .512 of  $X_6$ ).

**IMPACT OF MULTICOLLINEARITY** It is noteworthy that even with the third variable in the regression equation, multicollinearity is held to a minimum. The lowest tolerance value is for  $X_{12}$  (.916), indicating that only 8.4 percent of variance of  $X_{12}$  is accounted for by the other two variables. This pattern of variables entering the stepwise procedure should be expected, however, when viewed in light of the factor analysis done in Chapter 3. From those results, we see that the three variables now in the equation ( $X_9$ ,  $X_6$ , and  $X_{12}$ ) were each members of different factors in that analysis. Because variables within the same factor exhibit a high degree of multicollinearity, it would be expected that when

one variable from a factor enters a regression equation, the odds of another variable from that same factor entering the equation are rather low (and if it does, the impact of both variables will be reduced due to multicollinearity).

**LOOKING AHEAD** At this stage in the analysis, only three variables ( $X_7$ ,  $X_{11}$ , and  $X_{18}$ ) have the statistically significant partial correlations necessary for inclusion in the regression equation. What happened to the other variables' predictive power? By reviewing the bivariate correlations of each variable with  $X_{19}$  in Table 5.1, we can see that of the 13 original independent variables, three variables had non-significant bivariate correlations with the dependent variable ( $X_8$ ,  $X_{15}$ , and  $X_{17}$ ). Thus  $X_{10}$ ,  $X_{13}$ ,  $X_{14}$ , and  $X_{16}$  all have significant bivariate correlations, yet their partial correlations are now non-significant. For  $X_{16}$ , the high bivariate correlation of .522 was reduced markedly by high multicollinearity (tolerance value of .426, denotes that less than half of original predictive power remaining). For the other three variables,  $X_{10}$ ,  $X_{13}$ , and  $X_{14}$ , their lower bivariate correlations (.305, -.208, and .178) have been reduced by multicollinearity just enough to be non-significant.

At this stage, we will skip to the final regression model and detail the entry of the final two variables ( $X_7$  and  $X_{11}$ ) in a single stage for purposes of conciseness.

**Stepwise Estimation: Fourth and Fifth Variables ( $X_7$  and  $X_{11}$ ) are Added** The final regression model (Table 5.6) is the result of two more variables ( $X_7$  and  $X_{11}$ ) being added. For purposes of conciseness, we omitted the details involved in adding  $X_7$  and will focus on the final regression model with both variables included.

**Table 5.6** Example Output: Step 5 of HBAT Multiple Regression Example

<b>Step 5—Variable Entered: <math>X_{11}</math> Product Line</b>											
Multiple R										.889	
Coefficient of Determination ( $R^2$ )										.791	
Adjusted $R^2$										.780	
Standard error of the estimate										.559	
<b>Analysis of Variance</b>											
	<b>Sum of Squares</b>		<b>df</b>	<b>Mean Square</b>		<b>F</b>		<b>Sig.</b>			
Regression	111.205		5	22.241		71.058		.000			
Residual	29.422		94	.313							
Total	140.628		99								
<b>Variables Entered into the Regression Model</b>											
<b>Variables Entered</b>	<b>Regression Coefficients</b>			<b>Statistical Significance</b>			<b>Correlations</b>		<b>Collinearity Statistics</b>		
	B	Std. Error	Beta	t	Sig.		Zero-order	Partial	Part	Tolerance	VIF
(Constant)	-.1151	.500		-.2303	.023						
$X_9$ Complaint Resolution	.319	.061	.323	5.256	.000	.603	.477	.248	.588	1.701	
$X_6$ Product Quality	.369	.047	.432	7.820	.000	.486	.628	.369	.728	1.373	
$X_{12}$ Salesforce Image	.775	.089	.697	8.711	.000	.500	.668	.411	.347	2.880	
$X_7$ E-Commerce	-.417	.132	-.245	-.3162	.002	.283	-.310	-.149	.370	2.701	
$X_{11}$ Product Line	.174	.061	.192	2.860	.005	.551	.283	.135	.492	2.033	
<b>Variables Not Entered into the Regression Model</b>											
	<b>Statistical Significance</b>				<b>Collinearity Statistics</b>						
	Beta In	t	Sig.		Partial Correlation	Tolerance				VIF	
$X_8$ Technical Support	-.009	-.187	.852		-.019	.961				1.041	
$X_{10}$ Advertising	-.009	-.162	.872		-.017	.698				1.432	
$X_{13}$ Competitive Pricing	-.040	-.685	.495		-.071	.667				1.498	
$X_{14}$ Warranty & Claims	-.023	-.462	.645		-.048	.901				1.110	
$X_{15}$ New Products	.002	.050	.960		.005	.989				1.012	
$X_{16}$ Order & Billing	.124	1.727	.088		.176	.423				2.366	
$X_{17}$ Price Flexibility	.129	1.429	.156		.147	.272				3.674	
$X_{18}$ Delivery Speed	.138	1.299	.197		.133	.197				5.075	

**OVERALL MODEL FIT** The final regression model with five independent variables ( $X_9$ ,  $X_6$ ,  $X_{12}$ ,  $X_7$ , and  $X_{11}$ ) explains almost 80 percent of the variance of customer satisfaction ( $X_{19}$ ). The adjusted  $R^2$  of .780 indicates no overfitting of the model and that the results should be generalizable from the perspective of the ratio of observations to variables in the equation (20:1 for the final model). Also, the standard error of the estimate has been reduced to .559, which means that at the 95 percent confidence level ( $\pm 1.96 \times$  standard error of the estimate), the margin of error for any predicted value of  $X_{19}$  can be calculated at  $\pm 1.1$ .

**ESTIMATED COEFFICIENTS** The five regression coefficients, plus the constant, are all significant at the .05 level, and all except the constant are significant at the .01 level. The next section (Stage 5) provides a more detailed discussion of the regression coefficients and beta coefficients as they relate to interpreting the variate.

**IMPACT OF MULTICOLLINEARITY** The impact of multicollinearity, even among just these five variables, is substantial. Of the five variables in the equation, three of them ( $X_{12}$ ,  $X_7$ , and  $X_{11}$ ) have tolerance values less than .50 indicating that over one-half of their variance is accounted for by the other variables in the equation. Moreover, these variables were the last three to enter in the stepwise process.

If we examine the zero-order (bivariate) and partial correlations, we can see more directly the effects of multicollinearity. For example,  $X_{11}$  has the third highest bivariate correlation (.551) among all 13 variables, yet multicollinearity (tolerance of .492) reduces it to a partial correlation of only .135, making it a marginal contributor to the regression equation. In contrast,  $X_{12}$  has a bivariate correlation (.500) that even with high multicollinearity (tolerance of .347) still has a partial correlation of .411. Thus, multicollinearity will always affect a variable's contribution to the regression model, but must be examined to assess the actual degree of impact.

If we take a broader perspective, the variables entering the regression equation correspond almost exactly to the factors derived in Chapter 3.  $X_9$  and  $X_6$  are each members of separate factors, with multicollinearity reducing the partial correlations of other members of these factors to a nonsignificant level.  $X_{12}$  and  $X_7$  are both members of a third factor, but multicollinearity caused a change in the sign of the estimated coefficient for  $X_7$  (see a more detailed discussion in Stage 5). Finally,  $X_{11}$  did not load on any of the factors, but was a marginal contributor in the regression model.

The impact of multicollinearity as reflected in the factor structure becomes more apparent in using a stepwise estimation procedure and will be discussed in more detail in Stage 5. Even apart from issues in explanation, however, multicollinearity can have a substantial impact on the overall predictive ability of any set of independent variables.

**LOOKING AHEAD** As noted earlier, the regression model at this stage consists of the five independent variables with the addition of  $X_{11}$ . Examining the partial correlations of variables not in the model at this stage (see Table 5.6), we see that none of the remaining variables have a significant partial correlation at the .05 level needed for entry. Moreover, all of the variables in the model remain statistically significant, avoiding the need to remove a variable in the stepwise process. Thus, no more variables are considered for entry or exit and the model is finalized.

**An Overview of the Stepwise Process** The stepwise estimation procedure is designed to develop a regression model with the fewest number of statistically significant independent variables and maximum predictive accuracy. In doing so, however, the regression model can be markedly affected by issues such as multicollinearity. Moreover, the researcher relinquishes control over the formation of the regression model and runs a higher risk of decreased generalizability. The following section provides an overview of the estimation of the stepwise regression model discussed earlier from the perspective of overall model fit. Issues relating to interpretation of the variate, other estimation procedures, and alternative model specifications will be addressed in subsequent sections.

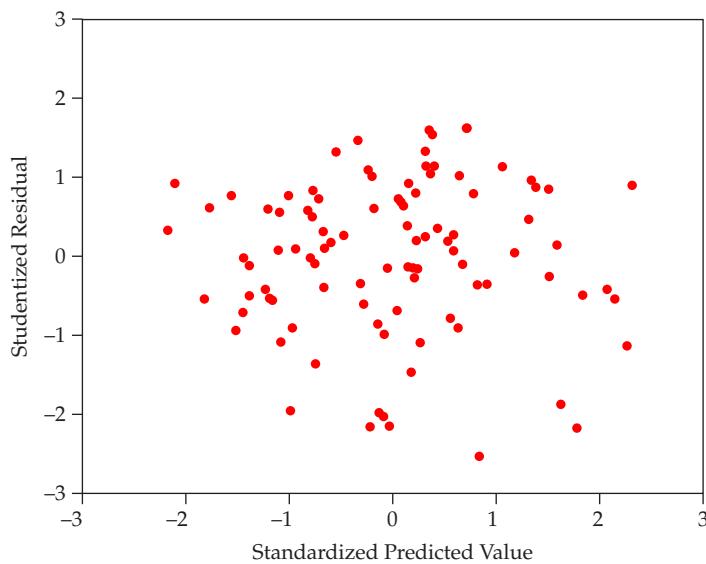
Table 5.7 provides a step-by-step summary detailing the measures of overall fit for the regression model used by HBAT in predicting customer satisfaction. Each of the first three variables added to the equation made substantial contributions to the overall model fit, with substantive increases in the  $R^2$  and adjusted  $R^2$  while also decreasing the standard error of the estimate. With only the first three variables, 75 percent of the variation in customer satisfaction is explained with a confidence interval of  $\pm 1.2$ . Two additional variables are added to arrive at the final model, but these variables, although statistically significant, make much smaller contributions. The  $R^2$  increases by 3 percent and the confidence interval decreases to  $\pm 1.1$ , an improvement of .1. The relative impacts of each variable will be discussed in Stage 5, but the stepwise procedure highlights the importance of the first three variables in assessing overall model fit.

**Table 5.7** Model Summary of Stepwise Multiple Regression Model

Model Summary									
Step	Overall Model Fit			Std. Error of the Estimate	$R^2$ Change	$R^2$ Change Statistics			Significance of $R^2$ Change
	$R$	$R^2$	Adjusted $R^2$			F Value of $R^2$	Change	df1	df2
1	.603	.364	.357	.955	.364	56.070	1	98	.000
2	.738	.544	.535	.813	.180	38.359	1	97	.000
3	.868	.753	.745	.602	.208	80.858	1	96	.000
4	.879	.773	.763	.580	.020	8.351	1	95	.005
5	.889	.791	.780	.559	.018	8.182	1	94	.005

Step 1:  $X_9$  Complaint Resolution  
Step 2:  $X_9$  Complaint Resolution,  $X_6$  Product Quality  
Step 3:  $X_9$  Complaint Resolution,  $X_6$  Product Quality,  $X_{12}$  Salesforce Image  
Step 4:  $X_9$  Complaint Resolution,  $X_6$  Product Quality,  $X_{12}$  Salesforce Image,  $X_7$  E-Commerce Activities  
Step 5:  $X_9$  Complaint Resolution,  $X_6$  Product Quality,  $X_{12}$  Salesforce Image,  $X_7$  E-Commerce Activities,  $X_{11}$  Product Line

Note: Constant (intercept term) included in all regression models.



**Figure 5.22**  
Analysis of Standardized Residuals

In evaluating the estimated equation, we considered statistical significance. We must also address two other basic issues: (1) meeting the assumptions underlying regression and (2) identifying the influential data points. We consider each of these issues in the following sections.

**Evaluating the Variate for the Assumptions of Regression Analysis** To this point, we examined the individual variables for meeting the assumptions required for regression analysis in Step 3. However, we must also evaluate the variate for meeting these assumptions as well. The assumptions to examine are linearity, homoscedasticity, independence of the residuals, and normality. The principal measure used in evaluating the regression variate is the residual—the difference between the actual dependent variable value and its predicted value. For comparison, we use the studentized residuals, a form of standardized residuals (see Key Terms).

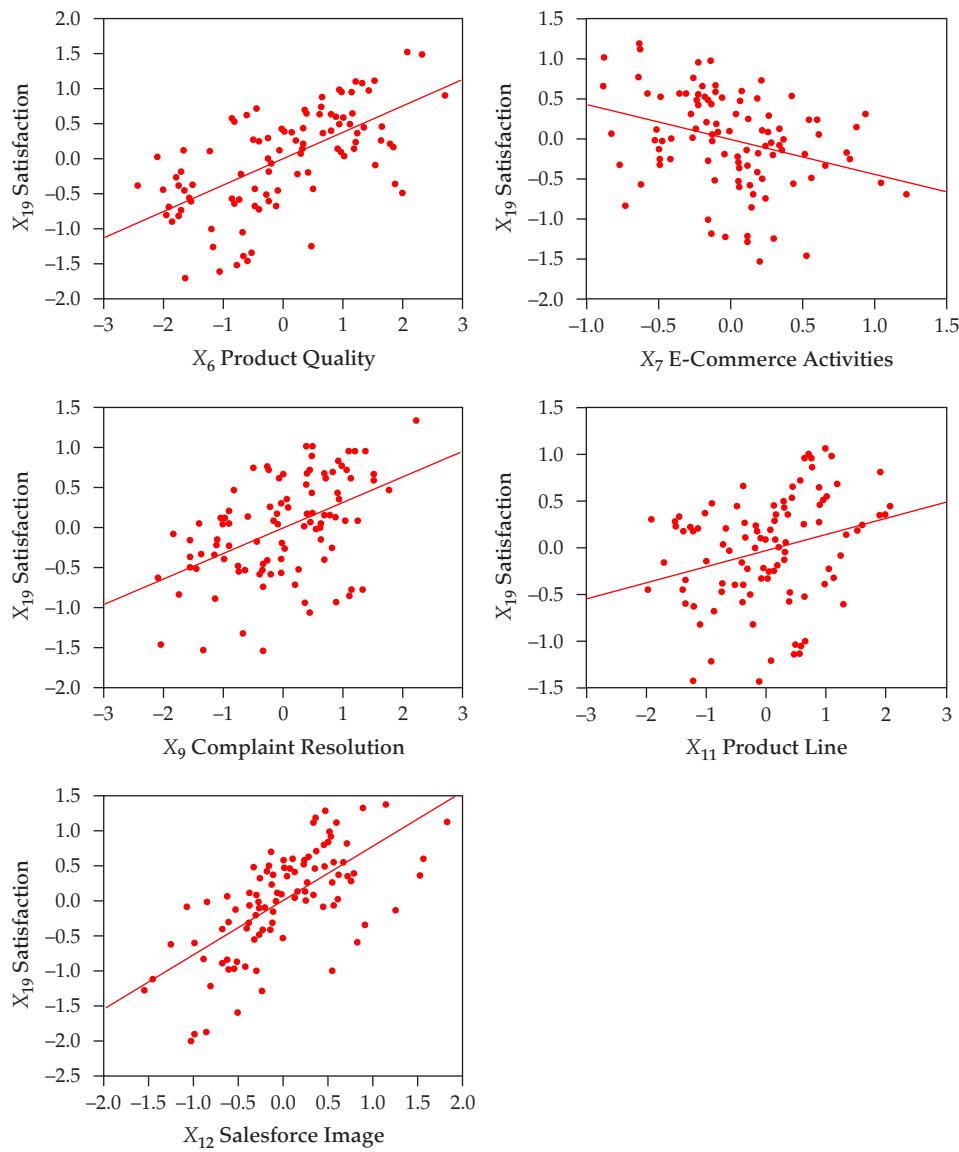
The most basic type of residual plot is shown in Figure 5.22, the studentized residuals versus the predicted values. As we can see, the residuals fall within a generally random pattern, similar to the null plot in Figure 5.11a. However, we must make specific tests for each assumption to check for violations.

**LINEARITY** The first assumption, linearity, will be assessed through an analysis of residuals (testing the overall variate) and partial regression plots (for each independent variable in the analysis).

Figure 5.22 does not exhibit any nonlinear pattern to the residuals, thus ensuring that the overall equation is linear. We must also be certain, when using more than one independent variable, that each independent variable's relationship is also linear to ensure its best representation in the equation. To do so, we use the partial regression plot for each independent variable in the equation. In Figure 5.23 we see that the relationships for  $X_6$ ,  $X_9$ , and  $X_{12}$  are reasonably well defined; that is, they have strong and significant effects in the regression equation. Variables  $X_7$  and  $X_{11}$  are less well defined, both in slope and scatter of the points, thus explaining their lesser effect in the equation (evidenced by the smaller coefficient, beta value, and significance level). For all five variables, no nonlinear pattern is shown, thus meeting the assumption of linearity for each independent variable.

**HOMOSCEDASTICITY** The next assumption deals with the constancy of the residuals across values of the independent variables. Our analysis is again through examination of the residuals (Figure 5.22), which shows no pattern of

**Figure 5.23**  
Standardized Partial Regression Plots



increasing or decreasing residuals. This finding indicates homoscedasticity in the multivariate (the set of independent variables) case. Moreover, no pattern of heteroscedasticity seen in Figure 5.23.

**INDEPENDENCE OF THE RESIDUALS** The third assumption deals with the effect of carryover from one observation to another, thus making the residual not independent. When carryover is found in such instances as time series data, the researcher must identify the potential sequencing variables (such as time in a time series problem) and plot the residuals by this variable. For example, assume that the identification number represents the order in which we collect our responses. We could plot the residuals and see whether a pattern emerges.

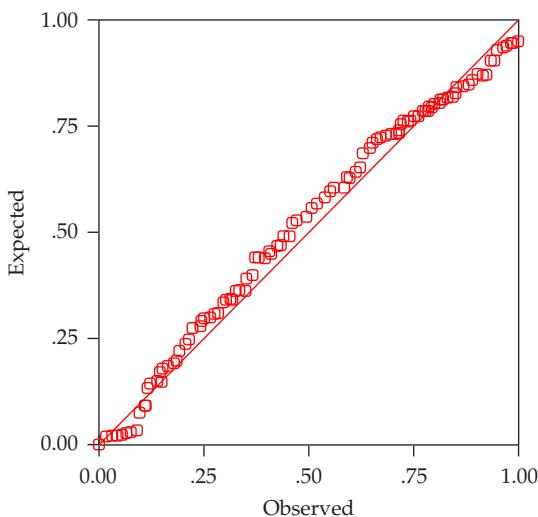
In our example, several variables, including the identification number and each independent variable, were tried and no consistent pattern was found. We must use the residuals in this analysis, not the original dependent variable values, because the focus is on the prediction errors, not the relationship captured in the regression equation. Also, there were no omitted variables that suggested dependency among the observations.

**NORMALITY** The final assumption we will check is normality of the error term of the variate with a visual examination of the normal probability plots of the residuals. As shown in Figure 5.24, the values fall along the diagonal with no substantial or systematic departures; thus, the residuals are considered to represent a normal distribution. The regression variate is found to meet the assumption of normality.

**APPLYING REMEDIES FOR ASSUMPTION VIOLATIONS** After testing for violations of the four basic assumptions of multivariate regression for both individual variables and the regression variate, the researcher should assess the impact of any remedies on the results.

In the examination of individual variables in Chapter 2, the only remedies needed were the transformations of  $X_6$ ,  $X_7$ ,  $X_{12}$ ,  $X_{13}$ ,  $X_{16}$ , and  $X_{17}$  in terms of meeting the normality assumption. A set of differing transformations were used, including the squared term ( $X_6$  and  $X_{16}$ ), logarithm ( $X_7$ ), cubed term ( $X_{13}$ ), and inverse ( $X_{16}$ ). Only in the case of  $X_{12}$  did the transformation not achieve normality. If we substitute these variables for their original values and re-estimate the regression equation with a stepwise procedure, we achieve almost identical results. The same variables enter the equation with no substantive differences in either the estimated coefficients or overall model fit as assessed with  $R^2$  and standard error of the estimate. The independent variables not in the equation still show non-significant levels for entry—even those that were transformed. Thus, in this case, the remedies for violating the assumptions improved the prediction slightly but did not alter the substantive findings.

While the visual examination of the residuals did not suggest substantive issues with heteroscedasticity, the model was estimated with heteroscedasticity-consistent standard errors (HCSE) for comparative purposes. Table 5.8 contains the original standard errors along with the HCSE estimates. As noted earlier, the HCSE process only impacts the standard errors. As can be seen in Table 5.8, estimation of the HCSE adjustments does not change the standard



**Figure 5.24**  
Normal Probability Plot: Standardized Residuals

**Table 5.8** Corrections for Potential Heteroscedasticity: HCSE Standard Errors

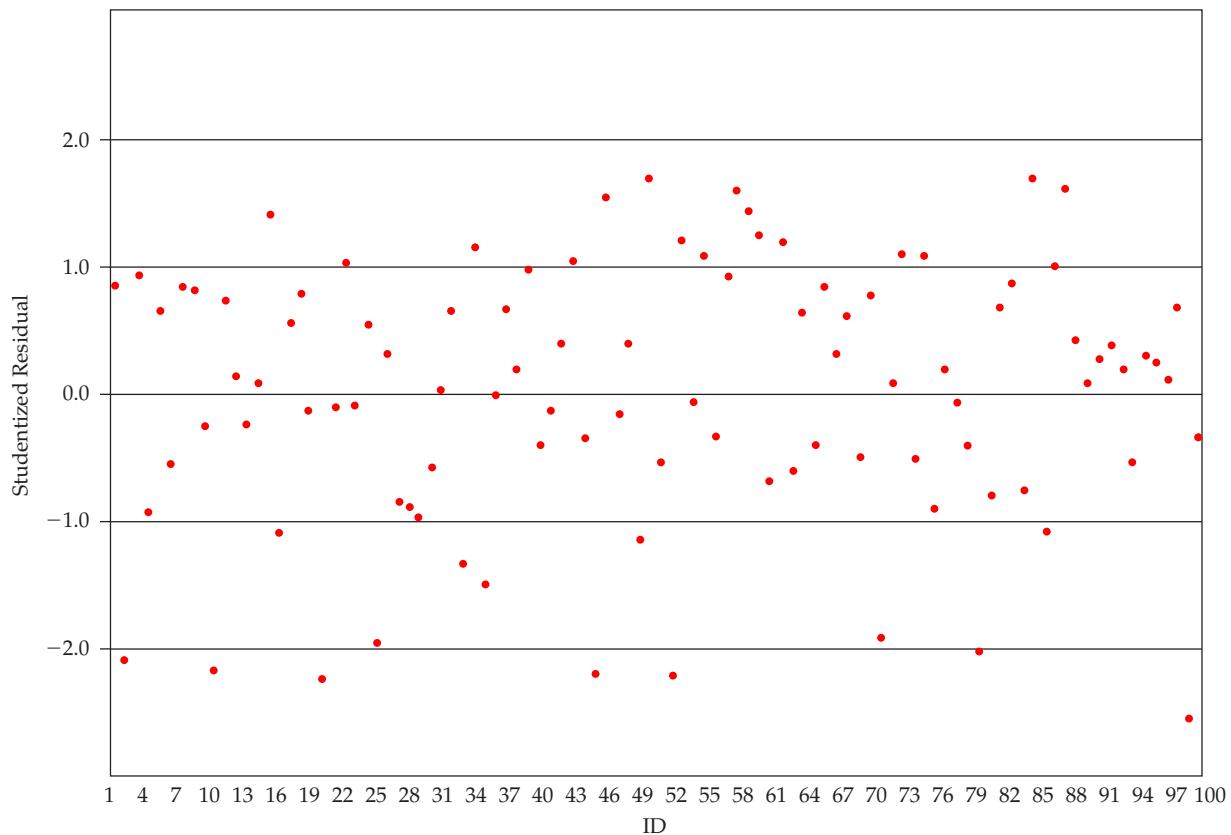
Variable	Original Standard Error			Heteroscedasticity Consistent Standard Error		
	Standard Error	t Value	Significance	Standard Error	t Value	Significance
Intercept	0.49984	-2.3	0.0235	0.45082	-2.55	0.0123
X <sub>6</sub>	0.04719	7.82	< .0001	0.04566	8.08	< .0001
X <sub>7</sub>	0.13192	-3.16	0.0021	0.11992	-3.48	0.0008
X <sub>9</sub>	0.06068	5.26	< .0001	0.05675	5.62	< .0001
X <sub>11</sub>	0.06095	2.86	0.0052	0.05265	3.31	0.0013
X <sub>12</sub>	0.08898	8.71	< .0001	0.09886	7.84	< .0001

errors or significance levels in any substantive fashion. This would be expected with the low to moderate levels of multicollinearity in this set of variables, but the differences would be more marked if multicollinearity levels were higher. We will see an example of these effects in a later example.

**Identifying Outliers as Influential Observations** For our final analysis, we attempt to identify any observations that are influential (having a disproportionate impact on the regression results) and determine whether they should be excluded from the analysis. Although more detailed procedures are available for identifying outliers as influential observations, we address the use of residuals in identifying outliers in the following section.

**ANALYSIS OF RESIDUALS** The most basic diagnostic tool involves the residuals and identification of any outliers—that is, observations not predicted well by the regression equation that have large residuals. Figure 5.25 shows the studentized residuals for each observation. Because the values correspond to *t* values, upper and lower limits can be set once the

**Figure 5.25**  
Plot of Studentized Residuals



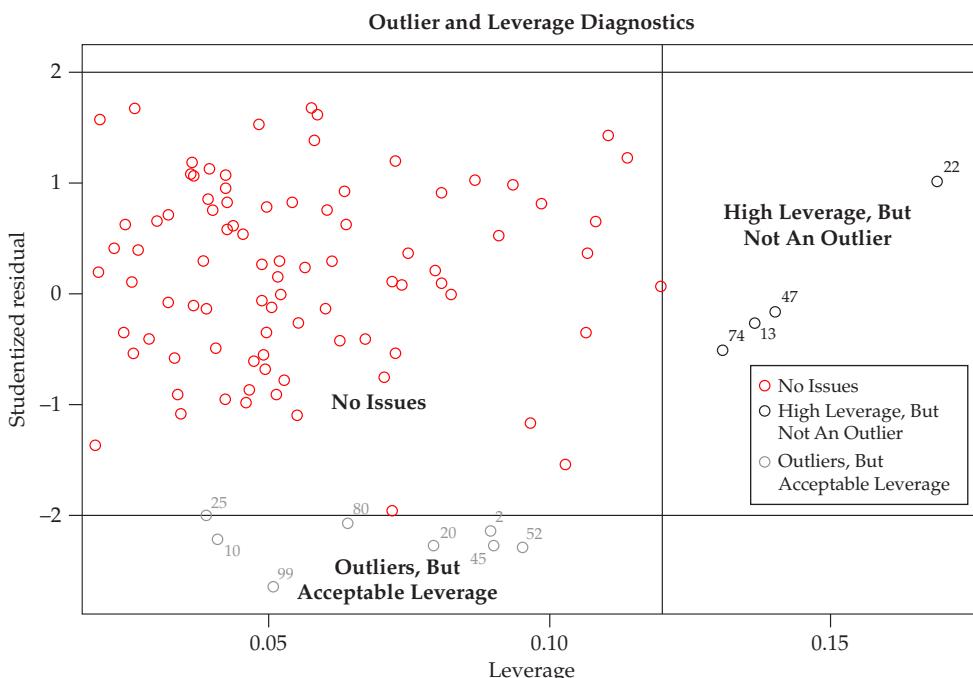
desired confidence interval has been established. Perhaps the most widely used level is the 95 percent confidence interval ( $\alpha = .05$ ). The corresponding  $t$  value is 1.96, thus identifying statistically significant residuals as those with residuals greater than this value (1.96). Seven observations can be seen in Figure 5.25 (2, 10, 20, 45, 52, 80, and 99) to have significant residuals and thus be classified as outliers. Outliers are important because they are observations not represented by the regression equation for one or more reasons, any one of which may be an influential effect on the equation that requires a remedy.

**Analysis of Partial Regression Plots** Examination of the residuals also can be done through the partial regression plots (see Figure 5.23). These plots help to identify influential observations for each independent–dependent variable relationship. Consistently across each graph in Figure 5.23, the points at the lower portion are those observations identified as having high negative residuals (observations 2, 10, 20, 45, 52, 80, and 99 in Figure 5.25). These points are not well represented by the relationship and thus affect the partial correlation as well.

**Analysis of Residual vs Leverage Plot** A final graphical analysis is the Residual versus Leverage Plot (see Figure 5.26) which combines the residual analysis discussed above with leverage values, which characterize the observations on the variate variables (i.e., the set of independent variables). We see two distinct patterns of potential influential observations. The first are those observations identified as outliers (i.e., standardized residuals outside the 1.96 threshold). In this instance, eight observations have negative residuals outside this range. The second pattern is four observations that have high leverage values (i.e., beyond the threshold of  $2p \div n = .12$ ). While there are outliers and high leverage observations, no observations have both characteristics. Thus, further analysis of the measures of overall influence or influence on specific estimated parameters is needed.

**ANALYSIS OF MEASURES OF OVERALL AND INDIVIDUAL INFLUENCE** To complement the residual and leverage diagnostics in identifying influential observations, an overall measure of influence (DFFITS) and the influences on the model parameter estimates (DFBETAs) are also considered. Table 5.9 contains all of the observations which exceeded the thresholds for standardized residuals ( $\pm 1.96$ ), leverage ( $2p \div n = .12$ ), DFFITS ( $2 \times \sqrt{p \div n} = .49$ ) or any of the DFBETAs ( $\pm .2$ ).

**Figure 5.26**  
Diagnosing Influential Observations: Residual Versus Leverage Plot



**Table 5.9** Diagnostic Measures for Identifying Influential Observations

ID	Standardized		Leverage	DFFITS	intercept	$X_6$	DFBETA			
	Residual						$X_7$	$X_9$	$X_{11}$	$X_{12}$
2	x			x				x		x
10	x					x	x			
13		x								
20	x			x	x	x		x	x	x
22		x			x				x	
25	x					x				
35			x				x	x	x	x
43					x		x			
45	x			x				x		x
46						x				
47		x								
49							x			x
52	x			x				x		x
53							x			
59			x			x	x			x
60						x		x	x	
71			x		x	x			x	x
74		x								
80	x			x			x	x		x
85							x			
87					x	x		x		
88							x			
99	x			x	x					x

As seen in the residual versus leverage plot earlier, no observation was both an outlier and high leverage. Moreover, only one of the high leverage observations (#22) had any other impact and that was on three of the estimated parameters. But several of the outliers had substantially more impact (#20 had high DFFITS as well as impacting five of the six estimated parameters, while #35 and #71 also had high DFFITS and impacted four of the estimated parameters). Four additional observations (#59, 60, 80 and 87) impacted three parameters, but only two of those had high DFFITS. So for purposes of this analysis, the three observations with high DFFITS and impacts on more than three parameters) were deemed influential. It is interesting to note that observation 35 was neither a high leverage nor an outlier, although the residual value was  $-1.53$ . Thus, extending the analysis beyond just the residual versus leverage plot did identify an additional influential observation.

**DELETION OF INFLUENTIAL OBSERVATIONS** Based on the influential analysis, three observations (#20, 35 and 71) are deleted and the results compared to the original results. In viewing the overall model fit measures, the results are slightly improved, but that would be expected since all three of these observations were outliers. What is more impactful, however, is the change in the final model variate, where  $X_{11}$  in the original model is replaced by  $X_{18}$  in the reduced sample model. In the original model,  $X_{11}$  was the final variable added and  $X_{18}$  had the next highest  $t$  value for entry, but it was not significant (see Table 5.6). Likewise, in the reduced sample results,  $X_{18}$  was the final variable added and  $X_{11}$  the variable with highest non-significant  $t$  value for entry. Thus, we see that deleting these three observations shifted the marginal variables from significant to non-significant in the two sets of results. Another impact was to reduce the coefficient for  $X_9$  even though it was the first variable entered in both models. Here its collinearity with  $X_{18}$  had an impact in the reduced sample model.

**Table 5.10 Impact of Deleting Three Influential Observations**

Model Results	Original Sample	Sample Minus Influential Observations
R	0.889	0.905
R Square	0.791	0.819
Adjusted R Square	0.78	0.809
Std. Error of the Estimate	0.559	0.527
Estimated Coefficients		
Intercept	-1.151	-1.939
$X_9$ Complaint Resolution	0.319	0.19
$X_6$ Product Quality	0.369	0.472
$X_{12}$ Salesforce Image	0.775	0.833
$X_7$ E-Commerce	-0.417	-0.47
$X_{11}$ Product Line	0.174	
$X_{18}$ Delivery Speed		0.42

**SUMMARY** These results do not necessarily support either retention or deletion of the observations, that is an issue for further analysis. What is important, however, is demonstrating the impact of deleting even a few observations can have on the results. Researchers are cautioned when approaching the issue of influential observations and the decision to eliminate observations from the sample. Results from both the original sample and any reduced sample should always be compared to understand the full implications of such an action.

## STAGE 5: INTERPRETING THE REGRESSION VARIATE

With the model estimation completed, the regression variate specified, and the diagnostic tests that confirm the appropriateness of the results administered, we can now examine our predictive equation based on five independent variables ( $X_6$ ,  $X_7$ ,  $X_9$ ,  $X_{11}$ , and  $X_{12}$ ).

**Interpretation of the Regression Coefficients** The first task is to evaluate the regression coefficients for the estimated signs, focusing on those of unexpected direction.

The section of Table 5.6 headed “Variables Entered into the Regression Equation” yields the prediction equation from the column labeled “Regression Coefficient: B.” From this column, we read the constant term (-1.151) and the coefficients (.319, .369, .775, -.417, and .174) for  $X_9$ ,  $X_6$ ,  $X_{12}$ ,  $X_7$ , and  $X_{11}$ , respectively. The predictive equation would be written:

$$Y = -1.151 + .319X_9 + .369X_6 + .775X_{12} + (-.417)X_7 + .174X_{11}$$

*Note:* The coefficient of  $X_7$  is included in parentheses to avoid confusion due to the negative value of the coefficient.

With this equation, the expected customer satisfaction level for any customer could be calculated if that customer's evaluations of HBAT are known. For illustration, let us assume that a customer rated HBAT with a value of 6.0 for each of these five measures. The predicted customer satisfaction level for that customer would be:

$$\begin{aligned} \text{Predicted Customer} &= -1.151 + (.319 \times 6) + (.369 \times 6) + (.775 \times 6) + ((-.417) \times 6) \\ &\quad + (.174 \times 6) \\ &= -1.151 + 1.914 + 2.214 + 4.650 - 2.502 + 1.044 \\ &= 6.169 \end{aligned}$$

We first start with an interpretation of the constant. It is statistically significant (significance = .023), thus making a substantive contribution to the prediction. However, because in our situation it is highly unlikely that any respondent would have zero ratings on all the HBAT perceptions, the constant merely plays a part in the prediction process and provides no insight for interpretation.

In viewing the regression coefficients, the sign is an indication of the relationship (positive or negative) between the independent and dependent variables. All of the variables except one have positive coefficients. Of particular note is the reversed sign for  $X_7$  (E-Commerce), suggesting that an increase in perceptions on this variable has a negative impact on predicted customer satisfaction. All the other variables have positive coefficients, meaning that more positive perceptions of HBAT (higher values) increase customer satisfaction.

Does  $X_7$ , then, somehow operate differently from the other variables? In this instance, the bivariate correlation between  $X_7$  and customer satisfaction is positive, indicating that when considered separately,  $X_7$  has a positive relationship with customer satisfaction, just as the other variables. We will discuss in the following section the impact of multicollinearity on the reversal of signs for estimated coefficients.

**Assessing Variable Importance** In addition to providing a basis for predicting customer satisfaction, the regression coefficients also provide a means of assessing the relative importance of the individual variables in the overall prediction of customer satisfaction. In any interpretation of the regression variate, the researcher must be aware of the impact of multicollinearity. As discussed earlier, highly collinear variables can distort the results substantially or make them quite unstable and thus not generalizable. We will first examine the levels of multicollinearity in the stepwise solution where hopefully the stepwise approach has minimized the level of multicollinearity. We will then examine the various measures of variable importance to identify the varying levels of importance for this set of variables. We should note that the same process will be undertaken in a later examination of the confirmatory approach where all 13 variables are entered into the equation and multicollinearity is expected to be much more impactful.

**Measuring the Degree and Impact of Multicollinearity** Two measures are available for testing the impact of collinearity: (1) calculating the tolerance and VIF values and (2) using the condition indices and decomposing the regression coefficient variance online. The tolerance value is 1 minus the proportion of the variable's variance explained by the other independent variables. Thus, a high tolerance value indicates little collinearity, and tolerance values approaching zero indicate that the variable is almost totally accounted for by the other variables (high multicollinearity). The variance inflation factor is the reciprocal of the tolerance value; thus we look for small VIF values as indicative of low correlation among variables.

**DIAGNOSING MULTICOLLINEARITY** In our example, tolerance values for the variables in the equation range from .728 ( $X_6$ ) to .347 ( $X_{12}$ ), indicating a wide range of multicollinearity effects (see Table 5.6). Likewise, the VIF values range from 1.373 to 2.701. Even though none of these values indicate levels of multicollinearity that should seriously distort the regression variate, we must be careful even with these levels to understand their effects, especially on the stepwise estimation process. The following section will detail some of these effects on both the estimation and interpretation process.

A second approach to identifying multicollinearity and its effects is through the decomposition of the coefficient variance. Table 5.11 shows both the condition indices and the variance proportions of each coefficient. Using the two step process, we first identify any condition indices exceeding (at a minimum) 30 and find that all of the condition indices are below this threshold. This is expected since the stepwise procedure minimizes multicollinearity by including variables based on the partial correlations that account for collinearity with variables in the model. We would expect, however, to perhaps see more multicollinearity when variables are entered in a confirmatory fashion into the model.

**IMPACTS DUE TO MULTICOLLINEARITY** Although multicollinearity is not so high that the researcher must take corrective action before valid results can be obtained, multicollinearity still has impact on the estimation process, particularly on the composition of the variate and the estimated regression coefficients.

If you examine the bivariate correlations, after  $X_9$  (the first variable added to the regression variate in the stepwise process), the second-highest correlation with the dependent variable is  $X_{18}$  (Delivery Speed), followed by  $X_{11}$  (Product Line), and  $X_{16}$  (Order & Billing). Yet due to collinearity with  $X_9$ , the second variable entered was  $X_6$ , only the sixth highest bivariate correlation with  $X_{19}$ .

**Table 5.11** Condition Indexes and Variance Decompositions for Assessing Multicollinearity

Number	Eigenvalue	Condition Index	(Constant)	Variance Proportions				
				$X_9$	$X_6$	$X_{12}$	$X_7$	$X_{11}$
1	5.858	1.000	0.00	0.00	0.00	0.00	0.00	0.00
2	0.073	8.935	0.00	0.02	0.04	0.06	0.04	0.09
3	0.037	12.661	0.02	0.38	0.24	0.00	0.00	0.01
4	0.015	19.668	0.12	0.41	0.08	0.01	0.06	0.78
5	0.010	24.543	0.65	0.05	0.53	0.27	0.05	0.04
6	0.007	28.647	0.21	0.14	0.11	0.65	0.84	0.08

The impacts of multicollinearity are seen repeatedly throughout the estimation process, such that the final set of five variables entered into the regression model ( $X_6$ ,  $X_7$ ,  $X_9$ ,  $X_{11}$ , and  $X_{12}$ ) represent the first, sixth, fifth, eighth, and third highest correlations with the dependent variable, respectively. Variables with the second highest correlation ( $X_{18}$  at .577) and fourth highest correlation ( $X_{16}$  at .522) are never entered into the regression model. Does their exclusion mean they are unimportant? Are they lacking in impact? If a researcher went only by the estimated regression model, multicollinearity would cause serious problems in interpretation. What happened is that  $X_{16}$  and  $X_{18}$  are highly correlated with  $X_9$ , to such an extent that they have little unique explanatory power apart from that shared with  $X_9$ . Yet by themselves, or if  $X_9$  was not allowed to enter the model, they would be important predictors of customer satisfaction. The extent of multicollinearity among these three variables is evidenced in Chapter 3, where these three variables were found to all form one of the four factors arising from the HBAT perceptions.

In addition to affecting the composition of the variate, multicollinearity has a distinct impact on the signs of the estimated coefficients. In this situation, it relates to the collinearity between  $X_{12}$  (Salesforce Image) and  $X_7$  (E-Commerce). As noted in our earlier discussion about multicollinearity, one possible effect is the reversal of sign for an estimated regression coefficient from the expected direction represented in the bivariate correlation. Here, the high positive correlation between  $X_{12}$  and  $X_7$  (correlation = .792) causes the sign for the regression coefficient for  $X_7$  to change from positive (in the bivariate correlation) to a negative sign. If the researcher did not investigate the extent of multicollinearity and its impact, the inappropriate conclusion might be drawn that increases in E-Commerce activities decrease customer satisfaction.

Thus, the researcher must understand the basic relationships supported by the conceptual theory underlying the original model specification and make interpretation based on this theory, not just on the estimated variate. The researcher must never allow an estimation procedure to dictate the interpretation of the results, but instead must understand the issues of interpretation accompanying each estimation procedure. For example, if all 13 independent variables are entered into the regression variate, the researcher must still contend with the effects of collinearity on the interpretation of the coefficients, but in a different manner than if stepwise were used.

**Measures of Variable Importance** The impacts of multicollinearity discussed above not only influence the estimation process, but also the ability to discern the total impact of each independent variable in the regression model, whether it be for purposes of prediction or explanation. We will first examine a series of measures directly related to the regression results—bivariate correlations, squared semi-partial correlations, regression weights and standardized regression weights. Then a series of additional measures will be reviewed which extend beyond just the single regression model—structure correlations, commonality, dominance analysis and relative weights. Table 5.12 provides these measures for the five variables in the final stepwise model. Each of these measures provides some unique insights into the importance of the independent variables in the regression results. We will also examine the results of all possible subsets regression to compare the final regression model to any number of alternatives.

**BIVARIATE CORRELATIONS** The first measure is the bivariate correlation between each independent variable and the dependent variable  $X_{19}$ . While we know that multicollinearity reduces this relationship once multiple variables enter the regression variate, these correlations still provide the most fundamental perspective on the relationships with

**Table 5.12** Stepwise Regression Results: Measures of Variable Importance

Measure of Variable Importance	$X_6$	$X_7$	$X_9$	$X_{11}$	$X_{12}$
Bivariate Correlation with $X_{19}$	0.4863	0.2827	<b>0.6033</b>	0.5505	0.5002
Squared Semi-partial (Part) correlation	0.1361	0.0223	0.0615	0.0182	<b>0.1689</b>
Regression weight	0.3690	-0.4171	0.3190	0.1744	<b>0.7751</b>
Standardized regression weight (Beta)	0.4323	-0.2452	0.3234	0.1924	<b>0.6974</b>
Structure correlation (with predicted value)	0.5469	0.3180	<b>0.6784</b>	0.6191	0.5625
Commonality					
Unique	0.1361	0.0223	0.0615	0.0182	<b>0.1689</b>
Shared	0.1004	0.0577	<b>0.3024</b>	0.2849	0.0813
Total	0.2365	0.0799	<b>0.3639</b>	0.3031	0.2502
Dominance Analysis:					
Average Predictor Contributions					
Overall and to Models of Each Size					
Overall	0.1850	0.0541	0.1885	0.1478	<b>0.2154</b>
Models of 1 variable	0.2365	0.0799	<b>0.3639</b>	0.3031	0.2502
Models of 2 variables	0.2124	0.0741	<b>0.2525</b>	0.2141	0.2415
Models of 3 variables	0.1827	0.0562	0.1633	0.1340	<b>0.2203</b>
Models of 4 variables	0.1573	0.0381	0.1012	0.0696	<b>0.1961</b>
Models of 5 variables	0.1361	0.0223	0.0615	0.0182	<b>0.1689</b>
Relative Weights	0.1831	0.0478	0.1812	0.1315	<b>0.2051</b>

Note: Bold values indicate highest value.

the dependent variable. Yet while many would accept these correlations as the “truth” which must be reflected in the model results, we have also discussed issues such as mediation and confounding (see Chapter 6 for a more complete discussion) may give rise to prescribing these as “causal” factors. As such, the researcher should always understand the bivariate correlations as the “first look” at the relationships. In this example, the highest bivariate correlation is for  $X_9$ , which is the reason it enters the stepwise process as the first variable. But it is closely followed by three other variables ( $X_{11}$ ,  $X_{12}$ , and  $X_6$ ), with only  $X_7$  a somewhat lower correlation. As expected given the stepwise approach used for variable selection, all of the variables have substantial bivariate correlations with the dependent variable.

**SQUARED SEMI-PARTIAL CORRELATIONS** The squared semi-partial correlations represent the amount that  $R^2$  would decrease if the variable were eliminated from the regression variate. Thus, they quantify in some measure the unique explanatory impact of each variable that is also reflected in the regression and standardized regression weights. As we can see in Table 5.12,  $X_{12}$  has the highest value, followed closely by  $X_6$ . The other three variables have markedly lower values. Of note is the lower value for  $X_9$  that had the highest bivariate correlation, but its unique impact is diminished by multicollinearity. We should also note that the sum of the squared semi-partial correlations is roughly only 40 percent, which about half of the models  $R^2$  value of 79 percent. The difference is the amount of shared variance among the independent variables. So as we can see, even when a stepwise approach is used to reduce multicollinearity, the amount of shared explanation is substantial.

**REGRESSION AND BETA WEIGHTS** In this situation, all the variables are expressed on the same scale, but we will use the beta coefficients for comparison between independent variables. Using the beta coefficients from Table 5.12, the researcher can make direct comparisons among the variables to determine their relative importance in the regression variate based on their *unique explanatory impact*. For our example,  $X_{12}$  (Salesforce Image) was the most important, followed by  $X_6$  (Product Quality),  $X_9$  (Complaint Resolution),  $X_7$  (E-Commerce), and finally  $X_{11}$  (Product Line). This is the same pattern seen with the squares semi-partial correlations. With a steady decline in size of the beta

coefficients across the variables, it is difficult to categorize variables as high, low, or otherwise. However, viewing the relative magnitudes does indicate that, for example,  $X_{12}$  (Salesforce Image) shows a more marked effect (three times as much) than  $X_{11}$  (Product Line). Thus, to the extent that salesforce image can be increased uniquely from other perceptions, it represents the most direct way, *ceterus paribus*, of increasing customer satisfaction.

**STRUCTURE COEFFICIENTS** The next measure of importance is the structure coefficient, which is the correlation of each variable with the *predicted value* of  $X_{19}$ , not the actual value. In this sense it most directly reflects the relationship between each independent variable and the predicted outcome. But since this is a bivariate correlation, it reflects both unique and shared relationships with the predicted value. The ordering of the variables also follows that of the bivariate correlations with the dependent variable, with  $X_9$  having the highest value and  $X_7$  the lowest.

**RELATIVE WEIGHTS** The next measure of variable importance involves a data transformation to (a) identify uncorrelated predictors of the dependent variable and (b) relate the original independent variables to these uncorrelated predictors to quantify each independent variable's importance. While expressed in terms of model fit (i.e.,  $R^2$ ), it should be noted that relative weights do not distinguish between unique and shared effect. In our example (see Table 5.12),  $X_{12}$  has the highest value, followed closely by  $X_6$  and then  $X_9$ . There is relatively little difference among these three variables, thus reflecting their role as the major impacts on  $X_{19}$ .

**COMMONALITY ANALYSIS** While the measures discussed to this point have only dealt with the estimated regression model, the final two measures extend past this single model and evaluate the independent variables in a series of models to assess each variable's impact across this entire set. In the case of commonality and dominance analysis (see next section) all possible subsets regression is used to provide a complete range of potential models involving this set of variables. Commonality analysis focused on the distinction between unique and shared variance across all of the possible models. As we can see in Table 5.12,  $X_{12}$  has the greatest level of unique variance explained, but  $X_9$  has the greatest level of shared variance explained. The difference between these two variables is a markedly different pattern. The  $X_{12}$  has .1689 in unique variance and .0813 in shared variance, while  $X_{12}$  has the opposite pattern (.0615 in unique variance and .3024 in shared variance). This markedly higher level of shared variance compared to all of the other variables results in  $X_9$  also having the highest overall impact. The distinction between unique and shared variation across all of the potential model specifications gives the researcher a much better understanding of the source of impact for each independent variable.

**DOMINANCE ANALYSIS** This final measure also employs all possible subsets regression, but focuses on variable importance both overall and by the number of variables in the model. The overall measure is an average across all possible models as to variable impact, but the method also provides average impacts by model size. In our situation we can see a distinct pattern that supports the earlier findings of  $X_9$  being more heavily weighted towards shared impact while  $X_{12}$  is more oriented towards unique impact. This is reflected by  $X_9$  having the largest impacts in the one and two variable models, where the effects of multicollinearity are the smallest. But as the model sizes increase and thus the effects of multicollinearity increase, the largest impact shifts to  $X_{12}$  where its unique effects persist. Thus, these results differ from commonality analysis in that  $X_{12}$  was deemed the most impactful overall based on its higher levels of impact in more complex model specifications.

**ALL POSSIBLE SUBSETS REGRESSION** While all possible subsets regression does not directly reflect variable importance, it is useful to assess the relative model fit of alternative model specifications and thus the role of variables in those models. Table 5.13 shows the results of the top ten models as ranked by adjusted  $R^2$ . As would be expected in this situation, the final regression model is ranked first. But several models are very close in terms of  $R^2$  and the other measures of overall fit. When we examine the confirmatory approach of including all the independent variables in the model, the all possible subsets regression results will present a number of alternatives to the five-variable stepwise solution.

**SUMMARY** The set of measures of variable importance highlight several points regarding the set of independent variables. First, most of the measures of variable importance directly associated with the regression model point to  $X_{12}$  because it has the largest unique impact of all of the independent variables. This is reflected in the squared

**Table 5.13** All Possible Subsets Regression Results: Top Ten Models (Based on  $R^2$ ) for Stepwise Solution

Model Index	Number of Variables in the Model	Variables in Model	Adjusted R-Square	R-Square	Mallows C(p)	AIC	BIC	MSE	Standard Error of the Estimate
1	5	$X_6 X_7 X_9 X_{11} X_{12}$	0.7797	0.7908	6.0000	-110.342	-107.585	0.31300	0.55947
2	4	$X_6 X_7 X_9 X_{12}$	0.763	0.7726	12.1817	-103.997	-102.211	0.33666	0.58023
3	4	$X_6 X_9 X_{11} X_{12}$	0.7588	0.7685	13.9979	-102.235	-100.622	0.34265	0.58536
4	3	$X_6 X_9 X_{12}$	0.7448	0.7526	19.1640	-97.5712	-96.3948	0.36244	0.60203
5	4	$X_6 X_7 X_{11} X_{12}$	0.7179	0.7293	31.6300	-86.5735	-86.4151	0.40074	0.63304
6	3	$X_6 X_{11} X_{12}$	0.6903	0.6997	42.9097	-78.2113	-78.4153	0.43986	0.66322
7	4	$X_7 X_9 X_{11} X_{12}$	0.6401	0.6547	65.1505	-62.2323	-64.0396	0.51118	0.71497
8	3	$X_9 X_{11} X_{12}$	0.6155	0.6272	75.5105	-56.5673	-58.1163	0.54616	0.73902
9	3	$X_7 X_{11} X_{12}$	0.6103	0.6221	77.7649	-55.2304	-56.8558	0.55351	0.74398
10	4	$X_6 X_7 X_9 X_{11}$	0.6060	0.6219	79.8847	-53.1598	-55.6119	0.55973	0.74815

semi-partial correlations, regression and beta weights and the relative weights. But when we employ methods that examine the variables in the context of a range of potential models that can be specified with this set of variables, then the role of unique versus shared impact becomes more pronounced, and  $X_9$  is deemed most impactful due to its higher levels of shared impact. Neither of the results is definitive, and the researcher should consider both aspects—unique impact and shared impact—and determine how they align with the objectives of the analysis when interpreting the regression variate.

**SUMMARY** Interpretation of the regression variate is of critical importance in achieving the objective of explanation and enabling the researcher to provide insights into how the relationships between the variate and the dependent variable is apportioned across the independent variables. This becomes a key element when the researcher is attempting to provide insights into the elements that give rise to the outcome, whether the research context is academic or organizational. The influence of multicollinearity on the impact of the independent variables and how their impact can be reflected in shared versus unique components is the reason for the additional measures of variable importance that look beyond just the estimated model results.

## STAGE 6: VALIDATING THE RESULTS

The final task facing the researcher involves the validation process of the regression model. The primary concern of this process is to ensure that the results are generalizable to the population and not specific to the sample used in estimation. The most direct approach to validation is to obtain another sample from the population and assess the correspondence of the results from the two samples. In the absence of an additional sample, the researcher can assess the validity of the results in several approaches, including an assessment of the adjusted  $R^2$  or estimating the regression model on two or more subsamples of the data (see Table 5.14).

**Adjusted Coefficient of Determination** Examining the adjusted  $R^2$  value reveals little loss in predictive power when compared to the  $R^2$  value (.780 versus .791, see Table 5.14), which indicates a lack of overfitting that would be shown by a more marked difference between the two values. Moreover, with five variables in the model, it maintains an adequate ratio of observations to variables in the variate.

**Split-Sample Estimation** A second approach is to divide the sample into two subsamples, estimate the regression model for each subsample, and compare the results. Table 5.14 contains the stepwise models estimated for two subsamples of 50 observations each. Comparison of the overall model fit demonstrates a high level of similarity of the

**Table 5.14** Split-Sample Validation of Stepwise Estimation

Overall Model Fit		Sample 1	Sample 2
Multiple R		.910	.888
Coefficient of Determination ( $R^2$ )		.828	.788
Adjusted $R^2$		.808	.769
Standard error of the estimate		.564	.529

## Analysis of Variance

	SAMPLE 1					SAMPLE 2				
	Sum of Squares	df	Mean Square	F	Sig.	Sum of Squares	df	Mean Square	F	Sig.
Regression	67.211	5	13.442	2.223	.000	46.782	4		41.747	.000
Residual	14.008	44	.318			12.607	45	.280		
Total	81.219	49				59.389	49			

## Variables Entered into the Stepwise Regression Model

Variables	SAMPLE 1						SAMPLE 2					
	Regression Coefficients			Statistical Significance			Regression Coefficients			Statistical Significance		
	Entered	B	Std. Error	Beta	t	Sig.	B	Std. Error	Beta	t	Sig.	
(Constant)		−1.413	.736		−1.920	.061	−.689	.686		−1.005	.320	
$X_{12}$ Salesforce Image		1.069	.151	.916	7.084	.000	.594	.105	.568	5.679	.000	
$X_6$ Product Quality		.343	.066	.381	5.232	.000	.447	.062	.518	7.170	.000	
$X_7$ E-Commerce		−.728	.218	−.416	−3.336	.002	−.349	.165	−.212	−2.115	.040	
$X_{11}$ Product Line		.295	.078	.306	3.780	.000						
$X_{16}$ Order & Billing		.285	.115	.194	2.473	.017						
$X_9$ Complaint Resolution							.421	.070	.445	5.996	.000	

results in terms of  $R^2$ , adjusted  $R^2$ , and the standard error of the estimate. Yet in comparing the individual coefficients, some differences do appear. In sample 1,  $X_9$  did not enter in the stepwise results as it did in sample 2 and the overall sample. Instead,  $X_{16}$ , highly collinear with  $X_9$ , entered. Moreover,  $X_{12}$  had a markedly greater beta weight in sample 1 than found in the overall results. In the second sample, four of the variables entered as with the overall results, but  $X_{11}$ , the least forceful variable in the overall results, did not enter the model. The omission of  $X_{11}$  in one of the subsamples confirms that it was a marginal predictor, as indicated by the low beta and  $t$  values in the overall model.

**Calculating the PRESS Statistic** An alternative to a split sample or cross-validation is the PRESS (predicted residual error sum of squares) statistic, which is a variation of the jackknife or hold-one-out approach. As described earlier, a separate regression equation is estimated for each time a different observation is “held out” and then its predicted value is from the model based on all of the other observations in the sample. In this manner each observation has no impact on the model estimation used to make its predicted value.

The research can compare the PRESS value to the residual sum of squares from the original model to determine the degree to which there is an increase. In this example the original residual sum of squares is 29.42 out of a total sum of squares of 140.63, resulting in an  $R^2$  of .7908. We can calculate a comparable measure, the coefficient of prediction, using the PRESS value of 33.53 and obtain a value of .7662, quite close to the original  $R^2$ . From this we can conclude that the model is generalizable and we would expect to achieve comparable levels of model fit on other samples of observations.

**Summary** The concept of validation is extremely important for models oriented to either prediction or explanation. In both cases the degree of generalizability is at stake and the researcher must be assured that the model would produce comparable results on other samples of comparable observations. While the PRESS statistic is quite useful in overall model validation, only some form of split sample or cross-validation allows for assessment of the individual variables as well, which is of critical importance if explanation is an objective of the analysis.

## Evaluating Alternative Regression Models

The stepwise regression model examined in the previous discussion provided a solid evaluation of the research problem as formulated. However, it represents only one of the variable selection options available to the researcher. The researcher is always well served in evaluating alternative regression models in the search of additional explanatory power and confirmation of earlier results. In this section, we examine three additional regression models: (1) a model including all 13 independent variables in a confirmatory approach to assess the other option in variable selection, (2) a model substituting summated scales developed in Chapter 3 for the multicollinear independent variables to address the advantages/disadvantages of using composite measures and (3) a model adding a nonmetric variable ( $X_3$ , Firm Size) through the use of a dummy variable.

### CONFIRMATORY REGRESSION MODEL

A primary alternative to the stepwise regression estimation method in variable selection is the confirmatory approach, whereby the researcher specifies the independent variable to be included in the regression equation. In this manner, the researcher retains complete control over the regression variate in terms of both prediction and explanation. This approach is especially appropriate in situations of replication of prior research efforts or for validation purposes.

In this situation, the confirmatory perspective involves the inclusion of all 13 perceptual measures as independent variables. These same variables are considered in the stepwise estimation process, but in this instance all are directly entered into the regression equation at one time. Here the researcher can judge the potential impacts of multicollinearity on the selection of independent variables and the effect on overall model fit from including all thirteen variables.

The primary comparisons between the stepwise and confirmatory procedures involve (1) examination of the overall model fit of each procedure, (2) the interpretations drawn from each set of results, (3) a review the multicollinearity diagnostics given the confirmatory approach, and finally (4) examining the all possible subsets regression method to compare competing models to the confirmatory and stepwise models.

**Impact on Overall Model Fit** The results in Table 5.15 are similar to the final results achieved through stepwise estimation (see Table 5.6), with two notable exceptions:

**DECREASE IN MODEL FIT** Even though more independent variables are included, the overall model fit decreases. Whereas the coefficient of determination increases (.889 to .897) because of the additional independent variables, the adjusted  $R^2$  decreases slightly (.780 to .774), which indicates the inclusion of several independent variables that were non-significant in the regression equation. Although they contribute to the overall  $R^2$  value, they detract from the adjusted  $R^2$ . This change illustrates the role of the adjusted  $R^2$  in comparing regression variates with differing numbers of independent variables.

**Table 5.15** Multiple Regression Results Using a Confirmatory Estimation Approach with All 13 Independent Variables

Confirmatory Specification with 13 Variables	
Multiple R	.897
Coefficient of Determination ( $R^2$ )	.804
Adjusted $R^2$	.774
Standard error of the estimate	.566

## Analysis of Variance

	Sum of Squares	df	Mean Square	F	Sig.
Regression	113.044	13	8.696	27.111	.000
Residual	27.584	86	.321		
Total	140.628	99			

## Variables Entered into the Regression Model

	Variables Entered	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
	(Constant)	-1.336	1.120		-1.192	.236					
X <sub>6</sub>	Product Quality	.377	.053	.442	7.161	.000	.486	.611	.342	.598	1.672
X <sub>7</sub>	E-Commerce	-.456	.137	-.268	-3.341	.001	.283	-.339	-.160	.354	2.823
X <sub>8</sub>	Technical Support	.035	.065	.045	.542	.589	.113	.058	.026	.328	3.047
X <sub>9</sub>	Complaint Resolution	.154	.104	.156	1.489	.140	.603	.159	.071	.207	4.838
X <sub>10</sub>	Advertising	-.034	.063	-.033	-.548	.585	.305	-.059	-.026	.646	1.547
X <sub>11</sub>	Product Line	.362	.267	.400	1.359	.178	.551	.145	.065	.026	37.978
X <sub>12</sub>	Salesforce Image	.827	.101	.744	8.155	.000	.500	.660	.389	.274	3.654
X <sub>13</sub>	Competitive Pricing	-.047	.048	-.062	-.985	.328	-.208	-.106	-.047	.584	1.712
X <sub>14</sub>	Warranty & Claims	-.107	.126	-.074	-.852	.397	.178	-.092	-.041	.306	3.268
X <sub>15</sub>	New Products	-.003	.040	-.004	-.074	.941	.071	-.008	-.004	.930	1.075
X <sub>16</sub>	Order & Billing	.143	.105	.111	1.369	.175	.522	.146	.065	.344	2.909
X <sub>17</sub>	Price Flexibility	.238	.272	.241	.873	.385	.056	.094	.042	.030	33.332
X <sub>18</sub>	Delivery Speed	-.249	.514	-.154	-.485	.629	.577	-.052	-.023	.023	44.004

**INCREASE IN SEE** Another indication of the overall poorer fit of the confirmatory model is the increase in the standard error of the estimate (SEE) from .559 to .566, which illustrates that overall  $R^2$  should not be the sole criterion for predictive accuracy because it can be influenced by many factors, one being the number of independent variables.

**Impact on Variate Interpretation** The other substantive difference is in the regression variate, where multicollinearity affects the number and strength of the significant variables.

**FEWER SIGNIFICANT VARIABLES** First, only three variables ( $X_6$ ,  $X_7$ , and  $X_{12}$ ) are statistically significant, whereas the stepwise model contains two more variables ( $X_9$  and  $X_{11}$ ). In the stepwise model,  $X_{11}$  was the least significant variable, with a significance level of .005. When the confirmatory approach is used, the multicollinearity with other variables (as indicated by its tolerance value of .026) renders it non-significant. The same happens to  $X_9$ , which was the first variable entered in the stepwise solution, but it now has a non-significant coefficient in the confirmatory model. Again, multicollinearity had a sizeable impact, reflected in its tolerance value of .207—an 80 percent overlap with other variables in the model.

**INCREASE IN MULTICOLLINEARITY** The impact of multicollinearity on other variables not in the stepwise model is also substantial. In the confirmatory approach, three variables ( $X_{11}$ ,  $X_{17}$ , and  $X_{18}$ ) have tolerance values under .05 (with corresponding VIF values of 33.3, 37.9, and 44.0), meaning that 95 percent or more of their variance is accounted for by the other HBAT perceptions. In such situations, it is practically impossible for these variables to be significant predictors. Six other variables have tolerance values under .50, indicating that the regression model variables account for more than half of the variance in these variables.

Given the high levels of multicollinearity, the additional collinearity diagnostics of condition indices and decomposition of variance were performed. As shown in Table 5.16, six of the condition indices exceed the threshold of 30 to merit further examination. Only the last condition index, however, had two or more variables with values exceeding .90. In this instance the variables were  $X_{11}$ ,  $X_{17}$ , and  $X_{18}$ , the same variables that had tolerance/VIF values indicating high multicollinearity. If warranted, these variables could be examined for combination in some form of composite or perhaps only a subset used in further analyses.

Another implication of multicollinearity is inflation of standard errors and the potential for heteroscedasticity. To address this concern, HCSE estimates were obtained to assess whether they impacted any of the significance tests. As seen in Table 5.17, the estimation of HCSE standard errors resulted in one additional variable ( $X_{11}$ ) becoming statistically significant ( $p = .014$ ). Thus, the impact of multicollinearity and heteroscedasticity on the standard errors was substantial. We should note, however, that  $X_9$ , the variable with a substantial shared impact, was still not significant in the confirmatory model.

Given that multicollinearity provided for the creation of four well-developed factors in Chapter 3, here the inclusion of all variables creates issues in estimation and interpretation. We will explore the use of summated scales as a replacement for the individual variables in a later section.

**Measures of Variable Importance** The final issue facing the analyst using a confirmatory approach is the assessment of variable importance in the face of possible multicollinearity. Even with adjustments with the HCSE estimates, only four of the thirteen variables are statistically significant even though the model  $R^2$  is 80 percent. The set of measures for variable importance discussed earlier can be employed to assess as best possible the impacts (shared and unique) for each of the independent variables. Rather than pursue this approach however, the next section illustrates the use of summated scales to avoid these sets of multicollinear variables that make interpretation difficult. The interested reader is encouraged to apply the measures discussed earlier for purposes of explanation using the original variables.

**Summary** The confirmatory approach provides the researcher with control over the regression variate, but at the possible cost of a regression equation with poorer prediction and explanation if the researcher does not closely examine the results. The two approaches to variable selection—user-controlled (confirmatory) versus software-controlled (sequential approaches)—both have strengths and weaknesses that should be considered in their use, but the prudent researcher will employ both approaches in order to address the strengths of each.

## USE OF SUMMATED SCALES AS REMEDIES FOR MULTICOLLINEARITY

Up to this point all of the regression models have used the full set of 13 independent variables, either in a confirmatory or stepwise model. But the researcher has another option available in variable specification—the creation of composite measures, such as summated scales, to replace the original independent variables and hopefully mitigate to some extent the potentially high levels of multicollinearity.

**Table 5.16** Collinearity Diagnostics for Confirmatory Model with All 13 Variables

Number	Eigenvalue	Condition Index	Proportion of Variation													
			Intercept	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$	$X_{16}$	$X_{17}$	$X_{18}$
1	13.48216	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.14409	9.67308	0.000	0.011	0.003	0.031	0.001	0.027	0.002	0.004	0.027	0.002	0.000	0.000	0.003	0.000
3	0.09208	12.10019	0.000	0.001	0.004	0.055	0.020	0.002	0.001	0.003	0.018	0.003	0.018	0.025	0.001	0.001
4	0.08148	12.86311	0.000	0.004	0.005	0.030	0.001	0.016	0.000	0.005	0.001	0.000	0.000	0.064	0.001	0.000
5	0.0643	14.48068	0.000	0.019	0.009	0.049	0.002	0.254	0.001	0.010	0.029	0.002	0.000	0.005	0.005	0.000
6	0.04786	16.78455	0.003	0.067	0.004	0.069	0.001	0.174	0.001	0.000	0.152	0.000	0.171	0.001	0.001	0.000
7	0.03102	20.8462	0.001	0.049	0.113	0.002	0.001	0.392	0.000	0.091	0.045	0.000	0.062	0.000	0.001	0.000
8	0.01913	26.5457	0.001	0.305	0.010	0.003	0.007	0.031	0.015	0.001	0.237	0.000	0.001	0.108	0.006	0.002
9	0.0123	33.10879	0.003	0.092	0.000	0.008	0.036	0.035	0.000	0.003	0.149	0.011	0.004	0.724	0.022	0.003
10	0.00965	37.38604	0.035	0.098	0.117	0.083	0.263	0.000	0.008	0.155	0.184	0.048	0.031	0.000	0.012	0.002
11	0.00724	43.15705	0.011	0.008	0.589	0.184	0.031	0.040	0.000	0.500	0.053	0.148	0.036	0.000	0.001	0.000
12	0.00539	50.00918	0.049	0.090	0.042	0.237	0.635	0.019	0.009	0.082	0.035	0.173	0.002	0.040	0.006	0.023
13	0.00292	67.90061	0.427	0.253	0.085	0.247	0.000	0.001	0.003	0.075	0.057	0.611	0.000	0.093	0.001	0.011
14	0.00037354	189.98195	0.471	0.003	0.019	0.001	0.002	0.010	<b>0.961</b>	0.070	0.011	0.002	0.011	0.003	<b>0.941</b>	<b>0.958</b>

Note: Bold values indicate patterns of high multicollinearity.

**Table 5.17** HCSE Estimates for Confirmatory Model with 13 Variables

	Original Standard Error estimates			HCSE Estimates		
	Standard Error	t value	Significance	Standard Error	t value	Significance
<b>Intercept</b>	1.12031	-1.19	0.2364	0.99482	-1.34	0.1829
<b>x6</b>	0.05271	7.16	< .0001	0.04954	7.62	< .0001
<b>x7</b>	0.13651	-3.34	<b>0.0012</b>	0.11256	-4.05	<b>0.0001</b>
<b>x8</b>	0.06492	0.54	0.5891	0.05883	0.6	0.5511
<b>x9</b>	0.1036	1.49	0.1401	0.09424	1.64	0.1052
<b>x10</b>	0.06283	-0.55	0.5853	0.05416	-0.64	0.5269
<b>x11</b>	0.26669	1.36	0.1778	0.14449	2.51	<b>0.014</b>
<b>x12</b>	0.10146	8.15	< .0001	0.1005	8.23	< .0001
<b>x13</b>	0.0482	-0.98	0.3275	0.03583	-1.32	0.1888
<b>x14</b>	0.12553	-0.85	0.3965	0.10979	-0.97	0.3326
<b>x15</b>	0.03953	-0.07	0.9409	0.03219	-0.09	0.9275
<b>x16</b>	0.10452	1.37	0.1746	0.10275	1.39	0.1674
<b>x17</b>	0.27249	0.87	0.385	0.15732	1.51	0.1341
<b>x18</b>	0.5141	-0.48	0.6291	0.28609	-0.87	0.3862

Note: Significant values in bold

In this section we will examine four regression models that illustrate the basic methods available by combining options from both variable specification and variable selection. We will review the models using all thirteen independent variables that employed confirmatory and stepwise variable selection as well as estimate two additional models—substituting summated scales for most of the independent variables and then using the confirmatory and stepwise approaches to variable selection. Comparison of these four models provides coverage of all of the basic model types and highlights the advantages and disadvantages of each combination. Models 1 and 2 use the original set of thirteen variables, while Models 3 and 4 use the summated scales. Table 5.18 shows the correspondence of summated scales and individual variables as well as the bivariate correlations of each variable with  $X_{19}$ . The application of exploratory factor analysis in Chapter 3 is the basis for creating the four summated scales, each representing two or three of the original variables. Three variables ( $X_{11}$ ,  $X_{15}$ , and  $X_{17}$ ) were not included in the final factor results and thus are included here as individual variables in Models 3 and 4 to make equivalent comparisons.

**Overall Model Fit** The first comparison across models is in terms of model fit. Models 1 and 2, using all thirteen independent variables, show a substantially higher model fit than the regression models using the summated scales— $R^2$  of about 80 percent versus 64 percent respectively. This is expected since the summated scales primarily represent the shared variation among the variables in each scale and in doing so decrease the multicollinearity among the summated scales to levels much lower than found among the individual variables (i.e., highest VIF among summated sales is 1.15). The result is a reduction in multicollinearity among the summated scales, but with each scale representing the collinearity of the variables within the scale.

**Regression Coefficients** As discussed earlier, Models 1 and 2 using all of the individual independent variables had only a small number of variables achieve statistical significance. Model 1, using the confirmatory approach, had three significant variables while the stepwise model had five significant variables. Yet nine of the thirteen variables had significant bivariate correlations with  $X_{19}$ . What about the variables that were not significant – does this mean they are not important? Especially troubling is that in Model 1 none of the variables contained in Summated Scale 1 achieved significance, yet all of them had significant bivariate correlations (ranking first, second and fourth among all variables). Since summated scales are based on collinearity among the variables in the scale, this multicollinearity may have diminished the predictive relationship for each of these variables enough such that none were significant.

**Table 5.18** Comparison of Model Options for Variable Specification and Variable Selection

Independent Variables		Independent Variable Specification				Summarized Scales			
		Original Variables: $X_6$ to $X_{18}$		Variable Selection		Variable Selection		Model 4:	
Summarized Scales	Independent Variable	Rank of Bivariate Correlation With $X_{19}$	Model Fit		Model 1: Confirmatory	Model 2: Stepwise	Model 3: Confirmatory	Model 4: Stepwise	
			$R^2$	Adj. $R^2$					
Summarized Scale 1	$X_9$	1			0.774	0.780	0.622	0.6278	
	$X_{16}$	4			0.154	0.319	0.614	0.601	
	$X_{18}$	2			-0.249				
Summarized Scale 2	$X_7$	8			<b>-0.456</b>	<b>-0.417</b>	<b>0.633</b>	<b>0.608</b>	
	$X_{10}$	7			-0.034				
	$X_{12}$	5			<b>0.827</b>	<b>0.775</b>			
Summarized Scale 3	$X_8$	11			0.035		-0.032		
	$X_{14}$	10			-0.107				
Summarized Scale 4	$X_6$	6			<b>0.377</b>	<b>0.369</b>	<b>0.344</b>	<b>0.430</b>	
	$X_{13}$	9			-0.047				
Not in Summarized Scale	$X_{11}$	3			0.362	<b>0.174</b>	0.067		
	$X_{15}$	12			-0.003		0.015		
	$X_{17}$	13			0.238		-0.094		

Note: Significant values in bold.

This conclusion is reinforced since in Model 2 the stepwise approach found one of the variables ( $X_9$ ) to be significant. Viewing the results of Models 3 and 4 demonstrates the benefits of using summated scales when facing high levels of multicollinearity among the independent variables. Three of the four summated scales were significant in both models, and the other scale (#3) was based on two variables, neither of which had significant bivariate correlations with  $X_{19}$ . These three factors represent eight of the nine variables with significant bivariate correlations, something not achieved in either Models 1 or 2.

**Summary** Managing the variate through variable specification and variable selection options provides the researcher with a number of alternatives that can identify relationships for specific variables as well as the broader composite measures. By examining a number of model forms, the researcher can gain a clear insight into not only the estimated relationships, but also the underlying relationships among variables that impact the estimated models.

## INCLUDING A NONMETRIC INDEPENDENT VARIABLE

The prior discussion focused on the confirmatory estimation method as an alternative for possibly increasing prediction and explanation, but the researcher also should consider the possible improvement from the addition of nonmetric independent variables. As discussed in an earlier section and in Chapter 2, nonmetric variables must be modified (re-coded) to be directly included in the regression equation. The process involves the creation of a series of new variables (dummy variables) that represent the separate categories of the nonmetric variable.

In this example, the variable of firm size ( $X_3$ ), which has the two categories (large and small firms), will be included in the stepwise estimation process. The variable is already coded in the appropriate form, with large firms (500 or more employees) coded as 1 and small firms as 0. The variable can be directly included in the regression equation to represent the difference in customer satisfaction between large and small firms, given the other variables in the regression equation. Specifically, because large firms have the value 1, small firms act as the reference category. The coding is developed to facilitate interpretation of the new dummy variable in the regression model, with the larger number (1) representing the large firms and the smaller number (0) representing the smaller firms (the reference group).

The regression coefficient is interpreted as the value for large firms compared to small firms. A positive coefficient indicates that large firms have higher customer satisfaction than small firms, whereas a negative coefficient indicates that small firms have higher customer satisfaction. The amount of the coefficient represents the difference in customer satisfaction between the means of the two groups, controlling for all other variables in the model.

Table 5.19 contains the results of the addition of  $X_3$  in a stepwise model, where it was added with the five variables that formed the stepwise model earlier in this section (see Table 5.6). Examination of the overall fit statistics indicates minimal improvement, with all of the measures ( $R^2$ , adjusted  $R^2$ , and SEE) increasing over the original stepwise model (see Table 5.6).

When we examine the regression coefficients, note that the coefficient for  $X_3$  is .271 and is significant at the .03 level. The positive value of the coefficient indicates that large firms, given their characteristics on the other five independent variables in the equation, still have a customer satisfaction level that is about a quarter point higher (.271) on the 10-point customer satisfaction question. The use of  $X_3$  increased the prediction only slightly. From an explanatory perspective, though, we now know that large firms enjoy higher customer satisfaction.

This example illustrates the manner in which the researcher can add nonmetric variables to the metric variables in the regression variate and improve both prediction and explanation.

## A MANAGERIAL OVERVIEW OF THE RESULTS

The regression results, including the evaluation of the confirmatory model, estimation of the additional models using summated scales as substitutes for the individual variables, and the addition of the nonmetric variable, all assist in addressing the basic research question: What affects customer satisfaction? In formulating a response, the researcher must consider two aspects: prediction and explanation.

**Table 5.19** Multiple Regression Results Adding  $X_3$  (Firm Size) as an Independent Variable by Using a Dummy Variable

Stepwise Regression with Transformed Variables	
Multiple R	.895
Coefficient of Determination ( $R^2$ )	.801
Adjusted $R^2$	.788
Standard error of the estimate	.548

#### Analysis of Variance

	Sum of Squares	df	Mean Square	F	Sig.
Regression	112.669	6	18.778	62.464	.000
Residual	27.958	93	.301		
Total	140.628	99			

#### Variables Entered into the Regression Model

Variables Entered	Regression Coefficients			Statistical Significance		Correlations			Collinearity Statistics		
	B	Std. Error	Beta	t	Sig.	Zero- order	Partial	Part	Tolerance	VIF	
	(Constant)	-1.250	.492		-2.542	.013					
$X_9$	Complaint Resolution	.300	.060	.304	4.994	.000	.603	.460	.231	.576	1.736
$X_6$	Product Quality	.365	.046	.427	7.881	.000	.486	.633	.364	.727	1.375
$X_{12}$	Salesforce Image	.701	.093	.631	7.507	.000	.500	.614	.347	.303	3.304
$X_7$	E-Commerce	-.333	.135	-.196	-2.473	.015	.283	-.248	-.114	.341	2.935
$X_{11}$	Product Line	.203	.061	.224	3.323	.001	.551	.326	.154	.469	2.130
$X_3$	Firm Size	.271	.123	.114	2.207	.030	.229	.223	.102	.798	1.253

In terms of prediction, the regression models all achieve high levels of predictive accuracy. The amount of variance explained equals about 80 percent and the expected error rate for any prediction at the 95 percent confidence level is about 1.1 points. In this type of research setting, these levels, augmented by the results supporting model validity, provide the highest levels of assurance as to the quality and accuracy of the regression models as the basis for developing business strategies.

In terms of explanation, all of the estimated models arrived at essentially the same results: three strong influences ( $X_{12}$ , Salesforce Image;  $X_6$ , Product Quality; and  $X_9$ , Complaint Resolution). Increases in any of these variables result in increases in customer satisfaction. For example, an increase of 1 point in the customer's perception of Salesforce Image ( $X_{12}$ ) will result in an average increase of at least seven-tenths (.701) of a point on the 10-point customer satisfaction scale. Similar results are seen for the other variables. Moreover, at least one firm characteristic, firm size, demonstrated a significant effect on customer satisfaction. Larger firms have levels of satisfaction about a quarter of a point (.271) higher than the smaller firms. These results provide management with a framework for developing strategies for improving customer satisfaction. Actions directed toward increasing the perceptions of HBAT can be justified in light of the corresponding increases in customer satisfaction.

The impact of two other variables ( $X_7$ , E-Commerce;  $X_{11}$ , Product Line) on customer satisfaction is less certain. Even though these two variables were included in the stepwise solution, their combined explained variance was only .038 out of an overall model  $R^2$  of .791. Both variables were not significant in the confirmatory model. Moreover,  $X_7$  had the reversed sign in the stepwise model, which, although due to multicollinearity, still represents a result contrary to managerial strategy development. As a result, the researcher should consider

reducing the influence allotted to these variables and even possibly omit them from consideration as influences on customer satisfaction.

In developing any conclusions or business plans from these results, the researcher should also note that the three major influences ( $X_{12}$ ,  $X_6$ , and  $X_9$ ) are primary components of the perceptual dimensions identified through exploratory factor analysis in Chapter 3. When these dimensions are represented by summated scales, three of the four dimensions are shown to be significantly related to customer satisfaction. These dimensions, which represent broad measures of customers' perceptions of HBAT, should also be considered in any conclusions. To state that only these three specific variables are influences on customer satisfaction would be a serious mis-statement of the more complex patterns of collinearity among variables. Thus, these variables are better viewed as representatives of the perceptual dimensions, with the other variables in each dimension also considered in any conclusions drawn from these results.

Management now has an objective analysis that confirms not only the specific influences of key variables, but also the perceptual dimensions that must be considered in any form of business planning regarding strategies aimed at affecting customer satisfaction.

This chapter provided an overview of the fundamental concepts underlying multiple regression analysis. Multiple regression analysis can describe the relationships among two or more intervally-scaled variables and is much more powerful than simple regression with a single independent variable. This chapter helps you to do the following:

**Determine when regression analysis is the appropriate statistical tool in analyzing a problem.** Multiple regression analysis can be used to analyze the relationship between a single dependent (criterion) variable and several independent (predictor) variables. The objective of multiple regression analysis is to use the several independent variables whose values are known to predict the single dependent value. Multiple regression is a dependence technique. To use it you must be able to divide the variables into dependent and independent variables, and both the dependent and independent variables are metric. Under certain circumstances, it is possible to include nonmetric data either as independent variables (by transforming either ordinal or nominal data with dummy variable coding) or the dependent variable (by the use of a binary measure in the specialized technique of logistic regression). Thus, to apply multiple regression analysis: (1) the data must be metric or appropriately transformed, and (2) before deriving the regression equation, the researcher must decide which variable is to be dependent and which remaining variables will be independent.

**Understand how regression helps us make both predictions and explanations using the least squares concept.** One objective of regression analysis is to predict a single dependent variable from the knowledge of one or more independent variables. Before estimating the regression equation, we must calculate the baseline against which we will compare the predictive ability of our regression models. The baseline should represent our best prediction without the use of any independent variables. In regression, the baseline predictor is the simple mean of the dependent variable. Because the mean will not perfectly predict each value of the dependent variable, we must have a way to assess predictive accuracy that can be used with both the baseline prediction and the regression models we create. The customary way to assess the accuracy of any prediction is to examine the errors in predicting the dependent variable. Although we might expect to obtain a useful measure of prediction accuracy by simply adding the errors, this approach is not possible because the errors from using the mean value always sum to zero. To overcome this problem, we square each error and add the results together. This total, referred to as the sum of squared errors (SSE), provides a measure of prediction accuracy that will vary according to the amount of prediction errors. The objective is to obtain the smallest possible sum of squared errors as our measure of prediction accuracy. Hence, the concept of least squares enables us to achieve the highest accuracy possible.

Another objective of regression analysis is explanation where the variate and the individual values contained in the variate are examined for the relationship to the dependent measure. Regression weights or coefficients for each independent variable provide a formal basis for assessing the change in the dependent measure for each one unit change in the independent variable. While the regression coefficients are “scaled” by the nature of each independent variable, other measures, such as standardized coefficients, provide some basis for assessing the relative impact of each independent variable. A series of other measures can assess a variable’s impact in terms of its unique impact and that shared with other variables due to multicollinearity among variables.

**Use dummy variables with an understanding of their interpretation.** A common situation faced by researchers is the desire to utilize nonmetric independent variables. Many multivariate techniques assume metric measurement for both independent and dependent variables. When the dependent variable is measured as a dichotomous (0, 1) variable, either discriminant analysis or a specialized form of regression (logistic regression), is appropriate. When the independent variables are nonmetric and have two or more categories, we can create dummy variables that act as replacement independent variables. Each dummy variable represents one category of a nonmetric independent variable, and any nonmetric variable with  $k$  categories can be represented as  $k - 1$  dummy variables. Thus, nonmetric variables can be converted to a metric format for use in most multivariate techniques.

**Be aware of the assumptions underlying regression analysis and how to assess them.** Improvements in predicting the dependent variable are possible by adding independent variables and even transforming them to represent nonlinear relationships. To do so, we must make several assumptions about the relationships between the dependent and independent variables that affect the statistical procedure (least squares) used for multiple regression. The basic issue is to know whether in the course of calculating the regression coefficients and predicting the dependent variable the assumptions of regression analysis have been met. We must know whether the errors in prediction are the result of the absence of a relationship among the variables or caused by some characteristics of the data not accommodated by the regression model. The assumptions to be examined include linearity of the phenomenon measured, constant variance of the error terms, independence of the error terms, and normality of the error term distribution. The assumptions underlying multiple regression analysis apply both to the individual variables (dependent and independent) and to the relationship as a whole. Once the variate has been derived, it acts collectively in predicting the dependent variable, which necessitates assessing the assumptions not only for individual variables, but also for the variate. The principal measure of prediction error for the variate is the residual—the difference between the observed and predicted values for the dependent variable. Plotting the residuals versus the independent or predicted variables is a basic method of identifying assumption violations for the overall relationship.

**Understand the implications of managing the variate and its implications on the regression results.** Managing the variate involves two decisions that the researcher can employ to achieve the objectives of explanation and/or prediction. The first decision area is variable specification, where the researcher must make the fundamental choice between using the individual variables or some form of composite, such as factor scores or summated scales, based on exploratory factor analysis. While this decision may be driven just by the sheer numbers of variables encountered in some situations today, it primarily reflects the researcher’s choice for dealing with multicollinearity among the independent variables. Using the individual variables reflects a choice to deal with multicollinearity in other ways (e.g., variable selection approaches) to achieve the most detailed effects from the independent variables. A choice for composite measures greatest detail reflects the decision to minimize multicollinearity effects, but also rely in the interpretation stage on making assessments as to the meaning of the composites versus the independent variables.

The second decision involves variable selection for model estimation. In the confirmatory approach the researcher strictly controls the variables to be included in the estimated model. In a software-controlled approach the researcher uses algorithms to determine which independent variables are ultimately included in the regression model. This decision is related to how the researcher elected to approach variable specification since the use of composite measures reduces the need for software-controlled approach.

The authors strongly recommend that researchers carefully consider both decisions as they are quite impactful in the resulting model. Researcher may wish to compare various combinations of these decisions, as was done when

comparing regression models using individual variables to those using composite measures. In any instance, the researcher should always understand the implications for decisions in each of these areas.

**Address the implications of user- versus software-controlled variable selection and explain the options available in software-controlled variable selection.** In managing the variate, researcher must choose how variables are to be included in the final model. One choice is for the researcher to specify the exact set of independent variables such that the regression model is essentially used in a confirmatory approach. This approach, referred to as *simultaneous regression*, includes all the variables at the same time. The researcher may also utilize a combinatorial approach -- a generalized search process across all possible combinations of independent variables. The best-known procedure is all-possible-subsets regression, which is exactly as the name suggests. All possible combinations of the independent variables are examined, and the best-fitting set of variables is identified.

The other option is to employ some form of software-controlled process, where the researcher may use the estimation technique to “pick and choose” among the set of independent variables with either sequential or constrained processes. The most popular sequential search method is stepwise estimation, which enables the researcher to examine the contribution of each independent variable to the regression model. There are also constrained approaches which place some constraints on the estimated coefficients and LASSO will even reduce coefficients to zero to effectively eliminate some variables from the analysis. All of these options are designed to assist the researcher in finding the best regression model using different approaches.

**Interpret the results of regression and variable importance, especially in light of multicollinearity.** The regression variate must be interpreted by evaluating the estimated regression coefficients for their explanation of the dependent variable. The researcher must evaluate not only the regression model that was estimated, but also the potential independent variables that were omitted if a sequential search or combinatorial approach was employed. In those approaches, multicollinearity may substantially affect the variables ultimately included in the regression variate. The researcher can start this process by examining the bivariate correlation that are independent of the estimated model as the fundamental relationships that should be reflected in the model results. Then using the model results, the researcher can assess the estimated coefficients which provides some perspective in the impact of each variable. The estimated regression coefficients, or beta coefficients, represent both the type of relationship (positive or negative) and the strength of the relationship between independent and dependent variables in the regression variate. The sign of the coefficient denotes whether the relationship is positive or negative, whereas the value of the coefficient indicates the change in the dependent value each time the independent variable changes by one unit. Additional measures of variable importance extend the analysis of impact to assessing not only the unique impact which is reflected in the regression coefficients, but also the shared impact among independent variables that is due to multicollinearity. These measures, such as commonality analysis, dominance analysis and relative weights, all attempt to reflect a variable's total impact. If some form of software-controlled procedures is used, the researcher must also evaluate the potential impact of omitted variables to ensure that the managerial significance is evaluated along with statistical significance.

Prediction is an integral element in regression analysis, both in the estimation process as well as in forecasting situations. Regression involves the use of a variate to estimate a single value for the dependent variable. This process is used not only to calculate the predicted values in the estimation procedure, but also with additional samples for validation or forecasting purposes. When prediction is the primary focus, understanding variable importance many times becomes less important.

**Apply the diagnostic procedures necessary to assess influential observations.** Influential observations include all observations that have a disproportionate effect on the regression results. The three basic types of influentials are as follows: (1) Outliers—observations that have large residual values and can be identified only with respect to a specific regression model (2) Leverage points—observations that are distinct from the remaining observations based on their independent variable values, and (3) Influential observations—all observations that have a disproportionate effect on the regression results. Influentials, outliers, and leverage points are based on one of four conditions: (1) *An error in observations or data entry:* Remedy by correcting the data or deleting the case, (2) *A valid but exceptional observation that is explainable by an extraordinary situation:* Remedy by deletion of the case unless variables reflecting

the extraordinary situation are included in the regression equation., (3) *An exceptional observation with no likely explanation*: Presents a special problem because the researcher has no reason for deleting the case, but its inclusion cannot be justified either, suggesting analyses with and without the observations to make a complete assessment, and (4) *An ordinary observation in its individual characteristics but exceptional in its combination of characteristics*: Indicates modifications to the conceptual basis of the regression model and should be retained. The researcher may delete truly exceptional observations but avoid deleting observations that, although different, are representative of the population. When deleting influential observations, always test the sensitivity of the results by estimating the model by samples with and without the influential observations.

**Understand the benefits gained from the extended forms of regression, namely multilevel models and panel models.** While multiple regression can handle a wide array of research questions, two extensions of regression have emerged to more easily address two specifics of research problems. The first extension is multilevel models, developed to more easily handle hierarchical or nested data structures. Hierarchical data structures reflect common “grouping” effects that result from contextual factors faced in all facets of behavior. The impact of the contextual factor is to create dependencies among those impacted by the factor, such that they cannot be considered statistically independent as required by one of the fundamental assumptions of regression. The researcher can attempt to address these contextual effects within regression analysis, but multilevel analysis makes this much easier and provides a framework for assessing a range of potential effects while also accounting for the dependencies among observations within the groups. Thus, whether analyzing students grouped within classrooms, employees under a specific supervisor or residents in a particular neighborhood, multilevel analysis provides a coordinated analysis of the unique effects from the individuals as well as the impacts due to the contextual factors (e.g., classrooms, supervisors or neighborhoods).

The second extension, panel models, shares many of the technical aspects of multilevel models, but is oriented to another specific research situation—cross-sectional analyses of longitudinal or time-series data. It provides a means of accommodating a large cross-sectional data structure while also integrating a time-series component not generally addressed in regression analysis. Quite applicable to many forms of longitudinal secondary data (e.g., firm or brand sales over time, economic performance of countries or even individual's repeated decisions), panel models have developed a sophisticated analytical approach to addressing many of the statistical challenges faced in these types of research situations.

This chapter provides a fundamental presentation of how regression works and what it can achieve. Familiarity with the concepts presented will provide a foundation for regression analyses the researcher might undertake and help you to better understand the more complex and detailed technical presentations in other textbooks on this topic.

How might regression analysis be performed differently when facing a research question using Big Data?

How do the objectives of explanation and prediction differ as well as overlap?

How does sample size impact statistical power and generalizability?

What options do the creation of additional variables provide in addition to data transformations?

What are mediation and moderation? What characteristics do they share and how do they differ?

Why is it important to examine the assumptions of linearity and homoscedasticity when using

regression? What are the potential remedies for violations of each?

Could you find a regression equation that would be acceptable as statistically significant and yet offer no acceptable interpretational value to management? What would be the underlying reason and how might it be explained?

Are influential cases always omitted? Give examples of occasions when they should or should not be omitted.

What is multicollinearity? Is it “Good” or “Bad” when found within the variables in a regression variate.? How should it be addressed?

How would you explain the relative importance of the independent variables used in a regression equation?

Differentiate between using information directly from the analysis as well as additional measures of variable importance.

What is the difference in interpretation between regression coefficients associated with interval-scale independent variables and dummy-coded (0, 1) independent variables?

What are the differences between interaction effects and correlated independent variables? Do any of these differences affect your interpretation of the regression equation?

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 Aguinis H., J. C. Beaty, R. J. Boik, and C. A. Pierce. 2005. Effect Size and Power in Assessing Moderating Effects of Categorical Variables Using Multiple Regression: A 30-Year Review. *Journal of Applied Psychology* 90: 94–107.
- 2 Aguinis, H., R. K. Gottfredson, and H. Joo. 2013. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods* 16: 270–301.
- 3 Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19: 716–23.
- 4 Alkharusi, H. 2012. Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding. *International Journal of Education* 4: 202–10.
- 5 Baltagi, B. 2008. *Econometric Analysis of Panel Data*. New York: Wiley.
- 6 Banerjee, S., B. P. Carlin, and A. E. Gelfan. 2014. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: CRC Press.
- 7 Barcikowski, R. S. 1981. Statistical Power with Group Mean as the Unit of Analysis. *Journal of Educational Statistics* 6: 267–85.
- 8 Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data*, 3rd edn. New York: Wiley.
- 9 Beck, Nathaniel. 2001. Time-Series–Cross-Section Data: What Have We Learned in the Past Few Years? *Annual Review of Political Science* 4: 271–93.
- 10 Beckstead, J. W. 2012. Isolating and Examining Sources of Suppression and Multicollinearity in Multiple Linear Regression. *Multivariate Behavioral Research* 47: 224–46.
- 11 Bell, A., and K. Jones. 2015. Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods* 3: 133–53.
- 12 Belsley, D. A., E. Kuh, and R. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
- 13 Blalock, H. M. 1984. Contextual-Effects Models: Theoretical and Methodological Issues. *Annual Review of Sociology* 10: 353–72.
- 14 Bliese, P. D., and P. J. Hanges. 2004. Being Both Too Liberal and Too Conservative: The Perils of Treating Grouped Data as Though They Were Independent. *Organizational Research Methods* 7: 400–17.
- 15 BMDP Statistical Software, Inc. 1991. *SOLO Power Analysis*. Los Angeles: BMDP.
- 16 Bock, R. D. (ed.). 2014. *Multilevel Analysis of Educational Data*. Amsterdam: Elsevier.
- 17 Bollen, K. A., and J. E. Brand. 2010. A General Panel Model with Random and Fixed Effects: A Structural Equations Approach. *Social Forces* 89: 1–34.
- 18 Box, G. E. P., and D. R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society B* 26: 211–43.
- 19 Boyd, L. H., and G. R. Iversen. 1979. *Contextual Analysis: Concepts and Statistical Techniques*. Belmont, CA: Wadsworth.
- 20 Bryk, A. S., and S. W. Raudenbush. 1988. Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model. *American Journal of Education* 97: 65–108.
- 21 Bryk, A. S., and S. W. Raudenbush. 1989. Methodology for Cross-Level Organizational Research. In S. B. Bacharach (ed.), *Research in the Sociology of Organizations*, Vol. 1, Greenwich, CT: JAI Press, pp. 233–73.
- 22 Bryk, A. S., and S. W. Raudenbush. 1992. *Hierarchical Linear Models*. Newbury Park, CA: Sage.
- 23 Budescu D. V. 1993. Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression. *Psychological Bulletin* 114: 542–51.
- 24 Budescu D. V., and R. Azen. 2004. Beyond Global Measures of Relative Importance: Insights from Dominance Analysis. *Organizational Research Methods* 7: 341–50.
- 25 Capraro, R. M. and M. M. Capraro, 2001. Commonality Analysis: Understanding Variance Contributions to Overall Canonical Correlation Effects of Attitude Toward Mathematics on Geometry Achievement. *Multiple Linear Regression Viewpoints* 27: 16–23.
- 26 Charlton, C., J. Rasbash, W. J. Browne, M. Healy, and B. Cameron. 2017. *MLwiN Version 3.00*. Centre for Multilevel Modelling, University of Bristol.

- 27 Chen, C. 2002. Robust Regression and Outlier Detection with the ROBUSTREG Procedure. In *Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*, Paper 265-27.
- 28 Clark, T. S., and D. A. Linzer. 2015. Should I Use Fixed or Random Effects? *Political Science Research and Methods* 3: 399–408.
- 29 Cohen, J., Stephen G. West, Leona Aiken, and P. Cohen. 2002. *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences*, 3rd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 30 Cortina J. M. 1993. Interaction, Nonlinearity, and Multicollinearity: Implications for Multiple Regression. *Journal of Management* 19: 915–22.
- 31 Cortina J. M., and R. S. Landis. 2009. When Small Effect Sizes Tell a Big Story, and When Large Effect Sizes Don't. In C. E. Lance and R. J. Vandenberg (eds.), *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences*. New York: Routledge, pp. 287–308.
- 32 Courville, T., and B. Thompson. 2001. Use of Structure Coefficients in Published Multiple Regression Articles:  $\beta$  is Not Enough. *Educational and Psychological Measurement* 61, 229–48.
- 33 Daniel, C., and F. S. Wood. 1999. *Fitting Equations to Data*, 2nd edn. New York: Wiley-Interscience.
- 34 De Leeuw, J., and I. Kreft. 1986. Random Coefficient Models for Multilevel Analysis. *Journal of Educational Statistics* 11: 57–85.
- 35 Deadrick, D. L., N. Bennett, and C. J. Russell. 1997. Using Hierarchical Linear Modeling to Examine Dynamic Performance Criteria Over Time. *Journal of Management* 23: 745–57.
- 36 Diez-Roux, A. V. 1998. Bringing Context Back into Epidemiology: Variables and Fallacies in Multilevel Analysis. *American Journal of Public Health* 88: 216–22.
- 37 DiPrete, T. A., and J. D. Forristal. 1994. Multilevel Models: Methods and Substance. *Annual Review of Sociology* 20: 331–57.
- 38 Duncan, C., K. Jones, and G. Moon. 1998. Context, Composition and Heterogeneity: Using Multilevel Models in Health Research. *Social Science and Medicine* 46: 97–117.
- 39 Edwards, J. R., and L. S. Lambert. 2007. Methods for Integrating Moderation and Mediation: A General Analytical Framework Using Moderated Path Analysis. *Psychological Methods* 12: 1–22.
- 40 Elhorst, J. P. 2014. Spatial Panel Models. In *Handbook of Regional Science*, Berlin: Springer, pp. 1637–52.
- 41 Finch, H., and J. Bolin. 2017. *Multilevel Modeling Using Mplus*. Boca Raton, FL: CRC Press.
- 42 Finch, W. H., J. E. Bolin, and K. Kelley. 2014. *Multilevel Modeling Using R*. Boca Raton, FL: CRC Press.
- 43 Ganzach, Y. 1997. Misleading Interaction and Curvilinear Terms. *Psychological Methods* 2: 235–47.
- 44 Ganzach, Y. 1998. Nonlinearity, Multicollinearity and the Probability of Type II Error in Detecting Interaction. *Journal of Management* 24: 615–22.
- 45 Gelman, A., and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Vol. 1. New York: Cambridge University Press.
- 46 Goldstein, H. 1995. *Multilevel Statistical Models*. New York: Halsted Press.
- 47 Goldstein, H. 2011. *Multilevel Statistical Models*, Wiley Series in Probability and Statistics, Vol. 922. New York: Wiley.
- 48 Goldstein, H., J. Rasbash, M. Yang, G. Woodhouse, H. Pan, D. Nuttal, and S. Thomas. 1993. A Multilevel Analysis of School Examination Results. *Oxford Review of Education* 19: 425–33.
- 49 Greene, W. H. 2018. *Econometric Analysis*, 8th edn. Harlow: Pearson.
- 50 Griffin, M. A. 1997. Interaction Between Individuals and Situations; Using HLM Procedures to Estimate Reciprocal Relationships. *Journal of Management* 23: 759–73.
- 51 Gunasekara, F. I., K. Richardson, K. Carter, and T. Blakely. 2013. Fixed Effects Analysis of Repeated Measures Data. *International Journal of Epidemiology* 43: 264–69.
- 52 Halaby, C. N. 2004. Panel Models in Sociological Research: Theory into Practice. *Annual Review of Sociology* 30: 507–44.
- 53 Hausman, Jerry A. 1978. Specification Tests in Econometrics. *Econometrica* 46:1251–71.
- 54 Hayes, A. F. 2015. An Index and Test of Linear Moderated Mediation. *Multivariate Behavioral Research* 50: 1–22.
- 55 Hayes, A. F., and J. Matthes. 2009. Computational Procedures for Probing Interactions in OLS and Logistic Regression: SPSS and SAS Implementations. *Behavior Research Methods* 41: 924–36.
- 56 Hayes, A. F., and L. Cai. 2007. Using Heteroskedasticity-Consistent Standard Error Estimators in OLS Regression: An Introduction and Software Implementation. *Behavior Research Methods* 39: 709–22.
- 57 Heck, R. and S. Thomas. 2000. *An Introduction to Multilevel Modeling Techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 58 Heck, R. H., S. L. Thomas, and L. N. Tabata. 2013. *Multilevel and Longitudinal Modeling with IBM SPSS*. New York: Routledge.
- 59 Hoerl, A. E., and Kennard, R. W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12: 55–67.
- 60 Hofmann, D. A. 1997. An Overview of the Logic and Rationale of Hierarchical Linear Models. *Journal of Management* 23: 723–44.
- 61 Hox, J. 2002. *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 62 Hox, J. J., M. Moerbeek, and R. van de Schoot. 2017. *Multilevel Analysis: Techniques and Applications*, 3rd edn. New York: Routledge.
- 63 Jaccard, J., R. Tursi, and C. K. Wan. 2003. *Interaction Effects in Multiple Regression*, 2nd edn. Beverly Hills, CA: Sage.
- 64 Johnson, J. W. 2000. A Heuristic Method for Estimating the Relative Weight Of Predictor Variables in Multiple Regression. *Multivariate Behavioral Research* 35: 1–19.

- 65 Johnson, J. W. 2004. Factors Affecting Relative Weights: The Influence of Sampling and Measurement Error. *Organizational Research Methods* 7: 283–99.
- 66 Johnson, J. W., and J. M. LeBreton. 2004. History and Use of Relative Importance Indices in Organizational Research. *Organizational Research Methods* 7: 238–57.
- 67 Johnson, R. A., and D. W. Wichern. 2002. *Applied Multivariate Statistical Analysis*, 5th edn. Upper Saddle River, NJ: Prentice Hall.
- 68 Kish, L. 1965. *Survey Sampling*. New York: Wiley.
- 69 Kraha, A., H. Turner, K. Nimon, L. R. Zientek, and R. K. Henson. 2012. Tools to Support Interpreting Multiple Regression in the Face of Multicollinearity. *Frontiers in Psychology* 3: 1–16.
- 70 Kreft, I. G., I. Kreft, and J. de Leeuw. 1998. *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage.
- 71 Lazarsfeld, P. F., and H. Menzel. 1961. On the Relation Between Individual and Collective Properties. In P. Lazarsfeld and H. Menzel (eds.), *Complex Organizations: A Sociological Reader*, New York: Holt, Rinehart and Winston, pp. 422–40.
- 72 LeBreton, J. M., and S. Tonidandel. 2008. Multivariate Relative Importance: Extending Relative Weight Analysis to Multivariate Criterion Spaces. *Journal of Applied Psychology* 93: 329–45.
- 73 LeBreton, J. M., S. Tonidandel, and D. V. Krasikova. 2013. Residualized Relative Importance Analysis: A Technique for the Comprehensive Decomposition of Variance in Higher Order Regression Models. *Organizational Research Methods* 16: 449–73.
- 74 Liao, H., and A. Chuang. 2004. A Multilevel Investigation of Factors Influencing Employee Service Performance and Customer Outcomes. *Academy of Management Journal* 47: 41–58.
- 75 Long, J. S., and L. H. Ervin. 2000. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *American Statistician* 54: 217–24.
- 76 Longford, N. T. 1993. *Random Coefficient Models*. New York: Oxford University Press.
- 77 Maas, C. J., and J. J. Hox. 2005. Sufficient Sample Sizes for Multilevel Modeling. *Methodology* 1: 86–92.
- 78 MacKinnon, D. P., J. L. Krull, and C. M. Lockwood. 2000. Equivalence of the Mediation, Confounding and Suppression Effect. *Prevention Science* 1: 173–81.
- 79 MacKinnon, J. G., and H. White. 1985. Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics* 29: 305–25.
- 80 Mason, C. H., and W. D. Perreault, Jr. 1991. Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research* 28: 268–80.
- 81 Mason, W. M., G. Y. Wong, and B. Entwistle. 1983. Contextual analysis through the multilevel linear model. In S. Leinhardt (ed.), *Sociological Methodology*, San Francisco, CA: Jossey-Bass, pp. 72–103.
- 82 Mood, A. M. 1971. Partitioning Variance in Multiple Regression Analyses as a tool for Developing Learning Models. *American Educational Research Journal* 8: 191–202.
- 83 Mosteller, F., and J. W. Tukey. 1977. *Data Analysis and Regression*. Reading, MA: Addison-Wesley.
- 84 Nathans, L. L., F. L. Oswald, and K. Nimon. 2012. Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. *Practical Assessment, Research and Evaluation*, 17: 1–19.
- 85 Neter, J., M. H. Kutner, W. Wassermann, and C. J. Nachtsheim. 1996. *Applied Linear Regression Models*, 3rd edn. Homewood, IL: Irwin.
- 86 Newton, R. G. and Spurrell, D. 1967. A Development of Multiple Regression for the Analysis of Routine Data. *Applied Statistics* 16: 51–64.
- 87 Nimon, K. F., and F. L. Oswald. 2013. Understanding the Results of Multiple Linear Regression: Beyond Standardized Regression Coefficients. *Organizational Research Methods* 16: 650–74.
- 88 O'Brien, R. M. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality and Quantity* 41: 673–90.
- 89 Paterson, L., and H. Goldstein. 1991. New Statistical Methods for Analysing Social Structures: An Introduction to Multilevel Models. *British Educational Research Journal* 17: 387–93.
- 90 Pickett, K. E., and M. Pearl. 2001. Multilevel Analyses of Neighbourhood Socioeconomic Context and Health Outcomes: A Critical Review. *Journal of Epidemiology and Community Health* 55: 111–22.
- 91 Rabe-Hesketh, S., and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.
- 92 Raudenbush, S. W. 1988. Educational Applications of Hierarchical Linear Models: A Review. *Journal of Educational Statistics* 13: 85–116.
- 93 Raudenbush, S. W., A. S. Bryk, and R. Congdon. 2013. *HLM 7.01 for Windows*. Skokie, IL: Scientific Software International, Inc.
- 94 Ray-Mukherjee, J., K. Nimon, S. Mukherjee, D. W. Morris, R. Slotow, and M. Hamer. 2014. Using Commonality Analysis in Multiple Regressions: A Tool to Decompose Regression Effects in the Face of Multicollinearity. *Methods in Ecology and Evolution* 5: 320–8.
- 95 Reichwein Zientek, L., and B. Thompson. 2006. Commonality Analysis: Partitioning Variance to Facilitate Better Understanding of Data. *Journal of Early Intervention* 28: 299–307.
- 96 Rice, N., and Leyland, A. 1996. Multilevel Models: Applications to Health Data. *Journal of Health Services Research* 1: 154–64.
- 97 Robinson, W. S. 1950. Ecological Correlations and the Behavior of Individuals. *American Sociological Review* 15: 351–7.
- 98 Rosopa, P. J., M. M. Schaffer, and A. N. Schroeder. 2013. Managing Heteroscedasticity in General Linear Models. *Psychological Methods* 18: 335–51.

- 99 Rousseeuw, P. J., and A. M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: Wiley.
- 100 Rousseeuw, P. J., and A. M. Leroy. 2003. *Robust Regression and Outlier Detection*. New York: Wiley.
- 101 Rousseeuw, P. J., and A. M. Leroy. 2005. *Robust Regression and Outlier Detection*, Wiley Series in Probability and Statistics Vol. 589. New York: Wiley.
- 102 Sampson, R. J., S. W. Raudenbush, and F. Earls. 1997. Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy. *Science* 277(5328): 918–24.
- 103 Schoen, J. L., J. A. DeSimone, and L. R. James. 2011. Exploring Joint Variance Between Independent Variables and a Criterion: Meaning, Effect, and Size. *Organizational Research Methods* 14: 674–95.
- 104 Schwarz, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics* 6: 461–4.
- 105 Seber, G. A. F. 2004. *Multivariate Observations*. New York: Wiley.
- 106 Sharma, S., R. M. Durand, and O. Gur-Arie. 1981. Identification and Analysis of Moderator Variables. *Journal of Marketing Research* 18: 291–300.
- 107 Singer, J. D. 1987. An Intraclass Correlation Model for Analyzing Multilevel Data. *Journal of Experimental Education* 55: 219–28.
- 108 Snijders, T., and R. J. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd edn. Thousand Oaks, CA: Sage.
- 109 Snijders, Tom A. B. 2005. Power and Sample Size in Multilevel Linear Models. In B. S. Everitt and D. C. Howell (eds.), *Encyclopedia of Statistics in Behavioral Science, Volume 3*, Chichester: Wiley, pp. 1570–73.
- 110 Soyer, E., and R. M. Hogarth. 2012. The Illusion of Predictability: How Regression Statistics Mislead Experts. *International Journal of Forecasting* 28: 695–711.
- 111 Thomas, D. R., B. D. Zumbo, E. Kwan, and L. Schweitzer. 2014. On Johnson's (2000) Relative Weights Method for Assessing Variable Importance: A Reanalysis. *Multivariate Behavioral Research* 49: 329–38.
- 112 Thomas, S. L., and R. H. Heck. 2001. Analysis of Large-Scale Secondary Data in Higher Education Research: Potential Perils Associated with Complex Sampling Designs. *Research in Higher Education* 42: 517–40.
- 113 Thompson, D. M., D. H. Fernald, and J. W. Mold. 2012. Intraclass Correlation Coefficients Typical of Cluster-Randomized Studies: Estimates from the Robert Wood Johnson Prescription for Health Projects. *Annals of Family Medicine* 10: 235–40.
- 114 Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B* 58: 267–88.
- 115 Tibshirani, R. 2011. Regression Shrinkage and Selection via the Lasso: A Retrospective. *Journal of the Royal Statistical Society Series B* 73: 273–82.
- 116 Tonidandel, S., and J. M. LeBreton. 2010. Determining the Relative Importance of Predictors in Logistic Regression: An Extension of Relative Weights Analysis. *Organizational Research Methods* 13: 767–81.
- 117 Tonidandel, S., and J. M. LeBreton. 2011. Relative Importance Analysis: A Useful Supplement to Regression Analysis. *Journal of Business and Psychology* 26: 1–9.
- 118 Tonidandel, S., and J. M. LeBreton. 2015. RWA web: A Free, Comprehensive, Web-Based, and User-Friendly Tool for Relative Weight Analyses. *Journal of Business and Psychology* 30: 207–16.
- 119 Tonidandel, S., J. M. LeBreton, and J. W. Johnson. 2009. Determining the Statistical Significance of Relative Weights. *Psychological Methods* 14: 387–99.
- 120 Varian, H. R. 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives* 28: 3–27.
- 121 Vrieze, S. I. 2012. Model Selection and Psychological Theory: A Discussion of the Differences Between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods* 17: 228–43.
- 122 Wang, J., H. Xie, and J. F. Fisher. 2011. *Multilevel Models: Applications Using SAS*. Berlin: Walter de Gruyter.
- 123 Weisberg, S. 1985. *Applied Linear Regression*, 2nd edn. New York: Wiley.
- 124 White, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48: 817–38.
- 125 Wilkinson, L. 1975. Tests of Significance in Stepwise Regression. *Psychological Bulletin* 86: 168–74.
- 126 Wilson, Sven E., and Daniel M. Butler. 2007. A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications. *Political Analysis* 15: 101–23.
- 127 Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
- 128 Yoo, W., R. Mayberry, S. Bae, K. Singh, K., Q. (Peter) He, and J. W. Lillard. 2014. A Study of Effects of Multi-Collinearity in the Multivariable Analysis. *International Journal of Applied Science and Technology* 4: 9–19.
- 129 Zhang, C. H., and J. Huang. 2008. The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression. *Annals of Statistics* 1567–94.

# 6

# MANOVA: Extending ANOVA

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

Explain the difference between the univariate null hypothesis of ANOVA and the multivariate null hypothesis of MANOVA.

Discuss the advantages of a multivariate approach to significance testing compared to the more traditional univariate approaches.

State the assumptions for the use of MANOVA.

Discuss the different types of test statistics that are available for significance testing in MANOVA.

Describe the purpose of post hoc tests in ANOVA and MANOVA.

Interpret interaction results when more than one independent variable is used in MANOVA.

Describe the purpose of multivariate analysis of covariance (MANCOVA).

Describe the uses and insights gained from incorporating mediation and moderation effects.

Identify situations in which causal inferences may be strengthened even in non-randomized studies.

## Chapter Preview

Multivariate analysis of variance (MANOVA) is an extension of analysis of variance (ANOVA) to accommodate more than one dependent variable. It is a dependence technique that measures the differences for two or more metric dependent variables based on a set of categorical (nonmetric) variables acting as independent variables. ANOVA and MANOVA can be stated in the following general forms:

$$\begin{array}{c} \text{Analysis of Variance} \\ Y_1 = X_1 + X_2 + X_3 + \cdots + X_n \\ (\text{metric}) \qquad \qquad (\text{nonmetric}) \end{array}$$

$$\begin{array}{c} \text{Multivariate Analysis of Variance} \\ Y_1 + Y_2 + Y_3 + \cdots + Y_n = X_1 + X_2 + X_3 + \cdots + X_n \\ (\text{metric}) \qquad \qquad (\text{nonmetric}) \end{array}$$

Like ANOVA, MANOVA is concerned with differences between groups (or experimental treatments). ANOVA is termed a *univariate procedure* because we use it to assess group differences on a single metric dependent variable. MANOVA is termed a *multivariate procedure* because we use it to assess group differences across multiple metric

dependent variables simultaneously. In MANOVA, each treatment group is observed on two or more dependent variables.

The concept of multivariate analysis of variance was introduced more than 70 years ago by Wilks [125]. However, it was not until the development of appropriate test statistics with tabled distributions and the more recent widespread availability of software programs to compute these statistics that MANOVA became a practical tool for researchers.

Both ANOVA and MANOVA are particularly useful when used in conjunction with experimental designs. That is, research designs in which the researcher directly controls or manipulates one or more independent variables to determine the effect on the dependent variable(s). ANOVA and MANOVA provide the tools necessary to judge the observed effects (i.e., whether an observed difference is due to a treatment effect or to random sampling variability). However, MANOVA has a role in non-experimental designs (e.g., survey research) where groups of interest (e.g., gender, purchaser/non-purchaser) are defined and then the differences on any number of metric variables (e.g., attitudes, satisfaction, purchase rates) are assessed for statistical significance.

## Key Terms

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

**Alpha ( $\alpha$ )** Significance level associated with the statistical testing of the differences between two or more groups. Typically, small values, such as .05 or .01, are specified to minimize the possibility of making a *Type I error*.

**Analysis of variance (ANOVA)** Statistical technique used to determine whether samples from two or more groups come from populations with equal means (i.e., Do the group means differ significantly?). Analysis of variance examines one dependent measure, whereas multivariate analysis of variance compares group differences on two or more dependent variables.

**A priori test** See *planned comparison*.

**Assumption of the strongly ignorable treatment assignment** A condition for making causal inferences, it requires that all confounding be eliminated from the treatment and control groups. Random assignment controls for this assumption and propensity score models attempt to adjust the treatment and control groups so that they meet it as well.

**ATE** Average treatment effect of the treatment represents the expected difference in the outcome between treatment and control groups.

**ATT** Average treatment effect for the treated represents the difference between those who were treated versus those that were not. The ATT does not equal the ATE.

**Balance** Comparison on a *covariate/confound* between the *treatment* and *control* groups. Balance is achieved when there is no statistical significance between the two groups on a covariate/confound, with each tested separately for balance.

**Balanced design** Experimental design with an equal number of respondents in each cell. Also see *unbalanced design*.

**Beta ( $\beta$ )** See *Type II error*.

**Blocking factor** Characteristic of respondents in the ANOVA or MANOVA that is used to reduce within-group variability by becoming an additional *factor* in the analysis. Most often used as a control variable (i.e., a characteristic not included in the analysis but one for which differences are expected or proposed). By including the blocking factor in the analysis, additional groups are formed that are more homogeneous and increase the chance of showing significant differences on the *main effect* of interest. As an example, assume that customers were asked about their buying intentions for a product and the independent measure used was age. Prior experience showed that substantial variation in buying intentions for other products of this type was also due to gender. Then gender could be added as a further factor so that each age category was split into male and female groups with greater within-group homogeneity. Used to address *nuisance factors*.

**Bonferroni inequality** Approach for adjusting the selected *alpha* level to control for the overall *Type I error* rate when performing a series of separate tests. The procedure involves calculating a new *critical value* by dividing the proposed *alpha* rate by the number of statistical tests to be performed. For example, if a .05 *significance level* is desired for a series of five separate tests, then a rate of .01 (.05 ÷ 5) is used in each separate test.

**Box's M test** Statistical test for the equality of the variance–covariance matrices of the dependent variables across the groups. It is especially sensitive to the presence of non-normal variables. Use of a conservative *significance level* (i.e., .01 or less) is suggested as an adjustment for the sensitivity of the statistic.

**Causal effects** The difference (*main effect*) between *treatment* and *control* groups can be attributed solely to the treatment effect, with no confounds or other influences. See *causal inference*.

**Causal inference** The process of estimating *causal effects*, whether by *randomized experiments* or other means (e.g., *propensity score models*).

**Complete mediation** When the *indirect effect of mediation* “replaces” the original *main effect* such that the mediated main effect is nonsignificant.

**Confound** An omitted variable from the analysis which is related to both *factor/treatment* and *outcome* such that it acts as another “cause” for the *main effect*. Its omission calls into question whether the treatment → outcome relationship is valid or if the confound is the true explanation.

**Contrast** Procedure for investigating specific group differences of interest in conjunction with ANOVA and MANOVA (e.g., comparing group mean differences for a specified pair of groups).

**Control group** Group of respondents identical to the *treatment* group except that they did not receive the treatment.

**Controlled experiment** *Experiment* in which respondents are randomly assigned to *treatment* or *control group* to ensure that *causal inferences* can be made.

**Counterfactual** When participants are exposed to either the *treatment* or *control*, they obviously cannot be exposed to both. The counterfactual represents the outcome of each participant that is “missing” (e.g., the outcome of a participant in the treatment group would have had if they were in the control group or vice versa).

**Covariates, or covariate analysis** Use of regression-like procedures to remove extraneous (nuisance) variation in the dependent variables due to one or more uncontrolled metric independent variables (covariates). The covariates are assumed to be linearly related to the dependent variables. After adjusting for the influence of covariates, a standard ANOVA or MANOVA is carried out. This adjustment process (known as ANCOVA or MANCOVA) usually allows for more sensitive tests of treatment effects.

**Critical value** Value of a test statistic (*t* test, *F* test) that denotes a specified *significance level*. For example, 1.96 denotes a .05 significance level for the *t* test with large sample sizes.

**Cross-sectional study** See *observational study*.

**DAG (directed acyclic graphs)** Conceptual framework for identifying *causal effects* among the presence of *confounds*, both measured and unmeasured. The DAG is a graphical representation of the network of effects that impact the causal effect of interest.

**Discriminant function** Dimension of difference or discrimination between the groups in the MANOVA analysis. The discriminant function is a *variate* of the dependent variables.

**Disordinal interaction** Form of *interaction effect* among independent variables that invalidates interpretation of the *main effects* of the treatments. A disordinal interaction is exhibited graphically by plotting the means for each group and having the lines intersect or cross. In this type of interaction the mean differences not only vary, given the unique combinations of independent variable levels, but the relative ordering of groups changes as well.

**Effect size** Standardized measure of group differences used in the calculation of statistical *power*. Calculated as the difference in the group means divided by the standard deviation, it is then comparable across research studies as a generalized measure of effect (i.e., differences in group means).

**Experiment** Research approach typically characterized by (a) random assignment of respondents to *treatment* and *control groups*, and (b) researcher manipulation of the *treatment*.

**Experimental design** Research plan in which the researcher directly manipulates or controls one or more independent variables (see *treatment* or *factor*) and assesses their effect on the dependent variables. Associated with the scientific method, it is gaining in popularity in business and the social sciences. For example, respondents are shown separate advertisements that vary systematically on a characteristic, such as different appeals (emotional versus rational) or types of presentation (color versus black-and-white) and are then asked their attitudes, evaluations, or feelings toward the different advertisements.

**Experimentwide error rate** The combined or overall error rate that results from performing multiple *t* tests or *F* tests that are related (e.g., *t* tests among a series of correlated variable pairs or a series of *t* tests among the pairs of categories in a multi-chotomous variable).

**External validity** Extent to which the results of an analysis can be generalized to other contexts (e.g., situations, populations).

**Factor** Nonmetric independent variable, also referred to as a *treatment* or experimental variable.

**Factorial design** Design with more than one *factor* (treatment). Factorial designs examine the effects of several factors simultaneously by forming groups based on all possible combinations of the levels (values) of the various treatment variables.

**Field experiment** Implementation of a *controlled experiment* in a “natural” setting (i.e., outside a controlled laboratory environment) in an attempt to increase *external validity*, but also raising threats to *internal validity*.

**General linear model (GLM)** Extension of the linear model that can accommodate multiple dependent and independent variables. As such, it can perform a wide range of multivariate techniques, including multiple regression, canonical correlation, ANOVA, MANOVA and discriminant analysis.

**Generalized linear model (GLZ)** Generalized estimation procedure based on three components: (1) a *variate* formed by the linear combination of independent variables, (2) a probability distribution specified by the researcher based on the characteristics of the dependent variables, and (3) a *link function* that denotes the connection between the variate and the probability distribution.

**Homogeneity of regression effect** Assumption of *covariance analysis* where *covariates* are assumed to have the same relationship (i.e., slope coefficient) with all dependent variables.

**Hotelling's  $T^2$**  Test to assess the statistical significance of the difference on the means of two or more variables between two groups. It is a special case of MANOVA used with two groups or levels of a treatment variable.

**Independence** Critical assumption of ANOVA or MANOVA that requires that the dependent measures for each respondent be totally uncorrelated with the responses from other respondents in the sample. A lack of independence severely affects the statistical validity of the analysis unless corrective action is taken.

**Indirect effect** The effect in *mediation* that “transmits” the main effect through the mediating variable and is formed by two interrelated relationships: (a) treatment → mediator, and (b) mediator → outcome.

**Instructional manipulation check** Questions inserted into a research design to assess if respondents are following instructions and attentive to the task. An example would be the question in which respondent is given the answer, so an incorrect answer would suggest inattention to questions.

**Instrumental variable** Variable related solely to the treatment that is used in observational studies to “control” for confounds and allowing for making causal inferences.

**Interaction effect** In *factorial designs*, the joint effects of two *treatment* variables in addition to the individual *main effects*. It means that the difference between groups on one treatment variable varies depending on the level of the second treatment variable. For example, assume that respondents were classified by income (three levels) and gender (males versus females). A significant interaction would be found when the differences between males and females on the independent variable(s) varied substantially across the three income levels.

**Internal validity** Degree to which the research design and analysis is correctly performed. Particularly important in *experimental research* as it deals with direct threats (e.g., *confounds*) to making *causal inferences*.

**Inverse probability of treatment weighting (IPTW)** Process of weighting individual observations to adjust for the covariate set rather than *matching* or *stratification*.

**Intervening effect** See *indirect effect*.

**Link function** A primary component of the *generalized linear model (GLZ)* that specifies the transformation between the variate of independent variables and the specified probability distribution. In MANOVA (and regression) the identity link can be used with a normal distribution, corresponding to our statistical assumptions of normality.

**Main effect** The individual effect of each *treatment* variable on the dependent variable(s).

**Manipulation check** Question(s) focused on ensuring that respondents correctly perceived or experienced the *treatment*. Not directed toward the outcome, but only used to assess if the treatment performs as expected.

**Matching** Method of creating comparable treatment and control groups by “*matching*” observations from one group with observations from the other group with equal or very similar propensity scores.

**Mediator** An intervening variable that attempts to explain “Why” a treatment → outcome relationship occurs. The mediator is assumed to be “caused” by the *treatment* and the mediator then “causes” the outcome in a causal chain. In this manner it becomes an alternative representation of the *main effect*.

**Moderator** The strength and/or direction of a main effect varies between different values of a third variable—the moderator.

**Model overlap** Comparison of the range of *propensity scores* between *treatment* and *control* groups.

**Multivariate normal distribution** Generalization of the univariate normal distribution to the case of  $p$  variables. A multivariate normal distribution of sample groups is a basic assumption required for the validity of the significance tests in MANOVA (see Chapter 2 for more discussion of this topic).

**Natural experiment** A non-random form of experimental research where the treatment occurs naturally (e.g., natural disaster, etc.).

**Nuisance factor** A characteristic of respondents or the research context which may increase the variability of the outcome, but has no relevance to the research question. Can be accounted for through use of a *blocking factor* or *covariate*.

**Null hypothesis** Hypothesis with samples that come from populations with equal means (i.e., the group means are equal) for either a dependent variable (univariate test) or a set of dependent variables (multivariate test). The null hypothesis can be accepted or rejected depending on the results of a test of statistical significance.

**Observational study** Non-random research design common to many research contexts. Primary limitations are that respondent selection is not controlled by the researcher and *confounds* are very difficult to identify and then account for in the analysis.

**Ordinal interaction** Acceptable type of *interaction effect* in which the magnitudes of differences between groups vary but the groups’ relative positions remain constant. This type of interaction is graphically represented by plotting mean values and observing non-parallel lines that do not intersect.

**Orthogonal** Statistical independence or absence of association. Orthogonal *variates* explain unique variance, with no variance explanation shared between them. Orthogonal *contrasts* are *planned comparisons* that are statistically independent and represent unique comparisons of group means.

**Outcome** Term used in *controlled experimental setting* to represent the dependent variable.

**Partial mediation** When the *indirect effect* of mediation is significant but does not completely “replace” the original main effect such that the mediated main effect is still significant as well. Also see *complete mediation*.

**Pillai's criterion** Test for multivariate differences similar to *Wilks' lambda*.

**Planned comparison** *A priori test* that tests a specific comparison of group mean differences. These tests are performed in conjunction with the tests for *main* and *interaction effects* by using a *contrast*.

**Post hoc test** Statistical test of mean differences performed after the statistical tests for *main effects* have been performed. Most often, post hoc tests do not use a single *contrast*, but instead test for differences among all possible combinations of groups. Even though they provide abundant diagnostic information, they do inflate the overall *Type I error* rate by performing multiple statistical tests and thus must use strict confidence levels.

**Potential outcomes** Framework for *causal inferences* developed by Rubin which provides an approach for achieving the comparability for estimating *causal effects* between *treatment* and *control* groups in *non-experimental settings* (e.g., observational data).

**Power** Probability of identifying a treatment effect when it actually exists in the sample. Power is defined as  $1 - \beta$  (see *beta*). Power is determined as a function of the statistical significance level ( $\alpha$ ) set by the researcher for a *Type I error*, the sample size used in the analysis, and the *effect size* being examined.

**PROCESS macro** Software available for IBM SPSS and SAS that provides template-driven models of mediation, moderation and even combinations of mediation/moderation.

**Propensity score** A single value that represents the set of covariates/*confounders* impacting the *causal effect*. Most often calculated as the predicted probability from a *propensity scoring model* and used to "match" observations from *treatment* and *control* groups.

**Propensity scoring model** Technique in which all potential confounders are combined into a single variate through a logistic regression model to generate a *propensity score*. The propensity score is then used to adjust *treatment* and *control* groups so that they are equivalent across all the confounders, meet the *strongly ignorable treatment assignment assumption* and allow for *causal inferences* to be made.

**Quasi-experiment** Similar to a controlled experiment except it lacks assignment to the two groups through randomization.

**Repeated measures** Use of two or more responses from a single individual in an *ANOVA* or *MANOVA* analysis. The purpose of a repeated measures design is to control for individual-level differences that may affect the within-group variance. Repeated measures represent a lack of *independence* that must be accounted for in a special manner in the analysis.

**Replication** Repeated administration of an experiment with the intent of validating the results in another sample of respondents.

**Roy's greatest characteristic root (gcr)** Statistic for testing the null hypothesis in *MANOVA*. It tests the first *discriminant function* of the dependent variables for its ability to discern group differences.

**Significance level** See *alpha*.

**Sobel test** A parametric statistical test for the significance of the *indirect effect in mediation*. Criticized for low power, researchers have suggested using bootstrapping as an alternative test of significance.

**Sphericity assumption** Found in repeated measures *ANOVA* designs, assumes that differences between all possible pairs of levels of the independent variable are equal. *MANOVA* can be used when violations occur.

**Stable unit treatment value assumption (SUTVA)** Assumption applicable to all causal inferences that the treatment assignment process does not impact the outcomes (i.e., one person's assignment does not impact another person's outcomes).

**Stepdown analysis** Test for the incremental discriminatory power of a dependent variable after the effects of other dependent variables have been taken into account. Similar to stepwise regression or discriminant analysis, this procedure, which relies on a specified order of entry, determines how much an additional dependent variable adds to the explanation of the differences between the groups in the *MANOVA* analysis.

**Stratification** Formation of subgroups of observations that have similar values on the *propensity scores*. The number of strata can vary based on the degree of equivalence desired in each stratum.

**t statistic** Test statistic that assesses the statistical significance between two groups on a single dependent variable (see *t test*).

**t test** Test to assess the statistical significance of the difference between two sample means for a single dependent variable. The *t test* is a special case of *ANOVA* for two groups or levels of a treatment variable.

**Treatment** Independent variable (*factor*) that a researcher manipulates to see the effect (if any) on the dependent variables. The treatment variable can have several levels. For example, different intensities of advertising appeals might be manipulated to see the effect on consumer believability.

**Type I error** Probability of rejecting the null hypothesis when it should be accepted, that is, concluding that two means are significantly different when in fact they are the same. Small values of *alpha* (e.g., .05 or .01), also denoted as  $\alpha$ , lead to rejection of the null hypothesis that population means are equal, and acceptance of the alternative hypothesis that population means are not equal.

**Type II error** Probability of failing to reject the null hypothesis when it should be rejected, that is, concluding that two means are not significantly different when in fact they are different. Also known as the *beta* ( $\beta$ ) *error*.

**Unbalanced design** When the cell sizes formed by the treatment(s) are unequal in size. Most easily occurs in noncontrolled settings such as *observational studies* or *natural experiments*.

**U statistic** See *Wilks' lambda*.

**Variate** Linear combination of variables. In MANOVA, the dependent variables are formed into variates in the discriminant function(s).

**Vector** Set of real numbers (e.g.,  $X_1, \dots, X_n$ ) that can be written in either columns or rows. Column vectors are considered conventional, and row vectors are considered transposed. Column vectors and row vectors are shown as follows:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad X^T = [X_1 \ X_2 \ \cdots \ X_n]$$

Column vector    Row vector

**Wilks' lambda** One of the four principal statistics for testing the null hypothesis in MANOVA. Also referred to as the maximum likelihood criterion or *U statistic*.

## Re-Emergence of Experimentation

Experimentation has long been associated with the physical sciences and is a foundational principle of the scientific method of research. Yet while the social sciences were a less frequent user of experimentation in past years, that trend has now been reversed. First, the traditional randomized laboratory experiment found its way into areas such as consumer behavior and behavioral economics, and recent years have seen an explosion in the use of field experiments in many disciplines [85, 17, 30]. Academic researchers are increasingly looking to all forms of experimental methods to address an ever-widening range of research questions.

Along with this increased use in academia, the business sector has embraced experimentation, particularly field experiments, in the era of Big Data [2]. The rise of massive datasets of customer information and the ready access to customer behavior data through digital means make applications of experimentation accessible to a wide range of firms. Examples include Capital One's use of tens of thousands of experiments a year or Facebook's embrace of experimentation [106]. Of particular focus has been customer-oriented studies given the access to customer data (e.g., the gaming industry [90]), but it applies to all business sectors, even public organizations [6]. See Chapter 1 for a broader overview of Big Data and the impact on all areas of analytics.

## Experimental Approaches Versus Other Multivariate Methods

ANOVA and MANOVA are perhaps most widely associated with experiments and their numerous variations across every discipline. The fundamental characteristic across all types of **experiments** is the treatment→outcome relationship, where a **factor/treatment(s)** is conceptualized to “cause” a specific outcome (i.e., cause-and-effect). We use the term “cause” in a rather general sense to establish at least an ordering to the variables in the relationship. There is a vast research tradition of **experimental design** and it is beyond the scope of this chapter to address all of its facets covered in a number of fields of research [25, 8, 104, 87]. We generally associate a “hypothesis” with this treatment→outcome relationship, although it may or may not conform to the principles of causation. But as we will discuss in a later section on causal inference, it is not the statistical techniques that confer causation, but the conceptual and theoretical development and specific research design that are necessary before causality can be assessed.

What distinguishes research with an experimental emphasis from many other multivariate techniques discussed in this text is the focus on a single or small number of “causes” to establish a very specific “effect.” Certainly the interdependence techniques (exploratory factor analysis and cluster analysis) have quite different objectives, but even multiple regression, discriminant analysis, logistic regression, and structural equation modeling, which all focus on prediction and explanation, rely on a much more complex variate with multiple independent variables.

There are methods such as sequential variable selection and variable importance measures to provide assistance in assessing the impact of the variate elements. But these methods are not employed in experimental designs since the focus is on a much smaller number of independent variables, many times trying to isolate the effect of a single variable.

## MANOVA: Extending Univariate Methods for Assessing Group Differences

Multivariate techniques are often extensions of univariate techniques, as in the case for multiple regression, which extended simple bivariate regression (with only one independent variable) to a multivariate analysis where two or more independent variables could be used. A similar situation is found in analyzing group differences. These procedures are classified as univariate not because of the number of independent variables (known as treatments or factors), but instead because of the number of dependent variables. In multiple regression, the terms *univariate* and *multivariate* refer to the number of independent variables, but for ANOVA and MANOVA the terminology applies to the use of single or multiple dependent variables.

The univariate techniques for analyzing group differences are the *t test* (two groups) and **analysis of variance (ANOVA)** for two or more groups. The multivariate equivalent procedures are the Hotelling's  $T^2$  and multivariate analysis of variance, respectively. The relationships between the univariate and multivariate procedures are shown in Figure 6.1.

The *t* test and Hotelling's  $T^2$  are portrayed as specialized cases in that they are limited to assessing only two groups (categories) for an independent variable. Both ANOVA and MANOVA can also handle the two group situations as well as address analyses where the independent variables have more than two groups. A review of both the *t* test and ANOVA are available in the Basic Stats appendix available in the online resources at the text's websites.

### MULTIVARIATE PROCEDURES FOR ASSESSING GROUP DIFFERENCES

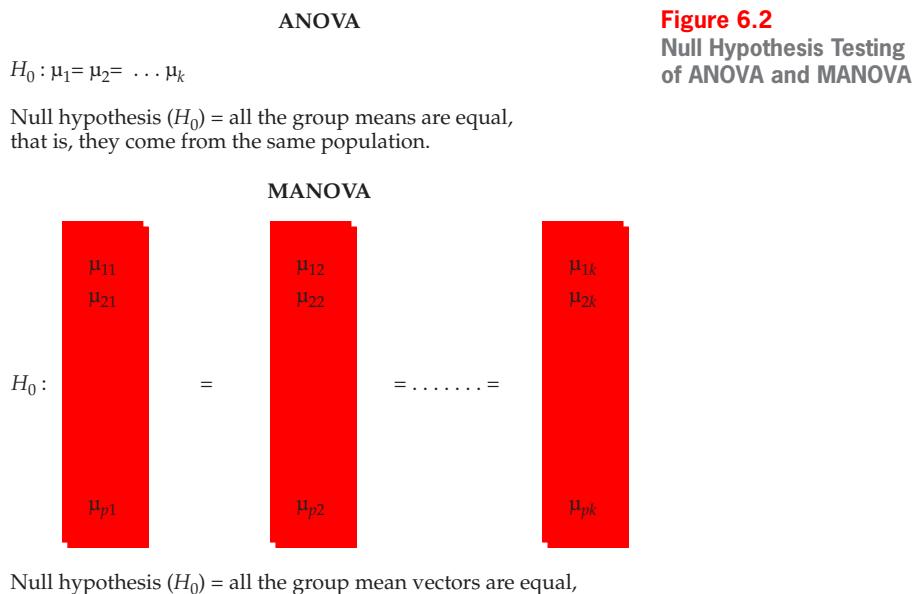
As statistical inference procedures, both the univariate techniques (*t* test and ANOVA) and their multivariate extensions (Hotelling's  $T^2$  and MANOVA) are used to assess the statistical significance of differences between groups. In the *t* test and ANOVA, the **null hypothesis** tested is the equality of a single dependent variable means across groups. In the multivariate techniques, the null hypothesis tested is the equality of **vectors** of means on multiple dependent variables across groups. The distinction between the hypotheses tested in ANOVA and MANOVA is illustrated in Figure 6.2. In the univariate case, a single dependent measure is tested for equality across the groups. In the multivariate case, a variate is tested for equality. The concept of a **variate** has been instrumental in our discussions of the previous multivariate techniques and is covered in detail in Chapter 1.

In MANOVA, the researcher actually has two variates – one for the dependent variables and another for the independent variables. The dependent variable variate is of more interest because the metric-dependent measures can be combined in a linear combination as we have already seen in multiple regression and discriminant analysis. The unique aspect of MANOVA is that the *variate for the dependent variables optimally combines the multiple dependent measures into a single value that maximizes the differences across groups*.

**Figure 6.1**

Extending the Univariate Methods (*t* Test and ANOVA) to the Multivariate Analysis

Number of Groups in Independent Variable	Number of Dependent Variables	
	One (Univariate)	Two or More (Multivariate)
Two Groups (Specialized Case)	<i>t</i> test	Hotelling's $T^2$
Two or More Groups (Generalized Case)	Analysis of variance (ANOVA)	Multivariate analysis of variance (MANOVA)



$\mu_{pk}$  = means of variable  $p$ , group  $k$

**Figure 6.2**  
Null Hypothesis Testing  
of ANOVA and MANOVA

**The Two-Group Case: Hotelling's  $T^2$**  Assume that researchers were interested in both the appeal and purchase intent generated by two advertising messages. If only univariate analyses were used, the researchers would perform separate  $t$  tests on the ratings of both the appeal of the messages and the purchase intent generated by the messages. Yet the two measures are interrelated; thus, what is really desired is a test of the differences between the messages on both variables collectively. Here is where **Hotelling's  $T^2$** , a specialized form of MANOVA and a direct extension of the univariate  $t$  test, can be used.

**CONTROLLING FOR TYPE I ERROR RATE** Hotelling's  $T^2$  provides a statistical test of the variate formed from the dependent variables, which produces the greatest group difference. It also addresses the problem of inflating the **Type I error** rate that arises when making a series of  $t$  tests of group means on several dependent measures. It controls this inflation of the Type I error rate by providing a single overall test of group differences across all dependent variables at a specified  $\alpha$  level.

How does Hotelling's  $T^2$  achieve these goals? Consider the following equation for a variate of the dependent variables:

$$C = W_1X_1 + W_2X_2 + \dots + W_nX_n$$

where:

$C$  = composite or variate score for a respondent

$W_i$  = weight for dependent variable  $i$

$X_i$  = dependent variable  $i$

In our example, the ratings of message appeal are combined with the purchase intentions to form the composite. For any set of weights, we could compute composite scores for each respondent and then calculate an ordinary  **$t$  statistic** for the difference between groups on the composite scores. However, if we can find a set of weights that gives the maximum value for the  $t$  statistic for this set of data, these weights would be the same as the discriminant function between the two groups (as shown in Chapter 7). The maximum  $t$  statistic that results from the composite scores produced by the discriminant function can be squared to produce the value of Hotelling's  $T^2$  [43].

The computational formula for Hotelling's  $T^2$  represents the results of mathematical derivations used to solve for a maximum  $t$  statistic (and, implicitly, the most discriminating linear combination of the dependent variables).

It is equivalent to saying that if we can find a discriminant function for the two groups that produces a significant  $T^2$ , the two groups are considered different across the mean vectors. Figure 6.3 is a graphical portrayal of the test for group differences for the  $t$  test and Hotelling's  $T^2$ . As we will discuss later, MANOVA is statistically equivalent to discriminant analysis and readers familiar with graphical portrayals of discriminant analysis will see the similarities.

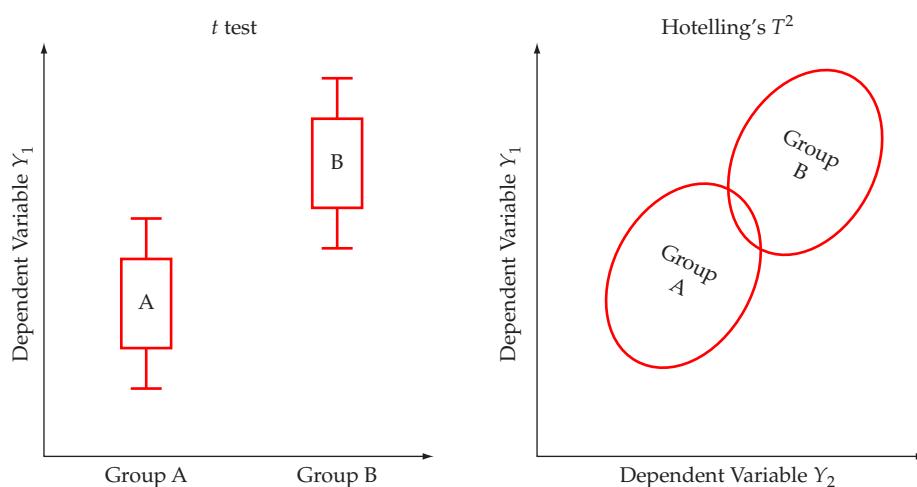
**STATISTICAL TESTING** How does Hotelling's  $T^2$  provide a test of the hypothesis of no group difference on the vectors of mean scores? Just as the  $t$  statistic follows a known distribution under the null hypothesis of no treatment effect on a single dependent variable, Hotelling's  $T^2$  follows a known distribution under the null hypothesis of no treatment effect on any of a set of dependent measures. This distribution turns out to be an  $F$  distribution with  $p$  and  $N_1 + N_2 - 2 - 1$  degrees of freedom after adjustment (where  $p$  = the number of dependent variables). To get the **critical value** for Hotelling's  $T^2$ , we find the tabled value for  $F_{\text{crit}}$  at a specified  $\alpha$  level and compute  $T_{\text{crit}}^2$  as follows:

$$T_{\text{crit}}^2 = \frac{p(N_1 + N_2 - 2)}{N_1 + N_2 - p - 1} \times F_{\text{crit}}$$

**The K-Group Case: MANOVA** Just as ANOVA was an extension of the  $t$  test, MANOVA can be considered an extension of Hotelling's  $T^2$  procedure. We estimate dependent variable weights to produce a variate score for each respondent that is maximally different across all of the groups. Many of the same analysis design issues discussed for ANOVA apply to MANOVA, but the method of statistical testing differs markedly from that of ANOVA.

**ANALYSIS DESIGN** All of the issues of analysis design applicable to ANOVA (number of levels per factor, number of factors, etc.) also apply to MANOVA. Moreover, the number of dependent variables and the relationships among these dependent measures raise additional issues that will be discussed later. MANOVA enables the researcher to assess the impact of multiple independent variables on not only the individual dependent variables, but on the dependent variables collectively as well.

**STATISTICAL TESTING** In the case of two groups, once the variate is formed, the procedures of ANOVA are basically used to identify whether differences exist. With three or more groups (either by having a single independent variable with three levels or by using two or more independent variables), the analysis of group differences becomes more closely allied to discriminant analysis (see Chapter 7). For three or more groups, just as in discriminant analysis, multiple variates of the dependent measures are formed. The first variate, termed a **discriminant function**, specifies a set of weights that maximize the differences between groups, thereby maximizing the  $F$  value. The maximum  $F$  value itself



**Figure 6.3**  
Extending the  $t$  Test to  
Hotelling's  $T^2$

Note: Letters A and B designate group means in  $t$  test and group centroids in Hotelling's  $T^2$

enables us to compute directly what is called **Roy's greatest characteristic root (gcr)** statistic, which allows for the statistical test of the first discriminant function. The greatest characteristic root statistic can be calculated as [43]:

$$\text{Roy's gcr} = (k - 1)F_{\max}/(N - k)$$

To obtain a single test of the hypothesis of no group differences on this first vector of mean scores, we could refer to tables of Roy's gcr distribution. Just as the  $F$  statistic follows a known distribution under the null hypothesis of equivalent group means on a single dependent variable, the gcr statistic follows a known distribution under the null hypothesis of equivalent group mean vectors (i.e., group means are equivalent on a set of dependent measures). A comparison of the observed gcr to Roy's  $\text{gcr}_{\text{crit}}$  gives us a basis for rejecting the overall null hypothesis of equivalent group mean vectors.

Any subsequent discriminant functions are **orthogonal**; they maximize the differences among groups based on the remaining variance not explained by the prior function(s). Discriminant functions are created for significance testing of each effect—for each independent variable (main effect) and each interaction created when two or more independent variables are included in the analysis. Thus, in some instances, the test for differences between groups involves not just the first variate score but also a set of variate scores that are evaluated simultaneously. In these cases, a set of multivariate tests is available (e.g., Wilks' lambda, Pillai's criterion), and each test is designed for specific situations of testing these multiple variates. Again, these are the same multivariate tests used in discriminant analysis.

**DIFFERENCES BETWEEN MANOVA AND DISCRIMINANT ANALYSIS** We noted earlier that in statistical testing MANOVA employs a discriminant function, which is the variate of dependent variables that maximizes the difference between groups. The question may arise: What is the difference between MANOVA and discriminant analysis? In some aspects, MANOVA and discriminant analysis are mirror images:

- The dependent variables in MANOVA (a set of metric variables) are the independent variables in discriminant analysis,
  - The single nonmetric dependent variable of discriminant analysis becomes the independent variable in MANOVA.
- Moreover, both use the same methods in forming the variates and assessing the statistical significance between groups [51]. The differences, however, center around the objectives of the analyses and the role of the nonmetric variable(s).
- Discriminant analysis employs a single nonmetric variable as the dependent variable. The categories of the dependent variable are assumed as given, and the independent variables are used to form variates that maximally differ between the groups formed by the dependent variable categories. Discriminant analysis does not have the ability to use more than one nonmetric variable as the dependent measure. Moreover, emphasis is placed on detecting which of the metric independent variables provide discrimination between the groups, which are assumed as fixed.
  - MANOVA uses the set of given metric variables as dependent variables and the objective becomes finding groups of respondents that exhibit differences on the set of dependent variables. The groups of respondents are not prespecified as they were in discriminant analysis. Instead, the researcher uses one or more independent variables (nonmetric variables) to form groups. MANOVA, even while forming these groups, still retains the ability to assess the impact of each nonmetric variable separately.

So we can see that while the two techniques use the same underlying statistical model, their focus differs. In discriminant analysis, the groups formed by the nonmetric dependent measure are assumed as given and interest is in the set of metric independent variables for their ability to discriminate among the groups. In MANOVA, however, the set of metric variables used as dependent measures are assumed given and interest is in finding the nonmetric variable(s) that form groups with the greatest differences on the set of dependent measures.

## A Hypothetical Illustration of MANOVA

A simple example can illustrate the benefits of using MANOVA while also illustrating the use of two independent variables to assess differences on two dependent variables.

Assume that HBAT's advertising agency identified two characteristics of HBAT's advertisements (product type being advertised and customer status), which they thought caused differences in how people evaluated the advertisements. They asked the research department to develop and execute a study to assess the impact of these characteristics on advertising evaluations.

### ANALYSIS DESIGN

In designing the study, the research team defined the following elements relating to factors used, the dependent variables, and sample size:

- *Factors.* Two factors (independent variables) were defined as representing Product Type and Customer Status. For each factor, two levels were also defined: product type (product 1 versus product 2) and customer status (current customer versus ex-customer). In combining these two variables, we get four distinct groups of respondents as shown in Figure 6.4.
- *Dependent variables.* Evaluation of the HBAT advertisements used two variables (ability to gain attention and persuasiveness) measured with a 10-point scale.
- *Sample.* Respondents were shown one of the hypothetical advertisements (one for each combination of ability to gain attention and persuasiveness) and asked to rate it on the two dependent measures (see Figure 6.4).

### DIFFERENCES FROM DISCRIMINANT ANALYSIS

Although MANOVA constructs the variate and analyzes differences in a manner similar to discriminant analysis, the two techniques differ markedly in how the groups are formed and analyzed. Let us use the following example to revisit and illustrate these differences:

- With discriminant analysis, we could only examine the differences between the set of four groups (forming a single nonmetric variable with levels for each combination of the two independent variables), without making the distinction as to a group's characteristics (product type or customer status). The researcher would be able to determine whether the variate significantly differed only across the groups, but could not assess which characteristics of the groups related to these differences.
- With MANOVA, however, the researcher analyzes the differences in the groups while also assessing whether the differences are due to product type, customer type, or both. Thus, MANOVA focuses the analysis on the composition of the groups based on their characteristics (the independent variables).

**Figure 6.4**

Analysis Design with Two Factors and Two Dependent Measures

		Factor 1: Product Type	
Factor 2:		Product 1	Product 2
Customer Status	Current Customer	Group 1	Group 3
	Ex-Customer	Group 2	Group 4
		DV: Attention	DV: Attention
		DV: Persuasiveness	DV: Persuasiveness
		DV: Attention	DV: Attention
		DV: Persuasiveness	DV: Persuasiveness

MANOVA enables the researcher to propose a more complex research design by using any number of independent nonmetric variables (within limits) to form groups and then look for significant differences in the dependent variable variate associated with specific nonmetric variables.

## FORMING THE VARIATE AND ASSESSING DIFFERENCES

With MANOVA we combine multiple dependent measures into a single variate that will then be assessed for differences across one or more independent variables. Let us see how a variate is formed and used in our example.

Assume for this example that the two dependent measures (attention and persuasion) were equally weighted when summed into the variate value (variave total = score<sub>ability to gain attention</sub> + score<sub>persuasiveness</sub>). This first step is similar to discriminant analysis (see example in Chapter 7) by providing a single composite value with the variables weighted to achieve maximum differences among the groups. Here the independent variables (product type and customer status) formed the groups (i.e., the dependent variable in discriminant analysis) and the two dependent measures were combined in the variate (i.e., like the independent variables in discriminant analysis). It is this process that provides the variate composed of the two dependent variables (attention and persuasion) that is compared across the four groups formed by the independent variables.

With the variate formed from the two dependent variables, we can now calculate variate means for each of the four groups (formed as the combination of the two independent variables) as well as the overall variate mean for each level. From Figure 6.5 we can draw several conclusions:

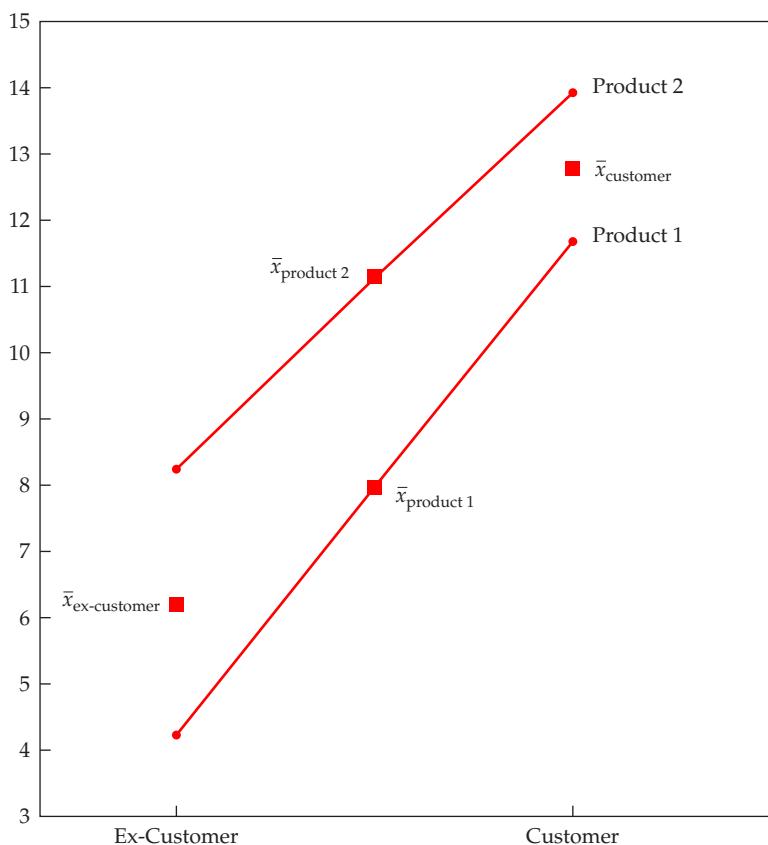
- *Overall group differences.* The four group means for the composite variable totals (i.e., 4.25, 8.25, 11.75, and 14.0 = Total column) vary significantly between each group, and are quite different from each other. MANOVA and discriminant analysis both have a statistical test of this conclusion. But this is where discriminant analysis is limited. Since discriminant analysis is limited to a single nonmetric dependent variable, we can only assess the differences in the composite variate and the two variables it represents across the four groups, but with no insight as to how the two variables that formed the four group variable (i.e., product type and customer status) individually contributed to these differences.
- *Group differences by independent variable.* MANOVA, however, goes beyond analyzing only the differences across groups by assessing whether product type and/or customer status created groups with these differences. In MANOVA product type and customer status are now the independent variables and are treated separately

**Figure 6.5**

Respondent Data for the Hypothetical Example of MANOVA

Customer Type/Product Line	Product 1				Product 2			
	ID	Attention	Purchase	Total	ID	Attention	Purchase	Total
Ex-customer	1	1	3	4	5	3	4	7
$\bar{x}_{\text{attention}} = 3.00$	2	2	1	4	6	4	3	7
$\bar{x}_{\text{purchase}} = 3.25$	3	2	3	5	7	4	5	9
$\bar{x}_{\text{total}} = 6.25$	4	3	2	5	8	5	5	10
Average		2.0	2.25	4.25		4.0	4.25	8.25
Customer	9	4	7	11	13	6	7	13
$\bar{x}_{\text{attention}} = 6.00$	10	5	6	11	14	7	8	15
$\bar{x}_{\text{purchase}} = 6.875$	11	5	7	12	15	7	7	14
$\bar{x}_{\text{total}} = 12.875$	12	6	7	13	16	8	6	14
Average		5.0	6.75	11.75		7.0	7.0	14.0

Values are responses on a 10-point scale (1 = Low, 10 = High).



**Figure 6.6**  
Graphical Display of Group  
Means of Variate (Total) for  
Hypothetical Example

rather than having to be combined into a single nonmetric dependent variable in discriminant analysis. Keeping them as separate independent variables allows for the evaluation of each variable separately (product 1 versus product 2 and customer versus ex-customer—denoted by the symbol ■), which are shown in Figure 6.6 along with the means of the four groups (the two lines connect the groups—ex-customer and customer—for product 1 and product 2). If we look at product type (ignoring distinctions as to customer status), we can see a mean value of 8.0 for users of product 1 versus a mean value of 11.125 for users of product 2. Likewise, for customer status, ex-customers had a mean value of 6.25 and customers a mean value of 12.875. Visual examination suggests that both category means show significant differences, with the differences for customer type ( $12.875 - 6.25 = 6.625$ ) greater than that for product ( $11.125 - 8.00 = 3.125$ ).

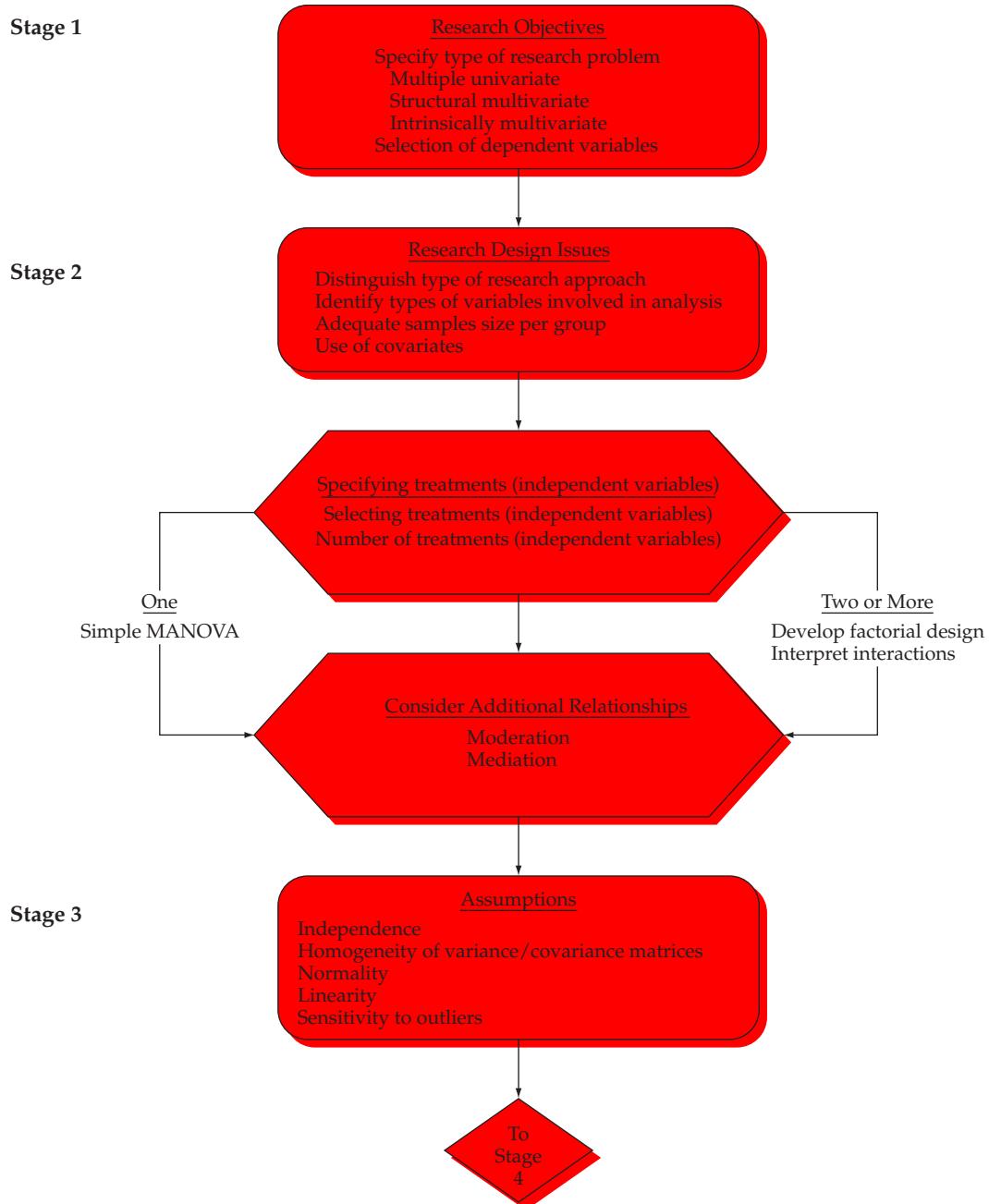
By being able to represent these independent variable category means separately in the analysis, MANOVA not only shows that overall differences between the four groups do occur (as was done with discriminant analysis), but also that both customer type and product type contribute significantly to forming these different groups. Therefore, both characteristics “cause” significant differences, a finding not possible with discriminant analysis.

## A Decision Process for MANOVA

The process of performing a multivariate analysis of variance is similar to that found in many other multivariate techniques, so it can be described through the six-stage model-building process described in Chapter 1. The process begins with the specification of research objectives. It then proceeds to a number of design issues facing a multivariate analysis and then an analysis of the assumptions underlying MANOVA. With these issues addressed,

the process continues with estimation of the MANOVA model and the assessment of overall model fit. When an acceptable MANOVA model is found, then the results can be interpreted in more detail. The final step involves efforts to validate the results to ensure generalizability to the population. Figure 6.7 (Stages 1–3) and Figure 6.10 (Stages 4–6, shown later in the text) provide a graphical portrayal of the process, which is discussed in detail in the following sections.

**Figure 6.7**  
Stages 1–3 in the Multivariate Analysis of Variance (MANOVA) Decision Diagram



## Stage 1: Objectives of MANOVA

The selection of MANOVA is based on the desire to analyze a dependence relationship represented as the differences in a set of dependent measures across a series of groups formed by one or more categorical independent measures. As such, MANOVA represents a powerful analytical tool suitable to a wide array of research questions. Whether used in actual or quasi-experimental situations (i.e., field settings or survey research for which the independent measures are categorical), MANOVA can provide insights into not only the nature and predictive power of the independent measures, but also the interrelationships and differences seen in the set of dependent measures.

### WHEN SHOULD WE USE MANOVA?

With the ability to examine several dependent measures simultaneously, the researcher can gain in several ways from the use of MANOVA. Here we discuss the issues in using MANOVA from the perspectives of controlling statistical accuracy and efficiency while still providing the appropriate forum for testing multivariate questions.

**Control of Experimentwide Error Rate** The use of separate univariate ANOVAs or *t* tests can create a problem when trying to control the **experimentwide error rate** [49]. For example, assume that we evaluate a series of five dependent variables by separate ANOVAs, each time using .05 as the significance level. Given no real differences in the dependent variables, we would expect to observe a significant effect on any given dependent variable five percent of the time. However, across our five separate tests, the probability of a Type I error lies somewhere between five percent, if all dependent variables are perfectly correlated, and 23 percent ( $1 - .95^5$ ), if all dependent variables are uncorrelated. Thus, a series of separate statistical tests leaves us without control of our effective overall or experimentwide Type I error rate (i.e., the probability of at least one false rejection of the hypothesis of no differences). If the researcher desires to maintain control over the experimentwide error rate and at least some degree of correlation is present among the dependent variables, then MANOVA is appropriate.

**Differences Among a Combination of Dependent Variables** A series of univariate ANOVA tests also ignores the possibility that some composite (linear combination) of the dependent variables may provide evidence of an overall group difference that may go undetected by examining each dependent variable separately. Individual tests ignore the correlations among the dependent variables, and in the presence of multicollinearity among the dependent variables, MANOVA will be more powerful than the separate univariate tests in several ways:

- MANOVA may detect *combined* differences not found in the univariate tests.
- If multiple variates are formed, then they may provide *dimensions* of differences that can distinguish among the groups better than single variables.
- If the number of dependent variables is kept relatively low (five or fewer), the statistical power of the MANOVA tests equals or exceeds that obtained with a single ANOVA [20].

The considerations involving sample size, number of dependent variables, and statistical power are discussed in a subsequent section.

### TYPES OF MULTIVARIATE QUESTIONS SUITABLE FOR MANOVA

The advantages of MANOVA versus a series of univariate ANOVAs extend past the statistical issues in its ability to provide a single method of testing a wide range of differing multivariate questions. Throughout the text, we emphasize the interdependent nature of multivariate analysis. MANOVA has the flexibility to enable the researcher to select the test statistics most appropriate for the question of concern. Hand and Taylor [42] have classified multivariate problems into three categories, each of which employs different aspects of MANOVA in its resolution. These three categories are multiple univariate, structured multivariate, and intrinsically multivariate questions.

**Multiple Univariate Questions** A researcher studying multiple univariate questions identifies a number of separate dependent variables (e.g., age, income, education of consumers) that are to be analyzed separately but needs some control over the experimentwide error rate. In this instance, MANOVA is used to assess whether an overall difference is found between groups, and then the separate univariate tests are employed to address the individual issues for each dependent variable.

**Structured Multivariate Questions** A researcher dealing with structured multivariate questions gathers two or more dependent measures that have specific relationships among them. A common situation in this category is repeated measures, where multiple responses are gathered from each subject, perhaps over time or in a pretest–posttest exposure to some stimulus, such as an advertisement. Here MANOVA provides a structured method for specifying the comparisons of group differences on a set of dependent measures while maintaining statistical efficiency.

**Intrinsically Multivariate Questions** An intrinsically multivariate question involves a set of dependent measures in which the principal concern is how they differ *as a whole* across the groups. Differences on individual dependent measures are of less interest than their collective effect. One example is the testing of multiple measures of response that should be consistent, such as attitudes, preference, and intention to purchase, all of which relate to differing advertising campaigns. The full power of MANOVA is utilized in this case by assessing not only the overall differences but also the differences among combinations of dependent measures that would not otherwise be apparent. This type of question is served well by MANOVA's ability to detect multivariate differences, even when no single univariate test shows differences.

## SELECTING THE DEPENDENT MEASURES

In identifying the questions appropriate for MANOVA, it is important also to discuss the development of the research question, specifically the selection of the dependent measures. A common problem encountered with MANOVA is the tendency of researchers to misuse one of its strengths—the ability to handle multiple dependent measures—by including variables without a sound conceptual or theoretical basis. The fundamental reason for conceptual support is that MANOVA cannot distinguish whether a dependent variable is suitable for inclusion, only if there are actual differences among groups. This is seen in both extremes when the results indicate that a subset of the dependent variables has the ability to influence interpretation of the overall differences among groups. If some of the dependent measures with the strong differences are not really appropriate for the research question, then “false” differences may lead the researcher to draw incorrect conclusions about the set as a whole.

Likewise, if the set of dependent variables have been selected only because of their differences across groups, and other dependent variables excluded because of a lack of differences, then the research question is also confounded. The researcher should always scrutinize the dependent measures and form a solid rationale for including them. Any ordering of the variables, such as possible sequential effects, should also be noted. MANOVA provides a special test, stepdown analysis, to assess the statistical differences in a sequential manner, much like the addition of variables to a regression analysis.

In addition to the set of dependent variables in the analysis, their level of correlation also must be considered. The recommended level of inter-correlation is between .4 and .6, with inter-correlations lower (or higher) creating problems in the analysis.

- If the inter-correlations are very high ( $>.7$  or  $.8$ ), they have a tendency to create redundancies and reduce the statistical efficiency. This may become especially impactful if multiple composite variates (i.e., discriminant functions) are formed when analyzing two or more independent variables. In such a situation, removing some of the highly correlated items or forming composite measures is suggested.
- If the correlations are very low ( $<.3$ ), the analyst should consider running separate ANOVAs and adjusting for the experiment-wide error (e.g., Bonferroni adjustment). This is due to a reduced power of the multivariate test versus the univariate tests.

## Decision Processes for MANOVA

MANOVA is an extension of ANOVA that examines the effect of one or more nonmetric independent variables on two or more metric dependent variables.

The focus of MANOVA and ANOVA is typically on a single or very small number of independent variables (i.e., a specific treatment→outcome relationship) with all other variables impacting the relationship accounted for in some manner.

In addition to the ability to analyze multiple dependent variables, MANOVA also has the advantages of:

- Controlling the experiment-wide error rate when some degree of intercorrelation among dependent variables is present.

- Providing more statistical power than ANOVA when the number of dependent variables is five or fewer.

- Nonmetric independent variables create groups between which the dependent variables are compared; many times the groups represent experimental variables or “treatment effects.”

- Researchers should include only dependent variables that have strong theoretical support.

In summary, the researcher should assess all aspects of the research question carefully, both in conceptual and statistical terms, thereby ensuring that MANOVA is applied in the correct and most powerful way. The following sections address many issues that have an impact on the validity and accuracy of MANOVA. But it is ultimately the responsibility of the researcher to employ the technique properly.

## Stage 2: Issues in the Research Design of MANOVA

MANOVA follows all of the basic design principles of ANOVA, yet in some instances the multivariate nature of the dependent measures requires a unique perspective. In the following section we will review the basic design principles and illustrate those unique issues arising in a MANOVA analysis. As noted earlier, ANOVA and MANOVA have long been associated with experimentation, but are commonly used in non-experimental situations as well. *The distinguishing characteristic in all of these applications is the interest in a single or small number of effects of the independent variables or treatments.* As such, even in non-experimental settings, we think in terms of experimental design principles (e.g., treatments and outcomes). And as we discuss in a later section, techniques are emerging to bring non-experimental data into much closer alignment with the experimental approach, thereby enabling causal inferences to be made.

We should note that we provide only an overview of the broader principles of experimental design and the various approaches available to the researcher. Whether it be the type of experimental approach (e.g., controlled experiment versus other methods) or the various types of effects that can be accounted for in the research design and analysis plan, interested researchers are strongly encouraged to review more complete discussions of these topics.

## TYPES OF RESEARCH APPROACHES

The techniques of ANOVA and MANOVA are applicable in a wide range of research settings. While they are not totally mutually exclusive, we can identify several distinct types of situations in which the principles of experimentation are applied.

**Experimental: Randomization of groups** A key element of any experimental approach is the use of randomization in assigning participants into the experimental groups. Randomization provides comparability between the experimental groups on all variables except the treatment. The correct use of randomization is the strongest element in any claim of a research design to assess causality of the treatment→outcome relationship. There are two experimental approaches that include randomization.

**CONTROLLED EXPERIMENT** A **controlled experiment** is the “typical” type of experiment with two groups—one group of respondents exposed to the treatment of interest (the independent variable) as well as a **control group**, a group of respondents that were not exposed to the treatment. Both groups participate in the experiment under identical conditions which attempt to eliminate any external influences as well. Causation, as we will discuss in more detail in the later section on causal inference, can be inferred if we can be assured that the two groups do not differ in any way except for exposure to the treatment. With equivalence of the two groups, then we can draw the conclusion that any differences between the two groups is due only to the treatment since the two groups are equal on all other characteristics, both observed and unobserved. Randomization is used to establish this equivalence between the two groups by randomly assigning respondents to either the treatment or control group. The important characteristic is that there is no systematic or non-random effect on how the two groups are formed. The presence of a systematic effect, as we will see in other forms of experiments, can introduce differences into the groups and bias the group differences, thus distorting the causal effect.

**FIELD EXPERIMENT** While controlled experiments provide a setting in which causation is most easily supported, a primary criticism is that they lack external validity—they do not represent realistic conditions for the basic relationship to operate (e.g., experiments about consumer behavior conducted in a lab, voter turnout in an election or leadership behaviors by supervisors). The **field experiment** is an attempt to conduct the controlled experiment in a more realistic or natural context rather than the strictly controlled setting. The problem that the more realistic setting also increases the chance for possible exposure to external factors and thus contamination of the causal effect. Even with these caveats, this form of experimentation is gaining widespread use across multiple disciplines [56, 57, 58].

**Experimental: Non-randomization** While randomization is always desired, there are numerous situations in which random assignment is not possible, either due to the nature of the research question (e.g., impact of smoking on health) or data collection (e.g., data already available in non-experimental form). The lack of randomization is a threat to **internal validity** (i.e., the equivalence of the two groups) and thus may introduce bias into any differences between the groups. The researcher is still interested in analyses using the experimental approach, but cannot directly make causal claims. Attempts to control the possible effects of a non-randomized design are discussed in a later section.

**QUASI-EXPERIMENT** A **quasi-experiment** is similar to a controlled experiment except it does not employ randomization for the assignment process to the two groups. Instead, the researcher explicitly controls the assignment process. While attempting to control for comparability of the groups in making the treatment assignment, it cannot be assured to be random and thus also has the problems of equivalence and comparability.

**NATURAL EXPERIMENT** In some situations the assignment to groups is determined by nature (i.e., weather-related event exposes some individuals to an event) or by other factors which appear random, thus the term **natural experiment**. But even though exposure to the treatment may seem random, the natural experiment does not provide a means for assessing causal inference as the researchers cannot be assured of the equivalence of the groups. These types of situations are more like observational studies and really differ only in that the exposure to the treatment occurs outside the control of the researcher, thus making it appear to be random.

### **Non-Experimental**

**OBSERVATIONAL STUDY** Many times, an **observational study**, also known as a **cross-sectional study**, is analyzed with techniques associated with the experimental approach (e.g., ANOVA or MANOVA). But the use of these techniques does not qualify this approach as an experimental study nor can any causal inference be made directly. Yet many times data is already available or it is not possible to assign groups to treatment versus control (i.e., ethically we can't assign some respondents to be smokers and others not). With no ability to apply randomization to control for factors creating biases in the estimated effects, this approach has little basis for making causal inferences and the researcher addresses the research questions using the natural group differences in the sample. Recent advances, however, have developed a range of statistical methods that utilize the respondents from observational data to form groups comparable to random assignment. These methods will be discussed briefly in a later section on making causal inferences.

**Selecting the Research Approach** As seen in our discussion, there is a broad range of approaches available to researchers today, each with its own specific advantages and disadvantages. Many times the research context dictates the approach that is even feasible. It is beyond the scope of this text to provide a full discussion of the trade-offs inherent in each, but researchers should be pleased that all of these approaches have research frameworks that allow for the critical examination of research hypotheses. Even the observational study, with the development of methods for causal inference (e.g., propensity scores, inverse probability weighting and instrumental variables), can now address questions originally thought the purview of only the controlled experiment. We discuss these approaches in a subsequent section.

## TYPES OF VARIABLES IN EXPERIMENTAL RESEARCH

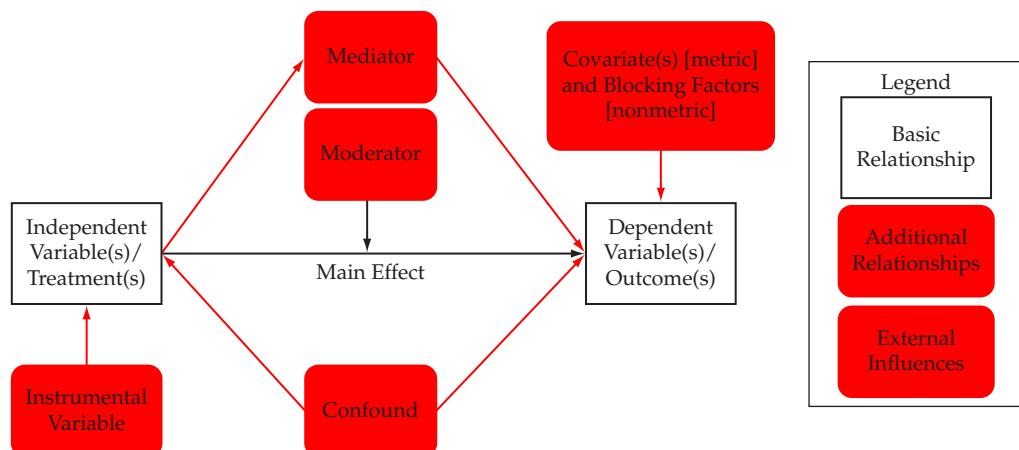
Given the wide range of research contexts for ANOVA/MANOVA, the basic elements have taken many labels across contexts and disciplines. Figure 6.8 provides an overview of many of the types of variables and relationships encountered in these types of research contexts. For purposes of simplification, we have divided them into three classes: basic relationships, extended relationships and external influences.

**Basic Relationship** The basic relationship is the primary interest of the research—i.e., the impact of the independent variable/treatment in “causing” the dependent variable/outcome. The independent variable acts like its counterparts in other techniques—the reason/“cause” for the associated outcomes. In experimental research the independent variable is known as a **treatment** or **factor** and is manipulated by the researcher. As noted earlier, the independent variable/treatment is always a nonmetric variable, such that it relates in many instances to discrete actions (e.g., one group given a drug treatment and another a placebo). While many times we think of experiments as having only a single treatment, there may be multiple treatments in an experimental design. Yet most often the number is much more limited (two or three at the most) than other multivariate techniques (e.g., multiple regression) since a larger number of treatments creates a quite complicated experimental design.

The dependent variable is many times known as the **outcome** and represents the expected effect due to the different levels of the independent variable/treatment. While we cannot always attribute causality to an experimental approach, we do expect the impact of the treatment to be reflected in differences in the outcome variable.

These differences in the outcome between groups of respondents formed by the independent variable/treatment is the **main effect**—a measure of the impact of the independent variable on the outcome(s) and a test of the hypothesis of interest in the analysis. Thus, there is a separate main effect for each independent variable or treatment in the analysis.

**Figure 6.8**  
Variable Types and Relationships in ANOVA/MANOVA Models



For our purposes we will use the term independent variable and treatment somewhat interchangeably, although in practice the term independent variable is more associated with non-experimental approaches and treatments with those of an experimental focus (i.e., actual manipulation of the treatment). This distinction is more important in the interpretation of the results and whether or not we can attempt to attribute causality to the main effect. But the distinction does not substantially impact the basic elements of the analytical technique.

**Additional Relationships** The additional relationships are researcher-controlled additions to the basic model that either provide for greater validity (i.e., by controlling for additional variables) or examine relationships that supplement the main effect. These relationships all address additional aspects of the main effect and thus require the same strong conceptual support. We will discuss each briefly here and then elaborate on their use in a later section.

The two additional relationships that add to the explanation of the main effect are mediation and moderation. A **mediator** is a third variable that acts as a conduit for the main effect, either partially or in full. In this regard, the mediator provides a “why” explanation of the main effect. A **moderator** changes the main effect across values of the moderator (i.e., interacts with treatment variable). In this case, the moderator details “how” the main effect works in different contexts (i.e., the values of the moderator). In a simple sense it is a test of generalizability—does the main effect change in strength and/or direction when viewed in different situations.

The final types of additional relationships to be considered are covariates and blocking factors. **Covariates** are metric variables that are related to the outcome, but not to the treatment, and are used to control for external factors. As will be discussed later, this independence from the treatment is an important feature of the covariate, otherwise it will diminish the main effect. A **blocking factor** is a nonmetric variable, just like a treatment, that is included solely to reduce the unexplained variation in the outcome and “strengthen” the main effect. We are not interested in the blocking factor per se, but add it to the analysis to reduce “noise” that weakens the main effect.

**External Influences** The final two variables reflect relationships that impact the ability to generate causal inferences from the analysis. A **confound** is a variable, which if omitted, “confounds” the main effect because it is related to both treatment and outcome (i.e., it is another “cause” that could explain the main effect). The presence of a confound casts doubt on the observed main effect, either in part or in total. It is the primary “threat” to causal inference and is a concern for any experimental approach. As we will discuss in a later section, specialized models have been developed to address confounds specifically and hopefully account for their effects. Finally, an **instrumental variable** is a variable utilized solely to assist in causal inference in the presence of confound variables. Related only to the treatment variable, an instrumental variable is many times difficult to identify since they must be exogenous – causally unrelated to all variables in the model except the treatment variable.

It is important to understand the distinction between a mediator and a confound. They both “explain” the treatment effect and have the same statistical impact on the results. But a variable is termed a mediator or confound based on the conceptual model, not any statistical criteria. If it is intrinsic to the causal process (i.e., transmits the effect) then it is a mediator. If the variable is external to the causal process and acts as an additional “cause,” then it is a confound.

**Which Variables to Include in the Experimental Design** Obviously, every experimental approach will include the basic relationship (main effect), with one or more treatments and an outcome(s). The inclusion of additional relationships reflects the additional research questions (i.e., mediation and/or moderation) that are of interest along with those variables for which the researcher wishes to control for (covariates and/or blocking factors). Every experiment must consider the potential impact of confounds as they can have a substantive impact on the basic relationship. The extent to which these additional relationships are required to meet the criteria for causality is primarily due to the research approach taken. Thus, all analyses must consider the types of variables to be included to accurately assess the research question and draw the desired inferences.

## SAMPLE SIZE REQUIREMENTS—OVERALL AND BY GROUP

MANOVA, like all of the other multivariate techniques, can be markedly affected by the sample size used. A number of basic issues arise concerning the sample sizes needed in MANOVA.

**Overall Sample Size** What differs most for MANOVA (and the other techniques assessing group differences such as the *t* test and ANOVA) is that the sample size requirements relate to individual group sizes and not the total sample per se. Yet while there is no required overall sample size, the number of respondents can start to increase dramatically as the number of treatments and number of levels of treatments increases. We will illustrate some of these considerations in the next section.

### Group/Cell Sample Size

**MINIMUM CELL SIZE** As a bare minimum, the sample in each cell (group) must be greater than the number of dependent variables. Although this concern may seem minor, the inclusion of just a small number of dependent variables (from 5 to 10) in the analysis may place a bothersome constraint on data collection. This problem is particularly prevalent in field experimentation or survey research, where the researcher has less control over the achieved sample.

**RECOMMENDED CELL SIZE** As a practical guide, a recommended minimum cell size is 20 (and preferably 30) observations [18]. Again, remember this quantity is per group, necessitating fairly large overall samples even for relatively simple analyses. In our earlier example of advertising messages, we had only two factors, each with two levels, but this analysis would require 80 observations for an adequate analysis. If the recommended minimum group sample size cannot be achieved, even small increases in cell size are beneficial. For example, for a three group design with seven respondents per cell and a moderate effect size (.50), statistical power is .50. Increasing the cell size to fourteen respondents increases the power to the desired level of .80 [69].

**NUMBER OF DEPENDENT VARIABLES** As noted earlier, each cell size must exceed the number of dependent variables. But even meeting that requirement, the increases in the number of dependent variables requires increases in the sample size required to maintain statistical power. We will defer our discussion of sample size and power until a later section, but as an example, required samples sizes increase by almost 50 percent as the number of dependent variables goes from two to just six.

**EQUAL VERSUS UNEQUAL CELL SIZES** Researchers should strive to maintain equal or approximately equal sample sizes per group. A **balanced design** has equal cell sizes for all cells, while an **unbalanced design** has varying cell sizes across the different cells. Although computer programs can easily accommodate unequal group sizes, the objective is to ensure that an adequate sample size is available for all groups. The smallest cell size is the basis for power, so increasing larger cells has no negative impact, but it also has limited positive impact. Also, balanced designs provide orthogonal estimates of the effects of the treatments [26]. This should lead the researcher to always be aware of the cell sizes, particularly in field studies where control over the achieved cell sizes may be limited. Finally, when cell sizes are very different (e.g., one group 20 and the other group 80), this difference can bias the test of statistical significance between the groups.

## FACTORIAL DESIGNS—TWO OR MORE TREATMENTS

Many times the researcher wishes to examine the effects of several independent variables or treatments rather than using only a single treatment in either the ANOVA or MANOVA tests. This capability is a primary distinction between MANOVA and discriminant analysis in being able to determine the impact of multiple independent variables in forming the groups with significant group differences. Moreover, including multiple independent variables allows for testing of interactions between treatments (see later discussion). An analysis with two or more independent variables (treatments or factors) is called a **factorial design**. In general, a design with  $n$  treatments is called an  $n$ -way factorial design.

**Selecting Treatments** The most common use of factorial designs involves those research questions that relate two or more nonmetric independent variables to a set of dependent variables. In these instances, the independent variables are specified in the design of the experiment or included in the design of the field experiment or survey questionnaire.

**BASIC MODEL TREATMENTS** As discussed throughout the chapter, a **treatment** or **factor** is a nonmetric independent variable with a defined number of levels (categories). Each level represents a different condition or characteristic that affects the dependent variable(s). The treatment(s) represent the theoretical constructs of interest in the main effect. In most controlled experiments, the treatments are manipulated so as to ensure exactly what the respondent is reacting toward. In an experiment these treatments and levels are designed by the researcher and administered in the course of the experiment. In field or survey research, they are characteristics of the respondents gathered by researcher and then included in the analysis.

When treatments are manipulated you should include a **manipulation check**, which is a variable(s) assessing if the respondent perceived the manipulation correctly. For example, if the experiment manipulates the type of ad (e.g., humorous vs. non-humorous), then the manipulation check would determine if the respondent interpreted the humorous ad as humorous, if that is the ad they were exposed to, and if the respondent interpreted the non-humorous ad as lacking humor, if that is the ad they saw. The manipulation check is not to ensure that the outcome was impacted, but solely to ensure that the manipulation worked as planned. Manipulation checks are many times insightful when the main effect is not supported, since the reasons may be either (a) an ineffective manipulation or (b) a failure of the treatment to create a difference. Respondents may be eliminated from the analysis if they fail a manipulation check, but researchers should be cautious in this regard so as to not impact (a) effects of the treatment as well as (b) sample size.

Successful manipulation checks help ensure that the reason for the results are truly the treatments and not their implementation. Manipulation checks are also recommended when mediation relationships are considered (see discussion in subsequent section) since the treatment impacts not only the outcome, but the mediator as well. One form of manipulation check that has gained in popularity is the **instructional manipulation check** which is a question or questions inserted to assess if the respondents are following instructions and providing sufficient attention to the actual questions [91]. It takes many forms (e.g., a Likert statement in which the respondent is told which response to answer) and can focus on any portion of the experiment to help support internal validity and identify any respondents that should be eliminated.

**CONTROLS AS TREATMENTS** In some instances treatments are used in addition to those involved in the main effects. Two common uses are as blocking factors or controls for confounds (see next section). A **blocking factor** is a nonmetric characteristic used to segment the sample to account for variability other than the treatment(s). A blocking factor represents conditions (e.g., method of data collection) or characteristics of the respondents (e.g., geographic location, gender, etc.) that potentially create differences in the dependent measures. Even though they are not independent variables of interest to the study and are independent of the treatment, neglecting them ignores potential sources of difference that, left unaccounted for, may obscure some results of interest to the study. As such, a blocking factor is many times termed a **nuisance factor**—a factor that may have variability we wish to control. Ideally blocking factors are known before an experiment is administered and can be integrated into the design, resulting in a randomized block design where randomization occurs within levels of the blocks. But blocking factors can also be applied post hoc to account for this source of variability in the outcome and hopefully strengthen the main effect.

Assume in our earlier advertising example we discovered that males in general reacted differently than females to the advertisements. If gender is then used as a blocking factor, we can evaluate the effects of the independent variables separately for males and females. Hopefully, this approach will make the effects more apparent than when we assume they both react similarly by not making a distinction on gender. The effects of message type and customer type can now be evaluated for males and females separately, providing a more precise test of their individual effects.

As discussed earlier, confounds are “common causes” that impact both treatment and outcome, thus becoming an alternative explanation for the main effect. To achieve a correct estimate of the main effect, the researcher must control for the confound(s). The most direct approach is randomization, which “breaks” the naturally occurring correlation of confound and treatment in the two groups. But when that is not possible, then the confounding variable can be included in the analysis to “control” for its effect. If the confound is a nonmetric variable (e.g., graduated college or not), then it enters the analysis in the same manner as the treatments. In this approach we are controlling for the confound by stratifying the sample across the confound group. If it is a metric variable, then it will be entered as a covariate (discussed in later section).

Thus, any nonmetric characteristic can be incorporated directly into the analysis to account for its impact on the dependent measures. However, if the variables you wish to control for are metric, they can be included as covariates, which are discussed in the next section.

**Number of Treatments** One of the advantages of multivariate techniques is the use of multiple variables in a single analysis. For MANOVA, this feature relates to both the number of dependent as well as independent variables that can be analyzed concurrently. As already discussed, the number of dependent variables affects the sample sizes required and other issues. But does the number of treatments (i.e., independent variables) affect the required sample size? Although ANOVA and MANOVA can analyze several treatments at the same time, several considerations relate to the number of treatments in an analysis.

**NUMBER OF CELLS FORMED** Perhaps the most limiting issue involving multiple treatments is the number of cells (groups) formed. As discussed in our earlier example, the number of cells is the product of the number of levels for each treatment. For example, if we had two treatments with two levels each and one treatment with four levels, a total of 16 cells ( $2 \times 2 \times 4 = 16$ ) would be formed. Maintaining a sufficient sample size for each cell (assuming 20 respondents per cell) would then require a total sample of 320.

When applied to survey or field experimentation data, however, increasing the number of cells becomes much more problematic. Because surveys or field research are generally not able to administer the survey individually to each cell of the design, the researcher must plan for a large enough overall sample to fill each cell to the required minimum. The proportions of the total sample in each cell most likely vary widely (i.e., some cells would be much more likely to occur than others), especially as the number of cells increases. In such a situation, the researcher must plan for an even larger sample size than the size determined by multiplying the number of cells by the minimum per cell. Let's look back to our earlier example to illustrate this problem.

Assume that we have a simple two-factor design with two levels for each factor ( $2 \times 2$ ). If this four-cell design were a controlled experiment, the researcher would be able to randomly assign 20 respondents per cell for an overall sample size of 80. What then if it is a field survey? If it were equally likely that respondents would fall into each cell, then the researcher could get a total sample of 80 and each cell should have a sample of 20. Such tidy proportions and samples rarely happen. What if one cell was thought to represent only 10 percent of the population? If we use a total sample of 80, then this cell would be expected to have a sample of only 8. Thus, if the researcher wanted a sample of 20 even for this small cell, the overall sample would have to be increased to 200.

Unless sophisticated sampling plans are used to ensure the necessary sample per cell, increasing the number of cells (thus the likelihood of unequal population proportions across the cells) will necessitate an even greater sample size than in a controlled experiment. Failure to do so would create situations in which the statistical properties of the analysis could be markedly diminished. This issue arises even in controlled experiments where constraints may limit a full factorial design or even suggest several ANOVAs versus MANOVA [21].

**CREATION OF INTERACTION EFFECTS** Any time more than one treatment is used, **interaction effects** are created. The interaction term represents the joint effect of two or more treatments. In simple terms, it means that the difference between groups of one treatment depends on the values of another treatment [76]. Let us look at a simple example.

**Figure 6.9**  
Interaction Terms for Factorial Models (Two, Three and Four Treatments)

Treatments	Interaction Terms		
	Two-Way	Three-Way	Four-Way
A, B	$A \times B$		
A, B, C	$A \times B$	$A \times B \times C$	
	$A \times C$		
	$B \times C$		
A, B, C, D	$A \times B$	$A \times B \times C$	$A \times B \times C \times D$
	$A \times C$		
	$A \times D$	$A \times B \times D$	
	$B \times C$		
	$B \times D$	$B \times C \times D$	
	$C \times D$		
		$A \times C \times D$	

Assume that we have two treatments: region (East versus West) and customer status (customer and non-customer). First, assume that on the dependent variable (attitude toward HBAT) customers score 15 points higher than non-customers. However, an interaction of region and customer status would indicate that the amount of the difference between customer and non-customer depended on the region of the customer. For example, when we separated the two regions, we might see that customers from the East scored 25 points higher than non-customers in the East, while in the West the difference was only 5 points. In both cases the customers scored higher, but the amount of the difference depended on the region. This outcome would be an interaction of the two treatments.

Interaction terms are created for each combination of treatment variables. Two-way interactions are treatments taken two at a time. Three-way interactions are combinations of three treatments, and so on. The number of treatments determines the number of interaction terms possible. Figure 6.9 shows the interactions created for two, three, and four independent variables:

We will discuss the various types of interaction terms and their interpretation in Stage 5, but the researcher must be ready to interpret and explain the interaction terms, whether significant or not, depending on the research question.

Obviously, the sample size considerations are of most importance, but the researcher should not overlook the implications of interaction terms. Besides using at least one degree of freedom for each interaction, the interactions directly impact how the effects for treatments can be interpreted.

## USING COVARIATES—ANCOVA AND MANCOVA

We discussed earlier the use of a blocking factor to control for influences on the dependent variable that are not part of the research design yet need to be accounted for in the analysis. It enables the researcher to control for nonmetric variables, but what about metric variables? One approach would be to convert the metric variable into a nonmetric variable (e.g., median splits, etc.), but this process is generally deemed unsatisfactory because much of the information contained in the metric variable is lost in the conversion. A second approach is to include the metric variables as **covariates**. These variables can extract extraneous influences from the dependent variable, thus increasing the within-group variance ( $MS_W$ ). The process follows two steps:

- 1 Procedures similar to linear regression are employed to remove variation in the dependent variable associated with one or more covariates.
- 2 A conventional analysis is carried out on the adjusted dependent variable. In a simplistic sense, it becomes an analysis of the regression residuals once the effects of the covariate(s) are removed.

When used with ANOVA, the analysis is termed *analysis of covariance* (ANCOVA) and the simple extension of the principles of ANCOVA to multivariate (multiple dependent variables) analysis is termed MANCOVA.

**Objectives of Covariance Analysis** The objective of the covariate is to eliminate any effects that (1) affect only a portion of the respondents or (2) vary among the respondents. Similar to the uses of a blocking factor, **covariate analysis** can achieve two specific purposes:

- 1 To eliminate some systematic error outside the control of the researcher that can bias the results
- 2 To account for differences in the responses due to unique characteristics of the respondents

In experimental settings, most systematic bias can be eliminated by the random assignment of respondents to various treatments. But in non-experimental research, such controls are not possible. For example, in testing advertising, effects may differ depending on the time of day or the composition of the audience and their reactions. Moreover, personal differences, such as attitudes or opinions, may affect responses, but the analysis does not include them as a treatment factor. The researcher uses a covariate to take out any differences due to these factors before the effects of the experiment are calculated.

**Selecting Covariates** An effective covariate is one that is *highly correlated with the dependent variable(s) but not correlated with the treatment(s)*. Let us examine why. Variance in the dependent variable forms the basis of our error term. In randomized experimental settings they are most often used to reduce any “noise” in the outcome measure. In nonrandomized settings they may act as a control variable to make the groups as comparable as possible on all factors except the treatment(s).

**UNCORRELATED WITH TREATMENT(S)** If the covariate is correlated with the dependent variable and *not* the treatment(s), we can explain some of the variance with the covariate (through linear regression), leaving a smaller residual (unexplained) variance in the dependent variable. This residual variance provides a smaller error term ( $MS_W$ ) for the  $F$  statistic and thus a more efficient test of treatment effects. The amount explained by the uncorrelated covariate would not have been explained by the independent variable anyway (because the covariate is not correlated with the independent variable). Thus, the test of the independent variable(s) is more sensitive and powerful.

**CORRELATED WITH TREATMENT(S)** However, if the covariate is correlated with the treatment(s), then the covariate will explain some of the variance that could have been explained by the treatment and reduce its effects. Because the covariate is extracted first, any variation associated with the covariate is not available for the treatment. In the case of a substantial correlation with the treatment, then the covariate is accounting for a substantive confounding variable, and should be “controlled for” so as to more accurately assess the main effect.

Thus, it is critical that the researcher ensure that the correlation of the covariates and independent variable(s) is small enough such that the reduction in explanatory power from reducing the variance that could have been explained by the independent variable(s) is less than the decrease in unexplained variance attributable to the covariates.

**Number of Covariates** A common question involves how many covariates to add to the analysis. Although the researcher wants to account for as many extraneous effects as possible, too large a number will reduce the statistical efficiency of the procedures. A rule of thumb [52] dictates that the maximum number of covariates is determined as follows:

$$\text{Maximum number of covariates} = (.10 \times \text{Sample size}) - (\text{Number of groups} - 1)$$

For example, for a sample size of 100 respondents and 5 groups, the number of covariates should be less than 6 [ $6 = .10 \times 100 - (5 - 1)$ ]. However, for only two groups, the analysis could include up to nine covariates.

The researcher should always attempt to minimize the number of covariates, while still ensuring that effective covariates are not eliminated, because in many cases, particularly with small sample sizes, they can markedly improve the sensitivity of the statistical tests.

**Assumptions for Covariance Analysis** Two requirements for use of an analysis of covariance are the following:

- 1 The covariates must have some relationship (correlation) with the dependent measures.
- 2 The covariates must have a **homogeneity of regression effect**, meaning that the covariate(s) have equal effects on the dependent variable across the groups. In regression terms, it implies equal coefficients for all groups.

Statistical tests are available to assess whether this assumption holds true for each covariate used. If either of these requirements is not met, then the use of covariates is inappropriate.

## MODELING OTHER RELATIONSHIPS BETWEEN TREATMENT AND OUTCOME

Recent years have seen an increased emphasis on elaborate conceptual models and increasingly sophisticated analytical procedures that have shifted the focus solely from the main effect to a set of additional relationships—notably mediation and moderation. As noted earlier, **mediation** is a third variable that “links” the treatment and outcome. Many times it is conceptualized as the variable that transmits the main effect to the outcome. **Moderation** is when the direction and/or strength of the main effect varies based on values of a third variable, i.e., an interaction. And as we will see later, these two relationships can be combined into an extended framework where we can assess if the direction and/or strength of a mediation is contingent on a variable separate from the mediation effect (i.e., moderated mediation) or identify the variable that transmits the moderation effect to the main effect (i.e., mediated moderation).

We should note that mediation and moderation are conceptual extensions of the theoretical models used for the main effect [122]. They should never be approached from a strictly empirical perspective and require the same level of conceptual support. Ultimately the failure to view mediation and moderation in the basic conceptual model may lead to seriously compromised results. Moreover, the emphasis on conceptual support reinforces the fact that statements as to causality come from the research design, not the statistical techniques.

**Mediation** Mediation is based on effects being transmitted from the treatment to the outcome through the mediator. As noted earlier, this is not a statistical relationship, but a conceptual relationship that can be verified by statistical analysis. The viability of mediation rests first and foremost on the conceptual model, then the nature of the study design, and finally on the statistical tests. Any mediation effect assumes that the causal ordering is correct, as the actual statistical results are similar in nature to other types of effects.

Mediation explains the process of “why” and “how” a main effect occurs [47, 55, 65]. In simple terms, the treatment causes the mediator, and in turn, the mediator causes the outcome. This mediation effect is commonly known as an **indirect effect** or **intervening effect**. The depiction of mediation as similar to the line of dominoes falling one after another [22] highlights the necessity of the mediation effect being both correct in causal ordering and the transmission of the effect from treatment to mediator to outcome.

The original analytical framework for examining mediation was proposed by Baron and Kenny [5] and Judd and Kenny [62]. A first step in this initial framework was to assess three constituent bivariate relationships of a mediation effect: treatment → outcome, treatment → mediator and mediator → outcome (refer back to Figure 6.8). Only if all three bivariate relationships are significant does the researcher move to the next step, which is assessing the mediated main effect by controlling for/partialling out the mediator (e.g., enter both mediator and treatment as effects on the outcome). While simplistic in design and applied across many disciplines, it still has generated substantial debate and research. From a more conceptual perspective, issues have been raised involving the necessity of the first step (e.g., does the main effect have to be significant) as well as subsequent tests on the type of mediation (full versus partial) and the significance tests for the indirect effect [130]. There are also more specific analytical issues such as statistical power anomalies [65], differences in results when testing through the Baron and Kenny approach versus structural equation modeling [60], and the ability to infer causal effects outside of a controlled experiment [78]. The interested reader is directed toward any number of special issues within various disciplines [e.g., 68, 70] or excellent texts/articles/chapters on the issues [80, 44, 77, 66, 72, 81, 61].

**Moderation** The second type of additional relationship is moderation, where the strength and/or direction of a main effect varies between different values of the moderator [5, 31, 44]. A classic moderator is gender, where the main effect is tested to see if it is the same for males versus females. Whereas mediation addresses the question “why,” moderation addresses the question “when.” Many researchers are expanding the domain of moderation testing and highlighting the additional insights gained [117].

Moderation is referred to many times as an interaction effect, which we also discussed in factorial designs earlier as a necessary condition for assessing the nature of the main effect. The moderation effect is not a statistical requirement, but instead an element of the conceptual model. While it shares statistical equivalence with a two-way interaction, only a select subset (if any) of the total set of interactions in a factorial design could be considered moderators as they may lack the requisite conceptual support. Moreover, a moderator is distinct from a covariate since the covariate requires the homogeneity of regression assumption, which is inconsistent with the basic characteristic of a moderator having different slopes.

**Selecting Mediation or Moderation** Both mediation and moderation pose fundamentally different questions. But is the researcher limited to addressing only one or the other, or can both be addressed concurrently? While there are procedures for estimating mediated moderation and moderated mediation effects, researchers generally approach mediation and moderation separately. We will discuss the two separate approaches in Stage 4 and provide interested readers with resources to address the more complicated effects.

## MANOVA COUNTERPARTS OF OTHER ANOVA DESIGNS

Many types of ANOVA designs exist and are discussed in standard experimental design texts [67, 86, 94]. Every ANOVA design has its multivariate counterpart; that is, any ANOVA on a single dependent variable can be extended to MANOVA designs. To illustrate this fact, we would have to discuss each ANOVA design in detail. Clearly, this type of discussion is not possible in a single chapter because entire books are devoted to the subject of ANOVA designs. For more information, the reader is referred to more statistically oriented texts [3, 15, 24, 33, 35, 9, 11, 20, 25].

## A SPECIAL CASE OF MANOVA: REPEATED MEASURES

We discussed a number of situations in which we wish to examine differences on several dependent measures. A special situation of this type occurs when the same respondent provides several measures, such as test scores over time, and we wish to examine them to see whether any trend emerges. Without special treatment, however, we would be violating the most important assumption, independence. Special ANOVA and MANOVA models, termed **repeated measures** models, account for this dependence and still ascertain whether any differences occurred across individuals for the set of dependent variables. The within-person perspective is important so that each person is placed on equal footing.

For example, assume we were assessing improvement on test scores over the semester. We must account for the earlier test scores and how they relate to later scores, and we might expect to see different trends for those with low versus high initial scores. Thus, we must match each respondent's scores when performing the analysis. The differences we are interested in become how much each person changes, not necessarily the changes in group means over the semester.

The use of ANOVA models for repeated measures analysis requires that the variances of the differences between all possible pairs of the treatment levels are equal. This is known as the **sphericity assumption**. If this is violated, then MANOVA provides a means of analysis where each repeated measure is treated separately. Panel analysis (see Chapter 5) also provides a framework for this type of research question. We do not address the details of repeated measures models in this text because it is a specialized form of MANOVA. The interested reader is referred to any number of excellent treatments on the subject [3, 15, 24, 33, 35, 34, 43, 89, 113].

## Research Design of MANOVA

A number of research approaches are applicable to ANOVA/MANOVA analysis:

Randomized designs, such as controlled experiments and field experiments, are most closely associated with these techniques.

Nonrandom designs, such as natural experiments and quasi-experiments, apply some of the experimental design principles to focus on specific effects.

Even observational studies provide data which can be analyzed with specialized techniques to allow for valid inferences of specific effects.

While the primary focus is on the main effect (i.e., the relationship between treatment/independent variable and the dependent/outcome variable), additional variables and relationships are also considered:

Additional explanatory effects such as moderation and mediation.

Control for external influences through covariates and blocking factors.

External variables such as confounds and instrumental variables.

Cells (groups) are formed by the combination of independent variables; for example, a three-category nonmetric variable (e.g., low, medium, high) combined with a two-category nonmetric variable (e.g., gender of male versus female) will result in a  $3 \times 2$  design with six cells (groups).

Sample size per group is a critical design issue:

Minimum sample size per group must be greater than the number of dependent variables.

The recommended minimum cell size is 20 observations per cell (group).

Researchers should try to have balanced designs (equal sample sizes per cell, i.e., group).

Manipulation checks provide the researcher with objective measures of the “success” of the manipulation and perception of the treatment and also degree to which respondents are following instructions and maintaining engagement.

Covariates and blocking variables are effective ways of controlling for external influences on the dependent variables that are not directly represented in the independent variables.

An effective covariate is one that is highly correlated with the dependent variable(s) but not correlated with the independent variables.

The maximum number of covariates in a model should be  $(.10 \times \text{Sample size}) - (\text{Number of groups} - 1)$ .

Mediation provides a supplementary explanation for “Why” a main effect might occur:

Requires strong conceptual support before any empirical testing

The type of mediation (complete or partial) is based on the extent to which the indirect effect (i.e., causal path through the mediator) accounts for>equals to the original main effect.

Moderation addresses external validity in terms of whether the main effect generalizes to the population or must be viewed as moderated (i.e., main effect varies in strength and/or direction based on a third variable). An example of moderation would be if the main effect varied based on gender.

## Stage 3: Assumptions of ANOVA and MANOVA

The univariate test procedures of ANOVA described in this chapter are valid (in a statistical sense) if it is assumed that the dependent variable is normally distributed, the groups are independent in their responses on the dependent variable, and variances are equal for all treatment groups. Some evidence [86, 126], however, indicates that  $F$  tests in ANOVA are robust with regard to these assumptions except in extreme cases.

For the multivariate test procedures of MANOVA to be valid, three assumptions must be met:

- Observations must be independent.
- Variance–covariance matrices must be equal for all treatment groups.
- The set of dependent variables must follow a multivariate normal distribution (i.e., any linear combination of the dependent variables must follow a normal distribution) [43].

In addition to the strict statistical assumptions, the researcher must also consider several issues that influence the possible effects—namely, the linearity and multicollinearity of the variate of dependent variables, and the impact of outliers.

## INDEPENDENCE

The most basic, yet most serious, violation of an assumption comes from a lack of **independence** among observations, meaning that the responses in each cell (group) are correlated with each other by some characteristic of the group. Violations of this assumption can occur as easily in experimental as well as non-experimental situations and increase the Type I error, making it easier to find significant differences. Any number of extraneous and unmeasured effects can affect the results by creating dependence within the groups, but two of the most common violations of independence follow:

- Time-ordered effects (serial correlation) occurring if measures are taken over time, even from different respondents.
- Gathering information in group settings, so that a common experience (such as a noisy room or confusing set of instructions) would cause a subset of individuals (those with the common experience) to have answers that are somewhat correlated.

Although no tests provide an absolute certainty of detecting all forms of dependence, the researcher should explore all possible effects and correct for them if found. One potential solution is the use of a blocking factor for the potential clustering effect in the research design and then randomize within each block as well. Another solution is application of hierarchical, multilevel or cluster sampling methods, each of which attempts to correct for the correlation within each cell. We discuss multilevel models in Chapter 5. In either case, or when dependence is suspected, the researcher should use a stricter level of significance (.01 or even lower).

## EQUALITY OF VARIANCE-COVARIANCE MATRICES

The second assumption of MANOVA is the equivalence of covariance matrices across the groups. Here we are concerned with substantial differences in the amount of variance of one group versus another for the dependent variables (similar to the problem of heteroscedasticity in multiple regression and homogeneity of variance in ANOVA). In MANOVA, with multiple dependent variables, the interest is in the variance–covariance matrices of the dependent measures for each group.

The variance equivalence test is a very “strict” test because instead of equal variances for a single variable in ANOVA, the MANOVA test examines all elements of the covariance matrix of the dependent variables. For example, for five dependent variables, the five correlations and 10 covariances are all tested for equality across the groups. As a result, increases in the number of dependent variables and/or the number of cells/groups in the analysis make the test more sensitive to finding differences and thus influence the significance levels used to determine if a violation has occurred.

**Testing for Equality of the Variance-Covariance Matrices** MANOVA programs conduct the test for equality of covariance matrices—typically the **Box’s M test**—and provide significance levels for the test statistic that indicate the likelihood of differences between the groups. Thus, the researcher is looking for *nonsignificant* differences between the groups, and the observed significance level of the test statistic is considered acceptable if it is less significant

than the threshold value for comparison. For example, if a .01 level was considered the threshold level for indicating violations of the assumption, values greater than .01 (e.g., .02) would be considered acceptable because they indicate no differences between groups, whereas values less than .01 (e.g., .001) would be problematic because they indicate that significant differences were present. Since the Box's M test very often indicates significant differences in the variance-covariance matrices if the .05 level is used, a value of .01 is often a better rule of thumb.

Given the sensitivity of the Box's M test to the size of the covariance matrices and the number of groups in the analysis, even simple research designs (four to six groups) with a small number of dependent variables will want to use a very different guideline for the level of significant differences (e.g., a significance level  $< .01$  rather than  $< .05$ ) when assessing whether differences are present. As the design complexity increases, even more conservative levels of significance may be considered acceptable.

**When the Assumption is Not Met** The sensitivity of the Box's M test results in many situations in which significant differences between the covariance matrices are indicated. In these instances, the researcher has several options for proceeding with the analysis.

**CHECK FOR NORMALITY** The Box's M test is especially sensitive to departures from normality [43, 110]. Thus, one should always check for univariate normality of all dependent measures before performing this test. Fortunately, a violation of this assumption has minimal impact if the groups are of approximately equal size (i.e., Largest group size  $\div$  Smallest group size  $< 1.5$ ).

**VARIABLE TRANSFORMATION** Apply one of the many variance-stabilizing transformations available (see Chapter 2 for a discussion of these approaches) and retest to see whether the problem is remedied. This approach becomes more difficult as the number of dependent variables increases, both in transforming the appropriate variables and interpretation of results, particularly interaction terms.

**EQUAL CELL SIZES** In most cases with balanced or relatively equal cell sizes violations of this assumption are mitigated. This reinforces the importance of a research design that strives for balanced designs as well as highlights the difficulties that may arise in non-controlled experimental approaches.

**UNEQUAL CELL SIZES** If unequal variances persist after transformation and the group sizes differ markedly, the researcher should make adjustments for their effects in the interpretation of the significance levels of both main and interaction effects. First, one must ascertain which group has the largest variance. This determination is easily made either by examining the variance–covariance matrix or by using the determinant of the variance–covariance matrix, which is provided by all statistical programs. In both measures high values indicate greater variance.

- If the larger variances are found with the larger group sizes, the alpha level is overstated, meaning that differences should actually be assessed using a somewhat lower value (e.g., use .03 instead of .05).
- If the larger variance is found in the smaller group sizes, then the reverse is true. The power of the test has been reduced, and the researcher should increase the significance level.

The researcher has a number of options to address violations of the assumption, ranging from specific variable transformations to adjustment of significance levels. Consistent in all cases, however, is the impact of balanced versus unbalanced designs and the role they play in this issue.

## **NORMALITY**

The last assumption for MANOVA concerns normality of the dependent measures. In the strictest sense, the assumption is that all the variables are multivariate normal. A **multivariate normal distribution** assumes that the joint effect of two variables is normally distributed. Even though this assumption underlies most multivariate techniques, no direct test is available for multivariate normality. Therefore, most researchers test for univariate normality of each

## MANOVA/ANOVA Assumptions

For the multivariate test procedures used with MANOVA to be valid:

Observations must be independent.

Variance–covariance matrices must be equal (or comparable) for all treatment groups.

The dependent variables must have a multivariate normal distribution.

Multivariate normality is assumed, but many times hard to assess; univariate normality does not guarantee multivariate normality, but if all variables meet the univariate normality requirement, then departures from multivariate normality are inconsequential.

F tests are generally robust if violations of these assumptions are modest.

Outliers can have substantial influence on the results and should be identified and potentially eliminated.

variable. Although univariate normality does not guarantee multivariate normality, if all variables meet this requirement, then any departures from multivariate normality are usually inconsequential.

Violations of this assumption have little impact with larger sample sizes, just as is found with ANOVA. Violating this assumption primarily creates problems in applying the Box's M test, but transformations can correct these problems in most situations. For a discussion of transforming variables, refer to Chapter 2. With moderate sample sizes, modest violations can be accommodated as long as the differences are due to skewness and not outliers.

## LINEARITY AND MULTICOLLINEARITY AMONG THE DEPENDENT VARIABLES

In addition to the three assumptions impacting the statistical results, analysts must also consider the linearity of the relationships between the outcomes (and any covariates, if included). The researcher is again encouraged first to examine the data, this time assessing the presence of any nonlinear relationships. If these exist, then the decision can be made whether they need to be incorporated into the dependent variable set, at the expense of increased complexity but greater representativeness. Chapter 2 addresses such tests.

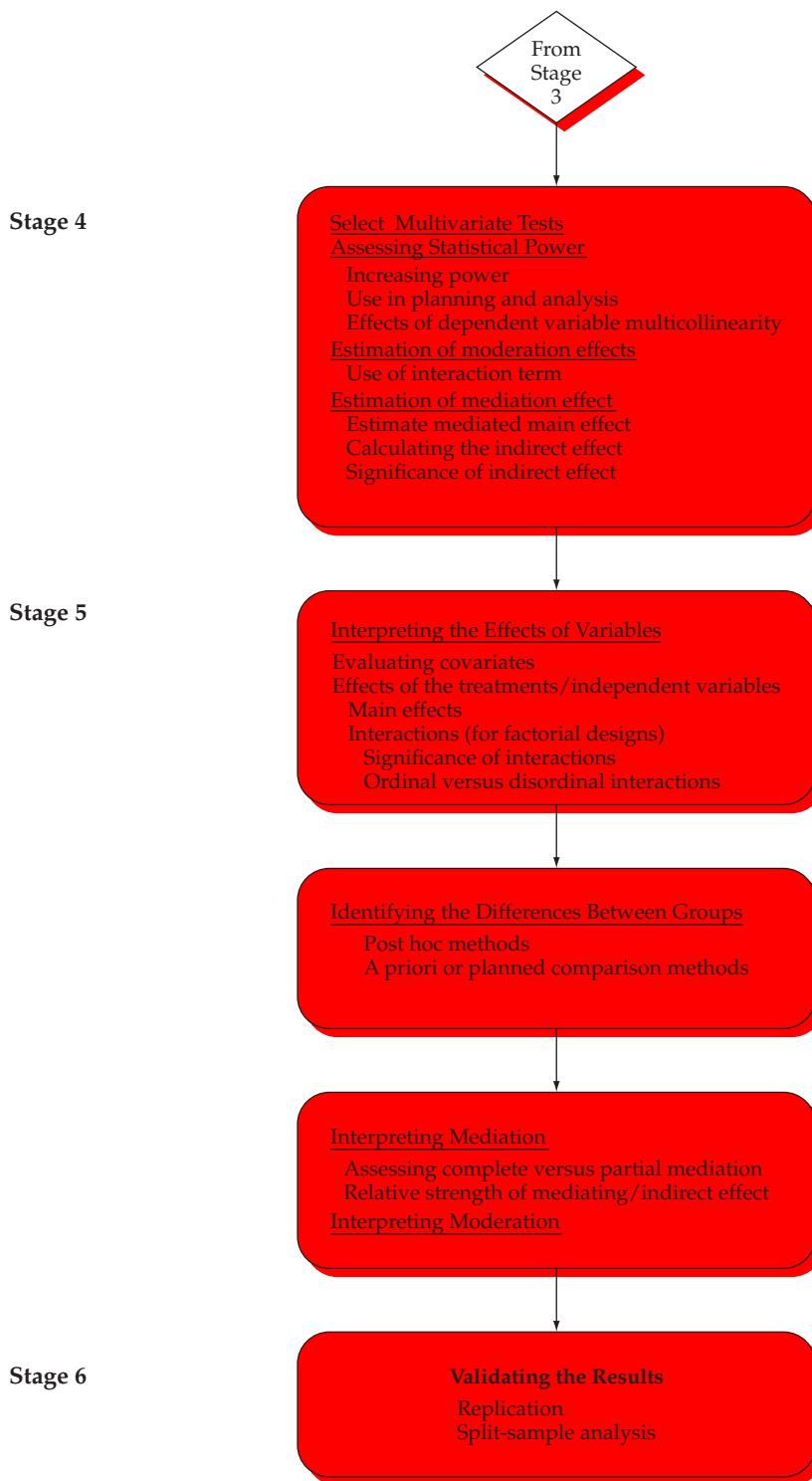
In addition to the linearity requirement, the dependent variables should not have high multicollinearity (discussed in Stage 2), which indicates redundant dependent measures and decreases statistical efficiency. We discuss the impact of multicollinearity on the statistical power of the MANOVA in the next section.

## SENSITIVITY TO OUTLIERS

Finally, in addition to the impact of heteroscedasticity discussed earlier, MANOVA (and ANOVA) are especially sensitive to outliers and their impact on the Type I error. The researcher is strongly encouraged first to examine the data for outliers and eliminate them from the analysis, if at all possible, because their impact will be disproportionate in the overall results.

## Stage 4: Estimation of the MANOVA Model and Assessing Overall Fit

Once the MANOVA analysis is formulated and the assumptions tested for compliance, the assessment of significant differences among the groups formed by the treatment(s) can proceed (see Figure 6.10). Estimation procedures based on the general linear model are becoming more common and the basic issues will be addressed. With the estimated model, the researcher then can assess the differences in means based on the test statistics most appropriate for the



**Figure 6.10**  
Stages 4–6 in the Multivariate Analysis of Variance (MANOVA) Decision Diagram

study objectives. Moreover, in any situation, but especially as the analysis becomes more complex, the researcher must evaluate the power of the statistical tests to provide the most informed perspective on the results obtained.

## ESTIMATION WITH THE GENERAL LINEAR MODEL

The traditional means of calculating the appropriate test statistics for ANOVA and MANOVA were established more than 70 years ago [125]. But it is useful to note that ANOVA/MANOVA in addition to multiple regression and its variants, plus discriminant analysis, are all forms of the **general linear model (GLM)**. The GLM is a more general version of the basic linear model in that it can accommodate a wide range of dependence models (e.g., multiple regression, ANOVA and MANOVA). In many instances procedures specific to each model (e.g., multiple regression, ANOVA) have gained popularity because of their specialization to each technique. But in some instances, such as with IBM SPSS and SAS, the GLM procedure is used rather than a specific MANOVA procedure. Note that the GLM differs from the generalized linear model (GLZ), which allows error distributions other than the normal distribution. For further discussion on these two model forms see Chapter 1.

## MEASURES FOR SIGNIFICANCE TESTING

In our discussions of the similarity of MANOVA to discriminant analysis we referred to the greatest characteristic root and the first discriminant function, and these terms imply that multiple discriminant functions may act as variates of the dependent variables. The number of functions used to test the significance of any effect (for treatments or interactions if present) is defined by the smaller of  $(k - 1)$  or  $p$  where  $k$  is the number of groups and  $p$  is the number of dependent variables. Thus, any measure for testing the statistical significance of group differences in MANOVA may need to consider differences across multiple discriminant functions.

As we discussed with discriminant analysis (Chapter 7), researchers use a number of statistical criteria to apply **significance tests** relating to the differences across dimensions of the dependent variables. The most widely used measures are:

- **Roy's greatest characteristic root (gcr)**, as the name implies, measures the differences on only the first discriminant function among the dependent variables. This criterion provides advantages in power and specificity of the test but makes it less useful in certain situations where all dimensions should be considered. Roy's gcr test is most appropriate when the dependent variables are strongly interrelated on a single dimension, but it is also the measure most likely to be severely affected by violations of the assumptions.
- **Wilks' lambda** (also known as the **U statistic**) is many times referred to as the *multivariate F* and is commonly used for testing overall significance between groups in a multivariate situation. Unlike Roy's gcr statistic, which is based on the first discriminant function, Wilks' lambda considers all the discriminant functions; that is, it examines whether groups are somehow different without being concerned with whether they differ on at least one linear combination of the dependent variables. Although the distribution of Wilks' lambda is complex, good approximations for significance testing are available by transforming it into an *F* statistic [94].
- **Pillai's criterion** and **Hotelling's T<sup>2</sup>** are two other measures similar to Wilks' lambda because they consider all the characteristic roots and can be approximated by an *F* statistic.

With only two groups, all of the measures are equivalent. Differences occur as the number of discriminant functions increase. Rules of Thumb 6-4 identify the measure(s) best suited to differing situations.

## STATISTICAL POWER OF THE MULTIVARIATE TESTS

In simple terms for MANOVA, **power** is the probability that a statistical test will identify a treatment's effect if it actually exists. Power can also be expressed as one minus the probability of a **Type II error** or **beta ( $\beta$ )** error (i.e.,  $\text{Power} = 1 - \beta$ ). Statistical power plays a critical role in any MANOVA analysis because it is used both in the planning process (i.e., determining necessary sample size) and as a diagnostic measure of the results, particularly

## Selecting a Statistical Measure

The preferred measure is the one that is most immune to violations of the assumptions underlying MANOVA and yet maintains the greatest power. Each measure is preferred in differing situations:

Pillai's criterion or Wilks' lambda is the preferred measure when the basic design considerations (adequate sample size, no violations of assumptions, approximately equal cell sizes) are met.

Pillai's criterion is considered more robust and should be used if sample size decreases, unequal cell sizes appear, or homogeneity of covariances is violated.

Roy's gcr is a more powerful test statistic if the researcher is confident that all assumptions are strictly met and the dependent measures are representative of a single dimension of effects.

In a vast majority of the situations, all of the statistical measures provide similar conclusions.

When faced with conflicting conditions, however, statistical measures can be selected that meet the situation faced by the researcher.

when nonsignificant effects are found. The following sections first examine the impacts on statistical power and then address issues unique to utilizing power analysis in a MANOVA design. The reader is also encouraged to review the discussion of power in Chapter 1.

**Impacts on Statistical Power** The level of power for any of the four statistical criteria—Roy's gcr, Wilks' lambda, Hotelling's  $T^2$ , or Pillai's criterion—is based on three considerations: the alpha ( $\alpha$ ) level, the effect size of the treatment, and the sample size of the groups. Each of these considerations is controllable in varying degrees in a MANOVA design and provides the researcher with a number of options in managing the power in order to achieve the desired level of power in the range of .80 or above.

**STATISTICAL SIGNIFICANCE LEVEL ALPHA ( $\alpha$ )** As discussed in Chapter 1, power is inversely related to the **alpha ( $\alpha$ )** level selected. Many researchers assume that the significance level is fixed at some level (e.g., .05), but it actually is a judgment by the researcher as to where to place the emphasis of the statistical testing. Many times the other two elements affecting power (effect size and sample size) are already specified or the data have been collected, thus the alpha level becomes the primary tool in defining the power of an analysis.

By setting the alpha level required to denote statistical significance, the researcher is balancing the desire to be strict in what is deemed a significant difference between groups while still not setting the criterion so high that differences cannot be found.

**Increasing Alpha** (i.e.,  $\alpha$  becomes more conservative, such as moving from .05 to .01) reduces the chances of accepting differences as significant when they are not really significant. However, doing so decreases power because being more selective in what is considered a statistical difference also increases the difficulty in finding a significant difference.

**Decreasing Alpha** (e.g.,  $\alpha$  moves from .05 to .10) is considered many times as being “less statistical” because the researcher is willing to accept smaller group differences as significant. However, in instances where effect sizes or sample sizes are smaller than desired it may be necessary to be less concerned about accepting these false positives and decreasing the alpha level to increase power. One such example is when making multiple comparisons. To control experiment-wide error rate, the alpha level is increased for each separate comparison. However, to make several comparisons and still achieve an overall rate of .05 may require strict levels (e.g., .01 or less) for each separate comparison, thus making it hard to find significant differences (i.e., lower

power). Here the researcher may increase the overall alpha level to allow a more reasonable alpha level for the separate tests.

The researcher must always be aware of the implications of adjusting the alpha level, because the overriding objective of the analysis is not only avoiding Type I errors but also identifying the treatment effects if they do indeed exist. If the alpha level is set too stringently, then the power may be too low to identify valid results. The researcher should try to maintain an acceptable alpha level with power in the range of .80. For a more detailed discussion of the relationships between Type I and Type II errors and power, see Chapter 1.

**EFFECT SIZE** How does the researcher increase power once an alpha level is specified? The primary tool at the researcher's disposal is the sample size of the groups. Before we assess the role of sample size, however, we need to understand the impact of **effect size**, which is a standardized measure of group differences, typically expressed as the differences in group means divided by their standard deviation. This formula leads to several generalizations:

- As would be expected, all other things equal, larger effect sizes have more power (i.e., are easier to find) than smaller effect sizes.
- The magnitude of the effect size has a direct impact on the power of the statistical test. For any given sample size, the power of the statistical test will be higher the larger the effect size. Conversely, if a treatment has a small expected effect size, it is going to take a much larger sample size to achieve the same power as a treatment with a large effect size.

Researchers are always hoping to design experiments with large effect sizes. However, when used with field research, researchers must "take what they get" and thus be aware of the possible effect sizes when planning their research as well as when analyzing results.

**SAMPLE SIZE** With the alpha level specified and the effect size identified, the final element affecting power is the sample size. In many instances, this element is the most controllable by the researcher. As discussed before, increased sample size generally reduces sampling error and increases the sensitivity (power) of the test. Other factors discussed earlier (alpha level and effect size) also affect power, and we can draw some generalizations for ANOVA and MANOVA designs:

- In analyses with group sizes of fewer than 30 members, obtaining desired power levels can be quite problematic. If effect sizes are small, then the researcher may be required to decrease alpha (e.g., .05 to .10) to obtain desired power.
- Increasing sample sizes in each group has noticeable effects until group sizes of approximately 150 are reached, and then the increase in power slows markedly.
- Remember that large sample sizes (e.g., 400 or larger) reduce the sampling error component to such a small level that most small differences are regarded as statistically significant. When the sample sizes do become large and statistical significance is indicated, the researcher must examine the power and effect sizes to ensure not only statistical significance but practical significance as well.

**UNIQUE ISSUES WITH MANOVA** The ability to analyze multiple dependent variables in MANOVA creates additional constraints on the power in a MANOVA analysis. The reduction of power for the MANOVA tests may result in a nonsignificant multivariate test, but significant univariate tests. One source [70] of published tables presents power in a number of common situations for which MANOVA is applied. However, we can draw some general conclusions from examining a series of conditions encountered in many research designs. Figure 6.11 provides an overview of the sample sizes needed for various levels of analysis complexity. A review of the table leads to some general points.

- Increasing the number of dependent variables requires increased sample sizes to maintain a given level of power. The additional sample size needed is more pronounced for the smaller effect sizes.

**Figure 6.11**

Sample Size Requirements per Group for Achieving Statistical Power of .80 in MANOVA

Effect Size	NUMBER OF GROUPS											
	3				4				5			
	Number of Dependent Variables				Number of Dependent Variables				Number of Dependent Variables			
2	4	6	8	2	4	6	8	2	4	6	8	
Very large	13	16	18	21	14	18	21	23	16	21	24	27
Large	26	33	38	42	29	37	44	46	34	44	52	58
Medium	44	56	66	72	50	64	74	84	60	76	90	100
Small	98	125	145	160	115	145	165	185	135	170	200	230

Source: J. Läuter. 1978. Sample Size Requirements for the  $T^2$  Test of MANOVA (Tables for One-Way Classification). *Biometrical Journal* 20: 389–406.

- For small effect sizes, the researcher must be prepared to engage in a substantial research effort to achieve acceptable levels of power. For example, to achieve the suggested power of .80 when assessing small effect sizes in a four-group design, 115 subjects per group are required if two dependent measures are used. The required sample size increases to 185 per group if eight dependent variables are considered.

As we can see, the advantages of utilizing multiple dependent measures come at a cost in our analysis. As such the researcher must always balance the use of more dependent measures versus the benefits of parsimony in the dependent variable set that occur not only in interpretation but in the statistical tests for group differences as well.

**CALCULATING POWER LEVELS** To calculate power for ANOVA analyses, published sources [18, 111] as well as computer programs are now available. The methods of computing the power of MANOVA, however, are much more limited [41, 116]. The researcher can also approximate the power level by choosing the smallest effect size among the outcome measures. Fortunately, most computer programs provide an assessment of power for the significance tests and enable the researcher to determine whether power should play a role in the interpretation of the results.

In terms of published material for planning purposes, little exists for MANOVA because many elements affect the power of a MANOVA analysis. The available methods of computing the power of MANOVA are much more limited to specific programs (e.g., G Power [29]) or procedures within statistical packages (e.g., GLMPOWER in SAS [14]). The researcher, however, should utilize the tools available for ANOVA and then make adjustments described to approximate the power of a MANOVA design.

**Using Power in Planning and Analysis** The estimation of power should be used both in planning the analysis and in assessing the results. In the planning stage, the researcher determines the sample size needed to identify the estimated effect size. In many instances, the effect size can be estimated from prior research or reasoned judgments, or even set at a minimum level of practical significance. In each case, the sample size needed to achieve a given level of power with a specified alpha level can be determined.

By assessing the power of the test criteria after analysis is completed, the researcher provides a context for interpreting the results, especially if significant differences are not found. The researcher must first determine whether the achieved power is sufficient (.80 or above). If not, can the analysis be reformulated to provide more power? A possibility includes some form of blocking treatment or covariate analysis that will make the test more efficient by accentuating the effect size. If the power was adequate and statistical significance was not found for a treatment effect, then most likely the effect size for the treatment was too small to be of statistical or practical significance.

**The Effects of Dependent Variable Multicollinearity on Power** Up to this point we discussed power from a perspective applicable to both ANOVA and MANOVA. In MANOVA, however, the researcher must also consider

the effects of multicollinearity of the dependent variables on the power of the statistical tests [114]. The researcher, whether in the planning or analysis stage, must consider the strength and direction of the correlations as well as the effect sizes of the dependent variables. If we classify variables by their effect sizes as strong or weak, then several patterns emerge [20].

- First, if the correlated variable pair is made up of either strong-strong or weak-weak variables, then the greatest power is achieved when the correlation between variables is highly negative. This result suggests that MANOVA is optimized by adding dependent variables with high negative correlations. For example, rather than including two redundant measures of satisfaction, the researcher might replace them with correlated measures of satisfaction and dissatisfaction to increase power.
- When the correlated variable pair is a mixture (strong-weak), then power is maximized when the correlation is high, either positive or negative.
- One exception to this general pattern is the finding that using multiple items to increase reliability results in a net gain of power, even if the items are redundant and positively correlated.

**Review of Power in MANOVA** One of the most important considerations in a successful MANOVA is the statistical power of the analysis. Even though researchers engaged in experiments have much more control over the three elements that affect power, they must be sure to address the issues raised in the preceding sections or potential problems that reduce power below the desired value of .80 can easily occur. In field research, the researcher is faced not only with less certainty about the effect sizes in the analysis, but also the lack of control of group sizes and the potentially small group sizes that may occur in the sampling process. Thus, issues in the design and execution of field research discussed in Stage 2 are critical in a successful analysis as well.

## ESTIMATING ADDITIONAL RELATIONSHIPS: MEDIATION AND MODERATION

Up to this point we have discussed issues directly relating to the main effect. But the researcher may also be interested in two additional relationships—mediation and moderation. These relationships add insight into “why” the treatment impacted the outcome (mediation) and “when” the treatment effect occurred (moderation). Moreover, these effects may be viewed in a more general framework where they are combined (e.g., moderated mediation or mediated moderation) [28]. We should note that the following discussion uses an ANOVA context with a single outcome variable for simplicity, but in practice the only change for MANOVA is the substitution of the discriminant function for the outcome measures for the single outcome measure. This will be addressed later in our discussion

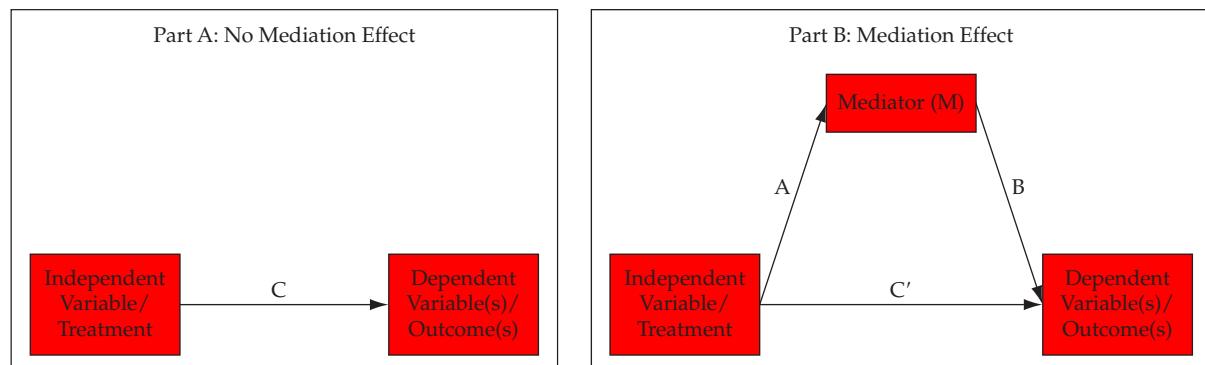
**Mediation** Mediation involves three bivariate relationships depicted in Figure 6.12. The unmediated or original main effect is C as shown in Part A. When we add the mediation effect, two additional relationships are added (Part B): A is the relationship of the treatment and the mediator (M) and B is the relationship of the mediator and the outcome. When these two relationships are added, then we term the mediated main effect C' to distinguish it from the unmediated relationship.

**ESTIMATING THE MEDIATION MODEL** Baron and Kenny [5] proposed a simple process which requires each of the first three steps to be statistically significant and then estimating the mediating effect in step 4:

- 1 Estimate path C: ensure that a significant relationship exists between treatment and outcome;
- 2 Estimate path A: ensure that a significant relationship exists between the treatment and the mediator (M);
- 3 Estimate path B: ensure that a significant relationship exists between mediator and outcome.

Each of these steps is estimated separately, either by correlational analysis or simple univariate regression. Then step 4 involves a multivariate regression controlling for the mediation effect:

**Figure 6.12**  
Mediation of the Main Effect



- 4 Estimate path  $C'$ : estimate the mediated effect by adding the mediator as another independent variable in a multiple regression, so that now the relationship  $C'$  is the treatment → outcome relationship “controlling for” the effect attributed to the mediator (B).

The significance of  $C'$  determines whether the mediation is complete or partial. **Complete mediation** occurs when  $C'$  is nonsignificant, meaning that after we account for the impact of the mediator (B) the original treatment → outcome relationship disappears. **Partial mediation** is when the impact of the mediator does not account for all of the original treatment → outcome relationship and  $C'$  is still significant.

**CALCULATING THE INDIRECT EFFECT** While we can now determine if there is complete or partial mediation, calculating the size of the mediating effect is important, especially if it is only partial mediation. The impact of the mediator is termed the indirect effect – the effect of the treatment that is transmitted through the mediator. In simple terms the indirect is just the product of the effects of A and B. From our results, we calculate the indirect effect as  $A \times B$ , using the estimated coefficients from steps 2 and 3 above.

This leads to the decomposition of the original treatment → outcome relationship into two parts: the mediated effect of treatment still remaining and the indirect effect through the mediator. It can be expressed as:

$$C = C' + A \times B$$

**SIGNIFICANCE OF THE INDIRECT EFFECT** A final step is to calculate the statistical significance of the indirect effect. While we know that both A and B are significant from steps 2 and 3, we do not know the significance of their combined effect. Sobel [107] developed a simple significance test using information about the estimated coefficients and their standard errors. The equation is:

$$Z = \frac{A \times B}{\sqrt{(A^2 \times s_b^2) + (B^2 \times s_a^2)}}$$

where  $A$  = estimated coefficient of treatment → mediator

$B$  = estimated coefficient of mediator → outcome

$s_a$  = standard error of A

$s_b$  = standard error of B

We then evaluate  $Z$  as a  $Z$  value (e.g., larger than 1.96 is significant at the .05 level).

The Sobel test is known for its low statistical power [79] and requires a normal distribution. Therefore, bootstrapping is recommended as a more accurate assessment of the indirect effect. Bootstrapped estimates are available in programs such as the PROCESS macro (see next section) and structural equation models used to estimate mediation effects.

**Moderation** The test for moderation is a test of interaction, equivalent to that discussed in Chapter 5 on multiple regression. Testing for moderation is a three-step process:

- 1 Calculate the interaction term between treatment and moderator.
- 2 Estimate the moderated relationship, with three independent variables (treatment, moderator and interaction term of treatment  $\times$  moderator).
- 3 Moderation is indicated by a statistically significant coefficient for the interaction term.

The coefficient for the interaction term represents the change in the coefficient/slope for the treatment when the moderator changes one unit, with a more detailed discussion in Stage 5. The interpretation of the moderation effect depends to some extent on the measurement characteristics (metric versus nonmetric) of both the moderator and outcome measure. The integration of mediation and moderation in a more general framework provides a useful perspective on interpreting moderation [28], while more direct methods of interpretation are also available (e.g., [93]).

**The PROCESS Macro** The procedures for estimating the basic mediation and moderation effects are straightforward and simple to implement in any software package. But as the nature of the effects becomes more complex (e.g., moderated mediation or such) or multiple mediators or moderators are proposed, the typical researcher will find it more difficult to correctly specify the estimation model. To this end the PROCESS macro developed by Andrew Hayes [44] for either IBM SPSS or SAS provides the researcher with a template-driven set of models for estimating mediation, moderation and a wide range of combined effects. The most recent version provides over 70 different models of varying effects and the researcher must only select the appropriate model, specify the variables involved and the macro will provide the appropriate results. A number of more advanced measures of both mediation and moderation effects are presented, along with such features as bootstrapping options for assessing statistical significance of more complex effects. We encourage researchers interested in mediation/moderation to review the materials provided with the PROCESS macro not only as an analytical aid, but a framework for specifying the exact set of mediation, moderation or even combined effects desired. A major limitation of the PROCESS macro is that it is useful only with multiple regression, and structural equation modeling may be the better approach for testing mediation and moderation.

**Mediation and Moderation in MANOVA** Up to this point we have discussed moderation and mediation more in conceptual terms rather than specifically oriented towards MANOVA. While we have introduced the concepts in a simple manner since they are applicable in so many research situations, there is no reason why they cannot be applied to MANOVA in most contexts. For moderation, creating the interaction term and introducing the mediator are straightforward in any MANOVA program. The complications arise with mediation, since the multiple outcome measures suggest either a series of mediation tests for each outcome or computation of some composite value to represent the outcomes in a single mediation test. Researchers can certainly perform the separate mediation tests, but are precluded from combining them in some fashion for an overall effect. To represent the outcome as a single composite measure, researchers might look to interpretive measures developed by Grice and Iwasaki [39], which demonstrate procedures for calculating these composite measures that could then be used in a single mediation test.

## MANOVA Estimation

The four most widely used measures for assessing statistical significance between groups on the independent variables are:

- Roy's greatest characteristic root
- Wilks' lambda
- Pillai's criterion
- Hotelling's  $T^2$ .

In most situations the results/conclusions will be the same across all four measures, but in some unique instances the results will differ between the measures.

Maintaining adequate statistical power is critical:

Power in the .80 range for the selected alpha level is acceptable.

When the effect size is small, the researcher should use larger sample sizes per group to maintain acceptable levels of statistical power.

If dependent measures are very highly correlated ( $>.60$ ), consider eliminating one or more dependent measures to reduce collinearity or use some form of summary measure.

Mediation is estimated by entering the mediator as an additional effect to the main effect.

Indirect effect must be calculated and tested either with the Sobel test or through bootstrapping procedures.

Moderation is tested as a conventional interaction effect of the treatment with the moderator.

Software such as the PROCESS macro make many forms of mediation and moderation, plus combinations of mediation and moderation, available to the researcher.

## Stage 5: Interpretation of the MANOVA Results

Once the statistical significance of the treatments has been assessed, the researcher turns attention to examining the results to understand how each treatment affects the dependent measures. In doing so, a series of three steps should be taken:

- 1 Interpret the effects of covariates, if employed.
- 2 For each significant treatment effect (and interaction in a factorial design), assess the pattern of group differences on the dependent variables. In the case of interactions, distinguish between ordinal and disordinal interactions and their implications on interpreting the corresponding treatment effects.
- 3 Determine for each significant treatment effect which specific groups within that treatment exhibit significant differences on the outcome measures.
- 4 Examine each treatment effect to assess the relative importance of the outcome measures in the overall treatment effect.

We first examine the methods by which the significant covariates and dependent variables are identified, and then we address the methods by which differences among individual groups and dependent variables can be measured.

### EVALUATING COVARIATES

Covariates can play an important role by including metric variables into a MANOVA or ANOVA design. However, because covariates act as a control measure on the dependent variate, they must be assessed before the

treatments are examined. Having met the assumptions for applying covariates, the researcher can interpret the actual effect of the covariates on the dependent variate and their impact on the actual statistical tests of the treatments.

**Assessing Overall Impact** The most important role of the covariate(s) is the overall impact in the statistical tests for the treatments. The most direct approach to evaluating these impacts is to run the analysis with and without the covariates. Effective covariates will improve the statistical power of the tests and reduce within-group variance. If the researcher does not see any substantial improvement, then the covariates may be eliminated, because they reduce the degrees of freedom available for the tests of treatment effects. This approach also can identify those instances in which the covariate is too powerful and reduces the variance to such an extent that the treatments are all nonsignificant. Often this situation occurs when a covariate is included that is correlated with one of the independent variables and thus removes this variance, thereby reducing the explanatory power of the independent variable.

**Interpreting the Covariates** Because MANCOVA and ANCOVA are applications of regression procedures within the analysis of variance method, assessing the impact of the covariates on the dependent variables is quite similar to examining regression equations. If the overall impact is deemed significant, then each covariate can be examined for the strength of the predictive relationship with the dependent measures. If the covariates represent theoretically based effects, then these results provide an objective basis for accepting or rejecting the proposed relationships. In a practical vein, the researcher can examine the impact of the covariates and eliminate those with little or no effect.

## ASSESSING EFFECTS ON THE DEPENDENT VARIATE

With the impacts, if any, of the covariates accounted for in the analysis, the next step is to examine the impacts of each treatment (independent variable) on the dependent variables. In doing so, we will first discuss how to assess the differences attributable to each treatment, particularly in the presence of interactions. The presence of significant interactions requires that they be interpreted before any of the associated treatments can be interpreted.

**Main Effects of the Treatments** We already discussed the measures available to assess the statistical significance of a treatment. When a significant effect is found, we call it a **main effect**, meaning that significant differences between two or more groups are defined by the treatment. With two levels of the treatment, a significant main effect ensures that the two groups are significantly different. With three or more levels, however, a significant main effect *does not* guarantee that all three groups are significantly different, instead just that at least one significant difference is present between a pair of groups. As we will see in the next section, a wide array of statistical tests is available to assess which groups differ on both the variate and separate dependent variables.

So how do we portray a main effect? A main effect is typically described by the difference between groups on the dependent variables in the analysis. This is typically performed on each outcome measure separately since there is no overall value for the discriminant functions involved. Grice and Iwasaki [39] propose an approach that both interprets the discriminant functions in a manner similar to that performed in discriminant analysis, and then computes composite scores for each discriminant function that can be used to examine patterns across the groups. Researchers needing more insight into the nature of the differences reflected by the set of outcomes are encouraged to review this approach.

As an example, assume that gender had a significant main effect on a 10-point satisfaction scale. We could then look to the difference in means as a way of describing the impact. If the female group had a mean score of 7.5 and males had an average score of 6.0, we could state that the difference due to gender was 1.5. Thus, all other things equal, females would be expected to score 1.5 points higher than males.

To define a main effect in these terms, however, requires two additional analyses:

- 1 If the analysis includes more than one treatment, the researcher must examine the interaction terms to see whether they are significant and, if so, whether they allow for an interpretation of the main effects.
- 2 If a treatment involves more than two levels, then the researcher must perform a series of additional tests between the groups to see which pairs of groups are significantly different.

We will discuss the interpretation of interaction terms in the next section and then examine the types of statistical tests available for assessing group differences when the analysis involves more than two groups in the following section.

**Significance of the Interaction Terms** The interaction term represents the joint effect of two or more treatments. Any time a research design has two or more treatments, the researcher must first examine the interactions before any statement can be made about the main effects. First, we will discuss how to identify significant interactions. Then we will discuss how to classify significant interactions in order to interpret their impact on the main effects of the treatment variables. Several sources provide excellent reviews of interactions and their interpretation [119, 76].

Interaction effects are evaluated with the same criteria as main effects, namely both multivariate and univariate statistical tests and statistical power. Software programs provide a complete set of results for each interaction term in addition to the main effects. All of the criteria discussed earlier apply to evaluating interactions as well as main effects.

Statistical tests that indicate the interaction is nonsignificant can be interpreted as supporting the independent effects of each treatments. Independence in factorial designs means that the effect of one treatment (i.e., group differences) is the same for each level of the other treatment(s) and that the main effects can be interpreted directly. Here we can describe the differences between groups as constant when considered in combination with the second treatment. We will discuss interpretation of the main effect in a simple example in a later section.

If the interactions are deemed statistically significant, it is critical that the researcher identify the type of interaction (ordinal versus disordinal) because it has direct bearing on the conclusion that can be drawn from the results. As we will see in the next section, significant interactions can potentially confound any description of the main effects depending on their nature.

**Types of Significant Interactions** The statistical significance of an interaction term is made with the same statistical criteria used to assess the impact of main effects. Upon assessing the significance of the interaction term, the researcher must examine effects of the treatment (i.e., the differences between groups) to determine the type of interaction and the impact of the interaction on the interpretation of the main effect. Significant interactions can be classified into one of two types: ordinal or disordinal interactions.

**ORDINAL INTERACTIONS** When the effects of a treatment are not equal across all levels of another treatment, but the group difference(s) is always the same direction, we term this an **ordinal interaction**. In other words, the group means for one level are always greater/lower than another level of the same treatment no matter how they are combined with the other treatment.

Assume that two treatments (gender and age) are used to examine satisfaction. An ordinal interaction occurs, for example, when females are always more satisfied than males, but the amount of the difference between males and females differs by age group.

When significant interactions are ordinal, the researcher must interpret the interaction term to ensure that its results are acceptable conceptually. Here the researcher must identify where the variation in group differences occurs and how that variation relates to the conceptual model underlying the analysis. If so, then the effects of each treatment must be described in terms of the other treatments it interacts with. Tests of significance can be performed for each level to see if they are significantly different.

In the preceding example, we can make the general statement that gender does affect satisfaction in that females are always more satisfied than males. However, the researcher cannot state the difference in simple terms as could be done with a simple main effect. Rather the differences on gender must be described for each age category because the male/female differences vary by age. Moreover, significance tests for each level of gender (male versus female) across each level of age can reveal if the differences, while varying across levels, are also all significant or

not at each level. For example, the male/female difference may be significant for younger respondents, but not older respondents.

**DISORDINAL INTERACTIONS** When the differences between levels switch, depending on how they are combined with levels from another treatment, this is termed a **disordinal interaction**. Here the effects of one treatment are positive for some levels and negative for other levels of the other treatment.

In our example of examining satisfaction by gender and age, a disordinal interaction occurs when females have higher satisfaction than males in some age categories, but males are more satisfied in other age categories.

*If the significant interaction is deemed disordinal, then the main effects of the treatments involved in the interaction cannot be interpreted and the study should be redesigned.* This suggestion stems from the fact that with disordinal interactions, the main effects vary not only across treatment levels but also in direction (positive or negative). Thus, the treatments do not represent a consistent effect. Recent research has identified procedures for distinguishing between ordinal and disordinal interactions on a more empirical basis [124, 74].

The importance of assessing interactions before main effects is demonstrated with disordinal interactions. A disordinal interaction may be significant without either main effect being significant, which is not possible for a significant ordinal interaction. This occurs because the difference between group means, when viewed one treatment at a time, do not differ significantly. Thus, without assessing the interaction a researcher might contend that there were no main effects when actually they were present.

**An Example of Interpreting Interactions** Interactions represent the differences between group means when grouped by levels of another treatment variable. Even though we could interpret interactions by viewing a table of values, graphical portrayals are quite effective in identifying the type of interaction between two treatments. The result is a multiple line graph, with levels of one treatment represented on the horizontal axis. Each line then represents one level of the second treatment variable.

Figure 6.13 portrays each type of interaction using the example of interactions between two treatments: cereal shapes and colors. Cereal shape has three levels (balls, cubes, and stars) as does color (red, blue, and green). The vertical axis represents the mean evaluations (the dependent variable) of each group of respondents across the combinations of treatment levels. The X axis represents the three categories for color (red, blue, and green). The lines connect the category means for each shape across the three colors. For example, in the upper graph the value for red balls is about 4.0, the value for blue balls is about 5.0, and the value increases slightly to about 5.5 for green balls.

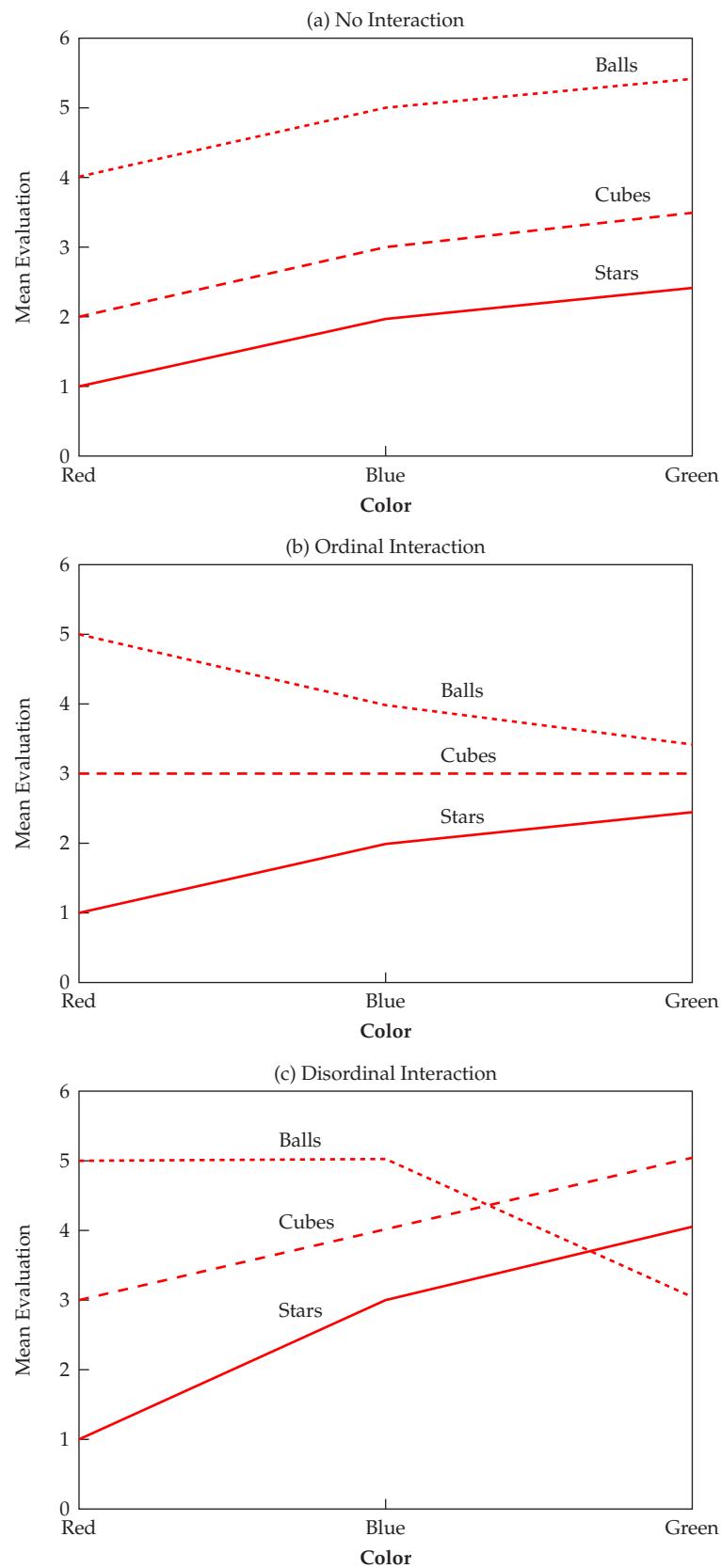
How do the graphs identify the type of interaction? As we will discuss, each of the three interactions has a specific pattern:

**NO INTERACTION** Shown by the parallel lines representing the differences of the various shapes across the levels of color (the same effect would be seen if the differences in color were graphed across the three types of shape). In the case of no interaction, the effects of each treatment (the differences between groups) are constant at each level and the lines are roughly parallel.

**ORDINAL INTERACTION** The effects of each treatment are not constant and thus the lines are not parallel. The differences for red are large, but they decline slightly for blue cereal and even more for green cereal. Thus, the differences by color vary across the shapes. The relative ordering among levels of shape are the same, however, with stars always highest, followed by the cubes and then the ball shapes. Statistical tests could assess whether the differences between shapes are significant at each color.

**DISORDINAL INTERACTION** The differences in color vary not only in magnitude but also in direction. This interaction is shown by lines that are not parallel and that cross between levels. The evaluation of balls is higher than cubes and stars for red and blue, but is evaluated lower than both for the color green.

The graphs complement the statistical significance tests [93] by enabling the researcher to quickly categorize the interaction, especially determining whether significant interactions fall into the ordinal or disordinal categories.



**Figure 6.13**  
Interaction Effects in Factorial Designs

## Interpreting Covariates and Interaction Effects

When covariates are involved in a MANOVA model:

Analyze the model both with and without the covariates.

If the covariates do not improve the statistical power or have no effect on the significance of the treatment effects, then they can be dropped from the final analysis.

Any time two or more independent variables (treatments) are included in the analysis, interactions must be examined before drawing conclusions about main effects for any independent variable.

If the interactions are not statistically significant, then main effects can be interpreted directly because the difference between treatments is considered constant across combinations of levels.

If the interaction is statistically significant and the differences are not constant across combinations of levels, then the interaction must be determined to be ordinal or disordinal.

Ordinal interactions mean that the direction of differences does not vary by level (e.g., males always less than females) even though the difference between males/females varies by level on the other treatment; in this case, the size of the main effect (e.g., males versus females) should only be described separately for each level of the other treatment.

Significant disordinal interactions occur when the direction of an observed main effect changes with the level of another treatment (e.g., males greater than females for one level and less than females for another level); disordinal interactions interfere with the interpretation of main effects.

## IDENTIFYING DIFFERENCES BETWEEN INDIVIDUAL GROUPS

Although the univariate and multivariate tests of ANOVA and MANOVA enable us to reject the null hypothesis that the groups' means are all equal, they do not pinpoint where the significant differences lie among more than two groups. Multiple *t* tests without any form of adjustment are not appropriate for testing the significance of differences between the means of paired groups because the probability of a Type I error increases with the number of intergroup comparisons made (similar to the problem of using multiple univariate ANOVAs versus MANOVA). Many procedures are available for further investigation of specific group mean differences of interest using different approaches to control Type I error rates across multiple tests.

**Multiple Univariate Tests Adjusting for the Experiment-wide Error Rate** Many times the simplest approach is to perform a series of univariate tests with some form of manual adjustment by the researcher to account for the experimentwide error rate. Researchers can make these adjustments based on whether the treatments involve two or more levels (groups).

**TWO-GROUP ANALYSES** Two-group treatments reduce to a series of *t* tests across the specified dependent measures. However, researchers should be aware that as the number of these tests increases, one of the major benefits of the multivariate approach to significance testing—control of the Type I error rate—is negated unless specific adjustments in the  $T^2$  statistic are made that control for the inflation of the Type I error.

If we wish to test the group differences individually for each of the dependent variables, we could use the square root of  $T_{\text{crit}}^2$  (i.e.,  $T_{\text{crit}}$ ) as the critical value needed to establish significance. This procedure would ensure that the probability of any Type I error across all the tests would be held to  $\alpha$  (where  $\alpha$  is specified in the calculation of  $T_{\text{crit}}^2$ ) [43].

***k*-GROUP ANALYSES** We could make similar tests for *k*-group situations by adjusting the  $\alpha$  level by the **Bonferroni inequality**, which adjusts the alpha level for the number of tests being made. The adjusted alpha level used in any separate test is defined as the overall alpha level divided by the number of tests [adjusted  $\alpha = (\text{overall } \alpha) / (\text{number of tests})$ ].

For example, if the overall error rate ( $\alpha$ ) is .05 and five statistical tests are to be made, then a Bonferroni adjustment would call for a .01 level to be used for each individual test.

**Structured Multigroup Tests** The procedures described in the previous section are best used in simple situations with a few tests being considered. If the researcher wants to systematically examine group differences across specific pairs for one or more dependent measures, more structured statistical tests should be used. In this section we will examine two types of tests:

- *Post hoc tests.* Tests of the dependent variables between all possible pairs of group differences that are tested after the data patterns are established.
- *A priori tests.* Tests planned from a theoretical or practical decision-making viewpoint prior to looking at the data.

The principal distinction between the two types of tests is that the post hoc approach tests all possible combinations, providing a simple means of group comparisons but at the expense of lower power. A priori tests examine only specified comparisons, so that the researcher must explicitly define the comparison to be made, but with a resulting greater level of power. Either method can be used to examine one or more group differences, although the a priori tests also provide the researcher with total control over the types of comparisons made between groups.

**POST HOC METHODS** Post hoc methods are widely used because of the ease in which multiple comparisons are made. Among the more common post hoc procedures are (1) the Scheffé method, (2) Tukey's honestly significant difference (HSD) method, (3) Tukey's extension of the Fisher least significant difference (LSD) approach, (4) Duncan's multiple-range test, and (5) the Newman–Keuls test.

Each method identifies which comparisons among groups (e.g., group 1 versus group 2, group 1 versus group 3, etc.) have significant differences. Although they simplify the identification of group differences, these methods all share the problem of having quite low levels of power for any individual test because they examine all possible combinations. These five post hoc or multiple-comparison tests of significance have been contrasted for power [110] and several conclusions can be drawn:

- Scheffé method is the most conservative with respect to Type I error, and the remaining tests are ranked in this order: Tukey HSD, Tukey LSD, Newman–Keuls, and Duncan.
- If the effect sizes are large or the number of groups is small, the post hoc methods may identify the group differences. However, the researcher must recognize the limitations of these methods and employ other methods if more specific comparisons can be identified.

A discussion of the options available with each method is beyond the scope of this chapter. Excellent discussions and explanations of these procedures can be found in other texts [52, 126].

**A PRIORI OR PLANNED COMPARISONS** The researcher can also make specific comparisons between groups by using a priori tests (also known as **planned comparisons**). This method is similar to the post hoc tests in the statistical methods for making the group comparisons, but differs in design and control by the researcher in three aspects:

- The researcher specifies which group comparisons are to be made versus testing the entire set, as done in the post hoc tests.
- Planned comparisons are more powerful because of the smaller number of comparisons, but more power is of little use if the researcher does not specifically test for correct group comparisons.
- Planned comparisons are most appropriate when conceptual bases can support the specific comparisons to be made. They should not be used in an exploratory manner, however, because they do not have effective controls against inflating the experimentwide Type I error levels.

The researcher specifies the groups to be compared through a **contrast**, which is a combination of group means that represents a specific planned comparison. Contrasts can be stated generally as:

$$C = W_1G_1 + W_2G_2 + \cdots + W_kG_k$$

where:

$C$  = contrast value

$W$  = weights

$G$  = group means

The contrast is formulated by assigning positive and negative weights to specify the groups to be compared while ensuring that the weights sum to 0.

For example, assume we have three group means ( $G_1$ ,  $G_2$ , and  $G_3$ ). To test for a difference between  $G_1$  and  $G_2$  (and ignoring  $G_3$  for this comparison), the contrast would be:

$$C = (1)G_1 + (-1)G_2 + (0)G_3$$

To test whether the average of  $G_1$  and  $G_2$  differs from  $G_3$ , the contrast is:

$$C = (.5)G_1 + (.5)G_2 + (-1)G_3$$

A separate  $F$  statistic is computed for each contrast.

In this manner, the researcher can create any comparisons desired and test them directly, but the probability of a Type I error for each a priori comparison is equal to  $\alpha$ . Thus, several planned comparisons will inflate the overall Type I error level. All the statistical packages can perform either a priori or post hoc tests for single dependent variables or the variate.

If the researcher wishes to perform comparisons of the entire dependent variate, extensions of these methods are available. After concluding that the group mean vectors are not equivalent, the researcher might be interested in whether any group differences occur on the composite dependent variate. A standard ANOVA  $F$  statistic can be calculated and compared to  $F_{\text{crit}} = (N - k)\text{gcr}_{\text{crit}}/(k - 1)$ , where the value of  $\text{gcr}_{\text{crit}}$  is taken from the gcr distribution with appropriate degrees of freedom. Many software packages have the ability to perform planned comparisons for the dependent variate as well as individual dependent variables.

## ASSESSING SIGNIFICANCE FOR INDIVIDUAL OUTCOME VARIABLES

Up to this time we have examined only the multivariate tests of significance for the collective set of outcome or dependent variables. What about each separate outcome? Does a significant difference with a multivariate test ensure that each outcome also is significantly different? Or does a nonsignificant effect mean that all outcomes also have nonsignificant differences? In both instances the answer is no. The result of a multivariate test of differences across the set of outcomes does not necessarily extend to each variable separately, just collectively. Thus, the researcher should always examine the multivariate results for the extent to which they extend to the individual outcome measures. The three most widely used approaches for examining the individual outcomes are discussed below.

**Univariate Significance Tests** The first step is to assess which of the outcomes contribute to the overall differences indicated by the statistical tests. This step is essential because a subset of variables in the set of outcomes may accentuate the differences, whereas another subset of outcomes may be nonsignificant or may mask the significant effects of the remainder.

Most statistical packages provide separate univariate significance tests for each outcome in addition to the multivariate tests, providing an individual assessment of each variable. The researcher can then determine how each individual outcome corresponds to the effects on the variate. To do so, the researcher should manually adjust the required significance level (e.g., Bonferroni adjustment) for each of the individual tests.

In using this approach, researchers should be cautious since using separate tests does not take into account the correlation among outcomes that underlie the multivariate test. In certain situations the univariate test can indicate counter-intuitive findings, such as the instance in which a significant overall test is shown to have no significant individual tests [50].

**Discriminant Analysis** We discussed earlier how MANOVA is the “reverse” of discriminant analysis and that they share the same multivariate significance tests of the estimated discriminant functions. In this instance, we use discriminant analysis to identify the outcome(s) that discriminate between groups on each treatment variable [9, 7, 121]. To do so, designate a treatment as the dependent variable in the discriminant analysis and the outcomes as the independent variables. Then the multivariate significance tests will (a) test for the significance of the set of outcomes for that treatment and (b) identify which of the outcomes are predictive of the difference. In this manner all of the outcomes can be assessed simultaneously while taking into account the correlations among them, something the univariate tests do not account for.

This approach requires, however, extensive analysis of the derived discriminant functions to assess which of the outcomes were included in each discriminant function. As a result, this approach is less widely used, although development of an index (Parallel Discriminant Ratio Coefficients) to assess the relative contribution of each independent variable may simplify the interpretation process [114].

**Stepdown Analysis** A procedure known as **stepdown analysis** [68, 110] may also be used to assess individually the differences of the dependent variables. This procedure involves computing a univariate  $F$  statistic for a dependent variable after eliminating the effects of other dependent variables preceding it in the analysis. The procedure is somewhat similar to stepwise regression, but here we examine whether a particular dependent variable contributes unique (uncorrelated) information on group differences. The stepdown results would be exactly the same as performing a covariate analysis, with the other preceding dependent variables used as the covariates.

A critical assumption of stepdown analysis is that the researcher must know the order in which the dependent variables should be entered, because the interpretations can vary dramatically given different entry orders. If the ordering has theoretical support, then the stepdown test is valid. Variables indicated to be nonsignificant are redundant with the earlier significant variables, and they add no further information concerning differences about the groups. The order of dependent variables may be changed to test whether the effects of variables are either redundant or unique, but the process becomes rather complicated as the number of dependent variables increases.

Tonidandel and LeBreton [115] have extended stepdown analysis through a variant of the relative importance weights developed for multiple regression. Relative importance weights are derived for each outcome and act as relative effect sizes, representing the proportion of the overall effect attributable to each outcome. Comparisons to

## Interpreting Differences Between Individual Groups and Individual Outcomes

When the independent variable has more than two groups, two types of procedures can be used to isolate the source of differences:

Post hoc tests examine potential statistical differences among all possible combinations of group means; post hoc tests have limited power and thus are best suited to identify large effects.

Planned comparisons are appropriate when *a priori* theoretical reasons suggest that certain groups will differ from another group or other groups; Type I error is inflated as the number of planned comparisons increases.

With the multivariate tests only addressing the differences of the collective set of dependent variables, researchers should also examine if those differences are found for each dependent variable separately.

Most widely used test is univariate ANOVA for each dependent measure

Discriminant analysis can provide insights as to dimensions of differences, particularly if there are more than two groups.

Stepdown analysis is powerful analysis if causal ordering is known for dependent variables.

the other approaches (univariate ANOVAs and discriminant analysis) show comparative advantages for this method [115] and when software code becomes more widely available this approach will provide researchers with another measure of relative importance.

## INTERPRETING MEDIATION AND MODERATION

In addition to the results directly from MANOVA, the two additional relationships (mediation and moderation) provide not only a statistical indicator of the impact of these third variables on the main effect, but also substantive insight into “How” and “When” issues. The following discussions focus on portraying the actual impact of these relationships (i.e., how large is the mediation effect or how much does the interaction change the group means).

**Mediation** The mediation relationship decomposes the original main effect into two elements: the mediated main effect and the indirect (mediation) effect. We can represent this as:

$$\begin{array}{ccc} C & = & C' + A*B \\ \text{Original} & & \text{Mediated} & \text{Indirect} \\ \text{main effect} & & \text{main effect} & \text{(mediation) effect} \end{array}$$

In interpreting the mediation effect, several questions must be addressed:

- 1 Was the indirect effect significant?
- 2 Is the mediated main effect significant?
- 3 If both mediated main effect and indirect effect are significant, what is their relative strength?

Let's now address each of these issues.

**SIGNIFICANCE OF INDIRECT MEDIATION EFFECT** The first issue relates to the significance of the indirect effect and, if significant, how that supports the research question. Statistical significance can be determined through the Sobel test, or as proposed more recently, the use of bootstrapping to determine confidence levels and significance of the effect.

In either approach the researcher is determining whether the indirect effect provides a substantive “Why” explanation for the original main effect. One should note that while there are statistical criteria, ultimately it is the conceptual support that provides the rationale for a mediation effect versus some alternative explanation (e.g., a confounder which is equivalent empirically but lacks the causal ordering of the mediation effect). Thus, as stated before, identifying a significant mediation effect requires first conceptual support and then statistical verification [84].

**SIGNIFICANCE OF MEDIATED MAIN EFFECT** In addition to the indirect effect, the significance of the mediated main effect provides a means to determine if the original main effect is completely explained/replaced by the indirect effect (i.e., complete mediation) or if a significant portion still remains (i.e., partial mediation). A nonsignificant mediated main effect empirically supports complete mediation and the ability of the indirect effect to provide a “Why” alternative explanation for the main effect.

But the researcher is more interested in the interpretation of both the mediated main effect and the indirect effect in combination. To this end, useful discussions of the various interpretations are available [128] and Zhao et al. even provide a flowchart approach to understanding how the resulting effects can be interpreted [130]. Recent research cautions researchers on making the distinctions of complete versus partial mediation due to a number of alternative impacts on the estimated effects [45]. Thus, there is a substantial and ever growing literature base on mediation across numerous disciplines and the interested reader is encouraged to explore their discipline for its treatment of mediation [e.g., 83].

**RELATIVE STRENGTH OF INDIRECT EFFECT** In the situation of partial mediation, with both the indirect and mediated main effects achieving statistical significance, the remaining questions involve the relative strength of the indirect effect versus either the total effect or the remaining mediated main effect. To this end, the PROCESS macro provides

three measures of relative impact: ratio of indirect to total effect, the ratio of the indirect effect to the direct effect and a measure of effect size for the indirect effect [44]. Bootstrapped estimates of the confidence intervals are also provided to assess their statistical significance. While there are now only characterizations of high, medium or low for these measures, they do provide some measures of relative size that can be used as common metrics for comparison purposes.

**COMBINED EFFECTS** Finally, there are combinations of mediation, notably moderated mediation and mediated moderation. While both of these are technically possible, Hayes contends the moderated mediation is the more appropriate and interpretable approach for examining these more complex effects [44]. As with other issues in mediation effects, the interested reader is encouraged to monitor advances in both the estimation and interpretation of these issues.

**Moderation** Moderation provides information concerning the extent to which a main effect may be contingent on a third variable—i.e., how the main effect varies for different values of the moderator. So it addresses the “When” question regarding a main effect. As with mediation, a moderating effect must have conceptual support since it is implemented in the same manner as the interaction terms accompanying factorial designs. Yet interactions are based on an empirical requirement for interpreting a main effect, but do not require any conceptual support for their calculation.

To better understand the interaction effect and how it reflects the group differences based on the moderator values, we can express a moderated model in linear model form (see Figure 6.14). In the unmoderated form,  $b_1$  represents the mean difference between treatment groups and the group mean for any group is calculated by adding in the constant. In the moderated situation, the moderator adds to both the intercept and group difference when the moderator effect is in effect. So when the moderator is zero (no interaction effect), then the group means are calculated similar to the unmoderated case ( $b_0$  versus  $b_0 + b_1$ ). But when the moderator is in effect, the moderator adds to both the intercept ( $b_2$ ) and the group differences ( $b_3$ ). The  $b_2$  term shifts both moderated groups, but the interaction effect ( $b_3$ ) is unique to the moderated treatment group (lower right cell) and creates the nonparallel lines characteristic of an interaction. We will illustrate these relationships in our later examples of applying moderation to one of the main effects of interest.

**SUMMARY** Mediation and moderation have become integral components of experimental research because of their ability to extend the research question beyond just the identification of an effect, but also through insights about the “Why” and “When” impacting the main effect. We have provided just an overview of the issues and applications of mediation and strongly recommend any research employing ANOVA/MANOVA models strongly consider mediation and moderation for their theoretical extensions of the treatment → outcome relationship.

## **Figure 6.14** Calculating Cell Means with a Moderation Effect

## Unmoderated Linear Model

$$Y = b_0 + b_1 X$$

## Moderated Linear Model

$$Y = b_0 + b_1 X + b_2 Z + b_3 ZX$$

$$Y = (b_0 + b_2 Z) + (b_1 + b_3 Z)X$$

combined  
intercept

combined  
difference

Calculating Cell Means With Moderation		
	Moderator	
	Z = 0	Z = 1
Treatment	X = 0	$b_0$
	X = 1	$b_0 + b_1$

## Interpreting Mediation and Moderation

Mediation is represented by the indirect effect formed from two causal paths: treatment → mediator and mediator → outcome.

Indirect effect represents the "Why" behind a main effect by providing a construct (i.e., the mediator) through which the treatment operates to create the effect in the outcome.

Complete mediation occurs when the mediation effect (indirect effect) totally accounts for the original main effect.

Partial mediation occurs when the indirect effect is significant, but there still remains a significant main effect even after accounting for the indirect effect.

Moderation occurs when the main effect varies based on different values of the moderator.

An example would be the main effect being different between males and females

Significant moderating effects exhibit patterns comparable to significant interactions (e.g., nonparallel lines)

## Stage 6: Validation of the Results

Analysis of variance techniques (ANOVA and MANOVA) were developed in the tradition of experimentation, with **replication** as the primary means of validation. The specificity of experimental treatments allows for a widespread use of the same experiment in multiple populations to assess the generalizability of the results. Although it is a principal tenet of the scientific method, in social science and business research, true experimentation is many times replaced with statistical tests in non-experimental situations such as survey research. The ability to validate the results in these situations is based on the replicability of the treatments. In many instances, demographic characteristics such as age, gender, income, and the like are used as treatments. These treatments may seem to meet the requirement of comparability, but the researcher must ensure that the additional element of randomized assignment to a cell is also met; however, many times in survey research randomness is not fully achieved.

For example, having age and gender be the independent variables is a common example of the use of ANOVA or MANOVA in survey research. In terms of validation, the researcher must be wary of analyzing multiple populations and comparing results as the sole proof of validity. Because respondents in a simple sense select themselves, the treatments in this case cannot be assigned by the researcher, and thus randomized assignment is impossible.

The researcher should strongly consider the use of covariates to control for other features that might be characteristic of the age or gender groups that could affect the dependent variables but are not included in the analysis.

Another issue is the claim of causation when experimental methods or techniques are employed. For our purposes here, the researcher must remember that in all research settings, including experiments, certain conceptual criteria (e.g., temporal ordering of effects and outcomes) must be established before causation may be supported. The single application of a particular technique used in an experimental setting does not ensure causation. The following section, however, provides some additional advanced statistical tools that can be used in situations where a controlled experiment is not possible, but causal inferences are desired.

## Advanced Issues: Causal Inference in Nonrandomized Situations

Up to this point in the chapter we have generally referred to the notion of causality in general terms. Yet the conditions in which a researcher can make truly causal statements is quite restricted. Identification of **causal effects** is the ultimate objective of most research, and an ideal setting would allow us to make such statements as: *When the treatment is applied,*

*the outcome always happens, or The treatment is the only reason for the observed change in the outcome.* While areas like the physical sciences many times follow these principles, most other areas, particularly those involving individuals (e.g., medical or behavioral sciences) would most likely be cautious in making these types of statements with their research.

So does this mean that only tightly controlled studies in the physical sciences can make causal statements? What about human behavior, which we do not feel follows the same rigorous rules as nature. And what about all of the non-experimental data that has been gathered over time and the tremendous influx of new data available in this age of Big Data? None of this data was collected in the tightly controlled experimental setting, yet it has potential value in being “real world” data that reflects actual outcomes experienced outside the laboratory setting. It is for these situations that a relatively new class of models are becoming available to help overcome the comparability issues of treatment and control groups found in observational studies.

In the following sections we will provide a brief overview of efforts to establish causality in the behavioral sciences. Particular emphasis will be placed on the conceptual framework introduced by Rubin [101] termed potential outcomes that underlies today’s notions about causality, especially in how to treat observational data.

## CAUSALITY IN THE SOCIAL AND BEHAVIORAL SCIENCES

The classic experimental design, the controlled randomized experiment, involves a number of very strict conditions: manipulation of the treatment by the researcher, random assignment of respondents to the treatment and control groups, and a tightly controlled setting to minimize any additional confounding influences on the experiment. If all of these conditions are met, a **causal inference** can be made where the observed differences between the treatment and control groups can be attributed to the treatment as a causal effect.

Research in the behavioral and social sciences, however, rarely utilizes the controlled randomized experiment and employs some form of non-experimental design. One fundamental reason is that in many instances the data is already available and analyzing what actually happened is considered more appropriate for many types of applied research settings. Also, many times it is not possible to manipulate the treatment among participants, such as in the case of risk factors (e.g., smoking or obesity), social impacts (e.g., segregation, poverty), personal choices (e.g., marriage at a young age), and personal characteristics (e.g., gender differences). Yet all of these relate to important research questions that impact all facets of our society. So the quest for making causal inferences in these areas remains strong, even though what constitutes causality in the social and behavioral sciences has long been a topic of discussion [48, 103, 92, 63, 108].

There have been two major research streams that have focused on bringing the conditions of causality to the social and behavioral sciences [123, 102, 23, 58]. The first involves Donald Campbell who focused on the internal validity of research designs—finding threats to causality and then remedies for those threats. The concept of confounds became instrumental in this work and the research design elements to eliminate them became the tenets of behavioral research [103, 13]. The second perspective is from Donald Rubin and is termed the potential outcomes approach [97, 98, 48]. In this approach, the ideal situation for making causal inference would be for the respondent to be both *exposed to the treatment* and *not exposed to the treatment* under exactly the same conditions. These are the potential outcomes, yet in reality we can only observe one of them (either exposure or no exposure) and the other is missing. The focus of this research was to make non-experimental data conform to the conditions of experimental research, primarily through procedures to create comparability among the observed treatment and control groups. If these two groups could be made as comparable as was achieved through the random assignment process, then causal inferences could be made. It is perhaps not surprising that Rubin is also considered one of the pioneers in missing data treatments and analysis.

While the work of Campbell and others has advanced research design dramatically, it is the work of Rubin that has had the biggest impact on the use of observational studies and the ability of researchers to potentially overcome the “correlation does not equal causation” caveat in much of social and behavioral research. The development of a process for processing and forming comparable treatment and control group in these non-experimental settings is the focus of the remainder of this section. We will first review in more detail the basic principles of the potential outcomes approach and then examine one of the more widely employed techniques, propensity score matching, and its application to observational data for making causal inferences.

## THE POTENTIAL OUTCOMES APPROACH

As noted earlier, Rubin's work is based on the basic premise that causal inference will always have a "missing data problem" in that both potential outcomes cannot be observed in what is termed the "fundamental problem of causal inference" [48]. We can view these **potential outcomes** (i.e., outcomes whether exposed to treatment or not) as counterfactuals of each other. A **counterfactual** is the "missing" outcome for each individual. If we could develop a method for "filling in" these counterfactuals, just as we do with missing data, then we could make causal inferences.

We can see this in the diagram below. If an individual could be observed for both potential outcomes, then all four cells would be filled. That is not possible, however, so to make a causal inference we need to generate counterfactuals (e.g., what would be the outcome for individuals assigned to the control group if they had received the treatment and vice versa for the treatment group).

		Potential Outcomes	
		Control	Treatment
Assignment Process	Control	Control Group (observed)	Missing
	Treatment	Missing	Treatment Group (observed)

Rubin's work does not treat randomized experiments differently from non-experimental designs. It just acknowledges that randomized experiments account for the potential outcomes and counterfactuals are generated in a specific and efficient manner. The random assignment process assumes that the resulting groups (i.e., treatment and control) are identical on any other characteristic influencing the causal effect. In a general sense, if the two samples are large enough, for each person in the control group there should be an equivalent person in the treatment group. If this is the case, then the treatment and control groups can act as the counterfactuals necessary for causal inference (e.g., the treatment group is equivalent to those control group participants who did not receive the treatment on all other characteristics). The only assumption in the process is the **stable unit treatment value assumption** (SUTVA) where the treatment does not change across individuals and the assignment process does not impact the outcomes (i.e., one person's treatment does not impact or interfere with another person's outcomes) [59].

## COUNTERFACTUALS IN NON-EXPERIMENTAL RESEARCH DESIGNS

While randomized experiments have always been considered the "gold standard" for causal inference, the many issues precluding the use of experiments discussed earlier have generated intense interest in overcoming these limitations and resulted in many excellent discussions on making causal inferences in non-experimental situations [46, 88, 48, 59]. It is beyond the scope of this discussion to address all of these issues, but several key issues emerge:

**Similarity of Treatment and Control Groups** Just as the random assignment process was developed to achieve similarity between groups, any causal inference in non-experimental settings requires treatment and control groups that are as similar as possible on all possible influences of the outcome, other than the treatment. If they are not similar, then we cannot be sure that our "apples to oranges" comparison is valid.

**Confounding Variables** As discussed earlier in Figure 6.8, a **confounder** is a third variable that relates to both treatment and outcome, but is not included in the analysis in some form. Many times termed an "unobserved common cause," the presence of confounders biases the estimate of the main effect. Confounders are a principal cause of endogeneity along with measurement error and simultaneity [4].

**Assumption of the Strongly Ignorable Treatment Assignment** The **assumption of the strongly ignorable treatment assignment**, similar to the notion of ignorability in missing data processes discussed in Chapter 2, implies that all confounding or sources of endogeneity can be eliminated if there is equivalence on the set of potential

confounding characteristics of the treatment and control groups. Confounding or endogeneity occurs when the assignment of the treatment is not independent of the outcome, and is due to omitted variables that relate to both the treatment and the outcome. This assumption allows for a procedure to provide ignorability of endogenous effects given the appropriate identification and treatment of the potential confounding effects.

**Balance** **Balance** refers to the equivalence of the treatment and control groups on a set of variables. For the random assignment process, tests for balance should demonstrate the equivalence of the treatment and control groups on any set of characteristics that might impact the outcome (e.g., potential confounders).

Rubin not only developed the framework of potential outcomes to provide a generalized context suitable to both experiments and non-experiments alike, but then proceeded to formally propose a procedure which could meet all of the above issues equally as well, in theory, as randomized experiments. The objective is simple in concept—correct for any biases introduced in the assignment process in the non-experimental setting by finding observations in the control group that are similar enough to observations in the treatment group to meet the ignorability assumption.

The following section will describe one such method—propensity scoring—which has become widely employed in the analysis of observational data in many disciplines. As we will discuss, it employs one of the other multivariate techniques covered in this text—logistic regression—as the mechanism of comparing two groups (treatment versus control) on a large number of variables that may act as confounders.

## PROPENSITY SCORE MODELS

The **propensity scoring model** in its many forms is one of the primary approaches for estimating the counterfactuals which meet the ignorability assumption [88, 40, 57]. Our discussion of propensity scoring models is not an endorsement per se of this method, but provides a discussion that extends another multivariate technique (logistic regression) into the domain of causal inference. Interested readers are encouraged to review the coverage of logistic regression in Chapter 8 if they intend to employ propensity scoring models.

**Popularity of Propensity Score Models** The concept of propensity score models has been widely supported from numerous perspectives [112, 50, 37, 98, 95] and has achieved widespread use in such areas as political science [56] and the health sciences [36, 120, 47]. Within the business disciplines, there have been numerous calls for increased use in management [75], information systems [16] and marketing [82], with accounting embracing the approach in a number of research areas [105, 71, 27, 109]. In all of these instances the research contexts address “real world” issues from a wide array of secondary data (survey data, census data, medical records, financial reports, etc.) to address any number of research questions. As a result, an entire range of data sources have now been made available for even more rigorous analytical approaches that provide the call for evidence-based decisions with more reliable information for decision-making.

**Developing and Applying a Propensity Score Model** While a propensity score model may appear to be an “empirical silver bullet” for solving issues in the use of non-experimental data, the researcher will find that the ultimate success or failure of the process is based on conceptual considerations, primarily the selection of the set of confounders to be included in the model. Once that set of confounders is identified, then logistic regression is used to estimate the propensity scores and these scores are then used to select (match) comparable observations from the treatment and control groups for further analysis. It should be noted that research design options can be utilized to increase the efficacy of these models and thus not rely solely on the adjustment process [96, 100, 54, 99, 103]. This entire process is covered in more detail with software examples [40] and in review articles detailing more detailed issues [112, 73, 75]. Common to these discussions are a six-step procedure for specifying and then applying a propensity score model to non-experimental data.

**STEP 1: SPECIFY COVARIATES** As noted several times, a key assumption of making causal inferences in non-experimental situations is the ability to meet the strong ignorability assumption. Variable selection is a critical phase in this process that must be driven by conceptual considerations since no form of adjustment can

overcome omission of relevant measureable confounders. As a general rule, inclusion can be made for any variable that is either a cause of a treatment or the outcome or both [118]. But as we will see, there are specific issues within these general recommendations. The variable types depicted in Figure 6.8 will be used to understand which of these variable types should be considered for inclusion. We should note that the literature on propensity score models uses the term covariates as a general category of all variables, of any type, to be included in the model. After our discussion here of various types, we will use this general term for the entire set of variables, regardless of their depiction in Figure 6.8.

**Confounding** The most fundamental type of variable for inclusion is the confounder, as it is a direct source of endogeneity that violates the ignorability assumption. Confounders, as shown in Figure 6.8, cause both the treatment and outcome and act as a “common cause” that must be included in the covariates of the propensity score model. A primary concern with confounders are those that are unmeasured or unobserved as while they impact the treatment→outcome, they cannot be included in the model. In this regard researchers are encouraged to expand their set of covariates to include variables which may be related to the unobserved confounders and thus incorporate some of their effect in the model. So while the researcher must be judicious in the confounders identified for inclusion, the bias resulting from excluding impactful confounders or the chance of not finding correlates of the unmeasured confounders encourages the research to err on the side of inclusion. Variables with weak relationships to the treatment or outcome will be minimized in the model results with little impact on the propensity scores.

**Covariates/Blocking Factors** These variables, related to only the outcome, can be included as they (a) provide increased precision without bias [10] and (b) provide additional adjustments for unmeasured confounders. In terms of estimation they operate as they did in ANOVA/MANOVA estimation to reduce variance in the outcome measure. They also can provide additional controls on any confounders that may operate through them in impacting the outcome measure.

**Instrumental Variables** The instrumental variable is related to the treatment but not the outcome. While some research has recommended inclusion, other research has urged caution. Recent research has demonstrated decreases in the efficiency of the causal effects, and in the presence of unmeasured confounding, increases the bias of the causal effects [127, 11]. Moreover, inclusion of variables which are highly related to the treatment have their own confounding effect as to what does the causal effect stem from (i.e., what is the “true” cause – the treatment or the variable that highly predicts the treatment) [112]. In addition to the inability to determine the “true” effect, it also makes it impossible to achieve balance on the highly related variable [73, 12]. As a result, inclusion of instrumental variables in the covariate set should be done only for specific conceptual reasons and then only after consideration of the potential negative effects.

**Mediators** Mediators are related to both treatment and outcome, but their causal relationship to the main effect calls for their exclusion as a model covariate. In general, any variable that is caused by the treatment should be excluded as they represent potential indirect effects of the treatment [38]. For purposes of estimating the causal effect we only need to measure the total effect and inclusion of indirect effects in the covariate set would bias downward the causal effect estimate.

**Moderator** The final type of variable is the moderator, a third variable which changes the magnitude and/or direction of the main effect. Given its relationship to both treatment and outcome it can be included in the covariate set, but only acting as a confounder, not a moderator. While recent research has explored moderation of the propensity score model [32], the most direct approach is to apply the moderating effect in the estimation procedures, such as done with doubly robust regression [235].

**Overall Selection Strategy** Unless limited by sample size, the researcher should specify a covariate set including all of the known confounders and other variables related to the outcome. Theory, domain knowledge and even field research should guide the selection process. Parsimony is not paramount in the covariate set since prediction is the sole purpose. This does not allow the research to indiscriminately include variables, as there are potential issues of

balance and thus ignorability that can result. But as noted earlier, excluding a relevant variable is much more impactful than including an irrelevant variable.

One quite useful framework for identifying all of the potential confounders is the use of the **DAG (directed acyclic graphs)** that formally details all causal relationships of interest [92, 118]. The DAG has become a topic of interest as a framework in which all of the “threats” to a causal effect can be identified. Pearl [92] has popularized this concept and developed an entire perspective on causality. While it is beyond the scope of this chapter to review this perspective, researchers interested in causal inferences must consider the method and underlying theory.

**STEP 2: ESTIMATE PROPENSITY SCORE MODEL** The fundamental purpose of the propensity score model is to fully represent the complete set of covariates which achieving balance (i.e., comparability) between the treatment and control groups. Specifying the complete set of confounds in the covariate set provides the information needed to satisfy the ignorability assumption. However, to meet this assumption the two groups must be equal on all of the covariates. As the number of covariates increase in number, finding a way to represent all of these covariates in a single value becomes essential. This is the role of the logistic regression model, where the covariate sets are the independent variables and the treatment variable is the dependent variable. Note that we are not predicting the outcome, but instead the treatment variable.

The predicted probability now becomes the **propensity score**, such that observations with similar predicted probability scores should have similar profiles on the covariate set while being in either the treatment or control group. The purpose of the propensity score model is not prediction of the treatment, but rather a method for estimating a variate from the covariates that can be balanced between the treatment and control groups. Thus, classification accuracy, model fit or even variable collinearity is not important, with the focus instead on achieving balance across the covariates [112]. A series of model estimation strategies are discussed in Guo and Fraser [40], with all aimed at achieving balance of the covariate set and model overlap discussed in the next step.

**STEP 3: ESTIMATE BALANCE OF COVARIATE SET AND MODEL OVERLAP** Before employing the propensity score in the adjustment process, it must be checked for model overlap and covariate balance. In each situation the focus is on the comparability of the treatment and control groups.

**Model Overlap** **Model overlap** is the comparison of the range of propensity scores between treatment and control groups. If the minimum and maximum propensity scores were identical for both treatment and control groups, then there would be complete overlap. But as is more often the case, one or the other groups has a minimum or maximum score that exceeds the other group. A high degree of overlap is critical since only observations with comparable propensity scores can be included in the analysis. Also, the lower the overlap the smaller the extent of comparison between the two groups, potentially affecting the validity of the causal estimate. At the highest or lowest propensity scores where there is no overlap between groups some observations may be eliminated before proceeding. While this reduces somewhat the external validity, it can improve the balance and matching done later in the analysis.

**Covariate Balance** As noted earlier, the sole objective of the logistic regression model is to develop a predicted probability that is based on a set of balanced covariates used as independent variables. To this end, the researcher evaluates the logistic model on the balance it achieves, where balance is the degree of equivalence between groups on each of the covariates. Balance is measured within subgroups of observations rather than a single overall measure. Most often the observations are divided into quintiles based on the propensity score (i.e., highest 20 percent of treatment and control group in first quintile, etc.). Then significance tests are performed on each of the covariates within each of the subgroups. When balance is not achieved, any number of actions may be taken, including combining or expanding subgroups, transformations of the covariate or the addition/deletion of covariate polynomials or interactions between the covariates [73]. Researchers must be cautious, however, to exclude irrelevant covariates as balancing on these covariates may impact the balance of other covariates.

After any changes for covariate imbalance, the logistic model is re-estimated and balance checks performed again. This iterative process continues until balance (i.e., nonsignificant differences between treatment and control groups) is achieved for each covariate in each subgroup. *Remember, covariate balance is the critical element in propensity score*

models since it is the balance/equivalence of the treatment and control groups across the entire set of covariates that meets the ignorability assumption and allows for causal estimates to be made.

**STEP 4: APPLICATION OF PROPENSITY SCORES** Once the propensity scores achieve model overlap and balance, they can be used to create comparability between the treatment and control groups. While the ideal situation would be to have exact matches where the observations are exactly equal on all covariates, that becomes unrealistic as the covariate set increases. So some form of matching must be employed to group together very similar observations. The three most widely used approaches for identifying equivalent observations are matching, stratification and weighting [40]. There are numerous options within each of these approaches, so we will only review the basic elements in each area and the interested reader can explore the details as needed. But each approach employs the same principle. That is, observations with equal propensity scores are now (a) assumed to be comparable across the entire covariate set, (b) so that direct comparisons between matched treatment and control members is an estimate of causal effect.

**Matching** Matching is perhaps the approach most commonly associated with propensity score models since it operates at the observation level. In this approach observations in each group are “matched” with observations from the other group with equal or very similar propensity scores. There are numerous methods which (a) vary how precisely the observations must match and (b) match observations in pairs (i.e., one to one) versus one observation versus all other that match (i.e., one to many). These and other options provide the researcher a wide range of approaches that can best fit the specific dataset under consideration. Note that this approach has the highest chance of sample reduction since observations from either the treatment or control group that cannot be matched must be eliminated from the analysis. Researchers should always monitor the extent of sample reduction using this approach.

**Stratification** A second approach involves stratification where subgroups are formed, similar to what was done in testing for balance. The number of strata can vary based on the degree of equivalence desired in each stratum. Some approaches employ a smaller number of strata (e.g., 5 to 10), while other approaches, such as coarsened exact matching (CEM) expands and contracts the strata to achieve the best balance among the treatment/control matches [53].

**Weighting** This final approach is to apply the propensity scores directly as weights in the linear model comparison of treatment versus control groups [112]. The most common method is to use the inverse of the propensity score (**inverse probability of treatment weighting**, IPTW). Observations are not grouped or stratified, with weighting now adjusting each treatment and control group for equivalence. The primary disadvantage of this approach is that very high or low propensity scores result in extremely large or small weights that may distort the analysis. Researchers are cautioned to examine the results for such impacts and potentially eliminate those cases with extreme weights.

**STEP 5: RECHECK BALANCE OF COVARIATES** Before proceeding to the calculation of the causal effect, the researcher should perform another balance check of the covariates among the observations retained in the analysis. If imbalance is found, then it must be corrected among the retained observations before proceeding to the final step of effect estimation.

**STEP 6: ESTIMATE CAUSAL EFFECT** With the adjustments to the treatment and control groups providing the necessary equivalence, the final step is to estimate the causal effect. We should note that at this stage this is nothing more than the main effect of treatment → control, but now with equivalent treatment and control groups. The group differences can be calculated within each matched group or strata and then combined by weighting by the relative size of the group/strata or directly estimated with weighting done by the IPTW. In any case, there are two types of effects that can be estimated: ATE and ATT [46, 88].

**ATE** The ATE is the average treatment effect of the treatment on the entire population and represents the expected difference in the outcome between treatment and control groups. This represents the effect most consistent with the main effect we have seen in ANOVA/MANOVA analyses.

**ATT** The ATT is the average treatment effect expected among those respondents in the treatment group. The ATT does not equal the ATE and has a quite different interpretation. See Morgan and Winship [88] for discussion of the differences between ATE and ATT.

**SUMMARY** The process of specifying the covariate set, estimating the propensity score model and then adjusting the treatment and control groups to obtain equivalence on the covariate set, involves a series of issues requiring both conceptual and empirical skills. The emergence of the technique across many social science and behavioral disciplines, however, has generated an extensive set of excellent resources for the researcher, ranging from complete texts devoted to the process [46, 88, 59, 40] as well as review articles within different domains (e.g., [112, 73, 59]). We encourage the interested research to explore software options within programs such as STATA, SAS and IBM SPSS as well as the many routines available in R. These will continue to expand as the technique develops and gains wider usage.

**Extension to MANOVA** The technical aspects of extending propensity score models to a MANOVA design would be relatively straightforward since the outcome is not involved in the estimation or adjustment stages. The outcome is critical, however, in the specification of the covariate set (Step 1). This is a difficult task with a single outcome and the task becomes even more challenging in accommodating multiple outcomes. One study did investigate the use of a single propensity score model on multiple outcomes, but it was not analyzed as a MANOVA design. It did, however, highlight the need for a unified set of covariates spanning all of the outcomes to be effective [129]. As a result, at this time, application of the MANOVA analysis is most likely in a randomized experimental setting where the equivalence of treatment and control groups can be managed through randomized assignment versus a propensity score approach.

## OVERVIEW

The concept of causal inference in non-experimental data has generated intense interest and usage in the social and behavioral sciences over the past decade. The introduction of the potential outcomes model by Rubin provided a framework for researchers in fields not amenable to randomized experiments to employ for making causal inferences from observational data. While the randomized experiment still remains the “gold standard” for making causal inferences, the addition of the approaches discussed here provides an exciting new research pathway for many research domains. We expect to see this topic expand in coverage across all research techniques and provide for even stronger research results across a wide array of research questions.

## Making Causal Inferences in Nonrandomized Situations

Controlled randomized experiments are considered the “gold standard” in estimating causal effects.

Estimating causal effects in non-experimental situations, such as observational data, has previously been thought impractical due to the inability to remove all of the confounding effects impacting the main effect.

Causal inferences are necessary outside the controlled experiment setting for many issues in which random assignment and manipulation of the treatment is not practical (e.g., health risks, naturally occurring phenomenon, etc.).

The development of the Potential Outcomes framework by Rubin established a general framework encompassing both randomized experiments and non-experimental data for the purposes of making causal inferences.

When participants are assigned, they can be in only treatment or control, not both. The counterfactual represents the outcomes expected of a participant if it was assigned to the other group (e.g., what would have happened to members of the control group if they were given the treatment).

The counterfactuals are always “missing” since a participant cannot be in both treatment and control. To make causal inferences appropriate counterfactuals must be formed.

Random assignment does this by assuming that the treatment and control group members are interchangeable (i.e., exactly equal on all characteristics other than the treatment variable).

Propensity scoring models are used to develop a propensity score that can be used to group treatment and control group members who also are highly equivalent on a set of variables.

Meeting the assumption of strongly ignorable treatment assignment (which random assignment does) means that the treatment assignment process (which participants are in treatment versus control) has no impact on the outcome. Violations of this assumption in non-experimental data are the result of sample selection issues or other factors/confounders that differ between treatment and outcome groups and also impact the outcome, thus producing biased estimates.

Confounding variables are those variables that impact both the treatment and control variables (i.e., act as a common cause), but are not included in the model. They must be controlled for in some manner so that treatment and control groups do not vary on any confounding variables. In this manner they become comparable to groups formed by random assignment.

The most critical step is the specification of all of the potential confounders to be included in the model. Omission of a relevant confounder risks the bias due to endogeneity in the causal effects. This is a conceptual, not an empirical issue.

While an unlimited number of variables should not be included in the set of confounders, it is best to include variables if not certain about their effect.

Do not include variables that are mediators (i.e., are caused by the treatment) or variables that are highly associated with the treatment, but not the outcome.

The logistic regression model is not evaluated on how well it “predicts” the treatment variables, but rather on its ability to estimate a variate across all of the confounders that also has balance between treatment and control groups.

The predicted probability value is used as the propensity score, a single value that can be used to group observations with similar profiles across the variables in the model.

The fundamental objective of the propensity score model is to obtain balance (i.e., equivalence) between the treatment and control groups across the variables included in the model. Balance is achieved when there is no significant difference on any model in the variable between treatment and control groups.

The propensity score is used to form equivalent groups of observations in the treatment and control groups by (a) matching, (b) stratification, or (c) weighting.

Matching involves matching observations from one group with observations from the opposite group based on equal or highly similar propensity scores.

Stratification involves forming subgroups of observations from the treatment and control groups which are similar on the propensity scores.

Weighting uses the inverse of the propensity score to weight individual observations before the causal effects are estimated.

Once the propensity scores are used to match, stratify or weight the observations to create equivalence between the treatment and control groups, average causal effects can be estimated.

ATE (Average treatment effect of the treatment) represents the expected difference in the outcome between treatment and control groups.

ATT (Average treatment effect for the treated) represents the expected difference among only those respondents who received the treatment.

The ATT does not equal the ATE.

## Summary

We discussed the appropriate applications and important considerations of MANOVA in addressing multivariate analyses with multiple dependent measures. Although considerable benefits stem from its use, MANOVA must be carefully and appropriately applied to the question at hand. Among the most notable advantages and disadvantages are:

### Advantages

- Allows for testing two or more related dependent variables in a single analysis while accounting for their multicollinearity.
- Controls for experiment-wide error rate that occurs from multiple ANOVA tests.
- Can have more statistical power than univariate tests.
- Can test for sequential orderings of dependent measures with stepdown analysis.

### Disadvantages

- Assumptions similar to discriminant analysis can be problematic.
- While overall multivariate tests are performed, tests of relative impact of individual dependent measures are less well defined.
- Requires increased sample size compared to separate ANOVAs.
- Overall results can be distorted by inclusion of inappropriate dependent measures.

MANOVA presents researchers a technique with flexibility and statistical power when focused on assessing a treatment→outcome relationship. We now illustrate the applications of MANOVA (and its univariate counterpart ANOVA) in a series of examples.

## Illustration of a MANOVA Analysis

Multivariate analysis of variance (MANOVA) affords researchers with the ability to assess differences across one or more nonmetric independent variables for a set of metric dependent variables. It provides a means for determining the extent to which groups of respondents (formed by their characteristics on the nonmetric independent variables) differ in terms of the dependent measures. Examining these differences can be done separately or in combination. In the following sections, we will detail the analyses necessary to examine two characteristics ( $X_1$  and  $X_5$ ) for their impact on a set of purchase outcomes ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ). First we will analyze each characteristic separately and then both in combination. The reader should note that an expanded version of HBAT (HBAT200 with a sample size of 200) is used in this analysis to allow for the analysis of a two-factor design. This data set is available in the online resources at the text's websites.

### RESEARCH SETTING

The HBAT customer relationship management (CRM) team is looking for strategic options for improving their ratings on the broad concept of customer experience. In strategy sessions, upper management identified two key elements of their CRM program they wish to investigate in further detail: their distribution system and their loyalty efforts which are based on tenure with HBAT as a customer.

**Strategic Objectives** Recent years have seen increased attention to the area of distribution systems. Fuelled by the widespread use of internet-based systems for channel integration and the cost savings being realized by improved logistical systems, upper management at HBAT is interested in assessing the current state of affairs in their distribution system, which utilizes both indirect (broker-based) and direct channels. In the indirect channel, products are sold to customers by brokers acting as both an external salesforce and even wholesalers in some instances. HBAT also employs a salesforce of its own; salespeople contact and service customers directly from both the corporate office and field offices. A concern has arisen that changes may be necessary in the distribution system, particularly focusing on the broker system that is perceived to not be performing well.

A second area of interest is a loyalty-based program that is a tiered program based on tenure as a customer. As customers move through the tiers, they are afforded more personalized services, discounts, etc. Upper management has great hope for improving customer perceptions and loyalty based on these efforts, especially in fostering long-term relationships with HBAT.

**Research Questions** To address these concerns, four research questions were posed:

- 1 What differences are present in customer satisfaction and other purchase outcomes between the two channels in the distribution system?
- 2 Is HBAT establishing better relationships with its customers over time, as reflected in customer satisfaction and other purchase outcomes?
- 3 What is the relationship between the distribution system and these relationships with customers in terms of the purchase outcomes?
- 4 For any significant main effects found for the distribution system ( $X_5$ ) with the purchase outcomes:
  - a Do they vary by firm size ( $X_3$ )?
  - b Does the percentage of a customer's purchases from HBAT ( $X_{22}$ ) play an intermediary role in the main effect on the purchase outcomes?

**Research Design** With the research questions defined, the research team now turns attention to defining the independent and dependent variables to be used and the ensuing sample size requirements.

**RESEARCH APPROACH** The necessity to examine existing programs and customers precludes a controlled experiment and instead must employ some type of field study. While a randomized field design would have been preferred, HBAT employed an observational study design with random selection of customers. While this method did not ensure balance among all of the cells (see Table 6.1 below), it was fairly well balanced and unequal cell sizes were not expected to create problems in the analysis.

**Table 6.1** Group Sizes for a Two-Factor Analysis Using the HBAT Data and HBAT200

Part A: HBAT (100 observations)					
		$X_5$ Distribution System			
		Indirect			
		Through Broker		Direct to Customer	Total
$X_1$	Less than 1 year	23		9	32
Customer	1 to 5 years	16		19	35
Type	More than 5 years	18		15	33
	Total	57		43	100
Part B: HBAT200 (200 observations)					
		$X_5$ Distribution System			
		Indirect			
		Through Broker		Direct to Customer	Total
$X_1$	Less than 1 year	52		16	68
Customer	1 to 5 years	25		39	64
Type	More than 5 years	31		37	68
	Total	108		92	200

**VARIABLES EMPLOYED** To examine these issues, researchers decided to employ MANOVA to examine the effects of  $X_5$  (Distribution System) and  $X_1$  (Customer Type) on three Purchase Outcome measures ( $X_{19}$ , Satisfaction;  $X_{20}$ , Likelihood of Recommending HBAT; and  $X_{21}$ , Likelihood of Future Purchase).

In selecting these three outcome measures, the management team was focused on including measures more fundamental to the CRM paradigm (e.g.,  $X_{19}$  Satisfaction) along with more behavioral measures ( $X_{20}$  Likely to recommend and  $X_{21}$  Likely to Purchase). From a strategic perspective, Satisfaction was deemed an objective fundamental to achieving success on Likely to Recommend and Likely to Purchase.

**CORRELATIONS OF OUTCOME VARIABLES** One of the advantages of MANOVA is its ability to correct for the correlations among the outcome variables. In this situation, the three outcomes had correlations ranging from .762 ( $X_{19}$  with  $X_{20}$ ) to .661 ( $X_{20}$  with  $X_{21}$ ). These correlations fall within the acceptable range for use in MANOVA, neither being too highly correlated or too low.

**SAMPLE** Although a sample size of 100 observations would be sufficient for either of the analyses of the individual variables, it would not be appropriate for addressing them in combination. A quick calculation of group sizes for this two-factor analysis (see Table 6.1, Part A) identified at least one group with fewer than 10 observations and several more with fewer than 20 observations.

Because these group sizes would not afford the ability to detect medium or small effect sizes with a desired level of statistical power (see Figure 6.11), a decision was made to gather additional responses to supplement the 100 observations already available. A second research effort added 100 more observations for a total sample size of 200. This new dataset is named HBAT200 and will be used for the MANOVA analyses that follow (see Table 6.1 Part B). Preliminary analyses indicated that the supplemented data set had the same basic characteristics as the HBAT, thus eliminating the need for additional examination of this new data to determine its basic properties.

## Example 1: Difference Between Two Independent Groups

To introduce the practical benefits of a multivariate analysis of group differences, we begin our discussion with one of the best-known designs: the two-group design in which each respondent is classified based on the levels (groups) of the treatment (independent variable). If this analysis was performed in an experimental setting, respondents would be assigned to groups randomly (e.g., depending on whether they see an advertisement or which type of cereal they taste). Many times, however, the groups are formed not by random assignment, but based instead on some characteristic of the respondent (e.g., age, gender, occupation).

In many research settings, however, it is also unrealistic to assume that a difference between any two experimental groups will be manifested in only a single dependent variable. For example, two advertising messages may not only produce different levels of purchase intent, but also may affect a number of other (potentially correlated) aspects of the response to advertising (e.g., overall product evaluation, message credibility, interest, attention).

Many researchers handle this multiple-outcome situation by repeated application of individual univariate  $t$  tests until all the dependent variables have been analyzed. However, this approach has serious deficiencies:

- Inflation of the Type I error rate over multiple  $t$  tests.
- Inability of paired  $t$  tests to detect differences among combinations of the dependent variables not apparent in univariate tests.

To overcome these problems, MANOVA can be employed to control the overall Type I error rate while still providing a means of assessing the differences on each dependent variable both collectively and individually.

## STAGE 1: OBJECTIVES OF THE ANALYSIS

The first step involves identifying the appropriate dependent and independent variables. As discussed earlier, HBAT identified the distribution system as a key element in its customer relationship strategy and first needs to understand the impact of the distribution system on customers.

**Research Question** HBAT is committed to strengthening its customer relationship strategy, with one aspect focused on distribution system. Concern has been raised about the differences due to distribution channel system ( $X_5$ ), which is composed of two channels: direct through HBAT's salesforce or indirect through a broker. Three purchase outcomes ( $X_{19}$ , Satisfaction;  $X_{20}$ , Likelihood of Recommending HBAT; and  $X_{21}$ , Likelihood of Future Purchase) have been identified as the focal issues in evaluating the impacts of the two distribution systems. The task is to identify whether any differences exist between these two systems across all or a subset of these purchase outcomes.

**Examining Group Profiles** Table 6.2 provides a summary of the group profiles on each of the purchase outcomes across the two groups (direct versus indirect distribution system). A visual inspection reveals that the direct distribution channel has the higher mean scores for each of the purchase outcomes. The task of MANOVA is to examine these differences and assess the extent to which these differences are significantly different, both individually and collectively.

## STAGE 2: RESEARCH DESIGN OF THE MANOVA

The principal consideration in the design of the two-group MANOVA is the sample size in each of the cells, which directly affects statistical power. Also, as is the case in most survey research, the cell sizes are unequal, making the statistical tests more sensitive to violations of the assumptions, especially the test for homogeneity of variance of the dependent variable. Both of these issues must be considered in assessing the research design using  $X_5$ .

As discussed earlier, the concern for adequate sample sizes across the entire MANOVA analysis resulted in the addition of 100 additional surveys to the original HBAT survey (see Table 6.2). Based on this larger dataset (HBAT200), 108 firms used the indirect broker system and 92 respondents used the direct system from HBAT.

These group sizes will provide more than adequate statistical power at an 80 percent probability to detect medium effect sizes and almost reach the levels necessary for identifying small effects sizes (see Figure 6.11). The result is a research design with relatively balanced group sizes and enough statistical power to identify differences at any managerially significant level.

## STAGE 3: ASSUMPTIONS IN MANOVA

The most critical assumptions relating to MANOVA are the independence of observations, homoscedasticity across the groups, and normality. Each of these assumptions will be addressed in regards to each of the purchase outcomes. Also of concern is the presence of outliers and their potential influence on the group means for the purchase outcome variables.

**Independence of Observations** The independence of the respondents was ensured as much as possible by the random sampling plan in the observational study. If the study had been done in an experimental setting, the random assignment of individuals would have ensured the necessary independence of observations.

**Table 6.2** Descriptive Statistics of Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) for Groups of  $X_5$  (Distribution System)

	$X_5$ Distribution System	Mean	Std. Deviation	N
$X_{19}$ Satisfaction	Indirect through broker	6.325	1.033	108
	Direct to customer	7.688	1.049	92
	Total	6.952	1.241	200
$X_{20}$ Likely to Recommend	Indirect through broker	6.488	.986	108
	Direct to customer	7.498	.930	92
	Total	6.953	1.083	200
$X_{21}$ Likely to Purchase	Indirect through broker	7.336	.880	108
	Direct to customer	8.051	.745	92
	Total	7.665	.893	200

**Homoscedasticity** A second critical assumption concerns the homogeneity of the variance–covariance matrices among the two groups. The first analysis assesses the univariate homogeneity of variance across the two groups. As shown in Table 6.3, univariate tests (Levene's test) for all three variables are nonsignificant (i.e., significance greater than .05). The next step is to assess the dependent variables collectively by testing the equality of the entire variance–covariance matrices between the groups. Again, in Table 6.3 the Box's M test for equality of the covariance matrices shows a nonsignificant value (.607), indicating no significant difference between the two groups on the three dependent variables collectively. Thus, the assumption of homoscedasticity is met for each individual variable separately and the three variables collectively.

**Correlation and Normality of Dependent Variables** Another test should be made to determine whether the dependent measures are significantly correlated. The most widely used test for this purpose is Bartlett's test for sphericity. It examines the correlations among all dependent variables and assesses whether, collectively, significant intercorrelation exists. In our example, a significant degree of intercorrelation does exist (significance = .000) (see Table 6.3).

The assumption of normality for the dependent variables ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) was examined in Chapter 2 and found to be acceptable. This supports the results of testing for the equality of the variance–covariance matrices between groups.

**Outliers** The final issue to be addressed is the presence of outliers. A simple approach that identifies extreme points for each group is the use of boxplots (see Figure 6.15). Examining the boxplot for each dependent measure shows few, if any, extreme points across the groups. When we examine these extreme points across the three dependent measures, no observation was an extreme value on all three dependent measures, nor did any observation have a value so extreme that it dictated exclusion. Thus, all 200 observations will be retained for further analysis.

#### STAGE 4: ESTIMATION OF THE MANOVA MODEL AND ASSESSING OVERALL FIT

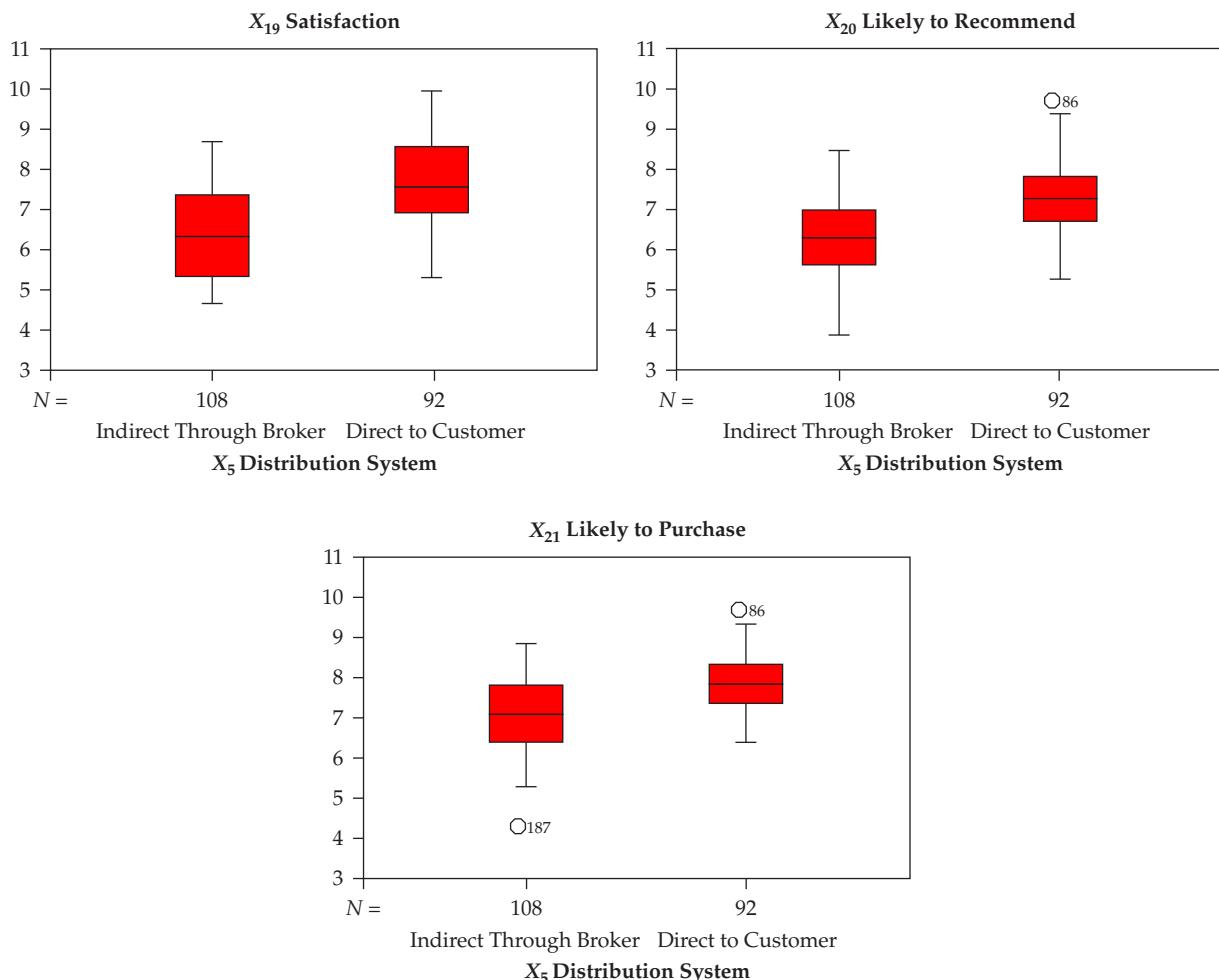
The next step is to assess whether the two groups exhibit statistically significant differences for the three purchase outcome variables, first collectively and then individually. To conduct the test, we first specify the maximum allowable

**Table 6.3** Multivariate and Univariate Measures for Testing Homoscedasticity of  $X_5$

<b>Multivariate Test of Homoscedasticity</b>				
<i>Box's Test of Equality of Covariance Matrices</i>				
Box's M	4.597			
F	.753			
df1	6			
df2	265275.824			
Sig.	.607			
<b>Univariate Tests of Homoscedasticity</b>				
<i>Levene's Test of Equality of Error Variances</i>				
Dependent Variable	F	df1	df2	Sig.
$X_{19}$ Satisfaction	.001	1	198	.978
$X_{20}$ Likely to Recommend	.643	1	198	.424
$X_{21}$ Likely to Purchase	2.832	1	198	.094
<b>Test for Correlation Among the Dependent Variables</b>				
<i>Bartlett's Test of Sphericity</i>				
Likelihood Ratio	.000			
Approx. Chi-Square	260.055			
df	5			
Sig.	.000			

**Figure 6.15**

Boxplots of Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) for Groups of  $X_5$  (Distribution System)



Type I error rate. In doing so, we accept that five times out of 100 we might conclude that the type of distribution channel has an impact on the purchase outcome variables when in fact it did not.

**Statistical Significance Testing** Having set the acceptable Type I error rate, we first use the multivariate tests to test the set of dependent variables for differences between the two groups and then perform univariate tests on each purchase outcome. Finally, power levels are assessed.

**MULTIVARIATE STATISTICAL TESTING** Table 6.4 contains the four most commonly used multivariate tests (Pillai's criterion, Wilks' lambda, Hotelling's  $T^2$  and Roy's greatest characteristic root). Each of the four measures indicates that the set of purchase outcomes have a highly significant difference (.000) between the two types of distribution channel. This confirms the group differences seen in Table 6.2 and the boxplots of Figure 6.15.

**UNIVARIATE STATISTICAL TESTS** Although we can show that the set of purchase outcomes differs across the groups, we also need to examine each purchase outcome separately for differences across the two types of distribution channel. Table 6.4 also contains the univariate tests for each individual purchase outcome. As we can see, all of the individual tests are also highly significant (significance = .000), indicating that each variable follows the same pattern of higher purchase outcomes (see Table 6.2) for those served by the direct distribution system (Direct Distribution customers

**Table 6.4** Multivariate and Univariate Tests for Group Differences in Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) Across Groups of  $X_5$  (Distribution System)

Multivariate Tests							
Statistical Test	Value	Hypothesis				Observed Power <sup>a</sup>	
		F	df	Error df	Sig.	$\eta^2$	Power <sup>a</sup>
Pillai's Criterion	.307	28.923	3	196	.000	.307	1.00
Wilks' Lambda	.693	28.923	3	196	.000	.307	1.00
Hotelling's $T^2$	.443	28.923	3	196	.000	.307	1.00
Roy's greatest characteristic root	.443	28.923	3	196	.000	.307	1.00

Univariate Tests (Between-Subjects Effects)							
Dependent Variable	Sum of Squares	df	Mean Square			Observed Power <sup>a</sup>	
			Square	F	Sig.	$\eta^2$	Power <sup>a</sup>
$X_{19}$ Satisfaction	92.300 <sup>b</sup>	1	92.300	85.304	.000	.301	1.00
$X_{20}$ Likely to Recommend	50.665 <sup>c</sup>	1	50.665	54.910	.000	.217	1.00
$X_{21}$ Likely to Purchase	25.396 <sup>d</sup>	1	25.396	37.700	.000	.160	1.00

Significance of Discriminant Functions							
Test of Function(s)	Wilks' Lambda			Structure Loadings on Function			
	Wilks' Lambda	Chi-square	df	Sig.	Dependent/Outcome Variables	Loading <sup>e</sup>	
1	.693	72.020	3	.000	$X_{19}$ Satisfaction	.986	
					$X_{20}$ Likely to Recommend	.791	
					$X_{21}$ Likely to Purchase	.656	

Roy–Bargman Stepdown Tests					
		Hypothesis MS	Error MS	Stepdown F	Significance
Variable Order ( $X_{19}$ , $X_{20}$ , $X_{21}$ )					
	$X_{19}$ Satisfaction	92.300	1.082	85.304	.000
	$X_{20}$ Likely to Recommend	.771	.494	1.561	.213
	$X_{21}$ Likely to Purchase	.026	.361	.072	.788
Variable Order ( $X_{21}$ , $X_{20}$ , $X_{19}$ )					
	$X_{21}$ Likely to Purchase	25.396	.674	37.700	.000
	$X_{20}$ Likely to Recommend	11.296	.610	18.521	.000
	$X_{19}$ Satisfaction	9.906	.469	21.127	.000

<sup>a</sup>Computed using alpha = .05.

<sup>b</sup> $R^2 = .301$  (Adjusted  $R^2 = .298$ ).

<sup>c</sup> $R^2 = .217$  (Adjusted  $R^2 = .213$ ).

<sup>d</sup> $R^2 = .160$  (Adjusted  $R^2 = .156$ ).

<sup>e</sup>Correlations between variables and standardized canonical discriminant functions.

have values of 7.688, 7.498, and 8.051 versus values of 6.325, 6.488, and 7.336 for Indirect Through Broker customers on  $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ , respectively).

**DISCRIMINANT ANALYSIS** One proposed method of assessing the individual outcome variables in terms of their differences across the treatment variable(s) is the use of discriminant analysis. The objective here is to “profile” the outcome variables in terms of their differences between groups of the treatment variable. This becomes particularly insightful when the treatment variable has three or more levels, since there is the possibility of two or more discriminant functions being extracted.

In the case of  $X_5$ , only one discriminant function is possible since there are only two levels. As shown in Table 6.4, the discriminant function differed significantly between the two groups, confirming the multivariate and univariate results discussed above. Moreover, we can see the structure loadings all indicating a fairly strong association with the discriminant variate, although the loadings do decrease in moving from  $X_{19}$  to  $X_{21}$ .

**STEPDOWN TESTS** A final approach to assessing the differences across the separate outcome measures was with stepdown analysis. This test is particularly useful if there is an “ordering” to the outcome measures. In this situation,  $X_{19}$  (Satisfaction) was thought to be a fundamental objective that was then very impactful in achieving  $X_{20}$  (Likely to Recommend) and  $X_{21}$  (Likely to Purchase). Stepdown analysis allows the researcher to apply a sequential process to assessing group differences, with each step accounting for the differences already accounted for by the prior outcomes included in the analysis. Obviously, if the outcome measures were uncorrelated this would be equivalent to the separate univariate tests. But as the outcomes become correlated, we can see how “distinctive” the differences are as each outcome measure is entered sequentially.

Two stepdown analyses were performed (see Table 6.4). The first had the order of Satisfaction, Likely to Recommend and finally Likely to Purchase. As the results show, once the differences on Satisfaction, which are significant, are accounted for there are not any incremental differences across the final two variables. This is indicative of the correlations among outcome measures and an ordering of the variables. If we look at the second stepdown analysis (Likely to Purchase, Likely to Recommend and Satisfaction), we see a very different pattern. Now, each sequential entry of an outcome measure still demonstrates significant differences.

These results demonstrate two important aspects of the stepdown tests. First, the ordering of the variables does matter, as we see quite different results depending on the order. Thus, an analyst must be cautious in drawing any substantive conclusions from a single stepdown test. Second, if there is a proposed ordering, the stepdown test can provide support as well. In this case, the results indicate that once we account for the differences on satisfaction, there are no significant differences on the remaining outcome measures. But if we reverse the order, accounting for the differences on Likely To Purchase does not represent all of the differences on the other two outcomes, which also have significant differences when they enter the analysis. As a result, the analyst can address more complex conceptual issues in addition to using the stepdown test as just a statistical verification of differences.

**Power Analysis** The power for the statistical tests was 1.0, indicating that the sample sizes and the effect size were sufficient to ensure that the significant differences would be detected if they existed beyond the differences due to sampling error.

## STAGE 5: INTERPRETATION OF THE RESULTS

The presence of only two groups eliminates the need to perform any type of post hoc tests. The statistical significance of the multivariate and univariate tests indicating group differences on the dependent variate (vector of means) and the individual purchase outcomes leads the researcher to an examination of the results to assess their logical consistency.

As noted earlier, firms using the direct type of distribution system scored significantly higher than those serviced through the broker-based indirect distribution channel. The group means shown in Table 6.2, based on

responses to a 10-point scale, indicate that the customers using the direct distribution channel are more satisfied ( $7.688 - 6.325 = 1.363$ ), more likely to recommend HBAT ( $7.498 - 6.488 = 1.01$ ), and more likely to purchase in the future ( $8.051 - 7.336 = .715$ ). These differences are also reflected in the boxplots for the three purchase outcomes in Figure 6.15.

## SUMMARY

These results confirm that the type of distribution channel does affect customer perceptions in terms of the three purchase outcomes, both individually and collectively. These statistically significant differences, which are of a sufficient magnitude to denote managerial significance as well, indicate that the direct distribution channel is more effective in creating positive customer perceptions on a wide range of purchase outcomes.

The stepdown analysis supports the ordering of Satisfaction → Likely to Recommend → Likely to Purchase as an appropriate decision-making framework. This results from the nature of the differences, in that accounting for the differences on Satisfaction is sufficient to account for the differences on the other outcomes, but the reverse does not hold. This supports a framework where Satisfaction is considered fundamental to achieving the subsequent behavioral actions.

## Example 2: Difference Between K Independent Groups

The two-group design (Example 1) is a special case of the more general  $k$ -group design. In the general case, each respondent is a member or is randomly assigned to one of  $k$  levels (groups) of the treatment (independent variable). In a univariate case, a single metric dependent variable is measured, and the null hypothesis is that all group means are equal (i.e.,  $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ ). In the multivariate case, multiple metric dependent variables are measured, and the null hypothesis is that all group vectors of mean scores are equal (i.e.,  $\nu_1 = \nu_2 = \nu_3 = \dots = \nu_k$ ), where  $\nu$  refers to a vector or set of mean scores.

In  $k$ -group designs in which multiple dependent variables are measured, many researchers proceed with a series of individual  $F$  tests (ANOVAs) until all the dependent variables have been analyzed. As the reader should suspect, this approach suffers from the same deficiencies as a series of  $t$  tests across multiple dependent variables; that is, a series of  $F$  tests with ANOVA:

- Results in an inflated Type I error rate.
- Ignores the possibility that some composite of the dependent variables may provide reliable evidence of overall group differences.

In addition, because individual  $F$  tests ignore the correlations among the independent variables, they use less than the total information available for assessing overall group differences.

MANOVA again provides a solution to these problems. MANOVA solves the Type I error rate problem by providing a single overall test of group differences at a specified  $\alpha$  level. It solves the composite variable problem by implicitly forming and testing the linear combinations of the dependent variables that provide the strongest evidence of overall group differences.

## STAGE 1: OBJECTIVES OF THE MANOVA

In the prior example, HBAT assessed its performance among customers based on which of the two distribution system channels ( $X_5$ ) were used. MANOVA was employed due to the desire to examine a set of three purchase outcome variables representing HBAT performance. A second research objective was to determine whether the three purchase outcome variables were affected by the length of their relationship with HBAT ( $X_1$ ), which coincided with their loyalty program efforts. The null hypothesis HBAT now wishes to test is that the three sample vectors of the mean scores (one vector for each category of customer relationship) are equivalent.

**Research Questions** In addition to examining the role of the distribution system, HBAT also has stated a desire to assess whether the differences in the purchase outcomes are attributable solely to the type of distribution channel or whether other nonmetric factors can be identified that show significant differences as well. HBAT specifically selected  $X_1$  (Customer Type) to determine whether the length of HBAT's relationship with the customer has any impact on these purchase outcomes.

**Examining Group Profiles** As can be seen in Table 6.5, the mean scores of all three purchase outcome variables increase as the length of the customer relationship increases. The question to be addressed in this analysis is the extent to which these differences as a whole can be considered statistically significant and if those differences extend to each difference between groups. In a second MANOVA analysis,  $X_1$  (Customer Type) is examined for differences in purchase outcomes.

## STAGE 2: RESEARCH DESIGN OF MANOVA

As was the situation in the earlier two-group analysis, sample size of the groups is a primary consideration in research design. Even when all instances of the group sizes far exceed the minimum necessary, the researcher should always be concerned with achieving the statistical power needed for the research question at hand.

Analysis of the impact of  $X_1$  now requires that we analyze the sample sizes for the three groups of length of customer relationship (less than one year, one to five years, and more than five years). In the HBAT sample, the 200 respondents are almost evenly split across the three groups with sample sizes of 68, 64, and 68 (see Table 6.5). These sample sizes, in conjunction with the three dependent variables, exceed the guidelines shown in Figure 6.11 to identify medium effect sizes (suggested sample sizes of 44 to 56), but fall somewhat short of the required sample size (98 to 125) needed to identify small effect sizes with a power of .80. Thus, any nonsignificant results should be examined closely to evaluate whether the effect size has managerial significance, because the low statistical power precluded designating it as statistically significant.

## STAGE 3: ASSUMPTIONS IN MANOVA

Having already addressed the issues of normality (see Chapter 2) and intercorrelation (Bartlett's test of sphericity in Table 6.3) of the dependent variables in the prior example, the only remaining concern rests in the homoscedasticity of the purchase outcomes across the groups formed by  $X_1$  and identification of any outliers. We first examine this homoscedasticity at the multivariate level (all three purchase outcome variables collectively) and then for each dependent variable separately. The multivariate test for homogeneity of variance

**Table 6.5** Descriptive Statistics of Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) for Groups of  $X_1$  (Customer Type)

	$X_1$ Customer Type	Mean	Std. Deviation	N
$X_{19}$ Satisfaction	Less than 1 year	5.729	.764	68
	1 to 5 years	7.294	.708	64
	More than 5 years	7.853	1.033	68
	Total	6.952	1.241	200
$X_{20}$ Likely to Recommend	Less than 1 year	6.141	.995	68
	1 to 5 years	7.209	.714	64
	More than 5 years	7.522	.976	68
	Total	6.953	1.083	200
$X_{21}$ Likely to Purchase	Less than 1 year	6.962	.760	68
	1 to 5 years	7.883	.643	64
	More than 5 years	8.163	.777	68
	Total	7.665	.893	200

of the three purchase outcomes is performed with the Box's M test, while the Levene test is used to assess each purchase outcome variable separately.

**Homoscedasticity** Table 6.6 contains the results of both the multivariate and univariate tests of homoscedasticity. The Box's M test indicates no presence of heteroscedasticity (significance = .069). In the Levene's tests for equality of error variances, two of the purchase outcomes ( $X_{20}$  and  $X_{21}$ ) showed nonsignificant results and confirmed homoscedasticity. In the case of  $X_{19}$ , the significance level was .001, indicating the possible existence of heteroscedasticity for this variable. However, given the relatively large sample sizes in each group, the relatively equal sizes across the groups, and the presence of homoscedasticity for the other two purchase outcomes, corrective remedies were not needed for  $X_{19}$ .

**Outliers** Examination of the boxplot for each purchase outcome variable (see Figure 6.16) reveals a small number of extreme points for each dependent measure (observation 104 for  $X_{19}$ ; observations 86, 104, 119, and 149 for  $X_{20}$ ; and observations 104 and 187 for  $X_{21}$ ). Only one observation had extreme values on all three dependent measures and none of the values were so extreme in any cases as to markedly affect the group values. Thus, no observations were classified as outliers designated for exclusion and all 200 observations were used in this analysis.

#### STAGE 4: ESTIMATION OF THE MANOVA MODEL AND ASSESSING OVERALL FIT

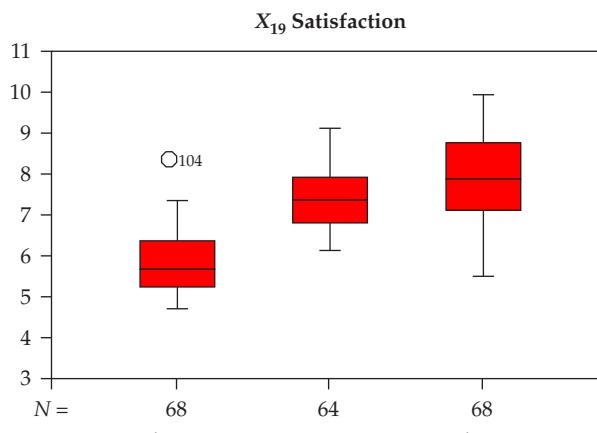
Using MANOVA to examine an independent variable with three or more levels reveals the differences across the levels for the dependent measures with the multivariate and univariate statistical tests illustrated in the earlier example. In these situations, the statistical tests are testing for a significant main effect, meaning that the differences between groups, when viewed collectively, are substantial enough to be deemed statistically significant. It should be noted that statistical significance of the main effect does not ensure that each group is also significantly different from each other group. Rather, separate tests described in the next section can examine which groups do exhibit significant differences.

**Statistical Significance Testing** All three dependent measures show a definite pattern of increasing as the length of the customer relationship increases (see Table 6.7 and Figure 6.16). We will first examine the multivariate tests and then the three approaches to assessing the differences of individual outcomes.

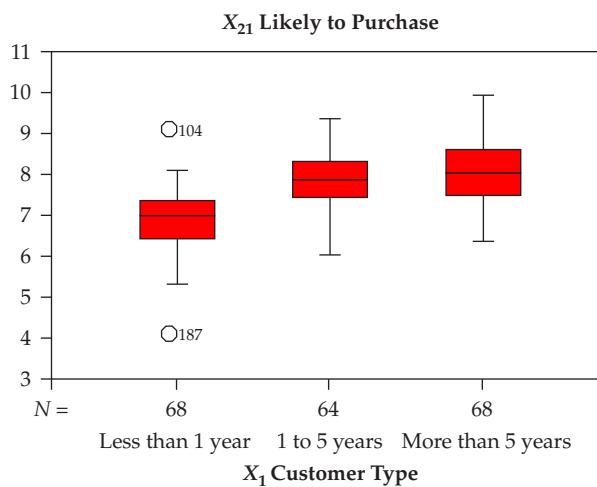
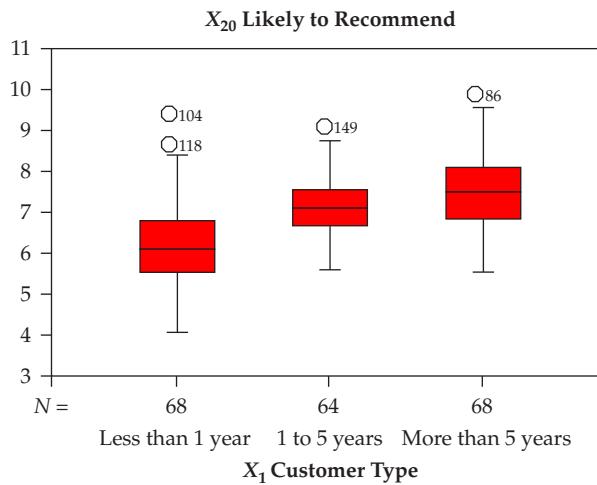
**MULTIVARIATE STATISTICAL TESTING** The first step is to utilize the multivariate tests and assess whether the set of purchase outcomes, which each individually seem to follow a similar increasing pattern as time increases, does vary in a

**Table 6.6** Multivariate and Univariate Measures for Testing Homoscedasticity of  $X_1$

<b>Multivariate Test of Homoscedasticity</b>				
<i>Box's Test of Equality of Covariance Matrices</i>				
Box's M		20.363		
F		1.659		
df <sub>1</sub>		12		
df <sub>2</sub>		186673.631		
Sig.		.069		
<b>Univariate Test of Homoscedasticity</b>				
<i>Levene's Test of Equality of Error Variances</i>				
Dependent Variable	F	df <sub>1</sub>	df <sub>2</sub>	Sig.
$X_{19}$ Satisfaction	6.871	2	197	.001
$X_{20}$ Likely to Recommend	2.951	2	197	.055
$X_{21}$ Likely to Purchase	.800	2	197	.451



**Figure 6.16**  
Boxplots of Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) for Groups of  $X_1$  (Customer Type)



**Table 6.7** Multivariate and Univariate Tests for Group Differences in Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) Across Groups of  $X_1$  (Customer Type)

Multivariate Tests							
Statistical Test	Value	F	Hypothesis df	Error df	Sig.	$\eta^2$	Observed Power <sup>a</sup>
Pillai's Criterion	.543	24.368	6	392	.000	.272	1.000
Wilks' Lambda	.457	31.103	6	390	.000	.324	1.000
Hotelling's T <sup>2</sup>	1.184	38.292	6	388	.000	.372	1.000
Roy's greatest characteristic root	1.183	77.280	3	196	.000	.542	1.000

Univariate Tests (Between-Subjects Effects)							
Dependent Variable	Type III Sum of Squares		Mean Square			Observed Power <sup>a</sup>	
			df		F	Sig.	$\eta^2$
$X_{19}$ Satisfaction	164.311 <sup>b</sup>	2	82.156	113.794	.000	.536	1.00
$X_{20}$ Likely to Recommend	71.043 <sup>c</sup>	2	35.521	43.112	.000	.304	1.00
$X_{21}$ Likely to Purchase	53.545 <sup>d</sup>	2	26.773	50.121	.000	.337	1.00

Significance of Discriminant Functions							
Wilks' Lambda						Structure Loadings on Function	
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.	Dependent/Outcome Variables	Loading <sup>e</sup>	
1 through 2	.457	153.285	6	.000	$X_{19}$ Satisfaction	.988	-.059
2	.999	.282	2	.869	$X_{20}$ Likely to Recommend	.656	.550
					$X_{21}$ Likely to Purchase	.608	.635

Roy-Bargman Stepdown Tests					
	Hypothesis MS		Error MS	Stepdown F	Significance
Variable Order ( $X_{19}$ , $X_{20}$ , $X_{21}$ )					
$X_{19}$ Satisfaction	82.156		.722	113.794	.000
$X_{20}$ Likely to Recommend	.062		.499	.125	.882
$X_{21}$ Likely to Purchase	.452		.358	1.264	.285
Variable Order ( $X_{21}$ , $X_{20}$ , $X_{19}$ )					
$X_{21}$ Likely to Purchase	26.773		.534	50.121	.000
$X_{20}$ Likely to Recommend	4.969		.620	8.015	.000
$X_{19}$ Satisfaction	12.895		.390	33.077	.000

<sup>a</sup>Computed using alpha = .05.

<sup>b</sup> $R^2 = .536$  (Adjusted  $R^2 = .531$ ).

<sup>c</sup> $R^2 = .304$  (Adjusted  $R^2 = .297$ ).

<sup>d</sup> $R^2 = .337$  (Adjusted  $R^2 = .331$ ).

<sup>e</sup>Correlations between variables and standardized canonical discriminant functions.

statistically significant manner (i.e., a significant main effect). Table 6.7 contains the four most commonly used multivariate tests and as we see, all four tests indicate a statistically significant difference of the collective set of dependent measures across the three groups.

**UNIVARIATE STATISTICAL TESTING** In addition to the multivariate tests, univariate tests for each dependent measure indicate that all three dependent measures, when considered individually, also have significant main effects. Thus, both collectively and individually, the three purchase outcomes ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) do vary at a statistically significant level across the three groups of  $X_1$ .

**DISCRIMINANT ANALYSIS AND STEPDOWN TESTS** The final two approaches for assessing the separate outcomes were comparable to those patterns seen in our discussion of Example 1 with  $X_5$ . This is expected given the correlated outcome

measures, so these results reinforce the conclusion drawn about the strength of the association between outcomes and the conceptual ordering of the three measures.

## STAGE 5: INTERPRETATION OF THE RESULTS

Interpreting a MANOVA analysis with an independent variable of three or more levels requires a two-step process:

- Examination of the main effect of the independent variable (in this case,  $X_1$ ) on the three dependent measures.
- Identifying the differences between individual groups for each of the dependent measures with either planned comparisons or post hoc tests.

The first analysis examines the overall differences across the levels for the dependent measures, whereas the second analysis assesses the differences between individual groups (e.g., group 1 versus group 2, group 2 versus group 3, group 1 versus group 3) to identify those group comparisons with significant differences.

**Assessing the Main Effect of  $X_5$**  All of the multivariate and univariate tests indicated a significant main effect of  $X_1$  (Customer Type) on each individual dependent variable as well as the set of the dependent variables when considered collectively. The significant main effect means that the dependent variable(s) do vary in significant amounts between the three customer groups based on length of customer relationship. As we can see in Table 6.5 and Figure 6.16 the pattern of purchases increases in each dependent measure as the customer relationship matures. For example, customer satisfaction ( $X_{19}$ ) is lowest (5.729) for those customers of less than 1 year, increasing (7.294) for those customers between 1 and 5 years until it reaches the highest level (7.853) for those customers of 5 years or more. Similar patterns are seen for the two other dependent measures.

**Making Post Hoc Comparisons** As noted earlier, a significant main effect indicates that the total set of group differences (e.g., group 1 versus group 2) are large enough to be considered statistically significant. It should also be noted that a significant main effect does not guarantee that every one of the group differences is also significant. We may find that a significant main effect is actually due to a single group difference (e.g., group 1 versus group 2) while all of the other comparisons (group 1 versus group 3 and group 2 versus group 3) are not significantly different.

The question becomes: How are these individual group differences assessed while maintaining an acceptable level of overall Type I error rate? This same problem is encountered when considering multiple dependent measures, but in this case in making comparisons for a single dependent variable across multiple groups. This type of question can be tested with one of the a priori procedures. If the contrast is used, a specific comparison is made between two groups (or sets of groups) to see whether they are significantly different. Another approach is to use one of the post hoc procedures that tests all group differences and then identifies those differences that are statistically significant.

Table 6.8 contains three post hoc comparison methods (Tukey HSD, Scheffé, and LSD) applied to all three purchase outcomes across the three groups of  $X_1$ . When we examine  $X_{19}$  (Satisfaction), we first see that even though the overall main effect is significant, the differences between adjacent groups are not constant. The difference between customers of less than 1 year and those of 1 to 5 years is  $-1.564$  (the minus sign indicates that customers of less than 1 year have the lower value). When we examine the group difference between customers of 1 to 5 years versus those of more than 5 years, however, the difference is reduced to  $-.559$  (about one-third of the prior difference).

The researcher is thus interested in whether both of these differences are significant, or only significant between the first two groups. When we look to the last three columns in Table 6.8, we can see that all of the separate group differences for  $X_{19}$  are significant, indicating that the difference of  $-.559$ , even though much smaller than the other group difference, is still statistically significant.

**Table 6.8** Post Hoc Comparisons for Individual Group Differences on Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) Across Groups of  $X_1$  (Customer Type)

Dependent Variable	Groups to Be Compared		Mean Difference Between Groups (I – J)		Statistical Significance of Post Hoc Comparison		
			Mean Difference	Standard Error	Tukey		
	Group I	Group J			HSD	Scheffé	LSD
$X_{19}$ Satisfaction							
	Less than 1 year	1 to 5 years	–1.564	.148	.000	.000	.000
	Less than 1 year	More than 5 years	–2.124	.146	.000	.000	.000
	1 to 5 years	More than 5 years	–.559	.148	.000	.001	.000
$X_{20}$ Likely to Recommend							
	Less than 1 year	1 to 5 years	–1.068	.158	.000	.000	.000
	Less than 1 year	More than 5 years	–1.381	.156	.000	.000	.000
	1 to 5 years	More than 5 years	–.313	.158	.118	.144	.049
$X_{21}$ Likely to Purchase							
	Less than 1 year	1 to 5 years	–.921	.127	.000	.000	.000
	Less than 1 year	More than 5 years	–1.201	.125	.000	.000	.000
	1 to 5 years	More than 5 years	–.280	.127	.071	.091	.029

When we examine the post hoc comparisons for the other two purchase outcomes ( $X_{20}$  and  $X_{21}$ ), a different pattern emerges. Again, the differences between the first two groups (less than 1 year and 1 to 5 years) are all statistically significant across all three post hoc tests. Yet when we examine the next comparison (customers of 1 to 5 years versus those of more than 5 years), two of the three tests indicate that the two groups are not different. In these tests, the purchase outcomes of  $X_{20}$  and  $X_{21}$  for customers of 1 to 5 years are not significantly different from those of more than 5 years. This result is contrary to what was found for satisfaction, in which this difference was significant.

When the independent variable has three or more levels, the researcher must engage in this second level of analysis in addition to the assessment of significant main effects. Here the researcher is not interested in the collective effect of the independent variable, but instead in the differences between specific groups. The tools of either planned comparisons or post hoc methods provide a powerful means of making these tests of group differences while also maintaining the overall Type I error rate.

## SUMMARY

The MANOVA analysis revealed both multivariate and univariate differences of all three outcomes across the different customer tenures groups. As expected, all three measures increased as the customer tenure increased. But while the gains in satisfaction were still positive across the years, the levels of Likely to Recommend and Likely to Purchase did not exhibit differences between the final two groups. Thus, all of the significant differences for these last two outcomes came between the first two groups (Less than 1 year versus 1 to 5 years). The management team may wish to investigate further why the continued increases in Satisfaction did not translate to increases on the other outcome measures in the later years.

## Example 3: A Factorial Design for MANOVA with Two Independent Variables

In the prior two examples, the MANOVA analyses have been extensions of univariate two- and three-group analyses. In this example, we explore a multivariate factorial design: two independent variables used as treatments to analyze differences of the set of dependent variables. In the course of our discussion, we assess the interactive or joint effects between the two treatments on the dependent variables separately and collectively.

## STAGE 1: OBJECTIVES OF THE MANOVA

In the previous multivariate research questions, HBAT considered the effect of only a single-treatment variable on the dependent variables. Here the possibility of joint effects among two or more independent variables must also be considered. In this way, the interaction between the independent variables can be assessed along with their main effects.

**Research Questions** The first two research questions we examined addressed the impact of two factors—distribution system and duration of customer relationship—on a set of purchase outcomes. In each instance, the factors were shown to have significant impacts (i.e., more favorable purchase outcomes for firms in the direct distribution system or those with longer tenure as an HBAT customer).

Left unresolved is a third question: How do these two factors operate when considered simultaneously? Here we are interested in knowing how the differences between distribution systems hold across the groups based on length of HBAT relationship. We saw that customers in the direct distribution system had significantly greater purchase outcomes (higher satisfaction, etc.), yet are these differences always present for each customer group based on  $X_1$ ? The following is just a sample of the types of question we can ask when considering the two variables together in a single analysis:

- Is the direct distribution system more effective for newer customers?
- Do the two distribution systems show differences for customers of 5 years or more?
- Is the direct distribution system always preferred over the indirect system across the customer groups of  $X_1$ ?

By combining both independent variables ( $X_1$  and  $X_5$ ) into a factorial design, we create six customer groups: the three groups based on length of their relationship with HBAT separated into those groups in each distribution system channel. Known as a  $3 \times 2$  design, the three levels of  $X_1$  separated for each level of  $X_5$  form a separate group for each customer type within each distribution system channel.

**Examining Group Profiles** Table 6.9 provides a profile of each group for the set of purchase outcomes. Many times a quicker and simpler perspective is through a graphical display. One option is to form a line chart, and we will illustrate this when viewing the interaction terms in a later section. We can also utilize boxplots to show not only the differences between group means, but the overlap of the range of values in each group. Figure 6.17 illustrates such a graph for  $X_{19}$  (Satisfaction) across the six groups of our factorial design. As we can see, satisfaction increases as the length of relationship with HBAT increases, but the differences between the two distribution systems are not always constant (e.g., they seem much closer for customers of 1 to 5 years).

The purpose of including multiple independent variables into a MANOVA is to assess their effects “contingent on” or “controlling for” the other variables. In this case, we can see how the length of the HBAT relationship changes in any way the more positive perceptions generally seen for the direct distribution system.

## STAGE 2: RESEARCH DESIGN OF THE MANOVA

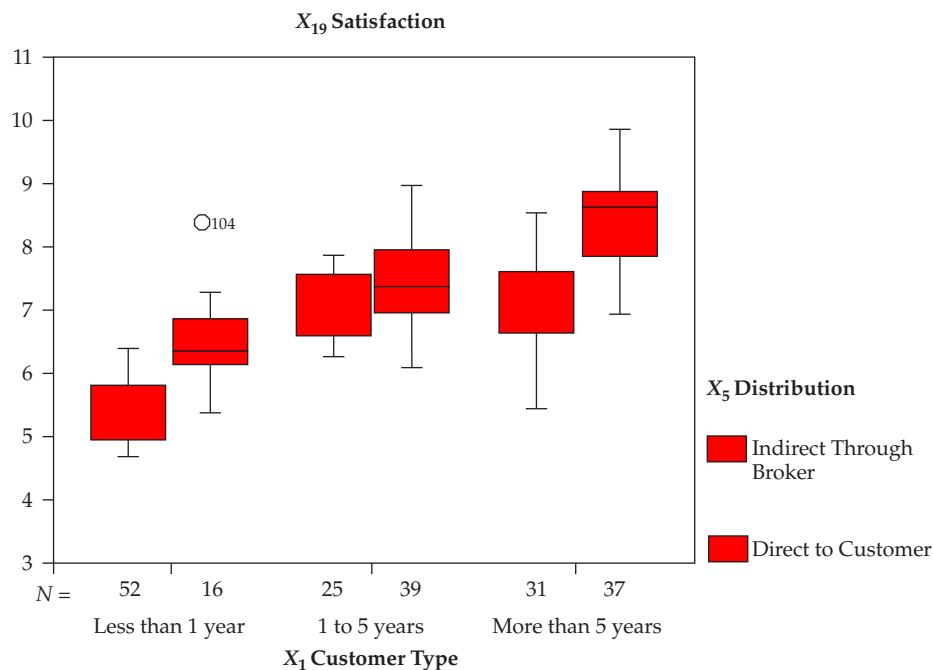
Any factorial design of two or more independent variables raises the issue of adequate sample size in the various groups. The researcher must ensure, when creating the factorial design, that each group has sufficient sample size for the following:

- To meet the minimum requirements of group sizes exceeding the number of dependent variables.
- To provide the statistical power to assess differences deemed practically significant.

**Sample Size Considerations** As noted in the previous section, this analysis is termed a  $2 \times 3$  design because it includes two levels of  $X_5$  (direct versus indirect distribution) and three levels of  $X_1$  (less than 1 year, 1 to 5 years, and more than 5 years). The issue of sample size per group was such a concern to HBAT researchers that the original HBAT survey of 100 observations was supplemented by 100 additional respondents just for this analysis (see more detailed discussion in the section preceding the examples). Even with the additional respondents, the sample of 200

**Table 6.9** Descriptive Statistics of Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) for Groups of  $X_1$  (Customer Type) by  $X_5$  (Distribution System)

Dependent Variable	$X_1$ Customer Type	$X_5$ Distribution System	Mean	Std. Deviation	N
$X_{19}$ Satisfaction	Less than 1 year	Indirect through broker	5.462	.499	52
		Direct to customer	6.600	.839	16
		Total	5.729	.764	68
	1 to 5 years	Indirect through broker	7.120	.551	25
		Direct to customer	7.405	.779	39
		Total	7.294	.708	64
	More than 5 years	Indirect through broker	7.132	.803	31
		Direct to customer	8.457	.792	37
		Total	7.853	1.033	68
	Total	Indirect through broker	6.325	1.033	108
		Direct to customer	7.688	1.049	92
		Total	6.952	1.241	200
$X_{20}$ Likely to Recommend	Less than 1 year	Indirect through broker	5.883	.773	52
		Direct to customer	6.981	1.186	16
		Total	6.141	.995	68
	1 to 5 years	Indirect through broker	7.144	.803	25
		Direct to customer	7.251	.659	39
		Total	7.209	.714	64
	More than 5 years	Indirect through broker	6.974	.835	31
		Direct to customer	7.981	.847	37
		Total	7.522	.976	68
	Total	Indirect through broker	6.488	.986	108
		Direct to customer	7.498	.930	92
		Total	6.953	1.083	200
$X_{21}$ Likely to Purchase	Less than 1 year	Indirect through broker	6.763	.702	52
		Direct to customer	7.606	.569	16
		Total	6.962	.760	68
	1 to 5 years	Indirect through broker	7.804	.710	25
		Direct to customer	7.933	.601	39
		Total	7.883	.643	64
	More than 5 years	Indirect through broker	7.919	.648	31
		Direct to customer	8.368	.825	37
		Total	8.163	.777	68
	Total	Indirect through broker	7.336	.880	108
		Direct to customer	8.051	.745	92
		Total	7.665	.893	200

**Figure 6.17**Boxplot of Purchase Outcome Measure ( $X_{19}$ ) for Groups of  $X_5$  (Distribution System) by  $X_1$  (Customer Type)

observations must be split across the six groups, hopefully in a somewhat balanced manner. The sample sizes per cell are shown above in Table 6.10.

**Adequacy of Statistical Power** The sample sizes in all but one of the cells provides enough statistical power to identify at least a large effect size with an 80 percent probability. However, the smaller sample size of 16 for customers of less than 1 year served by the direct distribution channel is of some concern. Thus, we must recognize that unless the effect sizes are substantial, the limited sample sizes in each group, even from this sample of 200 observations, may preclude the identification of significant differences. This issue becomes especially critical when examining nonsignificant difference in that the researcher should determine whether the nonsignificant result is due to insufficient effect size or low statistical power.

### STAGE 3: ASSUMPTIONS IN MANOVA

As with the prior MANOVA analyses, the assumption of greatest importance is the homogeneity of variance-covariance matrices across the groups. Meeting this assumption facilitates direct interpretation of the results without having to consider group sizes, level of covariances in the group, and so forth. Additional statistical assumptions related to the dependent variables (normality and correlation) have already been addressed in the

**Table 6.10** Sample Size per Cell for the Factorial Design of Customer Type ( $X_1$ ) and Distribution System ( $X_5$ )

$X_1$ Customer Type	$X_5$ Distribution System	
	Indirect	Direct
Less than 1 year	52	16
1 to 5 years	25	39
More than 5 years	31	37

prior examples. A final issue is the presence of outliers and the need for deletion of any observations that may distort the mean values of any group.

**Homoscedasticity** For this factorial design, six groups are involved in testing the assumption of homoscedasticity (see Table 6.11). The multivariate test (Box's M) has a nonsignificant value (.153), allowing us to accept the null hypothesis of homogeneity of variance–covariance matrices at the .05 level.

The univariate tests for the three purchase outcome variables separately are also all nonsignificant. With the multivariate and univariate tests showing nonsignificance, the researcher can proceed knowing that the assumption of homoscedasticity has been fully met.

**Outliers** The second issue involves examining observations with extreme values and the possible designation of observations as outliers with deletion from the analysis (see Figure 6.17 for  $X_{19}$ ). Interestingly enough, examination of the boxplots for the three purchase outcomes identifies a smaller number of observations with extreme values than found for  $X_1$  by itself. The dependent variable with the most extreme values was  $X_{21}$  with only three, whereas the other dependent measures had one and two extreme values. Moreover, no observation had extreme values on more than one dependent measure. As a result, all observations were retained in the analysis.

#### STAGE 4: ESTIMATION OF THE MANOVA MODEL AND ASSESSING OVERALL FIT

The MANOVA model for a factorial design tests not only for the main effects of both independent variables but also their interaction or joint effect on the dependent variables. The first step is to examine the interaction effect and determine whether it is statistically significant. If it is significant, then the researcher must confirm that the interaction effect is ordinal. If it is found to be disordinal, the statistical tests of main effects are not valid. But assuming a significant ordinal or a nonsignificant interaction effect, the main effects can be interpreted directly without adjustment.

**Assessing the Interaction Effect** Interaction effects can be identified both graphically and statistically. The most common graphical means is to create line charts depicting pairs of independent variables. As illustrated earlier in Figure 6.13, significant interaction effects are represented by non-parallel lines (with parallel lines denoting no interaction effect). If the lines depart from parallel but never cross in a significant amount, then the interaction is deemed ordinal. If the lines do cross to the degree that in at least one instance the relative ordering of the lines is reversed, then the interaction is deemed disordinal.

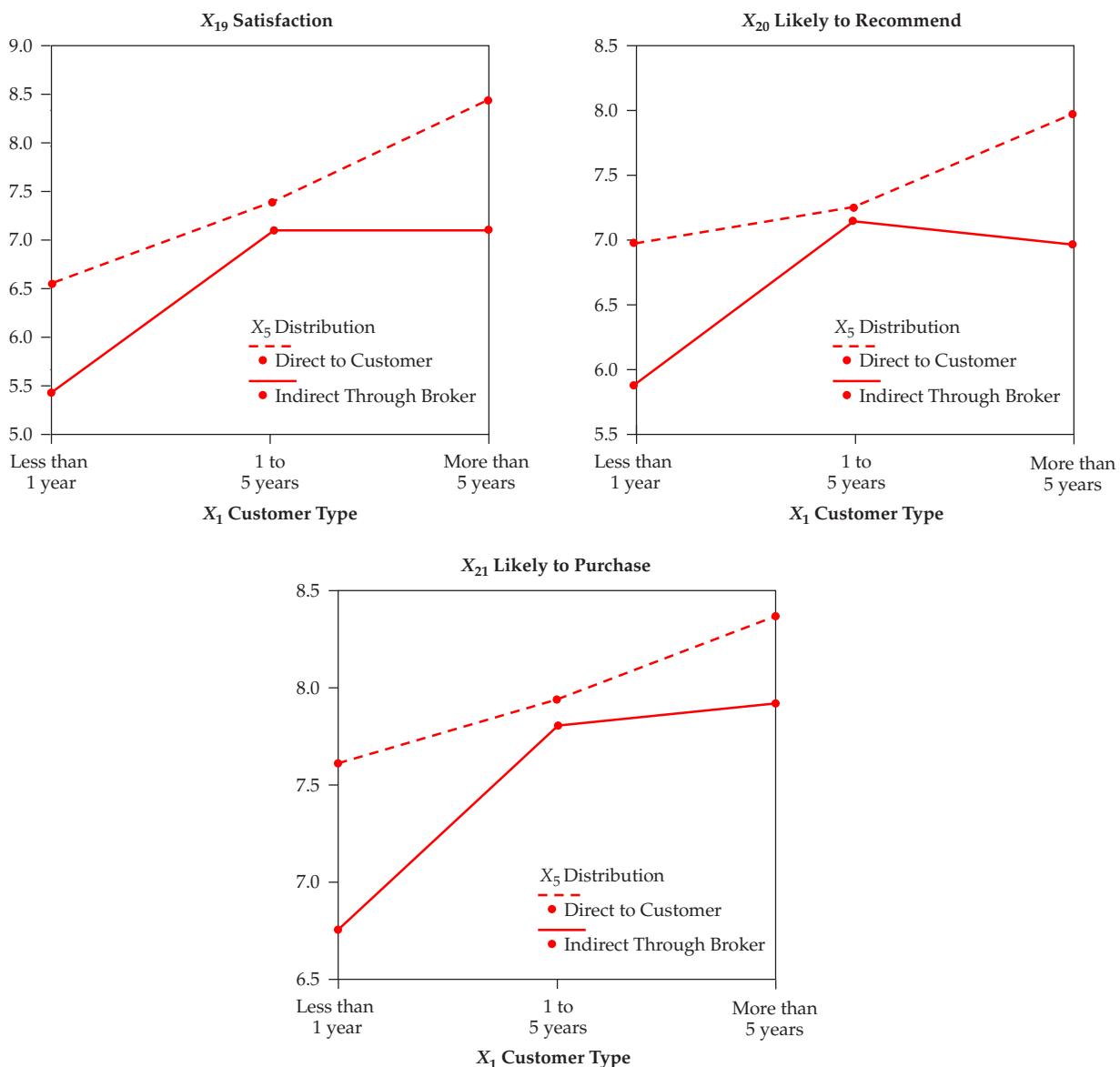
**Table 6.11** Multivariate and Univariate Measures for Testing Homoscedasticity Across Groups of  $X_1$  by  $X_5$

<b>Multivariate Tests for Homoscedasticity</b>				
<i>Box's Test of Equality of Covariance Matrices</i>				
Box's M	39.721			
F	1.263			
df1	30			
df2	33214.450			
Sig.	.153			
<b>Univariate Tests for Homoscedasticity</b>				
<i>Levene's Test of Equality of Error Variances</i>				
Dependent Variable	F	df1	df2	Sig.
$X_{19}$ Satisfaction	2.169	5	194	.059
$X_{20}$ Likely to Recommend	1.808	5	194	.113
$X_{21}$ Likely to Purchase	.990	5	194	.425

Figure 6.18 portrays each dependent variable across the six groups, indicating by the non-parallel pattern that an interaction may exist. As we can see in each graph, the middle level of  $X_1$  (1 to 5 years with HBAT) has a substantially smaller difference between the two lines (representing the two distribution channels) than the other two levels of  $X_1$ . We can confirm this observation by examining the group means from Table 6.9. Using  $X_{19}$  (Satisfaction) as an example, we see that the difference between direct and indirect distribution channels is 1.138 for customers of less than 1 year, which is quite similar to the difference between channels (1.325) for customers of greater than 5 years. However, for customers served by HBAT from 1 to 5 years, the difference between customers of the two channels is only (.285). Thus, the differences between the two distribution channels, although found to be significant in earlier examples, can be shown to differ (interact) based on how long the customer has been with HBAT. The interaction is deemed ordinal because in all instances the direct distribution channel has higher satisfaction scores.

**Figure 6.18**

Graphical Displays of Interaction Effects of  $X_1$  (Customer Type) by  $X_5$  (Distribution System) Across Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ )



**Testing the Interaction and Main Effects** In addition to the graphical means, interaction effects can also be tested in the same manner as main effects. Thus, the researcher can make a multivariate as well as univariate assessment of the interaction effect with the statistical tests described in earlier examples. Table 6.12 contains the MANOVA results for testing both the interaction and main effects.

**Table 6.12** Multivariate and Univariate Tests for Group Differences in Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) Across Groups of  $X_1$  by  $X_5$

Multivariate Tests							
Effect	Statistical Test	Value	F	Hypothesis df	Error df	Sig.	Observed Power <sup>a</sup>
$X_1$	Pillai's Criterion	.488	20.770	6	386	.000	.244
	Wilks' Lambda	.512	25.429	6	384	.000	.284
	Hotelling's $T^2$	.952	30.306	6	382	.000	.322
$X_5$	Roy's greatest characteristic root	.951	61.211	3	193	.000	.488
	Pillai's Criterion	.285	25.500	3	192	.000	.285
	Wilks' Lambda	.715	25.500	3	192	.000	.285
$X_1 \times X_5$	Hotelling's $T^2$	.398	25.500	3	192	.000	.285
	Roy's greatest characteristic root	.398	25.500	3	192	.000	.285
	Pillai's Criterion	.124	4.256	6	386	.000	.062
	Wilks' Lambda	.878	4.291	6	384	.000	.063
	Hotelling's $T^2$	.136	4.327	6	382	.000	.064
	Roy's greatest characteristic root	.112	7.194	3	193	.000	.101

<sup>a</sup>Computed using alpha = .05

Univariate Tests (Between-Subjects Effects)							
Effect	Dependent Variable	Sum of Squares	df	Mean Square	F	Sig.	Observed Power <sup>a</sup>
Overall	$X_{19}$ Satisfaction	210.999 <sup>b</sup>	5	42.200	85.689	.000	.688
	$X_{20}$ Likely to Recommend	103.085 <sup>c</sup>	5	20.617	30.702	.000	.442
	$X_{21}$ Likely to Purchase	65.879 <sup>d</sup>	5	13.176	27.516	.000	.415
$X_1$	$X_{19}$ Satisfaction	89.995	2	44.998	91.370	.000	.485
	$X_{20}$ Likely to Recommend	32.035	2	16.017	23.852	.000	.197
	$X_{21}$ Likely to Purchase	26.723	2	13.362	27.904	.000	.223
$X_5$	$X_{19}$ Satisfaction	36.544	1	36.544	74.204	.000	.277
	$X_{20}$ Likely to Recommend	23.692	1	23.692	35.282	.000	.154
	$X_{21}$ Likely to Purchase	9.762	1	9.762	20.386	.000	.095
$X_1 \times X_5$	$X_{19}$ Satisfaction	9.484	2	4.742	9.628	.000	.090
	$X_{20}$ Likely to Recommend	8.861	2	4.430	6.597	.002	.064
	$X_{21}$ Likely to Purchase	3.454	2	1.727	3.607	.029	.036

<sup>a</sup>Computed using alpha = .05

<sup>b</sup> $R^2 = .688$  (Adjusted  $R^2 = .680$ )

<sup>c</sup> $R^2 = .442$  (Adjusted  $R^2 = .427$ )

<sup>d</sup> $R^2 = .415$  (Adjusted  $R^2 = .400$ )

**TESTING INTERACTION EFFECTS** Testing for a significant interaction effect proceeds as does any other effect. First, the multivariate effects are examined and in this case all four tests show statistical significance. Then, univariate tests for each dependent variable are performed. Again, the interaction effect is also deemed significant for each of the three dependent variables. The statistical tests confirm what was indicated in the graphs: A significant ordinal interaction effect occurs between  $X_5$  and  $X_1$  for all three outcomes.

**ESTIMATING MAIN EFFECTS** If the interaction effect is deemed nonsignificant or even significant and ordinal, then the researcher can proceed to estimating the significance of the main effects for their differences across the groups. In those instances in which a disordinal interaction effect is found, the main effects are confounded by the disordinal interaction and tests for differences should not be performed.

With a significant ordinal interaction, we can proceed to assessing whether both independent variables still have significant main effects when considered simultaneously. Table 6.12 also contains the MANOVA results for the main effects of  $X_1$  and  $X_5$  in addition to the tests for the interaction effect already discussed. As we found when analyzing them separately, both  $X_1$  (Customer Type) and  $X_5$  (Distribution System) have a significant impact (main effect) on the three purchase outcome variables, both as a set and separately, as demonstrated by the multivariate and univariate tests. The results of the discriminant analysis and stepdown tests are not shown, but these produced comparable results to the univariate examples for both  $X_1$  and  $X_5$ .

The impact of the two independent variables can be compared by examining the relative effect sizes as shown by  $\eta^2$  (eta squared). The effect sizes for each variable are somewhat higher for  $X_1$  when compared to  $X_5$  on either the multivariate or univariate tests. For example, with the multivariate tests the eta squared values for  $X_1$  range from .244 to .488, but they are lower (all equal to .285) for  $X_5$ . Similar patterns can be seen on the univariate tests. This comparison gives an evaluation of practical significance separate from the statistical significance tests. When compared to either independent variable, however, the effect size attributable to the interaction effect is much smaller (e.g., multivariate eta squared values ranging from .062 to .101).

## STAGE 5: INTERPRETATION OF THE RESULTS

Interpretation of a factorial design in MANOVA is a combination of judgments drawn from statistical tests and examination of the basic data. The presence of an interaction effect can be assessed statistically, but the resulting conclusions are primarily based on the judgments of the researcher. The researcher must examine the differences for practical significance in addition to statistical significance. If specific comparisons among the groups can be formulated, then planned comparisons can be specified and tested directly in the analysis.

**Interpreting Interaction and Main Effects** Statistical significance may be supported by the multivariate tests, but examining the tests for each dependent variable provides critical insight into the effects seen in the multivariate tests. Moreover, the researcher may employ planned comparisons or even post hoc tests to determine the true nature of differences, particularly when significant interaction terms are found.

With the interaction and main effects found to be statistically significant by both the multivariate and univariate tests, interpretation is still heavily reliant on the patterns of effects shown in the values of the six groups (shown in Table 6.9 and Figure 6.18).

**INTERACTION OF  $X_1$  BY  $X_5$**  The non-parallel lines for each dependent measure notably portray the narrowing of the differences in distribution channels for customers of 1 to 5 years. Although the effects of  $X_1$  and  $X_5$  are still present, we do see some marked differences in these impacts depending on which specific sets of customers we examine. For example, for  $X_{20}$  the difference between Direct versus Indirect Distribution customers is 1.098 for customers of less than 1 year, decreases to only .107 for customers of 1 to 5 years, and then increases again to 1.007 for customers of more than 5 years. These substantial differences depending on the Customer Type illustrate the significant interaction effect.

**MAIN EFFECT OF  $X_1$**  Its main effect is illustrated for all three purchase outcomes by the upward sloping lines across the three levels of  $X_1$  on the X axis. Here we can see that the effects are consistent with earlier findings in that all three purchase outcomes increase favorably as the length of the relationship with HBAT increases. For example, again

examining  $X_{20}$ , we see that the overall mean score increases from 6.141 for customers of less than 1 year to 7.209 for customers of 1 to 5 years and finally to 7.522 for customers of more than 5 years.

**MAIN EFFECT OF  $X_5$**  The separation of the two lines representing the two distribution channels show us that the direct distribution channel generates more favorable purchase outcomes. Examining Figure 6.18 we see that for each dependent variable, the lines for customers with the direct distribution are greater than those served by the indirect system.

**Potential Covariates** The researcher also has an additional tool—adding covariates—to improve in the analysis and interpretation of the independent variables. The role of the covariate is to control for effects outside the scope of the MANOVA analysis that may affect the group differences in some systematic manner (see earlier discussion for more detail). A covariate is most effective when it has correlation with the dependent variables, but is relatively uncorrelated to the independent variables in use. In this way it can account for variance not attributable to the independent variables (due to its low correlation with them), but still reduce the amount of overall variation to be explained (the correlation with the dependent measures).

The HBAT researchers had limited options in choosing covariates for these MANOVA analyses. The only likely candidate was  $X_{22}$ , representing the customers' percentage of purchases coming from HBAT. The rationale would be to control for the perceived or actual dependence of firms on HBAT as represented in  $X_{22}$ . Firms with more dependence may react quite differently to the variables being considered.

However,  $X_{22}$  is a poor candidate for becoming a covariate even though it meets the criterion of being correlated with the dependent variables. Its fatal flaw is the high degree of differences seen on both  $X_1$  and  $X_5$ . These differences suggest that the effects of  $X_1$  and  $X_5$  would be severely confounded by the use of  $X_{22}$  as a covariate. Thus, no covariates will be utilized in this analysis.

## SUMMARY

The results reflected in both the main and interaction effects present convincing evidence that HBAT customers' postpurchase reactions are influenced by the type of distribution system and by the length of the relationship.

The direct distribution system is associated with higher levels of customer satisfaction, as well as likelihood to repurchase and recommend HBAT to others. Similarly, customers with longer relationships also report higher levels of all three dependent variables. The differences between the dependent variables are smallest among those customers who have done business with HBAT for 1 to 5 years.

The use of MANOVA in this process enables the researcher to control the Type I error rate to a far greater extent than if individual comparisons were made on each dependent variable. The interpretations remain valid even after the impact of other dependent variables has been considered. These results confirm the differences found between the effects of the two independent variables.

## Example 4: Moderation and Mediation

Having applied MANOVA to both single main effects as well as a factorial design, we now will investigate the use of mediation and moderation to extend the main effect of  $X_5$  (Distribution channel) in two aspects:

- Moderation: Does the main effect vary by firm size ( $X_3$ )?
- Mediation: Does the percentage of a customer's purchases from HBAT ( $X_{22}$ ) play an intermediary role in the main effect on the purchase outcomes?

In both instances the additional relationships will add insight into the “When” and “Why” questions concerning the effects of the distribution channel found earlier. Since we have already discussed this main effect in Example 1 and there no further analyses needed to apply moderation and mediation, the following sections will deal assume the reader is familiar with the conclusions detailed earlier regarding the main effect.

## MODERATION OF DISTRIBUTION SYSTEM ( $X_5$ ) BY FIRM SIZE ( $X_3$ )

The first additional relationship to be examined is moderation of the main effect of  $X_5$  (Distribution Channel) by  $X_3$  (Firm Size). This issue is of interest to the management team since there is the possibility that the operation and management of the distribution system may impact firms differently based on their size. While there is not any concerted effort to differentiate the operations by firm size, identification of such an effect would provide impetus for the management team to engage in further investigation.

**Group Profiles** As a first step the group means for the three outcomes were calculated including the moderator (see Table 6.13). As we can see, for example examining  $X_{19}$  Satisfaction, the differences on  $X_5$  between Indirect ( $X_5 = 0$ ) and Direct ( $X_5 = 1$ ) systems does seem to vary between Small Firms ( $7.128 - 6.211 = .917$ ) and Large Firms ( $8.449 - 6.406 = 2.043$ ). This would seem to indicate a possible interaction, with the MANOVA model being able to identify statistically significant moderating effects not only for separate outcomes, but also the outcomes collectively.

**Homoscedasticity** The only assumption that requires review is for homoscedasticity. Table 6.14 provides both the overall Box M test and tests of each outcome individually. The overall test shows nonsignificance, which indicates that the assumption of homoscedasticity is met. The individual tests generally concur, although  $X_{19}$  and  $X_{21}$  do show some indications individually of heteroscedasticity. With the overall test being nonsignificant and the group sizes being relatively large and equal, the research team can proceed without any further adjustments for this issue.

**Table 6.13** Descriptive Statistics of Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) for Groups of  $X_5$  (Distribution System) Moderated by  $X_3$  (Firm Size)

Moderator	Independent Variable		Outcomes		
	$X_5$ : Distribution System	$X_{19}$ : Satisfaction	$X_{20}$ : Likely to Recommend	$X_{21}$ : Likely to Purchase	Group Size
$X_3$ : Firm Size					
Small (0 to 499)	Indirect through broker	6.211	6.091	7.142	45
	Direct to customer	7.128	7.079	7.670	53
	Total	6.707	6.626	7.428	98
Large (500+)	Indirect through broker	6.406	6.771	7.475	63
	Direct to customer	8.449	8.067	8.569	39
	Total	7.187	7.267	7.893	102
Total	Indirect through broker	6.325	6.488	7.336	108
	Direct to customer	7.688	7.498	8.051	92
	Total	6.952	6.953	7.665	200

**Table 6.14** Multivariate and Univariate Measures for Testing Homoscedasticity Across Groups of  $X_5$  Moderated by  $X_3$

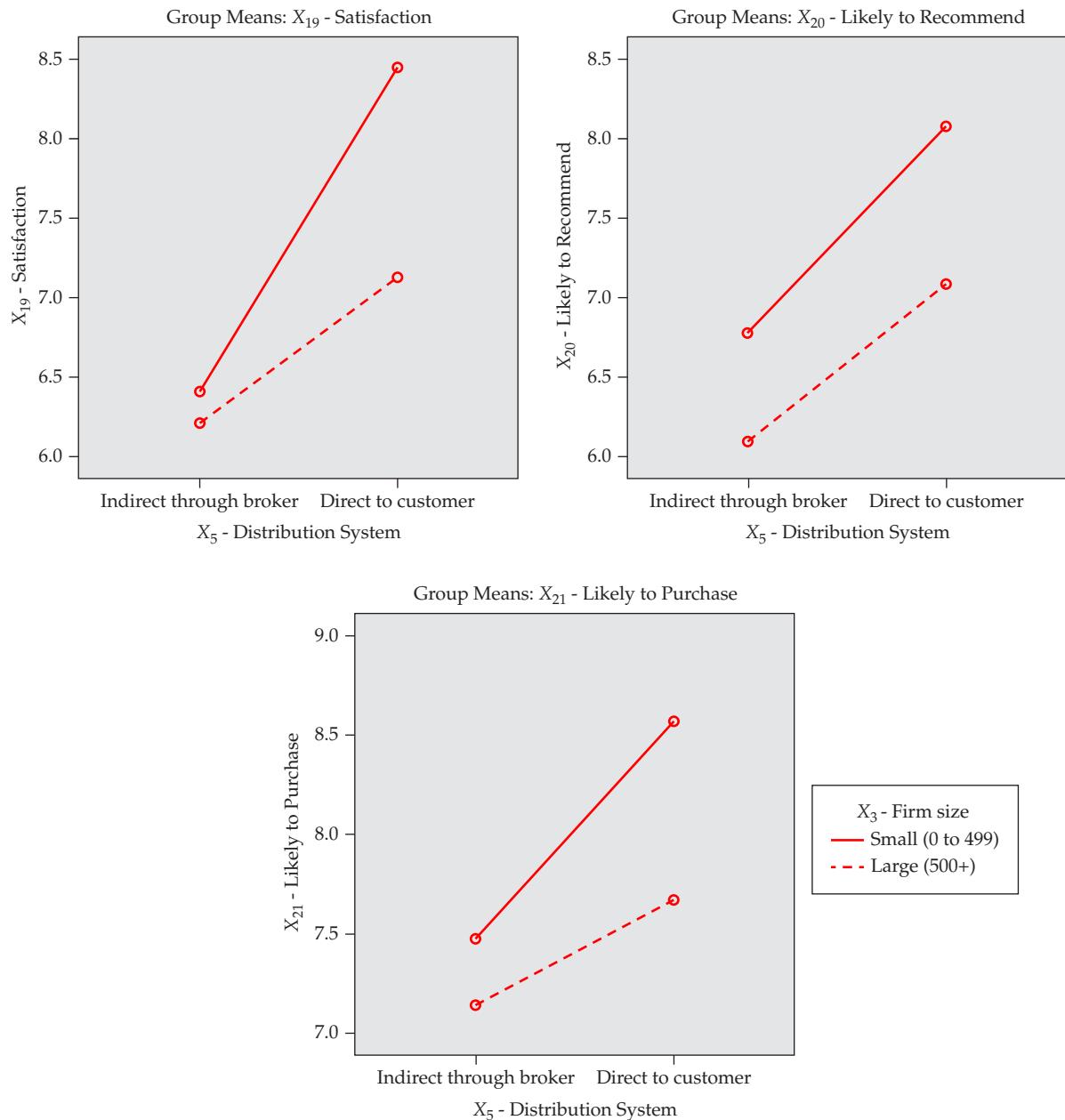
Multivariate Tests for Homoscedasticity				
Box's Test of Equality of Covariance Matrices				
Box's M	28.363			
F	1.530			
df1	18			
df2	106854.075			
Sig.	.069			
Univariate Tests for Homoscedasticity				
Levene's Test of Equality of Error Variances				
Dependent Variable	F	df1	df2	Sig.
$X_{19}$ Satisfaction	5.223	3	196	.002
$X_{20}$ Likely to Recommend	1.990	3	196	.117
$X_{21}$ Likely to Purchase	4.788	3	196	.003

**Statistical Significance Testing** The moderation effect was tested by adding  $X_5$  as an additional factor and creating the interaction term to represent the moderating effect. The multivariate and univariate results are shown in Table 6.15.

**EXAMINING INTERACTIONS** Before examining the statistical results, we need to examine the interactions to ensure that they all are ordinal so that any significant main effects can be interpreted correctly. Figure 6.19 plots all three outcomes with separate lines representing the two levels of the moderator ( $X_3$ , Firm Size). Both  $X_{19}$  and  $X_{21}$  have decidedly non-parallel lines, indicating possible interaction.  $X_{20}$  has markedly less difference between the slopes of the two lines. Also, all three interactions are ordinal (i.e., do not crossover) and thus all results can be directly interpreted.

**Figure 6.19**

Graphical Displays of Moderation Effects of  $X_3$  (Firm Size) on  $X_5$  (Distribution System) Across Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ )



**Table 6.15** Multivariate and Univariate Tests for Group Differences in Purchase Outcome Measures ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) Across Groups of  $X_5$  Moderated by  $X_3$ 

Multivariate Tests							
Effect	Statistical Test	Value	F	Hypothesis df	Error df	Sig.	Observed Power <sup>a</sup>
$X_5$	Pillai's Criterion	.394	20.770	3	194	.000	.394
	Wilks' Lambda	.606	25.429	3	194	.000	.394
	Hotelling's $T^2$	.651	30.306	3	194	.000	.394
	Roy's greatest characteristic root	.651	61.211	3	194	.000	.394
$X_3$	Pillai's Criterion	.210	25.500	3	194	.000	.210
	Wilks' Lambda	.790	25.500	3	194	.000	.210
	Hotelling's $T^2$	.265	25.500	3	194	.000	.210
	Roy's greatest characteristic root	.265	25.500	3	194	.000	.210
$X_5 \times X_3$	Pillai's Criterion	.097	4.256	3	194	.000	.097
	Wilks' Lambda	.903	4.291	3	194	.000	.097
	Hotelling's $T^2$	.107	4.327	3	194	.000	.097
	Roy's greatest characteristic root	.107	7.194	3	194	.000	.097

<sup>a</sup>Computed using alpha = .05.

Univariate Tests (Between-Subjects Effects)							
Effect	Dependent Variable	Sum of Squares	df	Mean Square	F	Sig.	Observed Power <sup>a</sup>
Overall	$X_{19}$ Satisfaction	132.472 <sup>b</sup>	3	44.157	49.721	.000	.432
	$X_{20}$ Likely to Recommend	84.720 <sup>c</sup>	3	28.240	37.238	.000	.363
	$X_{21}$ Likely to Purchase	46.471 <sup>d</sup>	3	15.490	27.035	.000	.293
$X_5$	$X_{19}$ Satisfaction	106.036	1	106.036	119.397	.000	.379
	$X_{20}$ Likely to Recommend	63.118	1	63.118	83.229	.000	.298
	$X_{21}$ Likely to Purchase	31.858	1	31.858	55.600	.000	.221
$X_3$	$X_{19}$ Satisfaction	27.810	1	27.810	31.314	.000	.138
	$X_{20}$ Likely to Recommend	33.671	1	33.671	44.399	.000	.185
	$X_{21}$ Likely to Purchase	18.369	1	18.369	32.058	.000	.141
$X_5 \times X_3$	$X_{19}$ Satisfaction	15.326	1	15.326	17.258	.000	.081
	$X_{20}$ Likely to Recommend	1.142	1	1.142	1.506	.221	.008
	$X_{21}$ Likely to Purchase	3.892	1	3.892	6.793	.010	.033

<sup>a</sup>Computed using alpha = .05.

<sup>b</sup> $R^2 = .432$  (Adjusted  $R^2 = .423$ ).

<sup>c</sup> $R^2 = .363$  (Adjusted  $R^2 = .353$ ).

<sup>d</sup> $R^2 = .293$  (Adjusted  $R^2 = .282$ ).

**MULTIVARIATE SIGNIFICANCE TESTS** The multivariate tests all indicate that both the main effect of  $X_5$  and the moderation effect of  $X_3$  (main effect and more importantly interaction) are significant. Examining the effect size ( $\eta^2$ ) indicates that while  $X_3$  is the largest, the interaction term is substantial and supports the significance test.

**UNIVARIATE SIGNIFICANCE TESTS** The univariate tests for the three outcomes presents a somewhat different set of results. The main effects of  $X_5$  and  $X_3$  are significant, but the interaction effect indicating moderation is only significant

for  $X_{19}$  and  $X_{21}$ . Examining the effect sizes reveals that the largest interaction effect occurs with  $X_{19}$ , then  $X_{21}$ , and finally an almost zero effect (.008) for  $X_{20}$ .

**Interpreting the Moderation Effect** Having demonstrated that the moderation effect does exist for at least two of the three outcomes ( $X_{19}$  and  $X_{21}$ ), the research team may wish to understand the size of this effect. We can obtain the parameter estimates for the linear models for each of the outcomes and quantify the size of the interaction effects. Table 6.16 provides the four estimated coefficients for the linear model in each outcome (they correspond to the example in Figure 6.14). Again using  $X_{19}$  as an example, we can see the size of the interaction effect (1.1252) and its differential impact on the difference between Indirect and Direct distribution channels (i.e., the right portion of the table where  $X_3 = 1$ ). The lack of a significant moderation effect is seen for  $X_{20}$  where the interaction term is relatively small (.3071), particularly when compared to the other effects. This is reflecting the small effect size discussed earlier. These match the group means seen in Table 6.13 and illustrate the source of those differences.

## SUMMARY

The test for the moderating effect of  $X_3$  (Firm Size) was supported for  $X_{19}$  (Satisfaction) and  $X_{21}$  (Likely to Purchase), although it did not moderate the main effect of  $X_5$  towards  $X_{20}$  (Likely To recommend). From these results, the management team can focus on the outcomes with significant moderation effects and also compare them to  $X_{20}$ , which did not exhibit interaction.

**Table 6.16** Group Means on Purchase Outcomes ( $X_{19}$ ,  $X_{20}$ , and  $X_{21}$ ) for  $X_5$  (Distribution System) Moderated by  $X_3$  (Firm Size)

Linear Model Estimates	Coefficients	$X_{19}$ : Satisfaction			
		Moderator: $X_3 = 0$	Moderator: $X_3 = 1$	Moderator: $X_3 = 0$	Moderator: $X_3 = 1$
Intercept ( $b_0$ )	6.211	6.211	6.211	6.211	6.211
Main effect ( $b_1$ )	0.9172		0.9172		0.9172
Moderator ( $b_2$ )	0.1952			0.1952	0.1952
Interaction ( $b_3$ )	1.1252				1.1252
Total		6.211	7.1282	6.4062	8.4486

Linear Model Estimates	Coefficients	$X_{20}$ : Likely to Recommend			
		Moderator: $X_3 = 0$	Moderator: $X_3 = 1$	Moderator: $X_3 = 0$	Moderator: $X_3 = 1$
Intercept ( $b_0$ )	6.0911	6.0911	6.0911	6.0911	6.0911
Main effect ( $b_1$ )	0.9881		0.9881		0.9881
Moderator ( $b_2$ )	0.6803			0.6803	0.6803
Interaction ( $b_3$ )	0.3071				0.3071
Total		6.0911	7.0792	6.7714	8.0666

Linear Model Estimates	Coefficients	$X_{21}$ : Likely to Purchase			
		Moderator: $X_3 = 0$	Moderator: $X_3 = 1$	Moderator: $X_3 = 0$	Moderator: $X_3 = 1$
Intercept ( $b_0$ )	7.1422	7.1422	7.1422	7.1422	7.1422
Main effect ( $b_1$ )	0.5276		0.5276		0.5276
Moderator ( $b_2$ )	0.3324			0.3324	0.3324
Interaction ( $b_3$ )	0.567				0.567
Total		7.1422	7.6698	7.4746	8.5692

Source: PROCESS Macro [44].

## MEDIATION OF DISTRIBUTION SYSTEM ( $X_5$ ) BY PURCHASE LEVEL ( $X_{22}$ )

The second additional relationship to the main effect of  $X_5$  (Distribution System) on the three outcomes is the potential mediating effect of  $X_{22}$  (Purchase Level). This question was the result of the management team's interest in the impact of supplier concentration of purchases on each of the three outcomes. Channel literature and thinking has always felt like more "dependency" of the customers on a seller would increase positive perceptions and hopefully behavior towards the seller. Thus, testing for the mediating effect of Purchase Level would investigate if it acted as a "transmission" effect for the effects of Distribution System.

**Estimation of Mediation by  $X_{22}$**  The use of MANOVA to test an overall mediation effect is not widespread in the literature, perhaps because of the interest on the individual effects rather than the collective effect and the more cumbersome estimation of effects. To this end we will test for mediation through two approaches. First, each of the outcomes will be evaluated separately for mediation. Then, principal components analysis will be used to form a composite measure of all three outcomes, which will then be used in a mediation analysis. While not strictly comparable to the MANOVA discriminant function, it should provide an acceptable substitute given the fairly high degree of correlation among the three outcomes.

**ESTIMATION OF SEPARATE MEDIATION TESTS** Each of the three outcomes was tested separately for mediation from  $X_5$  through  $X_{22}$  using the PROCESS macro [44]. Table 6.17 provides the set of mediation estimates, including the unmediated main effect of  $X_5$  on each outcome, the mediated main effect and the two paths of the mediation indirect effect ( $X_5 \rightarrow X_{22}$  and  $X_{22} \rightarrow$  outcome).

For all three outcomes the unmediated main effect of  $X_5$  on the outcome was significant. Although recent research has identified instances in which this effect did not need to be significant for mediation to occur, in all three instances it met the original requirement of Baron and Kenny. Also, for each outcome the two paths of the mediation effect were also significant, indicating that the mediation/indirect path should be significant as well. Finally, the significance level of the mediated main effect allows for categorization of the mediation as complete or partial.

The indirect effect was estimated from the bootstrapping process, which also generated the upper and lower confidence intervals. In all three outcomes the indirect effect was deemed significant at the 95 percent confidence interval.

**ESTIMATION OF COMPOSITE MEDIATION TESTS** A similar analysis was performed on a composite measure generated with principal components of the three outcome measures. All of the outcome measures had loadings of .88 or above on the factor and the factor accounted for 81 percent of the variance of the three outcome measures. Thus, the composite factor is a suitable representation of the combined outcome measures.

The mediation results for the composite (see Table 6.18) were comparable to those found for each of the individual outcomes. The unmediated main effect was significant as well as all of the effects from the mediated model. As for the individual outcome results, the indirect effect was also significant.

**Interpretation of the Mediation by  $X_{22}$**  The same general conclusions can be drawn from both mediation analyses of both the individual outcomes and the composite measure:

- There is a substantial mediation effect working through  $X_{22}$ , typically accounting for about 30 percent of the total effect of  $X_5$  on the outcome. The effect sizes are what would generally be characterized as medium in terms of their size.
- While there is a significant mediation effect, in all instances only partial mediation was achieved, since in each instance the mediated main effect was highly significant and still accounted for a majority of the overall effect.

These results provide support for the partial mediation of the effects of the distribution system on the purchase outcomes of interest. The research team can now develop plans to capitalize on this relationship as well as investigate other possible mediators to account for the remainder of the effect.

**Table 6.17 Mediation Effects of  $X_{22}$  (Purchase Level) on Main Effect of  $X_5$  (Distribution System) and Purchase Outcome Measures ( $X_{19}$ ,  $X_{21}$ , and  $X_{21}$ )**

Mediation: $X_{19}$ Satisfaction				
Mediation Effects	Parameter Estimates			
	Coefficient	Std. Error	t value	Signif.
C (unmediated main effect: $X_5 \rightarrow X_{19}$ )	1.363	0.1495	9.177	0.0000
C' (mediated main effect: $X_5 \rightarrow X_{19}$ )	0.9509	0.1124	8.4595	0.0000
A ( $X_5 \rightarrow X_{22}$ )	4.9275	1.2425	3.9658	0.0001
B ( $X_{22} \rightarrow X_{19}$ )	0.0836	0.0055	15.1855	0.0000
Indirect Effects <sup>a</sup>				
	Effect	Std. Error	Lower CI	Upper CI
Mediation (Indirect) Effects	0.4122	0.1086	0.2092	0.6339
Ratio of Indirect Effects to Total Effects	0.3024	0.0644	0.175	0.4291
Ratio of Indirect Effects to Direct Effects	0.4335	0.1361	0.2121	0.7517
Effect Size of Indirect Effect	0.1656	0.0454	0.0808	0.2569

<sup>a</sup> Bootstrapped estimates with 5,000 samples at confidence level of .95

Mediation: $X_{20}$ Likely to Recommend				
Mediation effects	Parameter Estimates			
	Coefficient	Std. Error	t value	Signif.
C (unmediated main effect: $X_5 \rightarrow X_{20}$ )	1.0099	0.1363	7.4073	0.0000
C' (mediated main effect: $X_5 \rightarrow X_{20}$ )	0.741	0.1246	5.9468	0.0000
A ( $X_5 \rightarrow X_{22}$ )	4.9275	1.2425	3.9658	0.0001
B ( $X_{22} \rightarrow X_{20}$ )	0.0545	0.0071	7.6345	0.0000
Indirect Effects <sup>a</sup>				
	Effect	Std. Error	Lower CI	Upper CI
Mediation (Indirect) Effects	0.2688	0.0755	0.1389	0.4403
Ratio of Indirect Effects to Total Effects	0.2662	0.0683	0.1503	0.4232
Ratio of Indirect Effects to Direct Effects	0.3628	0.1366	0.1769	0.7338
Effect Size of Indirect Effect	0.109	0.033	0.052	0.1816

<sup>a</sup> Bootstrapped estimates with 5,000 samples at confidence level of .95

Mediation: $X_{21}$ Likely to Purchase				
Mediation Effects	Parameter Estimates			
	Coefficient	Std. Error	t value	Signif.
C (unmediated main effect: $X_5 \rightarrow X_{21}$ )	0.7150	0.1165	6.1910	0.0000
C' (mediated main effect: $X_5 \rightarrow X_{21}$ )	0.5003	0.1018	4.9163	0.0000
A ( $X_5 \rightarrow X_{22}$ )	4.9275	1.2425	3.9658	0.0001
B ( $X_{22} \rightarrow X_{21}$ )	0.0436	0.0057	7.6784	0.0000
Indirect Effects <sup>a</sup>				
	Effect	Std. Error	Lower CI	Upper CI
Mediation (Indirect) Effects	0.2147	0.0595	0.1098	0.3441
Ratio of Indirect Effects to Total Effects	0.3003	0.0771	0.1666	0.4783
Ratio of Indirect Effects to Direct Effects	0.4291	0.1774	0.1999	0.9170
Effect Size of Indirect Effect	0.0875	0.0275	0.0396	0.1492

<sup>a</sup> Bootstrapped estimates with 5,000 samples at confidence level of .95.

Source: PROCESS Macro [44].

**Table 6.18** Mediation Effects of  $X_{22}$  (Purchase Level) on Main Effect of  $X_5$  (Distribution System) and Purchase Outcome Measures Composite

Mediation Effects	Mediation: Composite of $X_{19}$ , $X_{20}$ , and $X_{21}$			
	Coefficient	Std. Error	t value	Signif.
C (unmediated main effect: $X_5 \rightarrow$ Composite)	1.0504	0.1207	8.7052	0.0000
C' (mediated main effect: $X_5 \rightarrow$ Composite)	0.7458	0.0958	7.7847	0.0000
A ( $X_5 \rightarrow X_{22}$ )	4.9275	1.2425	3.9658	0.0001
B ( $X_{22} \rightarrow$ Composite)	0.0618	0.0051	12.0889	0.0000
Indirect Effects *				
	Effect	Std. Error	Lower CI	Upper CI
Mediation (Indirect) Effects	0.3046	0.0958	7.7847	0.0000
Ratio of Indirect Effects to Total Effects	0.2900	0.0629	0.1658	0.4162
Ratio of Indirect Effects to Direct Effects	0.4084	0.1293	0.1988	0.7130
Effect Size of Indirect Effect	0.1470	0.0401	0.0723	0.2336

\* Bootstrapped estimates with 5,000 samples at confidence level of .95.

Source: PROCESS Macro [44].

## SUMMARY

The ability of both moderation and mediation to provide additional insights into selected main effects was demonstrated in the above examples. The research team now has empirical support for the impact of firm size on the main effect as well as the partial mediating effect of purchase level. In both instances the effects were found to be consistent across outcomes (except for lack of moderation for  $X_{20}$ ). As such, more generalized approaches can be developed to cultivate these relationships as well as extend the analyses into other mediating and/or moderating effects.

## A Managerial Overview of the Results

HBAT researchers performed a series of ANOVAs and MANOVAs in an effort to understand how three purchase outcomes ( $X_{19}$ , Satisfaction;  $X_{20}$ , Likelihood of Recommending HBAT; and  $X_{21}$ , Likelihood of Future Purchase) vary across characteristics of the firms involved, such as distribution system ( $X_5$ ) and customer type ( $X_1$ ). In our discussion, we focus on the multivariate results as they overlap with the univariate results.

The first MANOVA analysis is direct: Does the type of distribution channel have an effect on the purchase outcomes? In this case the researcher tests whether the sets of mean scores (i.e., the means of the three purchase outcomes) for each distribution group are equivalent. After meeting all assumptions, we find that the results reveal a significant difference in that firms in the direct distribution system had more favorable purchase outcomes when compared to firms served through the broker-based model. Along with the overall results, management also needed to know whether this difference exists not only for the variate but also for the individual variables. Univariate tests revealed significant univariate differences for each purchase outcome as well. The significant multivariate and univariate results indicate to management that the direct distribution system serves customers better as indicated by the more favorable outcome measures. Thus, managers can focus on extending those benefits of the direct system while working on improving the broker-based distribution system.

The next MANOVA follows the same approach, but substitutes a new independent variable, customer type (i.e., the length of time the firm has been a customer), which has three groups (less than 1 year, 1 to 5 years, and more than 5 years). Once again, management focuses on the three outcome measures to assess whether significant differences are found across length of the customer relationship. Both univariate and multivariate tests show differences in the purchase outcome variables across the three groups of customers. Yet one question remains: Is each group different from the other?

Group profiles show substantial differences and post hoc tests indicate that for  $X_{19}$  (Satisfaction) each customer group is distinct from the other. For the remaining two outcome measures, groups 2 and 3 (customers of 1 to 5 years and customers more than 5 years) are not different from each other, although both are different from customers of less than 1 year. The implication is that for  $X_{20}$  and  $X_{21}$  the improvements in purchase outcomes are significant in the early years, but do not keep increasing beyond that period. From a managerial perspective, the duration of the customer relationship positively affects the firm's perceptions of purchase outcomes. Even though increases are seen throughout the relationship for the basic satisfaction measure, the only significant increase in the other two outcomes is seen after the first year.

The third example addresses the issue of the combined impact of these two firm characteristics ( $X_5$ , distribution system; and  $X_1$ , duration of the customer relationship) on the purchase outcomes. The three categories of  $X_1$  are combined with the two categories of  $X_5$  to form six groups. The objective is to establish whether the significant differences seen for each of the two firm characteristics when analyzed separately are also evident when considered simultaneously. The first step is to review the results for significant interactions: Do the purchase outcomes display the same differences between the two types of distribution systems when viewed by duration of the relationship? All three interactions were found to be significant, meaning that the differences between the direct and broker-based systems were not constant across the three groups of customers based on duration of the customer relationship. Examining the results found that the middle group (customers of 1 to 5 years) had markedly smaller differences between the two distribution systems than customers of either shorter or longer relationships. Although this pattern held for all three purchase outcomes and direct systems always were evaluated more favorably (maintaining ordinal interactions), HBAT must realize that the advantages of the direct distribution system are contingent on the length of the customer's relationship. Given these interactions, it was still found that each firm characteristic exhibited significant impacts on the outcome as was found when analyzed separately. Moreover, when considered simultaneously, the impact of each on the purchase outcomes was relatively even.

The three MANOVA analyses of both univariate and then factorial designs provided fundamental information about the main effects involved with the distribution system and tenure as a customer. While these are important findings, the application of moderation and mediation analyses for a more in-depth understanding of the effect of the distribution system was also demonstrated. The moderating effect of firm size provides the management team with insights as to the "When" (i.e., potential variability) of the main effect. And the mediation analysis supporting the role of purchase level as a partial explanation of "Why" the distribution system impacts the outcomes is useful as well. These results provide valuable perspectives on the nature of the distribution system's effect on these outcomes and provides an avenue for additional analyses as well.

In all of these situations, MANOVA enables HBAT managers to identify the significant effects of these firm characteristics on purchase outcomes, not only individually but also when combined.

Multivariate analysis of variance (MANOVA) is an extension of analysis of variance (ANOVA) to accommodate more than one dependent variable. It is a dependence technique that measures the differences for two or more metric dependent variables based on a set of categorical (nonmetric) variables acting as independent variables. This chapter helps you to do the following:

**Explain the difference between the univariate null hypothesis of ANOVA and the multivariate null hypothesis of MANOVA.** Like ANOVA, MANOVA is concerned with differences between groups (or experimental treatments). ANOVA is termed a univariate procedure because we use it to assess group differences on a single metric dependent variable. The null hypothesis is the means of the groups for a single dependent variable are equal (not statistically different). Univariate methods for assessing group differences are the  $t$  test (two groups) and analysis of variance (ANOVA) for two or more groups. The  $t$  test is widely used because it works with small group sizes and it is quite easy to apply and interpret. But its limitations include: (1) it only accommodates two groups; and (2) it can only assess one independent variable at a time. Although a  $t$  test can be performed with ANOVA, the  $F$  statistic has the ability to test

for differences between more than two groups as well as include more than one independent variable. Also, independent variables are not limited to just two levels, but instead can have as many levels (groups) as desired. MANOVA is considered a multivariate procedure because it is used to assess group differences across multiple metric dependent variables simultaneously. In MANOVA, each treatment group is observed on two or more dependent variables. Thus, the null hypothesis is the vector of means for multiple dependent variables is equal across groups. The multivariate procedures for testing group differences are the Hotelling  $T^2$  and multivariate analysis of variance, respectively.

**Discuss the advantages of a multivariate approach to significance testing compared to the more traditional univariate approaches.** As statistical inference procedures, both the univariate techniques ( $t$  test and ANOVA) and their multivariate extensions (Hotelling's  $T^2$  and MANOVA) are used to assess the statistical significance of differences between groups. In the univariate case, a single dependent measure is tested for equality across the groups. In the multivariate case, a variate is tested for equality. In MANOVA, the researcher actually has two variates, one for the dependent variables and another for the independent variables. The dependent variable variate is of more interest because the metric-dependent measures can be combined in a linear combination, as we have already seen in multiple regression and discriminant analysis. The unique aspect of MANOVA is that the variate optimally combines the multiple dependent measures into a single value that maximizes the differences across groups. To analyze data on multiple groups and variables using univariate methods, the researcher might be tempted to conduct separate  $t$  tests for the difference between each pair of means (i.e., group 1 versus group 2; group 1 versus group 3; and group 2 versus group 3). But multiple  $t$  tests inflate the overall Type I error rate. ANOVA and MANOVA avoid this Type I error inflation due to making multiple comparisons of treatment groups by determining in a single test whether the entire set of sample means suggests that the samples were drawn from the same general population. That is, both techniques are used to determine the probability that differences in means across several groups are due solely to sampling error.

**State the assumptions for the use of MANOVA.** The univariate test procedures of ANOVA are valid in a statistical sense if we assume that the dependent variable is normally distributed, the groups are independent in their responses on the dependent variable, and that variances are equal for all treatment groups. There is evidence, however, that  $F$  tests in ANOVA are robust with regard to these assumptions except in extreme cases. For the multivariate test procedures of MANOVA to be valid, three assumptions must be met: (1) observations must be independent, (2) variance-covariance matrices must be equal for all treatment groups, and (3) the set of dependent variables must follow a multivariate normal distribution. In addition to these assumptions, the researcher must consider two issues that influence the possible effects—the linearity and multicollinearity of the variate of the dependent variables.

**Understand how to interpret MANOVA results.** If the treatments result in statistically significant differences in the vector of dependent variable means, the researcher then examines the results to understand how each treatment impacts the dependent measures. Three steps are involved: (1) interpreting the effects of covariates, if included; (2) assessing which dependent variable(s) exhibited differences across the groups of each treatment; and (3) identifying if the groups differ on a single dependent variable or the entire dependent variate. When a significant effect is found, we say that there is a main effect, meaning that there are significant differences between the dependent variables of the two or more groups defined by the treatment. With two levels of the treatment, a significant main effect ensures that the two groups are significantly different. With three or more levels, however, a significant main effect does not guarantee that all three groups are significantly different, instead just that there is at least one significant difference between a pair of groups. If there is more than one treatment in the analysis, the researcher must examine the interaction terms to see if they are significant, and if so, do they allow for an interpretation of the main effects or not. If there are more than two levels for a treatment, then the researcher must perform a series of additional tests between the groups to see which pairs of groups are significantly different.

**Describe the purpose of post hoc tests in ANOVA and MANOVA.** Although the univariate and multivariate tests of ANOVA and MANOVA enable us to reject the null hypothesis that the groups' means are all equal, they do not pinpoint where the significant differences lie if there are more than two groups. Multiple  $t$  tests without any form of adjustment are not appropriate for testing the significance of differences between the means of paired groups because the probability of a Type I error increases with the number of intergroup comparisons made (similar to the problem of using multiple

univariate ANOVAs versus MANOVA). If the researcher wants to systematically examine group differences across specific pairs of groups for one or more dependent measures, two types of statistical tests should be used: post hoc and a priori. Post hoc tests examine the dependent variables between all possible pairs of group differences that are tested after the data patterns are established. A priori tests are planned from a theoretical or practical decision-making viewpoint prior to looking at the data. The principal distinction between the two types of tests is that the post hoc approach tests all possible combinations, providing a simple means of group comparisons but at the expense of lower power. A priori tests examine only specified comparisons, so that the researcher must explicitly define the comparison to be made, but with a resulting greater level of power. Either method can be used to examine one or more group differences, although the a priori tests also give the researcher control over the types of comparisons made between groups.

**Interpret interaction results when more than one independent variable is used in MANOVA.** The interaction term represents the joint effect of two or more treatments. Any time a research design has two or more treatments, the researcher must first examine the interactions before any statement can be made about the main effects. Interaction effects are evaluated with the same criteria as main effects. If the statistical tests indicate that the interaction is nonsignificant, this denotes that the effects of the treatments are independent. Independence in factorial designs means that the effect of one treatment (i.e., group differences) is the same for each level of the other treatment(s) and that the main effects can be interpreted directly. If the interactions are deemed statistically significant, it is critical that the researcher identify the type of interaction (ordinal versus disordinal), because this has direct bearing on the conclusion that can be drawn from the results. Ordinal interaction occurs when the effects of a treatment are not equal across all levels of another treatment, but the group difference(s) is always the same direction. Disordinal interaction occurs when the differences between levels “switch” depending on how they are combined with levels from another treatment. Here the effects of one treatment are positive for some levels and negative for other levels of the other treatment.

**Describe the purpose of multivariate analysis of covariance (MANCOVA).** Covariates can play an important role by including metric variables into a MANOVA or ANOVA design. However, since covariates act as a “control” measure on the dependent variate, they must be assessed before the treatments are examined. The most important role of the covariate(s) is the overall impact in the statistical tests for the treatments. The most direct approach to evaluating these impacts is to run the analysis with and without the covariates. Effective covariates will improve the statistical power of the tests and reduce within-group variance. If the researcher does not see any substantial improvement, then the covariates may be eliminated, because they reduce the degrees of freedom available for the tests of treatment effects. This approach also can identify those instances in which the covariate is “too powerful” and reduces the variance to such an extent that the treatments are all nonsignificant. Often this occurs when a covariate is included that is correlated with one of the independent variables and thus “removes” this variance, thereby reducing the explanatory power of the independent variable. Because MANCOVA and ANCOVA are applications of regression procedures within the analysis of variance method, assessing the impact of the covariates on the dependent variables is quite similar to examining regression equations. If the overall impact is deemed significant, then each covariate can be examined for the strength of the predictive relationship with the dependent measures. If the covariates represent theoretically-based effects, then these results provide an objective basis for accepting or rejecting the proposed relationships. In a practical vein, the researcher can examine the impact of the covariates and eliminate those with little or no effect.

It is often unrealistic to assume that a difference between experimental treatments will be manifested only in a single measured dependent variable. Many researchers handle multiple-criterion situations by repeated application of individual univariate tests until all the dependent variables are analyzed. This approach can seriously inflate Type I error rates, and it ignores the possibility that some composite of the dependent variables may provide the strongest evidence of group differences. MANOVA can solve both of these problems.

**Describe the uses and insights gained from incorporating mediation and moderation effects.** While the main effects are a primary interest in MANOVA, the use of mediation and moderation can provide support for an extended conceptual model in terms of both internal and external validity. Mediation can help address the “Why” question concerning how the treatment impacts the outcome. Understanding the mediation effect many times provides the researcher with additional pathways to achieving the outcome (i.e., influencing the mediator directly or through other

factors as well). Moderation addresses the “When” question which extends the external validity of the main effect. Is the main effect still operant under differing situations (i.e., different values of the moderator variable)? Questions of this nature help provide a better understanding of the potential variation in the main effect that is not possible otherwise. The researcher should note the similarity of adding a moderator and adding another treatment to create a factorial design. Both generate interaction terms and result in the same empirical results. For both mediation and moderation it is only through conceptual model support that their impacts can be determined.

**Identify situations in which causal inferences may be strengthened even in nonrandomized studies.** The methods of ANOVA and MANOVA have long been associated with experimental research, although they have been widely applied in non-experimental situations as well. In the non-experimental situation the researcher is faced with numerous challenges to making the types of causal inferences possible with experimental research. The development of causal inference techniques for observational data extends the range of research questions that can be addressed while still maintaining the strict conditions needed to assess causal effects. The potential outcomes framework of Rubin provides the conceptualization of how the effects of random assignment operate and, more importantly, how they can be extended to non-experimental/observational data. The propensity score model is the application of a logistic regression model to estimate a variate of variables representing confounders of the causal effect. Once the logistic regression model provides equivalence between the treatment and control groups, the predicted probability score can be used as a propensity score. Observations from the treatment or control group are “matched” to observations of the opposite group that have identical or very similar propensity scores. The process equates each matched set on the covariates, thus ensuring they are comparable and can be directly compared. Once all the cases are matched, comparisons between treatment and control groups are made for each matched set of observations and then combined into an overall effect. The result is a conceptually-based model whereby observations are matched so that they meet the necessary assumptions for causal inferences to be made. In this manner observational or other non-experimental data that was not obtained with random assignment can be analyzed and causal effects estimated.

What are the differences between MANOVA and discriminant analysis? What situations best suit each multivariate technique?

Design a two-way factorial MANOVA experiment. What are the different sources of variance in your experiment? What would a significant interaction tell you?

Besides the overall, or global, significance, at least three approaches to follow-up tests include (a) use of Scheffé contrast procedures; (b) stepdown analysis, which is similar to stepwise regression in that each successive  $F$  statistic is computed after eliminating the effects of the previous dependent variables; and (c) examination of the discriminant functions. Describe the practical advantages and disadvantages of each of these approaches.

How is statistical power affected by statistical and research design decisions? How would you design a study to ensure adequate power?

Describe some data analysis situations in which MANOVA and MANCOVA would be appropriate in your areas of interest. What types of uncontrolled variables or covariates might be operating in each of these situations?

Explain the appropriate application of mediation and moderation? What type of conceptual support is needed for each?

What are the unique effects on the main effect and how do they aid in a better understanding of the main effect?

When is causal inference applicable? What is required for the researcher to be assured that the application of these techniques does provide the ability to make causal inferences from observational/nonrandomized designs?

What types of variables should be included in the propensity score model?

How is the logistic regression model applied? What is the purpose of “predicting” the treatment variable?

What is meant by matching or stratifying and what role does that play in being able to make causal inferences?

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 Aguinis, H., J. R. Edwards, and K. J. Bradley. 2017. Improving our Understanding of Moderation and Mediation in Strategic Management Research. *Organizational Research Methods* 20: 665–85.
- 2 Anderson, E. T. (2011). A Step-By-Step Guide to Smart Business Experiments. *Development and Learning in Organizations: An International Journal* 25(6).
- 3 Anderson, T. W. 2003. *Introduction to Multivariate Statistical Analysis*, 3rd ed. New York: Wiley.
- 4 Antonakis, J., S. Bendahan, P. Jacquart, and R. Lalivé. 2010. On Making Causal Claims: A Review and Recommendations. *Leadership Quarterly* 21: 1086–120.
- 5 Baron, R. M., and D. A. Kenny. 1986. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology* 51: 1173–82.
- 6 Blom-Hansen, J., R. Morton, and S. Serritzew. 2015. Experiments in Public Management Research. *International Public Management Journal* 18: 151–70.
- 7 Borgen, F. H., and M. J. Seling. 1978. Uses of Discriminant Analysis Following MANOVA: Multivariate Statistics for Multivariate Purposes. *Journal of Applied Psychology* 63: 689–97.
- 8 Box, G. E., J. S. Hunter, and W. G. Hunter. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*, Vol. 2. New York: Wiley-Interscience.
- 9 Bray, J. H., and S. E. Maxwell. 1982. Analyzing and Interpreting Significant MANOVAs. *Review of Educational Research* 52: 340–67.
- 10 Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. 2006. Variable Selection for Propensity Score Models. *American Journal of Epidemiology* 163: 1149–56.
- 11 Brookhart, M. A., T. Stürmer, R. J. Glynn, J. Rassen, and S. Schneeweiss. 2010. Confounding Control in Healthcare Database Research: Challenges and Potential Approaches. *Medical Care* 48: S114–20.
- 12 Brooks, J. M., and R. L. Ohsfeldt. 2013. Squeezing the Balloon: Propensity Scores and Unmeasured Covariate Balance. *Health Services Research* 48: 1487–507.
- 13 Campbell, D. T. 1957. Factors Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin* 54: 297–312.
- 14 Casteloe, J. 2014. Power and Sample Size for MANOVA and Repeated Measures with the GLMPOWER Procedure. In *Proceedings of the SAS Global Forum 2014 Conference*. Cary, NC: SAS Institute Inc.
- 15 Cattell, R. B. (ed.). 1966. *Handbook of Multivariate Experimental Psychology*. Chicago: Rand McNally.
- 16 Chan, D. 2011. Advances in Statistical Analytical Strategies for Causal Inferences in the Social and Behavioural Sciences. *Information Knowledge Systems Management* 10: 261–78.
- 17 Chatterji, A. K., M. Findley, N. M. Jensen, S. Meier, and D. Nielson. 2016. Field Experiments in Strategy Research. *Strategic Management Journal* 37: 116–32.
- 18 Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 19 Cole, D. A., S. E. Maxwell, R. Arvey, and E. Salas. 1994. How the Power of MANOVA Can Both Increase and Decrease as a Function of the Intercorrelations Among the Dependent Variables. *Psychological Bulletin* 115: 465–74.
- 20 Cole, D. A., S. E. Maxwell, R. Avery, and E. Salas. 1994. How the Power of MANOVA Can Both Increase and Decrease as a Function of the Intercorrelations Among Dependent Variables. *Psychological Bulletin* 115: 465–74.
- 21 Collins, L. M., J. J. Dziak, and R. Li. 2009. Design of Experiments with Multiple Independent Variables: A Resource Management Perspective on Complete and Reduced Factorial Designs. *Psychological Methods* 14: 202–24.
- 22 Collins, L. M., J. W. Graham, and B. P. Flaherty. 1998. An Alternative Framework for Defining Mediation. *Multivariate Behavioral Research* 33: 295–313.
- 23 Cook, T. D., and P. M. Steiner. 2010. Case Matching and the Reduction of Selection Bias in Quasi-experiments: The Relative Importance of Pretest Measures of Outcome, of Unreliable Measurement, and of Mode of Data Analysis. *Psychological Methods* 15: 56–68.
- 24 Cooley, W. W., and P. R. Lohnes. 1971. *Multivariate Data Analysis*. New York: Wiley.
- 25 Cox, D. R. 1958. *Planning of Experiments*. New York: Wiley.
- 26 David G. Herr. 1986. On the History of ANOVA in Unbalanced, Factorial Designs: The First 30 Years. *American Statistician* 40: 265–70.
- 27 DeFond, M., D. H. Erkens, and J. Zhang. 2016. Do Client Characteristics Really Drive the Big N Audit Quality Effect? New Evidence from Propensity Score Matching. *Management Science* 63: 3628–49.
- 28 Fairchild, A. J., and D. P. MacKinnon. 2009. A General Model for Testing Mediation and Moderation Effects. *Prevention Science: The Official Journal of the Society for Prevention Research* 10: 87–99.

- 29 Faul, F., E. Erdfelder, A. Buchner, and A.-G. Lang. 2009. Statistical Power Analyses Using G\*Power 3.1: Tests for Correlation and Regression Analyses. *Behavior Research Methods* 41: 1149–60.
- 30 Floyd, E., and J. A. List. 2016. Using Field Experiments in Accounting and Finance. *Journal of Accounting Research* 54: 437–75.
- 31 Frazier, P. A., A. P. Tix, and K. E. Baron. 2004. Testing Moderator and Mediator Effects in Counselling Psychology. *Journal of Counselling Psychology* 51: 115–134.
- 32 Green, K. M., and E. A. Stuart. 2014. Examining Moderation Analyses in Propensity Score Methods: Application to Depression and Substance Use. *Journal of Consulting and Clinical Psychology* 82: 773–83.
- 33 Green, P. E. 1978. *Analyzing Multivariate Data*. Hinsdale, IL: Holt, Rinehart and Winston.
- 34 Green, P. E., and D. S. Tull. 1979. *Research for Marketing Decisions*, 3rd edn. Upper Saddle River, NJ: Prentice Hall.
- 35 Green, P. E., and J. Douglas Carroll. 1978. *Mathematical Tools for Applied Multivariate Analysis*. New York: Academic Press.
- 36 Greenland, S. 2000. Causal Analysis in the Health Sciences. *Journal of the American Statistical Association* 95: 286–9.
- 37 Greenland, S. 2000. Causal analysis in the health sciences. *Journal of the American Statistical Association* 95: 286–9.
- 38 Greenland, S. 2003. Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias. *Epidemiology* 14: 300–6.
- 39 Grice, J. W., and M. Iwasaki. 2007. A Truly Multivariate Approach to MANOVA. *Applied Multivariate Research* 12: 199–226.
- 40 Guo, S., and M. W. Fraser. 2015. *Propensity Score Analysis*, 2nd edn. Thousand Oaks, CA: Sage.
- 41 H. S. Steyn Jr. and S. M. Ellis. 2009. Estimating an Effect Size in One-Way Multivariate Analysis of Variance (MANOVA). *Multivariate Behavioral Research* 44: 106–29.
- 42 Hand, D. J., and C. C. Taylor. 1987. *Multivariate Analysis of Variance and Repeated Measures*. London: Chapman and Hall.
- 43 Harris, R. J. 2001. *A Primer of Multivariate Statistics*, 3rd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 44 Hayes, A. F. 2017. *An Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, 2nd edn. New York: Guilford Press.
- 45 Hayes, A. F., and N. J. Rockwood. 2016. Regression-Based Statistical Mediation and Moderation Analysis in Clinical Research: Observations, Recommendations, and Implementation. *Behaviour Research and Therapy* 98: 39–57.
- 46 Hernán, M. A., and J. M. Robins. 2010. *Causal Inference*. Boca Raton, FL: CRC Press.
- 47 Höfler, M. 2005. Causal Inference Based on Counterfactuals. *BMC Medical Research Methodology* 5: 28.
- 48 Holland, P. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association* 81: 945–60.
- 49 Hubert, C. J., and J. D. Morris. 1989. Multivariate Analysis Versus Multiple Univariate Analyses. *Psychological Bulletin* 105: 302–8.
- 50 Huberty, C. J., and J. D. Morris. 1989. Multivariate Analysis Versus Multiple Univariate Analyses. *Psychological Bulletin* 105: 302–8.
- 51 Huberty, C. J., and S. Olejnik. 2006. *Applied MANOVA and Discriminant Analysis*, 2nd edn. Hoboken, NJ: Wiley.
- 52 Huitema, B. 1980. *The Analysis of Covariance and Alternatives*. New York: Wiley.
- 53 Iacus, S. M., G. King, and G. Porro. 2012. Causal Inference Without Balance Checking: Coarsened Exact Matching. *Political Analysis* 20: 1–24.
- 54 Imai, K., G. King, and E. A. Stuart. 2008. Misunderstandings Between Experimentalists and Observationalists About Causal Inference. *Journal of the Royal Statistical Society: Series A* 171: 481–502.
- 55 Imai, K., L. Keele, D. Tingley, and T. Yamamoto. 2011. Unpacking the Black Box of Causality: Learning About Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review* 105: 765–89.
- 56 Imai, K., L. Keele, D. Tingley, and T. Yamamoto. 2011. Unpacking the Black Box of Causality: Learning About Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review* 105: 765–89.
- 57 Imbens, G. W. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics* 86: 4–29.
- 58 Imbens, G. W. 2010. An Economist's Perspective on Shadish (2010) and West and Thoemmes (2010). *Psychological Methods* 15: 47–55.
- 59 Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.
- 60 James, L. R., S. A. Mulaik, and J. M. Brett. 2006. A Tale of Two Methods. *Organizational Research Methods* 9: 233–44.
- 61 Jose, P. E. 2013. *Doing Statistical Mediation and Moderation*. New York: Guilford Press.
- 62 Judd, C. M., and D. A. Kenny. 1981. Process Analysis: Estimating Mediation in Treatment Evaluation. *Evaluation Review* 5: 602–19.
- 63 Kenny, D. A. 1979. *Correlation and Causality*. New York: Wiley.
- 64 Kenny, D. A. 2008. Reflections on Mediation. *Organizational Research Methods* 11: 353–58.
- 65 Kenny, D. A., and C. M. Judd. 2014. Power Anomalies in Testing Mediation. *Psychological Science* 25: 334–39.
- 66 Kenny, D. A., D. A. Kashy, and N. Bolger. 1998. Data Analysis in Social Psychology. In D. Gilbert, S. Fiske, and G. Lindzey (eds.), *The Handbook of Social Psychology*, Vol. 1, 4th edn. Boston, MA: McGraw-Hill., pp. 233–65.

- 67 Kirk, R. E. 1994. *Experimental Design: Procedures for the Behavioral Sciences*, 3rd edn. Belmont, CA: Wadsworth Publishing.
- 68 Koslowsky, M., and T. Caspy. 1991. Stepdown Analysis of Variance: A Refinement. *Journal of Organizational Behavior* 12: 555–9.
- 69 Kraemer, H. C., and S. Thiemann. 1987. *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage.
- 70 Läuter, J. 1978. Sample Size Requirements for the  $T^2$  Test of MANOVA (Tables for One-Way Classification). *Biometrical Journal* 20: 389–406.
- 71 Lawrence, A., M. Minutti-Meza, and P. Zhang. 2011. Can Big 4 Versus Non-Big 4 Differences in Audit-Quality Proxies Be Attributed to Client Characteristics? *Accounting Review* 86: 259–86.
- 72 LeBreton, J. M., J. Wu, and M. N. Bing. 2009. The Truth(s) on Testing for Mediation in the Social and Organizational Sciences. In C. E. Lance and R. J. Vandenberg (eds.), *Statistical and Methodological Myths and Urban Legends*. New York: Routledge, pp. 109–44.
- 73 Lee, J., and T. D. Little. 2017. A Practical Guide to Propensity Score Analysis for Applied Clinical Research. *Behaviour Research and Therapy* 98: 76–90.
- 74 Lee, S., M. K. Lei, and G. H. Brody. 2015. Confidence Intervals for Distinguishing Ordinal and Disordinal Interactions in Multiple Regression. *Psychological Methods* 20: 245–58.
- 75 Li, M. 2012. Using the Propensity Score Method to Estimate Causal Effects: A Review and Practical Guide. *Organizational Research Methods* 16: 188–226.
- 76 Lubin, A. 1961. The Interpretation of Significant Interaction. *Educational and Psychological Measurement* 21: 807–17.
- 77 MacKinnon, D. P. 2008. *Introduction to Statistical Mediation Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 78 MacKinnon, D. P., and A. G. Pirlott. 2015. Statistical Approaches for Enhancing Causal Interpretation of the M to Y Relation in Mediation Analysis. *Personality and Social Psychology Review* 19: 30–43.
- 79 MacKinnon, D. P., G. Warsi, and J. H. Dwyer. 1995. A Simulation Study of Mediated Effect Measures. *Multivariate Behavioral Research* 30: 41–62.
- 80 MacKinnon, D. P., S. Coxe, and A. N. Baraldi. 2012. Guidelines for the Investigation of Mediating Variables in Business Research. *Journal of Business and Psychology* 27: 1–14.
- 81 MacKinnon, D. P., Y. Kisbu-Sakarya, A. C. Gottschall. 2013a. Developments in Mediation Analysis. In T. D. Little (ed.), *The Oxford Handbook of Quantitative Methods*, Vol. 2. New York: Oxford University Press, pp. 338–60.
- 82 Manganaris, S., R. Bhasin, M. Reid, and K. B. Hermiz. 2010. Analyzing Causal Effects with Observational Studies for Evidence-Based Marketing at IBM. *Review of Marketing Science* 8: 1–19.
- 83 Mathieu, J. E., R. P. DeShon, and D. D. Bergh. 2008. Mediation Inferences in Organizational Research. *Organizational Research Methods* 11: 203–23.
- 84 Mayer, A., F. Thoemmes, N. Rose, R. Steyer, and S. G. West. 2014. Theory and Analysis of Total, Direct, and Indirect Causal Effects. *Multivariate Behavioral Research* 49: 425–42.
- 85 Meyer, Robert. 2017. Introduction to the Special Section on Field Experiments. *Journal of Marketing Research* 54: 138–9.
- 86 Meyers, J. L. 1979. *Fundamentals of Experimental Design*. Boston, MA: Allyn and Bacon.
- 87 Montgomery, D. C. 2017. *Design and Analysis of Experiments*. New York: Wiley.
- 88 Morgan, S. L., and C. Winship. 2014. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.
- 89 Morrison, D. F. 2002. *Multivariate Statistical Methods*, 4th edn. Belmont, CA: Duxbury Press.
- 90 Nair, H. S., S. Misra, W. J. Hornbuckle IV, R. Mishra, and A. Acharya. 2017. Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation. *Marketing Science* 36: 699–725.
- 91 Oppenheimer, D. M., T. Meyvis, and N. Davidenko. 2009. Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology* 45: 867–72.
- 92 Pearl, J. 2009. *Causality*. Cambridge: Cambridge University Press.
- 93 Preacher, K. J., P. J. Curran, and D. J. Bauer. 2006. Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis. *Journal of Educational and Behavioral Statistics* 31: 437–48.
- 94 Rao, C. R. 1978. *Linear Statistical Inference and Its Application*, 2nd edn. New York: Wiley.
- 95 Rosenbaum, P. R., and D. B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79: 516–24.
- 96 Rubin, D. B. (2007). The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials. *Statistics in Medicine* 26: 20–36.
- 97 Rubin, D. B. 1976. Inference and Missing Data (with Discussion). *Biometrika* 63: 581–92.
- 98 Rubin, D. B. 2005. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association* 100: 322–31.
- 99 Rubin, D. B. 2008. For Objective Causal Inference, Design Trumps Analysis. *Annals of Applied Statistics* 2: 808–40.

- 100 Rubin, D. B. 2008. For Objective Causal Inference, Design Trumps Analysis. *Annals of Applied Statistics* 2: 808–40.
- 101 Rubin, Donald. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66: 688–701.
- 102 Shadish, W. R. 2010. Campbell and Rubin: A Primer and Comparison of their Approaches to Causal Inference in Field Settings. *Psychological Methods* 15: 3–17.
- 103 Shadish, William R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Wadsworth Cengage Learning.
- 104 Shadish, William R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Wadsworth Cengage Learning.
- 105 Shipman, J. E., Q. T. Swanquist, and R. L. Whited. 2016. Propensity Score Matching in Accounting Research. *Accounting Review* 92: 213–44.
- 106 Sloan, L., and A. Quan-Haase (eds.) 2017. *The SAGE Handbook of Social Media Research Methods*. Newbury Park, CA: Sage.
- 107 Sobel, M. E. 1982. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. In S. Leinhardt (ed.), *Sociological Methodology 1982*. Washington DC: American Sociological Association, pp. 290–312.
- 108 Sobel, M. E. 1995. Causal Inference in the Social and Behavioral Sciences. In G. Arminger, C. C. Clogg, and M. E. Sobel, *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Springer, pp. 1–38.
- 109 Srivastava, R. P., T. J. Mock, K. V. Pincus, and A. M. Wright. 2012. Causal Inference in Auditing: A Framework. *Auditing* 31: 177–201.
- 110 Stevens, J. P. 1972. Four Methods of Analyzing Between Variations for the  $k$ -Group MANOVA Problem. *Multivariate Behavioral Research* 7: 442–54.
- 111 Stevens, J. P. 1980. Power of the Multivariate Analysis of Variance Tests. *Psychological Bulletin* 88: 728–37.
- 112 Stuart, E. A. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25: 1–21.
- 113 Tatsuoka, M. M. 1988. *Multivariate Analysis: Techniques for Education and Psychological Research*, 2nd edn. New York: Macmillan.
- 114 Thomas, D. R., and B. D. Zumbo. 1996. Using a Measure of Variable Importance to Investigate the Standardization of Discriminant Coefficients. *Journal of Educational and Behavioral Statistics* 21: 110–30.
- 115 Tonidandel, S., and J. M. LeBreton. 2013. Beyond Step-Down Analysis: A New Test for Decomposing the Importance of Dependent Variables in MANOVA. *Journal of Applied Psychology* 98: 469–77.
- 116 Vacha-Haase, T., and B. Thompson. 2004. How to Estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology* 51: 473–81.
- 117 Vancouver, J. B., and B. W. Carlson. 2015. All Things in Moderation, Including Tests of Mediation (at Least Some of the Time). *Organizational Research Methods* 18: 70–91.
- 118 VanderWeele, T. J., and I. Shpitser. 2011. A New Criterion for Confounder Selection. *Biometrics* 67: 1406–13.
- 119 VanderWeele, T. J., and M. J. Knol. 2014. A Tutorial on Interaction. *Epidemiologic Methods* 3: 33–72.
- 120 Ward, A., and P. J. Johnson. 2008. Addressing Confounding Errors When Using Non-experimental, Observational Data to Make Causal Claims. *Synthese* 163: 419–32.
- 121 Warner, R. M. 2012. *Applied Statistics: From Bivariate Through Multivariate Techniques*. Newbury Park, CA: Sage.
- 122 Wegener, D., and L. Fabrigar. 2000. Analysis and Design for Nonexperimental Data Addressing Causal and Noncausal Hypothesis. In H. T. Reis and C. M. Judd (eds.), *Handbook of Research Methods in Social and Personality Psychology*. New York: Cambridge University Press, pp. 412–50.
- 123 West, S. G., and F. Thoemmes. 2010. Campbell's and Rubin's Perspectives on Causal Inference. *Psychological Methods* 15: 18–37.
- 124 Widaman, K. F., J. L. Helm, L. Castro-Schilo, M. Pluess, M. C. Stallings, and J. Belsky. 2012. Distinguishing Ordinal and Disordinal Interactions. *Psychological Methods* 17: 615–22.
- 125 Wilks, S. S. 1932. Certain Generalizations in the Analysis of Variance. *Biometrika* 24: 471–94.
- 126 Winer, B. J., D. R. Brown, and K. M. Michels. 1991. *Statistical Principles in Experimental Design*, 3rd edn. New York: McGraw-Hill.
- 127 Wooldridge, J. M. 2016. Should Instrumental Variables be Used as Matching Variables? *Research in Economics* 70: 232–7.
- 128 Wu, A. D., and B. D. Zumbo. 2008. Understanding and Using Mediators and Moderators. *Social Indicators Research* 87: 367–92.
- 129 Wyss, R., C. J. Girman, R. J. LoCasale, M. A. Brookhart, and T. Stürmer. 2013. Variable Selection for Propensity Score Models When Estimating Treatment Effects on Multiple Outcomes: a Simulation Study. *Pharmacoepidemiology and Drug Safety* 22: 77–85.
- 130 Zhao, X., J. G. Lynch Jr, and Q. Chen. 2010. Reconsidering Baron and Kenny: Myths and Truths About Mediation Analysis. *Journal of Consumer Research* 37: 197–206.



# Dependence techniques – non-metric outcomes

**Multiple Discriminant Analysis**

**Logistic Regression: Regression  
with a Binary Dependent Variable**

## SECTION IV

### OVERVIEW

**A**s discussed in Section 3, dependence techniques are an essential element of multivariate analysis. While many types of research questions are based on metric outcomes and thus addressed by the techniques in Section 3, an entirely different class of problems are based on nonmetric outcomes. Many times termed classification problems, the nature of a nonmetric outcome requires that the technique generate a prediction of a group rather than a specific metric value. In this situation, the researcher is attempting to classify observations into groups. This is a very common type of problem, whether it be the prediction of a simple binary measure (e.g., Yes or No) or a multi-categorical value (e.g., customer type 1, 2 or 3). In either instance the researcher may be looking for explanation of the classification process, so interpretation of the variate also can be critical.

## CHAPTERS IN SECTION IV

Chapter 7, Multiple Discriminant Analysis, and Chapter 8, Logistic Regression: Regression with a Binary Dependent Variable, investigate this unique form of dependence relationship—a dependent variable that is not metric but rather is nonmetric. This classification into groups can be accomplished through either discriminant analysis or logistic regression, a variant of regression designed to specifically deal with nonmetric dependent variables. Discriminant analysis is particularly well suited for multi-categorical outcome variables, while logistic regression has many similarities with multiple regression in terms of variable interpretation and casewise diagnostics.

7

# Multiple Discriminant Analysis

# Chapter Preview

Multiple regression is undoubtedly the most widely used multivariate dependence technique. The primary basis for the popularity of regression has been its ability to predict and explain metric variables. But what happens when nonmetric dependent variables make multiple regression unsuitable? This chapter introduces a technique—discriminant analysis—that addresses the situation of a nonmetric dependent variable. In this type of situation, the researcher is interested in the prediction and explanation of the relationships that affect the category of the dependent variable in which an object is located, such as why a person is or is not a customer, or if a firm will succeed or fail. The two major objectives of this chapter are the following:

- 1 To introduce the underlying nature, philosophy, and conditions of multiple discriminant analysis.
  - 2 To demonstrate the application and interpretation of these techniques with an illustrative example.

Chapter 1 stated that the basic purpose of discriminant analysis is to estimate the relationship between a single nonmetric (categorical) dependent variable and a set of metric independent variables in this general form:

Multiple discriminant analysis has widespread application in situations in which the primary objective is to identify the group to which an object (e.g., person, firm, or product) belongs. Potential applications include predicting the success or failure of a new product, deciding whether a student should be admitted to graduate school, classifying students as to vocational interests, determining the category of credit risk for a person, or predicting whether a firm will be successful. In each instance, the objects fall into groups, and the objective is to predict and explain the bases for each object's group membership through a set of independent variables selected by the researcher.

A second technique—logistic regression—is also appropriate for handling research questions where the dependent variable is nonmetric. Logistic regression has most often been applied to situations with binary (two group) dependent variables (Yes/No, Purchase/Non-purchase, etc.), but has recently been adapted for applications with more than two possible dependent variable groups. The reader is encouraged to review the chapter on logistic regression (Chapter 8), because it presents many useful features in terms of interpretation of the impacts of the independent variables.

## Key Terms

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

**Analysis sample** Group of cases used in estimating the *discriminant function(s)*. When constructing *classification matrices*, the original sample is divided randomly into two groups, one for model estimation (the *analysis sample*) and the other for validation (the *holdout sample*).

**Box's M** Statistical test for the equality of the covariance matrices of the independent variables across the groups of the dependent variable. If the statistical significance does not exceed the critical level (i.e., nonsignificance), then the equality of the covariance matrices is supported. If the test shows statistical significance, then the groups are deemed different and the assumption is violated.

**Categorical variable** See *nonmetric variable*.

**Centroid** Mean value for the *discriminant Z scores* of all objects within a particular category or group. For example, a two-group discriminant analysis has two centroids, one for the objects in each of the two groups.

**Classification function** Method of classification in which a linear function is defined for each group. Classification is performed by calculating a score for each observation on each group's classification function and then assigning the observation to the group with the highest score. It differs from the calculation of the *discriminant Z score*, which is calculated for each *discriminant function*.

**Classification matrix** Means of assessing the predictive ability of the *discriminant function(s)* (also called a confusion, assignment, or prediction matrix). Created by cross-tabulating actual group membership with predicted group membership, this matrix consists of numbers on the diagonal representing correct classifications and off-diagonal numbers representing incorrect classifications.

**Cross-validation** Procedure of dividing the sample into two parts: the *analysis sample* used in estimation of the discriminant function(s) and the *holdout sample* used to validate the results. Cross-validation avoids the overfitting of the discriminant function by allowing its validation on a totally separate sample.

**Cutting score** Criterion against which each individual's *discriminant Z score* is compared to determine predicted group membership. When the analysis involves two groups, group prediction is determined by computing a single cutting score. Entities with discriminant Z scores below this score are assigned to one group, whereas those with scores above it are classified in the other group. For three or more groups, multiple discriminant functions are used, with a different cutting score for each function.

**Discriminant coefficient** See *discriminant weight*.

**Discriminant function** A variate of the independent variables selected for their discriminatory power used in the prediction of group membership. The predicted value of the discriminant function is the *discriminant Z score*, which is calculated for each object (person, firm, or product) in the analysis. It takes the form of the linear equation:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk}$$

where:

$Z_{jk}$  = discriminant Z score of discriminant function  $j$  for object  $k$

$a$  = intercept

$W_i$  = discriminant weight for independent variable  $i$

$X_{ik}$  = independent variable  $i$  for object  $k$

**Discriminant loadings** Measurement of the simple linear correlation between each independent variable and the *discriminant Z score* for each discriminant function; also called *structure correlations*. Discriminant loadings are calculated whether or not an independent variable is included in the discriminant function(s).

**Discriminant score** See *discriminant Z score*.

**Discriminant weight** Weight whose size relates to the discriminatory power of that independent variable across the groups of the dependent variable. Independent variables with large discriminatory power usually have large weights, and those with little discriminatory power usually have small weights. However, multicollinearity among the independent variables will cause exceptions to this rule. Also called the *discriminant coefficient*.

**Discriminant Z score** Score defined by the *discriminant function* for each object in the analysis and usually stated in standardized terms. Also referred to as the *Z score*, it is calculated for each object on each discriminant function and used in conjunction with the *cutting score* to determine predicted group membership. It is different from the *z score* terminology used for standardized variables.

**Extreme groups approach** Method of constructing a categorical dependent variable from a metric variable. First, the metric variable is divided into three categories. Then the extreme categories are used in the discriminant analysis and the middle category is not included in the analysis.

**Fisher's linear discriminant function** See *classification function*.

**Hit ratio** Percentage of objects (individuals, respondents, firms, etc.) correctly classified by the *discriminant function*. It is calculated as the number of objects in the diagonal of the *classification matrix* divided by the total number of objects. Also known as the *percentage correctly classified*.

**Holdout sample** Group of objects not used to compute the *discriminant function(s)*. This group is then used to validate the discriminant function with a separate sample of respondents. Also called the *validation sample*.

**Logistic regression** Special form of regression in which the dependent variable is a nonmetric, dichotomous (binary) variable. Although some differences exist, the general manner of interpretation is similar to linear regression.

**Maximum chance criterion** Measure of predictive accuracy in the *classification matrix* that is calculated as the percentage of respondents in the largest group. The rationale is that the best uninformed choice is to classify every observation into the largest group.

**Metric variable** Variable with a constant unit of measurement. If a metric variable is scaled from 1 to 9, the difference between 1 and 2 is the same as that between 8 and 9. A more complete discussion of its characteristics and differences from a *nonmetric* or *categorical variable* is found in Chapter 1.

**Nonmetric variable** Variable with values that serve merely as a label or means of identification, also referred to as a *categorical*, nominal, binary, qualitative, or taxonomic variable. The number on a football jersey is an example. A more complete discussion of its characteristics and its differences from a *metric variable* is found in Chapter 1.

**Optimal cutting score** *Discriminant Z score* value that best separates the groups on each discriminant function for classification purposes.

**Percentage correctly classified** See *hit ratio*.

**Polar extremes approach** See *extreme groups approach*.

**Potency index** Composite measure of the discriminatory power of an independent variable when more than one *discriminant function* is estimated. Based on *discriminant loadings*, it is a relative measure used for comparing the overall discrimination provided by each independent variable across all significant discriminant functions.

**Press's Q statistic** Measure of the classificatory power of the *discriminant function* when compared with the results expected from a chance model. The calculated value is compared to a critical value based on the chi-square distribution. If the calculated value exceeds the critical value, the classification results are significantly better than would be expected by chance.

**Proportional chance criterion** Another criterion for assessing the *hit ratio*, in which the average probability of classification is calculated considering all group sizes.

**Simultaneous estimation** Estimation of the *discriminant function(s)* where weights for all independent variables are calculated simultaneously; contrasts with *stepwise estimation* in which independent variables are entered sequentially according to discriminating power.

**Split-sample validation** See *cross-validation*.

**Stepwise estimation** Process of estimating the *discriminant function(s)* whereby independent variables are entered sequentially according to the discriminatory power they add to the group membership prediction.

**Stretched vector** Scaled vector in which the original vector is scaled to represent the corresponding *F* ratio. Used to graphically represent the *discriminant loadings* in a combined manner with the group *centroids*.

**Structure correlations** See *discriminant loadings*.

**Territorial map** Graphical portrayal of the *cutting scores* on a two-dimensional graph. When combined with the plots of individual cases, the dispersion of each group can be viewed and the misclassifications of individual cases identified directly from the map.

**Tolerance** Proportion of the variation in the independent variables not explained by the variables already in the model (function). It can be used to protect against multicollinearity. Calculated as  $1 - R_i^{2*}$ , where  $R_i^{2*}$  is the amount of variance of independent variable  $i$  explained by all of the other independent variables. A tolerance of 0 means that the independent variable under consideration is a perfect linear combination of independent variables already in the model. A tolerance of 1 means that an independent variable is totally independent of other variables already in the model.

**Validation sample** See *holdout sample*.

**Variate** Linear combination that represents the weighted sum of two or more independent variables that comprise the *discriminant function*. Also called linear combination or linear compound.

**Vector** Representation of the direction and magnitude of a variable's role as portrayed in a graphical interpretation of discriminant analysis results.

**Z score** See *discriminant Z score*.

## What Is Discriminant Analysis?

In attempting to choose an appropriate analytical technique, we sometimes encounter a problem that involves a categorical dependent variable and several metric independent variables. For example, we may wish to distinguish good from bad credit risks. If we had a metric measure of credit risk, then we could use multiple regression. In many instances we do not have the metric measure necessary for multiple regression. Instead, we are only able to determine whether someone is in a particular group (e.g., good or bad credit risk).

Discriminant analysis is the appropriate statistical techniques when the dependent variable is a **categorical** (nominal or **nonmetric**) **variable** and the independent variables are **metric variables**. In many cases, the dependent variable consists of two groups or classifications, for example, male versus female or high versus low. In other instances, more than two groups are involved, such as low, medium, and high classifications. Discriminant analysis is capable of handling either two groups or multiple (three or more) groups. When two classifications are involved, the technique is referred to as *two-group discriminant analysis*. When three or more classifications are identified, the technique is referred to as *multiple discriminant analysis (MDA)*. **Logistic regression** has most often been applied in its basic form to two groups, although other formulations can handle more groups. Logistic regression is covered in more detail in Chapter 8.

### THE VARIATE

Discriminant analysis involves deriving a **variate**. The discriminant variate is the linear combination of the two (or more) independent variables that will discriminate best between the objects (persons, firms, etc.) in the groups defined a priori. Discrimination is achieved by calculating the variate's weights for each independent variable to maximize the differences in the discriminant scores between the groups (i.e., the between-group variance relative to the within-group variance). The variate for a discriminant analysis, also known as the **discriminant function**, is derived from an equation much like that seen in multiple regression. It takes the following form:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \cdots + W_n X_{nk}$$

where:

$Z_{jk}$  = discriminant *Z* score of discriminant function  $j$  for object  $k$

$a$  = intercept

$W_i$  = discriminant weight for independent variable  $i$

$X_{ik}$  = independent variable  $i$  for object  $k$

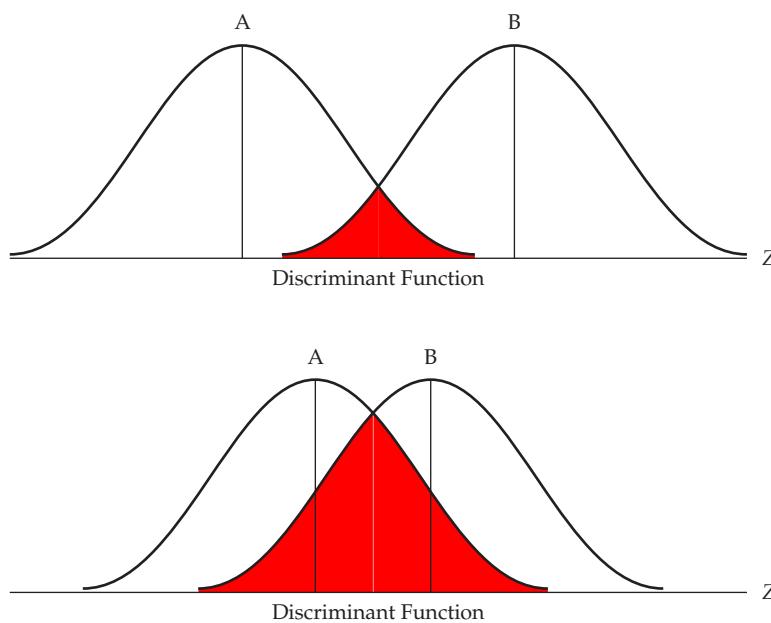
As with the variate in regression or any other multivariate technique, the **discriminant score**, also known as the **discriminant Z score** or **Z score**, is calculated for each object in the analysis (person, firm, etc.), and it is a summation of the values obtained by multiplying each independent variable by its discriminant weight. One of the unique features of discriminant analysis is that more than one discriminant function may be present in the data, resulting in each object possibly having more than one discriminant score. We will discuss what determines the number of discriminant functions later, but here we see that discriminant analysis has both similarities as well as unique elements when compared to other multivariate techniques.

## TESTING HYPOTHESES

Discriminant analysis is the appropriate statistical technique for testing the hypothesis that the group means of a set of independent variables for two or more groups are equal. By averaging the discriminant scores for all the individuals within a particular group, we arrive at the group mean. This group mean is referred to as a **centroid**. When the analysis involves two groups, there are two centroids; with three groups, there are three centroids; and so forth. The centroids indicate the most typical location of any member from a particular group, and a comparison of the group centroids shows how far apart the groups are in terms of a particular discriminant function.

The test for the statistical significance of the discriminant function is a generalized measure of the distance between the group centroids. It is computed by comparing the distributions of the discriminant scores for the groups. If the overlap in the distributions is small, the discriminant function separates the groups well. If the overlap is large, the function is a poor discriminator between the groups. Two distributions of discriminant scores shown in Figure 7.1 further illustrate this concept. The top diagram represents the distributions of discriminant scores for a function that separates the groups well, showing minimal overlap (the shaded area) between the groups. The lower diagram shows the distributions of discriminant scores on a discriminant function that is a relatively poor discriminator between groups A and B. The shaded areas of overlap represent the instances where misclassifying objects from group A into group B, and vice versa, can occur.

Multiple discriminant analysis is unique in one characteristic among the dependence relationships. If the dependent variable consists of more than two groups, discriminant analysis will calculate more than one discriminant



**Figure 7.1**  
Univariate Representation of  
Discriminant Z Scores

function. In fact, it will calculate  $NG - 1$  functions, where  $NG$  is the number of groups. Each discriminant function will calculate a separate discriminant Z score. In the case of a three-group dependent variable, each object (respondent, firm, etc.) will have a separate score for discriminant functions one and two, which enables the objects to be plotted in two dimensions, with each dimension representing a discriminant function. Thus, discriminant analysis is not limited to a single variate, as is multiple regression, but creates multiple variates representing dimensions of discrimination among the groups.

## Similarities to Other Multivariate Techniques

The application and interpretation of discriminant analysis is much the same as in regression analysis. That is, the discriminant function is a linear combination (variate) of metric measurements for two or more independent variables and is used to describe or predict a single dependent variable. The key difference is that discriminant analysis is appropriate for research problems in which the dependent variable is categorical (nominal or nonmetric), whereas regression is utilized when the dependent variable is metric. As discussed earlier, logistic regression is a variant of regression with many similarities except for the type of dependent variable.

As will be seen in later discussions, interpretation of the variate of independent variables, known as the discriminant function, is in some ways similar to interpretation of the factors/components in exploratory factor analysis (Chapter 3). Most comparable is the use of loadings as a means of understanding the relative impact of the independent variables on the discriminant function along with the possible rotation for improved interpretation when more than one discriminant function is estimated.

But the multivariate technique with the most similarities is multivariate analysis of variance (MANOVA) (see Chapter 6 for further discussion). Discriminant analysis can be seen as “reversing” a MANOVA analysis since both techniques are based on comparing metric versus nonmetric variables. In discriminant analysis, the single dependent variable is categorical with multiple metric independent variables. The opposite is true of MANOVA, which involves multiple metric dependent variables and one or more categorical/nonmetric independent variable(s). The model estimation process is similar, as evidenced by the same statistical measures of overall model fit in both techniques. Yet despite the similarities in types of variables analyzed, their primary objectives and accompanying analyses are distinct. In MANOVA, groups of cases are defined by a single categorical variable (single-factor design) or a combination of categorical variables (multi-factor design) in order to detect differences on the metric variables. The emphasis is on the impact of the categorical variable(s) in creating groups with differences on the metric variables. In discriminant analysis, the groups are specified in a single categorical variables and the emphasis is on identifying the set of metric variables providing significant differences between these groups. With the groups considered as given in discriminant analysis, it can also provide a classification function where cases are assigned to groups based on the metric variables.

## Hypothetical Example of Discriminant Analysis

Discriminant analysis is applicable to any research question with the objective of understanding group membership, whether the groups comprise individuals (e.g., customers versus non-customers), firms (e.g., profitable versus unprofitable), products (e.g., successful versus unsuccessful), or any other object that can be evaluated on a series of independent variables. To illustrate the basic premises of discriminant analysis, we examine two research settings, one involving two groups (purchasers versus non-purchasers) and the other three groups (levels of switching behavior).

### A TWO-GROUP DISCRIMINANT ANALYSIS: PURCHASERS VERSUS NON-PURCHASERS

Suppose Kitchenade wants to determine whether one of its new products—a new and improved food mixer—will be commercially successful. In carrying out the investigation, Kitchenade is primarily interested in identifying (if possible) those consumers who would purchase the new product versus those who would not. In statistical

terminology, Kitchenade would like to minimize the number of errors it would make in predicting which consumers would buy the new food mixer and which would not. To assist in identifying potential purchasers, Kitchenade devised rating scales on three characteristics—durability, performance, and style—to be used by consumers in evaluating the new product. Rather than relying on each scale as a separate measure, Kitchenade hopes that a weighted combination of all three characteristics would better predict purchase likelihood of consumers.

The primary objective of discriminant analysis is to develop a weighted combination of the three scales for predicting the likelihood that a consumer will purchase the product. In addition to determining whether consumers who are likely to purchase the new product can be distinguished from those who are not, Kitchenade would also like to know which characteristics of its new product are useful in differentiating likely purchasers from non-purchasers. That is, evaluations on which of the three characteristics of the new product best separate purchasers from non-purchasers?

For example, if the response “would purchase” is always associated with a high durability rating and the response “would not purchase” is always associated with a low durability rating, Kitchenade could conclude that the characteristic of durability distinguishes purchasers from non-purchasers. In contrast, if Kitchenade found that about as many persons with a high rating on style said they would purchase the food mixer as those who said they would not, then style is a characteristic that discriminates poorly between purchasers and non-purchasers.

**Identifying Discriminating Variables** To identify variables that may be useful in discriminating between groups (i.e., purchasers versus non-purchasers), emphasis is placed on group differences rather than measures of correlation used in multiple regression.

Table 7.1 lists the ratings of the new mixer on these three characteristics (at a specified price) by a panel of 10 potential purchasers. In rating the food mixer, each panel member is implicitly comparing it with products already on the market. After the product was evaluated, the evaluators were asked to state their buying intentions (“would purchase” or “would not purchase”). Five stated that they would purchase the new mixer and five said they would not.

**Table 7.1** Kitchenade Survey Results for the Evaluation of a New Consumer Product

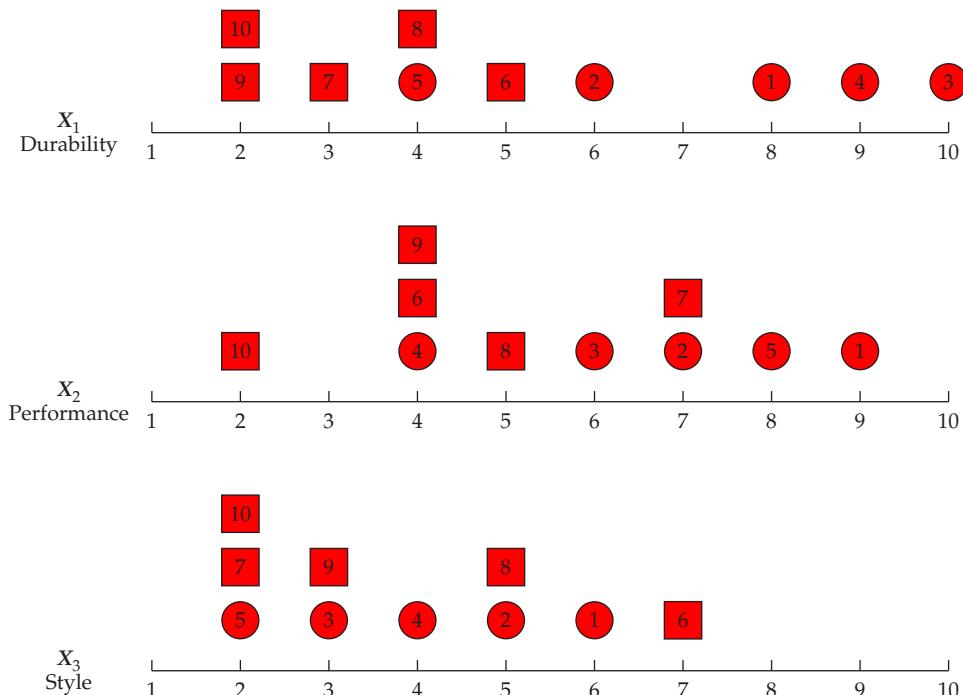
Groups Based on Purchase Intention	Evaluation of New Product*		
	X <sub>1</sub> Durability	X <sub>2</sub> Performance	X <sub>3</sub> Style
Group 1: Would purchase			
Subject 1	8	9	6
Subject 2	6	7	5
Subject 3	10	6	3
Subject 4	9	4	4
Subject 5	4	8	2
Group mean	7.4	6.8	4.0
Group 2: Would not purchase			
Subject 6	5	4	7
Subject 7	3	7	2
Subject 8	4	5	5
Subject 9	2	4	3
Subject 10	2	2	2
Group mean	3.2	4.4	3.8
Difference between group means	4.2	2.4	0.2

\*Evaluations are made on a 10-point scale (1 = very poor to 10 = excellent).

Examining Table 7.1 identifies several potential discriminating variables. First, a substantial difference separates the mean ratings of  $X_1$  (durability) for the “would purchase” and “would not purchase” groups (7.4 versus 3.2). As such, durability appears to discriminate well between the two groups and is likely to be an important characteristic to potential purchasers. In contrast, the characteristic of style ( $X_3$ ) has a much smaller difference of 0.2 between mean ratings ( $4.0 - 3.8 = 0.2$ ) for the “would purchase” and “would not purchase” groups. Therefore, we would expect this characteristic to be less discriminating in terms of a purchase decision. However, before we can make such statements conclusively, we must examine the distribution of scores for each group. Large standard deviations within one or both groups might make the difference between means nonsignificant and inconsequential in discriminating between the groups.

Because we have only 10 respondents in two groups and three independent variables, we can also look at the data graphically to determine what discriminant analysis is trying to accomplish. Figure 7.2 shows the 10 respondents on each of the three variables. The “would purchase” group is represented by circles and the “would not purchase” group by squares. Respondent identification numbers are inside the shapes:

- $X_1$  (Durability) had a substantial difference in mean scores, enabling us to almost perfectly discriminate between the groups using only this variable. If we established the value of 5.5 as our cutoff point to discriminate between the two groups, then we would misclassify only respondent 5, one of the “would purchase” group members. This variable illustrates the discriminatory power in having a large difference in the means for the two groups and a lack of overlap between the distributions of the two groups.
- $X_2$  (Performance) provides a less clear-cut distinction between the two groups. However, this variable does provide high discrimination for respondent 5, who was misclassified if we used only  $X_1$ . In addition, the respondents who would be misclassified using  $X_2$  are well separated on  $X_1$ . Thus,  $X_1$  and  $X_2$  might be used quite effectively in combination to predict group membership.
- $X_3$  (Style) shows little differentiation between the groups. Thus, by forming a variate of only  $X_1$  and  $X_2$ , and omitting  $X_3$ , a discriminant function may be formed that maximizes the separation of the groups on the discriminant score.



**Figure 7.2**  
Graphical Representation of 10 Potential Purchasers on Three Possible Discriminating Variables

**Calculating a Discriminant Function** With the three potential discriminating variables identified, attention shifts toward investigation of the possibility of using the discriminating variables in combination to improve upon the discriminating power of any individual variable. To this end, a variate can be formed with two or more discriminating variables to act together in discriminating between the groups.

Table 7.2 contains the results for three different formulations of a discriminant function, each representing different combinations of the three independent variables.

- The first discriminant function contains just  $X_1$ , equating the value of  $X_1$  to the discriminant  $Z$  score (also implying a weight of 1.0 for  $X_1$  and weights of zero for all other variables). As shown earlier, use of only  $X_1$ , the best discriminator, results in the misclassification of subject 5 as shown in Table 7.2, where four out of five subjects in group 1 (all but subject 5) and five of five subjects in group 2 are correctly classified (i.e., lie on the diagonal of the classification matrix). The percentage correctly classified is thus 90 percent (9 out of 10 subjects).
- Because  $X_2$  provides discrimination for subject 5, we can form a second discriminant function by equally combining  $X_1$  and  $X_2$  (i.e., implying weights of 1.0 for  $X_1$  and  $X_2$  and a weight of 0.0 for  $X_3$ ) to utilize each variable's unique discriminatory powers. Using a cutting score of 11 with this new discriminant function (see Table 7.2) achieves a perfect classification of the two groups (100% correctly classified). Thus,  $X_1$  and  $X_2$  in combination are able to make better predictions of group membership than either variable separately.
- The third discriminant function in Table 7.2 represents the actual estimated discriminant function ( $Z = -4.53 + .476X_1 + .359X_2$ ). Based on a cutting score of 0, this third function also achieves a 100 percent correct classification rate with the maximum separation possible between groups.

**Table 7.2** Creating Discriminant Functions to Predict Purchasers Versus Non-purchasers

Group	Calculated Discriminant Z Scores		
	Function 1: $Z = X_1$	Function 2: $Z = X_1 + X_2$	Function 3: $Z = -4.53 + .476X_1 + .359X_2$
<b>Group 1: Would purchase</b>			
Subject 1	8	17	2.51
Subject 2	6	13	.84
Subject 3	10	16	2.38
Subject 4	9	13	1.19
Subject 5	4	12	.25
<b>Group 2: Would not purchase</b>			
Subject 6	5	9	-.71
Subject 7	3	10	-.59
Subject 8	4	9	-.83
Subject 9	2	6	-2.14
Subject 10	2	4	-2.86
Cutting score	5.5	11	0.0

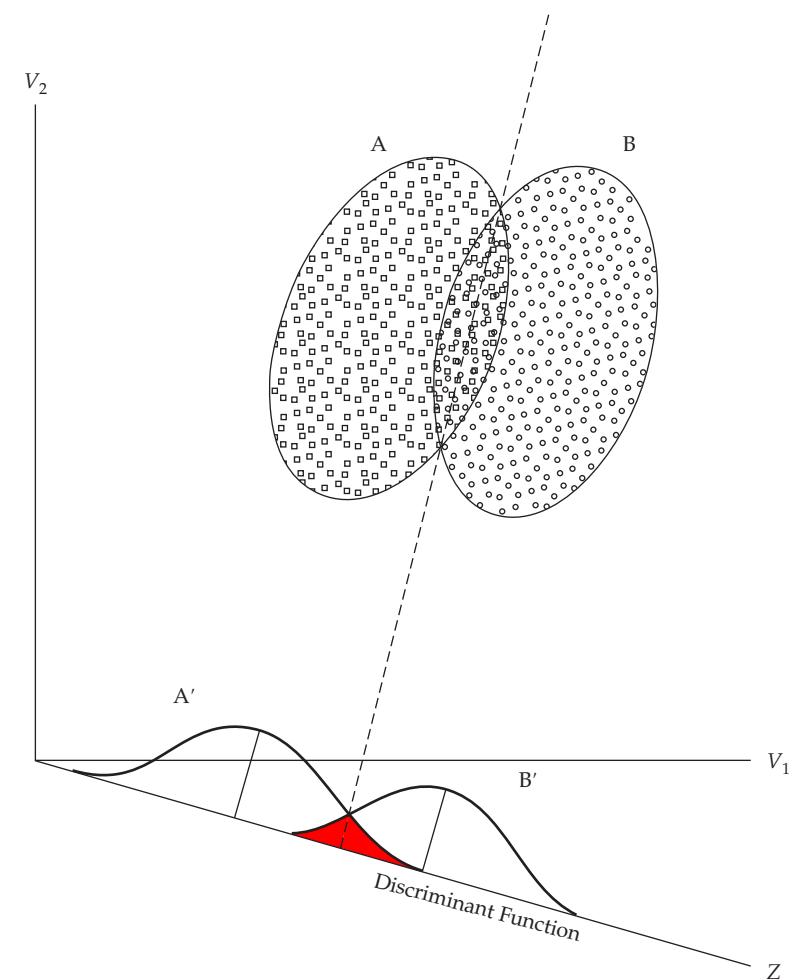
#### Classification Accuracy:

	Function 1: Predicted Group		Function 2: Predicted Group		Function 3: Predicted Group	
	Actual Group	Predicted Group	Actual Group	Predicted Group	Actual Group	Predicted Group
Actual Group	1	2	1	2	1	2
1: Would purchase	4	1	5	0	5	0
2: Would not purchase	0	5	0	5	0	5

As seen in this simple example, discriminant analysis identifies the variables with the greatest differences between the groups and derives a discriminant coefficient that weights each variable to reflect these differences. The result is a discriminant function that best discriminates between the groups based on a combination of the several independent variables.

**A Geometric Representation of the Two-Group Discriminant Function** A graphical illustration of another two-group analysis will help to further explain the nature of discriminant analysis [6]. Figure 7.3 demonstrates what happens when a two-group discriminant function is computed. Assume we have two groups, A and B, and two measurements,  $V_1$  and  $V_2$ , on each member of the two groups. We can plot in a scatter diagram of the association of variable  $V_1$  with variable  $V_2$  for each member of the two groups. In Figure 7.3 the small dots represent the variable measurements for the members of group B and the large dots those for group A. The ellipses drawn around the large and small dots would enclose some prespecified proportion of the points (discriminant score distributions), usually 95 percent or more in each group. If we draw a straight line through the two points at which the ellipses intersect and then project the line to a new Z axis, we can say that the overlap between the univariate distributions A' and B' (represented by the shaded area) is smaller than would be obtained by any other line drawn through the ellipses formed by the discriminant score scatterplots [6].

The important thing to note about Figure 7.3 is that the Z axis expresses the two-variable profiles of groups A and B as single numbers (discriminant scores). By finding a linear combination of the original variables  $V_1$  and  $V_2$ ,



**Figure 7.3**  
Graphical Illustration of Two-Group Discriminant

we can project the results as a discriminant function. For example, if the large and small dots are projected onto the new  $Z$  axis as discriminant  $Z$  scores, the result condenses the information about group differences (shown in the  $V_1V_2$  plot) into a set of points ( $Z$  scores) on a single axis, shown by distributions A' and B'.

To summarize, for a given discriminant analysis problem, a linear combination of the independent variables is derived, resulting in a series of discriminant scores for each object in each group. The discriminant scores are computed according to the statistical rule of maximizing the variance between the groups and minimizing the variance within them. If the variance between the groups is large relative to the variance within the groups, we say that the discriminant function separates the groups well.

## A THREE-GROUP EXAMPLE OF DISCRIMINANT ANALYSIS: SWITCHING INTENTIONS

The two-group example just examined demonstrates the rationale and benefit of combining independent variables into a variate for purposes of discriminating between groups. Discriminant analysis also has another means of discrimination—the estimation and use of multiple variates—in instances of three or more groups. These discriminant functions now become dimensions of discrimination, each dimension separate and distinct from the other. Thus, in addition to improving the explanation of group membership, these additional discriminant functions add insight into the various combinations of independent variables that discriminate between groups.

As an illustration of a three-group application of discriminant analysis, we examine research conducted by HBAT concerning the possibility of a competitor's customers switching suppliers. A small-scale pretest involved interviews of 15 customers of a major competitor. In the course of the interviews, the customers were asked their probability of switching suppliers on a three-category scale. The three possible responses were "definitely switch," "undecided," and "definitely not switch." Customers were assigned to groups 1, 2, or 3, respectively, according to their responses. The customers also rated the competitor on two characteristics: price competitiveness ( $X_1$ ) and service level ( $X_2$ ). The research issue is now to determine whether the customers' ratings of the competitor can predict their probability of switching suppliers. Because the dependent variable of switching suppliers was measured as a categorical (nonmetric) variable and the ratings of price and service are metric, discriminant analysis is appropriate.

**Identifying Discriminating Variables** With three categories of the dependent variable, discriminant analysis can estimate two discriminant functions, each representing a different dimension of discrimination.

Table 7.3 contains the survey results for the 15 customers, five in each category of the dependent variable. As we did in the two-group example, we can look at the mean scores for each group to see whether one of the variables discriminates well among all the groups. For  $X_1$ , price competitiveness, we see a rather large mean difference between group 1 and group 2 or 3 (2.0 versus 4.6 or 3.8).  $X_1$  may discriminate well between group 1 and group 2 or 3, but is much less effective in discriminating between groups 2 and 3. For  $X_2$ , service level, we see that the difference between groups 1 and 2 is very small (2.0 versus 2.2), whereas a large difference exists between group 3 and group 1 or 2 (6.2 versus 2.0 or 2.2). Thus,  $X_1$  distinguishes group 1 from groups 2 and 3, and  $X_2$  distinguishes group 3 from groups 1 and 2. As a result, we see that  $X_1$  and  $X_2$  provide different *dimensions* of discrimination between the groups.

**Calculating Two Discriminant Functions** With the potential discriminating variables identified, the next step is to combine them into discriminant functions that will utilize their combined discriminating power for distinguishing between groups.

To illustrate these dimensions graphically, Figure 7.4 portrays the three groups on each of the independent variables separately. Viewing the group members on any one variable, we can see that no variable discriminates well among all the groups. However, if we construct two simple discriminant functions, using just simple weights of 0.0 or 1.0, the results become much clearer. Discriminant function 1 gives  $X_1$  a weight of 1.0, and  $X_2$  a weight of 0.0.

**Table 7.3** HBAT Survey Results of Switching Intentions by Potential Customers

Groups Based on Switching Intention	Evaluation of Current Supplier *	
	X <sub>1</sub>	X <sub>2</sub>
	Price Competitiveness	Service Level
Group 1: Definitely switch		
Subject 1	2	2
Subject 2	1	2
Subject 3	3	2
Subject 4	2	1
Subject 5	2	3
Group mean	2.0	2.0
Group 2: Undecided		
Subject 6	4	2
Subject 7	4	3
Subject 8	5	1
Subject 9	5	2
Subject 10	5	3
Group mean	4.6	2.2
Group 3: Definitely not switch		
Subject 11		6
Subject 12	3	6
Subject 13	4	6
Subject 14	5	6
Subject 15	5	7
Group mean	3.8	6.2

\*Evaluations are made on a 10-point scale (1 = very poor to 10 = excellent).

Likewise, discriminant function 2 gives  $X_2$  a weight of 1.0, and  $X_1$  a weight of 0.0. The functions can be stated mathematically as:

$$\text{Discriminant function 1} = 1.0(X_1) + 0.0(X_2)$$

$$\text{Discriminant function 2} = 0.0(X_1) + 1.0(X_2).$$

These equations show in simple terms how the discriminant analysis procedure estimates weights to maximize discrimination.

With the two functions, we can now calculate two discriminant scores for each respondent. Moreover, the two discriminant functions provide the dimensions of discrimination.

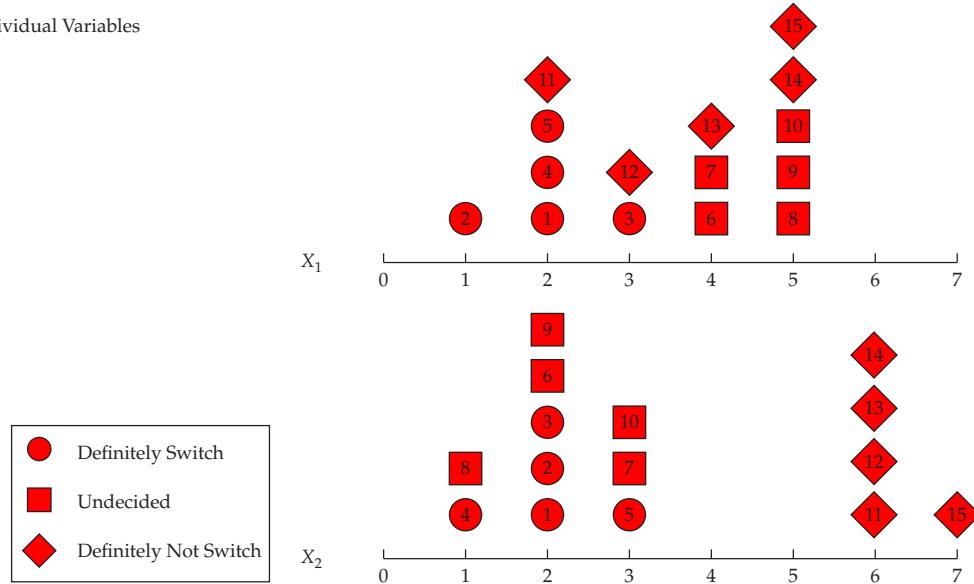
Figure 7.4 also contains a plot of each respondent in a two-dimensional representation. The separation between groups now becomes quite apparent, and each group can be easily distinguished. We can establish values on each dimension that will define regions containing each group (e.g., all members of group 1 are in the region less than 3.5 on dimension 1 and less than 4.5 on dimension 2). Each of the other groups can be similarly defined in terms of the ranges of their discriminant function scores.

In terms of dimensions of discrimination, the first discriminant function, price competitiveness, distinguishes between undecided customers (shown with a square) and those customers who have decided to switch (circles). But price competitiveness does not distinguish those who have decided not to switch (diamonds). Instead, the perception of service level, defining the second discriminant function, predicts whether a customer will decide not to switch versus whether a customer is undecided or determined to switch suppliers. The researcher can present to management the separate impacts of both price competitiveness and service level in making this decision.

**Figure 7.4**

Graphical Representation of Potential Discriminating Variables for a Three-Group Discriminant Analysis

(a) Individual Variables

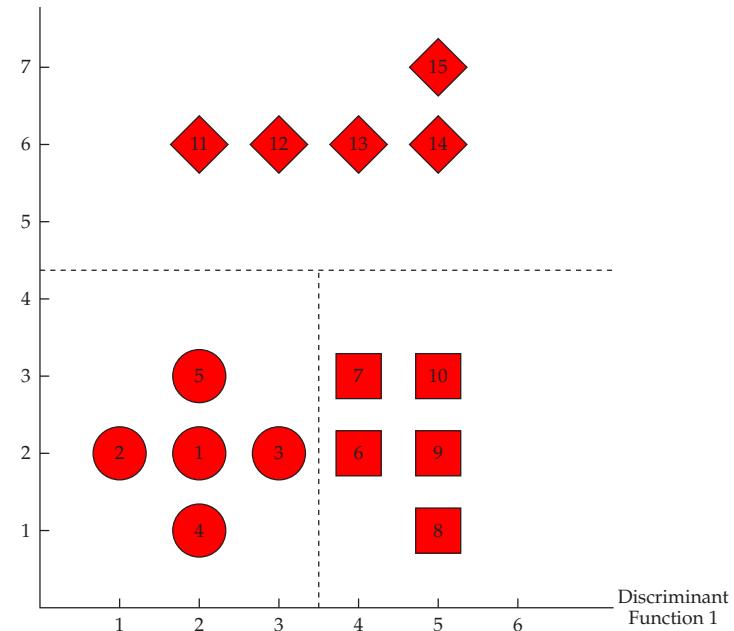


(b) Two-Dimensional Representation of Discriminant Functions

$$\text{Discriminant Function 1} = 1.0X_1 + 0X_2$$

$$\text{Discriminant Function 2} = 0X_1 + 1.0X_2$$

Discriminant Function 2



The estimation of more than one discriminant function, when possible, provides the researcher with both improved discrimination and additional perspectives on the features and the combinations that best discriminate among the groups. The following sections detail the necessary steps for performing a discriminant analysis, assessing its level of predictive fit, and then interpreting the influence of independent variables in making that prediction.

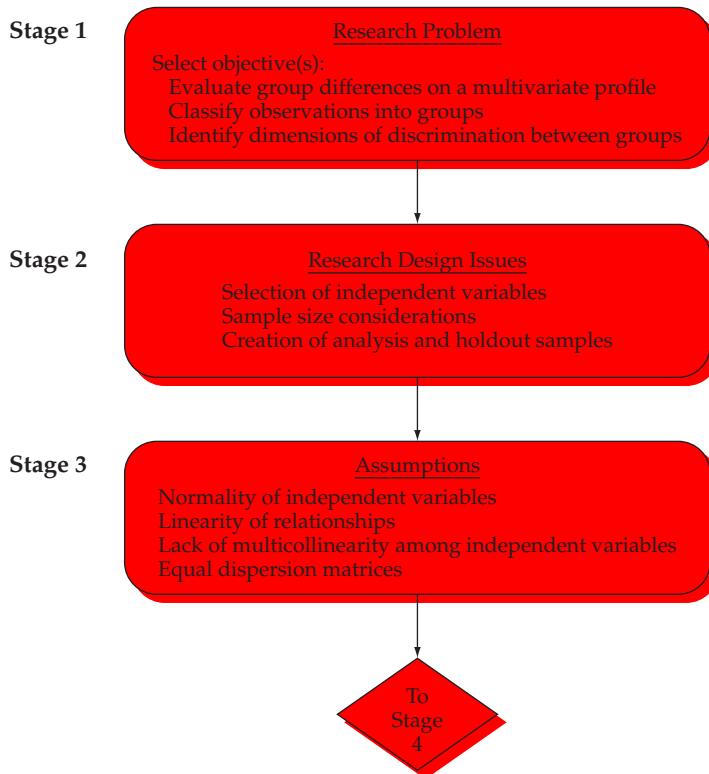
## The Decision Process for Discriminant Analysis

The application of discriminant analysis can be viewed from the six-stage model-building perspective introduced in Chapter 1 and portrayed in Figure 7.5 (stages 1–3) and Figure 7.6 (stages 4–6). As with all multivariate applications, setting the objectives is the first step in the analysis. Then the researcher must address specific design issues and make sure the underlying assumptions are met. The analysis proceeds with the derivation of the discriminant function and the determination of whether a statistically significant function can be derived to separate the two (or more) groups. The discriminant results are then assessed for predictive accuracy by developing a classification matrix. Next, interpretation of the discriminant function determines which of the independent variables contributes the most to discriminating between the groups. Finally, the discriminant function should be validated with a holdout sample. Each of these stages is discussed in the following sections.

### Stage 1: Objectives of Discriminant Analysis

A review of the objectives for applying discriminant analysis should further clarify its nature. Discriminant analysis can address any of the following research objectives:

- 1 Determining whether statistically significant differences exist between the average score profiles on a set of variables for two (or more) a priori defined groups
- 2 Determining which of the independent variables most account for the differences in the average score profiles of the two or more groups
- 3 Establishing the number and composition of the dimensions of discrimination between groups formed from the set of independent variables
- 4 Establishing procedures for classifying objects (individuals, firms, products, etc.) into groups on the basis of their scores on a set of independent variables



As noted in these objectives, discriminant analysis is useful when the researcher is interested either in understanding group differences or in correctly classifying objects into groups or classes. Discriminant analysis, therefore, can be considered either a type of profile analysis or an analytical predictive technique. In either case, the technique is most appropriate in situations with a single categorical dependent variable and several metrically scaled independent variables.

## DESCRIPTIVE PROFILE ANALYSIS

One primary function of discriminant analysis is to provide an objective assessment of differences between groups on a set of independent variables. In this situation, discriminant analysis is quite similar to multivariate analysis of variance (see Chapter 6 for a more detailed discussion of multivariate analysis of variance). For understanding group differences, discriminant analysis lends insight into the role of individual variables as well as defining combinations of these variables that represent dimensions of discrimination between groups. These dimensions are the collective effects of several variables that work jointly to distinguish between the groups. The use of sequential estimation methods also enables the analyst to identify subsets of variables with the greatest discriminatory power.

## CLASSIFICATION PURPOSES

Discriminant analysis also provides a basis for classifying not only the sample used to estimate the discriminant function but also any other observations that can have values for all the independent variables. In this way, the discriminant analysis can be used to classify other observations into the defined groups.

# Stage 2: Research Design for Discriminant Analysis

Successful application of discriminant analysis requires consideration of several issues. These issues include the selection of both dependent and independent variables, the sample size needed for estimation of the discriminant functions, and the division of the sample for validation purposes.

## SELECTING DEPENDENT AND INDEPENDENT VARIABLES

To apply discriminant analysis, the researcher first must specify which variables are to be independent measures and which variable is to be the dependent measure. Recall that the dependent variable is nonmetric and the independent variables are metric.

**The Dependent Variable** The researcher should focus on the dependent variable first. The number of dependent variable groups (categories) can be two or more, but these groups must be mutually exclusive and exhaustive. In other words, each observation can be placed into only one group. In some cases, the dependent variable may involve two groups (dichotomous), such as good versus bad. In other cases, the dependent variable may involve several groups (multichotomous), such as the occupations of physician, attorney, or professor.

**HOW MANY CATEGORIES IN THE DEPENDENT VARIABLE?** Theoretically, discriminant analysis can handle an unlimited number of categories in the dependent variable. As a practical matter, however, the researcher should select a dependent variable and the number of categories based on several considerations:

**Distinctive and Unique** In addition to being mutually exclusive and exhaustive, the dependent variable categories should be distinct and unique on the set of independent variables chosen. Discriminant analysis assumes that each group *should* have a unique profile on the independent variables used and thus develops the discriminant functions to maximally separate the groups based on these variables. Discriminant analysis does not, however, have a means of accommodating or combining categories that are not distinct on the independent variables. If two or more groups have quite similar profiles, discriminant analysis will not be able to uniquely profile each group, resulting in poorer

explanation and classification of the groups as a whole. As such, the researcher must select the dependent variables and its categories to reflect likely differences in the independent variables. An example will help illustrate this issue.

Assume the researcher wishes to identify differences among occupational categories based on a number of demographic characteristics (e.g., income, education, household characteristics). If occupations are represented by a small number of categories (e.g., blue-collar, white-collar, clerical/staff, and professional/upper management), then we would expect unique differences between the groups and that discriminant analysis would be best able to develop discriminant functions that would explain the group differences and successfully classify individuals into their correct category.

If the number of occupational categories was expanded, however, discriminant analysis may have a harder time identifying differences. For example, assume the professional/upper management category was expanded to the categories of doctors, lawyers, upper management, college professors, and so on. Although this expansion provides a more refined occupational classification, it would be much harder to distinguish between each of these categories on the demographic variables. The results would be poorer performance by discriminant analysis in both explanation and classification.

**Smaller Number** The researcher should also strive, all other things equal, for a smaller rather than larger number of categories in the dependent measure. It may seem more logical to expand the number of categories in search of more unique groupings, but expanding the number of categories presents more complexities in the profiling and classification tasks of discriminant analysis. If discriminant analysis can estimate up to  $NG - 1$  (number of groups minus one) discriminant functions, then increasing the number of groups expands the number of possible discriminant functions, increasing the complexity in identifying the underlying dimensions of discrimination reflected by each discriminant function as well as representing the overall effect of each independent variable.

As these two issues suggest, the researcher must always balance the desire to expand the categories for increased uniqueness versus the increased effectiveness in a smaller number of categories. The researcher should try and select a dependent variable with categories that have the maximum differences among all groups while maintaining both conceptual support and managerial relevance.

**CONVERTING METRIC VARIABLES** The preceding examples of categorical variables were true dichotomies (or multichotomies). In some situations, however, discriminant analysis is appropriate even if the dependent variable is not a true nonmetric (categorical) variable. We may have a dependent variable that is an ordinal or interval measurement that we wish to use as a categorical dependent variable. In such cases, we would have to create a categorical variable, and three approaches are the most commonly used:

**Response Values** The most common approach is to establish categories using the metric scale. For example, if we had a variable that measured the average number of cola drinks consumed per day, and the individuals responded on a scale from zero to eight or more per day, we could create an artificial trichotomy (three groups) by simply designating those individuals who consumed none, one, or two cola drinks per day as light users, those who consumed three, four, or five per day as medium users, and those who consumed six, seven, eight, or more as heavy users. Such a procedure would create a three-group categorical variable in which the objective would be to discriminate among light, medium, and heavy users of colas. Any number of categorical groups can be developed. Most frequently, the approach would involve creating two, three, or four categories. A larger number of categories could be established if the need arose.

**Extreme Groups Approach** When three or more categories are created, the possibility arises of examining only the extreme groups in a two-group discriminant analysis. The **extreme groups approach**, also known as the **polar extremes** approach, involves comparing only the extreme two groups and excluding the middle group from the discriminant analysis [15]. For example, the researcher could examine the light and heavy users of cola drinks and exclude the medium users. This approach can be used any time the researcher wishes to examine only the extreme groups. However, the researcher may also want to try this approach when the results of a regression analysis are not as good as anticipated. Such a procedure may be helpful because it is possible that group differences may appear even though regression results are poor. That is, the extreme groups approach with discriminant analysis can reveal

differences that are not as prominent in a regression analysis of the full data set [6]. Such manipulation of the data naturally would necessitate caution in interpreting one's findings.

**Mean or Median Split** In the past researchers have often developed two groups for analysis using either a mean or median split. For example, if the research includes a variable (or multi-item construct) that measures organizational commitment, then the highly committed group would be defined as those respondents exhibiting the mean (or median) or higher on the variable, and below the mean would be designated the group exhibiting low commitment. The mean and/or median split approaches, however, are arbitrary, non-scientific methods of defining groups. Instead, researchers should apply cluster analysis (see Chapter 4) to identify the natural grouping of respondents in which individuals in each group are the most similar within a particular group, and the most different from all other groups. Using cluster also overcomes the limitation of assuming that the sizes of the groups are equal, when in fact they often are not. Using the organizational commitment example, it is very unlikely that the high-commitment and low-commitment groups would be equal in size, which would be the result using either a mean or median split approach to developing groups using a metrically measured variable.

**The Independent Variables** After a decision has been made on the dependent variable, the researcher must decide which independent variables to include in the analysis. Independent variables usually are selected in two ways. The first approach involves identifying variables either from previous research, by conducting qualitative research, or from the theoretical model that is the underlying basis of the research question. The second approach is intuition—utilizing the researcher's knowledge and intuitively selecting variables for which no previous research or theory exists but that logically might be related to predicting the groups for the dependent variable.

In both instances, the most appropriate independent variables are those that differ across at least two of the groups of the dependent variable. Remember that the purpose of any independent variable is to present a unique profile of at least one group as compared to others. Variables that do not differ across the groups are of little use in discriminant analysis.

## SAMPLE SIZE

Discriminant analysis, like the other multivariate techniques, is affected by the size of the sample being analyzed. As discussed in Chapter 1, very small samples have so much sampling error that identification of all but the largest differences is improbable. Moreover, very large sample sizes will make all differences statistically significant, even though these same differences may have little or no managerial relevance. In between these extremes, the researcher must consider the impact of sample sizes on discriminant analysis, both at the overall level and on a group-by-group basis.

**Overall Sample Size** The first consideration involves the overall sample size. Discriminant analysis is quite sensitive to the ratio of sample size to the number of predictor variables. As a result, many studies suggest a ratio of 20 observations for each predictor variable. Although this ratio may be difficult to maintain in practice, the researcher must note that the results become unstable as the sample size decreases relative to the number of independent variables. The minimum size recommended is five observations per independent variable. Note that this ratio applies to all variables considered in the analysis, even if all of the variables considered are not entered into the discriminant function (such as in stepwise estimation).

**Sample Size per Category** In addition to the overall sample size, the researcher also must consider the sample size of each category. At a minimum, the smallest group size of a category must exceed the number of independent variables. As a practical guideline, each category should have at least 20 observations. Even when all categories exceed 20 observations, however, the researcher must also consider the relative sizes of the categories. Wide variations in the group sizes will impact the estimation of the discriminant function and the classification of observations. In the classification stage, larger groups have a disproportionately higher chance of classification. If the group sizes do

vary markedly, the researcher may wish to randomly sample from the larger group(s), thereby reducing their size to a level comparable to the smaller group(s). Always remember, however, to maintain an adequate sample size both overall and for each group.

## DIVISION OF THE SAMPLE

One final note about the impact of sample size in discriminant analysis. As will be discussed later in stage 6, the preferred means of validating a discriminant analysis is to divide the sample into two subsamples, one used for estimation of the discriminant function and another for validation purposes. In terms of sample size considerations, it is essential that each subsample be of adequate size to support conclusions from the results. As such, all of the considerations discussed in the previous section apply not only to the total sample, but also to each of the two subsamples (especially the subsample used for estimation). No hard-and-fast rules have been established, but it seems logical that the researcher would want at least 100 in the total sample to justify dividing it into the two groups.

**Creating the Subsamples** A number of procedures have been suggested for dividing the sample into subsamples. The usual procedure is to divide the total sample of respondents randomly into two subsamples. One of these subsamples, the **analysis sample**, is used to develop the discriminant function. The second, the **holdout sample**, is used to test the discriminant function. This method of validating the function is referred to as the **split-sample validation** or **cross-validation** [1, 4, 8, 14].

No definite guidelines have been established for determining the relative sizes of the analysis and holdout (or validation) subsamples. The most popular approach is to divide the total sample so that one-half of the respondents are placed in the analysis sample and the other half are placed in the holdout sample. However, no hard-and-fast rule has been established, and some researchers prefer a 60–40 or even 75–25 split between the analysis and the holdout groups, depending on the overall sample size.

When selecting the analysis and holdout samples, one usually follows a proportionately stratified sampling procedure. Assume first that the researcher desired a 50–50 split. If the categorical groups for the discriminant analysis are equally represented in the total sample, then the estimation and holdout samples should be of approximately equal size. If the original groups are unequal, the sizes of the estimation and holdout samples should be proportionate to the total sample distribution. For instance, if a sample consists of 50 males and 50 females, the estimation and holdout samples would have 25 males and 25 females. If the sample contained 70 females and 30 males, then the estimation and holdout samples would consist of 35 females and 15 males each.

**What If the Overall Sample Is Too Small?** If the sample size is too small to justify a division into analysis and holdout groups, the researcher has two options. First, develop the function on the entire sample and then use the function to classify the same group used to develop the function. This procedure results in an upward bias in the predictive accuracy of the function, but is certainly better than not testing the function at all. Second, several techniques discussed in stage 6 can perform a type of holdout procedure in which the discriminant function is repeatedly estimated on the sample, each time “holding out” a different observation. In this approach, much smaller sample sizes can be used because the overall sample need not be divided into subsamples.

## Stage 3: Assumptions of Discriminant Analysis

As with all multivariate techniques, discriminant analysis is based on a number of assumptions. These assumptions relate to both the statistical processes involved in the estimation and classification procedures and issues affecting the interpretation of the results. The following section discusses each type of assumption and the impacts on the proper application of discriminant analysis.

## IMPACTS ON ESTIMATION AND CLASSIFICATION

The key assumptions for deriving the discriminant function are multivariate normality of the independent variables and unknown (but equal) dispersion and covariance structures (matrices) for the groups as defined by the dependent variable [7, 9]. Although the evidence is mixed regarding the sensitivity of discriminant analysis to violations of these assumptions, the researcher must always understand the impacts on the results that can be expected. Moreover, if the assumptions are violated and the potential remedies are not acceptable or do not address the severity of the problem, the researcher should consider alternative methods (e.g., logistic regression described in Chapter 8).

**Identifying Assumption Violations** As discussed in Chapter 2, achieving univariate normality of individual variables will many times suffice to achieve multivariate normality. A number of tests for normality discussed in Chapter 2 are available to the researcher, along with the appropriate remedies, those most often being transformations of the variables.

The issue of equal dispersion of the independent variables (i.e., equivalent covariance matrices) is similar to homoscedasticity between individual variables (also discussed in Chapter 2). The most common test is the **Box's M** test assessing the significance of differences in the matrices between the groups. Here the researcher is looking for a *nonsignificant* probability level that would indicate there were no differences between the group covariance matrices. Given the sensitivity of the Box's M test, however, to the size of the covariance matrices and the number of groups in the analysis, researchers should use very conservative levels of significant differences (e.g., .01 rather than .05) when assessing whether differences are present. As the research design increases in sample size or terms of groups or number of independent variables, even more conservative levels of significance may be considered acceptable. The reader is also referred to the discussion of this issue in Chapter 6 where the implications in MANOVA are discussed.

**Impact on Estimation** Data not meeting the multivariate normality assumption can cause problems in the estimation of the discriminant function when there are substantive differences in group sizes or sample sizes are small. Remedies may be possible through transformations of the data to reduce the disparities among the covariance matrices. However, in many instances these remedies are ineffectual. In these situations, the models should be thoroughly validated. If the dependent measure is binary, logistic regression should be used if at all possible.

**Impact on Classification** Unequal covariance matrices also negatively affect the classification process. If the sample sizes are small and the covariance matrices are unequal, then the statistical significance of the estimation process is adversely affected. The more likely case is that of unequal covariances among groups of adequate sample size, whereby observations are overclassified into the groups with larger covariance matrices. This effect can be minimized by increasing the sample size and also by using the group-specific covariance matrices for classification purposes, but this approach mandates cross-validation of the discriminant results. Finally, quadratic classification techniques are available in many of the statistical programs if large differences exist between the covariance matrices of the groups and the remedies do not minimize the effect [5, 10, 12].

## IMPACTS ON INTERPRETATION

Another characteristic of the data that affects the results is multicollinearity among the independent variables. Multicollinearity, measured in terms of **tolerance**, denotes that two or more independent variables are highly correlated, so that one variable can be highly explained or predicted by the other variable(s) and thus it adds little to the explanatory power of the entire set. This consideration becomes especially critical when stepwise procedures are employed. The researcher, in interpreting the discriminant function, must be aware of the level of multicollinearity and its impact on determining which variables enter the stepwise solution. For a more detailed discussion of multicollinearity and its impact on stepwise solutions, see Chapter 5. The procedures for detecting the presence of multicollinearity are also addressed in Chapter 5.

## Discriminant Analysis Design

The dependent variable must be nonmetric, representing groups of objects that are expected to differ on the independent variables.

Choose a dependent variable that:

Best represents group differences of interest

Defines groups that are substantially different

Minimizes the number of categories while still meeting the research objectives.

In converting metric variables to a nonmetric scale for use as the dependent variable, consider using extreme groups to maximize the group differences.

Independent variables must identify differences between at least two groups to be of any use in discriminant analysis.

The sample size must be large enough to:

Have at least one more observation per group than the number of independent variables, but striving for at least 20 cases per group

Maximize the number of observations per variable, with a minimum ratio of five observations per independent variable

Have a large enough sample to divide it into estimation and holdout samples, each meeting the above requirements.

Assess the equality of covariance matrices with the Box's M test, but apply a conservative significance level of .01 and become even more conservative as the analysis becomes more complex with a larger number of groups and/or independent variables.

Examine the independent variables for univariate normality, because that is the most direct remedy for ensuring both multivariate normality and equality of covariance matrices.

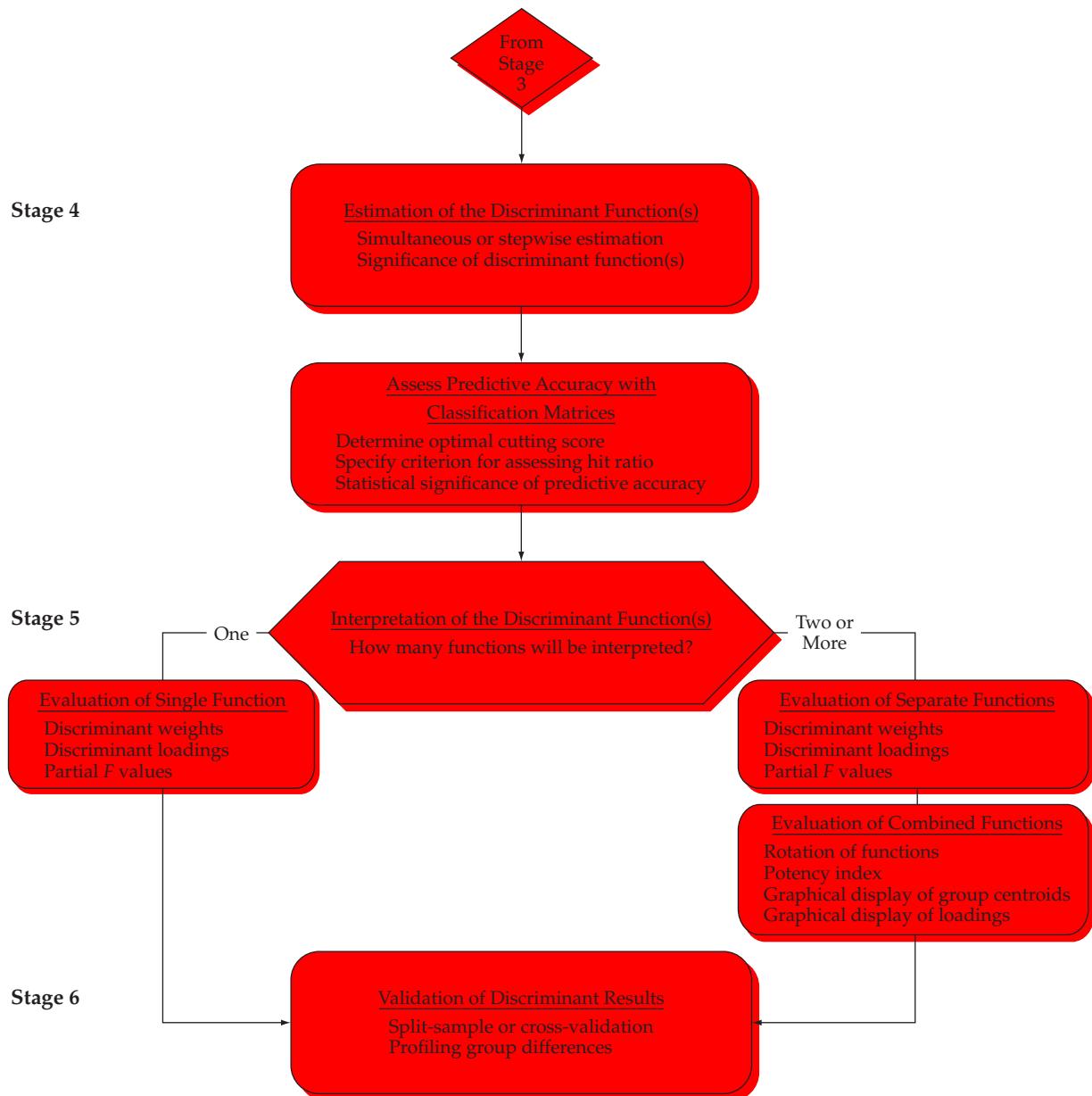
Multicollinearity among the independent variables can markedly reduce the estimated impact of independent variables in the derived discriminant function(s), particularly if a stepwise estimation process is used.

As with any of the multivariate techniques employing a variate, an implicit assumption is that all relationships are linear. Nonlinear relationships are not reflected in the discriminant function unless specific variable transformations are made to represent nonlinear effects. Finally, outliers can have a substantial impact on the classification accuracy of any discriminant analysis results. The researcher is encouraged to examine all results for the presence of outliers and to eliminate true outliers if needed. For a discussion of some of the techniques for assessing the violations in the basic statistical assumptions or outlier detection, see Chapter 2.

## Stage 4: Estimation of the Discriminant Model and Assessing Overall Fit

To derive the discriminant function, the researcher must decide on the method of estimation and then determine the number of functions to be retained (see Figure 7.6). With the functions estimated, overall model fit can be assessed in several ways. First, discriminant Z scores, also known as discriminant scores or Z scores, can be calculated for each object. Comparison of the group means (centroids) on the discriminant Z scores provides one measure of discrimination between groups. Predictive accuracy can be measured as the number of observations classified into the correct groups, with a number of criteria available to assess whether the classification process achieves practical or statistical significance. Finally, casewise diagnostics can identify the classification accuracy of each case and its relative impact on the overall model estimation.

**Figure 7.6**  
Stages 4–6 in the Discriminant Analysis Decision Diagram



## SELECTING AN ESTIMATION METHOD

The first task in deriving the discriminant function(s) is to choose the estimation method. In making this choice, the researcher must balance the need for control over the estimation process versus a desire for parsimony in the discriminant functions. The two methods available are the simultaneous (direct) method and the stepwise method, each discussed next.

**Simultaneous Estimation** Simultaneous estimation involves computing the discriminant function so that all of the independent variables are considered concurrently. Thus, the discriminant function is computed based upon the entire set of independent variables, regardless of the discriminating power of each independent variable. The simultaneous

method is appropriate when, for theoretical reasons, the researcher wants to include all the independent variables in the analysis and is not interested in seeing intermediate results based only on the most discriminating variables.

**Stepwise Estimation** Stepwise estimation is an alternative to the simultaneous approach. It involves entering the independent variables into the discriminant function one at a time on the basis of their discriminating power. The stepwise approach follows a sequential process of adding or deleting variables in the following manner:

- 1 Choose the single best discriminating variable.
- 2 Pair the initial variable with each of the other independent variables, one at a time, and select the variable that is best able to improve the discriminating power of the function in combination with the first variable.
- 3 Select additional variables in a like manner. Note that as additional variables are included, some previously selected variables may be removed if the information they contain about group differences is available in some combination of the other variables included at later stages.
- 4 Consider the process completed when either all independent variables are included in the function or the excluded variables are judged as not contributing significantly to further discrimination.

The stepwise method is useful when the researcher wants to consider a relatively large number of independent variables for inclusion in the function. By sequentially selecting the next best discriminating variable at each step, variables that are not useful in discriminating between the groups are eliminated and a reduced set of variables is identified. The reduced set typically is almost as good as—and sometimes better than—the complete set of variables.

The researcher should note that stepwise estimation becomes less stable and generalizable as the ratio of sample size to independent variable declines below the recommended level of 20 observations per independent variable. It is particularly important in these instances to validate the results in as many ways as possible.

## STATISTICAL SIGNIFICANCE

After estimation of the discriminant function(s), the researcher must assess the level of significance for the collective discriminatory power of the discriminant function(s) as well as the significance of each separate discriminant function. Evaluating the overall significance provides the researcher with the information necessary to decide whether to continue on to the interpretation of the analysis or if respecification is necessary. If the overall model is significant, then evaluating the individual functions identifies the function(s) that should be retained and interpreted.

**Overall Significance** In assessing the statistical significance of the overall model, different statistical criteria are applicable for simultaneous versus stepwise estimation procedures. In both situations, the statistical tests relate to the ability of the discriminant function(s) to derive discriminant Z scores that are significantly different between the groups.

**SIMULTANEOUS ESTIMATION** When a simultaneous approach is used, the measures of Wilks' lambda, Hotelling's trace, and Pillai's criterion all evaluate the statistical significance of the discriminatory power of the discriminant function(s). Roy's greatest characteristic root evaluates only the first discriminant function. For a more detailed discussion of the advantages and disadvantages of each criterion, see the discussion of significance testing in multivariate analysis of variance in Chapter 6.

**STEPWISE ESTIMATION** If a stepwise method is used to estimate the discriminant function, the Mahalanobis  $D^2$  and Rao's V measures are most appropriate. Both are measures of generalized distance. The Mahalanobis  $D^2$  procedure is based on generalized squared Euclidean distance that adjusts for unequal variances. The major advantage of this procedure is that it is computed in the original space of the predictor variables rather than as a collapsed version used in other measures. The Mahalanobis  $D^2$  procedure becomes particularly critical as the number of predictor variables increases, because it does not result in any reduction in dimensionality. A loss in dimensionality would cause a loss of information because it decreases variability of the independent variables. In general, Mahalanobis  $D^2$  is the preferred procedure when the researcher is interested in the maximal use of available information in a stepwise process.

**Significance of Individual Functions** If the number of groups is three or more, then the researcher must decide not only whether the discrimination between groups overall is statistically significant but also whether each of the estimated discriminant functions is statistically significant. As discussed earlier, discriminant analysis estimates one less discriminant function than there are groups. If three groups are analyzed, then two discriminant functions will be estimated; for four groups, three functions will be estimated; and so on. The computer programs all provide the researcher the information necessary to ascertain the number of functions needed to obtain statistical significance, without including discriminant functions that do not increase the discriminatory power significantly.

The conventional significance criterion of .05 or beyond is often used, yet some researchers extend the required significance level (e.g., .10 or more) based on the trade-off of cost versus the value of the information. If the higher levels of risk for including nonsignificant results (e.g., significance levels  $> .05$ ) are acceptable, discriminant functions may be retained that are significant at the .2 or even the .3 level.

If one or more functions are deemed not statistically significant, the discriminant model should be re-estimated with the number of functions to be derived limited to the number of significant functions. In this manner, the assessment of predictive accuracy and the interpretation of the discriminant functions will be based only on significant functions.

**Explanatory Power of Individual Discriminant Functions** In addition to measures of statistical significance, the amount of variance explained by each function can be calculated as the square of the canonical correlation for that function. Thus, if the canonical correlation is .50, then the discriminant function explains 25 percent ( $.50^2$ ) of the variation among groups. Comparable to effect sizes in other analyses such as ANOVA/MANOVA and regression, these measures provide the researcher which empirical measure for assessing both the absolute and relative explanatory power of the discriminant function(s).

## ASSESSING OVERALL MODEL FIT

Once the significant discriminant functions have been identified, attention shifts to ascertaining the overall fit of the retained discriminant function(s). This assessment involves three tasks:

- 1 Calculating discriminant Z scores for each observation
- 2 Evaluating group differences on the discriminant Z scores
- 3 Assessing group membership prediction accuracy.

The discriminant Z score is calculated for each discriminant function for every observation in the sample. The discriminant score acts as a concise and simple representation of each discriminant function, simplifying the interpretation process and the assessment of the contribution of independent variables. Groups can be distinguished by their discriminant scores and, as we will see, the discriminant scores can play an instrumental role in predicting group membership.

**Calculating Discriminant Z Scores** With the retained discriminant functions defined, the basis for calculating the discriminant Z scores has been established. As discussed earlier, the discriminant Z score of any discriminant function can be calculated for each observation by the following formula:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk}$$

where:

$Z_{jk}$  = discriminant Z score of discriminant function  $j$  for object  $k$

$a$  = intercept

$W_i$  = discriminant coefficient for independent variable  $i$

$X_{ik}$  = independent variable  $i$  for object  $k$

## Model Estimation and Model Fit

Although stepwise estimation may seem optimal by selecting the most parsimonious set of maximally discriminating variables, beware of the impact of multicollinearity on the assessment of each variable's discriminatory power.

Overall model fit assesses the statistical significance between groups on the discriminant Z score(s), but does not assess predictive accuracy.

With more than two groups, do not confine your analysis to only the statistically significant discriminant function(s), but consider if nonsignificant functions (with significance levels of up to .3) add explanatory power.

The discriminant Z score, a metric variable, provides a direct means of comparing observations on each function. Observations with similar Z scores are assumed more alike on the variables constituting this function than those with disparate scores. The discriminant function can be expressed with either standardized or unstandardized weights and values. The standardized version is more useful for interpretation purposes, but the unstandardized version is easier to use in calculating the discriminant Z score.

**Evaluating Group Differences** Once the discriminant Z scores are calculated, the first assessment of overall model fit is to determine the magnitude of differences between the members of each group in terms of the discriminant Z scores. A summary measure of the group differences is a comparison of the group centroids, the average discriminant Z score for all group members. A measure of success of discriminant analysis is its ability to define discriminant function(s) that result in significantly different group centroids. The differences between centroids are measured in terms of Mahalanobis  $D^2$  measure, for which tests are available to determine whether the differences are statistically significant. The researcher should ensure that even with significant discriminant functions, significant differences occur between each of the groups.

Group centroids on each discriminant function can also be plotted to demonstrate the results from a graphical perspective. Plots are usually prepared for the first two or three discriminant functions (assuming they are statistically significant functions). The values for each group show its position in reduced discriminant space (so called because not all of the functions and thus not all of the variance are plotted). The researcher can see the differences between the groups on each function; however, visual inspection does not totally explain what these differences are. Circles can be drawn enclosing the distribution of observations around their respective centroids to clarify group differences further, but this procedure is beyond the scope of this text (see Dillon and Goldstein [3]).

**Assessing Group Membership Prediction Accuracy** Given that the dependent variable is nonmetric, it is not possible to use a measure such as  $R^2$ , as is done in multiple regression, to assess predictive accuracy. Rather, each observation must be assessed as to whether it was correctly classified. In doing so, several major considerations must be addressed:

- The statistical and practical rationale for developing classification matrices
- Classifying individual cases
- Construction of the classification matrices
- Standards for assessing classification accuracy.

**WHY CLASSIFICATION MATRICES ARE DEVELOPED** The statistical tests for assessing the significance of the discriminant function(s) only assess the degree of difference between the groups based on the discriminant Z scores, but do not

indicate how well the function(s) predicts. These statistical tests suffer the same drawbacks as the classical tests of hypotheses. For example, suppose the two groups are deemed significantly different beyond the .01 level. Yet with sufficiently large sample sizes, the group means (centroids) could be virtually identical and still have statistical significance. To determine the predictive ability of a discriminant function, the researcher must construct classification matrices.

The **classification matrix** procedure provides a perspective on practical significance rather than statistical significance. With multiple discriminant analysis, the **percentage correctly classified**, also termed the **hit ratio**, reveals how well the discriminant function classified the objects. With a sufficiently large sample size in discriminant analysis, we could have a statistically significant difference between the two (or more) groups and yet correctly classify only 53 percent (when chance is 50%, with equal group sizes) [13]. In such instances, the statistical test would indicate statistical significance, yet the hit ratio would allow for a separate judgment to be made in terms of practical significance. Thus, we must use the classification matrix procedure to assess predictive accuracy beyond just statistical significance.

**CLASSIFYING INDIVIDUAL OBSERVATIONS** The development of classification matrices requires that each observation be classified into one of the groups of the dependent variable based on the discriminant function(s). The objective is to characterize each observation on the discriminant function(s) and then determine the extent to which observation in each group can be consistently described by the discriminant functions. There are two approaches to classifying observations, one employing the discriminant scores directly and another developing a specific function for classification. Each approach will be discussed in the following section as well as the importance of determining the role that the sample size for each group plays in the classification process.

**Cutting Score Calculation** Using the discriminant functions deemed significant, we can develop classification matrices by calculating the **cutting score** (also called the *critical Z value*) for each discriminant function. The cutting score is the criterion against which each object's discriminant score is compared to determine into which group the object should be classified. The cutting score represents the dividing point used to classify observations into groups based on their discriminant function score. The calculation of a cutting score between any two groups is based on the two group centroids (group mean of the discriminant scores) and the relative size of the two groups. The group centroids are easily calculated and provided at each stage of the stepwise process.

**Developing a Classification Function** As noted earlier, using the discriminant function is only one of two possible approaches to classification. The second approach employs a **classification function**, also known as **Fisher's linear discriminant function**. The classification functions, one for each group, are used strictly for classifying observations. In this method of classification, an observation's values for the independent variables are inserted in the classification functions and a classification score for each group is calculated for that observation. The observation is then classified into the group with the highest classification score.

**DEFINING PRIOR PROBABILITIES** The impact and importance of each group's sample size in the classification process is many times overlooked, yet is critical in making the appropriate assumptions in the classification process. Do the relative group sizes tell us something about the expected occurrence of each group in the population or are they just an artifact of the data collection process? Here we are concerned about the representativeness of the sample as it relates to representation of the relative sizes of the groups in the actual in the actual population, which can be stated as prior probabilities (i.e., the relative proportion of each group to the total sample).

The fundamental question is: Are the relative group sizes representative of the group sizes in the population? The default assumption for most statistical programs is equal prior probabilities; in other words, each group is assumed to have an equal chance of occurring even if the group sizes in the sample are unequal. If the researcher is unsure about whether the observed proportions in the sample are representative of the population proportions, the conservative approach is to employ equal probabilities. In some instances estimates of the prior probabilities may be available, such as from prior research. Here the default assumption of equal prior probabilities is replaced

with values specified by the researcher. In either instance, the actual group sizes are replaced based on the specified prior probabilities.

If, however, the sample was conducted randomly and the researcher feels that the group sizes are representative of the population, then the researcher can specify prior probabilities to be based on the estimation sample. Thus, the actual group sizes are assumed representative and used directly in the calculation of the cutting score (see the following discussion). In all instances, however, the researcher must specify how the prior probabilities are to be calculated, which affects the group sizes used in the calculation as illustrated.

For example, consider a holdout sample consisting of 200 observations, with group sizes of 60 and 140 that relate to prior probabilities of 30 percent and 70 percent, respectively. If the sample is assumed representative, then the sample sizes of 60 and 140 are used in calculating the cutting score. If, however, the sample is deemed not representative, the researcher must specify the prior probabilities. If they are specified as equal (50% and 50%), sample sizes of 100 and 100 would be used in the cutting score calculation rather than the actual sample sizes. Specifying other values for the prior probabilities would result in differing sample sizes for the two groups.

**Calculating the Optimal Cutting Score** The importance of the prior probabilities can be illustrated in the calculation of the “optimal” cutting score, which takes into account the prior probabilities through the use of group sizes. The basic formula for computing the **optimal cutting score** between any two groups is:

$$Z_{CS} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B}$$

where:

$Z_{CS}$  = optimal cutting score between groups A and B

$N_A$  = number of observations in group A

$N_B$  = number of observations in group B

$Z_A$  = centroid for group A

$Z_B$  = centroid for group B

With unequal group sizes, the optimal cutting score for a discriminant function is now the weighted average of the group centroids. The cutting score is weighted toward the smaller group, hopefully making for a better classification of the larger group.

If the groups are specified to be of equal size (prior probabilities defined as equal), then the optimum cutting score will be halfway between the two group centroids and becomes simply the average of the two centroids:

$$Z_{CE} = \frac{Z_A + Z_B}{2}$$

where:

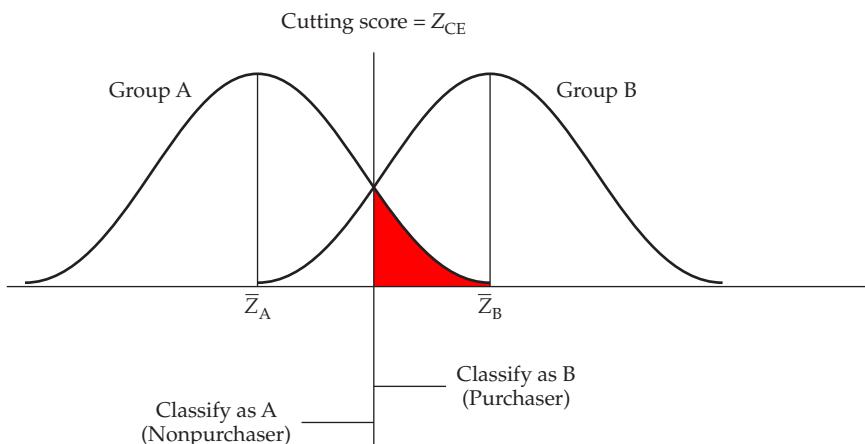
$Z_{CE}$  = critical cutting score value for equal group sizes

$Z_A$  = centroid for group A

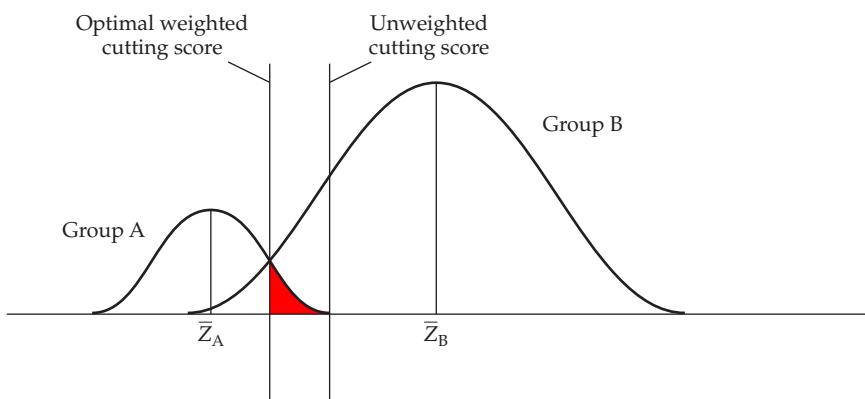
$Z_B$  = centroid for group B

Both of the formulas for calculating the optimal cutting score assume that the distributions are normal and the group dispersion structures are known.

The concept of an optimal cutting score for equal and unequal groups is illustrated in Figures 7.7 and 7.8, respectively. Both the weighted and unweighted cutting scores are shown. It is apparent that if group A is much smaller than group B, the optimal cutting score will be closer to the centroid of group A than to the centroid of group B. Also,



**Figure 7.7**  
Optimal Cutting Score with  
Equal Sample Sizes



**Figure 7.8**  
Optimal Cutting Score with  
Unequal Sample Sizes

if the unweighted cutting score was used, none of the objects in group A would be misclassified, but a substantial portion of those in group B would be misclassified.

**COSTS OF MISCLASSIFICATION** The optimal cutting score also must consider the cost of classifying an object into the wrong group. If the costs of misclassifying are approximately equal for all groups, the optimal cutting score will be the one that will misclassify the fewest number of objects across all groups. If the misclassification costs are unequal, the optimum cutting score will be the one that minimizes the costs of misclassification. More sophisticated approaches to determining cutting scores are discussed in Dillon and Goldstein [3] and Huberty et al. [11]. These approaches are based on a Bayesian statistical model and are appropriate when the costs of misclassification into certain groups are high, when the groups are of grossly different sizes, or when one wants to take advantage of a priori knowledge of group membership probabilities.

In practice, when calculating the cutting score, it is not necessary to insert the raw variable measurements for every individual into the discriminant function and to obtain the discriminant score for each person to use in computing the  $Z_A$  and  $Z_B$  (group A and B centroids). The computer program will provide the discriminant scores as well as the  $Z_A$  and  $Z_B$  as regular output. When the researcher has the group centroids and sample sizes, the optimal cutting score can be obtained by merely substituting the values into the appropriate formula.

**CONSTRUCTING CLASSIFICATION MATRICES** To validate the discriminant function through the use of classification matrices, the sample should be randomly divided into two groups. One of the groups (the analysis sample) is used to compute the discriminant function. The other group (the holdout or validation sample) is retained for use in developing the classification matrix.

The classification of each observation can be accomplished through either of the classification approaches discussed earlier. For the Fisher's approach, an observation is classified into the group with the largest classification function score. When using the discriminant scores and the optimal cutting score, the procedure is as follows:

Classify an individual into group A if  $Z_n < Z_{ct}$

or:

Classify an individual into group B if  $Z_n > Z_{ct}$

where:

$Z_n$  = discriminant Z score for the  $n$ th individual

$Z_{ct}$  = critical cutting score value

The results of the classification procedure are presented in matrix form, as shown in Table 7.4. The entries on the diagonal of the matrix represent the number of individuals correctly classified. The numbers off the diagonal represent the incorrect classifications. The entries under the column labeled "Actual Group Size" represent the number of individuals actually in each of the two groups. The entries at the bottom of the columns represent the number of individuals assigned to the groups by the discriminant function. The percentage correctly classified for each group is shown at the right side of the matrix, and the overall percentage correctly classified, also known as the hit ratio, is shown at the bottom. Chapter 8 discusses additional measures of predictive accuracy specific to logistic regression.

In our example, the number of individuals correctly assigned to group 1 is 22, whereas three members of group 1 are incorrectly assigned to group 2. Similarly, the number of correct classifications to group 2 is 20, and the number of incorrect assignments to group 1 is 5. Thus, the classification accuracy percentages of the discriminant function for the actual groups 1 and 2 are 88 and 80 percent, respectively. The overall classification accuracy (hit ratio) is 84 percent.

One final topic regarding classification procedures is the  $t$  test available to determine the level of significance for the classification accuracy. The formula for a two-group analysis (equal sample size) is:

$$t = \frac{p - .5}{\sqrt{\frac{.5(1.0 - .5)}{N}}}$$

where:

$p$  = proportion correctly classified

$N$  = sample size

This formula can be adapted for use with more groups and unequal sample sizes.

**ESTABLISHING STANDARDS OF COMPARISON FOR THE HIT RATIO** As noted earlier, the predictive accuracy of the discriminant function is measured by the hit ratio, which is obtained from the classification matrix. The researcher may ask, What

**Table 7.4** Classification Matrix for Two-Group Discriminant Analysis

		<i>Predicted Group</i>		<i>Actual Group</i>	<i>Percentage Correctly Classified</i>
<i>Actual Group</i>	<b>1</b>	<b>2</b>	<b>Size</b>		
1	22	3	25		88
2	5	20	25		80
Predicted group size	27	23	50		84 <sup>a</sup>

<sup>a</sup> Percent correctly classified = (Number correctly classified/Total number of observations)  $\times 100$   
 $= [(22 + 20)/50] \times 100 = 84\%$ .

is considered an acceptable level of predictive accuracy for a discriminant function? For example, is 60 percent an acceptable level, or should one expect to obtain 80 to 90 percent predictive accuracy? To answer this question, the researcher must first determine the percentage that could be classified correctly by *chance (without the aid of the discriminant function)*.

**Standards of Comparison for the Hit Ratio for Equal Group Sizes** When the sample sizes of the groups are equal, the determination of the chance classification is rather simple; it is obtained by dividing 1 by the number of groups. The formula is:

$$C_{\text{EQUAL}} = 1 \div \text{Number of groups}$$

For instance, for a two-group function the chance probability would be .50; for a three-group function the chance probability would be .33; and so forth.

**Standards of Comparison for the Hit Ratio for Unequal Group Sizes** The determination of the chance classification for situations in which the group sizes are unequal is somewhat more involved. Should we consider just the largest group, the combined probability of all different size groups, or some other standard? Let us assume that we have a total sample of 200 observations divided into holdout and analysis samples of 100 observations each. In the holdout sample, 75 subjects belong to one group and 25 to the other. We will examine the possible ways in which we can construct a standard for comparison and what each represents:

- *Maximum chance criterion.* In the simplest approach, we could arbitrarily assign all the subjects to the largest group. As such, the maximum chance criterion should be used when the sole objective of the discriminant analysis is to maximize the percentage correctly classified [13]. It is also the most conservative standard because it will generate the highest standard of comparison. However, situations in which we are concerned only about maximizing the percentage correctly classified are rare. Usually the researcher uses discriminant analysis to correctly identify members of all groups. In cases where the sample sizes are unequal and the researcher wants to classify members of all groups, the discriminant function defies the odds by classifying a subject in the smaller group(s). The maximum chance criterion does not take this fact into account [13].

In our simple example of a sample with two groups (75 and 25 people each), using this method would set a 75 percent classification accuracy, what would be achieved by classifying everyone into the largest group without the aid of any discriminant function. It could be concluded that unless the discriminant function achieves a classification accuracy higher than 75 percent, it should be disregarded because it has not helped us improve the prediction accuracy we could achieve without using any discriminant analysis at all.

- *Proportional chance criterion.* This is most appropriate when group sizes are unequal and the researcher wishes to correctly identify members of all of the groups, not just the largest group. The formula for this criterion is:

$$C_{\text{PRO}} = p^2 + (1 - p)^2$$

where:

$$p = \text{proportion of individuals in group 1}$$

$$1 - p = \text{proportion of individuals in group 2}$$

Using the group sizes from our earlier example (75 and 25), we see that the proportional chance criterion would be 62.5 percent [ $.75^2 + (1.0 - .75)^2 = .625$ ] compared with 75 percent. Therefore, in this instance, the actual prediction accuracy of 75 percent may be acceptable because it is above the 62.5 percent proportional chance criterion.

**Issues Concerning Sample Size in Standards of Comparison** An issue with either the maximum chance or proportional chance criteria is the sample sizes used for calculating the standard. Do you use the group sizes from the overall sample, the analysis/estimation sample, or the validation/holdout sample? A couple of suggestions:

- If the sample sizes of the analysis and estimation samples are each deemed sufficiently large (i.e., total sample of 100 with each group having at least 20 cases), derive separate standards for each sample.
- If the separate samples are not deemed sufficiently large, use the group sizes from the total sample in calculating the standards.
- Be aware of differing group sizes between samples when using the maximum chance criterion, because it is dependent on the largest group size. This guideline is especially critical when the sample size is small or when group size proportions vary markedly from sample to sample. It is another reason to be cautious in the use of the maximum chance criterion.

**Use With The Holdout Sample** These chance model criteria are useful only when computed with holdout samples (split-sample approach). If the individuals used in calculating the discriminant function are the ones being classified, the result will be an upward bias in the prediction accuracy. In such cases, both of these criteria would have to be adjusted upward to account for this bias.

**COMPARING THE HIT RATIO TO THE STANDARD** The question of “How high does classification accuracy have to be?” is crucial. If the percentage of correct classifications is significantly larger than would be expected by chance, the researcher can proceed in interpreting the discriminant functions and group profiles. However, if the classification accuracy is no greater than can be expected by chance, whatever differences appear to exist actually merit little or no interpretation; that is, differences in score profiles would provide no meaningful information for identifying group membership.

The question, then, is how high should the classification accuracy be relative to chance? For example, if chance is 50 percent (two-group, equal sample size), does a classification (predictive) accuracy of 60 percent justify moving to the interpretation stage? Ultimately, the decision depends on the cost relative to the value of the information. The cost-versus-value argument offers little assistance to the neophyte data researcher, but the following criterion is suggested: *The classification accuracy should be at least one-fourth greater than that achieved by chance.*

For example, if chance accuracy is 50 percent, the classification accuracy should be 62.5 percent ( $62.5\% = 1.25 \times 50\%$ ). If chance accuracy is 30 percent, the classification accuracy should be 37.5 percent ( $37.5\% = 1.25 \times 30\%$ ).

This criterion provides only a rough estimate of the acceptable level of predictive accuracy. The criterion is easy to apply with groups of equal size. With groups of unequal size, an upper limit is reached when the maximum chance model is used to determine chance accuracy. It does not present too great a problem, however, because under most circumstances, the maximum chance model would not be used with unequal group sizes.

**OVERALL VERSUS GROUP-SPECIFIC HIT RATIOS** To this point, we focused on evaluating the overall hit ratio across all groups in assessing the predictive accuracy of a discriminant analysis. The researcher also must be concerned with the hit ratio (percent correctly classified) for each separate group. If you focus solely on the overall hit ratio, it is possible that one or more groups, particularly smaller groups, may have unacceptable hit ratios while the overall hit ratio is acceptable. The researcher should evaluate each group's hit ratio and assess whether the discriminant analysis provides adequate levels of predictive accuracy both at the overall level as well as for each group.

**STATISTICALLY-BASED MEASURES OF CLASSIFICATION ACCURACY RELATIVE TO CHANCE** A statistical test for the discriminatory power of the classification matrix when compared with a chance model is **Press's Q statistic**. This simple measure compares the number of correct classifications with the total sample size and the number of groups. The calculated value is then compared with a critical value (the chi-square value for one degree of freedom at the desired confidence level). If it exceeds this critical value, then the classification matrix can be deemed statistically better than chance. The Q statistic is calculated by the following formula:

$$\text{Press's } Q = \frac{[N - (nK)]^2}{N(K - 1)}$$

where:

$N$  = total sample size

$n$  = number of observations correctly classified

$K$  = number of groups

For example, in Table 7.4, the Q statistic would be based on a total sample of  $N = 50$ ,  $n = 42$  correctly classified observations, and  $K = 2$  groups. The calculated statistic would be:

$$\text{Press's } Q = \frac{[50 - (42 \times 2)]^2}{50(2 - 1)} = 23.12$$

The critical value at a significance level of .01 is 6.63. Thus, we would conclude that in the example the predictions were significantly better than chance, which would have a correct classification rate of 50 percent.

This simple test is sensitive to sample size; large samples are more likely to show significance than small sample sizes of the same classification rate.

For example, if the sample size is increased to 100 in the example and the classification rate remains at 84 percent, the Q statistic increases to 46.24. If the sample size increases to 200, but retains the classification rate of 84 percent, the Q statistic increases again to 92.48. But if the sample size was only 20 and the misclassification rate was still 84 percent (17 correct predictions), the Q statistic would be only 9.8. Thus, examine the Q statistic in light of the sample size because increases in sample size will increase the Q statistic even for the same overall classification rate.

One must be careful in drawing conclusions based solely on this statistic, however, because as the sample sizes become larger, a lower classification rate will still be deemed significant.

## CASEWISE DIAGNOSTICS

The final means of assessing model fit is to examine the predictive results on a case-by-case basis. Similar to the analysis of residuals in multiple regression, the objective is to understand which observations (1) have been misclassified and (2) are not representative of the remaining group members. Although the classification matrix provides overall classification accuracy, it does not detail the individual case results. Also, even if we can denote which cases are correctly or incorrectly classified, we still need a measure of an observation's similarity to the remainder of the group.

**Misclassification of Individual Cases** When analyzing residuals from a multiple regression analysis, an important decision involves setting the level of residual considered substantive and worthy of attention. In discriminant analysis, this issue is somewhat simpler because an observation is either correctly or incorrectly classified. All computer programs provide information that identifies which cases are misclassified and to which group they were misclassified. The researcher can identify not only those cases with classification errors, but a direct representation of the type of misclassification error.

**Analyzing Misclassified Cases** The purpose of identifying and analyzing the misclassified observations is to identify any characteristics of these observations that could be incorporated into the discriminant analysis for improving predictive accuracy. This analysis may take the form of profiling the misclassified cases on either the independent variables or other variables not included in the model.

**PROFILING ON THE INDEPENDENT VARIABLES** Examining these cases on the independent variables may identify nonlinear trends or other relationships or attributes that led to the misclassification. Several techniques are particularly appropriate in discriminant analysis:

**Graphical Portrayal** A graphical representation of the observations is perhaps the simplest yet effective approach for examining the characteristics of observations, especially the misclassified observations. The most common approach is to plot the observations based on their discriminant  $Z$  scores and portray the overlap among groups and the misclassified cases. If two or more functions are retained, the optimal cutting points can also be portrayed to give what is known as a **territorial map** depicting the regions corresponding to each group.

Plotting the individual observations along with the group centroids, as discussed earlier, shows not only the general group characteristics depicted in the centroids, but also the variation in the group members. It is analogous to the areas defined in the three-group example at the beginning of this chapter, in which cutting scores on both functions defined areas corresponding to the classification predictions for each group.

**Empirical Measures** A direct empirical assessment of the similarity of an observation to the other group members can be made by evaluating the Mahalanobis  $D^2$  distance of the observation to the group centroid. Based on the set of independent variables, observations closer to the centroid have a smaller Mahalanobis  $D^2$  and are assumed more representative of the group than those farther away.

The empirical measure should be combined with a graphical analysis, however, because although a large Mahalanobis  $D^2$  value does indicate observations that are quite different from the group centroids, it does not always indicate misclassification. For example, in a two-group situation, a member of group A may have a large Mahalanobis  $D^2$  distance, indicating it is less representative of the group. However, if that distance is away from the group B centroid, then it would actually increase the chance of correct classification, even though it is less representative of the group. A smaller distance that places an observation between the two centroids would probably have a lower probability of correct classification, even though it is closer to its group centroid than the earlier situation.

Although no prespecified analyses are established, such as found in multiple regression, the researcher is encouraged to evaluate these misclassified cases from several perspectives in attempting to uncover the unique features they hold in comparison to their other group members.

## Assessing Model Fit and Predictive Accuracy

The classification matrix and hit ratio replace  $R^2$  as the measure of model fit:

Assess the hit ratio both overall and by group

If the estimation and analysis samples both exceed 100 cases and each group exceeds 20 cases, derive separate standards for each sample; if not, derive a single standard from the overall sample.

Multiple criteria are used for comparison to the hit ratio:

The maximum chance criterion for evaluating the hit ratio is the most conservative, giving the highest baseline value to exceed

Be cautious in using the maximum chance criterion in situations with overall samples less than 100 and/or group sizes under 20

The proportional chance criterion considers all groups in establishing the comparison standard and is the most popular

The actual predictive accuracy (hit ratio) should exceed any criterion value by at least 25 percent.

Analyze the misclassified observations both graphically (territorial map) and empirically (Mahalanobis  $D^2$ ).

## Stage 5: Interpretation of the Results

If the discriminant function is statistically significant and the classification accuracy is acceptable, the researcher should focus on making substantive interpretations of the findings. This process involves examining the discriminant functions to determine the relative importance of each independent variable in discriminating between the groups. Three methods of determining the relative importance have been proposed:

- 1 Standardized discriminant weights
- 2 Discriminant loadings (structure correlations)
- 3 Partial *F* values.

### DISCRIMINANT WEIGHTS

The traditional approach to interpreting discriminant functions examines the sign and magnitude of the standardized **discriminant weight** (also referred to as a **discriminant coefficient**) assigned to each variable in computing the discriminant functions. When the sign is ignored, each weight represents the relative contribution of its associated variable to that function. Independent variables with relatively larger weights contribute more to the discriminating power of the function than do variables with smaller weights. The sign denotes only that the variable makes either a positive or a negative contribution [3].

The interpretation of discriminant weights is analogous to the interpretation of beta weights in regression analysis and is therefore subject to the same criticisms. For example, a small weight may indicate either that its corresponding variable is irrelevant in determining a relationship or that it has been partialled out of the relationship because of a high degree of multicollinearity. Another problem with the use of discriminant weights is that they are subject to considerable instability. These problems suggest caution in using weights to interpret the results of discriminant analysis.

### DISCRIMINANT LOADINGS

**Discriminant loadings**, referred to sometimes as **structure correlations**, are increasingly used as a basis for interpretation because of the deficiencies in utilizing weights. Measuring the simple linear correlation between each independent variable and the discriminant function, the discriminant loadings reflect the variance that the independent variables share with the discriminant function. In that regard they can be interpreted like factor loadings in assessing the relative contribution of each independent variable to the discriminant function. (Chapter 3 further discusses factor-loading interpretation.)

One unique characteristic of loadings is that loadings can be calculated for all variables, whether they were used in the estimation of the discriminant function or not. This aspect is particularly useful when a stepwise estimation procedure is employed and some variables are not included in the discriminant function. Rather than having no way to understand their relative impact, loadings provide a relative effect of every variable on a common measure.

With the loadings, the primary question is: What value must loadings attain to be considered substantive discriminators worthy of note? In either simultaneous or stepwise discriminant analysis, variables that exhibit a loading of  $\pm .40$  or higher are considered substantive. With stepwise procedures, this determination is supplemented because the technique prevents nonsignificant variables from entering the function. However, multicollinearity and other factors may preclude a variable from entering the equation, which does not necessarily mean that it does not have a substantial effect. The researcher must be cautioned, however, that use of the discriminant loadings also is somewhat similar to univariate analyses, since the structure correlation is not controlled for any other variables in the analysis. So the discriminant loadings may overstate the “relative” importance of multicollinear variables in terms of actual discriminatory power when used simultaneously.

Discriminant loadings (like weights) may be subject to instability. Loadings are considered relatively more valid than weights as a means of interpreting the discriminating power of independent variables because of their correlational nature. The researcher still must be cautious when using loadings to interpret discriminant functions.

## PARTIAL F VALUES

As discussed earlier, two computational approaches—simultaneous and stepwise—can be utilized in deriving discriminant functions. When the stepwise method is selected, an additional means of interpreting the relative discriminating power of the independent variables is available through the use of partial *F* values. It is accomplished by examining the absolute sizes of the significant *F* values and ranking them. Large *F* values indicate greater discriminatory power. In practice, rankings using the *F* values approach are the same as the ranking derived from using discriminant weights, but the *F* values indicate the associated level of significance for each variable.

## INTERPRETATION OF TWO OR MORE FUNCTIONS

In cases of two or more significant discriminant functions, we are faced with additional problems of interpretation. First, can we simplify the discriminant weights or loadings to facilitate the profiling of each function? Second, how do we represent the impact of each variable across all functions? These problems are found both in measuring the total discriminating effects across functions and in assessing the role of each variable in profiling each function separately. We address these two questions by introducing the concepts of rotation of the functions, the potency index, and stretched vectors representations.

**Rotation of the Discriminant Functions** After the discriminant functions are developed, they can be rotated to redistribute the variance. The concept of rotation is more fully explained in Chapter 3. Basically, rotation preserves the original structure and the reliability of the discriminant solution while making the functions easier to interpret substantively. In most instances, the VARIMAX rotation is employed as the basis for rotation.

**Potency Index** Previously, we discussed using the standardized weights or discriminant loadings as measures of a variable's contribution to a discriminant function. When two or more functions are derived, however, a composite or summary measure is useful in describing the contributions of a variable across all significant functions. The **potency index** is a relative measure among all variables and is indicative of each variable's discriminating power [14]. It includes both the contribution of a variable to a discriminant function (its discriminant loading) and the relative contribution of the function to the overall solution (a relative measure among the functions based on eigenvalues). The composite is simply the sum of the individual potency indices across all significant discriminant functions. Interpretation of the composite measure is limited, however, by the fact that it is useful only in depicting the relative position (such as the rank order) of each variable, and the absolute value has no real meaning. The potency index is calculated by a two-step process:

**Step 1:** Calculate a potency value of each variable for each significant function. In the first step, the discriminating power of a variable, represented by the squared value of the unrotated discriminant loading, is “weighted” by the relative contribution of the discriminant function to the overall solution. First, the relative eigenvalue measure for each significant discriminant function is calculated simply as:

$$\text{Relative eigenvalue of discriminant function } j = \frac{\text{Eigenvalue of discriminant function } j}{\text{Sum of eigenvalues across all significant functions}}$$

The potency value of each variable on a discriminant function is then:

$$\text{Potency value of variable } i \text{ on function } j = (\text{Discriminant loading}_{ij})^2 \times \text{Relative eigenvalue of function } j$$

**Step 2:** Calculate a composite potency index across all significant functions. Once a potency value has been calculated for each function, the composite potency index for each variable is calculated as:

$$\text{Composite potency of variable } i = \frac{\text{Sum of potency values of variable } i \text{ across all significant discriminant functions}}{\text{Number of significant discriminant functions}}$$

The potency index now represents the total discriminating effect of the variable across all of the significant discriminant functions. It is only a relative measure, however, and its absolute value has no substantive meaning. An example of calculating the potency index is provided in the three-group example for discriminant analysis.

**Graphical Display of Discriminant Scores and Loadings** To depict group differences on the predictor variables, the researcher can use two different approaches to graphical display. The territorial map plots the individual cases on the significant discriminant functions to enable the researcher to assess the relative position of each observation based on the discriminant function scores. The second approach is to plot the discriminant loadings to understand the relative grouping and magnitude of each loading on each function. Each approach will be discussed in more detail in the following section.

**TERRITORIAL MAP** The most common graphical method is the territorial map, where each observation is plotted in a graphical display based on the discriminant function Z scores of the observations. For example, assume that a three-group discriminant analysis had two significant discriminant functions. A territorial map is created by plotting each observation's discriminant Z scores for the first discriminant function on the X axis and the scores for the second discriminant function on the Y axis. As such, it provides several perspectives on the analysis:

- Plotting each group's members with differing symbols allows for an easy portrayal of the distinctiveness of each group as well as its overlap with each other group.
- Plotting each group's centroids provides a means for assessing each group member relative to its group centroid. This procedure is particularly useful when assessing whether large Mahalanobis  $D^2$  measures lead to misclassification.
- Lines representing the cutting scores can also be plotted, denoting boundaries depicting the ranges of discriminant scores predicted into each group. Any group's members falling outside these boundaries are misclassified. Denoting the misclassified cases allows for assessing which discriminant function was most responsible for the misclassification as well as the degree to which a case is misclassified.

**VECTOR PLOT OF DISCRIMINANT LOADINGS** The simplest graphical approach to depicting discriminant loadings is to plot the actual rotated or unrotated loadings on a graph. The preferred approach would be to plot the rotated loadings. Similar to the graphical portrayal of factor loadings (see Chapter 3), this method depicts the degree to which each variable is associated with each discriminant function.

An even more accurate approach, however, involves plotting the loadings as well as depicting vectors for each loading and group centroid. A **vector** is merely a straight line drawn from the origin (center) of a graph to the coordinates of a particular variable's discriminant loadings or a group centroid. With a **stretched vector** representation, the length of each vector becomes indicative of the relative importance of each variable in discriminating among the groups. The plotting procedure proceeds in three steps:

**Step 1: Selecting Variables.** All variables, whether included in the model as significant or not, may be plotted as vectors. In this way, the importance of collinear variables that are not included, such as in a stepwise solution, can still be portrayed.

**Step 2: Stretching the Vectors.** Each variable's discriminant loadings are stretched by multiplying the discriminant loading (preferably after rotation) by its respective univariate  $F$  value. We note that vectors point toward the groups having the highest mean on the respective predictor and away from the groups having the lowest mean scores.

**Step 3: Plotting the Group Centroids.** The group centroids are also stretched in this procedure by multiplying them by the approximate  $F$  value associated with each discriminant function. If the loadings are stretched, the centroids must be stretched as well to plot them accurately on the same graph. The approximate  $F$  values for each discriminant function are obtained by the following formula:

$$F \text{ value}_{\text{Function}_i} = \text{Eigenvalue}_{\text{Function}_i} \left( \frac{N_{\text{Estimation Sample}} - NG}{NG - 1} \right)$$

where:

$$N_{\text{Estimation Sample}} = \text{sample size of estimation sample}$$

As an example, assume that the sample of 50 observations was divided into three groups. The multiplier of each eigenvalue would be  $(50 - 3) \div (3 - 1) = 23.5$ .

When completed, the researcher has a portrayal of the grouping of variables on each discriminant function, the magnitude of the importance of each variable (represented by the length of each vector), and the profile of each group centroid (shown by the proximity to each vector). Although this procedure must be done manually in most instances, it provides a complete portrayal of both discriminant loadings and group centroids. For more details on this procedure, see Dillon and Goldstein [3].

### WHICH INTERPRETIVE METHOD TO USE?

Several methods for interpreting the nature of discriminant functions have been discussed, both for single- and multiple-function solutions. Which methods should be used? The loadings approach is more valid than the use of weights and should be utilized whenever possible. The use of univariate and partial  $F$  values enables the researcher to use several measures and look for some consistency in evaluations of the variables. If two or more functions are estimated, then the researcher can employ several graphical techniques and the potency index, which aid in interpreting the multidimensional solution. The most basic point is that the researcher should employ all available methods to arrive at the most accurate interpretation.

## Stage 6: Validation of the Results

The final stage of a discriminant analysis involves validating the discriminant results to provide assurances that the results have external as well as internal validity. *With the propensity of discriminant analysis to inflate the hit ratio if evaluated only on the analysis sample, validation is an essential step.* In addition to validating the hit ratios, the researcher should use group profiling to ensure that the group means are valid indicators of the conceptual model used in selecting the independent variables.

### VALIDATION PROCEDURES

Validation is a critical step in any discriminant analysis because many times, especially with smaller samples, the results can lack generalizability (external validity). The most common approach for establishing external validity is the assessment of hit ratios. Validation can occur either with a separate sample (holdout sample) or utilizing a procedure that repeatedly processes the estimation sample. External validity is supported when the hit ratio of the selected approach exceeds the comparison standards that represent the predictive accuracy expected by chance (see earlier discussion).

**Utilizing A Holdout Sample** Most often the validation of the hit ratios is performed by creating a holdout sample, also referred to as the **validation sample**. The purpose of utilizing a holdout sample for validation purposes is to see how well the discriminant function works on a sample of observations not used to derive the discriminant function. This process involves developing a discriminant function with the analysis sample and then applying it to the holdout sample. The justification for dividing the total sample into two groups is that an upward bias will occur in the prediction accuracy of the discriminant function if the individuals used in developing the classification matrix are the same as those used in computing the function; that is, the classification accuracy will be higher than is valid when applied to the estimation sample.

Other researchers have suggested that even greater confidence could be placed in the validity of the discriminant function by following this procedure several times [14]. Instead of randomly dividing the total sample into analysis and holdout groups once, the researcher would randomly divide the total sample into analysis and holdout samples

several times, each time testing the validity of the discriminant function through the development of a classification matrix and a hit ratio. Then the several hit ratios would be averaged to obtain a single measure.

**Cross-validation** The cross-validation approach to assessing external validity is performed with multiple subsets of the total sample [2, 3]. The most widely used approach is the jackknife method. Cross-validation is based on the “leave-one-out” principle. The most prevalent use of this method is to estimate  $k - 1$  subsamples, eliminating one observation at a time from a sample of  $k$  cases. A discriminant function is calculated for each subsample and then the predicted group membership of the eliminated observation is made with the discriminant function estimated on the remaining cases. After all of the group membership predictions have been made, one at a time, a classification matrix is constructed and the hit ratio calculated.

Cross-validation is quite sensitive to small sample sizes. Guidelines suggest that it be used only when the smallest group size is at least three times the number of predictor variables, and most researchers suggest a ratio of 5:1 [11]. However, cross-validation may represent the only possible validation approach in instances where the original sample is too small to divide into analysis and holdout samples but still exceeds the guidelines already discussed. Cross-validation is also becoming more widely used as major computer programs provide it as a program option.

## PROFILING GROUP DIFFERENCES

Another validation technique is to profile the groups on the independent variables to ensure their correspondence with the conceptual bases used in the original model formulation. After the researcher identifies the independent variables that make the greatest contribution in discriminating between the groups, the next step is to profile the characteristics of the groups based on the group means. This profile enables the researcher to understand the character of each group according to the predictor variables.

For example, referring to the Kitchenade survey data presented in Table 7.1, we see that the mean rating on “durability” for the “would purchase” group is 7.4, whereas the comparable mean rating on “durability” for the “would not purchase” group is 3.2. Thus, a profile of these two groups shows that the “would purchase” group rates the perceived durability of the new product substantially higher than the “would not purchase” group.

Another approach is to profile the groups on a separate set of variables that should mirror the observed group differences. This separate profile provides an assessment of external validity in that the groups vary on both the independent variable(s) and the set of associated variables. This technique is similar in character to the validation of derived clusters described in Chapter 4.

## Interpreting and Validating Discriminant Functions

Discriminant loadings are the preferred method to assess the contribution of each variable to a discriminant function because they are:

A standardized measure of importance (ranging from 0 to 1)

Available for all independent variables whether used in the estimation process or not

Unaffected by multicollinearity.

Loadings exceeding  $\pm .40$  are considered substantive for interpretation purposes.

In case of more than one discriminant function, be sure to:

Use rotated loadings

Assess each variable's contribution across all the functions with the potency index.

The discriminant function must be validated either with a holdout sample or one of the “leave-one-out” procedures.

## A Two-Group Illustrative Example

To illustrate the application of two-group discriminant analysis, we use variables drawn from the HBAT database introduced in Chapter 1. This example examines each of the six stages of the model-building process to a research problem particularly suited to multiple discriminant analysis.

### STAGE 1: OBJECTIVES OF DISCRIMINANT ANALYSIS

Recall that one of the customer characteristics obtained by HBAT in its survey was a categorical variable ( $X_4$ ) indicating the region in which the firm was located: USA/North America or Outside North America. HBAT's management team is interested in any differences in perceptions between those customers located and served by their USA-based salesforce versus those outside the United States who are served mainly by independent distributors. Despite any differences found in terms of sales support issues due to the nature of the salesforce serving each geographic area, the management team is interested to see whether the other areas of operations (product line, pricing, etc.) are viewed differently between these two sets of customers. This inquiry follows the obvious need by management to always strive to better understand their customer, in this instance by focusing on any differences that may occur between geographic areas. If any perceptions of HBAT are found to differ significantly between firms in these two regions, the company would then be able to develop strategies to remedy any perceived deficiencies and develop differentiated strategies to accommodate the differing perceptions.

To do so, discriminant analysis was selected to identify those perceptions of HBAT that best distinguish firms in each geographic region.

### STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS

The research design stage focuses on three key issues: selecting dependent and independent variables, assessing the adequacy of the sample size for the planned analysis, and dividing the sample for validation purposes.

**Selection of Dependent and Independent Variables** Discriminant analysis requires a single nonmetric dependent measure and one or more metric independent measures that are affected to provide differentiation between the groups based on the dependent measure.

Because the dependent variable Region ( $X_4$ ) is a two-group categorical variable, discriminant analysis is the appropriate technique. The survey collected perceptions of HBAT that can now be used to differentiate between the two groups of firms. Discriminant analysis uses as independent variables the 13 perception variables from the database ( $X_6$  to  $X_{18}$ ) to discriminate between firms in each geographic area.

**Sample Size** Given the relatively small size of the HBAT sample (100 observations), issues of sample size are particularly important, especially the division of the sample into analysis and holdout samples (see discussion in next section).

The sample of 100 observations, when split into analysis and holdout samples of 60 and 40 respectively, barely meets the suggested minimum 5:1 ratio of observations to independent variables (60 observations for 13 potential independent variables) in the analysis sample. Although this ratio would increase to almost 8:1 if the sample were not split, it was deemed more important to validate the results rather than to increase the number of observations in the analysis sample.

The two group sizes of 26 and 34 in the estimation sample also exceed the minimum size of 20 observations per group. Finally, the two groups are comparable enough in size to not adversely impact either the estimation or the classification processes.

**Division of the Sample** Previous discussion emphasized the need for validating the discriminant function by splitting the sample into two parts, one used for estimation and the other validation. Any time a holdout sample is

used, the researcher must ensure that the resulting sample sizes are sufficient to support the number of predictors included in the analysis.

The HBAT database has 100 observations; it was decided that a holdout sample of 40 observations would be sufficient for validation purposes. This split would still leave 60 observations for estimation of the discriminant function. Moreover, the relative group sizes in the estimation sample (26 and 34 in the two groups) would allow for estimation without complications due to markedly different group sizes.

It is important to ensure randomness in the selection of the holdout sample so that any ordering of the observations does not affect the processes of estimation and validation. The control cards necessary for both selection of the holdout sample and performance of the two-group discriminant analysis are shown online.

### STAGE 3: ASSUMPTIONS OF DISCRIMINANT ANALYSIS

The principal assumptions underlying discriminant analysis involve the formation of the variate or discriminant function (normality, linearity, and multicollinearity) and the estimation of the discriminant function (equal variance and covariance matrices). How to examine the independent variables for normality, linearity, and multicollinearity is explained in Chapter 2. For purposes of our illustration of discriminant analysis, these assumptions are met at acceptable levels.

Most statistical programs have one or more statistical tests for the assumption of equal covariance or dispersion matrices addressed in Chapter 2. The most common test is Box's M (for more detail, see Chapter 2).

In this two-group example, the significance of differences in the covariance matrices between the two groups is .011. Even though the significance is less than .05 (in this test the researcher looks for values above the desired significance level), the sensitivity of the test to factors other than just covariance differences (e.g., normality of the variables and increasing sample size) makes this an acceptable level.

No additional remedies are needed before estimation of the discriminant function can be performed.

### STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT

The researcher has the choice of two estimation approaches (simultaneous versus stepwise) in determining the independent variables included in the discriminant function. Once the estimation approach is selected, the process determines the composition of the discriminant function subject to the requirement for statistical significance specified by the researcher.

The primary objective of this analysis is to identify the set of independent variables (HBAT perceptions) that maximally differentiates between the two groups of customers. If the set of perception variables was smaller or the objective was simply to determine the discriminating capabilities of the entire set of perception variables, with no regard to the impact of any individual perception, then the simultaneous approach of entering all variables directly into the discriminant function would be employed. But in this case, even with the knowledge of multicollinearity among the perception variables seen in performing exploratory factor analysis (see Chapter 3), the stepwise approach is deemed most appropriate. We should note, however, that multicollinearity may impact which variables enter into the discriminant function and thus require particular attention in the interpretation process.

**Assessing Group Differences** Let us begin our assessment of the two-group discriminant analysis by examining Table 7.5, which shows the group means for each of the independent variables, based on the 60 observations constituting the analysis sample.

In profiling the two groups, we can first identify five variables with the largest differences in the group means ( $X_6$ ,  $X_{11}$ ,  $X_{12}$ ,  $X_{13}$ , and  $X_{17}$ ). Table 7.5 also shows the Wilks' lambda and univariate ANOVA used to assess the significance between means of the independent variables for the two groups. These tests indicate that the five perception variables are also the only variables with significant univariate differences between the two groups. Finally, the minimum Mahalanobis  $D^2$  values are also given. This value is important because it is the measure used to select variables for entry in the stepwise estimation process. Because only two groups are involved, the largest  $D^2$  value

also has the most significant difference between groups (note that the same is not necessarily so with three or more groups, where large differences between any two groups may not result in the largest overall differences across all groups, as will be shown in the three-group example).

Examining the group differences leads to identifying five perception variables ( $X_6, X_{11}, X_{12}, X_{13}$ , and  $X_{17}$ ) as the most logical set of candidates for entry into the discriminant analysis. This marked reduction from the larger set of 13 perception variables reinforces the decision to use a stepwise estimation process.

To identify which of these five variables, plus any of the others, best discriminate between the groups, we must estimate the discriminant function.

**Estimation of the Discriminant Function** The stepwise procedure begins with all of the variables excluded from the model and then selects the variable that:

- 1 Shows statistically significant differences across the groups (.05 or less required for entry),
- 2 Provides the largest Mahalanobis distance ( $D^2$ ) between the groups.

This process continues to include variables in the discriminant function as long as they provide statistically significant additional discrimination between the groups beyond those differences already accounted for by the variables in the discriminant function. This approach is similar to the stepwise process in multiple regression (see Chapter 5), which adds variables with significant increases in the explained variance of the dependent variable. Also, in cases where two or more variables are entered into the model, the variables already in the model are evaluated for possible removal. A variable may be removed if high multicollinearity exists between it and the other included independent variables such that its significance falls below the significance level for removal (.10).

**STEPWISE ESTIMATION: ADDING THE FIRST VARIABLE  $X_{13}$**  From our review of group differences, we saw that  $X_{13}$  had the largest significant difference between groups and the largest Mahalanobis  $D^2$  (see Table 7.5). Thus,  $X_{13}$  is entered as the first variable in the stepwise procedure (see Table 7.6). Because only one variable enters in the discriminant model at this time, the significance levels and measures of group differences match those of the univariate tests.

**Table 7.5 Group Descriptive Statistics and Tests of Equality for the Estimation Sample in the Two-Group Discriminant Analysis**

Independent Variables	Dependent Variable Group		Test of Equality of Group Means*			Minimum Mahalanobis $D^2$	
	Means: $X_4$ Region		Wilks' Lambda			Minimum $D^2$	Between Groups
	Group 0: USA/ North America	Group 1: Outside North America	F Value	Significance			
$X_6$ Product Quality	(n = 26) 8.527	(n = 34) 7.297	.801	14.387	.000	.976	0 and 1
$X_7$ E-Commerce Activities	3.388	3.626	.966	2.054	.157	.139	0 and 1
$X_8$ Technical Support	5.569	5.050	.973	1.598	.211	.108	0 and 1
$X_9$ Complaint Resolution	5.577	5.253	.986	.849	.361	.058	0 and 1
$X_{10}$ Advertising	3.727	3.979	.987	.775	.382	.053	0 and 1
$X_{11}$ Product Line	6.785	5.274	.695	25.500	.000	1.731	0 and 1
$X_{12}$ Salesforce Image	4.427	5.238	.856	9.733	.003	.661	0 and 1
$X_{13}$ Competitive Pricing	5.600	7.418	.645	31.992	.000	2.171	0 and 1
$X_{14}$ Warranty & Claims	6.050	5.918	.992	.453	.503	.031	0 and 1
$X_{15}$ New Products	4.954	5.276	.990	.600	.442	.041	0 and 1
$X_{16}$ Order & Billing	4.231	4.153	.999	.087	.769	.006	0 and 1
$X_{17}$ Price Flexibility	3.631	4.932	.647	31.699	.000	2.152	0 and 1
$X_{18}$ Delivery Speed	3.873	3.794	.997	.152	.698	.010	0 and 1

\*Wilks' lambda (U statistic) and univariate F ratio with 1 and 58 degrees of freedom.

**Table 7.6** Results from Step 1 of Stepwise Two-Group Discriminant Analysis

Overall Model Fit		Value	F Value	Degrees of Freedom	Significance
Wilks' Lambda		.645	31.992	1, 58	.000
Variable Entered/Removed at Step 1					
Variable Entered	Minimum D <sup>2</sup>	Value	F	Significance	Between Groups
X <sub>13</sub> Competitive Pricing	2.171	31.992	.000		0 and 1

Note: At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

Variables in the Analysis After Step 1					
Variable	Tolerance	F to Remove		D <sup>2</sup>	Between Groups
X <sub>13</sub> Competitive Pricing	1.000	31.992			

Variables Not in the Analysis After Step 1					
Variable	Tolerance	Minimum Tolerance	F to Enter	Minimum D <sup>2</sup>	Between Groups
X <sub>6</sub> Product Quality	.965	.965	4.926	2.699	0 and 1
X <sub>7</sub> E-Commerce Activities	.917	.917	.026	2.174	0 and 1
X <sub>8</sub> Technical Support	.966	.966	.033	2.175	0 and 1
X <sub>9</sub> Complaint Resolution	.844	.844	1.292	2.310	0 and 1
X <sub>10</sub> Advertising	.992	.992	.088	2.181	0 and 1
X <sub>11</sub> Product Line	.849	.849	6.076	2.822	0 and 1
X <sub>12</sub> Salesforce Image	.987	.987	3.949	2.595	0 and 1
X <sub>14</sub> Warranty & Claims	.918	.918	.617	2.237	0 and 1
X <sub>15</sub> New Products	1.000	1.000	.455	2.220	0 and 1
X <sub>16</sub> Order & Billing	.836	.836	3.022	2.495	0 and 1
X <sub>17</sub> Price Flexibility	1.000	1.000	19.863	4.300	0 and 1
X <sub>18</sub> Delivery Speed	.910	.910	1.196	2.300	0 and 1

Significance Testing of Group Differences After Step 1 <sup>a</sup>		
USA/North America		
Outside North America	F	31.992
	Sig.	.000

<sup>a</sup>1, 58 degrees of freedom.

After X<sub>13</sub> enters the model (see Table 7.6), the remaining variables are evaluated on the basis of their incremental discriminating ability (group mean differences after the variance associated with X<sub>13</sub> is removed). Again, variables with significance levels greater than .05 are eliminated from consideration for entry at the next step.

Examining the univariate differences shown in Table 7.5 identifies X<sub>17</sub> (Price Flexibility) as the variable with the second most significant differences. Yet the stepwise process does not use these univariate results when the discriminant function has one or more variables. It calculates the D<sup>2</sup> values and statistical significance tests of group differences after the effect of the variable(s) in the models is removed (in this case only X<sub>13</sub> is in the model).

As shown in the last portion of Table 7.6, three variables (X<sub>6</sub>, X<sub>11</sub>, and X<sub>17</sub>) clearly met the .05 significance level criteria for consideration at the next stage. X<sub>17</sub> remains the next best candidate to enter the model because it has the highest Mahalanobis D<sup>2</sup> (4.300) and the largest F to enter value. However, other variables (e.g., X<sub>11</sub>) have substantial reductions in their significance level and the Mahalanobis D<sup>2</sup> from that shown in Table 7.5 due to the one variable in the model (X<sub>13</sub>).

**STEPWISE ESTIMATION: ADDING THE SECOND VARIABLE  $X_{17}$**  In step 2 (see Table 7.7),  $X_{17}$  enters the model as expected. The overall model is significant ( $F = 31.129$ ) and improves in the discrimination between groups as evidenced by the decrease in Wilks' lambda from .645 to .478. Moreover, the discriminating power of both variables included at this point is also statistically significant ( $F$  values of 20.113 for  $X_{13}$  and 19.863 for  $X_{17}$ ). With both variables statistically significant, the procedure moves to examining the variables not in the equation for potential candidates for inclusion in the discriminant function based on their incremental discrimination between the groups.

$X_{11}$  is the next variable meeting the requirements for inclusion, but its significance level and discriminating ability has been reduced substantially because of multicollinearity with  $X_{13}$  and  $X_{17}$  already in the discriminant function. Most noticeable is the marked increase in the Mahalanobis  $D^2$  from the univariate results in which each variable is considered

**Table 7.7** Results from Step 2 of Stepwise Two-Group Discriminant Analysis

Overall Model Fit		Degrees of Freedom		
	Value	F Value	Freedom	Significance
Wilks' Lambda	.478	31.129	2, 57	.000
Variable Entered/Removed at Step 2				
Variable Entered	Minimum $D^2$	F	Value	Significance
$X_{17}$ Price Flexibility	4.300	31.129	.000	0 and 1

Note: At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

Variables in the Analysis After Step 2				
Variable	Tolerance	F to Remove	$D^2$	Between Groups
$X_{13}$ Competitive Pricing	1.000	20.113	2.152	0 and 1
$X_{17}$ Price Flexibility	1.000	19.863	2.171	0 and 1

Variables Not in the Analysis After Step 2					
Variable	Tolerance	Minimum			
		Tolerance	F to Enter	$D^2$	Between Groups
$X_6$ Product Quality	.884	.884	.681	4.400	0 and 1
$X_7$ E-Commerce Activities	.804	.804	2.486	4.665	0 and 1
$X_8$ Technical Support	.966	.966	.052	4.308	0 and 1
$X_9$ Complaint Resolution	.610	.610	1.479	4.517	0 and 1
$X_{10}$ Advertising	.901	.901	.881	4.429	0 and 1
$X_{11}$ Product Line	.848	.848	5.068	5.045	0 and 1
$X_{12}$ Salesforce Image	.944	.944	.849	4.425	0 and 1
$X_{14}$ Warranty & Claims	.916	.916	.759	4.411	0 and 1
$X_{15}$ New Products	.986	.986	.017	4.302	0 and 1
$X_{16}$ Order & Billing	.625	.625	.245	4.336	0 and 1
$X_{18}$ Delivery Speed	.519	.519	4.261	4.927	0 and 1

Significance Testing of Group Differences After Step 2 <sup>a</sup>		
USA/North America		
Outside North America	F	32.129
	Sig.	.000

<sup>a</sup>2, 57 degrees of freedom.

separately. In the case of  $X_{11}$  the minimum  $D^2$  value increases from 1.731 (see Table 7.5) to 5.045 (see Table 7.7), indicative of a spreading out and separation of the groups by  $X_{13}$  and  $X_{17}$  already in the discriminant function. Note that  $X_{18}$  is almost identical in remaining discrimination power, but  $X_{11}$  will enter in the third step due to its slight advantage.

**STEPWISE ESTIMATION: ADDING A THIRD VARIABLE  $X_{11}$**  Table 7.8 reviews the results of the third step of the stepwise process, where  $X_{11}$  does enter the discriminant function. The overall results are still statistically significant and continue to improve in discrimination, as evidenced by the decrease in the Wilks' lambda value (from .478 to .438). Note however that the decrease was much smaller than found when the second variable ( $X_{17}$ ) was added to the discriminant function. With  $X_{13}$ ,  $X_{17}$ , and  $X_{11}$  all statistically significant, the procedure moves to identifying any remaining candidates for inclusion.

As seen in the last portion of Table 7.8, none of the remaining 10 independent variables pass the entry criterion for statistical significance of .05. After  $X_{11}$  was entered in the equation, both of the remaining variables that had

**Table 7.8** Results from Step 3 of Stepwise Two-Group Discriminant Analysis

Overall Model Fit				
	Value	F Value	Degrees of Freedom	Significance
Wilks' Lambda	.438	23.923	3, 56	.000
Variable Entered/Removed at Step 3				
	Minimum $D^2$	Value	F	Between Groups
$X_{11}$ Product Line	5.045	23.923	.000	0 and 1

Note: At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

Variables in the Analysis After Step 3				
Variable	Tolerance	F to Remove	$D^2$	Between Groups
$X_{13}$ Competitive Pricing	.849	7.258	4.015	0 and 1
$X_{17}$ Price Flexibility	.999	18.416	2.822	0 and 1
$X_{11}$ Product Line	.848	5.068	4.300	0 and 1

Variables Not in the Analysis After Step 3					
Variable	Tolerance	Minimum Tolerance	F to Enter	Minimum $D^2$	Between Groups
$X_6$ Product Quality	.802	.769	.019	5.048	0 and 1
$X_7$ E-Commerce Activities	.801	.791	2.672	5.482	0 and 1
$X_8$ Technical Support	.961	.832	.004	5.046	0 and 1
$X_9$ Complaint Resolution	.233	.233	.719	5.163	0 and 1
$X_{10}$ Advertising	.900	.840	.636	5.149	0 and 1
$X_{12}$ Salesforce Image	.931	.829	1.294	5.257	0 and 1
$X_{14}$ Warranty & Claims	.836	.775	2.318	5.424	0 and 1
$X_{15}$ New Products	.981	.844	.076	5.058	0 and 1
$X_{16}$ Order & Billing	.400	.400	1.025	5.213	0 and 1
$X_{18}$ Delivery Speed	.031	.031	.208	5.079	0 and 1

Significance Testing of Group Differences After Step 3 <sup>a</sup>		
USA/North America		
Outside North America	F	23.923
	Sig.	.000

<sup>a</sup>3, 56 degrees of freedom.

significant univariate differences across the groups ( $X_6$  and  $X_{12}$ ) have relatively little additional discriminatory power and do not meet the entry criterion. Thus, the estimation process stops with three variables ( $X_{13}$ ,  $X_{17}$ , and  $X_{11}$ ) constituting the discriminant function.

**SUMMARY OF THE STEPWISE ESTIMATION PROCESS** Table 7.9 provides the overall stepwise discriminant analysis results after all the significant variables are included in the estimation of the discriminant function. This summary table describes the three variables ( $X_{11}$ ,  $X_{13}$ , and  $X_{17}$ ) that were significant discriminators based on their Wilks' lambda and minimum Mahalanobis  $D^2$  values.

**Table 7.9** Summary Statistics for Two-Group Discriminant Analysis

Overall Model Fit: Canonical Discriminant Functions								
Function	Percent of Variance							
	Eigenvalue	Function %	Cumulative %	Canonical Correlation	Wilks' Lambda	Chi-Square	df	Significance
1	1.282	100	100	.749	.438	46.606	3	.000

Discriminant Function and Classification Function Coefficients					
Independent Variables	Discriminant Functions			Classification Functions	
	Unstandardized	Standardized	Group 0: USA/North America	Group 1: Outside North America	
$X_{11}$ Product Line	-.363	-.417	7.725	6.909	
$X_{13}$ Competitive Pricing	.398	.490	6.456	7.349	
$X_{17}$ Price Flexibility	.749	.664	4.231	5.912	
Constant	-3.752		-52.800	-60.623	

Structure Matrix <sup>a</sup>	
Independent Variables	Function 1
$X_{13}$ Competitive Pricing	.656
$X_{17}$ Price Flexibility	.653
$X_{11}$ Product Line	-.586
$X_7$ E-Commerce Activities*	.429
$X_6$ Product Quality*	-.418
$X_{14}$ Warranty & Claims*	-.329
$X_{10}$ Advertising*	.238
$X_9$ Complaint Resolution*	-.181
$X_{12}$ Salesforce Image*	.164
$X_{16}$ Order & Billing*	-.149
$X_8$ Technical Support*	-.136
$X_{18}$ Delivery Speed*	-.060
$X_{15}$ New Products*	.041

\*This variable not used in the analysis.

Group Means (Centroids) of Discriminant Functions <sup>a</sup>	
$X_4$ Region	Function 1
USA/North America	-1.273
Outside North America	.973

<sup>a</sup>Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions variables ordered by absolute size of correlation within function.

A number of different results are provided addressing both overall model fit and the impact of specific variables.

**Overall Model Fit** The multivariate measures of overall model fit are reported under the heading “Canonical Discriminant Functions.” Note that the discriminant function is highly significant (.000) and displays a canonical correlation of .749. We interpret this correlation by squaring it  $(.749)^2 = .561$ . Thus, 56.1 percent of the variance in the dependent variable ( $X_4$ ) can be accounted for (explained) by this model, which includes only three independent variables.

**Discriminant Function Coefficients** Both the unstandardized and standardized discriminant function coefficients are provided. The standardized function coefficients are one means of interpreting the impact of the independent variables, but are less preferred for interpretation purposes than the discriminant loadings. The unstandardized discriminant coefficients are used to calculate the discriminant Z scores that can be used in classification.

**Discriminant Loadings** The discriminant loadings are reported under the heading “Structure Matrix” and are ordered from highest to lowest by the size of the loading. The loadings are discussed later under the interpretation phase (Stage 5).

**Classification Function Coefficients** The classification function coefficients, also known as Fisher’s linear discriminant functions, are used in classification and are discussed later.

**Group Centroids** These values are the mean of the individual discriminant function scores for each group. Group centroids provide a summary measure of the relative position of each group on the discriminant function(s). In this case, Table 7.9 reveals that the group centroid for the firms in USA/North America (group 0) is  $-1.273$ , whereas the group centroid for the firms outside North America (group 1) is  $.973$ . To show that the overall mean is 0, multiply the number in each group by its centroid and add the result (e.g.,  $26 \times -1.273 + 34 \times .973 = 0.0$ ).

The overall model results are acceptable based on statistical and practical significance. However, before proceeding to an interpretation of the results, the researcher needs to assess classification accuracy and examine the casewise results.

**Assessing Classification Accuracy** With the overall model statistically significant and explaining 56 percent of the variation between the groups (see the preceding discussion and Table 7.9), we move to assessing the predictive accuracy of the discriminant function. In this example, we will illustrate the use of the discriminant scores and the cutting score for classification purposes. In doing so, we must complete three tasks:

- 1 Calculate the cutting score, the criterion against which each observation’s discriminant Z score is judged to determine into which group it should be classified.
- 2 Classify each observation and develop the classification matrices for both the analysis and the holdout samples.
- 3 Assess the levels of predictive accuracy from the classification matrices for both statistical and practical significance.

Although examination of the holdout sample and its predictive accuracy is actually performed in the validation stage, the results are discussed now for ease of comparison between estimation and holdout samples.

**CALCULATING THE CUTTING SCORE** The researcher must first determine how the prior probabilities of classification are to be determined, either based on the actual group sizes (assuming they are representative of the population) or specified by the researcher, most often specified as equal to be conservative in the classification process.

In this analysis sample of 60 observations, we know that the dependent variable consists of two groups, 26 firms located in the United States and 34 firms outside the United States. If we are not sure whether the population proportions are represented by the sample, then we should employ equal probabilities. However, because our sample of firms is randomly drawn, we can be reasonably sure that this sample does reflect the population

proportions. Thus, this discriminant analysis uses the sample proportions to specify the prior probabilities for classification purposes.

Having specified the prior probabilities, the optimum cutting score can be calculated. Because in this situation the groups are assumed representative, the calculation becomes a weighted average of the two group centroids (see Table 7.9 for group centroid values):

$$Z_{CS} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B} = \frac{(26 \times .973) + (34 \times -1.273)}{26 + 34} = -.2997$$

By substitution of the appropriate values in the formula, we can obtain the critical cutting score (assuming equal costs of misclassification) of  $Z_{CS} = -.2997$ .

**CLASSIFYING OBSERVATIONS AND CONSTRUCTING THE CLASSIFICATION MATRICES** Once the cutting score has been calculated, each observation can be classified by comparing its discriminant score to the cutting score.

The procedure for classifying firms with the optimal cutting score is as follows:

- Classify a firm as being in group 0 (United States/North America) if its discriminant score is less than  $-.2997$ .
- Classify a firm as being in group 1 (Outside the United States) if its discriminant score is greater than  $-.2997$ .

Classification matrices for the observations in both the analysis and the holdout samples were calculated, and the results are shown in Table 7.10. Table 7.11 contains the discriminant scores for each observation as well as the actual and predicted group membership values. Note that cases with a discriminant score less than  $-.2997$  have a predicted group membership value of 0, whereas those with a score above  $-.2997$  have a predicted value of 1. The analysis sample, with 86.7 percent prediction accuracy, is slightly higher than the 85.0 percent accuracy of the holdout sample, as anticipated. Moreover, the cross-validated sample achieved a prediction accuracy of 83.3 percent.

**Table 7.10** Classification Results for Two-Group Discriminant Analysis

Classification Results <sup>a, b, c</sup>		Predicted Group Membership			Total
		USA/North America	Outside North America		
Sample	Actual Group				
Estimation Sample	USA/North America	25	1	26	
	Outside North America	96.2%	3.8%		
Cross-validated <sup>d</sup>	USA/North America	7	27	34	
	Outside North America	20.6%	79.4%		
Holdout Sample	USA/North America	24	2	26	
	Outside North America	92.3	7.7		
	USA/North America	8	26	34	
	Outside North America	23.5	76.5		
	USA/North America	9	4	13	
	Outside North America	69.2	30.8		
	USA/North America	2	25	27	
	Outside North America	7.4	92.6		

<sup>a</sup>86.7% of selected original grouped cases (estimation sample) correctly classified.

<sup>b</sup>85.0% of unselected original grouped cases (validation sample) correctly classified.

<sup>c</sup>83.3% of selected cross-validated grouped cases correctly classified.

<sup>d</sup>Cross-validation is done only for those cases in the analysis (estimation sample). In cross-validation, each case is classified by the functions derived from all cases other than that case.

**Table 7.11** Group Predictions for Individual Cases in the Two-Group Discriminant Analysis

Case ID	Actual Group	Discriminant		Case ID	Actual Group	Discriminant	Predicted Group
		Z Score	Predicted Group				
<b>Analysis Sample</b>							
72	0	-2.10690	0	24	1	-.60937	0
14	0	-2.03496	0	53	1	-.45623	0
31	0	-1.98885	0	32	1	-.36094	0
54	0	-1.98885	0	80	1	-.14687	1
27	0	-1.76053	0	38	1	-.04489	1
29	0	-1.76053	0	60	1	-.04447	1
16	0	-1.71859	0	65	1	.09785	1
61	0	-1.71859	0	35	1	.84464	1
79	0	-1.57916	0	1	1	.98896	1
36	0	-1.57108	0	4	1	1.10834	1
98	0	-1.57108	0	68	1	1.12436	1
58	0	-1.48136	0	44	1	1.34768	1
45	0	-1.33840	0	17	1	1.35578	1
2	0	-1.29645	0	67	1	1.35578	1
52	0	-1.29645	0	33	1	1.42147	1
50	0	-1.24651	0	87	1	1.57544	1
47	0	-1.20903	0	6	1	1.58353	1
88	0	-1.10294	0	46	1	1.60411	1
11	0	-.74943	0	12	1	1.75931	1
56	0	-.73978	0	69	1	1.82233	1
95	0	-.73978	0	86	1	1.82233	1
81	0	-.72876	0	10	1	1.85847	1
5	0	-.60845	0	30	1	1.90062	1
37	0	-.60845	0	15	1	1.91724	1
63	0	-.38398	0	92	1	1.97960	1
43	0	.23553	1	7	1	2.09505	1
3	1	-1.65744	0	20	1	2.22839	1
94	1	-1.57916	0	8	1	2.39938	1
49	1	-1.04667	0	100	1	2.62102	1
64	1	-.67406	0	48	1	2.90178	1
<b>Holdout Sample</b>							
23	0	22.38834	0	25	1	1.47048	1
93	0	-2.03496	0	18	1	1.60411	1
59	0	-1.20903	0	73	1	1.61002	1
85	0	-1.10294	0	21	1	1.69348	1
83	0	-1.03619	0	90	1	1.69715	1
91	0	-.89292	0	97	1	1.70398	1
82	0	-.74943	0	40	1	1.75931	1
76	0	-.72876	0	77	1	1.86055	1
96	0	-.57335	0	28	1	1.97494	1
13	0	.13119	1	71	1	2.22839	1
89	0	.51418	1	19	1	2.28652	1
42	0	.63440	1	57	1	2.31456	1
78	0	.63440	1	9	1	2.36823	1
22	1	-2.73303	0	41	1	2.53652	1
74	1	-1.04667	0	26	1	2.59447	1
51	1	.09785	1	70	1	2.59447	1
62	1	.94702	1	66	1	2.90178	1
75	1	.98896	1	34	1	2.97632	1
99	1	1.13130	1	55	1	2.97632	1
84	1	1.30393	1	39	1	3.21116	1

**EVALUATING THE ACHIEVED CLASSIFICATION ACCURACY** Even though all of the measures of classification accuracy are quite high, the evaluation process requires a comparison to the classification accuracy in a series of chance-based measures. These measures reflect the improvement of the discriminant model when compared to classifying individuals without using the discriminant function. Given that the overall sample is 100 observations and group sizes in the holdout/validation sample are less than 20, we will use the overall sample to establish the comparison standards.

**Proportional Chance** The first measure is the **proportional chance criterion**, which assumes that the costs of misclassification are equal (i.e., we want to identify members of each group equally well). The proportional chance criterion is:

$$C_{\text{PRO}} = p^2 + (1 - p)^2$$

where:

$C_{\text{PRO}}$  = proportional chance criterion

$p$  = proportion of firms in group 0

$1 - p$  = proportion of firms in group 1

The group of customers located within the United States (group 0) constitutes 39.0 percent of the analysis sample (39/100), with the second group representing customers located outside the United States (group 1) forming the remaining 61.0 percent (61/100). The calculated proportional chance value is  $.524 (.390^2 + .610^2 = .524)$ .

**Maximum Chance** The **maximum chance criterion** is simply the percentage correctly classified if all observations were placed in the group with the greatest probability of occurrence. It reflects our most conservative standard and assumes no difference in cost of misclassification as well.

Because group 1 (customers outside the United States) is the largest group at 61.0 percent of the sample, we would be correct 61.0 percent of the time if we assigned all observations to this group. If we choose the maximum chance criterion as the standard of evaluation, our model should outperform the 61.0 percent level of classification accuracy to be acceptable.

To attempt to assure practical significance, the achieved classification accuracy must exceed the selected comparison standard by 25 percent. Thus, we must select a comparison standard, calculate the threshold, and compare the achieved hit ratio.

All of the classification accuracy levels (hit ratios) exceed 85 percent, which are substantially higher than the proportional chance criterion of 52.4 percent and the maximum chance criterion of 61.0 percent. All three hit ratios also exceed the suggested threshold of these values (comparison standard plus 25 percent), which in this case are 65.5 percent ( $52.4\% \times 1.25 = 65.5\%$ ) for the proportional chance and 76.3 percent ( $61.0\% \times 1.25 = 76.3\%$ ) for the maximum chance. In all instances (analysis sample, holdout sample, and cross-validation), the levels of classification accuracy are substantially higher than the threshold values, indicating an acceptable level of classification accuracy. Moreover, the hit ratio for individual groups is deemed adequate as well.

**Press's Q** The final measure of classification accuracy is **Press's Q**, which is a statistically-based measure comparing the classification accuracy to a random process.

From the earlier discussion, the calculation for the estimation sample is:

$$\text{Press's } Q_{\text{estimation sample}} = \frac{[60 - (52 \times 2)]^2}{60(2 - 1)} = 45.07$$

and the calculation for the holdout sample is:

$$\text{Press's } Q_{\text{holdout sample}} = \frac{[40 - (34 \times 2)]^2}{40(2 - 1)} = 19.6$$

In both instances, the calculated values exceed the critical value of 6.63. Thus, the classification accuracy for the analysis and, more important, the holdout sample exceeds at a statistically significant level the classification accuracy expected by chance.

**Casewise Diagnostics** In addition to examining the overall results, we can examine the individual observations for their predictive accuracy and identify specifically the misclassified cases. In this manner, we can find the specific cases misclassified for each group on both analysis and holdout samples as well as perform additional analysis profiling for the misclassified cases.

Table 7.11 contains the group predictions for the analysis and holdout samples and enables us to identify the specific cases for each type of misclassification tabulated in the classification matrices (see Table 7.10). For the analysis sample, the seven customers located outside the United States misclassified into the group of customers in the United States can be identified as cases 3, 94, 49, 64, 24, 53, and 32. Likewise, the single customer located in the United States but misclassified is identified as case 43. A similar examination can be performed for the holdout sample.

Once the misclassified cases are identified, further analysis can be performed to understand the reasons for their misclassification. In Table 7.12, the misclassified cases are combined from the analysis and holdout samples and

**Table 7.12** Profiling Correctly Classified and Misclassified Observations in the Two-Group Discriminant Analysis

Dependent Variable: <i>X</i> <sub>4</sub> Region	Group/Profile Variables	Mean Scores			t Test Statistical Significance
		Correctly Classified (n = 34)	Misclassified (n = 5)	Difference	
USA/North America					
	<i>X</i> <sub>6</sub> Product Quality	8.612	9.340	-.728	.000 <sup>b</sup>
	<i>X</i> <sub>7</sub> E-Commerce Activities	3.382	4.380	-.998	.068 <sup>b</sup>
	<i>X</i> <sub>8</sub> Technical Support	5.759	5.280	.479	.487
	<i>X</i> <sub>9</sub> Complaint Resolution	5.356	6.140	-.784	.149
	<i>X</i> <sub>10</sub> Advertising	3.597	4.700	-1.103	.022
	<i>X</i> <sub>11</sub> Product Line <sup>a</sup>	6.726	6.540	.186	.345 <sup>b</sup>
	<i>X</i> <sub>12</sub> Salesforce Image	4.459	5.460	-1.001	.018
	<i>X</i> <sub>13</sub> Competitive Pricing <sup>a</sup>	5.609	8.060	-2.451	.000
	<i>X</i> <sub>14</sub> Warranty & Claims	6.215	6.060	.155	.677
	<i>X</i> <sub>15</sub> New Products	5.024	4.420	.604	.391
	<i>X</i> <sub>16</sub> Order & Billing	4.188	4.540	-.352	.329
	<i>X</i> <sub>17</sub> Price Flexibility <sup>a</sup>	3.568	4.480	-.912	.000 <sup>b</sup>
	<i>X</i> <sub>18</sub> Delivery Speed	3.826	4.160	-.334	.027 <sup>b</sup>
Outside North America		(n = 52)	(n = 9)		
	<i>X</i> <sub>6</sub> Product Quality	6.906	9.156	-2.250	.000
	<i>X</i> <sub>7</sub> E-Commerce Activities	3.860	3.289	.571	.159 <sup>b</sup>
	<i>X</i> <sub>8</sub> Technical Support	5.085	5.544	-.460	.423
	<i>X</i> <sub>9</sub> Complaint Resolution	5.365	5.822	-.457	.322
	<i>X</i> <sub>10</sub> Advertising	4.229	3.922	.307	.470
	<i>X</i> <sub>11</sub> Product Line <sup>a</sup>	4.954	6.833	-1.879	.000
	<i>X</i> <sub>12</sub> Salesforce Image	5.465	5.467	-.002	.998
	<i>X</i> <sub>13</sub> Competitive Pricing <sup>a</sup>	7.960	5.833	2.126	.000
	<i>X</i> <sub>14</sub> Warranty & Claims	5.867	6.400	-.533	.007 <sup>b</sup>
	<i>X</i> <sub>15</sub> New Products	5.194	5.778	-.584	.291
	<i>X</i> <sub>16</sub> Order & Billing	4.267	4.533	-.266	.481
	<i>X</i> <sub>17</sub> Price Flexibility <sup>a</sup>	5.458	3.722	1.735	.000
	<i>X</i> <sub>18</sub> Delivery Speed	3.881	3.989	-.108	.714

Note: Cases from both analysis and validation samples included for total sample of 100.

<sup>a</sup>Variables included in the discriminant function.

<sup>b</sup>t test performed with separate variance estimates rather than pooled estimate because the Levene test detected significant differences in the variations between the two groups.

then compared to the correctly classified cases. The attempt is to identify specific differences on the independent variables that might identify either new variables to be added or common characteristics that should be considered.

The five cases (both analysis and holdout samples) misclassified among the United States customers (group 0) show significant differences on two of the three independent variables in the discriminant function ( $X_{13}$  and  $X_{17}$ ) as well as one variable not in the discriminant function ( $X_6$ ). For that variable not in the discriminant function, the profile of the misclassified cases is not similar to their correct group; thus, it is of no help in classification. Likewise, the nine misclassified cases of group 1 (outside the United States) show four significant differences ( $X_6$ ,  $X_{11}$ ,  $X_{13}$ , and  $X_{17}$ ), but only  $X_6$  is not in the discriminant function. We can see that here  $X_6$  works against classification accuracy because the misclassified cases are more similar to the incorrect group rather than the correct group.

The findings suggest that the misclassified cases may represent a distinct third group, because they share quite similar profiles across these variables more so than they do with the two existing groups. Management may analyze this group on additional variables or assess whether a geographic pattern among these misclassified cases justifies a new group.

Researchers should examine the patterns in both groups with the objective of understanding the characteristics common to them in an attempt at defining the reasons for misclassification.

## STAGE 5: INTERPRETATION OF THE RESULTS

After estimating the discriminant function, the next task is interpretation. This stage involves examining the function to determine the relative importance of each independent variable in discriminating between the groups, interpreting the discriminant function based on the discriminant loadings, and then profiling each group on the pattern of mean values for variables identified as important discriminating variables.

**Identifying Important Discriminating Variables** As discussed earlier, discriminant loadings are considered the more appropriate measure of discriminatory power, but we will also consider the discriminant weights for comparative purposes. The discriminant weights, either in unstandardized or standardized form, represent each variable's contribution to the discriminant function. However, as we will discuss, multicollinearity among the independent variables can impact the interpretation using only the weights.

Discriminant loadings are calculated for every independent variable, even for those not included in the discriminant function. Thus, discriminant weights represent the unique impact of each independent variable and are not restricted to only the shared impact due to multicollinearity. Moreover, because they are relatively unaffected by multicollinearity, they more accurately represent each variable's association with the discriminant score.

Table 7.13 contains the entire set of interpretive measures, including unstandardized and standardized discriminant weights, loadings for the discriminant function, Wilks' lambda, and the univariate  $F$  ratio. The original 13 independent variables were screened by the stepwise procedure, and three ( $X_{11}$ ,  $X_{13}$ , and  $X_{17}$ ) are significant enough to be included in the function. For interpretation purposes, we rank the independent variables in terms of their loadings and univariate  $F$  values—both indicators of each variable's discriminating power. Signs of the weights or loadings do not affect the rankings; they simply indicate a positive or negative relationship with the dependent variable.

**ANALYZING WILKS' LAMBDA AND UNIVARIATE  $F$**  The Wilks' lambda and univariate  $F$  values represent the separate or univariate effects of each variable, not considering multicollinearity among the independent variables. Analogous to the bivariate correlations of multiple regression, they indicate each variable's ability to discriminate among the groups, but only separately. To interpret any combination of two or more independent variables requires analysis of the discriminant weights or discriminant loadings as described in the following sections.

Table 7.13 shows that the variables ( $X_{11}$ ,  $X_{13}$ , and  $X_{17}$ ) with the three highest  $F$  values (and lowest Wilks' lambda values) were also the variables entered into the discriminant function. Two other variables ( $X_6$  and  $X_{12}$ ) also had significant discriminating effects (i.e., significant group differences), but were not included by the stepwise process in

**Table 7.13** Summary of Interpretive Measures for Two-Group Discriminant Analysis

Independent Variables	Discriminant Coefficients		Discriminant Loadings		Wilks' Lambda	Univariate F Ratio		
	Unstandardized	Standardized	Loading	Rank		F Value	Sig.	Rank
X <sub>6</sub> Product Quality	NI	NI	-.418	5	.801	14.387	.000	4
X <sub>7</sub> E-Commerce Activities	NI	NI	.429	4	.966	2.054	.157	6
X <sub>8</sub> Technical Support	NI	NI	-.136	11	.973	1.598	.211	7
X <sub>9</sub> Complaint Resolution	NI	NI	-.181	8	.986	.849	.361	8
X <sub>10</sub> Advertising	NI	NI	.238	7	.987	.775	.382	9
X <sub>11</sub> Product Line	-.363	-.417	-.586	3	.695	25.500	.000	3
X <sub>12</sub> Salesforce Image	NI	NI	.164	9	.856	9.733	.003	5
X <sub>13</sub> Competitive Pricing	-.398	.490	.656	1	.645	31.992	.000	1
X <sub>14</sub> Warranty & Claims	NI	NI	-.329	6	.992	.453	.503	11
X <sub>15</sub> New Products	NI	NI	.041	13	.990	.600	.442	10
X <sub>16</sub> Order & Billing	NI	NI	-.149	10	.999	.087	.769	13
X <sub>17</sub> Price Flexibility	.749	.664	.653	2	.647	31.699	.000	2
X <sub>18</sub> Delivery Speed	NI	NI	-.060	12	.997	.152	.698	12

NI = Not included in estimated discriminant function.

the discriminant function. This was due to the multicollinearity between these two variables and the three variables included in the discriminant function. These two variables added no incremental discriminating power beyond the variables already in the discriminant function. Interested readers are referred to a more complete discussion of multicollinearity and the stepwise estimation process in Chapter 5. All of the remaining variables had nonsignificant *F* values and correspondingly high Wilks' lambda values.

**ANALYZING THE DISCRIMINANT WEIGHTS** The discriminant weights are available in unstandardized and standardized forms. The unstandardized weights (plus the constant) are used to calculate the discriminant score, but can be affected by the scale of the independent variable (just like multiple regression weights). Thus, the standardized weights more truly reflect the impact of each variable on the discriminant function and are more appropriate than unstandardized weights when used for interpretation purposes. If simultaneous estimation is used, multicollinearity among any of the independent variables will impact the estimated weights. However, the impact of multicollinearity can be even greater for the stepwise procedure, because multicollinearity affects not only the weights but may also prevent a variable from even entering the equation.

Table 7.13 provides the standardized weights (coefficients) for the three variables included in the discriminant function. The impact of multicollinearity on the weights can be seen in examining  $X_{13}$  and  $X_{17}$ . These two variables have essentially equivalent discriminating power when viewed on the Wilks' lambda and univariate *F* tests. Their discriminant weights, however, reflect a markedly greater impact for  $X_{17}$  than  $X_{13}$  which, based on the weights, is now more comparable to  $X_{11}$ . This change in relative importance is due to the collinearity between  $X_{13}$  and  $X_{11}$ , which reduces the unique effect of  $X_{13}$ , thus reducing the discriminant weight as well.

**Interpreting The Discriminant Function Based On Discriminant Loadings** The discriminant loadings, in contrast to the discriminant weights, are less affected by multicollinearity and thus are more useful for interpretative purposes. Also, because loadings are calculated for all variables, they provide an interpretive measure even for variables not included in the discriminant function. An earlier rule of thumb indicated loadings above  $\pm .40$  should be used to identify substantive discriminating variables.

The loadings of the three variables entered in the discriminant function (see Table 7.13) are the three highest and all exceed  $\pm .40$ , thus warranting inclusion for interpretation purposes. Two additional variables ( $X_6$  and  $X_7$ ),

however, also have loadings above the  $\pm .40$  threshold. The inclusion of  $X_6$  is not unexpected, because it was the fourth variable with significant univariate discriminating effect, but was not included in the discriminant function due to multicollinearity (as was shown in Chapter 3, Exploratory Factor Analysis, where  $X_6$  and  $X_{13}$  formed a factor).  $X_7$ , however, presents another situation; it did not have a significant univariate effect. The combination of the three variables in the discriminant function created an effect that is associated with  $X_7$ , but  $X_7$  does not add any additional discriminating power. In this regard,  $X_7$  can be used to describe the discriminant function for profiling purposes even though it did not enter into the estimation of the discriminant function.

Interpreting the discriminant function and its discrimination between these two groups requires that the researcher consider all five of these variables. To the extent that they characterize or describe the discriminant function, they all represent some component of the function.

The three strongest effects in the discriminant function, which are all generally comparable based on the loading values, are  $X_{13}$  (Competitive Pricing),  $X_{17}$  (Price Flexibility), and  $X_{11}$  (Product Line).  $X_7$  (E-Commerce Activities) and the effect of  $X_6$  (Product Quality) can be added when interpreting the discriminant function. Obviously several different factors are being combined to differentiate between the groups, thus requiring more profiling of the groups to understand the differences.

With the discriminating variables identified and the discriminant function described in terms of those variables with sufficiently high loadings, the researcher then proceeds to profile each group on these variables to understand the differences between them.

**Profiling the Discriminating Variables** The researcher is interested in interpretations of the individual variables that have statistical and practical significance. Such interpretations are accomplished by first identifying the variables with substantive discriminatory power (see the preceding discussions) and then understanding what the differing group means on each variable indicated.

As described in Chapter 1, higher scores on the independent variables indicate more favorable perceptions of HBAT on that attribute (except for  $X_{13}$ , where lower scores are more preferable). Referring back to Table 7.5, we see varied profiles between the two groups on these five variables.

- Group 0 (customers in the USA/North America) has higher perceptions on three variables:  $X_6$  (Product Quality),  $X_{13}$  (Competitive Pricing), and  $X_{11}$  (Product Line).
- Group 1 (customers outside North America) has higher perceptions on the remaining two variables:  $X_7$  (E-Commerce Activities) and  $X_{17}$  (Price Flexibility).

In looking at these two profiles, we can see that the USA/North America customers have much better perceptions of the HBAT products, whereas those customers outside North America feel better about pricing issues and e-commerce. Note that  $X_6$  and  $X_{13}$ , both of which have higher perceptions among the USA/North America customers, form the *Product Value* factor developed in Chapter 3. Management should use these results to develop strategies that accentuate these strengths and develop additional strengths to complement them.

The mean profiles also illustrate the interpretation of signs (positive or negative) on the discriminant weights and loadings. The signs reflect the relative mean profile of the two groups. The positive signs, in this example, are associated with variables that have higher scores for group 1. The negative weights and loadings are for those variables with the opposite pattern (i.e., higher values in group 0). Thus, the signs indicate the pattern between the groups.

## STAGE 6: VALIDATION OF THE RESULTS

The final stage addresses the internal and external validity of the discriminant function. The primary means of validation is through the use of the holdout sample and the assessment of its predictive accuracy. In this manner, validity is established if the discriminant function performs at an acceptable level in classifying observations that

were not used in the estimation process. If the holdout sample is formed from the original sample, then this approach establishes internal validity and an initial indication of external validity. If another separate sample, perhaps from another population or segment of the population, forms the holdout sample, then this addresses more fully the external validity of the discriminant results.

In our example, the holdout sample comes from the original sample. As discussed earlier, the classification accuracy (hit ratios) for both the holdout sample and the cross-validated sample was markedly above the thresholds on all of the measures of predictive accuracy. As such, the analysis does establish internal validity. For purposes of external validity, additional samples should be drawn from relevant populations and the classification accuracy assessed in as many situations as possible.

The researcher is encouraged to extend the validation process through expanded profiling of the groups and the possible use of additional samples to establish external validity. Additional insights from the analysis of misclassified cases may suggest additional variables that could improve even more the discriminant model.

### A MANAGERIAL OVERVIEW

The discriminant analysis of HBAT customers based on geographic location (located within North America or outside) identified a set of perceptual differences that can provide a rather succinct and powerful distinction between the two groups. Several key findings include the following:

- Differences are found in a subset of only five perceptions, allowing for a focus on key variables and not having to deal with the entire set. The variables identified as discriminating between the groups (listed in order of importance) are  $X_{13}$  (Competitive Pricing),  $X_{17}$  (Price Flexibility),  $X_{11}$  (Product Line),  $X_7$  (E-Commerce Activities), and  $X_6$  (Product Quality).
- Results also indicate that firms located in the United States have better perceptions of HBAT than their international counterparts in terms of product value and the product line, whereas the non-North American customers have a more favorable perception of price flexibility and e-commerce activities. These perceptions may result from a better match between USA/North American buyers, whereas the international customers find the pricing policies conducive to their needs.
- The results, which are highly significant, provide the researcher the ability to correctly identify the purchasing strategy used based on these perceptions 85 percent of the time. Their high degree of consistency provides confidence in the development of strategies based on these results.
- Analysis of the misclassified firms revealed a small number of firms that seemed out of place. Identifying these firms may identify associations not addressed by geographic location (e.g., markets served rather than just physical location) or other firm or market characteristics that are associated with geographic location.

Thus, knowing a firm's geographic location provides key insights into their perceptions of HBAT and, more important, how the two groups of customers differ so that management can employ a strategy to accentuate the positive perceptions in their dealings with these customers and further solidify their position.

### A Three-Group Illustrative Example

To illustrate the application of a three-group discriminant analysis, we once again use the HBAT database. In the previous example, we were concerned with discriminating between only two groups, so we were able to develop a single discriminant function and a cutting score to divide the two groups. In the three-group example, it is necessary to develop two separate discriminant functions to distinguish among three groups. The first function separates one group from the other two, and the second separates the remaining two groups. As with the prior example, the six stages of the model-building process are discussed.

## STAGE 1: OBJECTIVES OF DISCRIMINANT ANALYSIS

HBAT's objective in this research is to determine the relationship between the firms' perceptions of HBAT and the length of time a firm has been a customer with HBAT. One of the emerging paradigms in marketing is the concept of a customer relationship, based on the establishment of a mutual partnership between firms over repeated transactions. The process of developing a relationship entails the formation of shared goals and values, which should coincide with improved perceptions of HBAT. Thus, the successful formation of a relationship should be seen by improved HBAT perceptions over time. In this analysis, firms are grouped on their tenure as HBAT customers. Hopefully, if HBAT has been successful in establishing relationships with its customers, then perceptions of HBAT will improve with tenure as an HBAT customer.

## STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS

To test this relationship, a discriminant analysis is performed to establish whether differences in perceptions exist between customer groups based on length of customer relationship. If so, HBAT is then interested in seeing whether the distinguishing profiles support the proposition that HBAT has been successful in improving perceptions among established customers, a necessary step in the formation of customer relationships.

**Selection of Dependent and Independent Variables** In addition to the nonmetric (categorical) dependent variables defining the groups of interest, discriminant analysis also requires a set of metric independent variables that are assumed to provide the basis for discrimination or differentiation between the groups.

A three-group discriminant analysis is performed using  $X_1$  (Customer Type) as the dependent variable and the perceptions of HBAT by these firms ( $X_6$  to  $X_{18}$ ) as the independent variables. Note that  $X_1$  differs from the dependent variable in the two-group example in that it has three categories in which to classify a firm's length of time being an HBAT customer (1 = less than 1 year, 2 = 1 to 5 years, and 3 = more than 5 years).

**Sample Size and Division of The Sample** Issues regarding sample size are particularly important with discriminant analysis due to the focus on not only overall sample size, but also on sample size per group. Coupled with the need for a division of the sample to provide for a validation sample, the researcher must carefully consider the impact of sample division on both samples in terms of the overall sample size and the size of each of the groups.

The HBAT database has a sample size of 100, which again will be split into analysis and holdout samples of 60 and 40 cases, respectively. In the analysis sample, the ratio of cases to independent variables is almost 5:1, the recommended lower threshold. More importantly, in the analysis sample, only one group, with 13 observations, falls below the recommended level of 20 cases per group. Although the group size would exceed 20 if the entire sample were used in the analysis phase, the need for validation dictated the creation of the holdout sample. The three groups are of relatively equal sizes (22, 13, and 25), thus avoiding any need to equalize the group sizes. The analysis proceeds with attention paid to the classification and interpretation of this small group of 13 observations.

## STAGE 3: ASSUMPTIONS OF DISCRIMINANT ANALYSIS

As was the case in the two-group example, the assumptions of normality, linearity, and collinearity of the independent variables have already been discussed at length in Chapter 2. The analyses performed in Chapter 2 indicated that the independent variables met these assumptions at adequate levels to allow for the analysis to continue without additional remedies. The remaining assumption, the equality of the variance/covariance or dispersion matrices, is also addressed in Chapter 2.

Box's M test assesses the similarity of the dispersion matrices of the independent variables among the three groups (categories). The test statistic indicated differences at the .09 significance level. In this case, the differences between

groups are nonsignificant and no remedial action is needed. Moreover, no impacts are expected on the estimation or classification processes.

## STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT

As in the previous example, we begin our analysis by reviewing the group means and standard deviations to see whether the groups are significantly different on any single variable. With those differences in mind, we then employ a stepwise estimation procedure to derive the discriminant functions and complete the process by assessing classification accuracy both overall and with casewise diagnostics.

**Assessing Group Differences** Identifying the most discriminating variables with three or more groups is more problematic than in the two-group situation. For three or more groups, the typical measures of significance for differences across groups (i.e., Wilks' lambda and the  $F$  test) only assess the overall differences and do not guarantee that each group is significant from the others. Thus, when examining variables for their overall differences between the groups, be sure to also address individual group differences.

Table 7.14 provides the group means, Wilks' lambda, univariate  $F$  ratios (simple ANOVAs), and minimum Mahalanobis  $D^2$  for each independent variable. Review of these measures of discrimination reveals the following:

- On a univariate basis, about one-half (7 of 13) of the variables display significant differences between the group means. The variables with significant differences include  $X_6$ ,  $X_9$ ,  $X_{11}$ ,  $X_{13}$ ,  $X_{16}$ ,  $X_{17}$ , and  $X_{18}$ .
- Although greater statistical significance corresponds to higher overall discrimination (i.e., the most significant variables have the lowest Wilks' lambda values), it does not always correspond to the greatest discrimination between all the groups.

**Table 7.14** Group Descriptive Statistics and Tests of Equality for the Estimation Sample in the Three-Group Discriminant Analysis

Independent Variables	Dependent Variable Group Means: $X_1$ Customer Type			Test of Equality of Group Means <sup>a</sup>			Minimum Mahalanobis $D^2$	
	Group 1: Less than 1 Year (n = 22)		Group 2: 1 to 5 Years (n = 13)		Group 3: More Than 5 Years (n = 25)		Wilks' Lambda	$F$ Value
$X_8$ Technical Support	4.959	5.615	5.376	.973	.782	.462	.023	2 and 3
$X_6$ Product Quality	7.118	6.785	9.000	.469	32.311	.000	.121	1 and 2
$X_7$ E-Commerce Activities	3.514	3.754	3.412	.959	1.221	.303	.025	1 and 3
$X_9$ Complaint Resolution	4.064	5.900	6.300	.414	40.292	.000	.205	2 and 3
$X_{10}$ Advertising	3.745	4.277	3.768	.961	1.147	.325	.000	1 and 3
$X_{11}$ Product Line	4.855	5.577	7.056	.467	32.583	.000	.579	1 and 2
$X_{12}$ Salesforce Image	4.673	5.346	4.836	.943	1.708	.190	.024	1 and 3
$X_{13}$ Competitive Pricing	7.345	7.123	5.744	.751	9.432	.000	.027	1 and 2
$X_{14}$ Warranty & Claims	5.705	6.246	6.072	.916	2.619	.082	.057	2 and 3
$X_{15}$ New Products	4.986	5.092	5.292	.992	.216	.807	.004	1 and 2
$X_{16}$ Order & Billing	3.291	4.715	4.700	.532	25.048	.000	.000	2 and 3
$X_{17}$ Price Flexibility	4.018	5.508	4.084	.694	12.551	.000	.005	1 and 3
$X_{18}$ Delivery Speed	3.059	4.246	4.288	.415	40.176	.000	.007	2 and 3

<sup>a</sup>Wilks' lambda ( $\lambda$  statistic) and univariate  $F$  ratio with 2 and 57 degrees of freedom.

- Visual inspection of the group means reveal that four of the variables with significant differences ( $X_{13}$ ,  $X_{16}$ ,  $X_{17}$ , and  $X_{18}$ ) only differentiate one group versus the other two groups [e.g.,  $X_{18}$  has significant differences only in the means between group 1 (3.059) versus groups 2 (4.246) and 3 (4.288)]. These variables play a limited role in discriminant analysis because they provide discrimination between only a subset of groups.
- Three variables ( $X_6$ ,  $X_9$ , and  $X_{11}$ ) provide some discrimination, in varying degrees, between all three groups simultaneously. One or more of these variables may be used in combination with the four preceding variables to create a variate with maximum discrimination.
- The Mahalanobis  $D^2$  value provides a measure of the degree of discrimination between groups. For each variable, the minimum Mahalanobis  $D^2$  is the distance between the two closest groups. For example,  $X_{11}$  has the highest  $D^2$  value, and it is the variable with the greatest differences between all three groups. Likewise,  $X_{18}$ , a variable with little differences between two of the groups, has a small  $D^2$  value. With three or more groups, the minimum Mahalanobis  $D^2$  is important in identifying the variable that provides the greatest difference between the two most similar groups.

All of these measures combine to help identify the sets of variables that form the discriminant functions as described in the next section. When more than one function is created, each function provides discrimination between sets of groups. In the simple example from the beginning of this chapter, one variable discriminated between groups 1 versus 2 and 3, whereas the other discriminated between groups 2 versus 3 and 1. It is one of the primary benefits arising from the use of discriminant analysis.

**Estimation of the Discriminant Function** The stepwise procedure is performed in the same manner as in the two-group example, with all of the variables initially excluded from the model. As noted earlier, the Mahalanobis distance should be used with the stepwise procedure in order to select the variable that has a statistically significant difference across the groups while maximizing the Mahalanobis distance ( $D^2$ ) between the two closest groups. In this manner, statistically significant variables are selected that maximize the discrimination between the most similar groups at each stage.

This process continues as long as additional variables provide statistically significant discrimination beyond those differences already accounted for by the variables in the discriminant function. A variable may be removed if high multicollinearity with independent variables in the discriminant function causes its significance to fall below the significance level for removal (.10).

**STEPWISE ESTIMATION: ADDING THE FIRST VARIABLE,  $X_{11}$**  The data in Table 7.14 show that the first variable to enter the stepwise model using the Mahalanobis distance is  $X_{11}$  (Product Line) because it meets the criteria for statistically significant differences across the groups and has the largest minimum  $D^2$  value (meaning it has the greatest separation between the most similar groups).

The results of adding  $X_{11}$  as the first variable in the stepwise process are shown in Table 7.15. The overall model fit is significant and each of the groups are significantly different, although groups 1 (less than 1 year) and 2 (1 to 5 years) have the smallest difference between them (see bottom section detailing group differences).

Of the variables not in the equation, only  $X_6$  (Product Quality) meets the significance level necessary for consideration. If added, the minimum  $D^2$  will now be between groups 1 and 2.

With the smallest difference between groups 1 and 2, the discriminant procedure will now select a variable that maximizes that difference while at least maintaining the other differences. If we refer back to Table 7.14, we see that four variables ( $X_9$ ,  $X_{16}$ ,  $X_{17}$ , and  $X_{18}$ ) all had significant differences, with substantial differences between groups 1 and 2. Looking in Table 7.15, we see that these four variables have the highest minimum  $D^2$  value, and in each case it is for the difference between groups 2 and 3 (meaning that groups 1 and 2 are not the most similar after adding that variable). Thus, adding any one of these variables would most affect the differences between groups 1 and 2, the pair that was most similar after  $X_{11}$  was added in the first step. The procedure will select  $X_{17}$  because it will create the greatest distance between groups 2 and 3.

**Table 7.15** Results from Step 1 of Stepwise Three-Group Discriminant Analysis

Overall Model Fit		Value	F Value	Degrees of Freedom	Significance
Wilks' Lambda		.467	32.583	2,57	.000
Variable Entered/Removed at Step 1					
			<i>F</i>		
Variable Entered		Minimum <i>D</i> <sup>2</sup>	Value	Significance	Between Groups
X <sub>11</sub> Product Line		.579	4.729	.000	Less than 1 year and 1 to 5 years

Note: At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

Variables in the Analysis After Step 1				
Variable	Tolerance	F to Remove	<i>D</i> <sup>2</sup>	Between Groups
X <sub>11</sub> Product Line	1.000	32.583	NA	NA

NA = Not applicable.

Variable	Tolerance	Minimum	<i>F</i>	Minimum	Between Groups
		Tolerance		to Enter	
X <sub>6</sub> Product Quality	1.000	1.000	17.426	.698	Less than 1 year and 1 to 5 years
X <sub>7</sub> E-Commerce Activities	.950	.950	1.171	.892	Less than 1 year and 1 to 5 years
X <sub>8</sub> Technical Support	.959	.959	.733	.649	Less than 1 year and 1 to 5 years
X <sub>9</sub> Complaint Resolution	.847	.847	15.446	2.455	1 to 5 years and more than 5 years
X <sub>10</sub> Advertising	.998	.998	1.113	.850	Less than 1 year and 1 to 5 years
X <sub>12</sub> Salesforce Image	.932	.932	3.076	1.328	Less than 1 year and 1 to 5 years
X <sub>13</sub> Competitive Pricing	.849	.849	.647	.599	Less than 1 year and 1 to 5 years
X <sub>14</sub> Warranty & Claims	.882	.882	2.299	.839	Less than 1 year and 1 to 5 years
X <sub>15</sub> New Products	.993	.993	.415	.596	Less than 1 year and 1 to 5 years
X <sub>16</sub> Order & Billing	.943	.943	12.176	2.590	1 to 5 years and more than 5 years
X <sub>17</sub> Price Flexibility	.807	.807	17.300	3.322	1 to 5 years and more than 5 years
X <sub>18</sub> Delivery Speed	.773	.773	19.020	2.988	1 to 5 years and more than 5 years

Significance Testing of Group Differences After Step 1 <sup>a</sup>				
X <sub>1</sub> – Customer Type	Less than 1 Year		1 to 5 Years	
	<i>F</i>	4.729		
1 to 5 years	Sig.	.034		
	<i>F</i>	62.893	20.749	
Over 5 years	Sig.	.000	.000	

<sup>a</sup>1, 57 degrees of freedom.

**STEPWISE ESTIMATION: ADDING THE SECOND VARIABLE, X<sub>17</sub>** Table 7.16 details the second step of the stepwise procedure: adding X<sub>17</sub> (Price Flexibility) to the discriminant function. The discrimination between groups increased, as reflected in a lower Wilks' lambda value and increase in the minimum *D*<sup>2</sup> (.467 to .288). The group differences, overall and individual, are still statistically significant. The addition of X<sub>17</sub> increased the differences between groups 1 and 2 substantially, such that now the two most similar groups are 2 and 3.

**STEPWISE ESTIMATION: ADDING THE THIRD AND FOURTH VARIABLES, X<sub>6</sub> AND X<sub>18</sub>** As noted previously, X<sub>6</sub> becomes the third variable added to the discriminant function. After X<sub>6</sub> is added, only X<sub>18</sub> exhibits a statistical significance across the groups (*Note*: The details of adding X<sub>6</sub> in step 3 are not shown for space considerations).

**Table 7.16** Results from Step 2 of Stepwise Three-Group Discriminant Analysis

Overall Model Fit				
	Value	F Value	Degrees of Freedom	Significance
Wilks' Lambda	.288	24.139	4, 112	.000
Variable Entered/Removed at Step 2				
		<i>F</i>		
Variable Entered	Minimum <i>D</i> <sup>2</sup>	Value	Significance	Between Groups
$X_{17}$ Price Flexibility	3.322	13.958	.000	1 to 5 years and more than 5 years

Note: At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

Variables in the Analysis After Step 2				
Variable	Tolerance	F to Remove	<i>D</i> <sup>2</sup>	Between Groups
$X_{11}$ Product Line	.807	39.405	.005	Less than 1 year and more than 5 years
$X_{17}$ Price Flexibility	.807	17.300	.579	Less than 1 year and 1 to 5 years

Variables Not in the Analysis After Step 2					
Variable	Tolerance	Minimum Tolerance	F to Enter	Minimum <i>D</i> <sup>2</sup>	Between Groups
$X_6$ Product Quality	.730	.589	24.444	6.071	Less than 1 year and 1 to 5 years
$X_7$ E-Commerce Activities	.880	.747	.014	3.327	Less than 1 year and 1 to 5 years
$X_8$ Technical Support	.949	.791	1.023	3.655	Less than 1 year and 1 to 5 years
$X_9$ Complaint Resolution	.520	.475	3.932	3.608	Less than 1 year and 1 to 5 years
$X_{10}$ Advertising	.935	.756	.102	3.348	Less than 1 year and 1 to 5 years
$X_{12}$ Salesforce Image	.884	.765	.662	3.342	Less than 1 year and 1 to 5 years
$X_{13}$ Competitive Pricing	.794	.750	.989	3.372	Less than 1 year and 1 to 5 years
$X_{14}$ Warranty & Claims	.868	.750	2.733	4.225	Less than 1 year and 1 to 5 years
$X_{15}$ New Products	.963	.782	.504	3.505	Less than 1 year and 1 to 5 years
$X_{16}$ Order & Billing	.754	.645	2.456	3.323	Less than 1 year and 1 to 5 years
$X_{18}$ Delivery Speed	.067	.067	3.255	3.598	Less than 1 year and 1 to 5 years

Significance Testing of Group Differences After Step 2 <sup>a</sup>				
$X_1$ Customer Type		Less than 1 Year	1 to 5 Years	
1 to 5 years	<i>F</i>	21.054		
	Sig.	.000		
More than 5 years	<i>F</i>	39.360	13.958	
	Sig.	.000	.000	

<sup>a</sup>2, 56 degrees of freedom.

The final variable added in step 4 is  $X_{18}$  (see Table 7.17), with the discriminant function now including four variables ( $X_{11}$ ,  $X_{17}$ ,  $X_6$ , and  $X_{18}$ ). The overall model is significant, with the Wilks' lambda declining to .127. Moreover, significant differences exist between all of the individual groups.

With these four variables in the discriminant function, no other variable exhibits the statistical significance necessary for inclusion and the stepwise procedure is completed in terms of adding variables. The procedure, however, also includes a check on the significance of each variable in order to be retained in the discriminant function. In this case, the "F to Remove" for both  $X_{11}$  and  $X_{17}$  is nonsignificant (.918 and 1.735, respectively), indicating that one or both are candidates for removal from the discriminant function.

**STEPWISE ESTIMATION: REMOVAL OF  $X_{17}$  AND  $X_{11}$**  When  $X_{18}$  is added to the model in the fourth step (see the preceding discussion),  $X_{11}$  had the lowest "F to Remove" value (.918), causing the stepwise procedure to eliminate that variable

**Table 7.17** Results from Step 4 of Stepwise Three-Group Discriminant Analysis

Overall Model Fit				
	Value	F Value	Degrees of Freedom	Significance
Wilks' Lambda	.127	24.340	8, 108	.000
Variable Entered/Removed at Step 4				
Variable Entered	Minimum D <sup>2</sup>	Value	F Significance	Between Groups
X <sub>18</sub> Delivery Speed	6.920	13.393	.000	Less than 1 year and 1 to 5 years

Note: At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

Variables in the Analysis After Step 4				
Variable	Tolerance	F to Remove	D <sup>2</sup>	Between Groups
X <sub>17</sub> Price Flexibility	.075	.918	6.830	Less than 1 year and 1 to 5 years
X <sub>6</sub> Product Quality	.070	1.735	6.916	Less than 1 year and 1 to 5 years
X <sub>18</sub> Delivery Speed	.680	27.701	3.598	1 to 5 years and more than 5 years
X <sub>18</sub> Delivery Speed	.063	5.387	6.071	Less than 1 year and 1 to 5 years

Variables Not in the Analysis After Step 4					
Variable	Tolerance	Minimum Tolerance	F to Enter	Minimum D <sup>2</sup>	Between Groups
X <sub>7</sub> E-Commerce Activities	.870	.063	.226	6.931	Less than 1 year and 1 to 5 years
X <sub>8</sub> Technical Support	.940	.063	.793	7.164	Less than 1 year and 1 to 5 years
X <sub>9</sub> Complaint Resolution	.453	.058	.292	7.019	Less than 1 year and 1 to 5 years
X <sub>10</sub> Advertising	.932	.063	.006	6.921	Less than 1 year and 1 to 5 years
X <sub>12</sub> Salesforce Image	.843	.061	.315	7.031	Less than 1 year and 1 to 5 years
X <sub>13</sub> Competitive Pricing	.790	.063	.924	7.193	Less than 1 year and 1 to 5 years
X <sub>14</sub> Warranty & Claims	.843	.063	2.023	7.696	Less than 1 year and 1 to 5 years
X <sub>15</sub> New Products	.927	.062	.227	7.028	Less than 1 year and 1 to 5 years
X <sub>16</sub> Order & Billing	.671	.062	1.478	7.210	Less than 1 year and 1 to 5 years

Significance Testing of Group Differences After Step 4 <sup>a</sup>				
X <sub>1</sub> Customer Type	Less than 1 Year		1 to 5 Years	
1 to 5 years	F	13.393		
	Sig.	.000		
More than 5 years	F	56.164	18.477	
	Sig.	.000	.000	

<sup>a</sup>4, 54 degrees of freedom.c

from the discriminant function in step 5 (details of this step 5 are omitted for space considerations). With now three variables in the discriminant function ( $X_{11}$ ,  $X_6$ , and  $X_{18}$ ), the overall model fit is still statistically significant and the Wilks' lambda increased only slightly to .135. All of the groups are significantly different. No variables reach the level of statistical significance necessary to be added to the discriminant function, and one more variable ( $X_{11}$ ) has an "F to Remove" value of 2.552, which indicates that it can also be removed from the function.

Table 7.18 contains the details of step 6 of the stepwise procedure where  $X_{11}$  is also removed from the discriminant function, with only  $X_6$  and  $X_{18}$  as the two variables remaining. Even with the removal of the second variable ( $X_{11}$ ), the overall model is still significant and the Wilks' lambda is quite small (.148). We should note that this two-variable model of  $X_6$  and  $X_{18}$  is an improvement over the first two-variable model of  $X_{11}$  and  $X_{17}$  formed in step 2 (Wilks' lambda is .148 versus the first model's value of .288 and all of the individual

**Table 7.18** Results from Step 6 of Stepwise Three-Group Discriminant Analysis

Overall Model Fit				
	Value	F Value	Degrees of Freedom	Significance
Wilks' Lambda	.148	44.774	4, 112	.000
Variable Entered/Removed at Step 6				
		<i>F</i>		
Variable Removed	Minimum <i>D</i> <sup>2</sup>	Value	Significance	Between Groups
<i>X</i> <sub>11</sub> Product Line	6.388	25.642	.000	Less than 1 year and 1 to 5 years

Note: At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

Variables in the Analysis After Step 6				
Variable	Tolerance	F to Remove	<i>D</i> <sup>2</sup>	Between Groups
<i>X</i> <sub>6</sub> Product Quality	.754	50.494	.007	1 to 5 years and more than 5 years
<i>X</i> <sub>18</sub> Delivery Speed	.754	60.646	.121	Less than 1 year and 1 to 5 years

Variables Not in the Analysis After Step 6					
Variable	Tolerance	Minimum Tolerance	F to Enter	Minimum <i>D</i> <sup>2</sup>	Between Groups
<i>X</i> <sub>7</sub> E-Commerce Activities	.954	.728	.177	6.474	Less than 1 year and 1 to 5 years
<i>X</i> <sub>8</sub> Technical Support	.999	.753	.269	6.495	Less than 1 year and 1 to 5 years
<i>X</i> <sub>9</sub> Complaint Resolution	.453	.349	.376	6.490	Less than 1 year and 1 to 5 years
<i>X</i> <sub>10</sub> Advertising	.954	.742	.128	6.402	Less than 1 year and 1 to 5 years
<i>X</i> <sub>11</sub> Product Line	.701	.529	2.552	6.916	Less than 1 year and 1 to 5 years
<i>X</i> <sub>12</sub> Salesforce Image	.957	.730	.641	6.697	Less than 1 year and 1 to 5 years
<i>X</i> <sub>13</sub> Competitive Pricing	.994	.749	1.440	6.408	Less than 1 year and 1 to 5 years
<i>X</i> <sub>14</sub> Warranty & Claims	.991	.751	.657	6.694	Less than 1 year and 1 to 5 years
<i>X</i> <sub>15</sub> New Products	.984	.744	.151	6.428	Less than 1 year and 1 to 5 years
<i>X</i> <sub>16</sub> Order & Billing	.682	.514	2.397	6.750	Less than 1 year and 1 to 5 years
<i>X</i> <sub>17</sub> Price Flexibility	.652	.628	3.431	6.830	Less than 1 year and 1 to 5 years

Significance Testing of Group Differences After Step 6 <sup>a</sup>			
<i>X</i> <sub>1</sub> Customer Type	Less than 1 Year		1 to 5 Years
1 to 5 years	<i>F</i>	25.642	
	Sig.	.000	
More than 5 years	<i>F</i>	110.261	30.756
	Sig.	.000	.000

<sup>a</sup>6, 52 degrees of freedom.

group differences are much greater). With no variables reaching the significance level necessary for addition or removal, the stepwise procedure terminates.

**SUMMARY OF THE STEPWISE ESTIMATION PROCESS** The estimated discriminant functions are linear composites similar to a regression line (i.e., they are a linear combination of variables). Just as a regression line is an attempt to explain the maximum amount of variation in its dependent variable, these linear composites attempt to explain the variations or differences in the dependent categorical variable. The first discriminant function is developed to explain (account for) the largest amount of variation (difference) in the discriminant groups. The second discriminant function, which is orthogonal and independent of the first, explains the largest percentage of the remaining (residual) variance after the variance for the first function is removed.

The information provided in Table 7.19 summarizes the steps of the three-group discriminant analysis, with the following results:

**Table 7.19** Summary Statistics for Three-Group Discriminant Analysis

Overall Model Fit: Canonical Discriminant Functions									
Function	Percent of Variance								
	Eigenvalue	Function %	Cumulative %	Canonical		Wilks' Lambda	Chi- Square	df	Significance
				Correlation	.148				
1	3.950	91.5	91.5	.893		107.932	4		.000
2	.365	8.5	100.0	.517	.733	17.569	1		.000

Discriminant Function and Classification Function Coefficients										
DISCRIMINANT FUNCTION										
Independent Variables	Unstandardized Discriminant Function		Standardized Discriminant Function		Classification Functions				Over 5 Years	
	Function 1	Function 2	Function 1	Function 2	Less than 1 Year	1 to 5 Years	Classification Functions			
	X <sub>6</sub> Product Quality	.308	1.159	.969	.622	14.382	15.510	18.753		
X <sub>18</sub> Delivery Speed	2.200	.584	1.021	-.533	25.487	31.185	34.401			
(Constant)	-10.832	-11.313			-91.174	-120.351	-159.022			

Structure Matrix										
Independent Variables	Unrotated Discriminant Loadings <sup>a</sup>				Rotated Discriminant Loadings <sup>b</sup>				Function 2	
	Function 1		Function 2		Function 1		Function 2			
	X <sub>9</sub> Complaint Resolution*	.572	-.470		.739		.039			
X <sub>16</sub> Order & Billing	.499	-.263		.546		.143				
X <sub>11</sub> Product Line*	.483	-.256		.529		.137				
X <sub>15</sub> New Products*	.125	-.005		.096		.080				
X <sub>8</sub> Technical Support*	.030	-.017		.033		.008				
X <sub>6</sub> Product Quality*	.463	.886		-.257		.967				
X <sub>18</sub> Delivery Speed	.540	-.842		.967		-.257				
X <sub>17</sub> Price Flexibility*	.106	-.580		.470		-.356				
X <sub>10</sub> Advertising*	.028	-.213		.165		-.138				
X <sub>7</sub> E-Commerce Activities*	-.095	-.193		.061		-.207				
X <sub>12</sub> Salesforce Image*	-.088	-.188		.061		-.198				
X <sub>14</sub> Warranty & Claims*	.030	-.088		.081		.044				
X <sub>13</sub> Competitive Pricing*	-.055	-.059		-.001		-.080				

<sup>a</sup>Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions variables ordered by absolute size of correlation within function.

<sup>b</sup>Pooled within-groups correlations between discriminating variables and rotated standardized canonical discriminant functions.

\*This variable is not used in the analysis.

Group Means (Centroids) of Discriminant Functions <sup>c</sup>		
X <sub>1</sub> Customer Type	Function 1	Function 1
Less than 1 year	-1.911	-1.274
1 to 5 years	.597	-.968
More than 5 years	1.371	1.625

<sup>c</sup>Unstandardized canonical discriminant functions evaluated at group means.

**Variables Included in the Discriminant Functions** Variables  $X_6$  and  $X_{18}$  are the two variables in the final discriminant function, although  $X_{11}$  and  $X_{17}$  were added in the first two steps and then removed after  $X_6$  and  $X_{18}$  were added. The unstandardized and standardized discriminant function coefficients (weights) and the structure matrix of discriminant loadings, unrotated and rotated, are also provided. Rotation of the discriminant loadings facilitates interpretation in the same way that factors were simplified for interpretation by rotation (see Chapter 3 for a more detailed discussion of rotation). We examine the unrotated and rotated loadings more fully in step 5.

Discrimination increased with the addition of each variable (as evidenced by decreases in Wilks' lambda) even though only two variables remained in the final model. By comparing the final Wilks' lambda for the discriminant analysis (.148) with the Wilks' lambda (.414) for the best result from a single variable,  $X_9$ , we see that a marked improvement is made using just two variables in the discriminant functions rather than a single variable.

**Overall Model Fit** The overall goodness-of-fit for the discriminant model is statistically significant and both functions are statistically significant as well. The first function accounts for 91.5 percent of the variance explained by the two functions, with the remaining variance (8.5%) due to the second function. The total amount of variance explained by the first function is  $.893^2$ , or 79.7 percent. The next function explains  $.517^2$ , or 26.7 percent, of the remaining variance (20.3%). Therefore, the total variance explained by both functions is 85.1 percent [ $79.7\% + (26.7\% \times .203)$ ] of the total variation in the dependent variable.

**Group Differences** Even though both discriminant functions are statistically significant, the researcher must always ensure that the discriminant functions provide differences among all of the groups. It is possible to have statistically significant functions, but have at least one pair of groups not be statistically different (i.e., not discriminated between). This problem becomes especially prevalent as the number of groups increases or a number of small groups are included in the analysis.

The last section of Table 7.18 provides the significance tests for group differences between each pair of groups (e.g., group 1 versus group 2, group 1 versus group 3, etc.). All pairs of groups show statistically significant differences, denoting that the discriminant functions created separation not only in an overall sense, but for each group as well. We also examine the group centroids graphically in a later section.

**Assessing Classification Accuracy** Because it is a three-group discriminant analysis model, two discriminant functions are calculated to discriminate among the three groups. Values for each case are entered into the discriminant model and linear composites (discriminant Z scores) are calculated. The discriminant functions are based only on the variables included in the discriminant model.

Table 7.19 provides the discriminant weights of both variables ( $X_6$  and  $X_{18}$ ) and the group means of each group on both functions (lower portion of the table). As we can see by examining the group means, the first function primarily distinguishes group 1 (Less than 1 year) from the other two groups (although a marked difference occurs between groups 2 and 3 as well), whereas the second function primarily separates group 3 (More than 5 years) from the other two groups. Therefore, the first function provides the most separation between all three groups, but is complemented by the second function, which discriminates best (1 and 2 versus 3) where the first function is weakest.

**Assessing Group Membership Prediction Accuracy** The final step of assessing overall model fit is to determine the predictive accuracy level of the discriminant function(s). This determination is accomplished in the same fashion as with the two-group discriminant model, by examination of the classification matrices and the percentage correctly classified (hit ratio) in each sample.

The classification of individual cases can be performed either by the cut-off method described in the two-group case or by using the classification functions (see Table 7.19) where each case is scored on each classification function and classified to the group with the highest score.

**Table 7.20** Classification Results for Three-Group Discriminant Analysis

		Predicted Group Membership				Total
		Less than		More than 5		
	Actual Group	1 Year	1 to 5 Years	Years		
Estimation Sample	Less than 1 year	21	1	0		22
		95.5	4.5	0.0		
	1 to 5 years	2	7	4		13
		15.4	53.8	30.8		
	More than 5 years	0	1	24		25
		0.0	4.0	96.0		
Cross-validated	Less than 1 year	21	1	0		22
		95.5	4.5	0.0		
	1 to 5 years	2	7	4		13
		15.4	53.8	30.8		
	More than 5 years	0	1	24		25
		0.0	4.0	96.0		
Holdout Sample	Less than 1 year	5	3	2		10
		50.0	30.0	20.0		
	1 to 5 years	1	9	12		22
		4.5	40.9	54.5		
	More than 5 years	0	0	8		8
		0.0	0.0	100.0		

Values in each cell represent actual observation counts and then percentage of actual group

<sup>a</sup>86.7% of selected original grouped cases correctly classified.

<sup>b</sup>55.0% of unselected original grouped cases correctly classified.

<sup>c</sup>86.7% of selected cross-validated grouped cases correctly classified.

Table 7.20 shows that the two discriminant functions in combination achieve a high degree of classification accuracy. The hit ratio for the analysis sample is 86.7 percent. However, the hit ratio for the holdout sample falls to 55.0 percent. These results demonstrate the upward bias that is likely when applied only to the analysis sample and not also to a holdout sample.

Both hit ratios must be compared with the maximum chance and the proportional chance criteria to assess their true effectiveness. The cross-validation procedure is discussed in step 6.

**Maximum Chance Criterion** Using the most conservative hit ratio, the maximum chance criterion is simply the hit ratio obtained if we assign all the observations to the group with the highest probability of occurrence. In the present sample of 100 observations, 32 were in group 1, 35 in group 2, and 33 in group 3. From this information, we can see that the highest probability would be 35 percent (group 2). The threshold value for the maximum chance ( $.35 \times 1.25$ ) is 43.74 percent.

**Proportional Chance Criterion** An alternative hit ratio which considers all group sizes is the proportional chance criterion, calculated by squaring the proportions of each group. In this three group example, the calculated value is 33.36 percent ( $.32^2 + .35^2 + .33^2 = .334$ ) and a threshold value of 41.7 percent ( $.334 \times 1.25 = 41.7\%$ ).

**Overall and Group-Specific Hit Ratios** The overall hit ratios for the analysis and holdout samples (86.7% and 55.0%, respectively) exceed both threshold values of 43.74 and 41.7 percent. In the estimation sample, all of the individual groups surpass both threshold values. In the holdout sample, however, group 2 has a hit ratio of only 40.9 percent, and it increased to only 53.8 percent in the analysis sample. These results show that group 2 should be the focus of

improving classification, possibly by the addition of independent variables or a review of classification of firms in this group to identify the characteristics of this group not represented in the discriminant function.

**Press's Q** The final measure of classification accuracy is Press's *Q*, calculated for both analysis and holdout samples. It tests the statistical significance that the classification accuracy is better than chance:

$$\text{Press's } Q_{\text{Estimation Sample}} = \frac{[ 60 - ( 52 \times 3 ) ]^2}{60( 3 - 1 )} = 76.8$$

And the calculation for the holdout sample is:

$$\text{Press's } Q_{\text{Holdout Sample}} = \frac{[ 40 - ( 22 \times 3 ) ]^2}{40( 3 - 1 )} = 8.45$$

Because the critical value at a .01 significance level is 6.63, the discriminant analysis can be described as predicting group membership better than chance.

When completed, we can conclude that the discriminant model is valid and has adequate levels of statistical and practical significance for all groups. The markedly lower values for the holdout sample on all the standards of comparison, however, support the concerns raised earlier about the overall and group-specific hit ratios.

**Casewise Diagnostics** In addition to the classification tables showing aggregate results, case-specific information is also available detailing the classification of each observation. This information can detail the specifics of the classification process or represent the classification through a territorial map.

**CASE-SPECIFIC CLASSIFICATION INFORMATION** A series of case-specific measures is available for identifying the misclassified cases as well as diagnosing the extent of each misclassification. Using this information, patterns among the misclassified may be identified.

Table 7.21 contains additional classification data for each individual case that was misclassified (similar information is also available for all other cases, but was omitted for space considerations). The basic types of classification information include the following:

**Group Membership** Both the actual and predicted groups are shown to identify each type of misclassification (e.g., actual membership in group 1, but predicted in group 2). In this instance, we see the eight cases misclassified in the analysis sample (verify by adding the off-diagonal values in Table 7.20) and the 18 cases misclassified in the holdout sample.

**Mahalanobis Distance to the Predicted Group Centroid** Denotes the proximity of these misclassified cases to the predicted group. Some observations, such as case 10, obviously are similar to the observations of the predicted group rather than their actual group. Other observations, such as case 57 (Mahalanobis distance of 6.041), are likely to be outliers in the predicted group as well as the actual group. The territorial map discussed in the next section graphically portrays the position of each observation and assists in interpretation of the distance measures.

**Discriminant Scores** The discriminant *Z* score for each case on each discriminant function provides a means of direct comparison between cases as well as a relative positioning versus the group means.

**Classification Probability** Derived from use of the discriminant classification functions, the probability of membership for each group is given. The probability values enable the researcher to assess the extent of misclassification. For example, two cases, 85 and 89, are the same type of misclassification (actual group 2, predicted group 3), but quite different in their misclassification when the classification probabilities are viewed. Case 85 represents a marginal misclassification, because the prediction probability for the actual group 2 was .462 and the incorrect predicted group 3 was only slightly higher (.529). This misclassification is in contrast to case 89, where the actual group probability

**Table 7.21** Misclassified Predictions for Individual Cases in the Three-Group Discriminant Analysis

Case ID	GROUP MEMBERSHIP		Mahalanobis Distance to Centroid <sup>a</sup>	DISCRIMINANT SCORES		CLASSIFICATION PROBABILITY		
	X <sub>1</sub> Actual	Predicted		Function 1	Function 2	Group 1	Group 2	Group 3
	Analysis/Estimation Sample							
10	1	2	.175	.81755	−1.32387	.04173	.93645	.02182
8	2	1	1.747	−.78395	−1.96454	.75064	.24904	.00032
100	2	1	2.820	−.70077	−.11060	.54280	.39170	.06550
1	2	3	2.947	−.07613	.70175	.06527	.28958	.64515
5	2	3	3.217	−.36224	1.16458	.05471	.13646	.80884
37	2	3	3.217	−.36224	1.16458	.05471	.13646	.80884
88	2	3	2.390	.99763	.12476	.00841	.46212	.52947
58	3	2	.727	.30687	−.16637	.07879	.70022	.22099
Holdout/Validation Sample								
97	1	2	1.180	−.41466	−.57343	.42296	.54291	.03412
25	1	2	1.723	−.18552	−2.02118	.40554	.59341	.00104
77	1	2	.813	.08688	−.22477	.13933	.70042	.16025
13	1	3	.576	1.77156	2.26982	.00000	.00184	.99816
96	1	3	3.428	−.26535	.75928	.09917	.27855	.62228
83	2	1	2.940	−1.58531	.40887	.89141	.08200	.02659
23	2	3	.972	.61462	.99288	.00399	.10959	.88641
34	2	3	1.717	.86996	.41413	.00712	.31048	.68240
39	2	3	.694	1.59148	.82119	.00028	.08306	.91667
41	2	3	2.220	.30230	.58670	.02733	.30246	.67021
42	2	3	.210	1.08081	1.97869	.00006	.00665	.99330
55	2	3	1.717	.86996	.41413	.00712	.31048	.68240
57	2	3	6.041	3.54521	.47780	.00000	.04641	.95359
62	2	3	4.088	−.32690	.52743	.17066	.38259	.44675
75	2	3	2.947	−.07613	.70175	.06527	.28958	.64515
78	2	3	.210	1.08081	1.97869	.00006	.00665	.99330
85	2	3	2.390	.99763	.12476	.00841	.46212	.52947
89	2	3	.689	.54850	1.51411	.00119	.03255	.96625

<sup>a</sup>Mahalanobis distance to predicted group centroid.

was .032 and the predicted probability for group 3 (the misclassified group) was .966. In both situations of a misclassification, the extent or magnitude varies widely.

The researcher should evaluate the extent of misclassification for each case. Cases that are obvious misclassifications should be selected for additional analysis (profiling, examining additional variables, etc.) discussed in the two-group analysis.

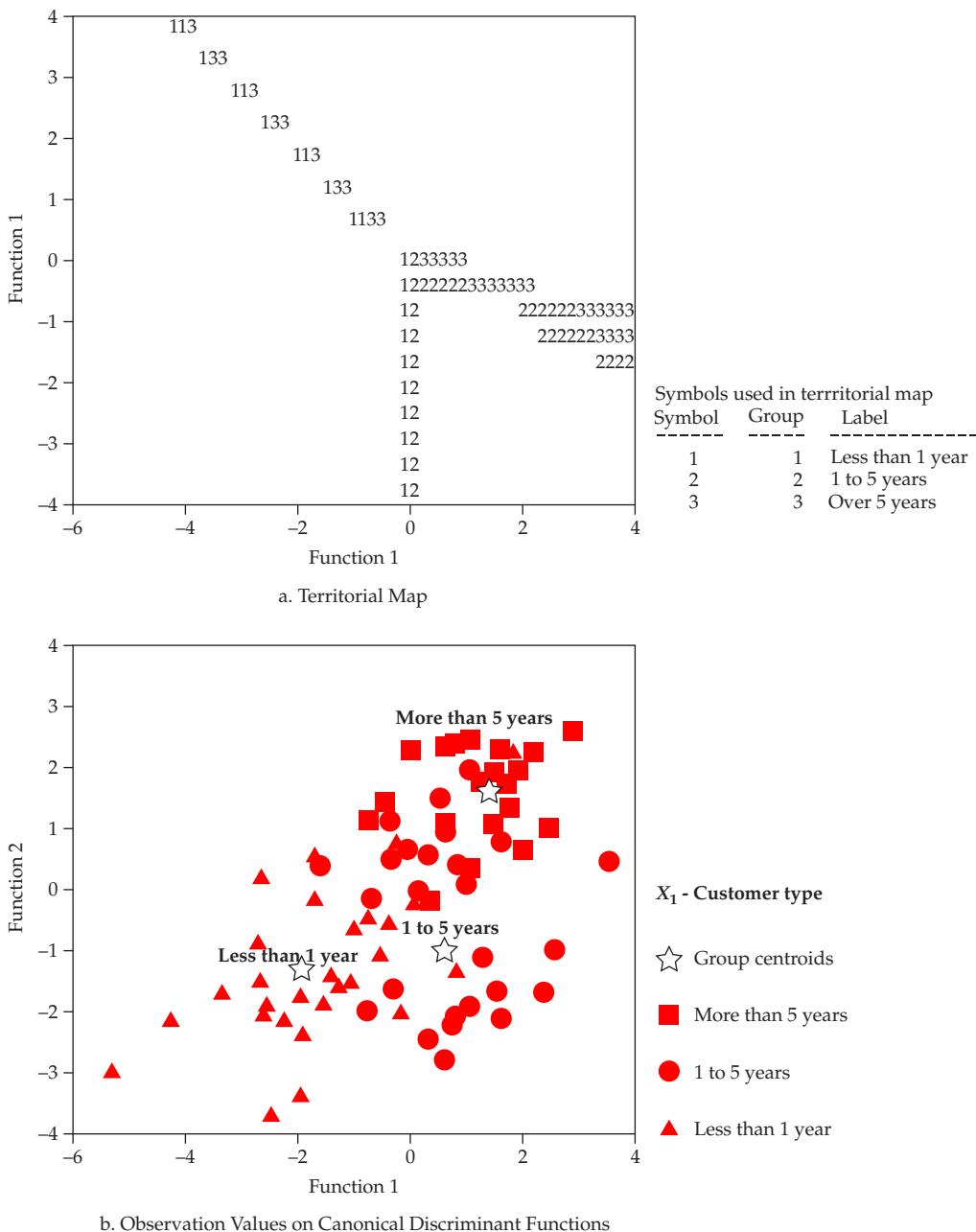
**TERRITORIAL MAP** The analysis of misclassified cases can be supplemented by the graphical examination of the individual observations by plotting them based on their discriminant Z scores.

Figure 7.9 plots each observation based on its two rotated discriminant Z scores with an overlay of the territorial map representing the boundaries of the cutting scores for each function. In viewing each group's dispersion around the group centroid, we can observe several findings:

- Group 3 (More than 5 years) is most concentrated, with little overlap with the other two groups as shown in the classification matrix where only one observation was misclassified (see Table 7.20).

**Figure 7.9**

Territorial Map and Plot of Individual Cases on Discriminant Functions



- Group 1 (Less than 1 year) is the least compact, but the range of cases does not overlap to any great degree with the other groups, thus making predictions much better than might be expected for such a diverse group. The only misclassified cases that are substantially different are case 10, which is close to the centroid for group 2, and case 13, which is close to the centroid of group 3. Both of these cases merit further investigation as to their similarities to the other groups.
- Both of these groups are in contrast to group 2 (1 to 5 years), which can be seen to have substantial overlap with group 3 and to a lesser extent with group 1 (Less than 1 year). This overlap results in the lowest levels of classification accuracy in both the analysis and holdout samples.

- The overlap that occurs between groups 2 and 3 in the center and right of the graph suggests the possible existence of a fourth group. Analysis could be undertaken to determine the actual length of time of customers, perhaps with the customers over 1 year divided into three groups instead of two.

The graphical portrayal is useful not only for identifying these misclassified cases that may form a new group, but also in identifying outliers. The preceding discussion identifies possible options for identifying outliers (case 57) as well as the possibility of group redefinition between groups 2 and 3.

## STAGE 5: INTERPRETATION OF THREE-GROUP DISCRIMINANT ANALYSIS RESULTS

The next stage of the discriminant analysis involves a series of steps in the interpretation of the discriminant functions.

- Calculate the loadings for each function and review the rotation of the functions for purposes of simplifying interpretation.
- Examine the contributions of the predictor variables: (a) to each function separately (i.e., discriminant loadings), (b) cumulatively across multiple discriminant functions with the potency index, and (c) graphically in a two-dimensional solution to understand the relative position of each group and the interpretation of the relevant variables in determining this position.

**Discriminant Loadings and Their Rotation** Once the discriminant functions are calculated, they are correlated with all the independent variables, even those not used in the discriminant function, to develop a structure (loadings) matrix. This procedure enables us to see where the discrimination would occur if all the independent variables were included in the model (i.e., if none were excluded by multicollinearity or lack of statistical significance).

**DISCRIMINANT LOADINGS** The unrotated loadings represent the association of each independent variable with each function, even if not included in the discriminant function. Discriminant loadings, similar to factor loadings described in Chapter 3, are the correlations between each independent variable and the discriminant score.

Table 7.19 contains the structure matrix of unrotated discriminant loadings for both discriminant functions. Selecting variables with loadings of .40 or above as descriptive of the functions, we see that function 1 has five variables exceeding .40 ( $X_9$ ,  $X_{18}$ ,  $X_{16}$ ,  $X_{11}$ , and  $X_6$ ), and four variables are descriptive of function 2 ( $X_6$ ,  $X_{18}$ ,  $X_{17}$ , and  $X_9$ ). Even though we could use these variables to describe each function, we are faced with the issue that three variables ( $X_9$ ,  $X_6$ , and  $X_{18}$ ) have double loadings (variables selected as descriptive of both functions). If we were to proceed with the unrotated loadings, each function would share more variables with the other than it would have as unique.

The lack of distinctiveness of the loadings with each variable descriptive of a single function can be addressed by rotation of the structure matrix, just as was done with factor loadings. For a more detailed description of the rotation process, see Chapter 3.

**ROTATION.** After the discriminant function loadings are calculated, they can be rotated to redistribute the variance (similar to rotation of factors more fully explained in Chapter 3). Basically, rotation preserves the original structure and reliability of the discriminant models while making them easier to interpret substantively.

The rotation of discriminant functions, however, is an option in many software programs. In IBM SPSS, for example, the rotated discriminant function coefficients can be obtained only through the use of command syntax rather than the “pull down” menus. Examples of using command syntax in IBM SPSS and the specific syntax used for discriminant analysis are provided online.

In the present application we chose the most widely used procedure of VARIMAX rotation. The rotation affects the function coefficients and discriminant loadings, as well as the calculation of the discriminant Z scores and the group centroids (see Table 7.19). Examining the rotated versus unrotated coefficients or

loadings reveals a somewhat more simplified set of results (i.e., loadings tend to separate into high versus low values instead of being midrange). The rotated loadings allow for much more distinct interpretations of each function:

- Function 1 is now described by three variables ( $X_{18}$ ,  $X_9$ , and  $X_{16}$ ) that comprised the *Postsale Customer Service* factor during exploratory factor analysis (see Chapter 3 for more details), plus  $X_{11}$  and  $X_{17}$ . Thus, customer service, plus product line and price flexibility, are the descriptors of function 1.
- Function 2 shows only one variable,  $X_6$  (Product Quality), that has a loading above .40 for the second function. Although  $X_{17}$  has a value just under the threshold (−.356), this variable has a higher loading on the first function, which makes it a descriptor of that function. Thus, the second function can be described by the single variable of Product Quality.

With two or more estimated functions, rotation can be a powerful tool that should always be considered to increase the interpretability of the results. In our example, each of the variables entered into the stepwise process was descriptive of one of the discriminant functions. What we must do now is assess the impact of each variable in terms of the overall discriminant analysis (i.e., across both functions).

**Assessing the Contribution of Predictor Variables** Having described the discriminant functions in terms of the independent variables—both those used in the discriminant functions and those not included in the functions—we turn our attention to gaining a better understanding of the impact of the functions themselves and then the individual variables.

**IMPACT OF THE INDIVIDUAL FUNCTIONS.** The first task is to examine the discriminant functions in terms of how they differentiate between the groups.

We start by examining the group centroids on the two functions as shown in Table 7.19. An easier approach is through viewing the territorial map (Figure 7.9):

- Examining the group centroids and the distribution of cases in each group, we see that function 1 primarily differentiates between group 1 versus groups 2 and 3, whereas function 2 distinguishes between group 3 versus groups 1 and 2.
- The overlap and misclassification of the cases of groups 2 and 3 can be addressed by examining the strength of the discriminant functions and the groups differentiated by each. Looking back to Table 7.19, function 1 was by far the most potent discriminator, and it primarily separated group 1 from the other groups. Function 2, which separated group 3 from the others, was much weaker in terms of discriminating power. It is not surprising that the greatest overlap and misclassification would occur between groups 2 and 3, which are differentiated primarily by function 2.

This graphical approach illustrates the differences in the groups due to the discriminant functions but it does not provide a basis for explaining these differences in terms of the independent variables.

To assess the contributions of the individual variables, the researcher has a number of measures to employ—discriminant loadings, univariate  $F$  ratios, and the potency index. The techniques involved in the use of discriminant loadings and the univariate  $F$  ratios were discussed in the two-group example. We will examine in more detail the potency index, a method of assessing a variable's contribution across multiple discriminant functions.

**POTENCY INDEX** The potency index is an additional interpretational technique quite useful in situations with more than one discriminant function. Even though it must be calculated “by hand,” it is very useful in portraying each individual variable's contribution across all discriminant functions.

The potency index reflects both the loadings of each variable and the relative discriminatory power of each function. The rotated loadings represent the correlation between the independent variable and the discriminant  $Z$  score. Thus, the squared loading is the variance in the independent variable associated with the discriminant function.

By weighting the explained variance of each function by the relative discriminatory power of the functions and summing across functions, the potency index represents the total discriminating effect of each variable across all discriminant functions.

Table 7.22 provides the details on calculating a potency index for each of the independent variables. Comparing the variables on their potency index reveals the following:

- $X_{18}$  (Delivery Speed) is the independent variable providing the greatest discrimination between the three types of customer groups.
- It is followed in impact by four variables not included in the discriminant function ( $X_9$ ,  $X_{16}$ ,  $X_{11}$ , and  $X_{17}$ ).
- The second variable in the discriminant function ( $X_6$ ) has only the sixth highest potency value.

Why does  $X_6$  have only the sixth highest potency value, even though it was one of the two variables included in the discriminant function?

- First, remember that multicollinearity affects stepwise solutions due to redundancy among highly multicollinear variables.  $X_9$  and  $X_{16}$  were the two variables highly associated with  $X_{18}$  (forming the Customer Service factor), thus their impact in a univariate sense, reflected in the potency index, was not needed in the discriminant function due to the presence of  $X_{18}$ .
- The other two variables,  $X_{11}$  and  $X_{17}$ , did enter through the stepwise procedure, but were removed once  $X_6$  was added, again due to multicollinearity. Thus, their greater discriminating power is reflected in their potency values even though they too were not needed in the discriminant function once  $X_6$  was added with  $X_{18}$  in the discriminant function.
- Finally,  $X_6$ , the second variable in the discriminant function, has a low potency value because it is associated with the second discriminant function, which has relatively little discriminating impact when compared to the first function. Thus, although  $X_6$  is a necessary element in discriminating among the three groups, its overall impact is less than those variables associated with the first function.

Remember that potency values can be calculated for all independent variables, even if not in the discriminant function(s), because they are based on discriminant loadings. The intent of the potency index is to provide for the interpretation in just such instances where multicollinearity or other factors may have prevented a variable(s) from being included in the discriminant function.

**AN OVERVIEW OF THE EMPIRICAL MEASURES OF IMPACT** As seen in the prior discussions, the discriminatory power of variables in discriminant analysis is reflected in many different measures, each providing a unique role in the interpretation of the discriminant results. By combining all of these measures in our evaluation of the variables, we can achieve a well-rounded perspective on how each variable fits into the discriminant results.

Table 7.23 presents the three preferred interpretive measures (rotated loadings, univariate  $F$  ratio, and potency index) for each of the independent variables. The results support the stepwise analysis, although several cases illustrate the impact of multicollinearity on the procedures and the results.

- Two variables ( $X_9$  and  $X_{18}$ ) have the greatest individual impact as evidenced by their univariate  $F$  values. However, because both are also highly associated (as evidenced by their inclusion on the Customer Service factor in Chapter 3), only one will be included in a stepwise solution. Even though  $X_9$  has a marginally higher univariate  $F$  value, the ability of  $X_{18}$  to provide better discrimination between all of the groups (as evidenced by its larger minimum Mahalanobis  $D^2$  value described earlier) made it a better candidate for inclusion. Thus,  $X_9$ , on an individual basis, has a comparable discriminating power, but  $X_{18}$  will be seen to work better in combination with other variables.
- Three additional variables ( $X_6$ ,  $X_{11}$ , and  $X_{16}$ ) are next highest in impact, but only one,  $X_6$ , is retained in the discriminant function. Note that  $X_{16}$  is highly correlated with  $X_{18}$  (both part of the Customer Service factor) and not included in the discriminant function, whereas  $X_{11}$  did enter the discriminant function, but was one of those variables removed after  $X_6$  was added.

**Table 7.22** Calculation of the Potency Indices for the Three-Group Discriminant Analysis

Independent Variables	Discriminant Function 1			Discriminant Function 2			Potency Index	
	Loading	Squared Loading	Relative Eigenvalue	Potency Value	Loading	Squared Loading	Eigenvalue	Value
X <sub>6</sub> Product Quality	-.257	.066	.915	.060	.967	.935	.085	.079
X <sub>7</sub> E-Commerce Activities	.061	.004	.915	.056	-.207	.043	.085	.004
X <sub>8</sub> Technical Support	.033	.001	.915	.001	.008	.000	.085	.000
X <sub>9</sub> Complaint Resolution	.739	.546	.915	.500	.039	.002	.085	.000
X <sub>10</sub> Advertising	.165	.027	.915	.025	-.138	.019	.085	.002
X <sub>11</sub> Product Line	.529	.280	.915	.256	.137	.019	.085	.002
X <sub>12</sub> Salesforce Image	.061	.004	.915	.004	-.198	.039	.085	.003
X <sub>13</sub> Competitive Pricing	-.001	.000	.915	.000	-.080	.006	.085	.001
X <sub>14</sub> Warranty & Claims	.081	.007	.915	.006	.044	.002	.085	.000
X <sub>15</sub> New Products	.096	.009	.915	.008	.080	.006	.085	.001
X <sub>16</sub> Order & Billing	.546	.298	.915	.273	.143	.020	.085	.002
X <sub>17</sub> Price Flexibility	.470	.221	.915	.202	-.356	.127	.085	.011
X <sub>18</sub> Delivery Speed	.967	.935	.915	.855	-.257	.066	.085	.006

Note: The relative eigenvalue of each discriminant function is calculated as the eigenvalue of each function (shown in Table 7.19 as 3.950 and .365 for discriminant functions I and II respectively) divided by the total of the eigenvalues (3.950 + .365 = 4.315).

**Table 7.23** Summary of Interpretive Measures for Three-Group Discriminant Analysis

	Rotated Discriminant Function Loadings			
	Function 1	Function 2	Univariate F Ratio	Potency Index
X <sub>6</sub> Product Quality	-.257	.967	32.311	.139
X <sub>7</sub> E-Commerce Activities	.061	-.207	1.221	.060
X <sub>8</sub> Technical Support	.033	.008	.782	.001
X <sub>9</sub> Complaint Resolution	.739	.039	40.292	.500
X <sub>10</sub> Advertising	.165	-.138	1.147	.027
X <sub>11</sub> Product Line	.529	.137	32.583	.258
X <sub>12</sub> Salesforce Image	.061	-.198	1.708	.007
X <sub>13</sub> Competitive Pricing	-.001	-.080	9.432	.001
X <sub>14</sub> Warranty & Claims	.081	.044	2.619	.006
X <sub>15</sub> New Products	.096	.080	.216	.009
X <sub>16</sub> Order & Billing	.546	.143	25.048	.275
X <sub>17</sub> Price Flexibility	.470	-.356	12.551	.213
X <sub>18</sub> Delivery Speed	.967	-.257	40.176	.861

- Finally, two variables ( $X_{17}$  and  $X_{13}$ ) had almost equal univariate effects, but only  $X_{17}$  had a substantial association with one of the discriminant functions (a loading of .470 on the first function). The result is that even though  $X_{17}$  can be considered descriptive of the first function and considered having an impact in discrimination based in these functions,  $X_{13}$  does not have any impact, either in association with these two functions or in addition once these functions are accounted for.
- All of the remaining variables had low univariate  $F$  values and low potency values, indicative of little or no impact in both a univariate and multivariate sense.

Of particular note is the interpretation of the two dimensions of discrimination. This interpretation can be done solely through examination of the loadings, but is complemented by a graphical display of the discriminant loadings, as described in the following section.

**GRAPHICAL DISPLAY OF DISCRIMINANT LOADINGS.** To depict the differences in terms of the predictor variables, the loadings and the group centroids can be plotted in reduced discriminant space. As noted earlier, the most valid representation is the use of stretched attribute vectors and group centroids.

Table 7.24 shows the calculations for stretching both the discriminant loadings (used for attribute vectors) and the group centroids. The plotting process always involves all the variables included in the model by the stepwise procedure (in our example,  $X_6$  and  $X_{18}$ ). However, we will also plot the variables not included in the discriminant function if their respective univariate  $F$  ratios are significant, which adds  $X_9$ ,  $X_{11}$ , and  $X_{16}$  to the reduced discriminant space. This procedure shows the importance of collinear variables that were not included in the final stepwise model, similar to the potency index.

The plots of the stretched attribute vectors for the rotated discriminant loadings are shown in Figure 7.10, which is based on the reduced space coordinates for both the five variables used to describe the discriminant functions and each of the groups (see Table 7.24). The vectors plotted using this procedure point to the groups having the highest mean on the respective independent variable and away from the groups having the lowest mean scores. Thus, interpretation of the plot in Figure 7.10 indicates the following:

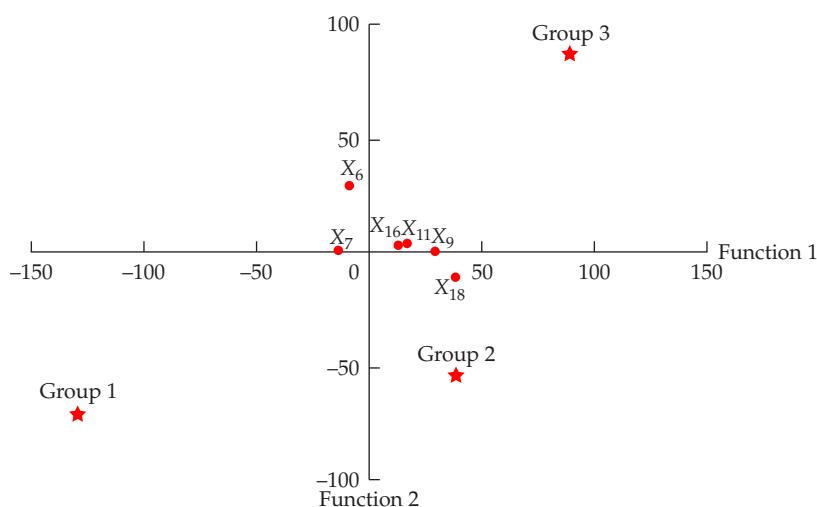
- As noted in the territorial map and analysis of group centroids, the first discriminant function distinguishes between group 1 versus groups 2 and 3, whereas the second discriminant function separates group 3 from groups 1 and 2.
- The correspondence of  $X_{11}$ ,  $X_{16}$ ,  $X_9$ , and  $X_{18}$  with the X axis reflects their association with the first discriminant function, but we see that only  $X_6$  is associated with the second discriminant function. The figure graphically illustrates the rotated loadings for each function and distinguishes the variables descriptive of each function.

**Table 7.24** Calculation of the Stretched Attribute Vectors and Group Centroids in Reduced Discriminant Space

Independent Variables	Rotated Discriminant Function Loadings		Univariate F Ratio	Reduced Space Coordinates	
	Function 1	Function 2		Function 1	Function 2
$X_6$ Product Quality	-.257	.967	32.311	-8.303	31.244
$X_7$ E-Commerce Activities <sup>a</sup>	.061	-.207	1.221		
$X_8$ Technical Support <sup>a</sup>	.033	.008	.782		
$X_9$ Complaint Resolution	.739	.039	40.292	29.776	1.571
$X_{10}$ Advertising <sup>a</sup>	.165	-.138	1.147		
$X_{11}$ Product Line	.529	.137	32.583	17.236	4.464
$X_{12}$ Salesforce Image <sup>a</sup>	.061	-.198	1.708		
$X_{13}$ Competitive Pricing <sup>a</sup>	-.001	-.080	9.432		
$X_{14}$ Warranty & Claims <sup>a</sup>	.081	.044	2.619		
$X_{15}$ New Products <sup>a</sup>	.096	.080	.216		
$X_{16}$ Order & Billing	.546	.143	25.048	13.676	3.581
$X_{17}$ Price Flexibility <sup>a</sup>	.470	-.356	12.551		
$X_{18}$ Delivery Speed	.967	-.257	40.176	38.850	-10.325

<sup>a</sup>Variables with nonsignificant univariate ratios are not plotted in reduced space.

	Group Centroids		Approximate F Value		Reduced Space Coordinates	
	Function 1	Function 2	Function 1	Function 2	Function 1	Function 2
Group 1: Less than 1 year	-1.911	-1.274	66.011	56.954	-126.147	-72.559
Group 2: 1 to 5 years	.597	-.968	66.011	56.954	39.408	-55.131
Group 3: More than 5 years	1.371	1.625	66.011	56.954	90.501	92.550



**Figure 7.10**  
Plot of Stretched Attribute (Variables) in Reduced Discriminant Space

## STAGE 6: VALIDATION OF THE DISCRIMINANT RESULTS

The hit ratios for the cross-classification and holdout matrices can be used to assess the internal and external validity, respectively, of the discriminant analysis. If the hit ratios exceed the threshold values on the comparison standards, then validity is established. As described earlier, the threshold values are 41.7 percent for the proportional chance criterion and 43.7 percent for the maximum chance criterion. The classification results shown in Table 7.20 provide the following support for validity:

Internal validity is assessed by the cross-classification approach, where the discriminant model is estimated by leaving out one case and then predicting that case with the estimated model. This process is done in turn for each

observation, such that an observation never influences the discriminant model that predicts its group classification.

As seen in Table 7.20, the overall hit ratio for the cross-classification approach of 86.7 substantially exceeds both standards, both overall and for each group. However, even though all three groups also have individual hit ratios above the standards, the group 2 hit ratio (53.8) is substantially less than that over the other two groups.

External validity is addressed through the holdout sample, which is a completely separate sample that uses the discriminant functions estimated with the analysis sample for group prediction.

In our example, the holdout sample has an overall hit ratio of 55.0 percent, which exceeds both threshold values, although not to the extent found in the cross-classification approach. Group 2, however, did not exceed either threshold value. When the misclassifications are analyzed, we see that more cases are misclassified into group 3 than correctly classified into group 2, which suggests that these misclassified cases be examined for the possibility of a redefinition of groups 2 and 3 to create a new group.

The researcher is also encouraged to extend the validation process through profiling the groups on additional sets of variables or applying the discriminant function to another sample(s) representative of the overall population or segments within the population. Moreover, analysis of the misclassified cases will help establish whether any additional variables are needed or whether the dependent group classifications need revision.

## A MANAGERIAL OVERVIEW

The discriminant analysis aimed at understanding the perceptual differences of customers based on their length of time as an HBAT customer. Hopefully, examining differences in HBAT perceptions based on tenure as a customer will identify perceptions that are critical to the development of a customer relationship, which is typified by those customers of long standing. Three customer groups were formed—less than 1 year, 1 to 5 years, and more than 5 years—and HBAT perceptions were measured on 13 variables. The analysis produced several major findings, both in terms of the types of variables that distinguished between the groups and the patterns of changes over time:

- First, there are two dimensions of discrimination between the three customer groups. The first dimension is typified by higher perceptions of customer service (Complaint Resolution, Delivery Speed, and Order & Billing), along with Product Line and Price Flexibility. In contrast, the second dimension is characterized solely in terms of Product Quality.
- Profiling the three groups on these two dimensions and variables associated with each dimension enables management to understand the perceptual differences among them.
  - *Group 1, customers of less than 1 year*, generally has the lowest perceptions of HBAT. For the three customer service variables (Complaint Resolution, Order & Billing, and Delivery Speed) these customers are lower than either other group. For Product Quality, Product Line, and Competitive Pricing, this group is comparable to group 2 (customers of 1 to 5 years), but still has lower perceptions than customers of more than 5 years. Only for Price Flexibility is this group comparable to the oldest customers, and both have lower values than the customers of 1 to 5 years. Overall, the perceptions of these newest customers follow the expected pattern of being lower than other customers, but hopefully improving as they remain customers over time.
  - *Group 2, customers of between 1 and 5 years*, has similarities to both the newest and oldest customers. On the three customer service variables, they are comparable to group 3 (customers of more than 5 years). For Product Quality, Product Line, and Competitive Pricing, their perceptions are more comparable to the newer customers (and lower than the oldest customers). They hold the highest perceptions of the three groups on Price Flexibility.
  - *Group 3, representing those customer of 5 years or more*, holds the most favorable perceptions of HBAT as would be expected. Although they are comparable to the customers of group 2 on the three customer service variables (with both groups greater than group 1), they are significantly higher than customers in the other two groups in terms of Product Quality, Product Line, and Competitive Pricing. Thus, this group represents

those customers that have the positive perceptions and have progressed in establishing a customer relationship through the strength of their perceptions.

- Using the three customer groups as indicators in the development of customer relationships, we can identify two stages in which HBAT perceptions change in this development process:
  - *Stage 1.* The first set of perceptions to change is that related to customer service (seen in the differences between groups 1 and 2). This stage reflects the ability of HBAT to positively affect perceptions with service-related operations.
  - *Stage 2.* A longer-run development is needed to foster improvements in more core elements (Product Quality, Product Line, and Competitive Pricing). When these changes occur, the customer hopefully becomes more committed to the relationship, as evidenced by a long tenure with HBAT.
- It should be noted that evidence exists that numerous customers make the transition through stage 2 more quickly than the 5 years as shown by the substantial number of customers who have been customers between 1 and 5 years, yet hold the same perceptions as those long-time customers. Thus, HBAT can expect that certain customers can move through this process possible quite quickly, and further analysis on these customers may identify characteristics that facilitate the development of customer relationships.

Thus, management is presented with managerial input for strategic and tactical planning from not only the direct results of the discriminant analysis, but also from the classification errors.

The underlying nature, concepts, and approach to multiple discriminant analysis have been presented. Basic guidelines for its application and interpretation were included to clarify further the methodological concepts. This chapter helps you to do the following:

**State the circumstances under which linear discriminant analysis should be used instead of multiple regression.** In choosing an appropriate analytical technique, we sometimes encounter a problem that involves a categorical dependent variable and several metric independent variables. Recall that the single dependent variable in regression was measured metrically. Multiple discriminant analysis is one of the appropriate statistical techniques when the research problem involves a single categorical dependent variable and several metric independent variables. In many cases, the dependent variable consists of two groups or classifications, for example, male versus female, high versus low, or good versus bad. In other instances, more than two groups are involved, such as low, medium, and high classifications. Discriminant analysis is capable of handling either two groups or multiple (three or more) groups. The results of a discriminant analysis can assist in profiling the intergroup characteristics of the subjects and in assigning them to their appropriate groups.

**Identify the major issues relating to types of variables used and sample size required in the application of discriminant analysis.** To apply discriminant analysis, the researcher first must specify which variables are to be independent measures and which variable is to be the dependent measure. The researcher should focus on the dependent variable first. The number of dependent variable groups (categories) can be two or more, but these groups must be mutually exclusive and exhaustive. After a decision has been made on the dependent variable, the researcher must decide which independent variables to include in the analysis. Independent variables are selected in two ways: (1) by identifying variables either from previous research or from the theoretical model underlying the research question, and (2) by utilizing the researcher's knowledge and intuition to select variables for which no previous research or theory exists but that logically might be related to predicting the dependent variable groups.

Discriminant analysis, like the other multivariate techniques, is affected by the size of the sample being analyzed. A ratio of 20 observations for each predictor variable is recommended. Because the results become unstable as the sample size decreases relative to the number of independent variables, the minimum size recommended is five observations per independent variable. The sample size of each group also must be considered. At a minimum, the smallest group size of a category must exceed the number of independent variables. As a practical guideline, each category should have at least 20 observations. Even if all categories exceed 20 observations, however, the researcher also must consider the relative sizes of the groups. Wide variations in the sizes of the groups will affect the estimation of the discriminant function and the classification of observations.

**Understand the assumptions underlying discriminant analysis in assessing its appropriateness for a particular problem.** The assumptions for discriminant analysis relate to both the statistical processes involved in the estimation and classification procedures and issues affecting the interpretation of the results. The key assumptions for deriving the discriminant function are multivariate normality of the independent variables and unknown (but equal) dispersion and covariance structures (matrices) for the groups as defined by the dependent variable. If the assumptions are violated, the researcher should understand the impact on the results that can be expected and consider alternative methods for analysis (e.g., logistic regression).

**Describe the two computation approaches for discriminant analysis and the method for assessing overall model fit.** The two approaches for discriminant analysis are the simultaneous (direct) method and the stepwise method. Simultaneous estimation involves computing the discriminant function by considering all of the independent variables at the same time. Thus, the discriminant function is computed based upon the entire set of independent variables, regardless of the discriminating power of each independent variable. Stepwise estimation is an alternative to the simultaneous approach. It involves entering the independent variables into the discriminant function one at a time on the basis of their discriminating power. The stepwise approach follows a sequential process of adding or deleting variables to the discriminant function. After the discriminant function(s) is estimated, the researcher must evaluate the significance or fit of the discriminant function(s). When a simultaneous approach is used, Wilks' lambda, Hotelling's trace, and Pillai's criterion all evaluate the statistical significance of the discriminatory power of the discriminant function(s). If a stepwise method is used to estimate the discriminant function, the Mahalanobis  $D^2$  and Rao's  $V$  measures are most appropriate to assess fit.

**Explain what a classification matrix is and how to develop one, and describe the ways to evaluate the predictive accuracy of the discriminant function.** The statistical tests for assessing the significance of the discriminant function(s) only assess the degree of difference between the groups based on the discriminant Z scores, but do not indicate how well the function(s) predicts. To determine the predictive ability of a discriminant function, the researcher must construct classification matrices. The classification matrix procedure provides a perspective on practical significance rather than statistical significance. Before a classification matrix can be constructed, however, the researcher must determine the cutting score for each discriminant function. The cutting score represents the dividing point used to classify observations into each of the groups based on discriminant function score. The calculation of a cutting score between any two groups is based on the two group centroids (group mean of the discriminant scores) and the relative size of the two groups. The results of the classification procedure are presented in matrix form. The entries on the diagonal of the matrix represent the number of individuals correctly classified. The numbers off the diagonal represent the incorrect classifications. The percentage correctly classified, also termed the *hit ratio*, reveals how well the discriminant function predicts the objects. If the costs of misclassifying are approximately equal for all groups, the optimal cutting score will be the one that will misclassify the fewest number of objects across all groups. If the misclassification costs are unequal, the optimum cutting score will be the one that minimizes the costs of misclassification. To evaluate the hit ratio, we must look at chance classification. When the group sizes are equal, determination of chance classification is based on the number of groups. When the group sizes are unequal, calculating chance classification can be done two ways: maximum chance and proportional chance.

**Tell how to identify independent variables with discriminatory power.** If the discriminant function is statistically significant and the classification accuracy (hit ratio) is acceptable, the researcher should focus on making substantive interpretations of the findings. This process involves determining the relative importance of each independent variable in discriminating between the groups. Three methods of determining the relative importance have been proposed: (1) standardized discriminant weights, (2) discriminant loadings (structure correlations), and (3) partial  $F$  values. The traditional approach to interpreting discriminant functions examines the sign and magnitude of the standardized discriminant weight assigned to each variable in computing the discriminant functions. Independent variables with relatively larger weights contribute more to the discriminating power of the function than do variables with smaller weights. The sign denotes whether the variable makes either a positive or a negative contribution. Discriminant loadings are increasingly used as a basis for interpretation because of the deficiencies in utilizing weights. Measuring the simple linear correlation between each independent variable and the discriminant function, the discriminant loadings reflect the variance that the independent variables share with the discriminant function. They can be interpreted like factor loadings in assessing the relative contribution of each independent variable to the discriminant function. When a stepwise estimation method is used, an additional means of interpreting the relative discriminating power of the independent variables is through the use of partial  $F$  values, which is accomplished by examining the absolute sizes of the significant  $F$  values and ranking them. Large  $F$  values indicate greater discriminatory power.

**Justify the use of a split-sample approach for validation.** The final stage of a discriminant analysis involves validating the discriminant results to provide assurances that the results have external as well as internal validity. In addition to validating the hit ratios, the researcher should use group profiling to ensure that the group means are valid indicators of the conceptual model used in selecting the independent variables. Validation can occur either with a separate sample (holdout sample) or utilizing a procedure that repeatedly processes the estimation sample. Validation of the hit ratios is performed most often by creating a holdout sample, also referred to as the validation sample. The purpose of utilizing a holdout sample for validation purposes is to see how well the discriminant function works on a sample of observations not used to derive the discriminant function. This assessment involves developing a discriminant function with the analysis sample and then applying it to the holdout sample.

Multiple discriminant analysis helps us to understand and explain research problems that involve a single categorical dependent variable and several metric independent variables. This technique can be used to profile the intergroup characteristics of the subjects and assign them to their appropriate groups. Potential applications to both business and non-business problems are numerous.

How would you differentiate among multiple discriminant analysis, regression analysis, logistic regression analysis, and analysis of variance?

What criteria could you use in deciding whether to stop a discriminant analysis after estimating the discriminant function(s)? After the interpretation stage?

What procedure would you follow in dividing your sample into analysis and holdout groups? How would you change this procedure if your sample consisted of fewer than 100 individuals or objects?

How would you determine the optimum cutting score?

How would you determine whether the classification accuracy of the discriminant function is sufficiently high relative to chance classification?

How does a two-group discriminant analysis differ from a three-group analysis?

Why should a researcher stretch the loadings and centroid data in plotting a discriminant analysis solution?

How does discriminant analysis handle the relationship of the dependent and independent variables?

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 2 Crask, M., and W. Perreault. 1977. Validation of Discriminant Analysis in Marketing Research. *Journal of Marketing Research* 14: 60–8.
- 3 Dillon, W. R., and M. Goldstein. 1984. *Multivariate Analysis: Methods and Applications*. New York: Wiley.
- 4 Frank, R. E., W. E. Massey, and D. G. Morrison. 1965. Bias in Multiple Discriminant Analysis. *Journal of Marketing Research* 2: 250–8.
- 5 Gessner, Guy, N. K. Maholtra, W. A. Kamakura, and M. E. Zmijewski. 1988. Estimating Models with Binary Dependent Variables: Some Theoretical and Empirical Observations. *Journal of Business Research* 16: 49–65.
- 6 Green, P. E., D. Tull, and G. Albaum. 1988. *Research for Marketing Decisions*. Upper Saddle River, NJ: Prentice Hall.
- 7 Green, P. E. 1978. *Analyzing Multivariate Data*. Hinsdale, IL: Holt, Rinehart and Winston.
- 8 Green, P. E., and J. D. Carroll. 1978. *Mathematical Tools for Applied Multivariate Analysis*. New York: Academic Press.
- 9 Harris, R. J. 2001. *A Primer of Multivariate Statistics*, 3rd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 10 Huberty, C. J. 1984. Issues in the Use and Interpretation of Discriminant Analysis. *Psychological Bulletin* 95: 156–71.
- 11 Huberty, C. J., J. W. Wisenbaker, and J. C. Smith. 1987. Assessing Predictive Accuracy in Discriminant Analysis. *Multivariate Behavioral Research* 22: 307–29.
- 12 Johnson, N., and D. Wichern. 2002. *Applied Multivariate Statistical Analysis*, 5th edn. Upper Saddle River, NJ: Prentice Hall.
- 13 Morrison, D. G. 1969. On the Interpretation of Discriminant Analysis. *Journal of Marketing Research* 6: 156–63.
- 14 Perreault, W. D., D. N. Behrman, and G. M. Armstrong. 1979. Alternative Approaches for Interpretation of Multiple Discriminant Analysis in Marketing Research. *Journal of Business Research* 7: 151–73.
- 15 Preacher, K. J., D. D. Rucker, R. C. MacCallum, and W. A. Nicewander. 2010. Use of the Extreme Groups Approach: A Critical Reexamination and New Recommendations. *Psychological Methods* 10: 178–92.

# 8

# Logistic Regression: Regression with a Binary Dependent Variable

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

State the circumstances under which logistic regression should be used instead of multiple regression or discriminant analysis.

Identify the types of variables used for dependent and independent variables in the application of logistic regression.

Describe the method used to transform binary measures into the likelihood and probability measures used in logistic regression.

Interpret the results of a logistic regression analysis and assessing predictive accuracy, with comparisons to both multiple regression and discriminant analysis.

Understand the strengths and weaknesses of logistic regression compared to discriminant analysis and multiple regression.

## Chapter Preview

Logistic regression is a specialized form of regression that is formulated to predict and explain a binary (two-group) categorical variable rather than a metric-dependent measure. The form of the logistic regression variate is similar to the variate in multiple regression. The variate represents a single multivariate relationship, with regression-like coefficients indicating the relative impact of each predictor variable. Moreover, casewise diagnostics (e.g., measures of influential observations) are also available. And since the dependent variable is categorical, we will also employ new measures of predictive accuracy.

The differences between logistic regression and discriminant analysis, another method for examining a categorical dependent variable, will also become apparent in our discussion of logistic regression's unique characteristics. Logistic regression has the advantage of being less affected than discriminant analysis when basic assumptions underlying statistical inference, particularly the normality of the variables and the inherent heteroscedasticity introduced by binary dependent measures, are not met. Logistic regression also can accommodate nonmetric independent variables through dummy-variable coding, just as regression can. Logistic regression is limited, however, to prediction of only a two-group dependent measure. While it can be extended to a multi-category dependent measure, situations for which three or more groups form the dependent measure, discriminant analysis is better suited in many instances in these multi-group situations.

While there are differences that many times favor logistic regression over discriminant analysis, many similarities also exist between the two methods. When the basic assumptions of both methods are met, they each give comparable

predictive and classificatory results and employ similar diagnostic measures. The reader is encouraged to examine Chapter 7, where we discuss in much more detail discriminant analysis as an alternative approach to situations with a nonmetric dependent variable, particularly a multi-category independent variable.

In Chapter 1, logistic regression was described as estimating the relationship between a single nonmetric (binary) dependent variable and a set of metric or nonmetric independent variables, in this general form:

$$Y_1 = X_1 + X_2 + X_3 + \dots + X_n$$

(binary nonmetric)      (nonmetric and metric)

As we will discuss later, the primary difference between logistic regression and regression is the transformation of the dependent measure in logistic regression. To provide a metric value for the dependent variable rather than just values of 1 or 0, the dependent variable is expressed as a probability based upon values of the independent variables. For example, assume you are attempting to identify the factors that are related to whether a current customer purchases a new product in the firm's product line or not. We know that overall 30 percent of the customers purchased the new product. But which customer characteristics impacted this purchase rate, if any? To examine gender, for example, we could now calculate the purchase rate for males (assume a 15 percent purchase rate is found) and then females (they have a purchase rate of 40%). We can see by comparing these probabilities that gender does seem to make some difference (i.e., females have a higher rate than males). We can do this for an entire set of characteristics and then, through the logistic regression model, assign weights to reflect the unique impact of each characteristic's impact on the purchase rate. These weights are comparable to the weights of multiple regression, but come in several forms that reflect various ways of expressing the effect.

Logistic regression has gained widespread application in situations involving a binary outcome (e.g., Yes/No). The prevalence of this outcome, coupled with the ease of use and robust estimation properties, have made logistic regression one of the most widely used techniques in the social sciences today. It has a long history in the fields of health care and social sciences, particularly as an extension of multiway frequency analysis [4, 5]. Indeed, use of the method has now been extended to almost every discipline, and it has been particularly impactful in business-related research [19, 1, 28].

But its use is not limited to just an explanatory objective, as it also provides a robust classification capability. As a result, the analyst not only can identify and understand the factors that drive a particular outcome (e.g., a loan default or not), but can also use logistic regression to predict the expected outcome for new loans and whether they should be approved or not. It is its use as a classifier that has made it such an instrumental and popular tool in data-driven decision making today. This capability also underlies much of the automated decisions in our digital domain today, such as real-time fraud protection on credit card purchases [8, 29, 26]. Thus, the simple binary outcome can be analyzed and then formalized for classification purposes across almost any situation.

## Key Terms

Before starting the chapter, review the key terms to develop an understanding of the concepts and terminology to be used.

Throughout the chapter the key terms appear in **boldface**. Other points of emphasis in the chapter and key term cross-references are *italicized*.

**Accuracy** Percentage of objects (individuals, respondents, firms, etc.) correctly classified by the logistic regression model. It is calculated as the number of objects in the diagonal of the *classification matrix* divided by the total number of objects. Also known as the *percentage correctly classified* or the *hit ratio*.

**Analysis sample** Group of cases used in estimating the *logistic regression* model. When constructing *classification matrices*, the original sample is divided randomly into two groups, one for model estimation (the analysis sample) and the other for validation (the *holdout sample*).

**AUC (area under the curve)** The area under the ROC curve which indicates the amount of discrimination for the estimated model. Higher values indicate better model discrimination and model fit, values of .5 indicate no discrimination and model fit no better than chance.

**Categorical variable** See *nonmetric variable*.

**C/CBAR** An *influence measure* indicative of an observation's impact on overall model fit, similar to Cook's distance measure in multiple regression.

**Chi-square difference** An *influence measure* that represents the amount the model chi-square value will decrease when the observation is omitted from the analysis.

**Classification matrix** Means of assessing the predictive ability of the logistic regression model. Created by cross-tabulating actual group membership with predicted group membership, this matrix consists of numbers on the diagonal representing correct classifications (*True Positives* and *True Negatives*) and off-diagonal numbers representing incorrect classifications (*False Positives* and *False Negatives*).

**Complete separation** A situation when an independent variable provides perfect prediction of the dependent measure. This creates a situation in which the model cannot be estimated and the variable must be eliminated from the analysis.

**Cross-validation** Procedure of dividing the sample into two parts: the *analysis sample* used in estimation of the logistic regression model and the *holdout sample* used to validate the results. Cross-validation avoids the overfitting of the logistic regression by allowing its validation on a totally separate sample.

**Deviance difference** An *influence measure* for an observation indicating the decrease in the model log likelihood value when that observation is omitted from the analysis.

**Deviance residual** A residual of an observation indicating its contribution to the  $-2 \log \text{likelihood}$  value of the estimated model.

**Dfbeta** An *influence measure* indicating the change in each estimated parameter if that observation is omitted from the analysis.

**Exponentiated logistic coefficient** Antilog of the *logistic coefficient*, which is used for interpretation purposes in logistic regression. The exponentiated coefficient minus 1.0 equals the percentage change in the *odds*. For example, an exponentiated coefficient of .20 represents a negative 80 percent change in the odds ( $.20 - 1.0 = -.80$ ) for each unit change in the independent variable (the same as if the odds were multiplied by .20). Thus, a value of 1.0 equates to no change in the odds and values above 1.0 represent increases in the predicted odds.

**False negative** One of the four cells in the *classification matrix* that denotes the number of observations that were positive, but incorrectly predicted to be negative.

**False positive** One of the four cells of the *classification matrix* that denotes the number of observations that were negative, but incorrectly predicted to be positive.

**Hit ratio** See *accuracy*.

**Holdout sample** Group of objects not used to compute the logistic regression model. This group is then used to validate the logistic regression model with a separate sample of respondents. Also called the *validation sample*.

**Hosmer and Lemeshow test** A chi-square based test of overall model predictive accuracy where actual and predicted outcomes are compared and a statistical significance test performed. Nonsignificance indicates close correspondence between actual and predicted, thus indicating a good predictive model.

**Influence measure** A measure of the impact of an individual observation either on the overall model fit or on any of the estimated model coefficients.

**Information value** An overall measure of each independent variable's ability to distinguish between the outcome categories based on *weight of evidence* for categories of the independent variable. It provides a method separate from the actual model estimates to assess the role of each independent variable to assist in both variable selection and relative comparison of independent variables in terms of their predictive impact.

**Leverage** An *influence measure* that depicts how "typical" an observation is based on the entire set of independent variables. High leverage observations are those that are most different from the centroid of the sample on the independent variables.

**Likelihood value** Measure used in *logistic regression* to represent the lack of predictive fit. Even though this method does not use the least squares procedure in model estimation, as is done in multiple regression, the likelihood value is similar to the sum of squared error in regression analysis.

**Logistic curve** An S-shaped curve formed by the *logit transformation* that represents the probability of an event. The S-shaped form is nonlinear, because the probability of an event must approach 0 and 1, but never fall outside these limits. Thus, although the midrange involves a linear component, the probabilities as they approach the lower and upper bounds of probability (0 and 1) must flatten out and become asymptotic to these bounds.

**Logistic regression** A special form of regression in which the dependent variable is a nonmetric, dichotomous (binary) variable. Although some differences exist, the general manner of interpretation is somewhat similar to linear regression.

**Logit analysis** See *logistic regression*.

**Logit transformation** Transformation of the values of the discrete binary dependent variable of *logistic regression* into an S-shaped curve (*logistic curve*) representing the probability of an event. This probability is then used to form the *odds ratio*, which acts as the dependent variable in logistic regression.

**Maximum chance criterion** Measure of predictive accuracy in the *classification matrix* that is calculated as the percentage of respondents in the largest group. The rationale is that the best uninformed choice is to classify every observation into the largest group.

**Maximum likelihood procedure** A method for estimating the parameters of a statistical model based on parameter values that maximize the likelihood function.

**Misclassification cost** A method of portraying the differing costs of making an incorrect prediction (i.e., cost of a false positive or a *false negative*). Can be integrated into the *classification matrix* to reflect these differential impacts across all outcomes.

**Negative predictive value (NPV)** Represents the probability of making a correct negative prediction—the percentage of negative predictions that are actually *true negative*.

**Nonmetric variable** Variable with values that serve merely as a label or means of identification, also referred to as a *categorical, nominal, binary, qualitative, or taxonomic* variable. The number on a football jersey is an example. A more complete discussion of its characteristics and its differences from a *metric variable* is found in Chapter 1.

**Odds** The ratio of the probability of an event occurring to the probability of the event not happening, which is used as a measure of the dependent variable in *logistic regression*.

**Original logistic coefficient** Estimated parameter from the logistic model that reflects the change in the logged odds value (logit) for a one unit change in the independent variable. It is similar to a regression weight or discriminant coefficient.

**Pearson residual** A residual for an observation that is the standardized difference between the actual outcome value (1 or 0) and the predicted probability.

**Percentage correctly classified** See *accuracy*.

**Positive predictive value (PPV)** Represents the probability of making a correct positive prediction—the percentage of positive predictions that are actually *true positive*.

**Probit** Alternative to the logistic function as the statistical model used in the estimation procedure. Generally provides results quite comparable to the *logit transformation* and while providing some advantages for instances of multi-category dependent measures, its estimated coefficients are somewhat more difficult to interpret.

**Proportional chance criterion** Another criterion for assessing the *hit ratio*, in which the average probability of classification is calculated considering all group sizes.

**Pseudo R<sup>2</sup>** A value of overall model fit that can be calculated for *logistic regression*; comparable to the  $R^2$  measure used in multiple regression.

**Quasi-complete separation** A situation where at least one cell related to an independent variable has zero observations, causing problems in estimation of the parameter estimate associated with that independent variable. While the parameter estimate may be inaccurate, the model can still be estimated, which is not the case for *complete separation*.

**Relative importance** A measure of contribution of each of the independent variable where the weights will sum to the coefficient of determination,  $R^2$ .

**ROC curve** A graphical means of portraying the trade-off between *sensitivity* (*true positive rate*) versus  $1 - \text{specificity}$  (*false positive rate*) for all possible cutoff values between 0 and 1.

**Sensitivity** Represents the *true positive* rate—percentage of actual positive outcomes that are predicted as positive.

**Specificity** Represents the *true negative* rate—percentage of actual negative outcomes that are predicted as negative.

**True negative** One of the four cells of the classification matrix that denotes the number of observations that were negative and were correctly predicted to be negative.

**True positive** One of the four cells of the classification matrix that denotes the number of observations that were positive and were correctly predicted to be positive.

**Validation sample** See *holdout sample*.

**Variate** Linear combination that represents the weighted sum of two or more independent variables that comprise the *discriminant function*. Also called linear combination or linear compound.

**Wald statistic** Test used in *logistic regression* for the significance of the *logistic coefficient*. Its interpretation is like the *F* or *t* values used for the significance testing of regression coefficients.

**Weight of evidence.** A measure of the impact of each category of a nonmetric independent variable in terms of distinguishing between the two outcome groups. When combined across all categories, it provides an overall measure termed *information value*.

**Youden index** A measure of overall predictive accuracy that is calculated as the sum of *sensitivity* and *specificity* minus 1. Higher values indicate better model fit with a maximum value of 1 and a minimum value of –1.

## What Is Logistic Regression?

Logistic regression, along with discriminant analysis, is the appropriate statistical technique when the dependent variable is a **categorical** (nominal or **nonmetric**) **variable** and the independent variables are metric or nonmetric variables. When compared to discriminant analysis, logistic regression is limited in its basic form to two groups for the dependent variable, although other formulations can handle more groups. It does have the advantage, however, of easily incorporating nonmetric variables as independent variables, much like in multiple regression.

In a practical sense, logistic regression may be preferred for two reasons. First, discriminant analysis relies on strictly meeting the assumptions of multivariate normality and equal variance–covariance matrices across groups—assumptions

that are not met in many social science research situations. Logistic regression is not as limited by these strict assumptions and is much more robust when these assumptions are not met, making its application appropriate in many situations. Second, even if the assumptions are met, many researchers prefer logistic regression because it is more similar to multiple regression. It has straightforward statistical tests, similar approaches to incorporating metric and nonmetric variables and nonlinear effects, and a wide range of diagnostics. Thus, for these and other more technical reasons, logistic regression is equivalent to two-group discriminant analysis and may be more suitable in many situations.

## The Decision Process for Logistic Regression

The application of logistic regression can be viewed from the six-stage model-building perspective introduced in Chapter 1. As with all multivariate applications, setting the objectives is Stage 1 in the analysis. In Stage 2, the researcher must address specific design issues, including specifying the dependent variable and ensuring an adequate samples size. Stage 3 involves making sure the underlying assumptions are met, even though logistic regression has fewer assumptions to meet than discriminant analysis or multiple regression. Stage 4 proceeds with the estimation of the probability of occurrence in each of the groups by use of the logistic curve as the underlying relationship. The binary measure is translated into the odds of occurrence and then a logit value (log of the odds) that acts as the dependent measure. The model form in terms of the independent variables is almost identical to multiple regression. Model fit is assessed much like discriminant analysis by first looking for statistical significance of the overall model and then determining predictive accuracy by developing a classification matrix. With an acceptable overall model achieved, Stage 5 focuses on interpretation of the variable coefficients. Given the unique nature of the transformed dependent variable, logistic coefficients are given in their “original” scale, which is in logarithmic terms, and a transformed scale, which is interpreted more like regression coefficients. Each form of the coefficient details a certain characteristic of the independent variable’s impact. Finally, in Stage 6, the logistic regression model should be validated with a holdout sample or other cross-validation procedures.

Each of these stages is discussed in the following sections. Our discussion focuses both on the characteristics of logistic regression while also identifying the differences between logistic regression and discriminant analysis or multiple regression. Thus, as mentioned earlier, the reader should also review Chapter 7 for many of the underlying principles of discriminant analysis models with nonmetric dependent variables and even Chapter 5 for the basics of multiple regression models.

### Stage 1: Objectives of Logistic Regression

Logistic regression, like discriminant analysis, is a statistically based technique in the general class of methods termed classifiers—models with the purpose of classifying objects into distinct groups based on the characteristics of the object. A common characteristic of both logistic regression and discriminant analysis is that group membership is known whereas in methods such as cluster analysis (see Chapter 4) group membership is not known. Among those techniques based on known group membership, the two methods are distinctive in that they are statistical models (see discussion of statistical models versus algorithms in Chapter 1) unlike other classification techniques like decision trees or support vector machines. As such, logistic regression and discriminant analysis have two complementary, but equally important objectives—explanation and prediction (i.e., classification in this case).

#### EXPLANATION

The first objective focuses on identifying the independent variables that impact group membership represented by the dependent variable. Here the focus is on the variate in terms of (a) specifying which object characteristics should be included as independent variables and (b) estimating the importance of each independent variable in explaining

group membership. Here the similarities are perhaps most apparent, as issues such as statistical significance of the independent variables as well as assessing their relative importance are crucial. The outcome is to provide the analyst with greater understanding or insight into the “reasons” for individual observations being in one group versus another.

## CLASSIFICATION

The second objective involves establishing a classification system based on the logistic model for determining group membership. Here the ultimate goal of prediction is not a specific metric value, like in multiple regression, but instead a method of placing each observation into a distinct category/group. The measure of predictive accuracy then becomes the correct classification (not  $R^2$ , as with multiple regression), with particular emphasis on various types of misclassification and their associated costs. Thus, as analytics becomes more embedded in decisions in almost every economic sector, predictive accuracy becomes increasingly important.

## Stage 2: Research Design for Logistic Regression

Logistic regression has several unique features that impact the research design. First is the unique nature of the binary dependent variable, which ultimately impacts the model specification and estimation. The second issue relates to sample size, both overall and within groups. The sample size is critical consideration due to the use of maximum likelihood as the estimation technique (it requires larger sample sizes than ordinary least squares, for example) as well as the sensitivity of logistic regression to small group sample sizes for both prediction (i.e., what is termed rare event analysis) and estimation (the “zero cells” effect). Moreover, the need for estimation and holdout samples, much like discriminant analysis, adds another requirement for maintaining an adequate sample size, both overall and within groups.

## REPRESENTATION OF THE BINARY DEPENDENT VARIABLE

Most classification techniques execute the classification prediction in two steps: (1) first by estimating one or more metric values that are then (2) used to assign an object to a specific group. In discriminant analysis, the nonmetric character of a dichotomous dependent variable is accommodated by making predictions of group membership based on discriminant Z scores. To do so, it calculates cutting scores to identify how to separate observations into groups, and then it assigns each observation to one of the groups.

**Assigning Binary Values** Logistic regression approaches the task of assigning binary values in a manner similar to discriminant analysis, but with logistic regression there are only two groups. The two groups of interest in logistic regression are represented as binary variables with values of 0 and 1. It does not matter which group is assigned the value of 1 versus 0, but this assignment (coding) must be noted for the interpretation of the coefficients.

- If the groups represent characteristics (e.g., gender), then either group can be assigned the value of 1 (e.g., females) and the other group the value of 0 (e.g., males). In such a situation, the coefficients would reflect the impact of the independent variable(s) on the likelihood of the person being female (i.e., the group coded as 1).
- If the groups represent outcomes or events (e.g., success or failure, purchase or non-purchase), the assignment of the group codes impacts interpretation as well. Assume that the group with success is coded as 1, with failure coded as 0. Then, the coefficients represent the impacts on the likelihood of success. Just as easily, the codes could be reversed (code of 1 now denotes failure) and the coefficients represent the forces increasing the likelihood of failure. This approach to coding is interpreted, therefore, as if the independent variables are predicting the group coded as 1.

Thus, just as we have seen when using the dummy variable approach to creating binary variables to represent multi-category nonmetric variables, the analyst must be aware of how the values are assigned and make sure they match the intent of the research problem.

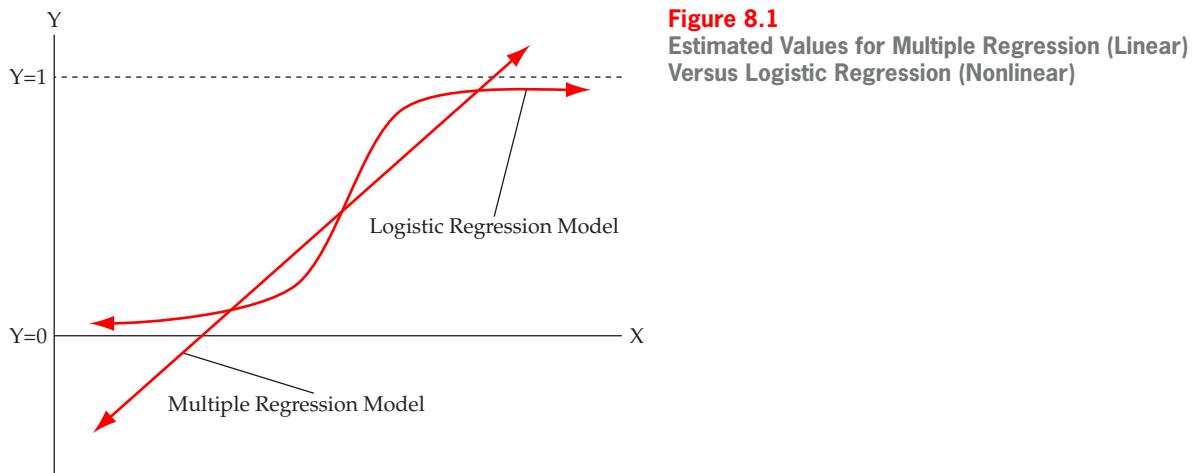
**Use of the Logistic Curve** Logistic regression, like most classification methods, is designed to predict a metric value, in this case the probability of an event occurring (i.e., the probability of an observation being in the group coded 1 versus the group coded 0). And although probability values are metric measures, there are fundamental differences between logistic regression and other models predicting metric outcomes, such as multiple regression. Because the binary dependent variable has only the values of 0 and 1, the predicted value (probability) must be bounded to fall within the same range. Yet the linear relationship from regression, even with additional terms of transformations for nonlinear effects, cannot guarantee that the predicted values will remain within the range of 0 and 1. We see the typical linear relationship portrayed in Figure 8.1, where it is impossible for a linear relationship to provide estimated values approaching zero and one without exceeding these values.

What is needed is an inherently nonlinear relationship, where at very low levels of the independent variable, the probability should approach 0, but never reach it. Likewise, as the independent variable increases, the predicted values should increase, but the probability should never exceed 1. This requires that the slope of the relationship change across values of the independent variable, starting almost horizontal, then increasing upwards towards values of one and then again starting to decrease so that at the highest levels of the independent variable the probability will approach 1.0 but never exceed it.

To define this relationship bounded by 0 and 1, logistic regression uses the **logistic curve** to represent the relationship between the independent and dependent variables (also shown in Figure 8.1). Using this form of the relationship allows for a direct estimation of a nonlinear relationship, although we will see in Stage 4 that probability values must be transformed before they can be used with the logistic function.

**Unique Nature of the Dependent Variable** The binary nature of the dependent variable (0 or 1) has properties that violate the assumptions of multiple regression. First, the error term of a discrete variable follows the binomial distribution instead of the normal distribution, thus invalidating all statistical testing based on the assumptions of normality. Second, the variance of a dichotomous variable is not constant, creating instances of heteroscedasticity as well. Moreover, neither violation can be remedied through transformations of the dependent or independent variables.

While logistic regression was developed specifically to deal with these issues, it does require differences in the assumptions underlying the model (see Stage 3) as well as estimating the variate, assessing goodness-of-fit, and interpreting the coefficients (discussion in Stages 4 and 5) when compared to multiple regression or discriminant analysis.



## SAMPLE SIZE

Logistic regression, like every other multivariate technique, must consider the size of the sample being analyzed. As discussed in Chapter 1, when dealing with statistical models, parameter estimates from very small samples have so much sampling error that identification of all but the largest differences is improbable. Very large sample sizes increase the statistical power so that any difference, whether practically relevant or not, will be considered statistically significant. Yet most research situations fall somewhere in between these extremes, meaning the researcher must consider the impact of sample sizes on the results, both at the overall level and on a group-by-group basis.

**Overall Sample Size** The first aspect of sample size is the overall sample size needed to adequately support estimation of the logistic model. One factor that distinguishes logistic regression from the other techniques is its use of maximum likelihood (MLE) as the estimation technique. MLE requires larger samples such that, all things being equal, logistic regression will require a larger sample size than multiple regression. For example, Hosmer and Lemeshow recommend sample sizes greater than 400 [21], although logistic regression is applied successfully in many situations which have smaller samples. Moreover, the researcher should strongly consider dividing the sample into analysis and holdout samples as a means of validating the logistic model (see a more detailed discussion in Stage 6). In making this split of the sample, the sample size requirements still hold for the analysis sample, thus increasing the initial overall sample size needed based on the model specification (number of parameters estimates, etc.).

**Sample Size Per Category of the Dependent Variable** While overall sample size is important, a more frequently encountered issue involves the sample size per group of the dependent variable. Many times logistic regression is employed for what is termed a rare event situation [22]—where the event of interest has a very low incidence rate (e.g., natural disasters, extreme economic conditions or other very infrequent natural or man-made events). While rare events most often are associated with events that have a very low number of occurrences, the techniques developed are also applicable to situations such as credit card fraud that have a very low rate of occurrence (e.g., 3 to 5% rate). While it is the frequency of the rare events that has the greatest impact, we will also examine issues associated with low rates of occurrence.

**LOW FREQUENCY OF OCCURRENCE** The primary consideration is that in samples of less than 200 there are potential biases in the estimated coefficients for the independent as well as the estimated probabilities of the rare events—they are typically underestimated [22]. So while a situation of 50 events in a sample of 1,000 would be problematic, 500 events in a sample of 10,000 would be quite appropriate for analysis. The recommended sample size for each group is at least 10 observations per estimated parameter. This is much greater than multiple regression, which had a minimum of five observations per parameter, and that was for the overall sample, not the sample size for each group, as seen with logistic regression.

There are several approaches for dealing with low frequency of occurrence. One approach when the overall sample is small is the use of exact logistic regression [18, 27, 14]. This method overcomes the biases associated with small sample sizes, but is also limited by computational issues with small samples (under 200) and a small number of independent variables, preferably binary. The other approach is some form of “penalized” estimation method in which inherent biases are estimated and corrected during estimation [22, 11]. These methods are becoming more available and provide a viable alternative to address the small sample size problem.

**LOW RATE OF OCCURRENCE** As discussed above, a low rate of occurrence is not inherently a problem unless the low rate is associated with a small sample size of events. As is many time mentioned, the “information is in the ones, not the zeros.” This allows for effective data management in that once a suitable sample of events is achieved, it is not necessary to include all of the non-events in the analysis. This form of data management, termed choice-based sampling or case-control design, enables the analyst to utilize as many of the events as possible and then select a random sample of the non-events [6, 31]. This approach is the rationale for the widely used approach of randomly selecting a sample of non-events to match the size of the sample of events, thus making the rate of event versus

non-event equal and eliminating any issues dealing with differing subsample sizes. This approach does not impact the estimated logistic coefficients, just the intercept term, and that can be adjusted as needed when making predictions back to the population.

**Impact of Nonmetric Independent Variables** A final consideration comes into play with the use of nonmetric independent variables. When they are included in the model, they further subdivide the sample into cells created by the combination of dependent and nonmetric independent variables. For example, a simple binary independent variable creates four groups when combined with the binary dependent variable. Although it is not necessary for each of these groups to meet the sample size requirements described above, the researcher must still be aware that if any one of these cells has a very small sample size then it is effectively eliminated from the analysis. All cell frequencies should be greater than 1 and a large majority (75% or more) should have frequencies greater than 5. The presence of small or empty cells may cause the logistic model to become unstable, reporting implausibly large logistic coefficients and odds ratios for independent variables. We will revisit this issue in Stage 4 when addressing quasi-separation and its impact on the estimation process.

## USE OF AGGREGATED DATA

To this point we have focused on individual observations, but it is also possible to analyze aggregated data. In aggregated form, we first identify each unique pattern of values for the set of nonmetric independent variables. For example, if there were three binary independent variables, there would be eight ( $2 \times 2 \times 2 = 8$ ) possible combinations resulting in a dataset with eight cases showing the values for each independent variable (e.g., 0, 0, 0; 0, 0, 1; ...; 1, 1, 1). For each pattern we also include the proportion of observations with that pattern that had a 1 for the dependent measure. In our example, each of the eight records would have values for one of the unique combinations and the proportion of events for observations with those independent variable values.

## Stage 3: Assumptions of Logistic Regression

The advantages of logistic regression compared to discriminant analysis and even multiple regression stem in large degree to the general lack of assumptions required in a logistic regression analysis. It does not require any specific distributional form of the independent variables and issues such as heteroscedasticity do not come into play as they did in discriminant analysis. Moreover, logistic regression does not require linear relationships between the independent variables and the dependent variables as does multiple regression. The method can address nonlinear effects even when exponential and polynomial terms are not explicitly added as additional independent variables because of the logistic relationship. Finally, it does require independence of observations and if present, this must be addressed by some form of hierarchical or clustered-sample approach as discussed in Chapter 5.

There is one implicit assumption underlying logistic regression that deserves attention—the **linearity of the logit** [20]. Just as we had in multiple regression, the transformed logistic model assumes a linear relationship between the logit and the independent variables, especially those of a continuous nature. In multiple regression a nonlinear relationship was not reflected in the regression coefficient and generally required transformations or polynomials to represent the nonlinearity (see Chapter 5 for more details). Likewise, logistic regression may encounter nonlinear relationships that diminish the ability of the coefficients to measure the variable's full impact. In logistic regression, however, determining the linearity is more difficult than seen in multiple regression since the scatterplots (either the bivariate scatterplots or residual plots) are not conducive to identifying nonlinearities in the relationships.

Given the inherent nonlinear relationship of the logistic function, what is the best method to examine nonlinearity? The simplest test is the Box-Tidwell procedure applied to each continuous independent variable. First, create an interaction term that is the product of the independent variable and its log value (independent variable \* log (independent variable)). The interaction term is similar to a polynomial term in regression. Then test for the significance of the interaction term when added to the model. A significant interaction indicates a nonlinear component and thus should be retained in the model.

## Logistic Regression Basics

Logistic regression is the preferred method for two-group (binary) dependent variables due to its robustness, ease of interpretation, and diagnostics.

Logistic regression has equally useful objectives in:

Explanation—providing estimates of the ability of a set of independent variables collectively and individual to distinguish between a binary outcome.

Classification—provide a means for classification of cases into the outcome groups and provide a range of diagnostic measures of predictive accuracy.

The logistic function is a means of directly analyzing an inherently nonlinear relationship between estimated probabilities and a binary outcome of either 0 or 1.

## Sample Size

Overall sample size should be 400 to achieve best results with maximum likelihood estimation, thus use with smaller samples sizes should be aware of less efficiency in estimation the model coefficients.

Sample size considerations more focused on the size of each outcome group, which should have 10 times the number of estimated model coefficients:

Particularly problematic are situations where the actual frequency of one outcome group is small (i.e., below 20). The actual size of the small group is more problematic than the low rate of occurrence.

Several approaches have been proposed for addressing what may be termed “rare event” situations.

Sample size requirements should be met in both the analysis and the holdout samples.

While normally associated with data at the individual level, logistic regression can analyze aggregated data if all the independent variables are nonmetric.

## Assumptions

Logistic regression does not require the assumptions of normality and homoscedasticity seen in both multiple regression and discriminant analysis.

The primary assumption is the independence of observations, which if violated requires some form of hierarchical/nested model approach.

An inherent assumption that should be addressed with the Box–Tidwell test is the linearity of the independent variables, especially continuous variables, with the outcome.

## Stage 4: Estimation of the Logistic Regression Model and Assessing Overall Fit

One of the unique characteristics of logistic regression is its use of the logistic relationship described earlier in both estimating the logistic model and establishing the relationship between dependent and independent variables. This procedure involves a unique transformation of the dependent variable, which impacts not only the estimation process, but also the resulting coefficients for the independent variables. And yet logistic regression shares approaches to assessing overall model fit with both discriminant analysis (i.e., use of classification matrices) and multiple regression (i.e.,  $R^2$  measures). The following sections discuss the estimation process, problematic issues frequently encountered and the various ways in which model fit is evaluated.

## ESTIMATING THE LOGISTIC REGRESSION MODEL

Logistic regression has a single variate composed of estimated coefficients for each independent variable, as found in multiple regression. However, this variate is estimated in a different manner. Logistic regression derives its name from the **logit transformation** used with the dependent variable, creating several differences in the estimation process (as well as the interpretation process discussed in a following section).

**Transforming the Dependent Variable** As shown earlier, the logit model uses the specific form of the logistic curve, which is S-shaped, to stay within the range of 0 to 1. To estimate a logistic regression model, this curve of predicted values is fitted to the actual data, just as was done with a linear relationship in multiple regression. However, because the actual data values of the dependent variables can only be either 1 or 0, the process is somewhat different.

Figure 8.2 portrays two hypothetical examples of fitting a logistic relationship to sample data. The actual data represent whether an event either happened or not by assigning values of either 1 or 0 to the outcomes (in this case a 1 is assigned when the event happened, 0 otherwise, but they could have just as easily been reversed). Observations are represented by the dots at either the top or bottom of the graph. These outcomes (happened or not) occur at each value of the independent variable (the X axis). In part (a), the logistic curve cannot fit the data well, because a number

X	Y
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1

**Figure 8.2**  
Examples of Fitting the Logistic Curve to Sample Data

X	Y
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
5.5	1
6.0	1
6.5	1
7.0	1
7.5	1
8.0	1
8.5	1
9.0	1
9.5	1
10.0	1

Copyright 2019 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

of values of the independent variable have both outcomes (1 and 0). In this case the independent variable does not distinguish between the two outcomes, as shown by the high overlap of the two groups. As a result, the logistic curve is “flatter,” indicating that across the range of the independent variables, the predicted probabilities ( $Y$  axis) were not that different in value. This corresponds to a regression line with minimal slope. In this situation we may find that the logistic curve provides insufficient ability to distinguish between 1 and 0. In that case, this independent variable could be found to have a nonsignificant relationship with the outcome.

However, in part (b), a much more well defined relationship is based on the independent variable. Lower values of the independent variable correspond to the observations with 0 for the dependent variable, whereas larger values of the independent variable correspond well with those observations with a value of 1 on the dependent variable. Thus, the logistic curve should be able to fit the data quite well. That is, there is a well-defined logistic curve that has both very high and very low probabilities, another indication of a strong relationship with the outcome.

But how do we predict group membership from the logistic curve? For each observation, the logistic regression technique predicts a probability value between 0 and 1. Plotting the predicted values for all values of the independent variable generates the curve shown in Figure 8.2. This predicted probability is based on the value(s) of the independent variable(s) and the estimated coefficients. Then, a cutoff value is selected (much more discussion on selecting the cut-off value is in the discussion of predictive accuracy). If the predicted probability is greater than the cut-off (e.g., generally .50 when the two groups are of equal size), then the prediction is that the outcome is 1 (the event happened); otherwise, the outcome is predicted to be 0 (the event did not happen). Let’s return to our example and see how it works.

In parts (a) and (b) of Figure 8.2, suppose that a cut-off value of 6.0 on the  $X$  axis (the independent variable) has been selected. In part (a), we can see that a number of observations of both groups fall on both sides of this value, resulting in a number of misclassifications. Thus, the predicted probabilities do very poorly in distinguishing between an event versus a non-event. In part (b), we make much better classifications (i.e., only a few misclassifications for both groups) when using the cut-off value of 6.0.

Thus, with an estimated logistic curve we can estimate the probability for any observation based on its values for the independent variable(s) and then predict group membership using a specified cut-off value. Once we have the predicted membership, we can create a classification matrix just as was done for discriminant analysis and assess predictive accuracy.

**Estimating the Coefficients** Where does the curve come from? In multiple regression, we estimate a linear relationship that best fits the data. In logistic regression, we follow the same process of predicting the dependent variable by a variate composed of the **logistic coefficient(s)** and the corresponding independent variable(s). What differs is that in logistic regression the predicted values can never be outside the range of 0 to 1. Although a complete discussion of the conceptual and statistical issues involved in the estimation process is beyond the scope of this text, several excellent sources with complete treatments of these issues are available [9, 24, 30, 2, 20, 16]. We can describe the estimation process in two basic steps as we introduce some common terms and provide a brief overview of the process.

**Transforming a Probability into Odds and Logit Values** Just as with multiple regression, logistic regression predicts a metric dependent variable, in this case probability values constrained to the range between 0 and 1. But how can we ensure that estimated values do not fall outside this range? The logistic transformation accomplishes this process in two steps. But before we discuss these transformations, let’s revisit how the probabilities are calculated.

**CALCULATING PROBABILITIES** As we illustrated in the introduction at the beginning of the chapter, we represent probabilities as the percentage of events,  $\text{Prob} = \text{Number of events} \div (\text{Number of events} + \text{Number of non-events})$ . So we can easily calculate the overall probability of an event across the sample as simple the percentage of observations with an event. At the individual observation level it is either 100 percent (the event occurred) or 0 percent (the event did not occur).

When we want to examine the relationship of an independent variable to this probability, we compute the probability for each subgroup of observations with different values of the independent variable (e.g., probability for males

at 15% versus females at 40%). Now, observations that are male have the associated probability of 15 percent versus observations of females at 40 percent. This continues as we add additional independent variables, similar to what we described when using aggregated data, but here for individual cases.

**RESTATING A PROBABILITY AS ODDS** In their original form, predicted probabilities are not constrained to values between 0 and 1. So, what if we were to restate the probability in a way that the new variable would always fall between 0 and 1? We restate it by expressing a probability as *odds*—the ratio of the probability of the two outcomes or events ( $Odds_i = Prob_i \div (1 - Prob_i)$ ). Odds are also many times stated as the relative frequency of events. For example, the odds of an event happening is the number of events divided by the number of non-events.

Let us use an example of the probability of success or failure to illustrate how the odds are calculated. In this example we have eight successes and two failures. We know that the probability of success is .80 and that the probability of the alternative outcome (i.e., failure) is .20 (.20 = 1.0 – .80). We can then state this as odds with either the number of events (many times stated as 8 to 2) or as the ratio of probabilities (.80 ÷ .20). In either case we know that the odds of success are 4.0, or that success is four times more likely to happen than failure. Conversely, we can state the odds of failure as .25 (.20 ÷ .80 or 2 to 8) meaning that failure happens at one-fourth the rate of success. Thus, no matter which outcome we look at (success or failure), we can state the probability as odds and vice versa, as shown below:

$$\text{Odds} = \frac{P}{1 - P}$$

or

$$P = \frac{\text{Odds}}{1 + \text{Odds}}$$

Using odds, any probability value is now stated in a metric variable that can be directly estimated. Any odds value can be converted back into a probability that falls between 0 and 1. As you can probably surmise, a probability of .50 results in odds of 1.0 (both outcomes have an equal chance of occurring). Odds less than 1.0 represent probabilities less than .50 and odds greater than 1.0 correspond to a probability greater than .50. Figure 8.3 illustrates a range of probabilities from zero to one and their associated odds. We have solved our problem of constraining the predicted values to within 0 and 1 by predicting the odds value and then converting it into a probability with a value of 0 to 1.

**CALCULATING THE LOGIT VALUE** The odds variable solves the problem of making probability estimates between 0 and 1, but we have another problem: How do we keep the odds values from going below 0, which is the lower limit of the odds (there is no upper limit). The solution is to compute what is termed the *logit value*, which is calculated by taking the logarithm of the odds. Odds less than 1.0 will have a negative logit value, odds ratios greater than 1.0 will have positive logit values, and the odds ratio of 1.0 (corresponding to a probability of .5) has a logit value of 0 (see Figure 8.3). Moreover, no matter how low the negative value gets, it can still be transformed by taking the antilog into an odds value greater than 0.

Probability	Odds <sup>a</sup>		Log Odds (Logit)
	Frequency	Value	
.00	0/10	.00	NC
.10	1/9	.111	-2.197
.30	3/7	.428	-.847
.50	5/5	1.000	.000
.70	7/3	2.333	.847
.90	9/1	9.000	2.197
1.00	10/0	NC	NC

**Figure 8.3**  
Correspondence of Probability, Odds  
and Log Odds

<sup>a</sup>Assuming 10 cases.

NC = Cannot be calculated.

With the logit value, we now have a metric variable that can have both positive and negative values but that can always be transformed back to a probability value that is between 0 and 1. This value now becomes the dependent variable of the logistic regression model. Note, however, that the logit, when converted back to a probability, can never actually reach either 0 or 1. We will see this problem in the next section when we discuss separation, which is the situation where we have a probability of either zero or one that cannot be transformed into a logit value.

**Model Estimation** Once we understand how to interpret the values of either the odds or logit measures, we can proceed to using them as the dependent measure in our logistic regression. The process of estimating the logistic coefficients is similar to that used in regression, although instead of using ordinary least squares as a means of estimating the model, the maximum likelihood method is used.

**USING MAXIMUM LIKELIHOOD FOR ESTIMATION** Multiple regression employs the method of least squares, which minimizes the sum of the squared differences between the actual and predicted values of the dependent variable. The nonlinear nature of the logistic transformation requires that another procedure, the **maximum likelihood procedure**, be used in an iterative manner to find the most likely estimates for the coefficients. Instead of minimizing the squared deviations (least squares), logistic regression maximizes the likelihood that an event will occur. The likelihood value, instead of the sum of squares, is then used when calculating a measure of overall model fit. Using this alternative estimation technique also requires that we assess model fit in different ways, as will be discussed in the next section.

**ESTIMATING THE COEFFICIENTS** The estimated coefficients for the independent variables are estimated using the logit value as the dependent measure to ensure that any predicted value (i.e., any logit value) can be transformed back to a probability that falls within zero and one. The model formulation is:

$$\text{Logit}_i = \ln\left(\frac{\text{prob}_{\text{event}}}{1 - \text{prob}_{\text{event}}}\right) = b_0 + b_1X_1 + \cdots + b_nX_n$$

As we will discuss in Stage 5, however, the coefficients estimated in this model formulation relate to impacts on a logged odds value, something that may be difficult to interpret. So we will see that if we transform this model formulation (see model formulation below) we can arrive at model coefficients that relate to changes in odds, which are more easy to interpret:

$$\text{Odds}_i = \left(\frac{\text{prob}_{\text{event}}}{1 - \text{prob}_{\text{event}}}\right) = e^{b_0 + b_1X_1 + \cdots + b_nX_n}$$

Both model formulations are equivalent, but whichever is chosen affects how the coefficients are interpreted. Many software programs provide the logistic coefficients in both forms, so the researcher must understand how to interpret each form. We will discuss interpretation issues in a later section.

This process can accommodate one or more independent variables, and the independent variables can be either metric or nonmetric (binary). As we will see later in our discussion of interpreting the coefficients, both forms of the coefficients reflect both direction and magnitude of the relationship, but are interpreted differently.

**Issues in Model Estimation** As discussed earlier, maximum likelihood estimation encounters estimation problems whenever the sample size in any group becomes very small. This can be encountered whenever one of the outcome groups is very small (i.e., termed rare events) or, when adding independent variables we create subgroups with very small numbers of observations. In these instances the estimation procedure may encounter problems generating accurate parameter estimates or standard errors.

But even more problematic are subgroups that have no observations (i.e., the probability of the event is zero). This is because, as we saw earlier, probabilities of zero (and one), when transformed to logit values, are values that cannot be calculated. So when the estimation procedure encounters these values, it cannot proceed in many instances. This problem is termed separation and can be seen in two forms discussed below.

**COMPLETE SEPARATION** The most problematic form of separation is **complete separation**, where the dependent variable is perfectly predicted by an independent variable. This would be a scatterplot (like those shown in Figure 8.2) where there was no overlap for the outcomes—i.e., they were completely separate on the independent variable. Now this might seem like an ideal independent variable, but remember that we predict logit values and when we have complete separation, all of the logit values for that variable are undefined since they are either zero or one. Figure 8.4 contains a simple cross-tabulation for a dichotomous independent variable with the outcome. Here we see that a value of zero on the independent variable is always associated with a one on the dependent variable, and vice versa for a value of 1 for the independent variable.

This is most often encountered with nonmetric independent variables, such as the example where males all had one outcome and females had the other outcome. But it could also found with continuous variables if one group had all values below a certain value and the other group had all values above that value (i.e., again, no overlap between groups on the independent variable). For example, distinguishing between adults and their pets based on weight could easily result in all of the adult weights being greater than the largest weight of any pet, thus complete separation.

When complete separation is encountered, the resulting model estimates for the offending variable are not possible. The only remedy is to collapse categories for multi-category variables (if possible) so that all of the probabilities are not just zero and one. If that is not possible, such as the case with dichotomous independent variables or continuous variables, then the variable should be eliminated from the analysis.

**QUASI-COMPLETE SEPARATION** A more frequently encountered form of separation is **quasi-complete separation**, where one or more of the groups defined by the nonmetric independent variable have counts of zero. An example is also shown in Figure 8.4. Many times termed the “zero cells” effect, it also presents a problem for the estimation procedure. In the case of quasi-complete separation, the estimation procedure may be completed, but many times the estimated parameter and standard error for the offending variable are very large (parameter estimates of five or more and standard errors much larger).

There are several remedies for addressing quasi-complete separation [17]. First, if possible, collapse categories and eliminate the zero cells. There are also alternative estimation procedures, such as the penalized methods [22, 11] or exact logistic models that were discussed earlier when addressing small cell counts. Finally, even though the estimated coefficient and standard error cannot be interpreted, retaining it in the model does not impact the other parameter estimates. For example, if it is deemed necessary for the offending variable to be included in the model, the other parameter estimates can still be reported and interpreted.

**Alternative Models: The Probit Model** It should be noted that there is an alternative model form to the logistic model that is almost equivalent. The **probit** model is a model form that is essentially equivalent to the logistic model in most instances of a binary outcome variable and its advantages become more apparent when the model is extended to a multi-category outcome measure. For binary outcomes logistic models are generally preferred because the model coefficients are somewhat easier to interpret. But many times the choice of logit or probit is dependent on the field of study and preferences established in that area of research [7].

**Figure 8.4**

Examples of Complete and Quasi-separation

Complete Separation			Quasi-Separation				
		Dependent Variable			Dependent Variable		
		0	1			0	1
Independent Variable	0	0	50	Independent Variable	0	10	40
	1	50	0		1	50	0

## Model Estimation

Uses the logged odds values instead of probabilities as the dependent value.

Maximum likelihood provides ability to estimate the logistic function that reflects the nonlinear relationship between predicted probabilities and the outcomes of 0 and 1.

For model estimation to work there must be some “overlap” of observations across values of the independent variable:

Complete separation – independent variable perfectly predicts outcome, yet cannot be included in model.

Quasi-complete separation – at least one cell created by addition of nonmetric variable has zero observations. Can be remedied by collapsing variable categories or specialized estimation methods.

Probit is an alternative to the logistic function with generally equivalent results. Although interpretation of probit coefficients is more difficult, it does perform well with multi-category outcome measures.

## ASSESSING THE GOODNESS-OF-FIT OF THE ESTIMATED MODEL

Before we can proceed to Stage 5 and model interpretation, the overall goodness-of-fit for the logistic regression model must be assessed and deemed acceptable. There are two primary methods for the evaluation of model fit. The first is to use an overall measure of statistical significance of the model fit and also “pseudo”  $R^2$  values, similar to that found in multiple regression. The second approach is to examine predictive accuracy where the ability of the model to correctly classify the outcome measure is computed in what is termed a classification matrix. The two approaches examine model fit from different perspectives, but should yield similar conclusions.

**Model Estimation Fit** The basic measure of how well the maximum likelihood estimation procedure fits is the **likelihood value**, similar to the sums of squares values used in multiple regression. Logistic regression measures model estimation fit with the value of  $-2$  times the log of the likelihood value, referred to as  $-2LL$  or  $-2 \log \text{likelihood}$ . The minimum value for  $-2LL$  is 0, which corresponds to a perfect fit ( $\text{likelihood} = 1$  and  $-2LL$  is then 0). Thus, the lower the  $-2LL$  value, the better the fit of the model. As will be discussed in the following section, the  $-2LL$  value can be used to compare equations for the change in fit or to calculate measures comparable to the  $R^2$  measure in multiple regression.

**BETWEEN MODEL COMPARISONS** The likelihood value can be compared between equations to assess the difference in predictive fit from one equation to another, with statistical tests for the significance of these differences. The basic approach follows three steps:

- 1 *Estimate a null model.* The first step is to calculate a null model, which acts as the baseline for making comparisons of improvement in model fit. The most common null model is one without any independent variables, which is similar to calculating the total sum of squares using only the mean in multiple regression. The logic behind this form of null model is that it can act as a baseline against which any model containing independent variables can be compared.
- 2 *Estimate the proposed model.* This model contains the independent variables to be included in the logistic regression model. Hopefully, model fit will improve from the null model and result in a lower  $-2LL$  value. Any number of proposed models can be estimated (e.g., models with one, two, and three independent variables can all be separate proposed models).
- 3  *$-2LL$  difference.* The final step is to assess the statistical significance of the  $-2LL$  value between the two models (null model versus proposed model). If the statistical tests support significant differences, then we can state that the set of independent variable(s) in the proposed model is significant in improving model estimation fit and that model as a whole is statistically significant.

We should note that any two proposed models can be compared. In these instances, the  $-2LL$  difference reflects the difference in model fit due to the different model specifications. For example, a model with two independent variables may be compared to a model with three independent variables to assess the improvement gained by adding one independent variable. In these instances, one model is selected to act as the null model and then compared against another model.

The chi-square test and the associated test for statistical significance are used to evaluate the reduction in the log likelihood value. However, these statistical tests are particularly sensitive to sample size (for small samples it is harder to show statistical significance, and vice versa, for large samples). Therefore, researchers must be particularly careful in drawing conclusions based solely on the significance of the chi-square test in logistic regression.

**GLOBAL NULL HYPOTHESIS TEST** In addition to testing for the difference from a base model, there is also a test for the significance of any of the estimated coefficients. This test, if significant, reflects that at least one of the estimated coefficients is significant. This is similar in nature to the overall model  $F$  test in multiple regression.

**PSEUDO  $R^2$  MEASURES** In addition to the statistical chi-square tests, several different "R<sup>2</sup>-like" measures have been developed and are presented in various statistical programs to represent overall model fit. These pseudo  $R^2$  measures are interpreted in a manner similar to the coefficient of determination in multiple regression. A **pseudo  $R^2$**  value can be easily derived for logistic regression similar to the  $R^2$  value in regression analysis [13]. The pseudo  $R^2$  for a logit model ( $R^2_{\text{LOGIT}}$ ) can be calculated as:

$$R^2_{\text{LOGIT}} = \frac{-2LL_{\text{null}} - (-2LL_{\text{model}})}{-2LL_{\text{null}}}$$

Just like its multiple regression counterpart, the logit  $R^2$  value ranges from 0.0 to 1.0. As the proposed model increases model fit, the  $-2LL$  value decreases. A perfect fit has a  $-2LL$  value of 0.0 and a  $R^2_{\text{LOGIT}}$  of 1.0.

Two other measures are similar in design to the pseudo  $R^2$  value and are generally categorized as pseudo  $R^2$  measures as well. The Cox and Snell  $R^2$  measure operates in the same manner, with higher values indicating greater model fit. However, this measure is limited in that it cannot reach the maximum value of 1, so Nagelkerke proposed a modification that had the range of 0 to 1. Both of these additional measures are interpreted as reflecting the amount of variation accounted for by the logistic model, with 1.0 indicating perfect model fit.

We should note that the different pseudo  $R^2$  measures vary widely in terms of magnitude and no one version has been deemed most preferred. For all of the pseudo  $R^2$  measures, however, the values tend to be much lower than for multiple regression models. This results from the outcome variable since we are trying to predict values that are only zero or one with probability values. Thus the analyst should always be aware of these limitations in using the pseudo- $R^2$  measure as an evaluation of overall model fit.

**A COMPARISON TO MULTIPLE REGRESSION** In discussing the procedures for assessing model fit in logistic regression, we made several references to similarities with multiple regression in terms of various measures of model fit. In Figure 8.5, we compare the concepts used in multiple regression and their counterparts in logistic regression.

As we can see, the concepts between multiple regression and logistic regression are similar. The basic approaches to testing overall model fit are comparable, with the differences arising from the estimation methods used in the two techniques.

**Figure 8.5**

Comparing the Primary Elements of Model Fit Between Multiple and Logistic Regression

Multiple Regression	Logistic Regression
Total sum of squares	$-2LL$ of base model
Error sum of squares	$-2LL$ of proposed model
Regression sum of squares	Difference of $-2LL$ for base and proposed models
$F$ test of model fit	Chi-square test of $-2LL$ difference
Coefficient of determination ( $R^2$ )	Pseudo $R^2$ measures

**Predictive Accuracy** Just as we borrowed the concept of  $R^2$  from regression as a measure of overall model fit, we will also use a measure from discriminant analysis (classification matrix) to assess overall predictive accuracy along with a chi-square-based measure of fit.

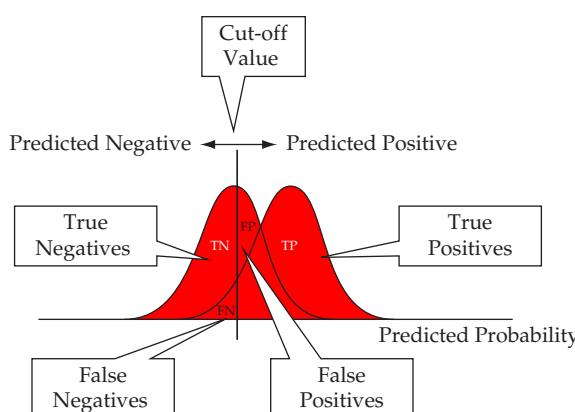
**SELECTING THE CUT-OFF VALUE** With logistic regression the analyst has the capability to select the cutoff value that maximizes the type of predictive accuracy to meet the research question. If misclassification costs are equal (or unspecified), then a misclassification of a positive outcome is just as impactful as a negative outcome misclassification. So the analyst can select the cutoff value that maximizes the overall predictive accuracy or the predictive accuracy of either positive or negative outcomes.

The default cut-off value is .50, which is applicable in many situations, especially when the group sizes of the dependent variable are equal. But when group sizes differ and/or research objectives dictate, the analyst may want to see if other cut-off values result in predictive accuracy more closely aligned with research objectives (e.g., emphasize correctly predicting positive outcomes or minimizing the error rate in positive predictions).

Whatever cut-off value is selected, the process of classification is straightforward. As depicted in Figure 8.6, all of the observations can be arrayed on their predicted probability. The cut-off value is specified and then all observations above the cut-off value are classified as positive and those below the cut-off as negative. Since we know the actual group membership, we can examine each outcome group and see how well it was predicted. In Figure 8.6 the positive outcomes are shown by the distribution to the right. For positive outcomes that were predicted to be positive, they are referred to as **true positives**. For the positive outcomes classified as negative, they are referred to as **false negatives**. Turning our attention to the actual negative outcomes, those that are correctly classified as negative (i.e., to the left of the cut-off value) are **true negatives** and the incorrectly classified negative outcomes are **false positives**.

As we can see from Figure 8.6, as the cut-off value is either increased or decreased, the percentage of cases in each of the four groups changes. For example, we can increase the percentage of True Positives, but we then also increase False Positives. The actual shapes of the distributions for each group dictate the magnitude of these trade-offs. What is needed are measures which reflect these tradeoffs and we will see that the classification matrix is a simple method for detailing each of these different predicted outcomes and that there are specific measures to describe each.

**CLASSIFICATION MATRIX** This classification matrix approach is simple in design—a cross-tabulation of the outcome variable with the predicted outcome. It measures how well group membership is predicted (both events and non-event) as well as the misclassifications (e.g., events predicted to be non-events). Figure 8.7 provides an example of the classification matrix along with the four groups we described earlier (true positives, true negatives, false positives and false negatives). We can also see that the column totals provide the total number of predicted outcomes for each group, while the row totals are the total number of actual outcomes. It should be noted that there will be different classification matrix values (true positives, true negatives, etc.) for each cut-off value used. Note that in the following discussion we will use the term Positive to represent the event of interest (dependent measure equals 1) and negative



**Figure 8.6**  
Applying the Cut-off Value for Making Predictions

to be a non-event (dependent measure equals zero). This is done solely for simplicity in comparing and contrasting the different measures.

**MEASURES OF PREDICTIVE ACCURACY** With the classification matrix complete, we can now calculate three types of measures—overall predictive accuracy, predictive accuracy of the actual outcomes and predictive accuracy of the predicted outcomes. The specific measures for each type, their calculations and description are shown in Figure 8.8 and discussed in the sections below.

**Overall Predictive Accuracy** The first type of predictive accuracy relates to the overall model, combining both positive and negative outcomes into a single measure. **Accuracy** is the number of true positive and true negatives divided by the total sample—the percentage of total cases correctly classified. In assessing discriminant analysis, this was termed the **hit ratio** or **percentage correctly classified**. While it does not distinguish between predictive accuracy of positives or negatives, it is a single composite reflecting overall model predictive accuracy.

**Youden index** A second measure is the **Youden index**, which is a combination of the True Positive rate and the True Negative rate minus 1. As we will discuss in the next section, the True Positive rate is termed Sensitivity and the True Negative Rate is termed Specificity. The maximum value is 1.0, with higher values indicating better overall predictive accuracy. Note that it can vary from the accuracy measure since it is the sum of the two rates (positive and negative) and does not account for different group sizes.

**Figure 8.7**  
Classification Table and Group Distributions Based on Predicted Probability

		Predicted Outcome		Total
		No	Yes	
Actual Outcome	No	True Negatives	False Positives	Total Actual No
	Yes	False Negative	True Positives	Total Actual Yes
	Total	Total Predicted No	Total Predicted Yes	Total Sample

**Figure 8.8**  
Measures of Predictive Accuracy: Overall, Actual Outcomes and Predicted Outcomes

Measure	Calculation	Description
<b>Overall Predictive Accuracy</b>		
Accuracy	$(TP + TN)/N$	Measure of overall classification accuracy for both states (Positive and Negative)
Youden Index (YI)	$SN + SP - 1$	Overall measure of predictive accuracy, with +1 as perfect prediction
<b>Predictive Accuracy of Actual Outcome</b>		
Sensitivity (SN)	$TP/(TP + FN)$	True Positive Rate – percentage of actual positives that are predicted as positive
Specificity (SP)	$TN/(TN + FP)$	True Negative Rate – percentage of actual negatives that are predicted as negative
<b>Predictive Accuracy of Predicted Outcome</b>		
Positive Predictive Value (PPV)	$TP/(TP + FP)$	Probability of Correct Positive Prediction – Percentage of positive predictions that are actually positive
Negative Predictive Value (NPV)	$TN/(TN + FN)$	Probability of Correct Negative Prediction – percentage of negative predictions that are actually negative
Legend: N = Total Sample TP = True Positives TN = True Negatives FP = False Positives FN = False Negatives		

As discussed in Chapter 7 on discriminant analysis, there are two chance-related standards of comparison for assessing our predictive accuracy value. The first is the **maximum chance criterion**, which is a naïve model—assign all cases to the largest group. For example, if the two groups were 75 percent and 25 percent, then the maximum chance criterion would be 75 percent. The **proportional chance criterion** is similar, but takes into account both groups. It is calculated as the sum of the percentage positive<sup>2</sup> plus the percentage negative<sup>2</sup>. So for our example of 75 percent and 25 percent, the proportional chance criterion would be  $.625 (.75^2 + .25^2 = .625)$ . These then become the values to which compare the model's accuracy value. Some analysts believe that the accuracy should exceed the chance-based measures by at least one-fourth. So in the case of a value of .625, the actual standard for comparison would be  $.78125 (.625 * 1.25)$ . Interested readers should refer to the more detailed discussion of these standards and others found in Chapter 7.

Even with the chance-based measures, acceptable levels of predictive accuracy are more within the judgment of the researcher versus some established standard. These measures do, however, provide an objective and consistent means of comparing different cutoff values in terms of their predictive accuracy.

**Predictive Accuracy of Actual Outcomes** Many times the focus of predictive accuracy is oriented toward the individual outcomes rather than the overall level of accuracy. To address this issue, measures of sensitivity and specificity have been developed (see Figure 8.8). **Sensitivity** is the true positive rate—the percentage of actual positive outcomes that were correctly predicted. **Specificity** is the true negative rate—the percentage of actual negative outcomes that were correctly predicted. So while accuracy was the overall level of predictive accuracy, sensitivity and specificity relate to levels of either the positive or negative outcomes.

The two measures now provide a means to understand the trade-offs between false positives and false negatives for any specific cut-off value. Increasing sensitivity works toward minimizing false negatives, and specificity does the same for false positives. From another perspective, a highly sensitive model rarely overlooks a positive (i.e., a high true positive rate) and a model with high specificity minimizes the false positives. While it would be useful to have high sensitivity and specificity, it is generally a trade-off between these measures that defines the final cut-off model. Note that the Youden index was calculated as the sum of sensitivity and specificity minus 1.0. Thus, maximizing the Youden index is preferred by some analysts as the appropriate balance between the two measures.

**Predictive Accuracy of Predicted Outcomes** Up to this point we have focused on the predictive accuracy of predicting the actual outcomes, either positive or negative. But we can also focus on the predictive accuracy of the prediction—of those we predicted as positive, how many were positive, and of those we predicted as negative how many were negative predictions. This differs from sensitivity and specificity in that the denominator is now the number of predictions, either positive or negative, versus the actual number of positive or negative cases. The **positive prediction value (PPV)** is the number of correct positive predictions divided by the total number of positive predictions (i.e., the true positives and the false positives). Likewise, the **negative prediction value (NPV)** is the number of correct negative predictions divided by the total number of negative predictions. Figure 8.8 contains the calculations and descriptions of these two measures.

The PPV and NPV allow the analyst to address slightly different questions—if you test positive or negative, how likely is it that the test is correct. In many situations, such as testing for medical conditions, emphasis may be on maximizing the PPV so as to not cause undue concern with a false diagnosis. But in doing so there is also the trade-off that this most likely results in an increase in false negatives—missing the condition if it is present. Many research questions, upon closer inspection, involve these types of tradeoffs between positive and negative outcomes. The concept of **misclassification costs** (the cost of a false positive or a false negative) is an approach to try and quantify these tradeoffs and assist in determining the right balance between PPV and NPV.

**Summary** The wide range of measures of predictive accuracy discussed above provide the analyst with a number of perspectives from which to evaluate a model, or more likely, compare between models. Figure 8.9 provides a simple illustration of the calculation of each of these measures of predictive accuracy. For example, this model and cut-off

**Figure 8.9**

Calculating Predictive Accuracy Measures for Actual and Predicted Outcomes

		Predicted		
		0	1	
Actual	0	20	3	Specificity = $20/(3 + 20) = .870$
	1	5	22	Sensitivity = $22/(22 + 5) = .815$
Negative Predictive Value:		Positive Predictive Value:		Accuracy: $(22 + 20)/50 = .840$
$20/(5 + 20) = .80$		$22/(22 + 3) = .880$		

value result in a better true negative rate (specificity) of .870 compared to the true positive rate (sensitivity) of .815. However, if we evaluate the PPV and NPV values, we see that there is a higher PPV value (.880), which indicates that if a prediction is positive, it is correct 88 percent of the time.

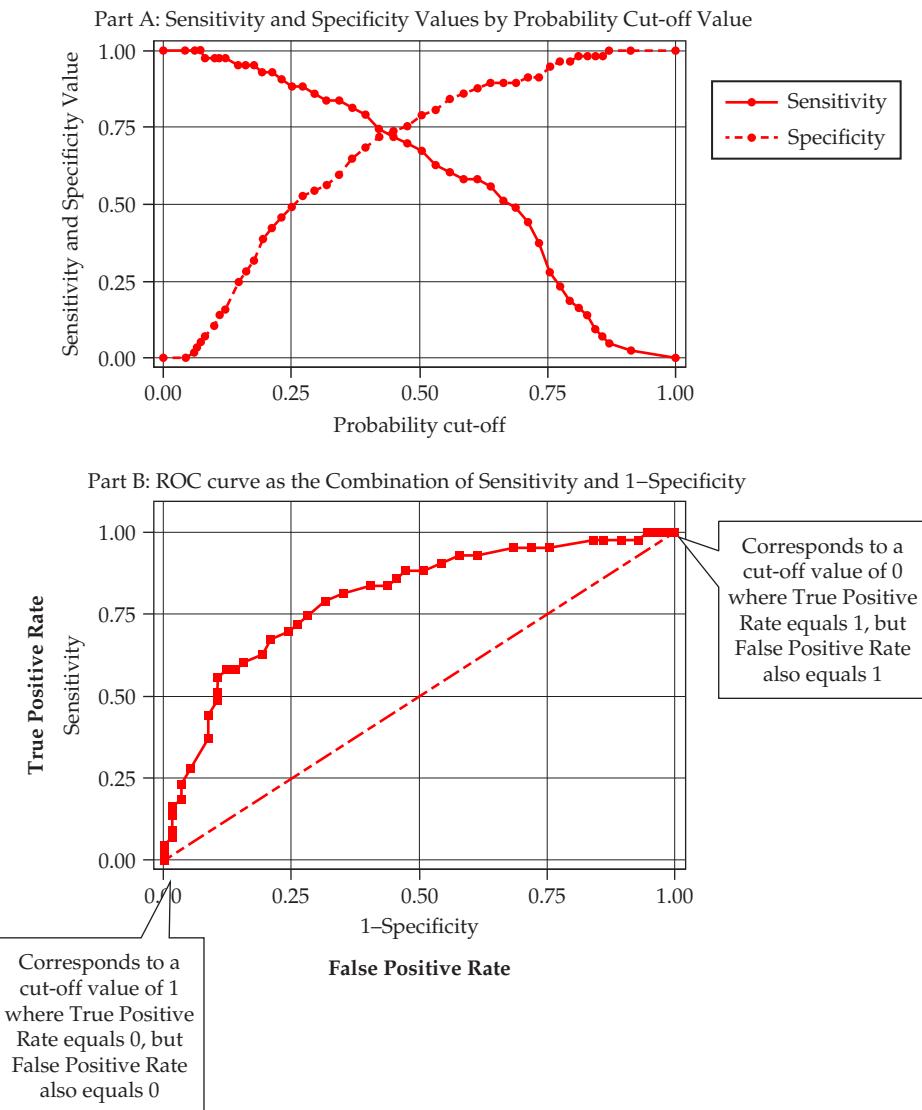
For each different cut-off value a new classification matrix and resulting set of predictive accuracy measures will be calculated. Once the appropriate measures have been selected, these measures can then be compared between several different cut-off values, to select the cut-off value which best meets the research objectives.

**ROC CURVE** We have discussed the trade-off between sensitivity and specificity as being a key comparison standard for predictive accuracy, plus the need for comparing between models to determine, if possible, the best trade-off to meet the research objectives. The **ROC curve** was developed to provide a graphical representation of this trade-off across the entire range of cut-off values and illustrate how well a model simultaneously predicts both positives and negatives. Developed initially in World War II to evaluate the ability of sonar to distinguish signal from noise (hence its name Receiver Operating Characteristic), the ROC curve plots sensitivity on the Y axis and 1 – specificity on the X axis. The result is, for each possible cut-off value, the true positive rate versus the false positive rate.

To illustrate how the ROC curve portrays the trade-off of sensitivity versus specificity, Figure 8.10 provides the sensitivity and specificity values across cut-off values ranging from 0 to 1 (see Part A) and the resulting ROC curve (Part B). In Part A we can see the trade-off between sensitivity and specificity as the cut-off value changes. This is most easily seen at the two extreme cut-off values, 0 and 1. At the far left of Part A we see that a cut-off value of 0 where all observations are classified as positive. This results in a sensitivity value of 1 (i.e., all positive observations are correctly classified), but also a sensitivity of 0 since no observations are classified as negative. The result on the ROC curve, shown in the upper right corner, is the point where sensitivity equals 1 and 1 – specificity equals 1. To the other extreme of a cut-off value of 1 (far right of Part A) we have all observations classified as negative. At this cut-off value the sensitivity is zero (no positives classified as positive) and the specificity is 1 (all the negatives are correctly predicted). The result is the point at the bottom left of the ROC curve, where the sensitivity is zero and 1 – specificity is also 0. All of the points in between are the sensitivity versus 1 – specificity values for the various cut-off values between 0 and 1.

So how do we interpret the ROC curve? First, the diagonal line across the ROC curve represents a null model, one predicting equally to chance. This line represents the lower bound of acceptability since we would never want to do worse than random chance and the objective is to be above this diagonal line as much as possible. The most preferable cut-off values are ones that move into the upper left of the chart, where the sensitivity (i.e., true positive rate) is high and the 1 – specificity (i.e., false negative rate) is low. The maximum Youden index, which combines sensitivity and specificity, corresponds to the cut-off value where the ROC curve is highest above the diagonal line. An excellent review of ROC curves and their interpretation is provided by [10].

Finally, the **AUC (area under the curve)** provides a value between 1 (perfect prediction) and .5 (a test no different from random chance). The AUC is an overall test of predictive accuracy, combining both the sensitivity and specificity across the entire range of cut-off values that might be used with a model. Typically values above .90 are considered excellent and decreasing in quality of predictive accuracy until reaching the minimum of .5. The AUC is particularly useful in comparing between different models, as illustrated in Figure 8.11. Models with the higher AUC



**Figure 8.10**  
Constructing and Interpreting a ROC Curve

Source: Adapted from [25].

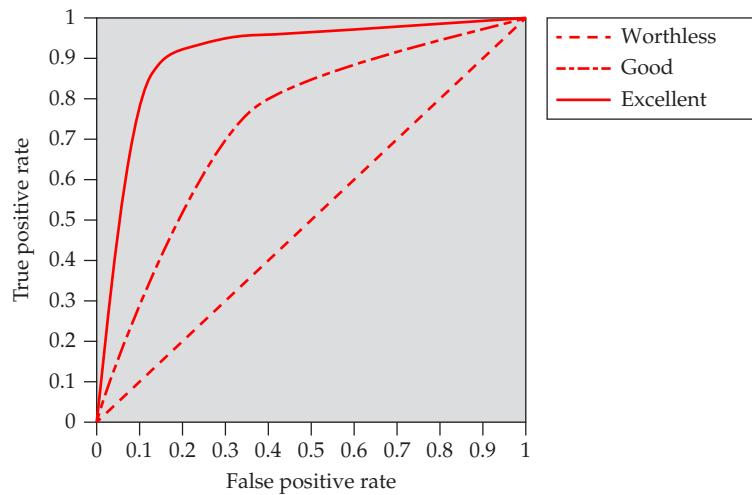
values are those extending towards the upper left portion of the graph and the desirable combinations of sensitivity and specificity.

**CONCORDANCE** A complementary measure to AUC is the **concordance (or c) statistic**, which is the probability that a randomly selected observation with a positive outcome will have a higher predicted probability than a randomly selected observation with a negative outcome. It is also comparable to a rank correlation between predicted probabilities of the outcome occurring and the observed response. Concordance is calculated by taking all possible pairs of observations of different outcomes (i.e., one outcome is positive and the other is negative). Then the percentage of pairs with a higher predicted probability of positive outcome versus negative outcome is the c statistic. That is, all subsets of the sample consisting of one subject who experienced the event of interest and one subject who did not experience the event of interest. The c-statistic is the proportion of such pairs in which the subject who experienced the event had a higher predicted probability of experiencing the event than the subject who did not

experience the event. The c statistic ranges from 1 (perfect prediction of positive outcomes) to zero. A value below 0.5 indicates a very poor model, with no better than chance in predicting the outcome. Values over 0.7 indicate a good model and values over 0.8 indicate a strong model [3]. We should note that the AUC is directly comparable to the c statistic as well ( $AUC = \text{percent concordant} + (.5 * \text{percent tied})$ ).

**CHI-SQUARE-BASED MEASURE** All of the measures of predictive accuracy discussed to this point have been measures derived from the classification matrix or model likelihood values. But none of them have had an associated statistical test of significance. **Hosmer and Lemeshow test** is a classification test of the statistical significance of the actual versus predicted outcomes [21]. Figure 8.12 illustrates application of the test to a sample of 200 observations where an event is an outcome of 1 and a non-event is an outcome of 0. The test consists of five steps:

- 1 **Divide subjects into deciles based on predicted probabilities.** The first decile contains the 10 percent of observations with lowest predicted probabilities, second decile the next lowest 10 percent of predicted



**Figure 8.11**  
Comparing Three Models with ROC Curves

**Figure 8.12**  
Partition Table for the Hosmer and Lemeshow Test

Group	Total	Average Predicted Probabilities		Outcome = 1		Outcome = 0	
		Outcome = 1	Outcome = 0	Observed	Expected	Observed	Expected
1	20	1.70%	98.30%	0	0.34	20	19.66
2	20	6.25%	93.75%	0	1.25	20	18.75
3	20	14.45%	85.55%	1	2.89	19	17.11
4	20	23.35%	76.65%	8	4.67	12	15.33
5	20	35.50%	64.50%	11	7.1	9	12.9
6	20	46.95%	53.05%	10	9.39	10	10.61
7	20	57.05%	42.95%	9	11.41	11	8.59
8	20	70.70%	29.30%	12	14.14	8	5.86
9	20	81.55%	18.45%	17	16.31	3	3.69
10	20	92.50%	7.50%	18	18.5	2	1.5

probabilities and so on. In our example, the 200 observations are divided into 10 groups of 20 each, with the first decile the 20 observations with the lowest predicted probabilities of an event.

- 2 **Compute the actual/observed events and non-events in each decile.** In our example the first decile had all 20 observations with a non-event. In the tenth decile, 18 observations had an event and 2 had a non-event.
- 3 **For each decile calculate the average predicted probabilities for both outcomes.** The average probability of an event and a non-event within each decile will sum to one. In our example the first decile had an average predicted probability of an event of 1.7 percent and an average predicted probability of a non-event of 98.3. The tenth decile had an average event probability of 92.5 percent and a non-event probability of 7.5 percent.
- 4 **Calculate the expected events and non-events in each decile.** The expected values are the decile average probabilities times the decile size, in this case 20 observations. So in the first decile the expected number of events is .34 ( $20 \times 1.7\%$ ) and the expected number of non-events is 19.66 ( $20 \times 98.3\%$ ).
- 5 **Compare the actual/observed and expected number of events and non-events with the chi-square statistic.** This test provides a comprehensive measure of predictive accuracy that is based not on the likelihood value, but rather on the actual prediction of the dependent variable. A nonsignificant value indicates a well-fitting model between actual and predicted events.

The appropriate use of this test requires a sample size of at least 50 cases to ensure that each class has at least five observations and generally an even larger sample because the number of predicted events should never fall below one. Also, the chi-square statistic is sensitive to sample size, enabling this measure to find small statistically significant differences when the sample size becomes large.

## OVERVIEW OF ASSESSING MODEL FIT

Logistic regression provides two approaches to assessing model fit—a model-based measure based upon the statistical process of model estimation and case-based measures of predictive accuracy derived from comparison of actual and predicted outcomes for each case. The model-based approach provides statistical tests of significance for the model overall and measures somewhat comparable to  $R^2$  from multiple regression. The case-based approaches are similar to those also used in discriminant analysis, although they provide more diagnostic perspectives (e.g., predictive accuracy of actual events versus predictive accuracy of predictions) than usually found in discriminant analysis. The analyst should consider both approaches for the perspectives they provide and hopefully a convergence of indications from these measures will provide the necessary support for the researcher in evaluating the overall model fit.

## CASEWISE DIAGNOSTICS

One of the benefits of logistic regression is its similarity to multiple regression in many areas, such as model specification and interpretation of the model coefficients discussed in the next section. An additional similarity is the availability of casewise diagnostics similar to those found in multiple regression, such as residuals and influential measures. Anyone using these diagnostics in multiple regression will make an easy transition to the measures provided by logistic regression. As such, we will refer the reader to Chapter 5 for a more detailed description of the basic concepts and approaches in using casewise diagnostics. Our discussion here will highlight the similarities and differences in these two types of measures—residuals and influential measures—that arise when having the unique outcome measure in logistic regression and the use of maximum likelihood estimation instead of ordinary least squares.

**Residuals** Residuals in logistic regression are similar to those found in multiple regression in that they represent the difference between the predicted and actual value of the dependent measure, but in logistic regression the binary dependent measure necessitates two forms of residuals. **Pearson residuals** are the difference between the observed values (1 or 0) and estimated probabilities that are standardized by dividing by the standard deviation of

the estimated probability. In large samples the standardized residuals should have a normal distribution. **Deviance residuals** are expressed in log likelihood terms ( $-2$  times the log of the difference in predicted probability and actual outcome), such that the sum of the squared deviance across all observations is equal to the  $-2 \log$  likelihood value for the model. Deviance for perfect fit (which can never occur since the probability will never exactly equal  $1$  or  $0$ ) is zero and values increase as the differences increase. Thus, in both types of residuals, the difference between the predicted probability and the actual outcome ( $1$  or  $0$ ) is restated. When using standardized residuals, observations with values greater than  $\pm 2$  should warrant further examination. Interpretation of either type of residual is similar, with larger values indicative of observations in which the predicted probability was substantially different than the actual outcome (i.e., a low predicted probability when the outcome was actually  $1$  or vice versa).

**Influential Measures** **Influential measures** provide a means of assessing the impact of individual observations on either overall model fit or specific model parameter estimates. As such, they extend beyond the information provided by residuals to address specific aspects of model estimation and interpretation. Chapter 5 provides a much more detailed discussion of characterizing the concept of influential observations and the various measures available.

In terms of the impact on overall model performance, there are three basic measures. First is **deviance difference**, which is the change in the deviance value as a result of deleting that observation. The second is the **chi-square difference**, which represents the change in model chi-square as a result of deleting the observation. The final is **C/CBAR** which are measures that are similar to Cooks distance, another measure of overall model fit change found in multiple regression. Each of these measures are best used in a comparative sense across all observations, with relatively large values indicating observations with markedly greater influence on the overall model fit.

In terms of estimated coefficients, logistic regression also provides **dfbeta** that represents the change in the estimated coefficient for each variable in the model when a specific observation is deleted from the analysis. Again, relatively large values denote those observations whose deletion from the analysis would cause a substantial change in the estimated coefficient. Observations that impact multiple coefficients, for example, may be candidates for further investigation.

The final influence measure is **leverage**, which is the degree to which each observation differs from the “average” observation’s profile on the independent variables. While not directly related to any specific aspect of model performance, it is a useful characterization of the “typicality” of an observation on the independent variables. High leverage values indicate observations that are markedly different on the set of independent variables. A useful tool is a scatterplot of leverage by residual values, such that observations with high residuals can be assessed on whether they represent unique/different profiles on the independent variables or not.

## SUMMARY

Logistic regression, while quite similar to multiple regression and discriminant analysis in many regards, has unique characteristics due to the binary dependent variable that give rise to differences in many aspects of model estimation, from the specification of the dependent measure to the method of estimation. The widespread use of logistic regression has led to substantial development of specialized measures of predictive accuracy that address not just overall model fit, but measures specific to each outcome. To this can be added the casewise diagnostics popularized in multiple regression allowing a focus on specific observations and their impact on model results. As a result, the analyst has a wide array of measures to assess a model’s overall fit to the data as well as for each outcome value and by observation.

## Stage 5: Interpretation of the Results

As discussed earlier, the logistic regression model results in coefficients for the independent variables much like regression coefficients and quite different from the loadings of discriminant analysis. Moreover, most of the diagnostics associated with multiple regression for influential observations are also available in logistic regression.

## Assessing Model Fit

Model estimation fit—degree to which the outcome measures coincide to predicted probabilities:

Model significance tests can be made with a chi-square test on the differences in the log likelihood values ( $-2LL$ ) between two models (e.g., estimated model and null model).

Pseudo  $R^2$  measures are comparable to those found in multiple regression, but generally have lower values.

Predictive accuracy—ability to classify observations into correct outcome group:

All predictive accuracy measures are based on the cut-off value selected for classification.

The final cut-off value selected should be based on comparison of predictive accuracy measures across cut-off values. While .5 is generally the default cut-off, other values may substantially improve predictive accuracy.

The classification matrix is a framework for multiple measures of predictive accuracy:

Accuracy: predictive accuracy of positives and negatives combined.

Sensitivity: true positive rate—percentage of positive outcomes correctly predicted.

Specificity: true negative rate—percentage of negative outcomes correctly predicted.

PPV (positive predictive value): percentage of positive predictions that are correct.

NPV (negative predictive value): percentage of negative predictions that are correct.

ROC curve is a graphical portrayal of the trade-off of sensitivity and specificity values across the entire range of cut-off values. A larger AUC (area under the curve) indicates better fit.

Hosmer and Lemeshow—the only statistical test of predictive accuracy; nonsignificance indicates well-fitting model.

Case-wise diagnostics—the impact of each observation on model fit and coefficients:

Both residuals (Pearson and deviance) reflect standardized differences between predicted probabilities and outcome value (0 and 1). Values above  $\pm 2$  merit further attention.

Influence measures reflect impact on model fit and estimated coefficients if an observation is deleted from the analysis.

What does differ from multiple regression, however, is the interpretation of the coefficients. Because the dependent variable has been transformed in the process described in the previous stage, the coefficients must be evaluated in a specific manner. The following discussion first addresses how the directionality and then magnitude of the coefficients are determined. Then, the differences in interpretation between metric and nonmetric independent are covered, just as was needed in multiple regression.

## TESTING FOR SIGNIFICANCE OF THE COEFFICIENTS

Logistic regression tests hypotheses about individual coefficients just as was done in multiple regression. In multiple regression, the statistical test was to see whether the coefficient was significantly different from 0. A coefficient of 0 indicates that the coefficient has no impact on the dependent variable. In logistic regression, we also use a statistical test to see whether the logistic coefficient is different from 0. Remember, however, in logistic regression using the logit as the dependent measure, a value of 0 corresponds to the odds of 1.00 or a probability of .50—values that indicate the probability is equal for each group (i.e., again no effect of the independent variable on predicting group membership).

In multiple regression, the  $t$  value is used to assess the significance of each coefficient. Logistic regression uses a different statistic, the **Wald statistic**. It provides the statistical significance for each estimated coefficient so that hypothesis testing can occur just as it does in multiple regression. If the logistic coefficient is statistically significant, we can interpret it in terms of how it impacts the estimated probability, and thus the prediction of group membership.

Estimation problems arising from quasi-complete separation are many times reflected in estimated coefficients with higher values and/or substantially higher standard errors. While many programs provide diagnostic warning when these issues arise, the analyst is always cautioned to view the estimated coefficients as potential indicators of these problems and resolve these issues if at all possible,

## INTERPRETING THE COEFFICIENTS

One of the advantages of logistic regression is that we need to know only whether an event (purchase or not, good credit risk or not, firm failure or success) occurred or not to define a dichotomous value as our dependent variable. When we analyze these data using the logistic transformation, however, the logistic regression and its coefficients take on a somewhat different meaning from those found in regression with a metric dependent variable. Similarly, discriminant loadings from a two-group discriminant analysis are interpreted differently from a logistic coefficient.

From the estimation process described earlier, we know that the coefficients ( $B_0, B_1, B_2, \dots, B_n$ ) are actually measures of the change in the ratio of the probabilities (the odds), not the probabilities directly. This requires some consideration by the analyst, especially in interpretation, as the changes are not in terms of probabilities, but rather the change in odds. Probability values closely correspond to their odds at low levels of probabilities, but they diverge sharply as probabilities increase. For example, a probability of 10 percent corresponds to odds of .111 while a probability of 50 percent equals odds of 1 and a probability of 80 percent equals 4. So increasing the odds by a certain amount relates to different changes in probability depending on the existing level of probability. This is also seen when we view the logistic curve itself. Increases of one unit of either high or low values of the independent variable results in relatively minor changes in probability (i.e., very little slope at the tails). But in the middle of the logistic curve, a one unit change in the independent variable shows marked change in probability values (i.e., the relatively steep slope in the middle section of the curve). We mention this since analysts must be aware of what the coefficients actually represent to accurately assess their impact.

The **original logistic coefficients** have an additional complication in their interpretation because they are expressed in terms of logarithms when we use the logit as the dependent measure. Thus, the original logistic coefficient reflects the change in the log of the odds. Thus, most computer programs also provide an **exponentiated logistic coefficient**, which is just a transformation (antilog) of the original logistic coefficient and now represents a change in odds. In this way, we can use either the original or exponentiated logistic coefficients for interpretation. The two types of logistic coefficient differ in that they reflect the relationship of the independent variable with the two forms of the dependent variable, as shown in Figure 8.13.

We will discuss in the next section how each form of the coefficient reflects both the direction and magnitude of the independent variable's relationship, but requires differing methods of interpretation.

**Directionality of the Relationship** The direction of the relationship (positive or negative) reflects the changes in the dependent variable associated with changes in the independent variable. A positive relationship means that an increase in the independent variable is associated with an increase in the predicted probability, and vice versa for a negative relationship. We will see that the direction of the relationship is reflected differently for the original and exponentiated logistic coefficients.

Logistic Coefficient	Reflects Changes in ...
Original	Logit (log of the odds)
Exponentiated	Odds

**Figure 8.13**  
Interpreting the Two Types of Logistic Coefficients

**INTERPRETING THE DIRECTION OF ORIGINAL COEFFICIENTS** The sign of the original coefficients (positive or negative) indicates the direction of the relationship, just as seen in regression coefficients. A positive coefficient increases the probability, whereas a negative value decreases the predicted probability, because the original coefficients are expressed in terms of logit values, where a value of 0.0 equates to an odds value of 1.0 and a probability of .50. Thus, negative numbers relate to odds less than 1.0 and probabilities less than .50.

**INTERPRETING THE DIRECTION OF EXPONENTIATED COEFFICIENTS** Exponentiated coefficients must be interpreted differently because they are the logarithms of the original coefficient. By taking the logarithm, we are actually stating the exponentiated coefficient in terms of odds, which means that exponentiated coefficients will not have negative values. Because the logarithm of 0 (no effect) is 1.0, an exponentiated coefficient of 1.0 actually corresponds to a relationship with no direction. Thus, exponentiated coefficients above 1.0 reflect a positive relationship and values less than 1.0 represent negative relationships.

**AN EXAMPLE OF INTERPRETATION** Let us look at a simple example to see what we mean in terms of the differences between the two forms of logistic coefficients. If  $B_i$  (the original coefficient) is positive, its transformation (exponentiated coefficient) will be greater than 1, meaning that the odds will increase for any positive change in the independent variable. Thus the model will have a higher predicted probability of occurrence. Likewise, if  $B_i$  is negative the exponentiated coefficient is less than 1.0 and the odds will be decreased. A coefficient of zero equates to an exponentiated coefficient value of 1.0, resulting in no change in the odds. A more detailed discussion of interpretation of coefficients, logistic transformation, and estimation procedures can be found in numerous texts [21, 24, 30].

**Magnitude of the Relationship of Metric Independent Variables** To determine how much the odds will change given a one-unit change in the independent variable, the numeric value of the coefficient must be evaluated. Just as in multiple regression, the coefficients for metric and nonmetric variables must be interpreted differently, because each reflects different impacts on the dependent variable.

For metric variables, the question is: How much will the estimated odds change for each unit change in the independent variable? In multiple regression, we knew that the regression coefficient was the slope of the linear relationship of the independent and dependent measures. A coefficient of 1.35 indicated that the dependent variable increased by 1.35 units each time that independent variable increased by one unit. In logistic regression, we know that we have a nonlinear relationship bounded between 0 and 1, so the coefficients are likely to be interpreted somewhat differently. Moreover, we have both the original and exponentiated coefficients to consider.

**ORIGINAL LOGISTIC COEFFICIENTS** Although most appropriate for determining the direction of the relationship, the original logistic coefficients are less useful in determining the magnitude of the relationship. They reflect the change in the logit (logged odds) value, a unit of measure not particularly understandable in depicting how much the odds or probabilities actually change.

**EXPONENTIATED LOGISTIC COEFFICIENTS** Exponentiated coefficients directly reflect the magnitude of the change in the odds value. Because they are exponents, they are interpreted slightly differently. Their impact is multiplicative, meaning that the coefficient's effect is not added to the dependent variable (the odds), but multiplied for each unit change in the independent variable. As such, an exponentiated coefficient of 1.0 denotes no change ( $1.0 \times$  independent variable = no change). This outcome corresponds to our earlier discussion, where exponentiated coefficients less than 1.0 reflect negative relationships and values above 1.0 denote positive relationships, while a value of 1 reflects no relationship.

**AN EXAMPLE OF ASSESSING MAGNITUDE OF CHANGE** Perhaps an easier approach to determine the amount of change in the odds from these values is as follows:

$$\text{Percentage change in odds} = (\text{Exponentiated coefficient}_i - 1.0) \times 100$$

Figure 8.14 illustrates how to calculate the change in odds due to a one-unit change in the independent variable for a range of exponentiated coefficients.

**Figure 8.14**

**Expressing Exponentiated Coefficients as a Percentage Change in Odds**

	Value				
Exponentiated Coefficient ( $e^b$ )	.20	.50	1.0	1.5	1.7
Exponentiated Coefficient - 1.0	-.80	-.50	0.0	.50	.70
Percentage change in odds	-80%	-50%	0%	50%	70%

If the exponentiated coefficient is .20, a one-unit change in the independent variable will reduce the odds by 80 percent (the same as if the odds were multiplied by .20). Likewise, an exponentiated coefficient of 1.5 denotes a 50 percent increase in the odds ratio.

A researcher who knows the existing odds and wishes to calculate the new odds value for a change in the independent variable can do so directly through the exponentiated coefficient as follows:

$$\begin{aligned}\text{New odds value} &= \text{Old odds value} \times \text{Exponentiated coefficient} \\ &\quad \times \text{Change in independent variable}\end{aligned}$$

Let us use a simple example to illustrate the manner in which the exponentiated coefficient affects the odds value. Assume that the odds are 1.0 (i.e., 50–50) when the independent variable has a value of 5.5 and the exponentiated coefficient is 2.35. We know that if the exponentiated coefficient is greater than 1.0, then the relationship is positive, but we would like to know how much the odds would change. If we expected that the value of the independent variable would increase 1.5 points to 7.0, we could calculate the following:

$$\text{New odds} = 1.0 \times 2.35 \times (7.0 - 5.5) = 3.525$$

Odds can be translated into probability values by the simple formula of:

$$\text{Probability} = \text{Odds}/(1 + \text{Odds})$$

Thus, the odds of 3.525 translate into a probability of 77.9 percent ( $3.25/(1 + 3.25) = .779$ ), indicating that increasing the independent variable by 1.5 points will increase the probability from 50 percent to 78 percent, an increase of 28 percent.

The nonlinear nature of the logistic curve is demonstrated, however, when we apply the same increase to the odds again. This time, assume that the independent variable increased another 1.5 points, to 8.5. Would we also expect the probability to increase by another 28 percent? It cannot, because that would make the probability greater than 100 percent ( $78\% + 28\% = 106\%$ ). Thus, the probability increase or decrease slows so that the curve approaches, but never reaches the two end points (0 and 1). In this example, another increase of 1.5 points creates a new odds value of 12.426, translating into odds of 92.6 percent, an increase of 14 percent. Note that in this case of increasing the probability from 78 percent, the increase in probability for the 1.5 increase in the independent variable is one-half (14%) of what it was for the same increase when the probability was 50 percent.

The result is that the researcher may find that exponentiated coefficients are quite useful not only in assessing the impact of an independent variable, but also in calculating the magnitude of the effects.

**Interpreting Magnitude for Nonmetric (Dummy) Independent Variables** As we discussed in multiple regression, dummy variables represent a single category of a nonmetric variable (see Chapter 5 for a more detailed discussion of dummy variables). As such, they are not like metric variables that vary across a range of values, but instead take on just the values of 1 or 0, indicating the presence or absence of a characteristic. As we saw in the preceding discussion for metric variables, the exponentiated coefficients are the best means of interpreting the impact of the dummy variable, but are interpreted differently from the metric variables.

Any time a dummy variable is used, it is essential to note the reference or omitted category. In a manner similar to the interpretation in regression, the exponentiated coefficient represents the relative level of the dependent variable for the represented group versus the omitted group. We can state this relationship as follows:

$$\text{Odds}_{\text{represented category}} = \text{Exponentiated coefficient} \times \text{Odds}_{\text{reference category}}$$

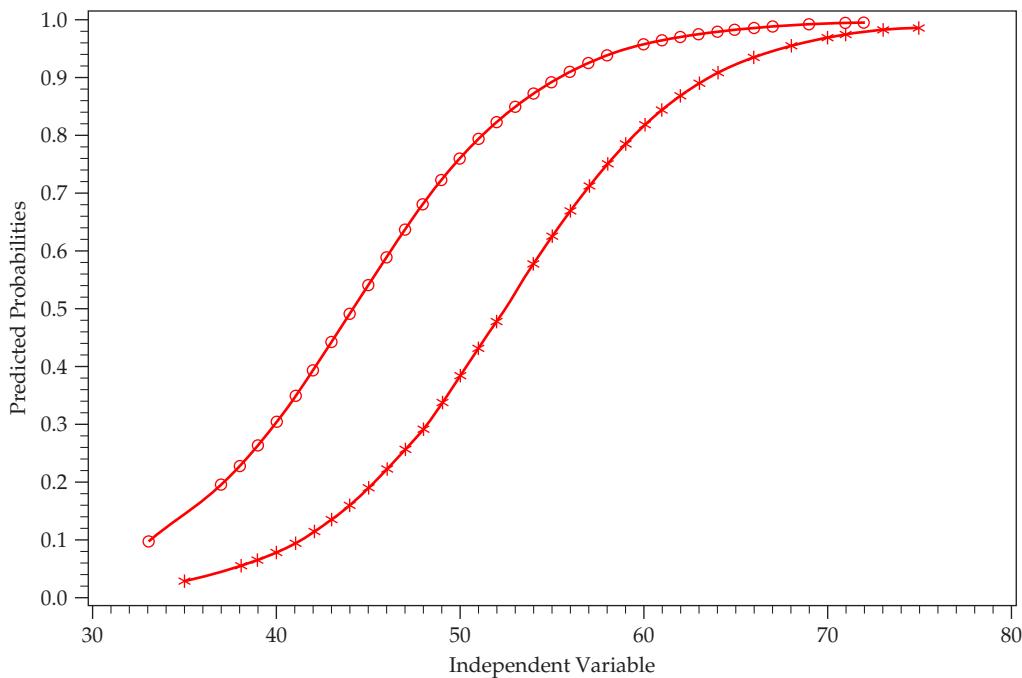
Let us use a simple example of two groups to illustrate these points. If the nonmetric variable is gender, the two possibilities are male and female. The dummy variable can be defined as representing males (i.e., value of 1 if male, 0 if female) or females (i.e., value of 1 if female, 0 if male). Whichever way is chosen, however, determines how the coefficient is interpreted. Let us assume that a 1 is given to females, making the exponentiated coefficient represent the percentage of the odds ratio of females compared to males. If the coefficient is 1.25, then females have 25 percent higher odds than males ( $1.25 - 1.0 = .25$ ). Likewise, if the coefficient is .80, then the odds for females are 20 percent less ( $.80 - 1.0 = -.20$ ) than males.

Figure 8.15 illustrates the impact of a nonmetric variable on the predicted probabilities. Just as we saw with binary variables in regression that generated predicted values that were parallel lines, here we see parallel logistic curves, varying by the impact of the binary measure.

**Other Measures of Variable Importance** As we have seen in multiple regression and other methods, a series of additional measures of variable importance have been developed. In the case of logistic regression, one of these (relative importance) is an extension of measures first derived for multiple regression, while the other (weight of evidence/information value) has been developed specifically for logistic regression. While both of these measures must be calculated externally to the major software packages, they do represent additional measures whose benefits may be realized in differing research situations or disciplines.

**RELATIVE IMPORTANCE** One of the major questions in interpretation of multiple regression results is finding a measure for determining importance of the independent variables on a metric which is easily understood and directly

**Figure 8.15**  
Impact of Binary Nonmetric Variable on Predicted Probabilities



comparable. **Relative importance**, in both regression and logistic regression, is a measure of contribution of each of the independent variable where these weights will sum to the coefficient of determination,  $R^2$ , and can be also be represented as a percentage of the model fit. As such, they represent a direct measure of relative effect sizes. Interested analysts are referred to [32], which provides an overview of the approach and links to macros for most software packages to enable calculating these values.

**WEIGHT OF EVIDENCE/INFORMATION VALUE** Another measure of variable importance is **weight of evidence** and its complementary measure **information value**. This is a measure that was developed primarily in the credit card industry [15] and has spread as an interpretation tool among other disciplines. Weight of evidence is a concept drawn from information theory that represents the degree to which a nonmetric variable distinguishes between the binary outcomes. Calculated for each variable separately, it indicates how well each category of the variable distinguishes among the outcomes. The weight of evidence values of the categories as combined to provide an overall measure of discrimination for an independent variable that is **information value**. The information value provides another measure for comparing between independent variables as to their relative value in distinguishing between the outcomes. Larger information values indicate that an independent variable is more informative about the outcome (i.e., provides more discrimination between the two outcomes).

In terms of assessing the information values themselves, information values less than 0.02 indicate that a variable is not predictive; 0.02 to 0.1 indicate weak predictive power; 0.1 to 0.3 indicate medium predictive power; and greater than 0.3 indicates strong predictive power [15]. Information value can also be calculated for metric independent variables by converting them to categorical variables, with the benefit of assessing the relative role each category/range of the metric variables to provide discrimination power for the outcome measures. It should be noted that weight of evidence and information value are calculated for each independent variable separately and thus do not reflect their final model importance which may be impacted by multicollinearity among the independent variables. As such, it is often also used as a variable selection measure among multiple independent variables for possible inclusion in the logistic model [23, 33].

**Multicollinearity** The issue of multicollinearity impacts the interpretation of coefficients in logistic regression in a manner similar to other multivariate methods, perhaps most comparably to multiple regression. High levels of multicollinearity reduce the unique impact of the independent variables involved and thus their estimated coefficients and standard errors. While not a problem for purposes of model fit, it does potentially confound interpretation. Most logistic regression programs do not provide collinearity diagnostics, but the analyst can easily use multiple regression models to obtain this information. Since multicollinearity is only based upon the independent variables in a model, the analyst can run the logistic model (dependent and independent variables) in a multiple regression procedure and obtain the collinearity diagnostics. While estimated coefficients are obviously not used, the VIF and tolerance measures are applicable. As with regular multiple regression, another approach is to calculate binary correlations among the independent variables and any bivariate correlation .50 or greater indicates the possibility of multicollinearity problems.

## CALCULATING PROBABILITIES FOR A SPECIFIC VALUE OF THE INDEPENDENT VARIABLE

In the earlier discussion of the assumed distribution of possible dependent variables, we described an S-shaped, or logistic, curve. To represent the relationship between the dependent and independent variables, the coefficients must actually represent nonlinear relationships between the dependent and independent variables. Although the transformation process of taking logarithms provides a linearization of the relationship, the researcher must remember that the coefficients actually represent different slopes in the relationship depending on the values of the independent variable. In this way, the S-shaped distribution can be estimated. If the researcher is interested in the slope of the relationship at various values of the independent variable, the coefficients can be calculated and the relationship assessed [13].

## OVERVIEW OF INTERPRETING COEFFICIENTS

The similarity of the coefficients to those found in multiple regression has been a primary reason for the popularity of logistic regression. As we have seen in the prior discussion, many aspects are quite similar, but the unique nature of the dependent variable (the odds ratio) and the logarithmic form of the variate (necessitating use of the exponentiated coefficients) requires a somewhat different approach to interpretation. The researcher, however, still has the ability to assess the direction and magnitude of each independent variable's impact on the dependent measure and ultimately the classification accuracy of the logistic model.

## Stage 6: Validation of the Results

The final stage of a logistic regression analysis involves ensuring the external as well as internal validity of the results. Although logistic regression is not as susceptible as discriminant analysis to "overfitting" the results, the process of validation is still essential, especially with smaller samples. The most common approach for establishing external validity is the assessment of hit ratios through either a separate sample (holdout sample) or utilizing a procedure that repeatedly processes the estimation sample. External validity is supported when the hit ratio of the selected approach exceeds the comparison standards that represent the predictive accuracy expected by chance (see discussion on predictive accuracy measures in Chapter 7).

The most common form of validation is through the creation of a **holdout sample**, also referred to as the **validation sample**, which is separate from the **analysis sample** used to estimate the model. The objective is to apply the logistic model to a totally separate set of respondents to assess the levels of predictive accuracy achieved. Because these cases were not used in the estimation process, they should provide insight into the generalizability of the logistic model.

A second approach is **cross-validation**, which uses a variant of the holdout sample where the test of external validity uses multiple subsets of the total sample. The most widely used approach is the jackknife method based on the "leave-one-out" principle. Typically the analysis is performed on  $k - 1$  subsamples, eliminating one observation at a time from a sample of  $k$  cases. The logistic model is computed for each sample and then the predicted group membership of the eliminated observation is computed. After all the subsamples have been analyzed, a classification matrix is constructed and the hit ratio calculated for the holdout cases in each subsample. Readers are encouraged to review Chapter 7 for more detail on the validation process in discriminant analysis.

### Interpreting Coefficients

Large values for original logistic coefficients and/or associated standard errors may indicate quasi-complete separation problems.

Coefficients are expressed in two forms: original and exponentiated to assist in interpretation.

Interpretation of the coefficients for direction and magnitude is as follows:

Direction can be directly assessed in the original coefficients (positive or negative signs) or indirectly in the exponentiated coefficients (less than 1 are negative, greater than 1 are positive).

Magnitude is best assessed by the exponentiated coefficient, with the percentage change in the dependent variable shown by:

$$\text{Percentage change} = (\text{Exponentiated coefficient} - 1.0) \times 100$$

## Validation

Split of the sample into estimation and holdout portions is encouraged to provide separate test of predictive accuracy.

If a holdout sample is precluded by small overall sample size, a cross-validation procedure (i.e., the “leave-one-out” or jackknife approach) is available.

# An Illustrative Example of Logistic Regression

Logistic regression is an attractive alternative to discriminant analysis whenever the dependent variable has only two categories. Its advantages over discriminant analysis include the following:

- 1 Less affected than discriminant analysis by the variance–covariance inequalities across the groups, a basic assumption of discriminant analysis.
- 2 Handles categorical independent variables easily, whereas in discriminant analysis the use of dummy variables created problems with the variance–covariance equalities.
- 3 Empirical results parallel those of multiple regression in terms of their interpretation and the casewise diagnostic measures available for examining residuals.

The following example, identical to the two-group discriminant analysis discussed in Chapter 7, illustrates these advantages and the similarity of logistic regression to the results obtained from multiple regression. As we will see, even though logistic regression has many advantages as an alternative to discriminant analysis, the researcher must carefully interpret the results due to the unique types of estimated coefficients along with aspects of how logistic regression handles the prediction of probabilities and group membership.

The issues addressed in the first three stages of the decision process are identical for the two-group discriminant analysis and logistic regression. We review them briefly here and refer the reader to Chapter 7 for more detail.

## STAGE 1: OBJECTIVES OF LOGISTIC REGRESSION

As discussed in Chapter 7, HBAT’s management team is interested in any differences in perceptions between those customers located and served by their US-based salesforce versus those outside the United States who are served mainly by independent distributors. Their specific focus is to identify any perceptions that differ in terms of their primary areas of operations (product line, pricing, etc.) between these two sets of customers. One objective is to identify which perception(s) provide distinctive differences so as to develop differentiated strategies between the customer groups. But a second objective, classification, is equally important as a means of developing support for the model among operational units. The ability of the management team to easily distinguish between customers even without knowing their geographic location makes for improved salesforce coordination. Moreover, identifying customers which are dissimilar to their contemporaries make allow for even more differentiated services.

## STAGE 2: RESEARCH DESIGN FOR LOGISTIC REGRESSION

The research design stage focuses on three key issues: selecting dependent and independent variables, assessing the adequacy of the sample size, both overall and by outcome group, and dividing the sample for validation purposes.

**Selection of Dependent and Independent Variables** Logistic regression analysis requires a single nonmetric dependent measure and either metric or nonmetric independent measures. The dependent variable is Region ( $X_4$ ) and the independent variables are the perceptions of HBAT. We should note that nonmetric variables will not be used as independent variables because of their impact on sample size as discussed in the next section.

**Sample Size** With a HBAT dataset of 100 observations, logistic regression will face issues relating to the ability to create a holdout sample as well as the lower power of the model to identify significant results. While the analysis and holdout samples of 60 and 40 respectively do fall below the recommended sample size, the size of the two outcomes in the estimation sample (34 and 26) provides an adequate sample size for model estimation. Moreover, no nonmetric independent variables will be used in the analysis, thus avoiding issues of quasi-complete separation due to cells with zero observations. It is important to ensure randomness in the selection of the holdout sample so that any ordering of the observations does not affect the processes of estimation and validation.

### STAGE 3: ASSUMPTIONS OF LOGISTIC REGRESSION

The only assumptions underlying logistic regression analysis involve the independence of observations, which is met in this sample by the nature of its research design. It is important to remember, however, that an underlying assumption is the “linearity of the logit” which forms the basis for the estimation of the logistic model. Given the inadequacy of graphical means for portraying nonlinearities, the Box Tidwell test is the most direct approach. As discussed earlier, this involves adding an interaction term representing the nonlinear effect of each independent variable and then assessing its significance. The test for nonlinear effects will be performed after model estimation to see if any variables and/or interaction terms should be added for improved model fit.

### STAGE 4: ESTIMATION OF THE LOGISTIC REGRESSION MODEL AND ASSESSING OVERALL FIT

Before model estimation begins, Table 8.1 provides univariate tests of each independent variable. A series of simple logistic regression models were estimated with each independent variable as the only variable in the model to provide the univariate significance tests. The univariate tests identify five of the 13 variables ( $X_6$ ,  $X_{11}$ ,  $X_{12}$ ,  $X_{13}$ , and  $X_{17}$ ) with statistically significant relationships with the dependent variable. We would expect one or more of these variables to be in our final model. We can also expect multicollinearity among these variables that can impact not only the estimated coefficients and standard errors, but also any sequential estimation process that is used. Prior analysis in Exploratory Factor Analysis (Chapter 3) gave some indications that multicollinearity may have an impact on model

**Table 8.1** Univariate Analysis of Independent Variables

Independent Variable	Univariate Logistic Model	
	Logit Coefficient	Significance
$X_6$ Product Quality	-.782	.001
$X_7$ E-Commerce Activities	.611	.159
$X_8$ Technical Support	-.215	.210
$X_9$ Complaint Resolution	-.184	.355
$X_{10}$ Advertising	.215	.376
$X_{11}$ Product Line	-.1145	.000
$X_{12}$ Salesforce Image	.852	.006
$X_{13}$ Competitive Pricing	1.129	.000
$X_{14}$ Warranty & Claims	-.240	.496
$X_{15}$ New Products	.131	.436
$X_{16}$ Order & Billing	-.078	.765
$X_{17}$ Price Flexibility	1.932	.000
$X_{18}$ Delivery Speed	-.135	.692

estimation and interpretation. First, among the five significant variables, both  $X_6$  and  $X_{13}$  were part of the Product Value factor derived by exploratory factor analysis. Thus, only one of those variables may be in a final model due to multicollinearity. Second, we also know that four factors were identified containing 10 of the 13 independent variables indicating some levels of multicollinearity among the entire set of variables. With these issues in mind, the collinearity diagnostics run in conjunction with multiple regression (see Chapter 5 for more detail) did not identify any levels of multicollinearity that required remediation. Thus, while logistic regression is affected by multicollinearity among the independent variables in a manner similar to discriminant analysis and regression analysis, it is not expected to be a serious problem in model estimation or interpretation.

While these five variables are the logical candidates for inclusion in the logistic regression variate because they demonstrate significant relationships, we need to assess all of the variables in some concurrent manner apart from the univariate tests. Logistic regression may include one or more of these variables in the model, as well as even other variables that do not have significant differences at this stage if they work in combination with other variables to significantly improve prediction. The possible nonlinear effects of the independent variables will also be examined. For purposes of illustration, the stepwise model will be fully evaluated and then the nonlinear effects will be tested to determine if there are substantive improvements to the model.

**Stepwise Model Estimation** A stepwise logistic regression model is estimated much like multiple regression in that a base model is first estimated to provide a standard for comparison (see earlier discussion for more detail). In multiple regression, the mean is used to set the base model and calculate the total sum of squares. In logistic regression, the same process is used, with the intercept used in the estimated model not to calculate the sum of squares, but instead to calculate the log likelihood value. From this model, the conditional relationship (i.e., accounting for variables in the model) for each variable can be established and the most discriminating variable chosen in a stepwise model according to the selection criteria. It should also be noted that the default cut-off value of .5 is used for predicting outcomes, but this will be re-examined once the final model is estimated.

**ESTIMATING THE BASE MODEL** Table 8.2 contains the base model results for the logistic regression analysis based on the 60 observations in the analysis sample. The log likelihood value ( $-2LL$ ) is 82.108. The score statistic, a measure of association used in logistic regression, is the measure used for selecting variables in the stepwise procedure. Several

**Table 8.2** Logistic Regression Base Model Results

Overall Model Fit: Goodness-of-Fit Measures		
		Value
$-2 \text{ Log Likelihood } (-2LL)$		
		82.108
Variables Not in the Equation		
Independent Variables	Score Statistic	Significance
$X_6$ Product Quality	11.925	.001
$X_7$ E-Commerce Activities	2.052	.152
$X_8$ Technical Support	1.609	.205
$X_9$ Complaint Resolution	.866	.352
$X_{10}$ Advertising	.791	.374
$X_{11}$ Product Line	18.323	.000
$X_{12}$ Salesforce Image	8.622	.003
$X_{13}$ Competitive Pricing	21.330	.000
$X_{14}$ Warranty & Claims	.465	.495
$X_{15}$ New Products	.614	.433
$X_{16}$ Order & Billing	.090	.764
$X_{17}$ Price Flexibility	21.204	.000
$X_{18}$ Delivery Speed	.157	.692

criteria can be used to guide entry: greatest reduction in the  $-2LL$  value, greatest Wald coefficient, or highest conditional probability. In our example, we employ the criteria of reduction of the log likelihood ratio.

In reviewing the score statistics of variables not in the model at this time, we see that the same five variables with statistically significant differences ( $X_6, X_{11}, X_{12}, X_{13}$ , and  $X_{17}$ ) are the only variables with significant score statistics in Table 8.2. Because the stepwise procedure selects the variable with the highest score statistic,  $X_{13}$  should be the variable added in the first step.

**STEPWISE ESTIMATION: ADDING THE FIRST VARIABLE,  $X_{13}$**  As expected,  $X_{13}$  was selected for entry in the first step of the estimation process (see Table 8.3). It corresponded to the highest score statistic across all 13 perception variables. The entry of  $X_{13}$  into the logistic regression model obtained a reasonable model fit, with pseudo  $R^2$  values ranging from .306 (pseudo  $R^2$ ) to .459 (Nagelkerke  $R^2$ ) and overall accuracy ratios of 73.3 percent and 75.0 percent for the analysis and holdout samples, respectively. We will examine all of the measures of overall model fit and predictive accuracy after the final stepwise model is determined.

**Table 8.3** Logistic Regression Stepwise Estimation: Adding  $X_{13}$  (Competitive Pricing)

Overall Model Fit: Goodness-of-Fit Measures					
CHANGE IN $-2LL$					
	From Base Model		From Prior Step		
	Value	Change	Significance	Change	Significance
–2 Log Likelihood ( $-2LL$ )	56.971	25.136	.000	25.136	.000
Cox and Snell $R^2$	.342				
Nagelkerke $R^2$	.459				
Pseudo $R^2$	.306				
	Value	Significance			
Hosmer and Lemeshow $\chi^2$	17.329	.027			

Variables in the Equation						
Independent						
Variable	B	Std. Error	Wald	df	Sig.	Exp(B)
$X_{13}$ Competitive Pricing	1.129	.287	15.471	1	.000	3.092
Constant	−7.008	1.836	14.570	1	.000	.001

B = logistic coefficient, Exp(B) = exponentiated coefficient.

Variables Not in the Equation		
Independent Variables	Score Statistic	Significance
$X_6$ Product Quality	4.859	.028
$X_7$ E-Commerce Activities	.132	.716
$X_8$ Technical Support	.007	.932
$X_9$ Complaint Resolution	1.379	.240
$X_{10}$ Advertising	.129	.719
$X_{11}$ Product Line	6.154	.013
$X_{12}$ Salesforce Image	2.745	.098
$X_{14}$ Warranty & Claims	.640	.424
$X_{15}$ New Products	.344	.557
$X_{16}$ Order & Billing	2.529	.112
$X_{17}$ Price Flexibility	13.723	.000
$X_{19}$ Delivery Speed	1.206	.272

**Table 8.3** (Continued)

		PREDICTED GROUP MEMBERSHIP <sup>a</sup>					
		ANALYSIS SAMPLE			HOLDOUT SAMPLE		
		$X_4$ Region			$X_4$ Region		
		USA/North America	Outside North America	Total	USA/North America	Outside North America	Total
Actual Group Membership	USA/North America	19 (73.1)	7	26	4 (30.8)	9	13
	Outside North America	9 (73.5)	25	34	1 (73.3)	26 (96.3)	27 (75.0)

<sup>a</sup> Values in parentheses are percentage correctly classified.

Examination of the results, however, identifies two reasons for considering an additional stage(s) to add variable(s) to the logistic regression model. First, three variables not in the current logistic model ( $X_{17}$ ,  $X_{11}$ , and  $X_6$ ) have statistically significant score statistics, indicating that their inclusion would significantly improve the overall model fit. Second, the overall accuracy for the holdout sample is good (75.0%), but one of the groups (USA/North America customers) has an unacceptably low level of predictive accuracy of 30.8 percent.

**STEPWISE ESTIMATION: ADDING THE SECOND VARIABLE,  $X_{17}$ .** Hopefully one or more steps in the stepwise procedure will result in the inclusion of all independent variables with significant score statistics as well as achieve acceptable levels of predictive accuracy (overall and group-specific) for both the analysis and holdout samples.

$X_{17}$ , with the highest score statistic after adding  $X_{13}$ , was selected for entry at step 2 (Table 8.4). Improvement in all measures of model fit ranged from a decrease in the  $-2LL$  value to the various  $R^2$  measures. More important from a model estimation perspective, however, none of the variables not in the equation had statistically significant change scores. Thus, the two-variable logistic model including  $X_{13}$  and  $X_{17}$  will be the final model in the stepwise procedure. We can then proceed to assessing in more detail overall model fit and predictive accuracy.

**Assessing Overall Model Fit** In making an assessment of the overall fit of a logistic regression model, we can draw upon three approaches: statistical measures of overall model fit improvement, pseudo  $R^2$  measures, and measures of classification accuracy. Each of these approaches will be examined for the one-variable and two-variable (final) logistic regression models from the stepwise procedure. We should note that while almost all of the measures shown here are common to all software packages, there are specific instances in which they differ. For example, IBM SPSS has direct estimation of a holdout sample classification matrix while SAS provides measures such as the global null hypothesis and concordance measures. While we combine all of these measures here to illustrate their use, the analyst should feel comfortable that each software package provides a comprehensive set of results with which to evaluate any model results.

**STATISTICAL MEASURES** The first statistical measure is the chi-square test for the change in the  $-2LL$  value from the base model, which is comparable to the overall  $F$  test in multiple regression. Smaller values of the  $-2LL$  measure indicate better model fit, and the statistical test is available for assessing the difference between both the base model and other proposed models (in a stepwise procedure, this test is always based on improvement from the prior step).

In the single-variable model (see Table 8.5), the  $-2LL$  value is reduced from the base model value of 82.108 to 59.971, a decrease of 25.136. This increase in model fit was statistically significant at the .000 level. In the two-variable model, the  $-2LL$  value decreased further to 39.960, resulting in significant decreases not only from the base model (42.148), but also a significant decrease from the one-variable model (17.011). Both of these improvements in model fit were significant at the .000 level.

**Table 8.4** Logistic Regression Stepwise Estimation: Adding  $X_{17}$  (Price Flexibility)

Overall Model Fit: Goodness-of-Fit Measures		CHANGE IN -2LL			
		From Base Model		From Prior Step	
	Value	Change	Significance	Change	Significance
-2 Log Likelihood (-2LL)	39.960	42.148	.000	17.011	.000
Cox and Snell $R^2$	.505				
Nagelkerke $R^2$	.677				
Pseudo $R^2$	.513				
	Value	Significance			
Hosmer and Lemeshow $\chi^2$	5.326	.722			

**Variables in the Equation**

Independent Variable	B	Std. Error	Wald	df	Sig.	Exp(B)
$X_{13}$ Competitive Pricing	1.079	.357	9.115	1	.003	2.942
$X_{17}$ Price Flexibility	1.844	.639	8.331	1	.004	6.321
Constant	-14.192	3.712	14.614	1	.000	.000

B = logistic coefficient, Exp(B) = exponentiated coefficient.

**Variables Not in the Equation**

Independent Variables	Score Statistic	Significance
$X_6$ Product Quality	.656	.418
$X_7$ E-Commerce Activities	3.501	.061
$X_8$ Technical Support	.006	.937
$X_9$ Complaint Resolution	.693	.405
$X_{10}$ Advertising	.091	.762
$X_{11}$ Product Line	3.409	.065
$X_{12}$ Salesforce Image	.849	.357
$X_{14}$ Warranty & Claims	2.327	.127
$X_{15}$ New Products	.026	.873
$X_{16}$ Order & Billing	.010	.919
$X_{18}$ Delivery Speed	2.907	.088

**Classification Matrix**

		Predicted Group Membership <sup>a</sup>			
		ANALYSIS SAMPLE		HOLDOUT SAMPLE	
		$X_4$ Region		$X_4$ Region	
Actual Group Membership		USA/ North America	Outside North America	USA/ North America	Outside North America
USA/ North America		25 (96.2)	1	Total 9 (69.2)	4 (25.0)
Outside North America		6	28 (82.4)	34 (88.3)	25 (92.6)
				Total 13	27 (85.0)

<sup>a</sup> Values in parentheses are percentage correctly classified (hit ratio).

The second measure of overall model fit is the test of the global null hypothesis that none of the estimated variable coefficients are significant. Comparable to the overall  $F$  test in regression, it provides a comparison of a model to the baseline or null model of only the intercept. Table 8.5 contains the three measures used to test the global null hypothesis (likelihood ratio, score coefficient and the Wald coefficient). As expected, the initial model with  $X_{13}$  was significantly different on all three measures and that difference increased substantially with the inclusion of  $X_{17}$ .

Both measures of overall model fit indicated a statistically significant improvement from the baseline or null model. However, as with many significance tests, we still lack some more direct assessment as to the magnitude of the model fit. In this regard we will examine two additional types of fit measures—pseudo  $R^2$  measures and measures of predictive accuracy.

**PSEUDO  $R^2$  MEASURES** Three available measures are comparable to the  $R^2$  measure in multiple regression: the Cox and Snell  $R^2$ , the Nagelkerke  $R^2$ , and a pseudo  $R^2$  measure based on the reduction in the  $-2LL$  value (see Table 8.6). For the one-variable logistic regression model, these values were .342, .459, and .306, respectively. In combination, they indicate that the one-variable regression model accounts for approximately one-third of the variation in the dependent measure. Although the one-variable model was deemed statistically significant on several overall measures of fit, these  $R^2$  measures are somewhat low for purposes of practical significance.

The two-variable model (see Table 8.4) has  $R^2$  values that are each over .50, indicating that the logistic regression model accounts for at least one-half of the variation between the two groups of customers. One would always want to improve these values, but this level is deemed practically significant in this situation, particularly given the fact that pseudo  $R^2$  measures for logistic regression are generally lower than those found in multiple regression. The  $R^2$  values of the two-variable model showed substantive improvement over the single-variable model and indicate good model fit, even when compared to the  $R^2$  values usually found in multiple regression. Coupled with the statistically based measures of model fit, the model is deemed acceptable in terms of both statistical and practical significance.

**CLASSIFICATION ACCURACY** The third type of measure for overall model fit involves the predictive accuracy of the model with particular emphasis on practical significance (i.e., how well actual outcomes are predicted by the model and how well the predictions actually work). The classification matrices, which are identical in nature to those used in discriminant analysis, represent the framework for the various measures of predictive accuracy. Yet before predictive accuracy can be assessed, the appropriate cut-off value used in predictions of group membership must be determined.

**Determining The Appropriate Cut-off Value** The classification matrices shown in Tables 8.3 and 8.4 used the default cut-off value of .50. This cut-off value may be appropriate given the relatively equal group sizes of 26 and 34 in the two outcome groups, but the analyst should examine the levels of predictive accuracy obtained with other cut-off

**Table 8.5** Statistical Measures of Overall Model Fit Improvement

		Step 1: $X_{13}$ entered		Step 2 (Final Model): $X_{17}$ entered	
Model Fit Improvement: Reduction in $-2 \log$ Likelihood (Base Model $-2LL$ 82.108)					
		Reduction in $-2LL$	Significance	Reduction in $-2LL$	Significance
–Log Likelihood		25.136	.000	17.011	.000
Testing the Global Null Hypothesis					
Test	Chi-Square	DF	Pr>ChiSq	Chi-Square	DF
Likelihood Ratio	25.1363	1	<.0001	42.1477	2
Score	21.3297	1	<.0001	31.3228	2
Wald	15.471	1	<.0001	14.1772	2
					Pr>Chisq
					<.0001
					<.0001
					0.0008

**Table 8.6** Pseudo  $R^2$  Measures of Model Fit

Pseudo $R^2$ Measure	Step 1: $X_{13}$ entered	Step 2 (Final Model): $X_{17}$ entered
Cox and Snell $R^2$	.342	.505
Nagelkerke $R^2$	.459	.677
Pseudo $R^2$	.306	.513

values. This is particularly important if the outcome has group sizes that are markedly dissimilar (e.g., the rare event situations discussed earlier). It should be noted that varying the cut-off value does not impact model estimation in any way since it is just applying the cutoff to the probability values generated by the model.

A wide range of cut-off values (ranging from 0 to 1.0 with smaller levels of difference around the default of .50) are shown in Table 8.7 along with the associated measures of predictive accuracy for each cut-off value for the final stepwise model. While there is no single measure that determines the cut-off value to select, the measures of accuracy (the overall percentage of correctly classified) and the Youden index (combination of sensitivity and specificity) best reflect overall model fit. As we can see, accuracy is highest within the range of .44 to .54, while the Youden index is maximized within the values of .50 to .54. Thus, the default cut-off value of .50 will be retained, but the analyst should take note of the range of cut-off values with maximum levels of predictive accuracy, even with very similar group sizes. This is an indication that other situations with more dissimilar group sizes may require a different cut-off value.

**Overall Predictive Accuracy** The two measures of overall predictive accuracy, accuracy and the Youden index, had maximum values of 88.3 and 78.5 percent respectively. Accuracy, similar to the hit ratio in discriminant analysis, can be compared to the values of 65.5 for the proportional chance criterion (the preferred measure) and 76.3 percent for the maximum chance criterion. The value of adding  $X_{17}$  to the initial model is shown by the accuracy values for the estimation and holdout sample were only 73.3 and 75.0 percent respectively. Thus, the additional variable provided a noticeable improvement in accuracy such that the final model exceeded both standards of comparison. If you are unfamiliar with the methods of calculating the proportional and maximum chance criterion measures, refer back to Chapter 7 regarding assessment of classification measurement accuracy.

**Sensitivity and Specificity** In addition to accuracy, sensitivity (true positive rate) and specificity (true negative rate) represent the group-specific measures that provide a more detailed perspective on predictive accuracy. These measures become even more important if the research situation requires a more focused approach on either of the two outcomes. For example, it may be deemed important to focus on improving the true positive rate, which also has the effect of reducing the false negative (predicting a true outcome to be negative) effect. Even if there is a focus on a specific outcome, it is critical to make sure that both sensitivity and specificity achieve acceptable levels, even if the overall level of predictive accuracy is acceptable.

An example of differences between levels of sensitivity and specificity are illustrated in the single-variable stepwise model for the holdout sample (see Table 8.3). Even though the accuracy level (85.0%) is greater than the proportional chance criterion and comparable to the maximum chance criterion, a significant problem appears in the specificity value (i.e., percent correctly classified for USA/North America customers), where only 30.8 percent of that group is correctly predicted. This level is below both standards and supports additional variables in the logistic model so as to increase that group-specific rate to acceptable levels. When  $X_{17}$  is added to the model, specificity increases to 69.2 percent, above the comparison value of 65.5 percent for the proportional chance criterion. With these improvements to both accuracy and sensitivity/specificity, the two-variable logistic regression model is deemed acceptable in terms of these measures of predictive accuracy.

**PPV and NPV** While sensitivity and specificity measure the ability to predict the actual outcome, they do not deal directly with the accuracy of the prediction. This is the role of PPV (positive predictive value) and NPV (negative predictive value), which are the percentage of the predictions, either positive or negative, that are accurate. So the PPV, for example, portrays the probability that the prediction of a positive outcome for an observation is correct. The focus of PPV and NPV is solely on the accuracy of the prediction, which has substantive value in the application of the logistic model for classification purposes.

For the cut-off value of .50, the PPV is 96.6 percent, where of the 29 positive predictions only one prediction was incorrect (see Table 8.7). The NPV is 80.6 percent, where of the 31 negative predictions 6 were incorrect. So while the sensitivity/specificity trade-off favored specificity (True Positive rate/sensitivity of 82.4 and True Negative rate/specificity of 96.2), the model produces a much more accurate test in terms of positive outcomes than negative outcomes.

**Table 8.7** Selecting the Cutoff Value for the Classification Matrix

Classification Table										Predictive Accuracy Measures				
Probability Level	Actual: No			Actual: Yes			Predicted: Yes	Predicted: True	Accuracy	Sensitivity	Specificity	Youden	PPV	NPV
	Predicted: No	Predicted: Yes	Predicted: False	Negative	Positive	False								
0	0	26	0	34	56.7%	100.0%	0.0%	0.0%	56.7%	0.0%	0.0%	0.0%	NC	NC
0.1	7	19	0	34	68.3%	100.0%	26.9%	26.9%	64.2%	64.2%	100.0%	100.0%	100.0%	100.0%
0.2	13	13	3	31	73.3%	91.2%	50.0%	41.2%	70.5%	70.5%	81.3%	81.3%	81.3%	81.3%
0.3	16	10	3	31	78.3%	91.2%	61.5%	52.7%	75.6%	75.6%	84.2%	84.2%	84.2%	84.2%
0.4	20	6	5	29	81.7%	85.3%	76.9%	62.2%	82.9%	82.9%	80.0%	80.0%	80.0%	80.0%
0.42	22	4	5	29	85.0%	85.3%	84.6%	69.9%	87.9%	87.9%	81.5%	81.5%	81.5%	81.5%
0.44	24	2	5	29	<b>88.3%</b>	85.3%	92.3%	77.6%	93.5%	93.5%	82.8%	82.8%	82.8%	82.8%
0.46	24	2	5	29	<b>88.3%</b>	85.3%	92.3%	77.6%	93.5%	93.5%	82.8%	82.8%	82.8%	82.8%
0.48	24	2	5	29	<b>88.3%</b>	85.3%	92.3%	77.6%	93.5%	93.5%	82.8%	82.8%	82.8%	82.8%
0.5	25	1	6	28	<b>88.3%</b>	82.4%	96.2%	<b>78.5%</b>	96.6%	96.6%	80.6%	80.6%	80.6%	80.6%
0.52	25	1	6	28	<b>88.3%</b>	82.4%	96.2%	<b>78.5%</b>	96.6%	96.6%	80.6%	80.6%	80.6%	80.6%
0.54	25	1	6	28	<b>88.3%</b>	82.4%	96.2%	<b>78.5%</b>	96.6%	96.6%	80.6%	80.6%	80.6%	80.6%
0.56	25	1	7	27	86.7%	79.4%	96.2%	75.6%	96.4%	96.4%	78.1%	78.1%	78.1%	78.1%
0.58	25	1	7	27	86.7%	79.4%	96.2%	75.6%	96.4%	96.4%	78.1%	78.1%	78.1%	78.1%
0.6	25	1	7	27	86.7%	79.4%	96.2%	75.6%	96.4%	96.4%	78.1%	78.1%	78.1%	78.1%
0.7	25	1	9	25	83.3%	73.5%	96.2%	69.7%	96.2%	96.2%	73.5%	73.5%	73.5%	73.5%
0.8	25	1	10	24	81.7%	70.6%	96.2%	66.7%	96.0%	96.0%	71.4%	71.4%	71.4%	71.4%
0.9	25	1	10	24	81.7%	70.6%	96.2%	66.7%	96.0%	96.0%	71.4%	71.4%	71.4%	71.4%
1	26	0	34	0	43.3%	0.0%	100.0%	0.0%	NC	NC	43.3%	43.3%	43.3%	43.3%

NC: Not calculated due to zero for either true positive or true negative.

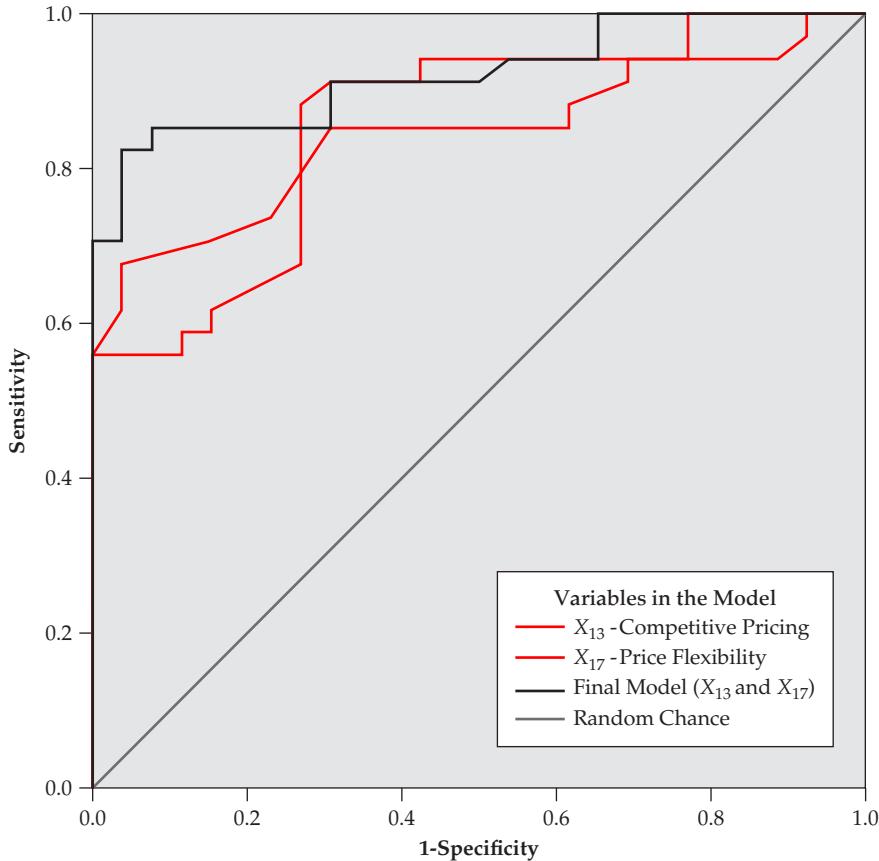
Bold indicates maximum values for measure.

**ROC Curve** The varying levels and trade-offs between sensitivity and specificity across the range of cut-off values in Table 8.7 is shown in graphical form in the ROC curve in Table 8.8. Here we see the specificity and 1 – specificity values for the complete range of cut-off values for a set of models. First, there are the curves for two univariate models, one each for  $X_{13}$  and  $X_{17}$ . These demonstrate the relative predictive accuracy for each variable separately. We can see that in some instances one model outperformed the other and they switched in other instances. As a result, the final model, which combines both variables, performs better than either single variable in almost all instances, especially at the highest levels of the cut-off value (to the left of the chart) where the false positive rate is lower (i.e., the  $x$  axis which is 1 - specificity).

Better fitting models are shown as being farther removed (i.e., towards the upper left of the chart) from the diagonal reference line representing a random chance model. The AUC measures this positioning of models, with higher values of the AUC indicating better fitting models. The final model has an AUC value of .9214, which is greater than either of the two univariate models. The AUC allows for easy comparison between models, as done here with the two univariate models versus the multivariate model. In this way the analyst can see a broader perspective of predictive accuracy versus just a single set of sensitivity and specificity values. We encouraged you to review Fawcett's article [10] for further detail on interpreting ROC curves.

**Hosmer and Lemeshow Test** None of the measures of predictive accuracy to this point provide a statistical test for the level of predictive accuracy. The Hosmer and Lemeshow measure is a statistical test of overall predictive accuracy based on the correspondence of actual and predicted values of the dependent variable. In this case, better model

**Table 8.8 Comparing the Independent Variables and Final Stepwise Model with the ROC Curve**



AUC values:  $X_{13}$  (.8495),  $X_{17}$  (.8535), Final Model (.9214), Random Chance (.5000).

fit is indicated by a smaller difference in the observed and predicted classification. Thus, well-fitting models have a small chi-square value that is nonsignificant.

The Hosmer and Lemeshow test shows statistical significance for the one-variable logistic model (.027 from Table 8.9), indicating that significant differences remain between the actual and expected values. The two-variable model, however, reduces the significance level to .722 (see Table 8.4), a nonsignificant value indicating that the model fit is acceptable. In comparing this statistical measure to the earlier measures of model fit, we do see differences. Both statistical measures of overall model fit indicated a well-fitting single variable model, but this was not supported by the statistical measure of predictive accuracy. This does not invalidate the model in any manner, but does highlight the various perspectives on model fit (i.e., model fit versus predictive accuracy) that are involved in assessing a logistic regression model.

**Concordance** One final measure of predictive accuracy is the concordance or c statistic (see Table 8.9). This measure details the percentage of all randomly formed pairs of observations (each pair has a positive outcome observation and a negative outcome observation) in which the predicted probability is higher for the positive outcome. Higher values indicate a well-fitting model across the entire set of observations. The concordance statistic is also comparable to the AUC measure of the ROC curve.

**SUMMARY** Across all of the basic types of measures of overall model fit, the two-variable model (with  $X_{13}$  and  $X_{17}$ ) demonstrates acceptable levels of both statistical and practical significances. With the overall model fit acceptable, we turn our attention to assessing the statistical tests of the logistic coefficients in order to identify the coefficients that have significant relationships affecting group membership. Before doing so, however, we will examine the casewise diagnostics to identify any observations that might exert undue influence on the model results.

**Casewise Diagnostics** The analysis of the misclassification of individual observations can provide further insight into possible improvements of the model. Casewise diagnostics such as standardized residuals and measures of influence are available in terms of both impact on overall model fit and individual model coefficients.

In the estimation sample, only seven cases were misclassified. The most direct measure of misclassification of individual cases is the standardized Pearson residual or standardized deviance. In terms of impact, the Dfbetas provide a measure of influence of each case on the estimated coefficients. Table 8.10 provides the Pearson residuals and the Dfbetas for  $X_{13}$  and  $X_{17}$  for each of the misclassified cases. Table 8.11 portrays these three influential measures for all of the cases with those exceeding the threshold ( $\pm 2$  for standardized residuals) or markedly higher than the

**Table 8.9** Hosmer and Lemeshow Test and Concordance Measures of Predictive Accuracy

Hosmer and Lemeshow Test	Step 1: $X_{13}$ Entered		Step 2 (Final Model): $X_{17}$ Entered	
	Chi-Square	Significance	Chi-Square	Significance
	17.329	.027	5.326	.722
Concordant Pairs	Percentage		Percentage	
	84.5%		92.1%	

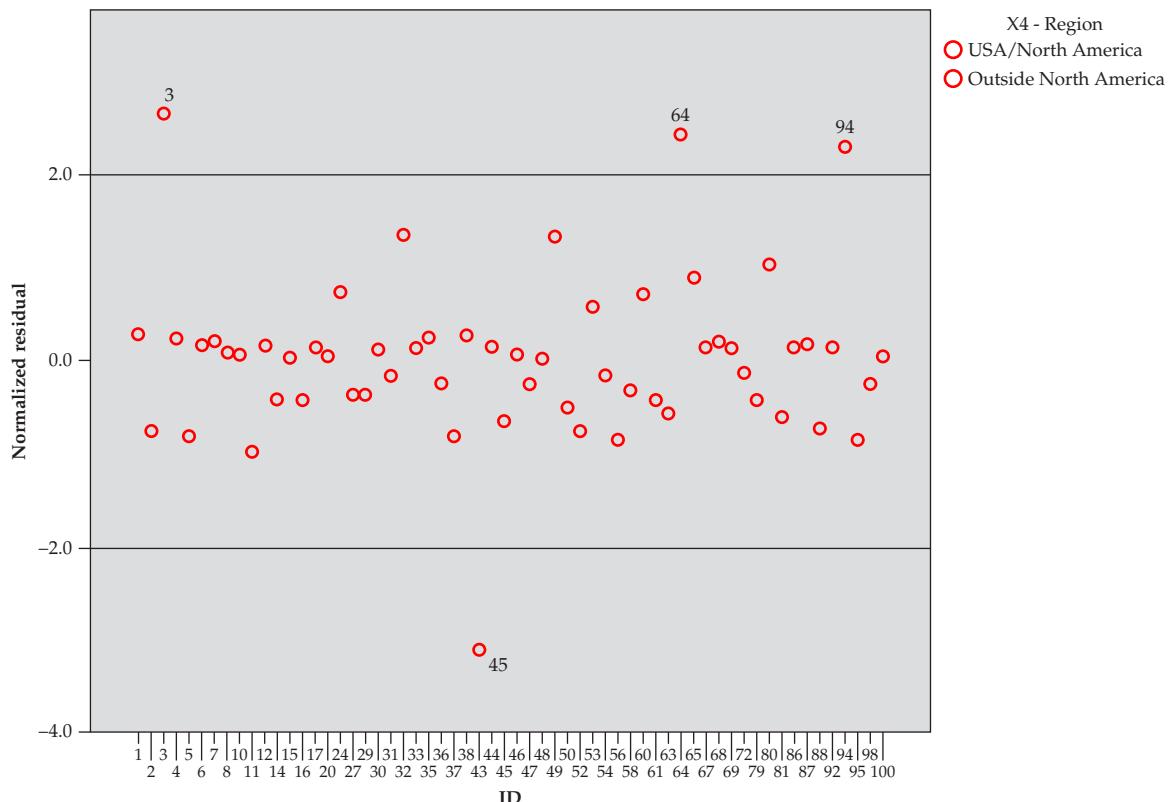
**Table 8.10** Casewise Diagnostics For Misclassified Cases

ID	$X_4$ , Region <sup>a</sup>	Predicted $X_4$	Pearson Residual	Normalized Residual	DFBETA for $X_{13}$	DFBETA for $X_{17}$
43	0	1	-0.90685	-3.12015	-0.18468	-0.33639
3	1	0	0.87653	2.66447	-0.19425	-0.04348
64	1	0	0.85632	2.44131	-0.03588	-0.29290
94	1	0	0.84127	2.30217	-0.17161	-0.00072
32	1	0	0.64928	1.36060	0.02221	-0.09600
49	1	0	0.64209	1.33939	0.03614	-0.11685
80	1	0	0.51462	1.02968	0.10015	-0.14607

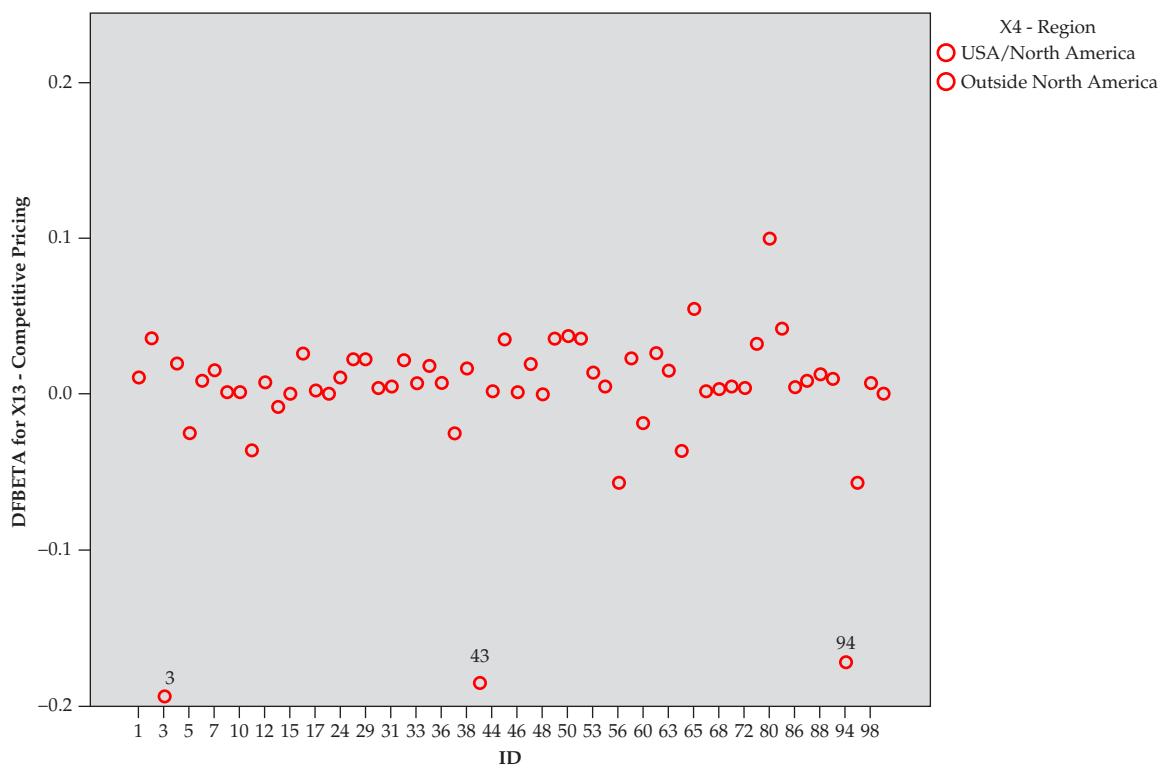
<sup>a</sup>Values of  $X_4$ : 0 = USA/North America, 1 = Outside North America.

**Table 8.11** Casewise Diagnostics for Final Stepwise Model ( $X_{13}$  and  $X_{17}$ )

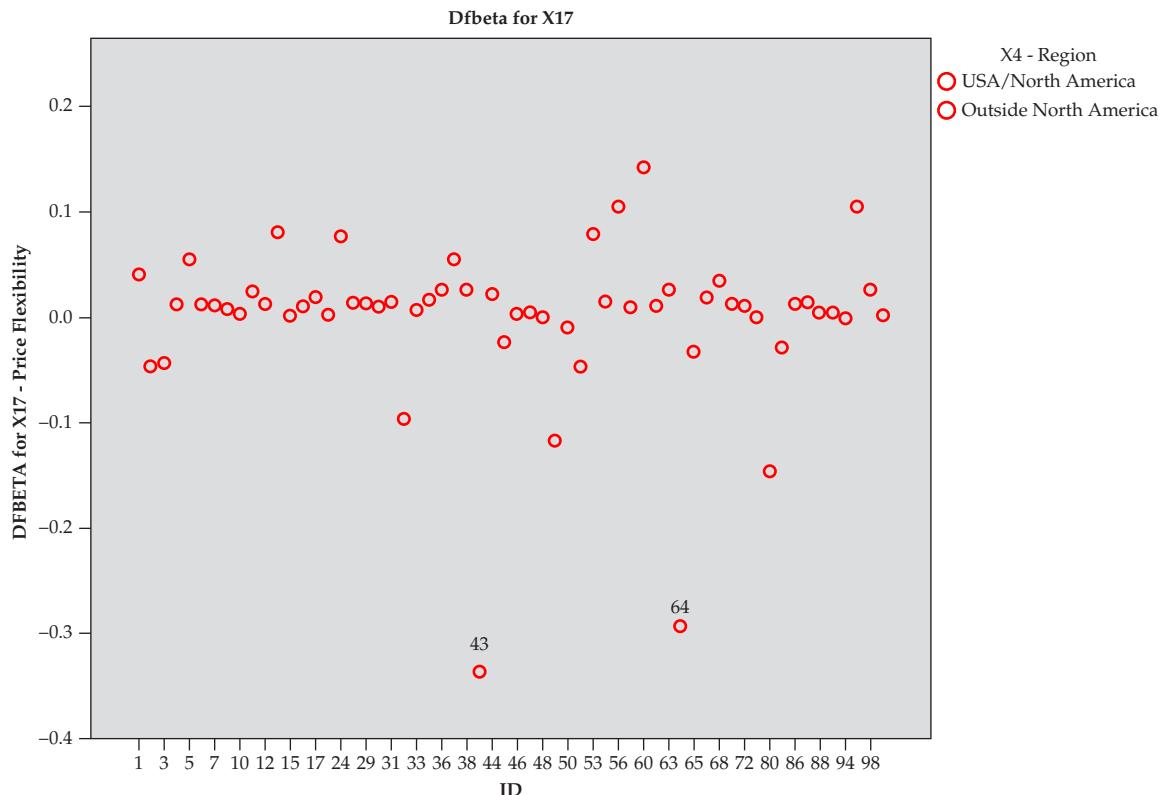
## Normalized Residuals



X4 - Region  
 ● USA/North America  
 ○ Outside North America

Dfbeta for  $X_{13}$ 

X4 - Region  
 ● USA/North America  
 ○ Outside North America

**Table 8.11 (Continued)**

remaining cases for the Dfbetas. As we can see in both tables, four cases (case IDs: 3, 43, 64, and 94) represent those cases with the highest values on at least two of the three measures. Also, note that case 43, the only misclassified case for USA/North America also had the highest residual as well as high values on both Dfbetas. These findings indicate that this case could be examined for possible data errors or even exclusion if it was found to not be representative of the USA/North American cases. Cases 3, 64, and 94 also had a very high residual indicating that their profiles on the independent variables was dissimilar to other cases in the Outside North America group and thus also candidates for further study. Given the low levels of misclassification, however, none of the cases are eliminated from the analysis.

## STAGE 5: INTERPRETATION OF RESULTS

The stepwise logistic regression procedure produced a variate quite similar to that of the two-group discriminant analysis, although with one less independent variable. We will examine the logistic coefficients to assess their statistical significance and then both the direction and impact each variable has on predicted probability and group membership.

**Statistical Significance of the Coefficients** The estimated coefficients for the two independent variables and the constant can also be evaluated for statistical significance. The Wald statistic is used to assess significance in a manner similar to the *t* test used in multiple regression. The logistic coefficients for  $X_{13}$  (1.079) and  $X_{17}$  (1.844) and the constant ( $-14.190$ ) are all significant at the .01 level, based on the statistical tests of the Wald statistic (see Table 8.4). No other variables would enter the model and achieve at least a .05 level of significance. Thus, the individual variables are significant and can be interpreted to identify the relationships affecting the predicted probabilities and subsequently group membership.

**Interpreting the Logistic Coefficients** The final logistic regression model included two variables ( $X_{13}$  and  $X_{17}$ ) with logistic regression coefficients of 1.079 and 1.844, respectively, and a constant of -14.190 (see Table 8.4). Comparing these results to the two-group discriminant analysis (see Chapter 7) reveals almost identical results, because discriminant analysis included three variables in the two-group model— $X_{13}$  and  $X_{17}$  along with  $X_{11}$ .

**DIRECTION OF THE RELATIONSHIPS** To assess the direction of the relationship of each variable, we can examine either the original logistic coefficients or the exponentiated coefficients. Let us start with the original coefficients. If you recall from our earlier discussion, we can interpret the direction of the relationship directly from the sign of the original logistic coefficients. In this case both variables have positive signs, indicating a positive relationship between both independent variables and predicted probability. As the values of either  $X_{13}$  or  $X_{17}$  increase, the predicted probability will increase, thus increasing the likelihood that a customer will be categorized as residing outside North America.

Turning our attention to the exponentiated coefficients, we should recall that values above 1.0 indicate a positive relationship and below 1.0 indicate a negative relationship. In our case, the values of 2.942 and 6.319 also indicate positive relationships.

**MAGNITUDE OF THE RELATIONSHIPS** The most direct method of assessing the magnitude of the change in probability due to each independent variable is to examine the exponentiated coefficients. As you recall, the exponentiated coefficients minus one equals the percentage change in odds (which is the same as the odds times the exponentiated coefficient). In our case, it means that an increase by one point increases the odds by 194 percent for  $X_{13}$  and 531 percent for  $X_{17}$  or multiplying the odds by 2.194 and 6.321 respectively. These numbers can exceed 100 percent because they are increasing the odds, not the probabilities themselves. The impacts are large because the constant term (-14.190) defines a starting point of almost zero for the probability values. Thus, large increases in the odds are needed to reach larger probability values. In terms of relative impact, we can see that  $X_{17}$  has slightly more than twice the impact on the odds per unit change than  $X_{13}$ .

**PREDICTING PROBABILITIES** Another approach in understanding how the logistic coefficients define probability is to calculate the predicted probability for any set of values for the independent variables. For the independent variables  $X_{13}$  and  $X_{17}$ , let us first use the group means for the two groups. In this manner, we can see what the predicted probability would be for a “typical” member of each group.

Table 8.12 shows the calculations for predicting the probability of the two group centroids. First, we calculate the logit value for each group centroid by inserting the group centroid values (e.g., 5.60 and 3.63 for group 0 on  $X_{13}$  and  $X_{17}$ , respectively) into the logit equation. Remember from Table 8.12 that the estimated weights were 1.079 and 1.844 for  $X_{13}$  and  $X_{17}$ , respectively, with a constant of -14.192. Thus, substitution of the group centroid values into this equation results in logit values of -1.452 (group 0) and 2.909 (group 1). Taking the antilog of the logit values results in odds of .234 and 18.332. Then, the probability of a group is calculated as its odds value over the sum of the odds for both groups. This results in the “typical” member of group 0 having a probability of being incorrectly assigned to group 1 of .189 (.189 = .234/(.234 + 18.332)) and the “typical” member of group 1 has a probability of .948 of being correctly assigned to group 1. This demonstrates the ability of the logistic model to create separation between the two group centroids in terms of predicted probability, resulting in the excellent classification results achieved for both analysis and holdout samples.

We can also plot the predicted probabilities across the range of values for each of independent variables along with the cases for each group (see Table 8.13). For both variables the relationships are well defined with relatively small confidence intervals. Here we can see how selecting different probability values for the cut-off values impact the predictive accuracy. For example, in the case of  $X_{13}$ , if we used a cut-off value of .5 (represented by a vertical line intersecting the curve at a value of .5) we would seem to misclassify relatively few of the upper outcome group ( $X_4 = 1$ ) compared to the lower group. If we were to increase the cut-off value above .5 then the correct classifications of the lower group would increase, but with an associated increase in misclassifications in the upper group.

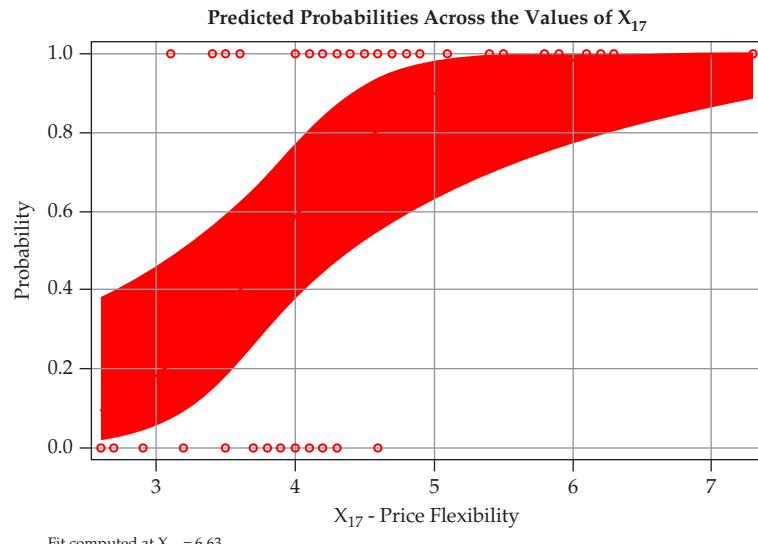
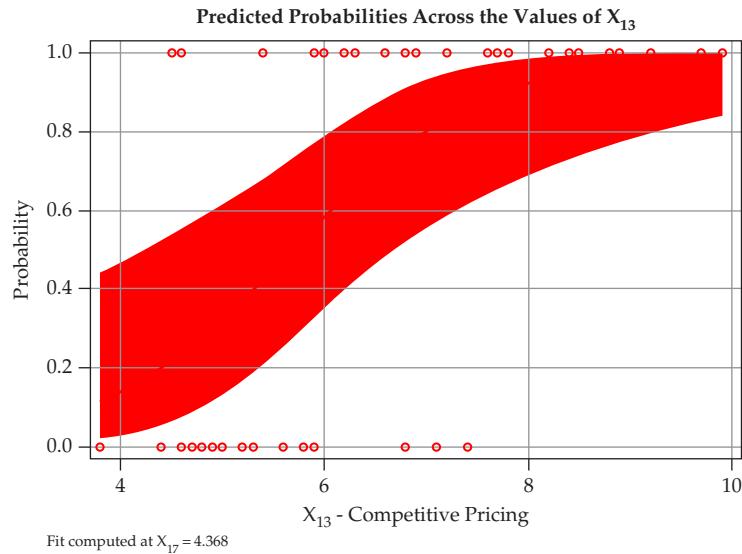
**Table 8.12** Calculating Estimated Probability Values for the Group Centroids of  $X_4$  Region

	$X_4$ (Region)	
	Group 0: USA/North America	Group 1: Outside North America
Centroid: $X_{13}$	5.60	7.42
Centroid: $X_{17}$	3.63	4.93
Logit Value <sup>a</sup>	-1.452	2.909
Odds <sup>b</sup>	.234	18.332
Probability <sup>c</sup>	.189	.948

<sup>a</sup>Calculated as: Logit =  $-14.190 + 1.079X_{13} + 1.844X_{17}$ .

<sup>b</sup>Calculated as: Odds =  $e^{\text{Logit}}$ .

<sup>c</sup>Calculated as: Probability = Odds/(1 + Odds).



Cases within each outcome group depicted on upper and lower axes.

**Table 8.13** Predicted Probability Values for  $X_{13}$  and  $X_{17}$  with Corresponding Cases in Each Outcome Group

**Assessing Nonlinear Effects** As a final step in interpreting the estimated coefficients, it is recommended to test for any additional nonlinear effects that can be added to the final model. We will first identify any nonlinear effects that would be significant additions to the univariate effects shown in Table 8.1. To do so we first computed the interaction term for each independent variable (i.e., independent variable \* log(independent variable)). The interaction terms are then added to each univariate model to test for significance. The presence of nonlinear effects is assessed both with (a) the significance test of the interaction term along with (b) a test of the incremental model fit from adding the interaction term so as to preclude any issues that might arise in the significance test due to multicollinearity of the independent variable and the interaction term.

As seen in Table 8.14, two of the independent variables exhibit potential nonlinear effects.  $X_6$ , one of the five independent variables with significant univariate effects, had a significant interaction term and significant improvement in model fit. Also,  $X_{16}$ , which did not have a significant univariate effect, exhibited a significant interaction term and model improvement. While neither of these variables was included in the final stepwise model, we will examine whether the inclusion of these nonlinear effects will improve the model at this stage. Each of these variables is tested for inclusion by adding the variable and its interaction term to the final model.

In the second section of Table 8.14 we can see the results of adding the two variables with nonlinear effects to the final model. For  $X_6$  the addition of the nonlinear effect resulted in both a nonsignificant interaction term and nonsignificant improvement in model fit. But  $X_{16}$  did exhibit significance both for the interaction term and the improvement in model fit. As a result,  $X_{16}$  and its interaction term can be added to the final model to determine if a significant and substantive improvement is achieved. In this example, the addition of the two variables did improve the model slightly, but the estimated coefficients were not interpretable due to the high multicollinearity between them. Given the emphasis on the ability to provide explanation about the relationships in the model,  $X_{16}$  and its nonlinear effect were not included in a final revised model. If the objectives of the research were focused primarily on classification, then  $X_{16}$  could be included in the model since the interpretability of the coefficients was less important.

**Summary** The estimated logistic coefficients define positive relationships for both independent variables and provide a means of assessing the impact of a change in either or both variables on the odds and thus the predicted probability. While the relationship of the independent variables are with odds or logit values, the ability to translate these to predicted probability values highlights why many researchers prefer logistic regression to discriminant

**Table 8.14** Testing for Nonlinear Effects of the Independent Variables

	Univariate Model			Addition to Final Stepwise Model		
	Significance of Interaction Term	Improvement in Model Fit Chi-Square	Significance	Significance of Interaction Term	Improvement in Model Fit Chi-Square	Significance
$X_6$ Product Quality	.016	8.095	.004	.064	4.622	.099
$X_7$ E-Commerce Activities	.337	1.018	.313			
$X_8$ Technical Support	.854	.034	.853			
$X_9$ Complaint Resolution	.202	1.871	.171			
$X_{10}$ Advertising	.254	1.356	.244			
$X_{11}$ Product Line	.144	2.712	.100			
$X_{12}$ Salesforce Image	.816	.053	.817			
$X_{13}$ Competitive Pricing	.559	.350	.554			
$X_{14}$ Warranty & Claims	.952	.004	.952			
$X_{15}$ New Products	.766	.090	.764			
$X_{16}$ Order & Billing	.060	5.274	.022	.046	7.045	.030
$X_{17}$ Price Flexibility	.214	1.904	.168			
$X_{18}$ Delivery Speed	.274	1.335	.248			

analysis when comparisons are made on the more useful information available from logistic coefficients versus discriminant loadings.

## STAGE 6: VALIDATION OF THE RESULTS

The validation of the logistic regression model is accomplished in this example through the same method used in discriminant analysis: creation of analysis and holdout samples. By examining the measures of predictive accuracy for the holdout sample (see Table 8.15), the researcher can assess the external validity and practical significance of the logistic regression model.

For the final two-variable logistic regression model, the accuracy measure (.850) exceeds both of the comparison standards (proportional chance and maximum chance criteria). Moreover, both sensitivity and specificity are sufficiently large, indicating acceptable predictive accuracy for each outcome group as well. Finally, both PPV and NPV are above .80 and provide reasonable levels of predictive ability for both positive and negative outcomes. The holdout sample results are particularly useful since they provide the primary evidence of external validity. These outcomes lead to the conclusion that the logistic regression model demonstrated sufficient external validity for complete acceptance of the results, as was found with the discriminant analysis model as well.

## A MANAGERIAL OVERVIEW

Logistic regression presents an alternative to discriminant analysis that may be more comfortable to many researchers due to its similarity to multiple regression. Given its robustness in the face of data conditions that can negatively affect discriminant analysis (e.g., unequal variance–covariance matrices), logistic regression is also the preferred estimation technique in many applications.

When compared to discriminant analysis, logistic regression provides comparable predictive accuracy with a simpler variate that used the same substantive interpretation, only with one less variable. From the logistic regression results, the researcher can focus on competitive pricing and price flexibility as the primary differentiating variables between the two groups of customers. The objective in this analysis is not to increase probability (as might be the case of analyzing success versus failure), yet logistic regression still provides a straightforward approach for HBAT to understand the relative impact of each independent variable in creating differences between the two groups of customers.

**Table 8.15 Predictive Accuracy Measures for the Holdout Sample**

Actual	Predicted		Specificity = $9/(9 + 4) = .692$
	0: USA/North America	1: Outside North America	
0: USA/North America	9	4	Specificity = $9/(9 + 4) = .692$
1 : Outside North America	2	25	Sensitivity = $22/(22 + 5) = .926$
	Negative Predictive Value: $9/(9 + 2) = .818$	Positive Predictive Value: $25/(25 + 4) = .862$	Accuracy: $(9 + 25)/40 = .850$

The researcher faced with a dichotomous dependent variable need not resort to methods designed to accommodate the limitations of multiple regression nor be forced to employ discriminant analysis, especially if its statistical assumptions are violated. Logistic regression addresses these problems and provides a method developed

to deal directly with this situation in the most efficient manner possible. Basic guidelines for their application and interpretation were included to clarify further the methodological concepts. This chapter helps you to do the following:

**State the circumstances under which logistic regression should be used instead of discriminant analysis or multiple regression.** In choosing an appropriate analytical technique, we sometimes encounter a problem that involves a categorical dependent variable and several metric independent variables. Logistic regression is the appropriate statistical technique when the research problem involves a single binary categorical dependent variable and several metric or nonmetric independent variables. Logistic regression is generally preferred over discriminant analysis when the dependent measure is binary given its minimal set of assumptions, and thus its robustness, in most situations. Moreover, the similarity in interpretation to multiple regression makes it easier for many researchers than the discriminant function(s) in the discriminant model.

**Identify the types of dependent and independent variables used in logistic regression.** Although logistic regression is limited to only a binary dependent measure, it does provide the ability to include both metric and nonmetric independent variables, much like multiple regression. This contrasts to discriminant analysis, which is limited in most situations to only metric independent variables.

**Interpret the results of a logistic regression analysis, with comparisons to both multiple regression and discriminant analysis.** The goodness-of-fit for a logistic regression model can be assessed in two ways: (1) using pseudo  $R^2$  values, similar to that found in multiple regression, and (2) examining predictive accuracy (i.e., the classification matrix in discriminant analysis). The two approaches examine model fit from different perspectives, but should yield similar conclusions. One of the advantages of logistic regression is that we need to know only whether an event occurred to define a dichotomous value as our dependent variable. When we analyze these data using the logistic transformation, however, the logistic regression and its coefficients take on a somewhat different meaning from those found in regression with a metric dependent variable. Similarly, discriminant loadings in discriminant analysis are interpreted differently from a logistic coefficient. The logistic coefficient reflects both the direction and magnitude of the independent variable's relationship, but requires differing methods of interpretation. The direction of the relationship (positive or negative) reflects the changes in the dependent variable associated with changes in the independent variable. A positive relationship means that an increase in the independent variable is associated with an increase in the predicted probability, and vice versa for a negative relationship. To determine the magnitude of the coefficient, or how much the probability will change given a one-unit change in the independent variable, the numeric value of the coefficient must be evaluated. Just as in multiple regression, the coefficients for metric and nonmetric variables must be interpreted differently, because each reflects different impacts on the dependent variable.

**Understand the strengths and weaknesses of logistic regression compared to discriminant analysis and multiple regression.** Although discriminant analysis can analyze any situation where the dependent variable is nonmetric, logistic regression is preferred for two reasons when the dependent variable is binary. First, discriminant analysis relies on strictly meeting the assumptions of multivariate normality and equal variance-covariance matrices across groups—assumptions that are not met in many situations. Logistic regression does not face these strict assumptions and is much more robust when these assumptions are not met, making its application appropriate in many situations. Second, even if the assumptions are met, many researchers prefer logistic regression, because it is similar to multiple regression. As such, it has straightforward statistical tests, similar approaches to incorporating metric and nonmetric variables and nonlinear effects, and a wide range of diagnostics. Logistic regression is equivalent to two-group discriminant analysis and may be more suitable in many situations.

Logistic regression is a valuable option in research problems that involve a single categorical dependent variable and several metric or nonmetric independent variables. Its relative strength comes in its ability to be flexible across multiple research settings, its robustness derived from a minimal set of underlying assumptions, and its similarity to multiple regression for purposes of interpretation. The result is a wide range of applications in both academic and practitioner contexts.

How would you differentiate among multiple discriminant analysis, regression analysis, logistic regression analysis, and analysis of variance?

When would you employ logistic regression rather than discriminant analysis? What are the advantages and disadvantages of this decision?

How does logistic regression handle the relationship of the dependent and independent variables?

What are the unique characteristics of interpretation in logistic regression?

Explain the concept of odds and why it is used in predicting probability in a logistic regression procedure.

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 Akinci, S., E. Kaynak, E. Atilgan, and S. Aksoy. 2007. Where Does the Logistic Regression Analysis Stand in Marketing Literature? A Comparison of the Market Positioning of Prominent Marketing Journals. *European Journal of Marketing* 47: 537–67.
- 2 Allison, P. D. 2012. *Logistic Regression Using SAS: Theory and Application*. Cary, NC: SAS Institute.
- 3 Austin, P. C., and E. W. Steyerberg. 2012. Interpreting the Concordance Statistic of a Logistic Regression Model: Relation to the Variance and Odds Ratio of a Continuous Explanatory Variable. *BMC Medical Research Methodology* 10.1186/1471-2288-12-82.
- 4 Bagley, S. C., H. White, and B. A. Golomb. 2001. Logistic Regression in the Medical Literature: Standards for Use and Reporting, With Particular Attention to One Medical Domain. *Journal of Clinical Epidemiology* 54: 979–85.
- 5 Bouwmeester, W., N. P. Zuithoff, S. Mallett, M. I. Geerlings, Y. Vergouwe, E. W. Steyerberg, and K. G. Moons. 2012. Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLoS Medicine* 9: e1001221.
- 6 Breslow, Norman E. 1996. Statistics in Epidemiology: The Case-Control Study. *Journal of the American Statistical Association* 91: 14–28.
- 7 Chen, G., and H. Tsurumi. 2010. Probit and Logit Model Selection. *Communications in Statistics—Theory and Methods* 40: 159–75.
- 8 Coussement, K., F. A. Van den Bossche, and K. W. De Bock. 2014. Data Accuracy's Impact on Segmentation Performance: Benchmarking RFM Analysis, Logistic Regression, and Decision Trees. *Journal of Business Research* 67: 2751–8.
- 9 Demaris, A. 1995. A Tutorial in Logistic Regression. *Journal of Marriage and the Family* 57: 956–68.
- 10 Fawcett, T. 2004. ROC Graphs: Notes and Practical Considerations for Researchers. *Machine Learning* 31: 1–38.
- 11 Firth, D. 1993. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* 80: 27–38.
- 12 Frank, R. E., W. E. Massey, and D. G. Morrison. 1965. Bias in Multiple Discriminant Analysis. *Journal of Marketing Research* 2: 250–58.
- 13 Gessner, Guy, N. K. Maholtra, W. A. Kamakura, and M. E. Zmijewski. 1988. Estimating Models with Binary Dependent Variables: Some Theoretical and Empirical Observations. *Journal of Business Research* 16: 49–65.
- 14 Greenland, S., J. A. Schwartzbaum, and W. D. Finkle. 2000. Problems Due to Small Samples and Sparse Data in Conditional Logistic Regression Analysis. *American Journal of Epidemiology* 151: 531–9.
- 15 Hababou, M., A. Y. Cheng, and R. Falk. 2006. Variable Selection in the Credit Card Industry. *NESUG Proceedings: Statistics and Pharmacokinetics*.
- 16 Harrell Jr, F. E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Berlin: Springer.
- 17 Heinze, G. 2006. A Comparative Investigation of Methods for Logistic Regression with Separated or Nearly Separated Data. *Statistics in Medicine* 25: 4216–26.
- 18 Hirji, K. F., C. R. Mehta, and N. R. Patel. 1987. Computing Distributions for Exact Logistic Regression. *Journal of the American Statistical Association* 82: 1110–7.
- 19 Hoetker, G. 2007. The Use of Logit and Probit Models in Strategic Management Research: Critical Issues. *Strategic Management Journal* 28: 331–43.
- 20 Hosmer Jr, D. W., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. New York: Wiley.

- 21 Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression*, 2nd edn. New York: Wiley.
- 22 King, Gary, and Langche Zeng. 2001. Logistic Regression in Rare Events Data. *Political Analysis* 9: 137–63.
- 23 Lin, A. Z. (2013). Variable Reduction in SAS by Using Weight of Evidence and Information Value. *SAS Global Forum Paper* No. 095-213.
- 24 Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables: Analysis and Interpretation*. Thousand Oaks, CA: Sage.
- 25 Lunt, Mark. 2017. *Modelling Binary Outcomes*. Retrieved from [http://personalpages.manchester.ac.uk/staff/mark.lunt/stats/7\\_Binary/text.pdf](http://personalpages.manchester.ac.uk/staff/mark.lunt/stats/7_Binary/text.pdf).
- 26 McCarty, J. A., and M. Hastak. 2007. Segmentation Approaches in Data-Mining: A Comparison of RFM, CHAID, and Logistic Regression. *Journal of Business Research* 60: 656–62.
- 27 Mehta, C. R., and N. R. Patel. 1995. Exact Logistic Regression: Theory and Examples. *Statistics in Medicine* 14: 2143–60.
- 28 Mojsilović, A., B. Ray, R. Lawrence, and S. Takriti. 2007. A Logistic Regression Framework for Information Technology Outsourcing Lifecycle Management. *Computers and Operations Research* 34: 3609–27.
- 29 Olson, D. L., and B. K. Chae. 2012. Direct Marketing Decision Support Through Predictive Customer Response Modeling. *Decision Support Systems* 54: 443–51.
- 30 Pampel, F. C. 2000. *Logistic Regression: A Primer*, Sage University Papers Series on Quantitative Applications in the Social Sciences No. 07–096. Newbury Park, CA: Sage.
- 31 Scott, A. J., and C. J. Wild. 1986. Fitting Logistic Models Under Case-Control or Choice Based Sampling. *Journal of the Royal Statistical Society Series B* 48: 170–82.
- 32 Tonidandel, S., and J. M. LeBreton. 2010. Determining the Relative Importance of Predictors in Logistic Regression: An Extension of Relative Weight Analysis. *Organizational Research Methods* 13: 767–81.
- 33 Zdravevski, E., P. Lameski, and A. Kulakov. 2011. Weight of Evidence as a Tool for Attribute Transformation in the Preprocessing Stage of Supervised Learning Algorithms. In *Neural Networks (IJCNN)*, The 2011 International Joint Conference on IEEE. pp. 181–8.



# Moving beyond the basics

**Structural Equation Modeling:**

**An Introduction**

**SEM: Confirmatory Factor Analysis**

**Testing Structural Equations Models**

**Advanced SEM Topics**

**13 Partial Least Squares Structural  
Equations Modeling (PLS-SEM)**

## SECTION V

### OVERVIEW

**B**ehavioral researchers often propose theoretical process models that suggest how some hypothetical factors may influence, or “cause,” other hypothetical factors. Testing causal sequences proves complicated and the final chapters of the book discuss statistical approaches that are a by-product of the need to test such models. This section provides a simple and concise introduction to a cutting-edge technique in multivariate analysis—structural equation modeling (SEM)—that has grown in popularity over the past 40 years as computational power increased with technology. The ability to simultaneously estimate multiple dependence relationships (similar to multiple regression equations) while also incorporating multiple measures for each concept (i.e., akin to factor

analysis) has been embraced across social sciences. This section provides the reader with a general understanding of the procedure, the knowledge of when it and how it can be applied, and the ability to apply this technique to basic problems.

## CHAPTERS IN SECTION V

Section V contains four chapters. Chapter 9 provides an overview of the analysis of covariance structures using structural equation modeling (SEM), a procedure allowing an omnibus test of a theory represented by a series of equations while accommodating a correction for measurement error directly in the estimation of a series of dependence relationships. It is the best multivariate procedure for testing both the construct validity and theoretical relationships among a set of concepts represented by multiple measured variables. Previous to the introduction of SEM, this process would require the application of several different statistical tools and the result would be a less satisfying examination of theory. We do not wish to underestimate the effort involved, but no researcher should avoid SEM solely for this reason because the principles of factor analysis and multiple regression form a basis for understanding SEM.

Following the basic overview, Chapter 10 is devoted to confirmatory factor analysis, which extends ideas presented earlier when we discussed exploratory factor analysis. Now, however, the researcher must take a more active role in developing and specifying a theory that will determine how many factors should exist among a set of variables and determines which variables relate to those factors. Just as importantly, the theory also dictates which variables are unrelated to factors, each other, or residual variance terms. By constraining models to dictate a lack of relationship, the analyst emphasizes the structure part of SEM and separates the statistical approach from more purely empirical approaches that allow everything to be inter-related. SEM provides a test of how well the theory fits the data and provides detailed results enabling a user to thoroughly examine construct validity for the entire measurement model.

Chapter 11 is devoted to the testing of theoretical relationships between the factors represented by multiple variables. The goal here is to test the structure of relationships among the factors. Therefore, it is conceptually similar to conducting regression analysis using a set of summated rating scales, each summated rating scale representing a factor that can be recovered with factor analysis. Using SEM, the researcher can assess the strength of relationships between any two factors more accurately because SEM will correct the relationship for measurement error. Furthermore, an overall test of fit is provided that enables the researcher to assess theoretical validity of the process represented by a theoretically justified model. Researchers also often theorize competing process models. Relative tests of fit provide an indication of which of a set of competing models is most valid.

Chapter 12 addresses several advanced topics in SEM, notably higher-order confirmatory factor analysis, testing relationships across groups, evaluating moderating and mediating relationships, time-dependent relationship, plus an introduction into Bayesian SEM. These issues extend the range of conceptual questions that SEM can address while maintaining the underlying foundation of measurement theory.

Chapter 13 describes another method of extracting parameter estimates for a series of relationships representing correspondence rules between measured variables and hypothetical constructs and among hypothetical constructs. Partial least squares structural equations models (PLS-SEM) utilizes principal components analysis and OLS regression techniques to develop parameter estimates for both a set of measurement and structural relationships. Although both SEM approaches provide estimates of relationships, PLS-SEM and covariance-based SEM are not interchangeable, being different in both purpose and application. Together, these final chapters provide the reader useful guidance to both approaches.

# 9

# Structural Equation Modeling: An Introduction

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

Understand the distinguishing characteristics of structural analysis.

Distinguish between variables and constructs.

Understand structural equation modeling and how it can be thought of as a combination of familiar multivariate techniques.

Know the basic conditions for causality and how SEM can help establish a cause-and-effect relationship.

Explain the types of relationships involved in SEM.

Understand that the objective of SEM is to explain covariance and determine the fit of a theoretical model.

Know how to represent a SEM model visually with a path diagram.

List the six stages of structural equation modeling and understand the role of theory in the process.

## Chapter Preview

One primary objective of multivariate techniques is to expand the researcher's explanatory ability. Multiple regression, exploratory factor analysis, multivariate analysis of variance, discriminant analysis, and the other techniques discussed in previous chapters all provide the researcher with powerful tools for addressing a wide range of managerial and theoretical questions. They also all share one common limitation: Each technique focuses on individual relationships. Even the techniques allowing for multiple dependent variables, such as multivariate analysis of variance, still are interpreted based on individual relationships between a dependent and independent variable.

All too often, however, the researcher is faced with a set of interrelated generalizations that together represent a theory. For example, what variables determine a retailer's image? How does that image combine with other variables to affect purchase decisions, shopping value, and satisfaction? Do shopping value and satisfaction drive a customer's long-term loyalty? Issues like these contain both managerial and theoretical importance. Yet none of the multivariate techniques we examined thus far enable us to address all these questions with one comprehensive technique. In other words, these techniques do not enable us to test the researcher's entire theory with a technique that considers all possible information. For this reason, we now examine the technique of structural equation modeling (SEM), an extension of several multivariate techniques we already studied, most notably factor analysis and multiple regression analysis.

As briefly described in Chapter 1, structural equation modeling can examine a series of dependence relationships simultaneously. SEM is particularly useful in testing theories that can be represented by multiple equations involving dependence relationships. In other words, if we believe that image creates loyalty because image first creates a

satisfied customer, then satisfaction is both a dependent and an independent variable in the same theory. Thus, a hypothesized dependent variable becomes an independent variable in a subsequent dependence relationship. None of the previous techniques in this book enable us to assess the merits of an hypothesized theoretical model while also assessing its measurement properties. SEM enables researchers to address a multi-variate, and indeed multi-equation, research problem with a single analysis. Note that this chapter as well as Chapters 10, 11, and 12 focus only on covariance-based SEM. In Chapter 13 we describe partial least squares structural equation modeling, referred to as variance-based SEM.

## Key Terms

Before starting the chapter, review the key terms to develop an understanding of the concepts and terminology used. Throughout the chapter the key terms appear in **boldface**. Other points of emphasis in the chapter and key term cross-references are *italicized*. Illustrative examples are in shaded text.

**Absolute fit indices** Measure of overall *goodness-of-fit* for both the *structural* and *measurement models* collectively. This type of measure does not make any comparison to a specified *null model* (*incremental fit measure*) or adjust for the number of parameters in the estimated model (*parsimonious fit measure*).

**All-available approach** Method for handling missing data that computes values based on all available valid observations. Also known as *pairwise deletion*.

**Badness-of-fit** An alternative perspective on *goodness-of-fit* in which larger values represent poorer fit. Examples include the root mean square error of approximation or the standardized root mean square residual.

**Causal inference** *Dependence relationship* of two or more variables in which the researcher clearly specifies that one or more variables cause or bring about an outcome represented by at least one other variable. It must meet the requirements for *causation*.

**Causation** Principle by which changes in an entity (independent variable/exogenous construct) bring about changes in an outcome entity (dependent/endogenous). Causal inferences require a sufficient degree of association (covariance) between the two variables, that the cause occurs before the effect, and that the observed relationship would not vanish when or if other potential causes were introduced. Strong theoretical support enhances empirical support of causation.

**Chi-square ( $\chi^2$ )** Statistical assessment of fit between the *observed* and theoretically implied *estimated* covariance matrices. Thus, it empirically assesses the accuracy of a proposed theoretical structure. It is an omnibus fit measure that has a direct assessment of statistical significance, and it forms the basis for most other *goodness-of-fit* indices.

**Chi-square difference statistic ( $\Delta\chi^2$ )** Competing, nested SEM theoretical models can be compared using this statistic, which is the simple difference between each model's  $\chi^2$  statistic. It has *degrees of freedom* equal to the difference in the models' degrees of freedom.

**Communality** Total amount of variance a *measured variable* has in common with the *constructs* upon which it theoretically loads. It represents the amount of variance in a measured variable explained by the factor it indicates. Sometimes referred to as squared multiple correlation or item-reliability. Also see *variance extracted* in the next chapter.

**Competing models strategy** Modeling strategy that compares the proposed model with plausible, alternative models. This approach is particularly relevant in *structural equation modeling*, because a model can be shown only to have acceptable fit, but acceptable fit alone does not guarantee that another model will not fit better or equally well.

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

**Confirmatory analysis** Use of a multivariate technique to test (confirm) a prespecified structure or relationship. For example, suppose we hypothesize that X causes M which in turn causes Y. If we empirically test this overall theoretical process and the implied relationships and nonrelationships, the test is a confirmatory analysis. It is the opposite of *exploratory analysis*.

**Confirmatory modeling strategy** Strategy that statistically assesses a single model for its fit to the observed data. This approach is actually less rigorous than the *competing models strategy*, because it does not consider alternative models that might fit better or equally well than the proposed model.

**Construct** Unobservable or *latent* concept that the researcher can define in conceptual terms but cannot measure directly or without error (see *measurement error*). A construct can be defined in varying degrees of specificity, ranging from quite narrow concepts to more complex or abstract concepts, such as intelligence or emotions. Latent constructs are approximately measured by multiple *indicator (measured) variables*.

**Construct validity** Extent to which a set of *measured variables* actually represents the theoretical *latent construct* they are designed to measure. Discussed in detail in Chapter 13.

**Degrees of freedom (df)** In SEM models, degrees of freedom are the number of nonredundant covariances/correlations (moments) in the input matrix minus the number of estimated coefficients. Each estimated (free) coefficient "uses up" a degree

of freedom. A model can never estimate more coefficients than the number of nonredundant correlations or covariances, meaning that zero is the lower bound for the degrees of freedom for any model.

**Dependence relationship** A regression type of relationship represented by a one-headed arrow flowing from an independent variable or *construct* to a dependent variable or construct. Typical dependence relationships in SEM connect constructs to measured variables and predictor (exogenous) constructs to outcome (endogenous) constructs. Dependence relationships imply causality.

**Endogenous constructs** *Latent*, multi-item equivalent to dependent variables. An endogenous construct is represented by a *variate* of dependent variables. In terms of a *path diagram*, one or more arrows lead into the endogenous construct.

**Estimated covariance matrix** Covariance matrix composed of the model-estimated covariances between all *indicator variables* involved in a SEM based on the equations that represent the hypothesized theoretical model. Typically abbreviated with  $\Sigma_k$ .

**Exogenous constructs** *Latent*, multi-item equivalent of independent variables. They are *constructs* determined by factors outside of the model.

**Exploratory analysis** Empirically oriented analysis that allows the multivariate technique to discover possible relationships. The opposite of *confirmatory analysis*, the researcher is not looking to confirm any relationships specified prior to the analysis, but instead lets the statistical technique define the nature of the relationships.

**Fit** See *goodness-of-fit*.

**Fixed parameter** Parameter that has a value specified by the researcher. Most often the value is specified as zero, indicating no relationship, and in a model, is represented by the absence of a connection.

**Free parameter** Parameter estimated by the structural equation program to represent the strength of a specified relationship. These parameters may occur in the *measurement model* (most often denoting loadings of *indicators* to *constructs*) as well as the *structural model* (relationships among constructs).

**Goodness-of-fit (GOF)** Measure indicating how well a specified model structure reproduces the covariance matrix among the *indicator* variables, alternatively, the accuracy of a proposed theory.

**Imputation** Process of estimating the missing data of an observation based on characteristics of non-missing data. One simple imputation method is substituting the variable mean using the non-missing observations and substituting it for missing values. See also *all-available*, *complete case*, and *model-based approaches* for missing data.

**Incremental fit indices** *Goodness-of-fit* indices that assesses how well a specified model fits relative to some alternative baseline model. Most commonly, the baseline model is a *null model* specifying that all *measured variables* are unrelated to each other. Complements the other two types of goodness-of-fit measures, the *absolute fit* and *parsimonious fit* measures.

**Indicator** Observed value (also called a *measured* or *manifest variable*) used to reflect a *latent construct* that cannot be measured directly. The researcher must specify which indicators are associated with each latent construct.

**Latent construct** A real phenomenon that cannot be measured directly but can be represented or measured by one or more variables (*indicators*). In combination, the indicators give a reasonably accurate measure of the latent construct when good psychometric properties exist.

**Latent factor** See *latent construct*.

**Latent variable** See *latent construct*.

**LISREL** The first widely used SEM software. The name is derived from LInear Structural RELations.

**LISREL Notation** A commonly used method of expressing SEM models that corresponds to matrices used by LISREL. The matrices, such as lambda, beta and gamma, represent specific components in an SEM model. Although the notation is specific to the LISREL program, the widespread use of LISREL has popularized the terminology when describing models and results.

**Manifest variable** See *measured variable*.

**Maximum likelihood estimation (MLE)** A robust estimation method commonly employed in structural equation models. An alternative to ordinary least squares, MLE is a procedure that yields consistent parameter estimates that are “most likely” to have produced the observed data.

**Measured variable** Observed (measured) value for a specific item or question, obtained either from respondents in response to questions (as in a questionnaire) or from some type of observation. Measured variables are used as the *indicators* of *latent constructs*. Same as *manifest variable*.

**Measurement error** Degree to which the indicator variables do not perfectly represent the *latent construct(s)* of interest. For all practical purposes, all constructs and all variables have some measurement error. However, the researcher’s objective is to minimize the amount of measurement error. SEM estimation corrects structural relationships for measurement error in order to provide more accurate estimates of the relationships between constructs.

**Measurement model** A theoretically derived *model* (1) specifying how the *indicators* correspond to latent *constructs* and (2) enabling an assessment of *construct validity*. The first of the two major steps in a complete *structural model* analysis discussed in more detail in Chapter 13.

**Measurement relationship** *Dependence relationship* between *indicators* or *measured variables* and their associated *construct(s)*.

A common specification depicts the construct “causing” or giving rise to the indicators, thus the arrows point from the construct to the indicators. An alternative specification reverses the relationship.

**Missing at random (MAR)** Classification of missing data applicable when missing values of  $Y$  depend on  $X$ , but not on  $Y$ . When missing data are MAR, observed data for  $Y$  are a truly random sample for the  $X$  values in the sample, but not a random sample of all  $Y$  values, due to missing values of  $X$ .

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

**Model** Representation and operationalization of a theory. A conventional model in SEM terminology consists of two parts. The first part is the *measurement model*, which represents theory explaining how *measured variables* come together to represent *constructs*. The second part is the *structural model*, which represents the processes through which constructs are associated with each other, often involving multiple *dependence relationships*. The model can be formalized in a *path diagram*.

**Model respecification** Modification of an existing *model* with estimated parameters to correct for inadequacies in the fit of a previously estimated model or to create a *competing model* for comparison using new data.

**Models-based approach** Replacement approach for missing data in which values for missing data are estimated based on all non-missing data for a given respondent. Most widely used methods are maximum likelihood estimation (ML) of missing values and EM, which involves maximum likelihood estimation of the means and covariances given missing data.

**Multicollinearity** Extent to which *constructs* or variables overlap with each other. As multicollinearity among predictors increases, it complicates the interpretation of regression effects because it biases estimates and makes it more difficult to ascertain the true effect of any single construct.

**Nested model** *Model* is nested within another model if it contains the same number of *constructs* and can be formed from the other model by altering the relationships. The most common form of nested model occurs when a single relationship is added to or deleted from another model. Thus, the model with fewer estimated relationships is nested within the more general model.

**Null model** Baseline or comparison standard used in *incremental fit indices*. The null model is hypothesized to be the simplest *model* that can be theoretically justified.

**Observed sample covariance matrix** Typical input matrix for SEM estimation composed of the observed variances and covariances for each *measured variable*. Typically abbreviated with a bold, capital letter  $S$  ( $\mathbf{S}$ ).

**Operationalizing a construct** Key process in the *measurement model* involving determination of the *measured variables* that will represent a *construct* and the way in which they will be measured.

**Parsimony fit indices** Measures of overall *goodness-of-fit* representing the degree of model fit per estimated coefficient. This measure attempts to correct for any overfitting of the *model* and evaluates the parsimony of the model compared to the goodness-of-fit. These measures complement the other two types of goodness-of-fit measures, the *absolute fit* and *incremental fit* measures.

**Path analysis** General term for an approach that employs simple bivariate correlations to estimate relationships in a SEM *model*. Path analysis seeks to determine the strength of the paths shown in *path diagrams*.

**Path diagram** A visual representation of a *model* and the complete set of relationships among the model's *constructs*. *Dependence relationships* are depicted by straight arrows, with the arrow emanating from the predictor variable and the arrowhead pointing to the dependent construct or variable. Curved arrows represent correlations between constructs or *indicators*, but no causation is implied.

**Reduced form equation** An equation predicting an endogenous variable/construct in a single equation using only and all exogenous constructs (or independent variables) involved in an analysis as predictors.

**Reliability** Measure of the degree to which the *indicators* of a *latent construct* are internally consistent with each other. The indicators of highly reliable *constructs* are highly interrelated, indicating that they all seem to measure the same thing. Individual item reliability can be computed as 1.0 minus the *measurement error variance*. Note that high reliability is a necessary, but not sufficient, condition for validity.

**Residual** The difference between the actual and estimated value for any relationship. In SEM analyses, residuals are the differences between the *observed* and *estimated covariance matrices*.

**Spurious relationship** A relationship that is false or misleading. A common occurrence in which a relationship can be spurious is when an omitted construct variable explains relates to both cause effect (i.e., relationship between original *constructs* becomes nonsignificant upon adding omitted construct).

**Structural equation modeling (SEM)** Multivariate technique combining aspects of factor analysis and multiple regression that enables the researcher to simultaneously examine a series of interrelated *dependence relationships* among the *measured variables* and *latent constructs (variates)*, as well as between several latent constructs.

**Structural model** Set of one or more *dependence relationships* linking the hypothesized model's *constructs*. The structural model is most useful in representing the interrelationships of variables between *constructs*.

**Structural relationship** *Dependence relationship* (regression type) specified between any two *latent constructs*. Structural relationships are represented with a single-headed arrow and suggest that one *construct* is dependent upon another. *Exogenous constructs* cannot be dependent on another construct. *Endogenous constructs* are dependent on either exogenous or endogenous constructs (see Chapter 11 for more detail).

**Theory** A systematic set of relationships providing a consistent and comprehensive explanation of phenomena. In practice, a theory is a researcher's attempt to specify the entire set of *dependence relationships* explaining a particular set of outcomes. A theory is a reasoned explanation, not just a precision, of each variable's correspondence to all others in the model.

**Variate** A linear combination of *measured variables* that represents a *latent construct*.

## What Is Structural Equation Modeling?

**Structural equation modeling (SEM)** is a family of statistical models that seeks to explain the relationships among multiple variables. In doing so, SEM examines the *structure* of interrelationships expressed in a series of equations, similar to a series of multiple regression equations. These equations depict all relationships among **constructs** (the dependent and independent variables) and variables involved in the analysis. Just as importantly, a theoretical structure specifies which variables and constructs are not likely to be related to one another. Constructs are unobservable, or **latent factors**, represented by multiple variables. In previous chapters, each multivariate technique has been classified either as an interdependence or dependence technique. SEM can be thought of as a unique combination of both types of techniques, because SEM's foundation lies in two familiar multivariate techniques: factor analysis and multiple regression analysis.

SEM is known by many names: covariance structure analysis, latent variable analysis, and sometimes users even refer to it by the name of the specialized software package used (e.g., a LISREL or AMOS model). SEM models are distinguished from traditional regression models in that they tend to involve:

- 1 Simultaneous estimation of multiple and interrelated dependence relationships
- 2 An ability to represent unobserved concepts in these relationships and account for measurement error in the estimation process
- 3 Defining a theoretical model to explain the entire set of relationships
- 4 Over-identifying assumptions (meaning variables are explained by a unique set of variables that does not include all possible relationships)

### ESTIMATION OF MULTIPLE INTERRELATED DEPENDENCE RELATIONSHIPS

One difference between SEM and other multivariate techniques is the use of separate relationships for each of a set of dependent variables. In simple terms, SEM estimates a series of separate, but interdependent, multiple regression equations simultaneously by specifying the **structural model** used by the statistical program. First, the researcher draws upon theory, prior experience, and the research objectives to distinguish which independent variables explain each dependent variable. Dependent variables in one relationship can become independent variables in subsequent relationships, giving rise to the interdependent nature of the structural model. Moreover, many of the same variables affect each of the dependent variables, but with differing effects. The structural model expresses these **dependence relationships** among independent and dependent variables, even when a dependent variable becomes an independent variable in other relationships.

The proposed relationships are then translated into a series of structural equations (similar to regression equations) for each dependent variable. This feature sets SEM apart from other multivariate techniques that accommodate multiple dependent variables—multivariate analysis of variance and canonical correlation—in that they allow only a single relationship between dependent and independent variables.

Structural equations are different from **reduced form equations**. A reduced form equation solves for a single endogenous construct (or dependent variable) in a single equation with all and only exogenous constructs (or independent variables) employed as predictors. Moreover, if a researcher is interested in more than one endogenous construct, a separate reduced form equation would be needed for each, which would again include all and only exogenous variables. In reduced form equations, endogenous constructs cannot predict other endogenous constructs. This differs from structural equations, which are parsimonious and include only the specific predictors, endogenous or exogenous, that are theoretically linked to the outcome construct. SEM solves structural equations

simultaneously, whereas reduced form equations are solved individually. Reduced form parameter estimates do not relate to any specific relationship or process among constructs, as in a structural model, but instead represent a total or composite estimate of the exogenous variable's impact on a single dependent variable. In this sense, structural equations represent theoretically or deductively derived models, while reduced form equations identify relationships inductively by including all predictors without regard to a prespecified theory.

## INCORPORATING LATENT VARIABLES NOT MEASURED DIRECTLY

SEM also has the ability to incorporate latent variables into the analysis. A **latent construct** (also termed a **latent variable**) is a hypothetical, unobserved concept that can be represented by observed or measured variables. Latent constructs are measured indirectly by examining multiple **measured variables**, sometimes referred to as **manifest variables**, or **indicators**. By any name, the indicators are variables that are directly assessing some specific aspect or concept. Most typically, particularly in psychometrics, the measured variables are individual survey item responses. But increasingly the measured variables included in SEM are obtained from digital Big Data (secondary or archival data) such as social media, mobile phone usage patterns, geographic movements (GPS), and so forth, as well as from data warehouses with organizational information, most often stored in the cloud.

**The Benefits of Using Latent Constructs** Yet why would we want to use a latent variable that we cannot measure directly instead of the direct measures provided by respondents? First, we can indeed represent latent, theoretical concepts using multiple indicators of a concept, which reduces the measurement error over relying on a single indicator. Second, SEM can account for the measurement error associated with latent constructs and correct for it to make more accurate statistical estimations of relationships among constructs.

**REPRESENTING THEORETICAL CONCEPTS** We introduced in Chapter 3 the notion that most concepts require multiple assessments for adequate representation. From a theoretical perspective, most concepts are relatively complex (e.g., patriotism, consumer confidence, self-identity) and have many aspects and/or dimensions. With complexity, the researcher tries to design the best items to measure the concept knowing that individuals may interpret any single item somewhat differently. The intent is for the collective set of questions to represent the concept better than any single item [13].

Moreover, the researcher must also be aware of measurement error that occurs with any form of measurement. Although we may be able to minimize it with physical concepts such as time (e.g., measurement with atomic clocks), any more theoretical or abstract concept is necessarily subject to measurement error. In its most basic form, measurement error is due to inaccurate representation of the concept. But, more importantly, measurement error occurs when respondents may be somewhat unsure about how to respond or may interpret the questions in a way that is different from what the researcher intended. Finally, it can result from a natural degree of inconsistency on the part of the respondent when we use multiple perspectives or items to measure the same concept. All of these situations give rise to measurement error. If we know the magnitude of the problem, we can incorporate the extent of the measurement error into the statistical estimation and improve our dependence model.

How do we represent concepts theoretically and quantify the amount of measurement error? SEM includes a **measurement model** that specifies the theoretical correspondence rules between measured and latent variables (constructs). The measurement model enables the computation of a proxy measure of the construct to represent any single independent or dependent construct with multiple items. By testing the fit of the theoretical measurement model against reality, one can assess the degree of measurement error present.

As an example, let us consider the following situation in developing a measurement model for HBAT. HBAT would like to determine which factors are influencing the job satisfaction of its employees. The dependent (outcome) variable is job satisfaction, and the two independent variables are how they feel about their supervisor and how they like their work environment. Each of these three variables can be defined as a latent construct. Each latent construct would be measured with several indicator variables. For example, how employees feel about their supervisor might be measured by the following three indicator variables: (1) My supervisor recognizes my potential; (2) My supervisor

helps me resolve problems at work; and (3) My supervisor understands that I have a life away from work. The researcher identifies the specific indicator variables associated with each construct, typically based on a combination of previous similar studies and the situation at hand. When SEM is applied, the researcher can assess the contribution of each indicator variable in representing its associated construct and measure how well the combined set of indicator variables represents the construct (reliability and validity). This is the measurement assessment component of SEM. After the constructs have met the required measurement standards, the theoretical model representing the way constructs are related to each other can be evaluated. This is the structural assessment component of SEM.

**IMPROVING STATISTICAL ESTIMATION** All the multivariate techniques reviewed in previous chapters overlook any measurement error present in our variables. As has been discussed, we know from both practical and theoretical perspectives that we cannot perfectly measure a concept and that some degree of **measurement error** is always present. For example, when asking about something as straightforward as household income, we know some people will answer incorrectly, either intentionally overstating or understating the amount or just not knowing it precisely. The answers contain measurement error, which affects the estimate of the structural coefficient between constructs.

**Reliability** is a measure of the degree to which a set of *indicators* of a *latent construct* is internally consistent based on how highly interrelated the indicators are with each other. In other words, reliability represents the extent to which multiple indicators all converge. Reliability does not guarantee, however, that the measures indicate only a single concept. We discuss this more in the next chapter. Generally, reliability is inversely related to measurement error. As reliability goes up, the relationships between a construct and the indicators are greater, meaning that the construct explains more of the variance in each indicator. In this way, high reliability is associated with lower measurement error. Note that the comments in this paragraph are associated with reflective measurement models, and not formative measurement models, both of which are discussed in Chapters 10 and 13.

Statistical theory tells us that a regression coefficient is actually composed of two elements: the *true* structural coefficient between the dependent and independent variable and the reliabilities. The impact of measurement error (and the corresponding reliability) can be illustrated by an expression of a regression coefficient as:

$$\beta_{yx} = f(B_x * \rho_x)$$

where  $\beta_{yx}$  is the observed regression coefficient,  $\beta_s$  is the true structural coefficient, and  $\rho_x$  is the reliability of the predictor variable. What SEM does is make an estimate of the true structural coefficient ( $\beta_s$ ) based on the estimated regression coefficient. This is a critical point, because *unless the reliability is 100 percent (i.e., no measurement error), the observed correlation (and resulting regression coefficient) will always underestimate the true relationship*. So SEM “corrects for” or “accounts for” the amount of measurement error in the variables (latent constructs) and estimates what the relationship would be if there was no measurement error. These are the estimates of the causal relationships in the structural model between constructs. The above example presumes measurement error only in the predictor. In reality, measurement error usually exists in both the outcome and the predictor.

Thus, the relationships we can estimate through regression models will always be weaker in the presence of measurement error (this makes sense when we think about it, because error can only detract from the true relationship). The equation means that relationships estimated with other multivariate procedures will underestimate the actual or true relationship because reliability can only take on values between 0 (meaning no reliability) and 1 (meaning 100 percent reliability). So, if one knows the reliability of measures and the observed regression coefficient, the true regression relationship can be found as a function of the observed regression coefficient divided by the square root of the product of the reliabilities for each construct—predictor and outcome. SEM offers the advantage of automatically applying such a correction. The parameter estimates are corrected for attenuation due to measurement error and should be more accurate than those found when using other approaches. Because the SEM relationship coefficients are corrected in this fashion, they will tend to be larger than coefficients obtained when multiple regression is used.

Although reliability is important, high reliability does not guarantee that a construct is measured accurately. That conclusion involves an assessment of validity, which is discussed in the next chapter. Reliability is a necessary, but not sufficient, condition for validity.

**Distinguishing Exogenous Versus Endogenous Latent Constructs** Recall that in multiple regression, multiple discriminant analysis, and MANOVA, it was important to distinguish between independent and dependent variables. Likewise, in SEM a similar distinction must be made. However, because we are now generally predicting latent constructs with other latent constructs, a different terminology is used.

**Exogenous constructs** are the latent, multi-item, equivalent of independent variables. As such, they use a **variate** of measures to represent the construct, which acts as an independent variable in the model. They are determined by factors outside of the model (i.e., they are not explained by any other construct or variable in the model), thus the term *independent*. SEM models are often depicted by a visual diagram. It is useful, therefore, to know how to identify an exogenous construct. Given that it is independent of any other construct in the model, visually an exogenous construct does not have any paths (single-headed arrows) from any other construct or variable going into it. We discuss the issues in constructing the visual diagram in the next section.

**Endogenous constructs** are the latent, multi-item equivalent to dependent variables. (i.e., a variate of individual dependent variables). These constructs are theoretically determined by factors within the model. Thus, they are dependent on other constructs, and this dependence is represented visually by a path to an endogenous construct from an exogenous construct (or from another endogenous construct, as we will see later).

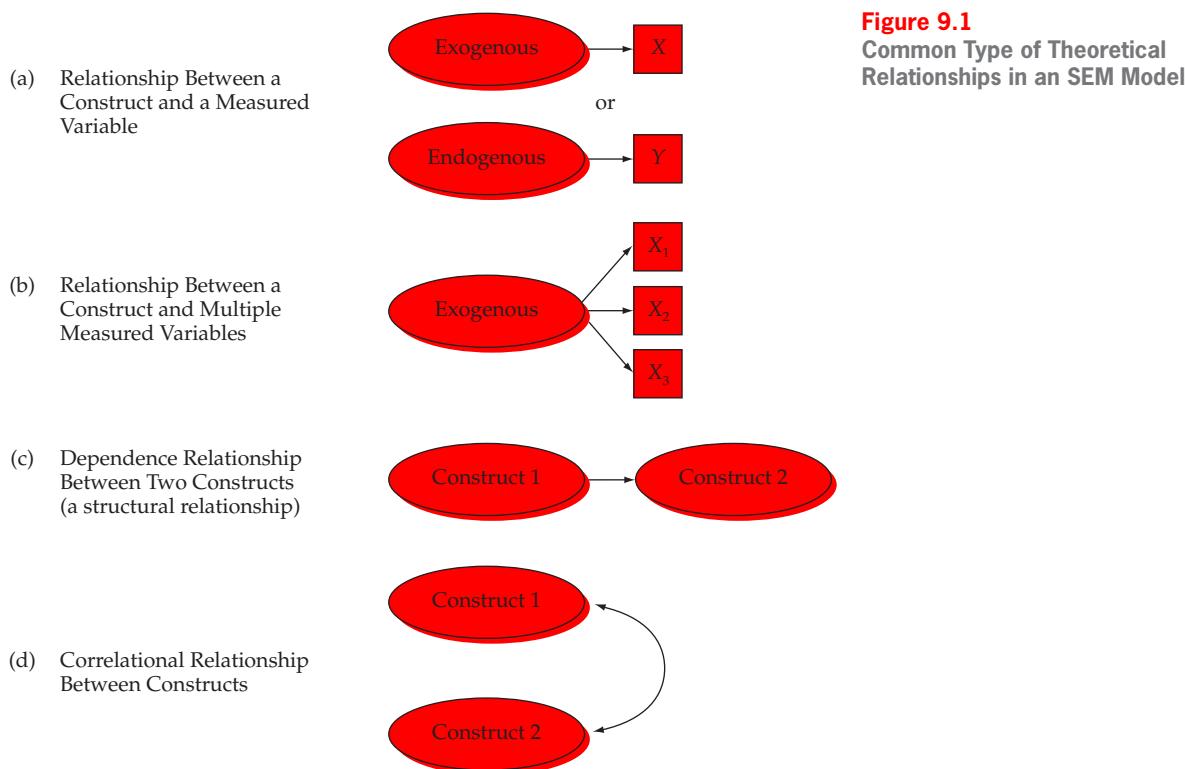
## DEFINING A MODEL

A **model** is a representation of a theory. **Theory** can be thought of as a systematic set of relationships providing a consistent and comprehensive *explanation* of phenomena. From this definition, we see that theory is not the exclusive domain of academia, but can be rooted in experience and practice obtained by observation of real-world behavior. A conventional model in SEM terminology consists of really two theories, the measurement model (representing how measured variables come together to represent constructs) and the structural model (showing how constructs are associated with each other). Chapter 10 is devoted to the first part of SEM, or the measurement model, whereas Chapter 11 addresses issues in the second part of SEM, or the structural model.

**Importance of Theory** A model should not be developed without some underlying theory. Theory is often a primary objective of academic research, but practitioners may develop or propose a set of relationships that are as complex and interrelated as any academically based theory. Thus, researchers from both academia and industry can benefit from the unique analytical tools provided by SEM. We will discuss in a later section specific issues in establishing a theoretical base for your SEM model, particularly as it relates to establishing causality. In all instances, SEM analyses should be dictated first and foremost by a strong theoretical base that portrays how things are related, and at least just as importantly, what things are not related.

**A Visual Portrayal of the Model** A complete SEM model consisting of measurement and structural models can be quite complex. Although all of the relationships can be expressed in path analysis notation, many researchers find it more convenient to portray a model in a visual form, known as a **path diagram**. This visual portrayal of the theoretical relationships employs specific conventions both for the constructs and measured variables as well as the associations and independences among them.

**DEPICTING THE CONSTRUCTS INVOLVED IN A STRUCTURAL EQUATIONS MODEL** Latent constructs are connected to corresponding measured variables with a **measurement relationship**. This is a type of dependence relationship (depicted by a straight arrow) between measured variables and constructs. In a typical SEM, the arrow is drawn from the latent constructs to the variables that are associated with the constructs. These variables are referred to as *indicators*, because no single variable can completely represent a construct, but it can be used as an indication of the construct. The researcher must justify the theoretical basis of the indicators, because SEM only examines the empirical characteristics of the variables. An alternative specification where the arrows point from the indicators toward the construct will be discussed later in this chapter and in more detail in Chapter 10. We will also discuss how to assess the quality



of the indicators of the constructs in a SEM model. Here, we focus on the basic principles in constructing a diagram of a measurement model:

- Constructs typically are represented by ovals or circles, and measured variables are represented by squares or rectangles.
- To assist in distinguishing the indicators for endogenous versus exogenous constructs, measured variables (indicators) for exogenous constructs are usually referred to as  $X$  variables, whereas endogenous construct indicators are usually referred to as  $Y$  variables.
- The  $X$  and/or  $Y$  measured variables are associated with their respective construct(s) by a single-headed straight arrow from the construct(s) to the measured variable.

Figure 9.1a illustrates the measurement relationship between a construct and one of its measured variables. Because constructs will likely be indicated by multiple measured variables, the more common depiction is as in Figure 9.1b. Remember that the indicators are labeled as either  $X$  or  $Y$ , depending on whether they are associated with an exogenous or endogenous construct, respectively.

**DEPICTING STRUCTURAL RELATIONSHIPS** A structural model involves specifying **structural relationships** between latent constructs. Specifying a relationship means that we either specify that a relationship exists by drawing an arrow, which if we do not specify a value for frees that path to be estimated, or we specify that no relationship exists by constraining a path to 0 by excluding any connection. Two types of relationships are possible among constructs: dependence relationships and correlational (covariance) relationships.

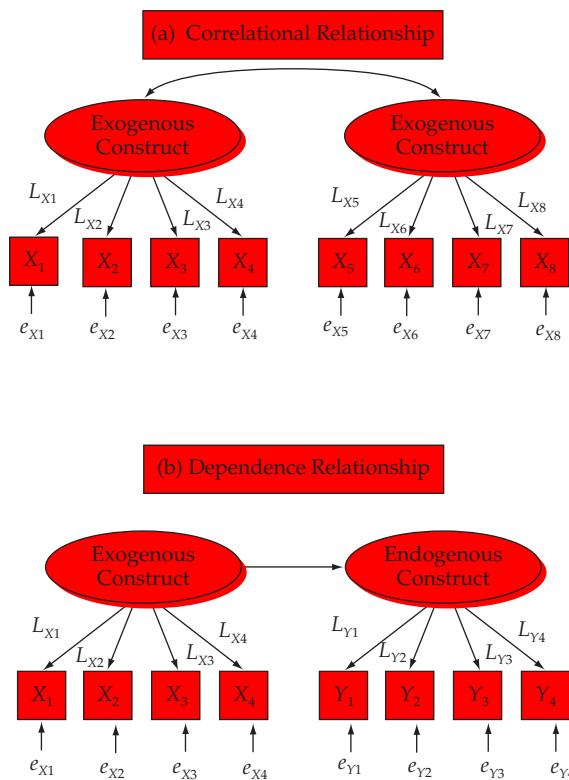
As we discussed earlier, measurement relationships are one form of dependence relationship between constructs to variables. The second form is a dependence relationship between constructs. Here the arrows point from the antecedent (independent variable) to the subsequent effect or outcome (dependent variable). This relationship is

depicted in Figure 9.1c. In a later section, we discuss issues involved in specifying causation, which is a special form of dependence relationship.

Specification of dependence relationships also determines whether a construct is considered exogenous or endogenous. Recall that an endogenous construct acts like a dependent variable, and any construct with a dependence path (arrow) pointing to it is considered endogenous. Whenever a construct is related to other constructs or variables, other than its own indicators, or error variance terms, it is in fact endogenous to some degree. An exogenous construct can only display correlational relationships with other exogenous constructs and acts like an independent variable in structural relationships with endogenous constructs.

In many instances, one may specify a simple correlation between exogenous constructs. This type of relationship does not imply dependence and is depicted by a two-headed arrow connection, as shown in Figure 9.1d. An exogenous construct cannot share this type of relationship with an endogenous construct. Only a dependence relationship can exist between exogenous and endogenous constructs.

**COMBINING MEASUREMENT AND STRUCTURAL RELATIONSHIPS** Figure 9.2 illustrates a simple SEM model incorporating both the measurement and structural relationships of two constructs with four indicators each. In Figure 9.2a, there is a correlational relationship between the two constructs, indicated by the curved arrow. The indicators (four on each construct) are labeled  $X_1$  to  $X_8$ . Figure 9.2b depicts a dependence relationship between the exogenous and endogenous construct. The two constructs retain their same indicators, but two changes distinguish it from the correlational relationship. First, the indicators of the exogenous constructs are denoted by  $X_1$  to  $X_4$ , whereas the endogenous indicators are  $Y_1$  to  $Y_4$ . The measured variables themselves did not change at all, just their designation in the model. Second, the single dependence relationship between the exogenous construct and the endogenous construct is depicted by the straight arrow between the constructs that replaces the curved arrow.



**Figure 9.2**  
Visual Representation of Measurement and Structural Model Relationships in a Simple SEM Model

The researcher determines whether constructs are exogenous or endogenous based on theory. A single SEM model most likely will contain both dependence and correlational relationships.

**How Well Does the Model Fit?** In contrast to OLS regression or other dependence techniques that seek to predict relationships in a single equation, the statistical goal of covariance-based SEM is to reproduce the observed covariance matrix of all measured variables needed to test a theory. Therefore, measures of predictive accuracy for other techniques (i.e.,  $R^2$  for multiple regression or PLS-SEM, classification accuracy in discriminant analysis, or statistical significance in MANOVA) are not the primary, statistical objective of covariance-based SEM. What is needed is a measure of fit or explanatory accuracy that reflects the overall model, not individual relationships. The researcher “accepts or rejects” the theoretical model tested with CB-SEM by assessing how closely the theoretical model fits the observed data.

Because the focus is on the entire theoretical model, CB-SEM relies on the observed covariance matrix among measured variables, which contains full information about how all variables correspond to each other. Model fit is determined by the resulting similarity between the observed covariance matrix and an estimated covariance matrix produced from the equations representing the proposed theoretical model. If the proposed theory creates equations that reproduce precisely the correspondence among measured variables (observed covariance matrix), then we can say the theory fits reality. Practically, fit also allows insight into relative fit, allowing the researcher to know one theory fits better than an alternative. We will discuss the process of estimating a covariance matrix from the proposed model, along with a number of measures of fit, in greater detail in later sections of this chapter as well as in Chapters 10 and 11.

## SEM and Other Multivariate Techniques

SEM is a multivariate technique based on variates explaining both the measurement and structural correspondences. In the measurement models, each set of indicators for a construct acts collectively (as a variate) to define the construct. In the structural model, constructs are related to one another in correlational and dependence relationships. SEM is most appropriate when the researcher has multiple constructs, each represented by several measured variables, and these constructs are distinguished based on whether they are exogenous or endogenous. In this sense, SEM is similar to other multivariate dependence techniques, such as MANOVA and multiple regression analysis. Moreover, the measurement model looks similar in form and function to exploratory factor analysis, because in fact it is very similar to exploratory factor analysis. We will discuss the similarities of SEM to both dependence and interdependence techniques in the following sections.

### SIMILARITY TO DEPENDENCE TECHNIQUES

SEM obviously is similar to multiple regression, one of the most widely used dependence techniques. Relationships for each individual endogenous construct can be expressed in a regression equation. The endogenous construct is the dependent variable, and the independent variables are the constructs, with arrows pointing to the endogenous construct. One principal difference in SEM is that a construct that acts as an independent variable in one relationship can be the dependent variable in another relationship. SEM then allows for all the relationships/equations to be estimated simultaneously.

SEM can also be used to represent other dependence techniques. Variations of the standard SEM models can be used to represent nonmetric, categorical variables, and even a MANOVA model can be examined using SEM. It enables the researcher to take advantage of SEM's ability to accommodate measurement error, for example, within a MANOVA context.

### SIMILARITY TO INTERDEPENDENCE TECHNIQUES

At first glance, the measurement model, associating measured variables with constructs, seems identical to exploratory factor analysis described in Chapter 3 where variables load on factors. Despite a great deal of similarity, such as the interpretation of the strength of the relationship of each variable to the construct (known as a *loading* in

exploratory factor analysis), one difference is critical. Exploratory factor analysis described in Chapter 3 is an **exploratory analysis** technique that searches for structure among variables by defining factors in terms of sets of variables. As a result, every variable loads on every factor.

SEM is a confirmatory procedure, meaning the opposite of an exploratory technique. The researcher *a priori* specifies which variables are associated with each construct and which variables are not associated with a construct or with other variables. Then loadings are estimated only where variables are associated with constructs. Typically, cross-loadings are constrained to 0 (no cross-loading paths). Exploratory factor analysis requires no such specification on the part of the researcher. In contrast, SEM requires complete specification of the measurement model. The non-specified but possible relationships help to over-identify the model.

The advantages of using multiple measures for a construct, discussed earlier and in Chapter 3, are realized through the measurement model in SEM. In this way, the estimation procedures for the structural model can include a direct correction for measurement error, as discussed earlier. By doing so, the relationships between constructs are estimated more accurately.

## THE EMERGENCE OF SEM

SEM's roots extend back to the first half of the twentieth century. SEM's development originated with the desires of genetics and economics researchers to be able to establish causal relationships between observed variables [8, 19, 55]. The mathematical complexity of SEM limited its application until computers and software became widely available. They enabled the two multivariate procedures of factor analysis and multiple regression to be combined. During the late 1960s and early 1970s, the work of Jöreskog and Sörbom led to simultaneous maximum likelihood estimation of a theory represented by relationships between latent constructs and measured indicator variables and among latent constructs (and the corresponding lack of relationships). This work culminated in the SEM program **LISREL** [27, 28, 29, 30]. LISREL was the first SEM software to gain widespread usage.

SEM's growth remained relatively slow during the 1970s and 1980s, in large part due to its perceived complexity. By 1994, however, more than 150 SEM articles were published in the academic social science literature. That number increased to more than 300 by 2000, and today SEM is "the dominant multivariate technique," followed by multiple regression, cluster analysis and MANOVA [23].

## The Role of Theory in Structural Equation Modeling

SEM should never be applied without a strong theoretical basis for specification of both the measurement and structural models. The following sections address some fundamental roles played by theory in SEM: (1) specifying relationships that define the model; (2) establishing causation, particularly when using cross-sectional data; and (3) the development of a modeling strategy.

### SPECIFYING RELATIONSHIPS

Although theory can be important in all multivariate procedures, it is critically important for SEM, because it is considered a **confirmatory analysis**. That is, it is useful for testing and potentially confirming theory. Theory is needed to specify what and how things are related and are not related to each other in both measurement and structural models. Theory provides a pattern of relationships and non-relationships that end up imposing a structure upon the data.

From a practical perspective, a theory-based approach to SEM is necessary, because all potential relationships and non-relationships must be specified by the researcher before a SEM model can be estimated. With other multivariate techniques, the researcher may have been able to specify a basic model and allow default values in the statistical programs to "fill in" the remaining estimation issues. Thus, when we stress the need for theoretical justification, we are emphasizing that SEM is a confirmatory method more about testing theory than exploring empirical relationships.

The relationships in a path diagram typically involve a combination of dependence and correlational relationships among exogenous and endogenous constructs. Any concepts not connected are theorized to be independent. The researcher can specify any sequence of relationships that make theoretical/logical sense derived from the research

question. The following examples illustrate how a sequence of relationships can involve both dependence and correlational elements, including some variables acting as both predictors and outcomes.

Figure 9.3 shows three examples of relationships depicted by path diagrams, along with the corresponding equations. Figure 9.3a shows a simple three-construct model. Both  $X_1$  and  $X_2$  are exogenous constructs related to the endogenous construct  $Y_1$ , and the curved arrow between  $X_1$  and  $X_2$  allows for intercorrelation (multicollinearity). We can show this model with a single equation with  $Y_1$  as a function of  $X_1$  and  $X_2$ , much as we did in our discussion of multiple regression.

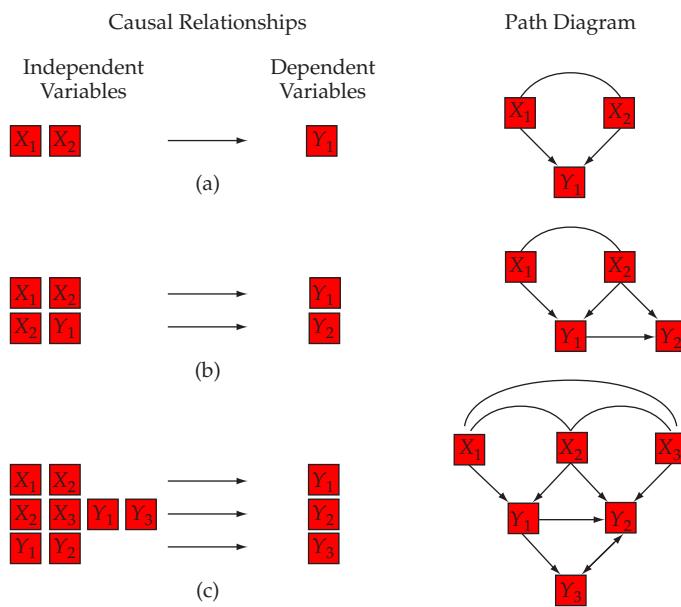
In Figure 9.3b, we add a second endogenous construct— $Y_2$ . Now, in addition to the model and equation shown in Figure 9.3a, we add a second equation showing the relationship between  $X_2$  and  $Y_1$  with  $Y_2$ . Here we can see the unique role played by SEM. We want to know the effects of  $X_1$  on  $Y_1$ , the effects of  $X_2$  on  $Y_1$ , which would be the same as frame a, but simultaneously we consider the effects of  $X_2$  and  $Y_1$  on  $Y_2$ .  $Y_1$  then serves as both independent and dependent variable. In the end, SEM will address how well the sequence of relationships explains fully all of the information indicating relationships among the variables. The full information about all interrelationships is captured in the covariance matrix.

The relationships become even more intertwined in Figure 9.3c, with three dependent constructs, each related to the others as well as to the independent constructs. A reciprocal relationship (two-headed, straight arrow) even occurs between  $Y_2$  and  $Y_3$ . This relationship is shown in the equations by  $Y_2$  appearing as a predictor of  $Y_3$  and  $Y_3$  appearing as a predictor of  $Y_2$ . It is not possible to express all the relationships in either Figure 9.3b or 9.3c in a single equation. Separate equations are required for each dependent construct. The need for a method that can estimate all the equations simultaneously can be addressed with SEM.

These examples are just a preview of the types of relationships that can be portrayed and then empirically examined through SEM. Given the ability for the models to become complex quite easily, it is even more important to use theory as a guiding factor to specification of both the measurement and structural models. Later in this chapter, as well as in Chapters 10 and 11 we will discuss the criteria by which the researcher can specify SEM models in more detail.

## ESTABLISHING CAUSATION

Perhaps the strongest type of theoretical inference a researcher can draw is a causal inference, which involves proposing that a dependence relationship actually is based on **causation**. A **causal inference** involves a hypothesized cause-and-effect relationship. If we understand the causal sequence among variables, then we can explain how some



**Figure 9.3**  
Representing Dependence and Correlational Relationships Through Path Diagrams

cause determines a given effect. In practical terms, the effect can be at least partially managed with some degree of certainty. So, dependence relationships can sometimes be theoretically hypothesized as causal. However, simply thinking that a dependence relationship is causal doesn't make it so. As such, we use the term *cause* with great care in SEM.

Let us consider HBAT's interest in job satisfaction as an example. If feeling positive about a supervisor can be proven to result in (cause) increased job satisfaction, then we know that higher job satisfaction can be achieved by improving how employees feel about their supervisor. Thus, company policies and training can focus on improving supervision approaches. If supervision is causally related as hypothesized, then the resulting improvements will increase employee job satisfaction.

Causal research designs traditionally involve an experiment with some controlled manipulation (e.g., a categorical independent variable as found in MANOVA or ANOVA). SEM models are typically used, however, in non-experimental situations in which the exogenous constructs are not experimentally controlled variables. This limits the researcher's ability to draw causal inferences, and SEM alone cannot establish causality. SEM can, however, treat dependence relationships as causal if four types of evidence (covariation, sequence, nonspurious covariation, and theoretical support) are reflected in the SEM model [26, 45].

**Covariation** Because causality means that a change in a cause brings about a corresponding change in an effect, systematic covariance (correlation) between the cause and effect is necessary, but not sufficient, to establish causality. Just as is done in multiple regression by estimating the statistical significance of coefficients of independent variables that affect the dependent variable, SEM can determine systematic and statistically significant covariation between constructs. Thus, statistically significant estimated paths in the structural model (i.e., relationships between constructs) provide evidence that covariation is present. Structural relationships between constructs are typically the paths corresponding to hypothesized links or sequences of connections among constructs. In addition, structure is imposed by omitting paths indicating nonrelationships.

**Sequence** A second requirement for causation is the temporal sequence of events. We use our earlier example as an illustration.

If improvements in supervision result in increased job satisfaction, then the changes in supervision cannot occur after the change in job satisfaction. If we picture many dominos standing in a row, and the first one is knocked down by a small ball, it may cause all the other dominos to fall. In other words, the ball hitting the first domino causes the other dominos to fall. If the ball is the cause of this effect, the ball must hit the first domino before the others fall. If the others have fallen before the ball strikes the first domino, then the ball cannot have caused the dominoes to fall. Thus, sequence in causation means that improvements in supervision must occur before job satisfaction increases if the relationship between the two variables is causal.

SEM cannot provide this type of evidence without a research design that involves either an experiment or longitudinal data. An experiment can provide this evidence, because the researcher maintains control of the causal variable through manipulations. Thus, the research first manipulates a variable and then observes the effect. Longitudinal data can provide this evidence, because the data enable us to account for the period in which events occur. A great deal of social science research relies on cross-sectional surveys. Measuring all of the variables at the same point in time does not provide a way of accounting for the time sequence. Thus, theory must be used to argue that the sequence of effects is from one construct to another.

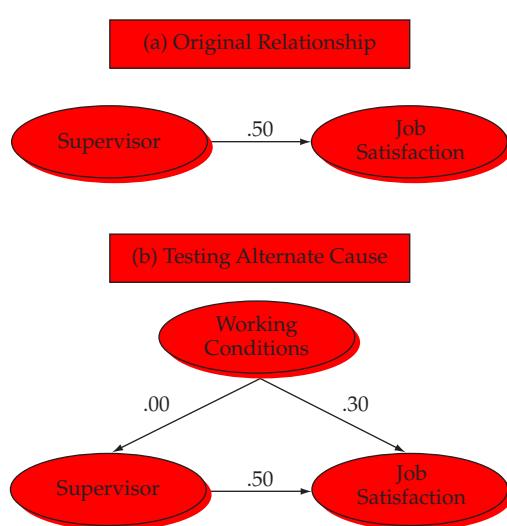
**Nonspurious Covariance** A **spurious relationship** is one that is false, misleading, or due to the lack of consideration of some other effect. A relationship is considered spurious when another event, not included in the original analysis, explains both the cause and effect. Simply put, the size and nature of the relationship between a true cause and the relevant effect should not be affected by including other constructs (or variables) in a model. Many anecdotes describe what can happen with spurious correlation. In addition, evidence suggesting correlated residual terms associated with the cause and effect, which can be assessed using SEM, also suggests a spurious relationship.

To illustrate, a significant correlation between ice cream consumption and the likelihood of drowning can be empirically verified. Is it safe, however, to say that eating ice cream causes drowning? If we account for some other potential cause (e.g., temperature is associated with increased ice cream consumption and more swimming), we would find no real relationship between ice cream consumption and drowning. Thus, we cannot say with any certainty that ice cream consumption causes the likelihood of drowning, even though they are significantly correlated. Thinking about the residuals from a regression predicting drowning with ice cream consumption alone, the residual (error variance) terms associated with drowning and ice cream consumption would be correlated. If indeed the residuals (error variance) are correlated and the effect is not modeled, the fit of the model will be diminished (we describe fit in more detail later). If the correlation between the residual terms is modeled, then causality is questioned. In fact, the notion of propensity scoring, introduced in an earlier chapter, is another way to determine the degree of causality in a relationship.

**THE IMPACT OF COLLINEARITY** Because a causal inference is supported when we can show that some other construct does not affect the relationship between the cause and effect, a lack of collinearity among the predictors (see Chapter 5 on multicollinearity) is desirable. When collinearity is not present, the researcher comes closest to reproducing the conditions that are present in an experimental design. These conditions include orthogonal, or uncorrelated, experimental predictor variables.

Unfortunately, most structural models involve multiple predictor constructs that exhibit **multicollinearity** with both other predictors and the construct. In these cases, making a causal inference is less certain. Therefore, in SEM models involving cross-sectional survey research, causal evidence is found when (1) the relationship between a cause and an effect remains constant when other predictor constructs are introduced into the model and (2) when the effect construct's error variance is independent (not related nor displaying high residual terms) [45, 51].

**TESTING FOR SPURIOUS RELATIONSHIPS** Figure 9.4 shows an example of testing for a nonspurious relationship with two SEM models. The first model specifies the proposed structural relationship between the two constructs. The second model incorporates the Alternative Cause construct as an additional predictor variable. If the estimated relationship between constructs found in the first model remains unchanged when the additional predictor is added (the second model), then the relationship is deemed nonspurious. However, if the structural relationship becomes nonsignificant in the second model because of the addition of the other predictors, then the relationship must be considered spurious. More than one additional construct may be added and the predicted causal structural relationships must remain consistent no matter how many constructs are added.



**Figure 9.4**  
Testing for a Nonspurious Relationship Between Constructs

In our employee job satisfaction example, we propose that supervisor perceptions influenced satisfaction (Figure 9.4a). One could argue, however, that how employees feel about their supervisor does not really determine their level of satisfaction with their job. An alternative explanation, for instance, is that good working conditions act as an alternative cause for both improved supervision and higher job satisfaction (Figure 9.4b). If the working conditions construct is measured along with the other constructs, and a relationship is specified between working conditions and both supervision and job satisfaction, then a SEM model can determine whether a relationship between the constructs exists or not. In our example, the estimated coefficient remains unchanged (.50), indicating that the relationship between supervisor and job satisfaction is nonspurious. If the estimated coefficient had become nonsignificant when working conditions is added to the model, then we would consider the relationship between supervisor and job satisfaction to be spurious.

**Theoretical Support** The final condition for causality is theoretical support, or a compelling rationale to support a cause-and-effect relationship. This condition emphasizes the fact that simply testing a SEM model and analyzing its results alone cannot establish causality. Theoretical support becomes especially important with cross-sectional data. A SEM model may suggest relationships between any correlated constructs (e.g., ice cream consumption and drowning statistics). But unless theory can be used to establish a causal ordering and a rationale for the observed covariance, the relationships remain simple association and should not be attributed with any further causal power.

Do employees' feelings about their supervisors cause job satisfaction? A theoretical justification for causation may exist in that as employees spend more time with their supervisors they become more familiar with their supervision approaches, which increases their understanding and reactions to the supervisor, and based on these experiences they become more satisfied with their job situation. Thus, a case can be made that more favorable feelings about supervisors cause increased job satisfaction.

Although SEM has often been referred to as causal modeling, causal inferences are only possible when evidence is consistent with the four conditions for causality already mentioned. SEM can provide evidence of systematic covariation and can help in demonstrating that a relationship is not spurious. If data are longitudinal, SEM can also help establish the sequence of relationships. However, it is up to the researcher to establish theoretical support. Thus, SEM is helpful in establishing a causal inference, but it cannot do it alone.

## DEVELOPING A MODELING STRATEGY

One of the most important concepts a researcher must learn regarding multivariate techniques is that no single correct way exists to apply them. In some instances, relationships are strictly specified, and the objective is a confirmation of the relationship. At other times, the relationships are loosely recognized, and the objective is the discovery of relationships. At each extreme, as well as the points in between, researchers must apply the multivariate technique in accordance with the research objectives.

The application of SEM follows this same tenet. Its flexibility provides researchers with a powerful analytical tool appropriate for many research objectives, which serve as guidelines in a modeling strategy. The use of the term *strategy* is designed to denote a plan of action toward a specific outcome. For our purposes, we define three distinct strategies in the application of SEM: confirmatory modeling strategy, competing models strategy, and model development strategy.

**Confirmatory Modeling Strategy** The most direct application of structural equation modeling is a **confirmatory modeling strategy**. The researcher specifies a specific theoretical model composed of a pattern of relationships and nonrelationships and then SEM assesses how well the model fits reality. The SEM approach is quite the opposite of exploratory approaches like stepwise regression or principal components analysis. If the proposed model has acceptable fit, the researcher has found support for that model. But, as we will discuss later, that model is just one of several different models that possibly have acceptable model fits. Perhaps a more insightful test can be achieved by comparing alternative theoretical models to find which model fits better than another.

**Competing Models Strategy** A **competing models strategy** is based on comparing one plausible theoretical estimated model with alternative theories by assessing relative fit. The strongest test of a proposed model is to identify and test competing models that represent truly different, but plausible, theories. When comparing these models, the researcher comes much closer to a test of competing theories, which is much stronger than a test of a single model in isolation. Philosophically, we may never know the absolute truth, but we can know that one theory is more truthful than another.

**Equivalent models** provide a second perspective on developing a set of comparative models. It has been shown that for a proposed structural equation model, at least one other model exists with the same number of parameters but with different relationships portrayed that fits at least as well as the proposed model.

**Model Development Strategy** The **model development strategy** differs from the prior two strategies in that, although a basic model framework is proposed, the purpose of the modeling effort is to improve this framework through modifications of the structural or measurement models. In many applications, theory can provide only a starting point for development of a theoretically justified model that can be empirically supported. Thus, the researcher must employ SEM not just to test the model empirically, but also to provide insights into its respecification.

One note of caution must be made. The researcher must be careful not to employ this strategy to the extent that the final model has acceptable fit but cannot be generalized to other samples or populations. Moreover, **model respecification** must always be done with theoretical support rather than just empirical justification. Models developed empirically should be verified with an independent sample as is the case with any exploratory or predictive approach.

## A Simple Example of SEM

The following example illustrates how SEM works. The example involves a sequence of relationships, multiple equations including equations because a dependent variable in one equation is an independent variable in another equation(s). This capability enables the researcher to model complex relationships in a way that is not possible with any of the other multivariate techniques discussed in this text. Another and perhaps more accurate way to think of the SEM process is to first imagine that every measured variable is related to every other measured variable. Then, constraints would be added that represented cases of independence, in other words, where relationships were pre-specified to be 0. The question then becomes whether the model can perform well even in the face of potentially many such constraints.

For simplicity, each construct in the following example is treated as a single variable. Thus, our example does not depict one of SEM's key strengths—the ability to employ multiple measures (the measurement model) to represent a construct through exploratory factor analysis. Chapter 10 discusses measurement theory and confirmatory factor analysis and will illustrate multiple item measurement in detail. For now, we focus only on the basic principles of model construction and estimating multiple relationships.

### THEORY

Theory must be the foundation of even the simplest of SEM models. Any sequence of multiple variables could be linked to one another in many ways. Perhaps some of the possible sequences of relationships would be complete nonsense. Theory should make the model plausible. The emphasis on representing dependence relationships necessitates that the researcher carefully details not only the number of constructs involved, but also the constraints necessary to represent only paths where relationships should exist. With these constructs in hand, model estimation can proceed.

To demonstrate how theory can be used to develop a model to test with SEM, let us use our example of employee job satisfaction, but expand it by adding a couple of more constructs. Two key research questions are: (1) what factors influence job satisfaction and (2) is job satisfaction related to employees' likelihood of looking for another job

(i.e., quitting their present job)? More specifically, HBAT management believes that favorable perceptions of supervision, coworkers, and working conditions will increase job satisfaction, which in turn will decrease the likelihood of searching for another job.

From their experiences, management developed a series of relationships they believe explain the process:

- Improved supervision leads to higher job satisfaction.
- Better work environment leads to higher job satisfaction.
- More favorable perceptions of coworkers lead to higher job satisfaction.
- Higher job satisfaction consequently leads to lower likelihood of job search.

These four relationships form the basis of how HBAT management believes they can reduce the likelihood of employees searching for another job. Additionally, the theory proposes that Supervision, Work Environment, and Coworkers do not relate directly to job search. Management would like to reduce job-searching activities because the cost of recruiting, hiring, and training new employees is very high.

The research team could use multiple regression, but that approach would only test part of this model, because regression is used to examine relationships between multiple independent variables and a single metric dependent variable. Given that the following theory involves more than a single dependent variable, the research team can use another technique that can examine relationships with more than a single dependent variable. In addition, the multiple equation approach recognizes that some effects on outcomes may be indirect and work through other constructs.

## SETTING UP THE STRUCTURAL EQUATION MODEL FOR PATH ANALYSIS

Once a theoretical sequence of effects and non-effects is specified, the researcher sets out to represent the model in a form suitable for analysis. First, constructs are identified as either being exogenous or endogenous. Then, the theoretical process can be portrayed visually in a path diagram, where straight arrows depict the impact of one construct on another. If causal effects are inferred, the arrows representing dependence relationships point from the cause to the subsequent effect. A construct with no arrow entering it is exogenous. A construct with an arrow(s) entering it is endogenous.

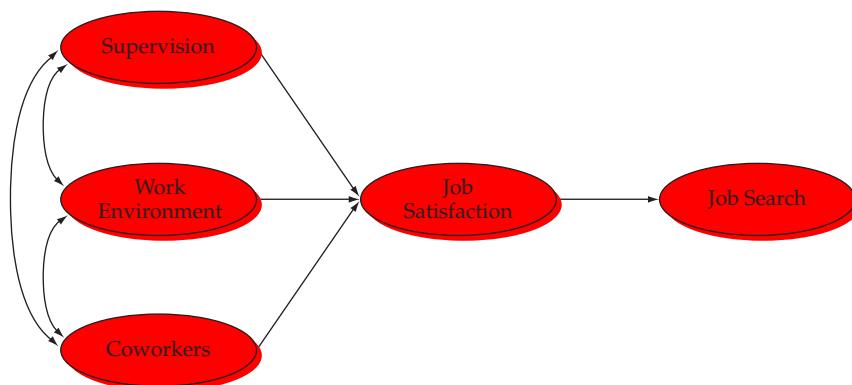
HBAT management proposes a theoretical model including five constructs: perceptions of supervision, work environment, and coworkers, along with job satisfaction and job search. An initial step is to identify which constructs are considered exogenous and which are endogenous. Remember that exogenous constructs are like independent variables, whereas endogenous are like dependent variables.

The supervision, work environment, and coworker constructs are identified as exogenous variables because they are not predicted by constructs within the model. Job search is clearly an endogenous variable, because it is represented as a dependent variable. But what about job satisfaction? It is dependent on the supervision, work environment, and coworker constructs, but it is also an independent variable because it is shown as influencing the job search construct. This is one of the unique and clearly beneficial characteristics of SEM—it can examine relationships (models) in which a construct operates as both an independent and dependent variable. From our model of the relationships, therefore, we can identify the types of constructs as shown below:

Exogenous Constructs	Endogenous Constructs
Supervision	Job Satisfaction
Work Environment	Job Search
Coworkers	

With the constructs specified as either exogenous or endogenous, the relationships can now be represented in a path diagram, as shown in Figure 9.5.

Note that one type of relationship also presented in Figure 9.5 was not expressed by the HBAT research team: the correlations among the exogenous constructs. Relationships among exogenous constructs are generally assumed unless there is a good theoretical reason to believe that the exogenous constructs are independent.



**Figure 9.5**  
Path Diagram of a Simple Structural Model

Typically, the assumption of correlated exogenous constructs involves the fact that additional variables not included in the model impact the exogenous variables (i.e., a common cause). In the case of exogenous variables, it is directly comparable to representing the multicollinearity discussed in multiple regression (see Chapter 4 for more detail). We have added these correlational relationships in our theoretical model because we expect the separate elements of managing HBAT employees (Supervision, Work Environment, and Coworkers) will be coordinated and based on consistent planning and execution. Moreover, including the interconstruct correlations between the exogenous variables often makes the estimates for the dependent relationships more reliable. However, relationships with endogenous factors are presumed not to exist unless and are constrained to zero unless some theoretical reasons dictates otherwise. We will discuss other reasons for adding this type of relationship in the following chapters. The research team can now collect data on the five constructs as a basis for evaluating the proposed theoretical model.

### THE BASICS OF SEM ESTIMATION AND ASSESSMENT

With the relationships and path diagram specified, researchers can now put them in a format suitable for analysis in SEM, estimate the strength of the relationships, and assess how well the data actually fits the model. In the example, we illustrate the basic procedures in each of these steps as we investigate the issues raised by employees' work environment and their job satisfaction and desire to engage in job search.

**Observed Covariance Matrix** SEM differs from multiple regression analysis in that it performs a covariance structure analysis rather than a variance decomposition analysis. As a result, SEM focuses on explaining covariation among all measured variables, which together form the **observed sample covariance matrix**. Although it may not always be obvious to the user, SEM programs can compute solutions using either a covariance matrix or a correlation matrix as input rather than using the individual data observations.

Correlation is just a special case of covariance. A correlation matrix is simply the covariance matrix when standardized variables are used (i.e., the standardized covariance matrix). The key at this point is to realize that we compute the *observed* covariance matrix from sample observations, just as we would compute a correlation matrix. It is not estimated, nor is it dependent on a model imposed by a researcher.

Let us revisit our example and see how the researchers would proceed after the model is defined.

To understand how data are input into SEM, think of the covariance matrix among the five variables. The observed covariance matrix would contain 25 values. The five diagonal values would represent the variance of each variable with 10 unique covariance terms. Because the covariance matrix is symmetric, the 10 unique terms would be repeated both above and below the diagonal. As a result, the number of unique values in the matrix is the five diagonal values (variances) plus the 10 unique off-diagonals (covariances), for a total of 15.

For example, suppose the sample involves individuals interviewed using a mall-intercept technique. The resulting covariance matrix is composed of the following values, with each construct simply abbreviated as S for Supervision, WE for Work Environment, CW for Coworkers, SAT for Job Satisfaction, and SRCH for Job Search (as in Figure 9.5). The matrix of unduplicated values would be as follows:

	S	WE	CW	SAT	SRCH										
Observed Covariance	Var (S)	Cov (S,WE)	Var (WE)	Cov (S,CW)	Cov (WE,CW)	Var (CW)	Cov (S,SAT)	Cov (WE,SAT )	Cov (CW,SAT )	Var (SAT)	Cov (S,SRCH)	Cov (WE,SRCH )	Cov (CW,SRCH)	Cov (SAT,SRCH)	Var (SRCH)

Actual values for this example are shown in Table 9.1a, the observed covariance matrix.

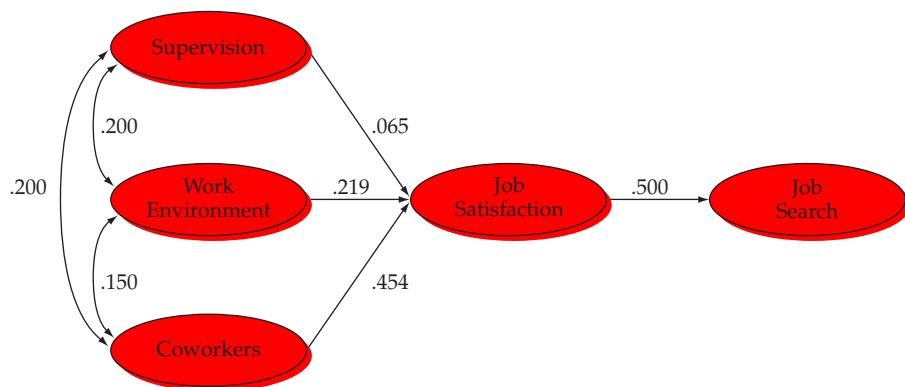
**Estimating and Interpreting Relationships** Prior to the widespread use of SEM programs, researchers found solutions for multiple equation models using a process known as **path analysis**. Path analysis uses bivariate correlations to estimate the relationships in a system of structural equations. This process estimates the strength of each structural relationship (a straight or curved arrow) in a path diagram. The actual mathematical procedure is briefly described in Appendix 9A.

Path analysis procedures provide estimates for each depicted relationship (arrow) in the model shown in Figure 9.6. These estimates are interpreted like regression coefficients if two separate equations are used—one to predict Job Satisfaction and a second to predict Job Search. But SEM does not keep each equation separate, and all estimates of relationships in both equations are computed at the same time using the information from all equations that make up the model. SEM also provides estimates of the correlational relationships between the exogenous constructs, which may be useful in our interpretation of the results, as well as directly influencing our assessment of the validity of the exogenous constructs.

With estimates for each path, an interpretation can be made of each relationship represented in the model. When statistical inference tests are applied, the researcher can assess the probability that the estimates are significant (i.e., not equal to zero). Moreover, the estimates can be used like regression coefficients to make estimates of the values of any construct in the model.

The relationships (paths) in the model shown in Figure 9.6 represent the research questions posed by the HBAT research team. When we look at the first three relationships (i.e., impact of Supervision, Work Environment and Coworkers on Job Satisfaction), we can see that the estimated coefficients are .065, .219, and .454, respectively. The sizes of these coefficients indicate that Coworkers has the biggest impact on job satisfaction, whereas Work

**Figure 9.6**  
An Estimated Structural Equation Model of Job Search



Environment is somewhat less and supervision has the smallest impact. Moreover, Job Satisfaction has a substantial impact on Job Search (.50) and provides evidence of that relationship as well.

Recall that regression coefficients can be used to compute predicted values for dependent variables. Those values were referred to as  $\hat{y}$ . Thus, for any particular values of the independent variables, an estimated value for the outcome can be obtained. In this case, where we treat constructs as variables, they would represent predicted values for endogenous constructs, or the outcome. The difference between the actual observed value for the outcome and  $\hat{y}$  is error. SEM also can provide estimated values for exogenous constructs when multiple variables are used to indicate the construct. This process will become clearer in the next chapters. Realize that several potential relationships between constructs have no path drawn, which means that the researcher does not expect a direct relationship between these constructs. For instance, no arrows are drawn between Supervision and Job Search, Work Environment and Job Search, or Coworkers and Job Search, which affects the equations for the predicted values.

In our model, if we take any observed values for Supervision, Work Environment and Coworkers, we can estimate a value for Job Satisfaction using the following equation:

$$\hat{y}_{JobSatisfaction} = .065(\text{Supervision}) + .219(\text{Work Environment}) + .454(\text{Coworkers})$$

Similarly, predicted values for Job Search can be obtained:

$$\hat{y}_{JobSearch} = .50(\text{Job Satisfaction})$$

This would represent a multiple equation prediction, because Job Satisfaction is also endogenous. Substituting the equation for Job Satisfaction into the equation for Job Search, we get:

$$\hat{y}_{JobSearch} = .50[.065(\text{Supervision}) + .219(\text{Work Environment}) + .454(\text{Coworkers})]$$

This illustrates, therefore, how path estimates in Figure 9.6 can be used to calculate estimated values for Job Satisfaction and Job Search. What is missing though is the big question of how well the proposed theoretical model represents reality. Here is where SEM enters the picture fully.

**Assessing Model Fit with the Estimated Covariance Matrix** The climatic step in SEM involves calculating an **estimated covariance matrix** that represents what the covariance matrix would be if the imposed structure, represented by the sequence of relationships and non-relationships, were true. The estimated covariance matrix is derived from the path estimates of the model. Then, SEM compares the estimated covariance matrix to the observed covariance matrix to test the fit of a theoretical model. Models that produce an estimated covariance matrix that is within sampling variation of the observed covariance matrix would be said to fit.

Let's look at one relationship (Work Environment and Job Satisfaction) to illustrate what happens. They involve both direct and indirect paths:

Direct path:

$$\text{Work Environment} \rightarrow \text{Job Satisfaction} = .219$$

Indirect paths:

$$\text{Work Environment} \rightarrow \text{Supervision} \rightarrow \text{Job Satisfaction} = .200 \times .065 = .013$$

$$\text{Work Environment} \rightarrow \text{Coworkers} \rightarrow \text{Job Satisfaction} = .150 \times .454 = .068$$

Total:

$$\text{Direct} + \text{Indirect} = .219 + .013 + .068 = .300$$

Thus, the estimated covariance between Work Environment and Job Satisfaction is .300, the sum of both the direct and indirect paths. The complete estimated covariance matrix is shown in Table 9.1b.

**Table 9.1** Observed, Estimated, and Residual Covariance Matrices

	Supervision	Work Environment	Coworkers	Job Satisfaction	Job Search
(A) Observed Covariance Matrix: ( $S$ )	Var (SP)	—	—	—	—
.20	Var (WE)	—	—	—	—
.20	.15	Var (CW)	—	—	—
.20	.30	.50	Var (JS)	—	—
-.05	.25	.40	.50	Var (JS)	—
(B) Estimated Covariance Matrix: ( $\Sigma$ )	—	—	—	—	—
.20	—	—	—	—	—
.20	.15	—	—	—	—
.20	.30	.50	—	—	—
.10	.15	.25	.50	—	—
(C) Residuals: Observed Minus Estimated Covariances	—	—	—	—	—
Supervision	—	—	—	—	—
Work Environment	.00	—	—	—	—
Coworkers	.00	.00	—	—	—
Job Satisfaction	.00	.00	.00	—	—
Job Search	-.15	.10	.15	.00	—

This example illustrates how the researcher's theory determines the estimated covariance matrix (and ultimately model fit) by the paths (and non-paths) specified in the model. In our example, if the Work Environment was not theoretically independent of the Supervision construct or Coworkers construct, and thus each possible path constrained to zero, then the estimated covariance would be different. Therefore, the researcher should note that each path added or constrained in the model ultimately controls how well the observed covariance matrix can be predicted. The identification of direct and indirect paths for each covariance is addressed in more detail in the Basic Stats appendix available online.

The last issue in assessing fit is the concept of a residual. The residuals in SEM models are the differences between each specific observed covariance and the corresponding estimated covariance. Thus, when we compare the observed and actual covariance matrices, any differences we detect are the residuals. The distinction with other multivariate techniques, especially multiple regression, is important. In those techniques, residuals reflected the errors in predicting individual observations ( $\hat{y} - y$ ). In SEM, predictions of individual observations are not the focus of the analysis. When a SEM program refers to residuals, it refers to the difference between the estimated and observed covariances for any pair of measured variables.

The matrix of residuals (the differences between the observed and estimated covariance matrices,  $|S - \Sigma_k|$ ) becomes the key driver in assessing the fit of a SEM model. The theoretical explanations represented by the proposed model are supported to the extent that the estimated covariance matrix ( $\Sigma_k$ ) is sufficiently close to the observed covariance matrix ( $S$ ). The smaller the residuals the closer the fit and the better the model matches reality. If the reader is familiar with cross-tabulation, it should be no surprise that a  $\chi^2$  statistic can be computed based on the difference between the two matrices. Later, we will use this statistic as the basic indicator of the goodness-of-fit of a theoretical model.

Comparing the observed and estimated covariance matrices shown in Table 9.1, some covariances are exactly predicted and some differences are found. For example, if you look in the first column of numbers in both matrices, you will see that the relationships between Supervision and Work Environment, and Coworkers and Job Satisfaction, are all predicted exactly. That is, they are all .20 in both the observed and estimated matrices. For other relationships, such as the relationship between Coworkers and Job Satisfaction, the estimated covariance (.25) is noticeably different from the observed covariance (.40).

## Structural Equation Modeling Introduction

No model should be developed for use with SEM without a plausible underlying theory, which is needed to develop:

Measurement model specification

Structural model specification.

Theoretical models can be represented by drawing a corresponding path diagram:

Dependence relationships are represented with single-headed directional arrows.

Correlational (covariance) relationships are represented with two-headed arrows.

Causal are the strongest type of inference made in applying multivariate statistics; therefore, they can be supported only when precise conditions for causality exist:

Covariance between the cause and effect

The cause must occur before the effect

Nonspurious association must exist between the cause and effect

Theoretical support exists for the relationship between the cause and effect.

Models developed with a model development strategy must be cross-validated with an independent sample.

The result is the residuals matrix (Table 9.1c). As we have noted, three residuals are not zero. Specifically, only the residuals for the relationships between the three exogenous constructs and Job Search are not zero (−.15, .10, and .15). These findings indicate that the SEM model does not perfectly explain the covariance between these constructs, and it could suggest that the researcher's theory is inadequate in explaining Job Search. But we need additional information before rejecting the proposed theory.

Notice that dependent variables in one relationship can easily be independent variables in another relationship (as with Job Satisfaction). No matter how large the path diagram gets or how many relationships are included, path analysis provides a way to analyze the set of relationships.

Fortunately, the researcher does not have to do all the calculations in path analysis manually because the software takes care of that. The researcher needs to understand the principles underlying SEM so that the implications of adding, constraining, or deleting potential paths or connections can be understood. The next two chapters explain how these procedures are implemented in testing measurement and structural theories, respectively.

## Six Stages in Structural Equation Modeling

Researchers are attracted to SEM because it provides a conceptually appealing way to test theory. If a researcher can express a theory in terms of relationships among measured variables and latent constructs (variates), then SEM will assess how well the theory *fits* reality as represented by data.

This section continues the discussion of SEM by describing a six-stage decision process (see Figure 9.7). This process varies slightly from that introduced in Chapter 1 in order to reflect the unique terminology and procedures of SEM. The six stages are as follows:

**Stage 1:** Defining individual constructs

**Stage 2:** Developing the overall measurement model

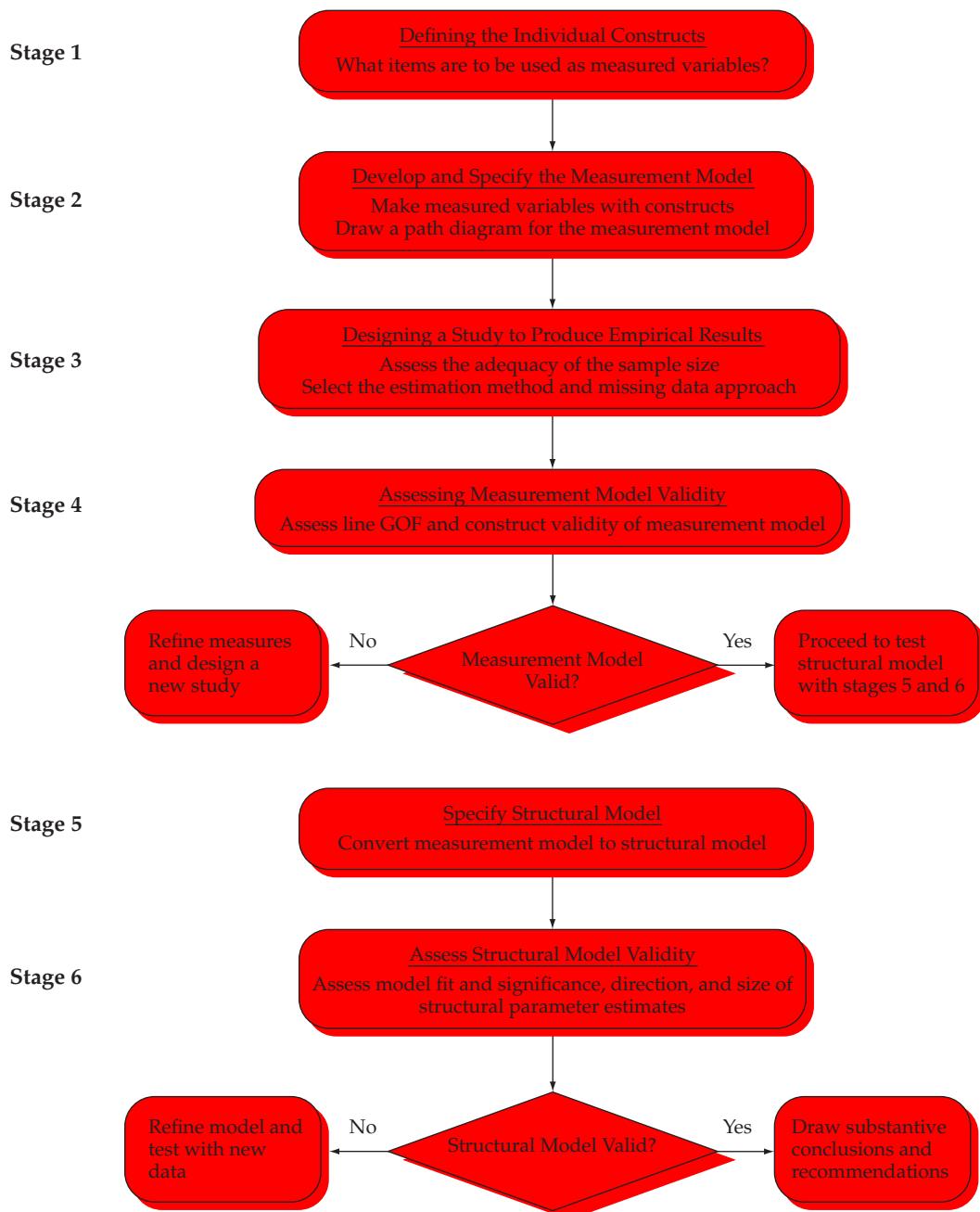
**Stage 3:** Designing a study to produce empirical results

**Stage 4:** Assessing the measurement model validity

**Stage 5:** Specifying the structural model

**Stage 6:** Assessing structural model validity

**Figure 9.7**  
Six-Stage Process for Structural Equation Modeling



The remainder of this chapter provides a brief overview and introduction of these six stages, which will also be discussed in greater detail over the next two chapters. Rather than include an HBAT example as an illustration of the technique here in this chapter, it will be introduced in the next chapter. The next two chapters are devoted to testing the measurement and structural models, respectively. Many SEM analyses involve testing both measurement theory (how the constructs are represented) and structural theory (how the constructs relate to each other). Chapter 10 covers the first four stages of SEM and Chapter 11 the remaining two stages.

## Stage 1: Defining Individual Constructs

A good measurement theory is a necessary condition to obtain useful results from SEM. Hypotheses tests involving the structural relationships among constructs will be no more reliable or valid than is the measurement model in explaining how these constructs are constructed. Researchers often have a number of established scales to choose from, each a slight variant from the others. But in other situations, the researcher is faced with the lack of an established scale and must develop a new scale or substantially modify an existing scale to the new context. In each case, how the researcher selects the items to measure each construct sets the foundation for the entire remainder of the SEM analysis. The researcher must invest significant time and effort early in the research process to make sure the measurement quality will enable valid conclusions to be drawn.

### OPERATIONALIZING THE CONSTRUCT

The process begins with a good theoretical definition of the constructs involved. This definition provides the basis for selecting or designing individual indicator items. A researcher **operationalizes a latent construct** by selecting its measurement scale items and scale type. In survey research, operationalizing a latent construct results in a series of scaled indicator items in a common format such as a Likert scale or a semantic differential scale. The definitions and items are derived from two common approaches.

**Scales from Prior Research** In many instances, constructs can be defined and operationalized as they were in previous research studies. Researchers may do a literature search on the individual constructs and identify scales that previously performed well. As we discussed in Chapter 3, compendiums of prior scales are available in numerous disciplines.

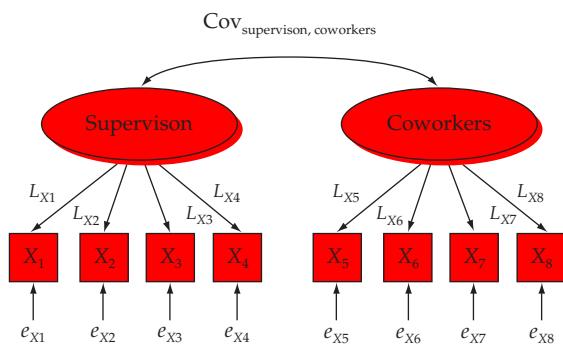
**New Scale Development** At times, research is needed to develop and validate a scale to measure a latent construct. This development is appropriate when a researcher is studying something that does not have a rich history of previous research or when existing scales are inappropriate for a given context. Also, a great deal of research exists for the sole purpose of developing psychometric scales capable of representing latent constructs. The term psychometrics refers to theory of and research directed toward quantitative and valid representation of latent psychological concepts. The general process for developing scale items is long and detailed. The essentials of this process are highlighted in the next chapter, but the reader is referred elsewhere for a more thorough discussion [13, 10, 20, 42].

### PRETESTING

Generally, when measures are either developed for a study or are taken from various sources, some type of pretest should be performed. The pretest should use respondents similar to those from the population to be studied so as to screen items for appropriateness. Pretesting is particularly important when scales are applied in specific contexts (e.g., purchase situations, industries, or other instances where specificity is paramount) or in contexts outside their normal use. Empirical testing of the pretest results is done in a manner identical to the final model analysis (see discussion on stage 4 later in this chapter). Items that do not behave statistically as expected may need to be refined or deleted to avoid these issues when the final model is analyzed.

## Stage 2: Developing and Specifying the Measurement Model

With the scale items specified, the research must now specify the measurement model. In this stage, each latent construct to be included in the model is defined and the measured indicator variables (items) are assigned to the corresponding latent constructs. Although this assignment is reflected by equations in reality, SEM software allows a simpler representation in the form of a diagram. Figure 9.8 represents a simple two-construct measurement model, with four indicators associated with each construct and a correlational relationship between constructs.



**Figure 9.8**  
Visual Representation of a Measurement Model

## SEM NOTATION

A key element in the path diagram is the labeling notation for indicators, constructs, and relationships between them. Each software program utilizes a somewhat unique approach, a standard convention developed with the widespread adoption of LISREL that is simply referred to as **LISREL notation**. LISREL notation is tied to the program's use of matrix notation, and thus it becomes unwieldy for those with no experience with LISREL. For purposes of this text, we will simplify our notation to be as generalizable as possible among all the SEM software programs, including AMOS, which is also widely used today. Given the widespread use of LISREL notation, however, we have developed a reference guide to LISREL notation (see Appendix 9B) along with a “conversion” between this notation and the LISREL notation for interested readers that are available online. Other software, such as M-Plus or lavaan, an R-based alternative, uses programming code somewhat like that presented here [32].

Table 9.2 lists the notation used in this text for the measurement and structural models. As discussed earlier, there are three types of relationships: measurement relationships between indicators/items and constructs; structural relationships between constructs; and correlational relationships between constructs. There are also two types of error terms, one related to individual indicators and the other to endogenous constructs.

Specification of the complete measurement model uses (1) measurement relationships for the items and constructs, (2) correlational relationships among the constructs, and (3) error terms for the items.

A basic measurement model can be illustrated as shown in Figure 9.8. The model has a total of 17 estimated parameters. The 17 free parameters include eight loading estimates, eight error estimates, and one between-construct correlation estimate (the two construct variances are fixed at 1). The estimate for each arrow linking a construct to a measured variable is an estimate of a variable's loading—the degree to which that item is related to the construct. This stage of SEM can be thought of as assigning individual variables to constructs. Visually, it answers the question, “Where should arrows be drawn or omitted linking constructs to variables?”

**Table 9.2** Notation for Measurement and Structural Models

Element	Symbol	Notation	Example
<i>Type of Indicator</i>			
Exogenous	X	X <sub>number</sub>	X <sub>1</sub>
Endogenous	Y	Y <sub>number</sub>	Y <sub>1</sub>
<i>Type of Relationship</i>			
Measurement (Loading)	L	L <sub>item</sub>	L <sub>X1</sub>
Structural (Path coefficient)	P	P <sub>outcome, predictor</sub>	P <sub>Job Sat, Sup</sub>
Correlational between Constructs	Cov	Cov <sub>construct1, construct2</sub>	Cov <sub>Sup, WE</sub>
<i>Error Terms</i>			
Indicators	e	e <sub>item</sub>	e <sub>X1</sub>
Constructs	E	E <sub>construct</sub>	E <sub>Job Search</sub>

Numerous possible paths are not specified. For example, no paths suggest correlations among indicator variables' error variances or loadings of indicators on more than one construct (cross-loadings). In the estimation process, these unspecified loadings are set (fixed or constrained) to the value of zero, meaning that they will not be estimated.

## CREATING THE MEASUREMENT MODEL

Specification of the measurement model can be a straightforward process, but a number of issues still must be addressed. Chapter 10 provides more detailed discussion of each issue. The types of questions are listed here:

- 1 Can we empirically support the validity and unidimensionality of the constructs? Essential points must be engaged in establishing the theoretical basis of the constructs and measures.
- 2 How many indicators should be used for each construct? What is the minimum number of indicators? Is there a maximum? What are the trade-offs for increasing or decreasing the number of indicators?
- 3 Should the measures be considered as portraying the constructs (meaning that they describe the construct) or seen as explaining the construct (such that we combine indicators into an index)? Each approach brings with it differing interpretations of what the construct represents.

The researcher, even with well-established scales, must still confirm the validity and unidimensionality in this specific context. In any scale development effort, issues as to the number of indicators and type of construct specification must be addressed. Researchers should always ensure that these issues are thoroughly examined, because any unresolved problems at this stage can affect the entire analysis, often in unseen ways.

## Stage 3: Designing a Study to Produce Empirical Results

With the basic model specified in terms of constructs and measured variables/indicators, the researcher must turn attention to issues involved with research design and estimation. Our discussion will focus on issues related to both research design and model estimation. In the area of research design, we will discuss (1) the type of data to be analyzed, either covariances or correlations; (2) the impact and remedies for missing data; and (3) the impact of sample size. In terms of model estimation, we will address model structure, the various estimation techniques available, and the current computer software being used.

### ISSUES IN RESEARCH DESIGN

As with any other multivariate technique, SEM requires careful consideration of factors affecting the research design necessary for a successful SEM analysis. SEM can be estimated with either covariances or correlations. Thus, the researcher must choose the appropriate type of data matrix for the research question being addressed. And even though the statistical issues of SEM estimation are discussed in the next section, here it is important to note that sample size and missing data can have a profound effect on the results no matter what method is used.

**Metric Versus Nonmetric Data** The observed or measured variables have traditionally been represented by metric data (interval or ratio). Metrically measured data is directly amenable to the calculation of covariances among items, as discussed earlier. Advances in the software programs, however, now allow for the use of many nonmetric data types (censored, binary, ordinal, or nominal). Differing types of variables can even be used as items for the same latent construct. The researcher must be careful to specify the type of data being used for each measured variable so that the appropriate measure of association can be calculated.

**Covariance Versus Correlation** Researchers conducting SEM analyses in the past debated over the use of a covariance versus correlation matrix as input. SEM was originally developed using covariance matrices (hence, it is referred to with the common name of *analysis of covariance structures*). Many researchers advocated the use of correlations as

a simpler form of analysis that was easier to interpret. The issue had practical significance, because for many years the input matrices were computed using a statistical routine outside the SEM program and then the matrix of correlations or covariances was used as input for the analysis. Today, most SEM programs can compute a model solution directly from raw data without the researcher computing a correlation or covariance matrix separately.

**INTERPRETATION** The key advantage of correlational input for SEM lies in the fact that the default parameter estimates are standardized, meaning they are not scale dependent. All estimated values must fall within the range  $-1.0$  to  $+1.0$ , making identification of inappropriate estimates easier than with covariances, which have no defined range. However, it is simple to produce these results from covariance input by requesting a standardized solution. As such, correlations hold no real advantage over the standardized results obtained using covariances.

**STATISTICAL IMPACT** The primary advantages of using covariances arise from statistical considerations. First, the use of correlations as input can at times lead to errors in standard error computations [12]. In addition, any time hypotheses concern questions related to the scale or magnitude of values (e.g., comparing means), then covariances must be used, because this information is not retained using correlations. Finally, any comparisons between samples require that covariances be used as input. Thus, covariances have distinct advantages in terms of their statistical properties versus correlations.

**CHOOSING BETWEEN COVARIANCES AND CORRELATIONS** In comparing the use of correlations versus covariances, we recommend using covariances whenever possible. Software makes the selection of one type versus another just a matter of selecting the type of data being computed. Covariance matrices provide the researcher with far more flexibility due to the relatively greater information content they contain.

**Missing Data** Just as with other multivariate procedures, the researcher must make several important decisions regarding missing data. Two questions must be answered concerning missing data to suitably address any problems it may create:

- 1 Is the missing data sufficient and nonrandom so as to cause problems in estimation or interpretation?
- 2 If missing data must be remedied, what is the best approach?

The reader is also referred back to Chapter 2, where a more complete discussion is provided on the methods of assessing the extent and pattern of missing data and the approaches to remedy missing data, if needed. For survey data in particular, the forced response options available in electronic questionnaires make missing data relatively uncommon. Nonetheless, if any missing data exists, it must be dealt with.

**EXTENT AND PATTERN OF MISSING DATA** Most notably, missing data must always be addressed if the missing data are in a nonrandom pattern or more than 10 percent of the data items are missing. Missing data are considered **missing completely at random (MCAR)** if the pattern of missing data for a variable does not depend on any other variable in the dataset or on the values of the variable itself [48]. If the pattern of missing data for a variable is related to other variables, but not related to its own values, then it is considered to be **missing at random (MAR)**. Chapter 2 provides much more detailed discussion on the procedures used in assessing the extent and pattern of missing data.

**MISSING DATA REMEDIES** Four basic methods are available for solving the missing data problem: the **complete case approach** (known as **listwise** deletion, where the observation is eliminated if missing data on any variable); the **all-available approach** (known as **pairwise** deletion, where all non-missing data are used); **imputation** techniques (e.g., mean substitution); and **model-based approaches**. Again, Chapter 2 provides a more detailed discussion of each of these options and their advantages and disadvantages. As with most missing data situations, listwise deletion and pairwise deletion have been the methods most widely employed, but both of these approaches do have substantial issues [1]. Model-based approaches extend past the simpler imputation approaches in that missing data are imputed (replaced) based on all available data for a given respondent. The most common approaches are the (1) maximum likelihood estimation of the missing values (ML), (2) the EM approach and (3) multiple imputation. Discussion of these imputation methods is not included in this chapter, but is available in a number of sources [11, 15] and in Chapter 2.

**SELECTING A MISSING DATA APPROACH** What is the best approach for handling missing data for SEM in general? We should first note that if missing data are random, less than 10 percent of observations, and the factor loadings are relatively high (.7 or greater), then any of the approaches are appropriate [14]. When missing data is more problematic than this, the first decision facing the researcher is whether to remedy the missing data problem before the estimation process or let the SEM software perform the missing data treatment.

Table 9.3 summarizes the strengths and weaknesses of each approach. When applying a remedy for missing data before estimation, the complete case approach (listwise deletion) becomes particularly problematic when samples and factor loadings are small. Conversely, the advantages of the model-based approaches become particularly apparent as sample sizes and factor loadings become generally smaller and/or the amount of missing data becomes larger. The all-available approach (pairwise deletion) is used most often when sample sizes exceed 250 and the total amount of missing data involved among the measured variables is below 10 percent. With this approach, the sample size ( $N$ ) should be set at the minimum (smallest) sample size available for any two covariances. The all-available approach has many good properties, but the user should be aware of the potential inflation of fit statistics when a modest or large amount of data are missing and factor loadings are large. All of these approaches allow the researcher to explicitly remedy the missing data before estimation and understand the implications of whichever approach is taken. Thus, if missing data are imputed, the imputed values are plugged in for the missing values and saved in a new dataset, from which, the covariance matrix used as input to SEM is computed [15].

Another option is FIML (full information maximum likelihood) imputation. Based on a regression-based maximum likelihood approach, FIML uses all the information in a dataset to substitute missing data points for each variable. Some variables may need much more substitution than others. The result is that a new dataset with no missing values becomes available, upon which an SEM model can be estimated as if no missing data were present to begin with. Many commercial statistical packages, including those devoted to SEM, facilitate FIML missing data

**Table 9.3** Some Advantages and Disadvantages of Different Missing Data Procedures

Method	Advantages	Disadvantages
Complete case (listwise)	$\chi^2$ shows little bias under most conditions. Effective sample size is known. Easy to implement using any program.	Increases the likelihood of non-convergence (SEM program cannot find a solution) unless factor loadings are high (> .6) and sample sizes are large (> 250). Increased likelihood of factor loading bias. Increased likelihood of bias in estimates of relationships among factors.
All-available (pairwise)	Fewer problems with convergence. Factor loading estimates relatively free of bias. Easy to implement using any program.	$\chi^2$ is biased upward when amount of missing data exceeds 10%, factor loadings are high, and sample size is high. Effective sample size is uncertain. Not as well known.
Model-based (ML/EM)	Fewer problems with convergence. $\chi^2$ shows little bias under most conditions. Least bias under conditions of random missing data.	Not available on older SEM programs. Effective sample size is uncertain for EM.
Full information maximum likelihood (FIML)	Remedy directly in estimation process. In most situations has less bias than other methods.	Researcher has no control over how missing data remedied. No knowledge how missing data impacts estimates. Typically only a subset of fit indices available.

Note: See Enders and Bandalos (2001) and Enders and Peugh (2004) for more detail. ML/EM have been combined based on the negligible differences between the results for the two (Enders and Peugh, 2004).

substitution [2,18]. Research has demonstrated that the model-based approach of FIML performs well compared to other methods presuming the data are missing at random [14], confirming the improvement found in the use of model-based approaches in most situations (see Chapter 2). Thus, the use of a model-based approach for missing data treatment is recommended when more than trivial amounts of missing data exist.

**Sample Size** SEM is often thought to require a larger sample relative to other multivariate approaches. Sample size, as in any other statistical method, provides a basis for the estimation of sampling error. Given the factor analytic basis for SEM, the reader can review the sample size discussions required for exploratory factor analysis (Chapter 3) [35]. Given that larger samples are usually more time consuming and expensive to obtain, the critical question in SEM involves how large a sample is needed to produce trustworthy results. Most importantly, the sample size required for any given statistic is a question secondary to the sample size required to generalize from a sample to a population. In almost all instances, the sample size requirement to infer to the population exceeds that for a specific statistic.

Five considerations affecting the required sample size for SEM include the following: (1) multivariate normality of the data, (2) estimation technique, (3) model complexity, (4) the amount of missing data, and (5) the average error variance among the reflective indicators.

**MULTIVARIATE NORMALITY** As data deviate more from the assumption of multivariate normality, then the ratio of observations to parameters needs to increase. A generally accepted ratio to minimize problems with deviations from normality is 10 respondents for each parameter estimated in the model. Although some estimation procedures are specifically designed to deal with non-normal data (asymptotic free), the researcher is always encouraged to provide sufficient sample size to allow for the sampling error's impact to be minimized, especially for non-normal data [54]. More importantly, a sample appropriate for generalizability is always needed, no matter the technique.

**ESTIMATION TECHNIQUE** The most common SEM estimation procedure is **maximum likelihood estimation (MLE)**. Simulation studies suggest that under ideal conditions, MLE provides valid and stable results for simple models with sample sizes as small as 50. As one moves away from conditions with very strong measurement and no missing data, minimum sample sizes to ensure stable MLE solutions increase when confronted with sampling error [34]. As an absolute minimum, SEM with maximum likelihood estimation is mathematically impossible with a sample size equal or less than the number of measured variables in a model.

**MODEL COMPLEXITY** Simpler models can be tested with smaller samples. In the simplest sense, more measured or indicator variables require larger samples. However, models can be complex in other ways that all require larger sample sizes:

- More constructs that require more parameters to be estimated.
- Constructs having less than three measured/indicator variables.
- Multigroup analyses require an adequate sample for each group.

The role of sample size is to produce more information and greater stability. Once a researcher has exceeded the absolute minimum size (one more observation than the number of observed covariances), larger samples mean less variability and increased stability in the solutions. Thus, model complexity leads to the need for larger samples.

**AVERAGE ERROR VARIANCE OF INDICATORS** Recent research indicates the concept of **communality** (see Chapter 3 for more details) is a more relevant way to approach the sample size issue. Communalities represent the average amount of variation among the measured/indicator variables explained by the measurement model. Standard measurement theories allow any single measured variable to load on only a single latent construct. The communality of an item can be directly calculated as the square of the standardized loading of a variable on its construct (see Chapter 10). Studies show that larger sample sizes are required as communalities become smaller (i.e., the unobserved constructs are not explaining as much variance in the measured items). Models containing multiple constructs with communalities less than .5 (i.e., standardized loading estimates less than .7) also require larger sizes for convergence and model stability [14]. The problem is much greater when models have constructs with only one or two items.

**SUMMARY ON SAMPLE SIZE** As SEM matures and additional research is undertaken on key research design issues, previous guidelines such as “always maximize your sample size” and “sample sizes of 300 are required” are not appropriate. Larger samples generally produce more stable solutions, particularly when data or measurement problems exist.

Based on the discussion of sample size, the following suggestions for minimum sample sizes are offered based on the model complexity and basic measurement model characteristics:

- Minimum sample size—100: Models containing five or fewer constructs, each with more than three items (observed variables), and with high item communalities (.6 or higher).
- Minimum sample size—150: Models with seven constructs or less, at least modest communalities (.5), and no underidentified constructs.
- Minimum sample size—300: Models with seven or fewer constructs, lower communalities (below .45), and/or multiple underidentified (fewer than three) constructs.
- Minimum sample size—500: Models with large numbers of constructs, some with lower communalities, and/or having fewer than three measured items.

In addition to these characteristics of the model being estimated, sample size should be increased in the following circumstances: (1) data deviates substantially from multivariate normality, (2) sample-intensive estimation techniques (e.g., ADF) are used, or (3) missing data exceeds 10 percent. Also, remember that group analysis requires that each group meet the sample size requirements just discussed. Finally, the researcher must remember that the sample size issue goes beyond being able to estimate a statistical model. The sample size, just as with any other statistical inference, must be adequate to represent the population of interest, which should be the overriding concern of the researcher. At the risk of being repetitive, inference to the population is the most important consideration in determining sample size.

## ISSUES IN MODEL ESTIMATION

In addition to the more general research design issues discussed in the prior section, SEM analysis has several unique issues as well. These issues relate to the model structure, estimation technique used, and computer program selected for the analysis.

**Model Structure** Among the most important step in setting up a SEM analysis is determining and communicating the theoretical model structure to the program. Path diagrams, like those used in prior examples, can be useful for this purpose. Knowing the theoretical model structure, the researcher can then specify the model parameters to be estimated. These models often include common SEM abbreviations denoting the type of relationship or variable referred to. As discussed earlier, LISREL notation is widely used as a notational form.

As we have mentioned many times, the researcher is responsible for specifying both the measurement and structural models. For every possible parameter (possible connection between variables), the researcher must decide if it is to be free or fixed. A **free parameter** is one to be estimated in the model and a value is produced by the SEM procedures. A **fixed parameter** is one in which the value is constrained to some value by the researcher. Most often a fixed parameter is constrained to a zero, implying independence between those variables (latent or observed). In a graphic model, any link that could be drawn but is omitted is a fixed parameter constrained to zero. At other times, the researcher may fix a parameter to a non-zero value. We will cover those situations later. In any event, the researcher specifies the complete SEM model in terms of free and fixed (constrained) linkages before estimating a model solution.

**Estimation Technique** Once the model is specified and the data collected, researchers choose the estimation method, the mathematical algorithm that will be used to identify estimates for each free parameter. Several options are available for obtaining a SEM solution.

Early attempts at structural equation model estimation were performed with ordinary least squares (OLS) regression. These efforts were quickly supplanted by MLE, which is consistent, more efficient and unbiased when the assumption

of multivariate normality as at least approximated. MLE is a flexible and robust approach to parameter estimation identifying the “most likely” parameter values to achieve the best model fit. The potential sensitivity of MLE to non-normality, however, created a need for alternative estimation techniques. Methods such as weighted least squares (WLS), generalized least squares (GLS), and asymptotically distribution free (ADF) estimation became available [21]. The ADF technique has received particular attention due to its insensitivity to non-normality of the data, but its requirement of rather large sample sizes limits its use.

All of the alternative estimation techniques have become more widely available as the computing power of the personal computer has increased, making them feasible for typical problems. MLE continues to be the most widely used approach and is the default in most SEM programs. In fact, it has proven fairly robust to violations of the normality assumption. Researchers compared MLE with other techniques, and it produced reliable results under many circumstances [43, 44, 49].

**Computer Programs** Several readily available statistical programs are convenient for performing SEM. Traditionally, the most widely used program for covariance-based SEM is LISREL (LInear Structural RELations) [9, 30]. LISREL is a flexible program that can be applied in numerous situations (i.e., cross-sectional, experimental, quasi-experimental, and longitudinal studies) and at one point became almost synonymous with structural equation modeling. AMOS (Analysis of Moment Structures) [3] is a program that has gained considerable popularity because, in addition to being a module in SPSS, it also is among the first SEM programs to rely heavily on a graphical interface for all functions so researchers never have to use any syntax commands or computer code. Mplus is a modeling program with multiple techniques that also has a graphical interface [41]. EQS (actually an abbreviation for *equations*) is another program that also can perform regression, factor analysis, and test structural models [6]. Finally, lavaan (latent variable analysis) is one of several SEM programs available through the R package.

## SEM Stages 1–3

When a model has scales borrowed or adapted from various sources reporting other research, a pretest using respondents similar to those from the population to be studied is recommended to screen items for appropriateness.

Pairwise deletion of missing cases (all-available approach) is a good alternative for handling missing data when the amount of missing data is less than 10 percent and the sample size is about 250 or more:

As sample sizes become small or when missing data exceed 10 percent, one of the imputation methods for missing data such as FIML becomes a good alternative for handling missing data.

When the amount of missing data becomes very high (15% or more), SEM may not be appropriate.

Covariance matrices provide the researcher with far more flexibility due to the relatively greater information content they contain and are the recommended form of input to SEM models.

The minimum sample size for a particular SEM model depends on several factors, including the model complexity and the communalities (average variance extracted among items) in each factor:

SEM models containing five or fewer constructs, each with more than three items (observed variables), and with high item communalities (.6 or higher), can be adequately estimated with samples as small as 50. But, remember more observations than the number of measured variables always are needed for the math to work.

When the number of factors is larger than six, some of which have fewer than three measured items as indicators, severe distributional problems exist, and multiple low communalities are present, sample size requirements become much greater.

The sample size must be sufficient to allow the model to run, but more important, it must adequately represent the population of interest.

Ultimately, the selection of a SEM program is based on researcher preference and availability. For most standard applications, the programs produce similar, although not always identical, substantive results. An appendix available online provides examples of the commands needed for several of these programs.

## Stage 4: Assessing Measurement Model Validity

With the measurement model specified, sufficient data collected, and the key decisions such as the estimation technique already made, the researcher comes to the most fundamental event in SEM testing: “Is the measurement model valid?” Measurement model validity depends on (1) establishing acceptable levels of goodness-of-fit for the measurement model (fit validity) and (2) finding other specific evidence of **construct validity**. Because we are focusing on the structural model in this simple example, we will defer the investigation of construct validity until discussed thoroughly in Chapter 10. The basics of fit though are the same whether testing a measurement or structural theory component.

**Goodness-of-fit (GOF)** indicates how well the user-specified model mathematically reproduces the observed covariance matrix among the indicator items (i.e., the similarity of the observed and estimated covariance matrices). Goodness of fit suggests how well the specified theoretical structure represents reality as represented by the data. The model must be able to account for all of the information about the data, meaning not only the variances but the covariances among measured variables as well. In the following sections, we first review some basic elements underlying all GOF measures, followed by discussions of numerous GOF heuristics that try to summarize the quality of fit in a single number rather than with a test of significance. Readers interested in more detailed and statistically based discussions are referred to Appendix 9C.

### THE BASICS OF GOODNESS-OF-FIT

Once a specified model is estimated, model fit compares the theory to reality by assessing the similarity of the estimated covariance matrix (theory) to reality (represented by the observed covariance matrix). If a researcher’s theory were perfect, the observed and estimated covariance matrices would be the same. The values of any GOF measure result from a mathematical comparison of these two matrices. The closer the values of these two matrices are to each other, the better the model is said to **fit**.

We start by examining **chi-square ( $\chi^2$ )**, the fundamental measure of statistical differences between the observed and estimated covariance matrices. Then the discussion focuses on calculating degrees of freedom, and finally, on how statistical inference is affected by sample size and the impetus that provides for alternative GOF measures.

**Chi-Square ( $\chi^2$ ) GOF** The difference in the observed and estimated covariance matrices (termed  $S$  and  $\Sigma_k$ , respectively) is the key value in assessing the GOF of any SEM model. The chi-square ( $\chi^2$ ) test is the only appropriate statistical test of the difference between matrices in SEM and is represented mathematically by the following equation:

$$\chi^2 = f[(N - 1)(\text{Observed sample covariance matrix} - \text{SEM estimated covariance matrix})]$$

or:

$$\chi^2 = f[(N - 1)(S - \Sigma_k)]$$

$N$  is the overall sample size. It should be noted that even if the differences in covariance matrices (i.e., residuals) remained constant, the  $\chi^2$  value increases as sample size increases. Likewise, the estimated covariance matrix is influenced by how many parameters are constrained (i.e., fixed) and the number of measured variables in the model, so the model degrees of freedom tend to increase the  $\chi^2$  GOF value. The  $\chi^2$  statistic provides the basis for most of the GOF heuristics discussed below.

**Degrees of Freedom (df)** As with other statistical procedures, **degrees of freedom** represent the amount of mathematical information available. Let's start by reviewing how the number of *df* are calculated. The net number of degrees of freedom for a SEM model is:

$$df = \frac{1}{2}[(p)(p + 1)] - k$$

where *p* is the total number of observed variables and *k* is the number of estimated (free) parameters. The fundamental difference in determining *df* in SEM comes in the first part of the calculation— $1/2[(p)(p + 1)]$ —which represents the number of covariance terms below the diagonal plus the variances on the diagonal. Thus, unlike other procedures, *df* are not derived at all from sample size as we saw in other multivariate techniques (e.g., in regression, *df* is the sample size minus number of estimated coefficients). Thus, degrees of freedom in SEM are based on the size of the covariance matrix, which is a square matrix with the number of rows (columns) equal to the number of indicators (measured variables) in the model. An important implication is that the researcher does not affect degrees of freedom through sample size, but we will see later how sample size does influence the use of chi-square as a GOF measure.

**Statistical Significance of  $\chi^2$**  The implied null hypothesis of SEM is that the observed sample and SEM estimated covariance matrices are equal, meaning that the model fits perfectly. The  $\chi^2$  value increases as differences (residuals) are found when comparing the two matrices. With the  $\chi^2$  test, we assess the statistical probability that the observed sample and SEM estimated covariance matrices are actually equal in a given population. This probability is the traditional *p*-value associated with parametric statistical tests. Note that there will nearly always be differences in the covariances of the two matrices, but statistically, good fit exists when the difference in the matrices is not statistically significant.

An important difference between SEM and other multivariate techniques also occurs in this statistical test for GOF. For other techniques we typically looked for a smaller *p*-value (less than .05) to show that a significant relationship existed. But with the  $\chi^2$  GOF test in SEM, we make inferences in a way that is in some ways exactly opposite. When we find a *p*-value for the  $\chi^2$  test to be small (statistically significant), it indicates that the two covariance matrices are statistically different and indicates lack of fit. So in SEM we look for a relatively small  $\chi^2$  value (and corresponding large *p*-value;  $> .05$ ), indicating no statistically significant difference between the two matrices, to support the idea that a proposed theory fits reality. Relatively small  $\chi^2$  values support the proposed theoretical model being tested.

We should note that the chi-square can also be used when comparing models, because the difference in chi-square between two models can be tested for statistical significance. Thus, if the researcher is expecting differences between models (e.g., differences in two models estimated for males and females), large  $\Delta\chi^2$  difference values would lend support that the models are different.

Chi-square ( $\chi^2$ ) is the fundamental statistical measure in SEM to quantify the differences between the covariance matrices. When used as a GOF measure, the comparison is between observed and predicted covariance matrices. Yet the actual assessment of GOF with a  $\chi^2$  value alone is complicated by several factors discussed in the next section. To provide alternative perspectives on model fit, researchers developed a number of alternative goodness-of-fit measures. The discussions that follow present the role of chi-square as well as the alternative measures.

## ABSOLUTE FIT INDICES

**Absolute fit indices** are a direct measure of how well the model specified by the researcher reproduces the observed data [31]. As such, they provide the most basic assessment of how well a researcher's theory fits the sample data. They do not explicitly compare the GOF of a specified model to any other model. Rather each model is evaluated independent of other possible models.

**$\chi^2$  Statistic** The most fundamental absolute fit index is the  $\chi^2$  statistic. It is the only statistically-based SEM fit measure [9] and is essentially the same as the  $\chi^2$  statistic used in cross-classification analysis between two nonmetric

measures. The one crucial distinction, however, is that when used as a GOF measure, the researcher interprets no differences between matrices (i.e., low  $\chi^2$  values) as support for the model as representative of the data.

The  $\chi^2$  GOF statistic has two mathematical properties that are problematic in its interpretation as a GOF measure. First, recall that the  $\chi^2$  statistic is a mathematical function of the sample size ( $N$ ) and the difference between the observed and estimated covariance matrices. As  $N$  increases, so does the  $\chi^2$  value, even if the differences between matrices do not change. Second, although perhaps not as obvious, the  $\chi^2$  statistic also is likely to be greater when the number of observed variables increases. Thus, all other things equal, just adding indicators to a model will cause the  $\chi^2$  values to increase and make it more difficult to achieve model fit.

Although the  $\chi^2$  test provides a test of statistical significance, these mathematical properties present trade-offs for the researcher. Although larger sample sizes are often desirable, just the increase in sample size itself will make it more difficult for those models to achieve a statistically insignificant GOF. Moreover, as more indicators are added to the model, because of either more constructs or better measurement of constructs, this will make it more difficult to obtain an insignificant  $\chi^2$ . One could argue that if more variables are needed to represent reality, then they should reflect a better fit, not a worse fit, as long as they produce valid measures. Thus, in some ways the mathematical properties of the  $\chi^2$  GOF test reduce the fit of a model for things that should not be detrimental to its overall validity.

For this reason, the  $\chi^2$  GOF test is often not used as the only GOF measure [52]. Researchers have developed many heuristic measures of fit to correct for the bias against large samples and increased model complexity. Several of these GOF indices are presented next. However, the  $\chi^2$  issues also impact many of these additional indices, particularly some of the absolute fit indices. This said, the  $\chi^2$  value for a model does summarize the fit of a model quite well and with experience the researcher can make educated judgments about models based on this result. In sum, the statistical test or resulting  $p$ -value is less meaningful particularly as sample sizes become large or the number of observed variables becomes large.

**Goodness-of-Fit Index (GFI)** The GFI was an early attempt to produce a fit statistic that was less sensitive to sample size. Even though  $N$  is not included in the formula, this statistic is still sensitive to sample size due to the effect of  $N$  on sampling distributions [36]. No statistical test is associated with the GFI, only guidelines to fit [53]. The possible range of GFI values is 0 to 1, with higher values indicating better fit. In the past, GFI values of greater than .90 typically were considered good. Others researchers argue that .95 should be used [24]. Development of other fit indices has led to a decline in usage of GFI.

**Root Mean Square Error of Approximation (RMSEA)** One of the most widely used measures that attempts to correct for the tendency of the  $\chi^2$  GOF test statistic to reject models with a large samples or a large number of observed variables is the root mean square error of approximation (RMSEA). This measure better represents how well a model fits a population, not just a sample used for estimation [25]. It explicitly tries to correct for both model complexity and sample size by including each in its computation. Lower RMSEA values indicate better fit.

The question of what is a “good” RMSEA value is debatable. Although previous research had sometimes pointed to a cut-off value of .05 or .08, more recent research points to the fact that drawing an absolute cut-off for RMSEA is inadvisable [17]. An empirical examination of several measures found that the RMSEA was best suited to use in a confirmatory or competing models strategy as samples become larger [47]. Large samples can be considered as consisting of more than 500 respondents. One key advantage to RMSEA is that a confidence interval can be constructed giving the range of RMSEA values for a given level of confidence. Thus, it enables us to report that the RMSEA is between 0.03 and 0.08, for example, with 95 percent confidence.

**Root Mean Square Residual (RMR) and Standardized Root Mean Residual (SRMR)** As discussed earlier, the error in prediction for each covariance term creates a residual. When covariances are used as input, the residual is stated in terms of covariances, which makes them difficult to interpret since they are impacted by the scale of the

indicators. But **standardized residuals** (SR) are directly comparable. The average SR value is 0, meaning that both positive and negative residuals can occur. Thus, a predicted covariance lower than the observed value results in a positive residual, whereas a predicted covariance larger than observed results in a negative residual. A common rule is to carefully scrutinize any standardized residual exceeding |4.0| (below -4.0 or above 4.0). Individual SRs enable a researcher to spot potential problems with a measurement model.

Standardized residuals are deviations of individual covariance terms and do not reflect overall model fit. What is needed is an “overall” residual value, and two measures have emerged in this regard. First is the root mean square residual (RMR), which is the square root of the mean of these squared residuals: an average of the residuals. Yet the RMR has the same problem as residuals in that they are related to the scale of the covariances. An alternative statistic is the standardized root mean residual (SRMR). This standardized value of RMR (i.e., the average standardized residual) is useful for comparing fit across models. Although no statistical threshold level can be established, the researcher can assess the practical significance of the magnitude of the SRMR in light of the research objectives and the observed or actual covariances or correlations [4]. Lower RMR and SRMR values represent better fit and higher values represent worse fits, which puts the RMR, SRMR, and RMSEA into a category of indices sometimes known as **badness-of-fit** measures in which high values are indicative of poor fit. A rule of thumb is that an SRMR over .1 suggests a problem with fit, although there are conditions that make the SRMR inappropriate that are discussed in a later section.

**Normed Chi-Square** This GOF measure is a simple ratio of  $\chi^2$  to the degrees of freedom for a model. Generally,  $\chi^2: df$  ratios on the order of 3:1 or less are associated with better-fitting models, except in circumstances with larger samples (greater than 750) or other extenuating circumstances, such as a high degree of model complexity. The normed chi-square is not a substitute for reporting the actual chi-square value and number of *df*.

**Other Absolute Indices** Most SEM programs provide the user with many different fit indices. In the preceding discussion, we focused more closely on those that are most widely used. But this is by no means an exhaustive list. For more information, the reader can refer to an extended discussion of these measures online, as well as the documentation associated with the specific SEM program used.

## INCREMENTAL FIT INDICES

**Incremental fit indices** differ from absolute fit indices in that they assess how well the estimated model fits relative to some alternative baseline model. The most common baseline model is referred to as a **null model**, one that assumes all observed variables are uncorrelated. It implies that no model specification could possibly improve the model, because it contains no multi-item factors (see Chapter 3) or relationships between them. This class of fit indices represents the improvement in fit by the specification of related multi-item constructs.

Most SEM programs provide multiple incremental fit indices as standard output. Different programs provide different fit statistics, however, so you may not find all of these in a particular SEM output. Also, they are sometimes referred to as comparative fit indices for obvious reasons. Listed below are some of the most widely used incremental fit measures, but the TLI and CFI are the most widely reported.

**Normed Fit Index (NFI)** The NFI is one of the original incremental fit indices. It is a ratio of the difference in the  $\chi^2$  value for the fitted model and a null model divided by the  $\chi^2$  value for the null model. It ranges between 0 and 1, and a model with perfect fit would produce an NFI of 1. One disadvantage is models that are more complex will necessarily have higher index values and artificially inflate the estimate of model fit. As a result, it is used less today in relation to either of the following incremental fit measures.

**Tucker Lewis Index (TLI)** The TLI conceptually similar to the NFI, but varies in that it is actually a comparison of the normed chi-square values for the null and specified model, which to some degree takes into account model complexity. However, the TLI is not normed, and thus its values can fall below 0 or above 1. Typically though, models

with good fit have values that approach 1, and a model with a higher value suggests a better fit than a model with a lower value.

**Comparative Fit Index (CFI)** The CFI is an incremental fit index that is an improved version of the normed fit index (NFI) [5, 7, 25]. The CFI is normed so that values range between 0 and 1, with higher values indicating better fit. Because the CFI has many desirable properties, including its relative, but not complete, insensitivity to model complexity. In fact, the CFI has become the most widely reported index to supplement the  $\chi^2$  and  $df$ .

**Relative Non-centrality Index (RNI)** The RNI also compares the observed fit resulting from testing a specified model to that of a null model. Like the other incremental fit indices, higher values represent better fit, and the possible values generally range between 0 and 1.

## PARSIMONY FIT INDICES

The third group of indices is designed specifically to provide information about which model among a set of competing models is best, considering its fit relative to its complexity. A **parsimony fit** measure is improved either by a better fit or by a simpler model. In this case, a simpler model is one with fewer estimated parameters paths. The parsimony ratio is the basis for these measures and is calculated as the ratio of degrees of freedom used by a model to the total degrees of freedom available [37].

Parsimony fit indices are conceptually similar to the notion of an adjusted  $R^2$  (discussed in Chapter 4) in the sense that they relate model fit to model complexity. More complex models are expected to fit the data better, so fit measures must be relative to model complexity before comparisons between models can be made. The indices are not useful in assessing the fit of a single model, but are quite useful in comparing the fit of two models, one more complex than the other.

The use of parsimony fit indices remains somewhat controversial. Some researchers argue that a comparison of competing models' incremental fit indices provides similar evidence and that we can take parsimony into account further in some other way. It is clear to say that a parsimony index can provide useful information in evaluating competing models, but that it should not be relied upon as the only fit measure. In theory, parsimony indices are a good idea. In practice, they tend to favor more parsimonious models to a large extent. When used, the PNFI is the most widely applied parsimony fit index.

**Adjusted Goodness of Fit Index (AGFI)** An adjusted goodness-of-fit index (AGFI) tries to take into account differing degrees of model complexity. It does so by adjusting GFI by a ratio of the degrees of freedom used in a model to the total degrees of freedom available. The AGFI penalizes more complex models and favors those with a minimum number of free paths. AGFI values are typically lower than GFI values in proportion to model complexity. No statistical test is associated with AGFI, only guidelines to fit [53]. As with the GFI, however, the AGFI is less frequently used in favor of the other indices which are not as affected by sample size and model complexity.

**Parsimony Normed Fit Index (PNFI)** The PNFI adjusts the normed fit index (NFI) by multiplying it times the parsimony ratio [40]. Relatively high values represent relatively better fit, so it can be used in the same way as the NFI. The PNFI takes on some of the added characteristics of incremental fit indices relative to absolute fit indices in addition to favoring less complex models. Once again, the values of the PNFI are meant to be used in comparing one model to another, with the highest PNFI value being most supported with respect to the criteria captured by this index.

## PROBLEMS ASSOCIATED WITH USING FIT INDICES

Ultimately, fit indices are used to establish the acceptability of any SEM model. Probably no SEM topic is more debated than what constitutes an adequate or good fit. The expanding collection of fit indices and the lack of consistent guidelines can tempt the researcher to "pick and choose" an index that provides the best fit evidence in

one specific analysis and a different index in another analysis. The researcher is faced with two basic questions in selecting a measure of model fit:

- 1 What are the best fit indices to objectively reflect a model's fit?
- 2 What are objective cut-off values suggesting good model fit for a given fit index?

Unfortunately, the answer to both questions is neither simple nor straightforward. Some researchers equate the search for answers to these questions with the “mythical Golden Fleece, the search for the fountain of youth, and the quest for absolute truth and beauty” [38]. Indeed many problems are associated with the pursuit of good fit. Following is a brief summary of the major issues found in various fit indices.

**Problems with the  $\chi^2$  Test** Perhaps the most clear and convincing evidence that a model's fit is adequate would be a  $\chi^2$  value with a  $p$ -value indicating no significant difference between the observed and estimated covariance matrices. For example, if a researcher used an error rate of five percent, then a  $p$ -value greater than .05 would suggest that the researcher's model is capable of reproducing the observed variables' covariance matrix—a “good” model fit.

But as we have discussed earlier, so many factors impact the  $\chi^2$  significance test that practically any result can be questioned. Does a nonsignificant  $\chi^2$  value always enable a researcher to say: “Case closed, we have good fit”? Not quite! Very simple models with small samples have a bias toward a nonsignificant  $\chi^2$  even though they do not meet other standards of validity or appropriateness. Likewise, there are inherent penalties in the  $\chi^2$  for larger sample sizes and larger numbers of indicator variables [5]. The result is that many models today are more complex and have sample sizes that make the  $\chi^2$  significance test less useful as a clear GOF measure that always separates good from poor models. Thus, no matter what the  $\chi^2$  test result, the researcher should complement it with other GOF indices. But just as important, the  $\chi^2$  value itself and the model degrees of freedom should always be reported [22, 49].

**Cut-off Values for Fit Indices: The Magic .90, or Is that .95?** Although we know we need to complement the  $\chi^2$  with additional fit indices, one question still remains no matter what index is chosen: What is the appropriate cut-off value for that index? For most of the incremental fit statistics, accepting models producing values of .90 became standard practice in the early 1990s. However, some scholars concluded that .90 was too low and could lead to false conclusions, and by the end of the decade .95 had become the recommended standard for indices such as the TLI and CFI [25]. In general, .95 somehow became the magic number indicating good-fitting models, even though no empirical evidence supported such a development.

Research has challenged the use of a single cut-off value for GOF indices, finding instead that a series of additional factors can affect the index values associated with acceptable fit. First, research using simulated data (for which the actual fit is known) provides counterarguments to these cut-off values and does not support .90 as a generally acceptable rule of thumb [25]. It demonstrates that at times even an incremental goodness-of-fit index above .90 would still be associated with a severely misspecified model. This suggests that cut-off values should be set higher than .90. Second, research continues to support the notion that model complexity unduly affects GOF indices, even with something as simple as just more indicators per construct [31]. Finally, the underlying distribution of data can influence fit indices [16]. In particular, as data become less appropriate for the particular estimation technique selected, the ability of fit indices to accurately reflect misspecification can vary. This issue seems to affect incremental fit indices more than absolute fit indices.

What has become clear is that no single “magic” value always distinguishes good models from bad models. GOF must be interpreted in light of the characteristics of the research. It is interesting to compare these issues in SEM to the general lack of concern for establishing a magic  $R^2$  number in multiple regression. If a magic minimum  $R^2$  value of .5 had ever been imposed, it would be just an arbitrary limit that would exclude potentially meaningful research. So we should be cautious in adopting one size fits all standards. It is simply not practical to apply a single set of cut-off rules that apply for all SEM models of any type.

## UNACCEPTABLE MODEL SPECIFICATION TO ACHIEVE FIT

Researchers sometimes test theory and sometimes pursue a good fit. The desire to achieve good fit should never compromise the theory being tested. Yet, in practice the pursuit of increasing model fit can lead to several poor practices in model specification [31, 33, 39]. In each of the following instances, a researcher may be able to increase fit, but only in a manner that compromises the theory test. Although each of these actions may be required in very specific instances, they should be avoided whenever possible, because each has the potential to unduly limit the ability of SEM to provide a true test of a model. Further, researchers learn not only from theory that is confirmed, but also from the areas where theoretical expectations are not confirmed [22].

One area of poor practices involves the number of items per construct. A common mistake is to reduce the number of items per construct to only two or three. Although doing so may improve model fit by reducing the total number of indicators, it very likely diminishes the construct's theoretical domain and ultimately its validity. The concept of multiple measures was to include as wide a range as possible of items that could measure the construct, not limit it to a very small subset of these items. An even more extreme action is to use a single item to represent a construct, necessitating an arbitrary specification of measurement error. Here the researcher circumvents the objective of the measurement model by providing the values for the indicator. Single items should only be used when the construct truly is and can be measured by a single item (e.g., a binary variable, such as purchase/no purchase, succeed/fail, or yes/no). Finally, a test of a measurement model should be performed with the full set of items. The parceling of items, where the full set of indicator variables (e.g., 15 indicators for a construct) is parceled into a small number of composite indicators (e.g., three composites of five items each), can reduce model complexity but may obscure the qualities of individual items. Thus, if parceling of items is performed, it should be employed after the entire set has been evaluated.

Another poor practice is to assess measurement model fit through a separate analysis for each construct instead of one analysis for the entire model with all constructs. This is an inappropriate use of the GOF indices, which need to examine the entire model at the same time, not a single construct at a time. The result is not only an incomplete test of the overall model, but a bias toward confirmation, because it is easier for single constructs to each meet the fit indices than it is for the entire set to achieve acceptable fit. Moreover, tests of discriminant validity and potential item cross-loadings (see Chapter 10) are impossible unless all of the constructs are tested collectively.

Finally, many model fit indices look more favorable with relatively small samples. While one might be tempted to use a small sample to improve the appearance of fit, such justification of a small sample runs counter to the need for use of as large a sample as possible or feasible to ensure representativeness and generalizability. Moreover, very small samples increase the chances for encountering statistical problems with model convergence and provide lower statistical power.

Occasionally, researchers may find merit in examining a CFA with only a single factor or be forced to rely on a small sample. However, improvement in fit is not an appropriate justification for any of these steps. Always remember that these procedures can interfere with the overall test of a measurement (or structural) model, and thus the theory remains untested until all measured variables and constructs are included in a single test.

## GUIDELINES FOR ESTABLISHING ACCEPTABLE AND UNACCEPTABLE FIT

A single cut-off for fit index values that distinguish good models from poor models across all situations cannot be offered. Consequently, we offer these *guides for usage, not rules that guarantee a correct model*. Thus, no specific value on any index can separate models into acceptable and unacceptable fits. The guidelines allow the analyst flexibility in application of fit criteria and even though for simplicity, one may want a yes or no answer, the best analyst still applies reason in evaluating the merit of a model.

**Use Multiple Indices of Differing Types** Typically, using three to four fit indices provides adequate evidence of model fit. Current research suggests a fairly common set of indices perform adequately across a wide range of situations, and the researcher need not report all GOF indices, because they are often redundant. However, the researcher should report at least one incremental index and one absolute index, in addition to the  $\chi^2$  value and the associated degrees of freedom, because using a single GOF index, even with a relatively high cut-off value, does not adequately

supplement the  $\chi^2$  GOF test alone [38]. Thus, reporting the  $\chi^2$  value and degrees of freedom, the CFI, and the RMSEA will often provide sufficient unique information to evaluate a model. The SRMR can replace RMSEA to also represent badness of fit (higher values mean relatively worse fit), whereas the others represent goodness of fit (higher values represent relatively better fit).

**Adjust the Index Cut-off Values Based on Model Characteristics** Table 9.4 provides guidelines for assessing fit indices in different situations. The guidelines are based primarily on simulation research that considers different sample sizes, model complexity, and degrees of error in model specification to examine how accurately various fit indices perform [25, 38]. One key point across the results is that *simpler models* and *smaller samples* should be subject to *more strict evaluation* than are more *complex models* with *larger samples*. Likewise, more *complex models* with *smaller samples* may require *somewhat less strict* criteria for evaluation with the multiple fit indices [50].

For example, based on a sample of 100 respondents and a four-construct model with only 12 total indicator variables, evidence of relatively good fit would include an insignificant  $\chi^2$  value, a CFI of at least .99, and a RMSEA of .08 or lower. It is extremely unrealistic, however, to apply the same guideline to an eight-construct model with 50 indicator variables tested with a sample of 2,000 respondents.

Remember that Table 9.4 is provided more to give the researcher an idea of how fit indices can be used practically than to suggest absolute rules separating good and bad fit. Moreover, it is worth repeating that even a model with a good fit must still meet the other criteria for validity discussed in subsequent chapters.

**Compare Models Whenever Possible** Although it is difficult to determine absolutely when a model is good or bad, it is much easier to determine that one model is better than another. The indices in Table 9.4 perform well in distinguishing the relative superiority of one model compared to another. A CFI of .96, for instance, indicates truly better fit than a similar model with a CFI of .88. A more in-depth discussion of competing models is described in Stage 6.

**The Pursuit of Better Fit at the Expense of Testing a True Model is Not a Good Trade-Off** Many model specifications can influence model fit. The researcher should be sure, therefore, that all model specifications are done to best approximate the theory to be tested rather than hopefully increase model fit. Remember, the idea is to test the theory accurately not to twist the theory to match the results.

**Table 9.4** Characteristics of Different Fit Indices Demonstrating Goodness-of-Fit Across Different Model Situations

		<i>N &lt; 250</i>			<i>N &gt; 250</i>		
No. of Stat.	vars. ( <i>m</i> )	<i>m</i> ≤ 12	12 < <i>m</i> < 30	<i>m</i> ≥ 30	<i>m</i> < 12	12 < <i>m</i> < 30	<i>m</i> ≥ 30
$\chi^2$	Insignificant	Significant	Significant	Insignificant	Significant	Significant	
	<i>p</i> -values expected	<i>p</i> -values even with good fit	<i>p</i> -values expected	<i>p</i> -values even with good fit	<i>p</i> -values expected	<i>p</i> -values expected	
CFI or TLI	.99 or better	.97 or better	Above .93	.96 or better	Above .94	Above .92	
RNI	May not diagnose misspecification well	.97 or better	Above .93	.96 or better, not used with <i>N</i> > 1,000	Above .94, not used with <i>N</i> > 1,000	Above .92, not used with <i>N</i> > 1,000	
SRMR	Biased upward, use other indices	.08 or less (with CFI of .95 or higher)	Less than .09 (with CFI above .93)	Biased upward; use other indices	.08 or less (with CFI above .94)	.08 or less (with CFI above .92)	
RMSEA	Values < .08 with CFI of = .99 or higher	Values < .08 with CFI of .97 or higher	Values < .08 with CFI above .93	Values < .07 with CFI of .96 or higher	Values < .07 with CFI of .94 or higher	Values < .07 with CFI of .92 or higher	

Note: *m* = number of observed variables; *N* applies to number of observations per group when applying CFA to multiple groups at the same time.

## Stage 5: Specifying the Structural Model

Specifying the measurement model (i.e., assigning indicator variables to the constructs they should represent) is a critical step in developing a SEM model. This activity is accomplished in Stage 2. Stage 5 involves specifying the structural model by assigning relationships from one construct to another based on the proposed theoretical model. Structural model specification focuses on using single-headed, directional arrows as shown in Figure 9.1c to show dependence relationships that represent structural hypotheses of the researcher's model. In other words, the researcher identifies the dependence relationships that are hypothesized to exist among the constructs, and each hypothesis represents a specific relationship that must be specified. The relationships show how one construct influences another and may be direct or indirect. Thus, the model displays all the dependence relationships that exist among the constructs. Sometimes, for convenience, researchers specify each specific link as a hypothesis, although the goal of SEM is to test theory and not individual relationships.

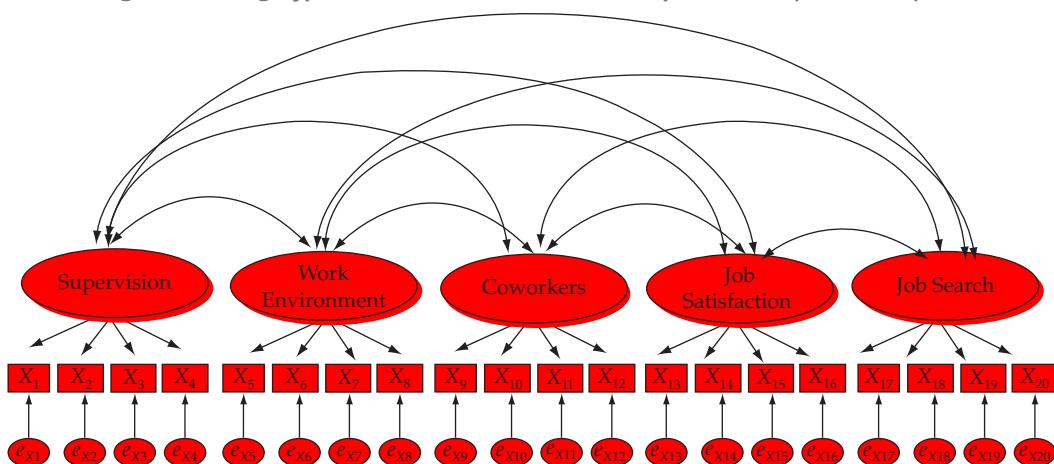
Now we return to the Job Search model discussed earlier in the chapter. We can specify the full measurement model, as shown in Figure 9.9, where there were no structural relationships among the constructs. All constructs were considered exogenous and correlated. This is also known as a confirmatory factor analysis (CFA) model.

In specifying a structural model, the researcher now carefully selects what are believed to be the key factors that influence Job Search. From their experience and judgment, the HBAT management team believes there is a strong reason to suspect that perceptions of supervision, work environment, and coworkers affect job satisfaction, which in turn affects job search. Based on theory the team proposes that environmental characteristics influence job satisfaction, and through job satisfaction, employees search behavior is determined. The theory can be expressed with the following direct structural relationships:

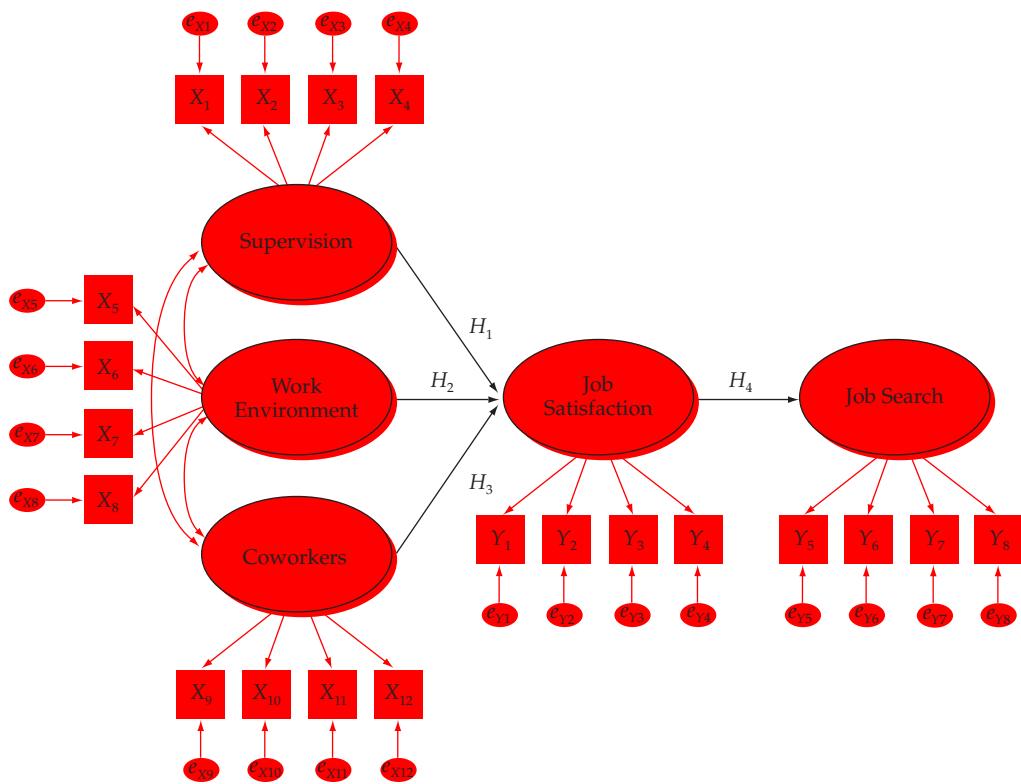
$H_1$ :	Supervision perceptions are positively related to Job Satisfaction.
$H_2$ :	Work Environment perceptions are positively related to customer share.
$H_3$ :	Coworkers perceptions are positively related to customer share.
$H_4$ :	Job Satisfaction is negatively related to Job Search.

These structural relationships are shown in Figure 9.10.  $H_1$  is specified with the arrow connecting supervision and job satisfaction. In a similar manner  $H_2$ ,  $H_3$ , and  $H_4$  are specified. The single-headed arrows showing the dependence relationship between constructs represents the structural part of the model. The constructs display the specified measurement structure (links to indicator variables) that would have already been tested in the confirmatory factor analysis stage. Any relationships among exogenous constructs are accounted for with correlational relationships

**Figure 9.9**  
A Path Diagram Showing Hypothesized Measurement Model Specification (CFA Model)



**Figure 9.10**  
A Path Diagram Showing Specified Hypothesized Structural Relationships and Measurement Specification



(curved, two-headed arrows). Thus, the three relationships among the two exogenous constructs are specified just as they were in the measurement model.

Another way to view the structural model is that “constraints” can be added to the measurement model. That is, specific structural paths replace the correlations between the constructs for each hypothesized relationship. In addition, every place where no relationship is hypothesized would represent a connection constrained to zero. With the exception of correlational relationships among exogenous constructs, no path is drawn between two constructs unless a direct, dependence relationship is hypothesized. Thus, all relationships not shown in the structural model are “constrained” to be equal to zero.

Although the focus in this stage is on the structural model, estimation of the SEM model requires that the measurement specifications be included as well. In this way, the path diagram represents both the measurement and structural part of SEM in one overall model. Thus, the path diagram in Figure 9.10 shows not only the complete set of constructs and indicators in the measurement model, but also imposes the structural relationships among constructs. The model is now ready for estimation. This becomes the test of the overall theory, including both the measurement relationships of indicators to constructs, as well as the hypothesized structural relationships among constructs.

## Stage 6: Assessing the Structural Model Validity

The final stage involves efforts to test the validity of the proposed theoretical structural model and examine the theoretical relationships embedded within that theory (e.g.,  $H_1$ – $H_4$  in our simple example). Realize that if the measurement model has not survived its tests of fit and other aspects of validity in Stage 4, Stages 5 and 6 are not advised because the results suggest the proposed measurement theory is flawed. If one does not achieve acceptable fit for the measurement model, model fit will not improve when constraints are added to represent the theoretical structural

model. Only when the measurement model is first validated and achieves acceptable model fit can we turn our attention to a test of the structural relationships.

Two key differences arise in testing the fit of a structural model relative to a measurement model. First, even though acceptable overall model fit must again be established, alternative or competing theoretical models are encouraged to support a model's superiority. Second, the parameter estimates for the structural relationships become a focus if fit is sufficient because they provide direct empirical evidence relating to hypothesized relationships implied by the proposed theoretical model.

**Structural Model GOF** The process of establishing the structural model's validity follows the general guidelines outlined in Stage 4. The observed data are still represented by the observed sample covariance matrix. It does not and should not change. However, a new SEM estimated covariance matrix is computed based on constraints added to the measurement model necessary to convert it to the proposed structural model. For example, some constructs that are presumed to be correlated in the measurement model may be considered independent theoretically. Thus, the relationship between the two would be constrained to zero and no path would link the two. This difference is a result of the structural relationships in the structural model. Therefore, for conventional SEM models, the  $\chi^2$  GOF statistic for the measurement model will be less than the  $\chi^2$  GOF for the structural model.

The overall fit can be assessed using the same criteria as the measurement model: using the  $\chi^2$  value for the structural model and at least one absolute index and one incremental index. These measures establish the validity of the structural model, but comparisons between the overall fit should also be made with the measurement model. Generally, the closer the structural model GOF comes to the measurement model, the better the structural model fit, because the measurement model fit provides an upper bound to the GOF of a conventional structural model.

## COMPETITIVE FIT

Competing models assessment is important in SEM. The primary objective is to ensure that the proposed model not only has acceptable model fit, but assessing whether one model outperforms a plausible alternative model. If not, then the alternative theoretical model is supported. Comparing models can be accomplished by assessing differences in incremental or parsimony fit indices along with differences in  $\chi^2$  GOF values for each model.

**Comparing Nested Models** A powerful test of alternative models is to compare models of similar complexity, yet representing varying theoretical relationships. A common approach is through **nested models**, where a model is nested within another model if it contains the same number of variables and can be formed from the other model by altering the relationships, such as either adding or deleting paths. Generally, competing nested SEM models are compared based on a **chi-square difference statistic ( $\Delta\chi^2$ )**. The  $\chi^2$  value from some baseline model (B) is subtracted from the  $\chi^2$  value of a lesser constrained, alternative nested model (A). Similarly, the difference in degrees of freedom is found, with one less degree of freedom for each additional path that is estimated. The following equation is used for computation:

$$\begin{aligned}\Delta\chi^2_{df} &= \chi^2_{df(B)} - \chi^2_{df(A)} \\ \Delta df &= df(B) - df(A)\end{aligned}$$

Because the difference of two  $\chi^2$  values is itself  $\chi^2$  distributed, we can test for statistical significance given a  $\Delta\chi^2$  difference value and the difference in degrees of freedom ( $\Delta df$ ). For example, for a model with one degree of freedom difference ( $\Delta df = 1$ , as when one additional path is added in model A), a  $\Delta\chi^2$  of 3.84 or better would be significant at the .05 level. The researcher would conclude that the model with one additional path (alternatively one could think of the model as having one less constraint) provides a better fit based on the significant reduction in the  $\chi^2$  GOF.

An example of a nested model in Figure 9.10 might be the addition of a structural path from the Supervision construct directly to the Job Search construct. This added path would reduce the degrees of freedom by one. The new model would be re-estimated and the  $\Delta\chi^2$  calculated. If it is larger than 3.84, then the researcher would conclude that the alternative model was a significantly better fit. Before the path is added, however, there must be theoretical support for the new relationship.

**Comparison to the Measurement Model** One useful comparison of models is between the CFA and the structural model fit. The structural model is composed of theoretical networks of relationships among constructs. In a conventional CFA (as will be detailed in the next chapter), all constructs are assumed to be related to all other constructs. As discussed earlier, a structural model will generally specify fewer relationships among constructs, because not every construct will be hypothesized to have a direct relationship with every other construct. In this sense, a structural model is more constrained (over-identified) than a measurement model, because more relationships are fixed to zero and not freely estimated. A way to think of this is that a structural model is formed from a measurement model by adding constraints. Adding a constraint cannot reduce the chi-square value. At best, if a relationship between constructs truly is zero, and the researcher constrains that relationship to zero by not specifying it in a structural model, the actual chi-square value will be unchanged by adding the constraint. When the two constructs truly are related, then adding the constraint will increase the actual chi-square value. Conversely, relaxing a constraint by including a relationship in the model should reduce the chi-square value or keep it the same.

As just described, adding or deleting paths (i.e., adding a path means a constraint has been relaxed and deleting a path means a constraint has been added) changes the degrees of freedom accordingly. Adding one constraint means the chi-square difference test will have one degree of freedom, adding two means the test will have two degrees of freedom, and so forth. When a measurement model and a structural model have approximately the same chi-square value, this means that the constraints added to form the structural model have not significantly added to the  $\chi^2$  value.

The overall  $\chi^2$  GOF for the Job Search example measurement model can be compared to the overall  $\chi^2$  GOF for the example's structural model shown in Figure 9.10. A  $\Delta\chi^2$  test can be used to compare these two models. The test would have  $\Delta df = 3$ , because three relationships that would be estimated in a CFA (measurement model test) are constrained to zero (i.e., not modeled) in the structural model. Specifically, no direct relationship from any exogenous construct (Supervision, Work Environment, or Coworkers) to the rightmost endogenous construct (Job Search) is modeled in this researcher's theory. If the  $\Delta\chi^2$  test with three degrees of freedom is insignificant, it would mean that constraining the measurement model (which includes all interconstruct covariances) by not allowing these three direct relationships did not significantly worsen fit. An insignificant  $\Delta\chi^2$  test between a measurement model and a structural model would generally provide supporting evidence for the proposed theoretical model.

Recently, an adjusted theoretical fit index (ATFI) has been proposed to quantify the comparison of measurement and structural model fit. The basic logic of the index, which mathematically compares the fits of the measurement and theoretical models, is captured in this equation [20]:

$$ATFI = \left( \frac{CFI_{CFA} - CFI_{TM}}{CFI_{CFA}} \right) \times \frac{DF_{TM}}{DF_{CFA}}$$

The equation takes the difference of the measurement model (CFA) and theoretical model CFIs as a ratio of the measurement model's fit and adjusts by a ratio of the degrees of freedom. Smaller values represent relatively better fits and a value of 0 would mean the measurement and theoretical fits are equal.

**Equivalent Models** Good fit statistics do not prove that a given theoretical model is the best way to explain the observed sample covariance matrix. As described earlier, any number of equivalent models may exist that offer equal or better fit to the same estimated covariance matrix. Therefore, any given model, even with good fit, is only one potential explanation. This means that good empirical fit does not prove that a given model is the

“only” true explanation. Favorable fit statistics are highly desirable, but many alternative models can provide an equivalent fit [46].

This issue further reinforces the need for building measurement models based on solid theory. Complex situations may produce many equivalent models. Yet, many models derived empirically may make little sense given the conceptual nature of the study’s variables. Thus, in the end empirical results provide some evidence of validity, but the researcher must provide theoretical evidence that is equally important in validating a model.

## TESTING STRUCTURAL RELATIONSHIPS

Good model fit is necessary but not sufficient to support all explanations represented by a structural theory. The researcher also must examine the individual parameter estimates that represent each link or “path.” A theoretical model’s unconstrained paths are considered valid to the extent that the parameter estimates are:

- 1 *Statistically significant and in the predicted direction.* That is, they are greater than zero for a positive relationship and less than zero for a negative relationship.
- 2 *Non-trivial.* Effect sizes should be checked for practical significance using the completely standardized loading estimates. The guideline here is the same as in other multivariate techniques. Coefficients can be statistically significant but practically meaningless, particularly as samples become large.

Therefore, the structural model shown in Figure 9.10 is considered acceptable only when it demonstrates acceptable model fit *and* the path estimates representing each of the four hypotheses are interpreted. The researcher also can examine the variance explained estimates for the endogenous constructs analogous to the analysis of  $R^2$  performed in multiple regression. More detail will be provided about procedures used in this stage in Chapter 10 and particularly in Chapter 11, including discussions on diagnostic measures for both the measurement and structural models.

## SEM Stages 4–6

As models become more complex, the likelihood of alternative models with equivalent fit increases

Multiple fit indices should be used to assess a model’s goodness-of-fit and should include:

- The  $\chi^2$  value and the associated  $df$ .
- One absolute fit index (i.e., GFI, RMSEA, or SRMR).
- One incremental fit index (i.e., CFI or TLI).
- One goodness-of-fit index (GFI, CFI, TLI, etc.).
- One badness-of-fit index (RMSEA, SRMR, etc.).

The ATFI provides a useful look at the relative fit of the theoretical structural and measurement models.

No single “magic” value for the fit indices separates good from poor models, and it is not practical to apply a single set of cut-off rules to all measurement models and, for that matter, to all SEM models of any type.

The quality of fit depends heavily on model characteristics, including sample size and model complexity:

Simple models with small samples should be held to strict fit standards; even an insignificant  $p$ -value for a simple model may not be meaningful.

More complex models with larger samples should not be held to the same strict standards, and so when samples are large and the model contains a large number of measured variables and parameter estimates, universal cut-off values of .95 on key GOF measures are unrealistic.

Several key learning objectives were provided for this chapter. These learning objectives together provide a basic overview of SEM. The basic overview should enable a better understanding of the more specific illustrations that follow in the next chapters.

**Understand the distinguishing characteristics of SEM.** SEM is a flexible approach to examining how things are related to each other. Therefore, SEM applications can appear quite different. However, three key characteristics of covariance-based SEM are (1) the estimation of multiple and interrelated dependence relationships and how well the overall model fits, (2) an ability to represent unobserved concepts in these relationships and correct for measurement error in the estimation process, (3) a focus on explaining the covariance among the measured items.

**Distinguish between variables and constructs.** The models typically tested using SEM involve both a measurement model and a structural model. Most of the multivariate approaches discussed in the previous chapters focused on analyzing variables directly. Variables are the actual items that are measured using a survey, observation, or some other measurement device. Variables are considered observable in the sense that we can obtain a direct measure of them. Constructs are unobservable or latent factors that are represented by a variate that consists of multiple variables. Simply put, multiple variables come together mathematically to represent a proxy of a construct. Constructs can be exogenous or endogenous. Exogenous constructs are the latent, multi-item equivalent of independent variables. They are constructs that are determined by factors outside of the model. Endogenous constructs are the latent, multi-item equivalent of dependent variables.

**Understand structural equation modeling and how it can be thought of as a combination of familiar multivariate techniques.** SEM can be thought of as a combination of exploratory factor analysis and multiple regression analysis. The measurement model part is similar to exploratory factor analysis in that it also demonstrates how measured variables load on a smaller number of factors (i.e., constructs). Several different regression analogies apply, but key among them is the fact that endogenous (outcome) constructs are predicted using multiple other constructs in the same way that independent variables predict dependent variables in multiple regression.

**Know the basic conditions for causality and how SEM can help establish a cause-and-effect relationship.** Theory can be defined as a systematic set of relationships providing a consistent and comprehensive explanation of a phenomenon. SEM has become the most prominent multivariate tool for testing behavioral theory. SEM's history grows out of the desire to test causal models. Theoretically, four conditions must be present to establish causality: (1) covariation, (2) temporal sequence, (3) non-spurious association, and (4) theoretical support. SEM can establish evidence of covariation through the tests of relationships represented by a model. SEM cannot, as a rule, demonstrate that cause occurred before the effect, because cross-sectional data are most often used in SEM. SEM models using longitudinal data can help demonstrate temporal sequentiality. Evidence of nonspurious association between a cause and effect can be supplied, at least in part, by SEM. If the addition of other alternative causes does not eliminate the relationship between the cause and effect, then the causal inference becomes stronger. Finally, theoretical support can only be supplied through reason. Empirical findings alone cannot render a relationship sensible. Thus, SEM can be useful in establishing causality, but simply using SEM on any given data does not mean that causal inferences can be established.

**Explain the basic types of relationships involved in SEM.** The four key theoretical relationship types in a SEM model are described in Figure 9.1, which also shows the conventional graphical representation of each type. The first shows relationships between latent constructs and measured variables. Latent constructs are represented with ovals and measured variables are represented with rectangles. The second shows simple covariation or correlation between constructs. It does not imply any causal sequence, and it does not distinguish between exogenous and endogenous constructs. These first two relationship types are fundamental in forming a measurement model. The third relationship type shows how an exogenous construct is related to an endogenous construct and can represent a causal inference in which the exogenous construct is a cause and the endogenous construct is an effect. The fourth relationship type shows how one endogenous construct is related to another. It can also represent a causal sequence from one endogenous construct to another.

**Understand that the objective of SEM is to explain covariance and how it translates into the fit of a model.** SEM is sometimes known as covariance structure analysis. The algorithms that perform SEM estimation have the

goal of explaining the observed covariance matrix of variables,  $S$ , using an estimated covariance matrix,  $\Sigma_k$ , calculated using the regression equations that represent the researcher's model. In other words, SEM is looking for a set of parameter estimates producing estimated covariance values that most closely match observed covariance values. The closer these values come, the better the model is said to fit. Fit indicates how well a specified model reproduces the covariance matrix among the measured items. The basic SEM fit statistic is the  $\chi^2$  statistic. However, its sensitivity to sample size and model complexity brought about the development of many other fit indices. Fit is best assessed using multiple fit indices. It is also important to realize that no magic values determine when a model is proved best on fit. Rather, the model context must be taken into account in assessing fit. Simple models with small samples should be held to different standards than more complex models tested with larger samples.

**Know how to represent a model visually using a path diagram.** The entire set of relationships that make up a SEM model can be represented visually using a path diagram. Each type of relationship is conventionally represented with a different type of arrow and abbreviated with a different character. Figure 9.10 depicts a path diagram showing both a measurement and a structural model. The inner portion represents the structural model. The outer portion represents the measurement model.

**List the six stages of structural equation modeling and understand the role of theory in the process.** Figure 9.7 lists the six stages in the SEM process. It begins with choosing the variables that will be measured. It concludes with assessing the overall structural model fit. It should also be emphasized that theory plays a key role in each step of the process. The goal of a SEM is to provide a test of a theory. Thus, without theory a true SEM test cannot be conducted.

As mentioned previously, this chapter does not include an extended HBAT example. Rather, a new HBAT example will be introduced in the next chapter. Over the next three chapters, it will illustrate the complete use of SEM to test relationships that will help HBAT management make key decisions.

What is the difference between a latent construct and a measured variable?

What are the distinguishing characteristics of covariance-based SEM?

Describe how the estimated covariance matrix in CB-SEM analysis ( $\Sigma_k$ ) can be computed. Why do we compare it to  $S$ ?

How is structural equation modeling similar to the other multivariate techniques discussed in the earlier chapters?

Explain what is a "constraint" in SEM and include a discussion of how constraints are depicted in a graphical SEM.

What is a theory? How is a theory represented in a SEM framework?

What is a spurious correlation? How might it be revealed using SEM?

What is fit?

What is the difference between an absolute and a relative fit index?

How does sample size affect structural equation modeling?

Why are no magic values available to distinguish good fit from poor fit across all situations?

Draw a path diagram with two exogenous constructs and one endogenous construct. The exogenous constructs are each measured by five items and the endogenous construct is measured by four items. Both exogenous constructs are expected to be related negatively to the endogenous construct.

For a further guidance on how to report quantitative results, including those from SEM, in academic papers, see Applebaum et al. (2018), "Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report," *American Psychologist*, 73 (1), 3–25. A list of suggested readings and other materials relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com)).

# Estimating Relationships Using Path Analysis

What was the purpose of developing the path diagram? Path diagrams are the basis for path analysis, the procedure for empirical estimation of the strength of each relationship (path) depicted in the path diagram. Path analysis calculates the strength of the relationships using only a correlation or covariance matrix as input. We will describe the basic process in the following section, using a simple example to illustrate how the estimates are actually computed.

The first step is to identify all relationships that connect any two constructs. Path analysis enables us to decompose the simple (bivariate) correlation between any two variables into the sum of the compound paths connecting these points. The number and types of compound paths between any two variables are strictly a function of the model proposed by the researcher.

A compound path is a path along the arrows of a path diagram that follow three rules:

After going forward on an arrow, the path cannot go backward again; but the path can go backward as many times as necessary before going forward.

The path cannot go through the same variable more than once.

The path can include only one curved arrow (correlated variable pair).

When applying these rules, each path or arrow represents a path. If only one arrow links two constructs (path analysis can also be conducted with variables), then the relationship between those two is equal to the parameter estimate between those two constructs. For now, this relationship can be called a direct relationship. If there are multiple arrows linking one construct to another as in  $X \rightarrow Y \rightarrow Z$ , then the effect of  $X$  on  $Z$  is equal to the product of the parameter estimates for each arrow and is termed an indirect relationship. This concept may seem quite complicated but an example makes it easy to follow:

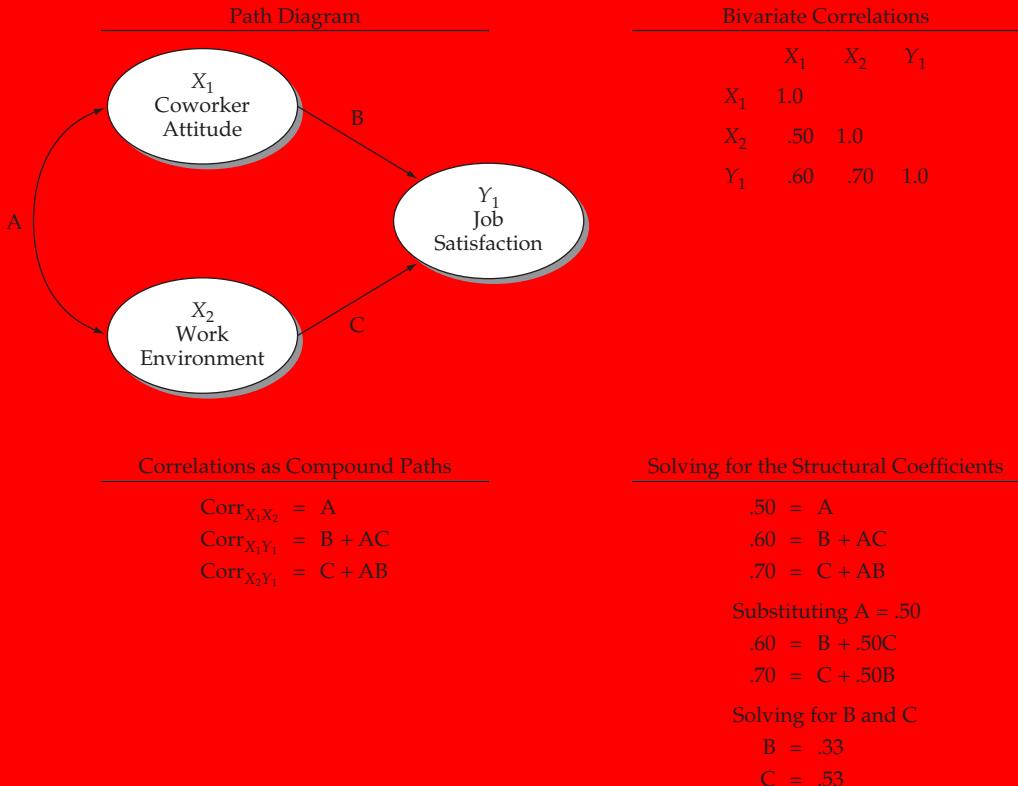
Figure 9A.1 portrays a simple model with two exogenous constructs ( $X_1$  and  $X_2$ ) causally related to the endogenous construct ( $Y_1$ ). The correlational path A is  $X_1$  correlated with  $X_2$ , path B is the effect of  $X_1$  predicting  $Y_1$ , and path C shows the effect of  $X_2$  predicting  $Y_1$ . The value for  $Y_1$  can be stated simply with a regression-like equation:

$$Y_1 = b_1X_1 + b_2X_2$$

We can now identify the direct and indirect paths in our model. For ease in referring to the paths, the causal paths are labeled A, B, and C.

<i>Direct Paths</i>	<i>Indirect Paths</i>
$A = X_1$ to $X_2$	
$B = X_1$ to $Y_1$	$AC = X_1$ to $Y_1$
$C = X_2$ to $Y_1$	$AB = X_2$ to $Y_1$

### Calculating Structural Coefficients with Path Analysis



With the direct and indirect paths now defined, we can represent the correlation between each construct as the sum of the direct and indirect paths.

The three unique correlations among the constructs can be shown to be composed of direct and indirect paths as follows:

$$\text{Corr}_{X_1 X_2} = A$$

$$\text{Corr}_{X_1 Y_1} = B + AC$$

$$\text{Corr}_{X_2 Y_1} = C + AB$$

First, the correlation of  $X_1$  and  $X_2$  is simply equal to  $A$ . The correlation of  $X_1$  and  $Y_1$  ( $\text{Corr}_{X_1 Y_1}$ ) can be represented as two paths:  $B$  and  $AC$ . The symbol  $B$  represents the direct path from  $X_1$  to  $Y_1$ , and the other path (a compound path) follows the curved arrow from  $X_1$  to  $X_2$  and then to  $Y_1$ . Likewise, the correlation of  $X_2$  and  $Y_1$  can be shown to be composed of two causal paths:  $C$  and  $AB$ .

Once all the correlations are defined in terms of paths, the values of the observed correlations can be substituted and the equations solved for each separate path. The paths then represent either the causal relationships between constructs (similar to a regression coefficient) or correlational estimates.

Using the correlations as shown in Figure 9A.1, we can solve the equations for each correlation (see Figure 9A.1) and estimate the causal relationships represented by the coefficients  $b_1$  and  $b_2$ .

We know that A equals .50, so we can substitute this value into the other equations. By solving these two equations, we get values of  $B(b_1) = .33$  and  $C(b_2) = .53$ . The actual calculations are shown in Figure 9A.1. This approach enables path analysis to solve for any causal relationship based only on the correlations among the constructs and the specified causal model.

As you can see from this simple example, if we change the path model in some way, the causal relationships will change as well. Such a change provides the basis for modifying the model to achieve better fit, if theoretically justified.

With these simple rules, the larger model can now be modeled simultaneously, using correlations or covariances as the input data. We should note that when used in a larger model, we can solve for any number of interrelated equations. Thus, dependent variables in one relationship can easily be independent variables in another relationship. No matter how large the path diagram gets or how many relationships are included, path analysis provides a way to analyze the set of relationships.

# SEM Abbreviations

The following guide will aid in the pronunciation and understanding of common SEM abbreviations. SEM terminology often is abbreviated with a combination of Greek characters and Roman characters to help distinguish different parts of a SEM model.

Symbol	Pronunciation	Meaning
$\xi$	xi (KSI or KZI)	A construct associated with measured X variables
$\lambda_x$	lambda "x"	A path representing the factor loading between a latent construct and a measured x variable
$\lambda_y$	lambda "y"	A path representing the factor loading between a latent construct and a measured y variable
$\Lambda$	capital lambda	A way of referring to a set of loading estimates represented in a matrix where rows represent measured variables and columns represent latent constructs
$\eta$	eta ("eight-ta")	A construct associated with measured Y variables
$\phi$	phi (fi)	A path represented by an arced two-headed arrow representing the covariation between one $\xi$ and another $\xi$
$\Phi$	capital phi	A way of referring to the covariance or correlation matrix between a set of $\xi$ constructs
$\gamma$	gamma	A path representing a causal relationship (regression coefficient) from a $\xi$ to an $\eta$
$\Gamma$	capital gamma	A way of referring to the entire set of $\gamma$ relationships for a given model
$\beta$	beta ("bay-ta")	A path representing a causal relationship (regression coefficient) from one $\eta$ construct to another $\eta$ construct
$\Beta$	capital beta	A way of referring to the entire set of $\beta$ relationships for a given model
$\delta$	delta	The error term associated with an estimated, measured x variable
$\theta_x$	theta ("they-ta") delta	A way of referring to the residual variances and covariances associated with the x estimates; the error variance items are the diagonal
$\varepsilon$	epsilon	The error term associated with an estimated, measured y variable
$\theta_\varepsilon$	theta-epsilon	A way of referring to the residual variances and covariances associated with the y estimates; the error variance items are the diagonal
$\zeta$	zeta ("zay-ta")	A way of capturing the covariation between $\eta$ construct errors
$\tau$	tau (rhymes with "now")	The intercept terms for estimating a measured variable
$\kappa$	kappa	The intercept terms for estimating a latent construct
$\chi^2$	chi (ki)-squared	The likelihood ratio

# Detail on Selected GOF Indices

The chapter describes how researchers developed many different fit indices that represent the GOF of a SEM model in different ways. Here, a bit more detail is provided about some of the key indices in an effort to provide a better understanding of just what information is contained in each.

If we think of  $F_k$  as the minimum fit function after a SEM model has been estimated using  $k$  degrees of freedom ( $S - \Sigma_k$ ), and we think of  $F_0$  as the fit function that would result if all parameters were zero (everything is unrelated to each other; no theoretical relationships), then we can define the GFI simply as:

$$\text{GFI} = 1 - \frac{F_k}{F_0}$$

A model that fits well produces a ratio of  $F_k/F_0$  that is quite small. Conversely, a model that does not fit well produces  $F_k/F_0$  that is relatively large because  $F_k$  would not differ much from  $F_0$ . This ratio works something like the ratio of SSE/SST discussed in Chapter 4. In the extreme, if a model failed to explain any true covariance between measured variables,  $F_k/F_0$  would be 1, meaning the GFI would be 0.

Computation of RMSEA is rather straightforward and provided here to demonstrate how statistics try to correct for the problems of using the  $\chi^2$  statistic alone.

$$\text{RMSEA} = \sqrt{\frac{(\chi^2 - df_k)}{(N - 1)}}$$

Note that the  $df$  are subtracted from the numerator in an effort to capture model complexity. The sample size is used in the denominator to take it into account. To avoid negative RMSEA values, the numerator is set to 0 if  $df_k$  exceeds  $\chi^2$ .

The general computational form of the CFI is:

$$\text{CFI} = 1 - \frac{(\chi^2_k - df_k)}{(\chi^2_N - df_N)}$$

Here, the subscript  $k$  represents values associated with the researcher's specified model or theory, that is, the resulting fit with  $k$  degrees of freedom. The subscript  $N$  denotes values associated with the statistical null model. Additionally, the equation is normed to values between 0 and 1—with higher values indicating better fit—by substituting an appropriate value (i.e., 0) if a  $\chi^2$  value is less than the corresponding degrees of freedom.

The equation for the TLI is provided here for comparison purposes:

$$\text{TLI} = \frac{\left[ \left( \frac{\chi^2_N}{df_N} \right) - \left( \frac{\chi^2_k}{df_k} \right) \right]}{\left[ \left( \frac{\chi^2_N}{df_N} \right) - 1 \right]}$$

Once again,  $N$  and  $k$  refer to the null and specified models, respectively. The TLI is not normed and thus its values can fall below 0 or above 1. It produces values similar to the CFI in most situations.

The parsimony ratio (PR) forms the basis for parsimony GOF measures [31]:

$$\text{PR} = \frac{df_k}{df_t}$$

As can be seen by the formula, it is the ratio of degrees of freedom used by a model to the total degrees of freedom available. Thus, other indices are adjusted by PR to form parsimony fit indices. Although parsimony fit indices can be useful, they tend to strongly favor the more parsimonious measures. These measures have existed for quite some time but are still not widely applied.

- 1 Allison, P. D. 2003. Missing Data Techniques for Structural Equations Models. *Journal of Abnormal Psychology* 112 (November): 545–56.
- 2 Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumacker (Eds.), *Advanced Structural Equation Modeling*. Mahwah, NJ: Lawrence Erlbaum Publishers.
- 3 Arbuckle, J. L. (2014). *Amos 7.0 User's Guide*. Chicago, IL: SPSS Inc.
- 4 Bagozzi, R. P., and Y. Yi. 1988. On the Use of Structural Equation Models in Experimental Designs. *Journal of Marketing Research* 26 (August): 271–84.
- 5 Bentler, P. M. 1990. Comparative Fit Indexes in Structural Models. *Psychological Bulletin* 107: 238–46.
- 6 Bentler, P. M. 2008. *EQS6 Structural Equations Program Manual*. Temple City, CA: Multivariate Software, Inc.
- 7 Bentler, P. M., and D. G. Bonnett. 1980. Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin* 88: 588–606.
- 8 Blalock, H. M. 1962. Four-Variable Causal Models and Partial Correlations. *American Journal of Sociology* 68: 182–94.
- 9 Byrne, B. 1998. *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 10 Churchill, G. A. 1979. A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research* 16 (February): 64–73.
- 11 Collins, L. M., J. L. Schafer, and C. M. Kam. 2001. A Comparison of Inclusive and Restrictive Strategies in Modern Missing-Data Procedures. *Psychological Methods* 6: 352–70.
- 12 Cudeck, R. 1989. Analysis of Correlation Matrices Using Covariance Structure Models. *Psychological Bulletin* 105: 317–27.
- 13 DeVellis, Robert. 1991. *Scale Development: Theories and Applications*. Thousand Oaks, CA: Sage.
- 14 Enders, C. K., and D. L. Bandalos. 2001. The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling* 8(3): 430–59.
- 15 Enders, C. K., and J. L. Peugh. 2004. Using an EM Covariance Matrix to Estimate Structural Equation Models with Missing Data: Choosing an Adjusted Sample Size to Improve the Accuracy of Inferences. *Structural Equations Modeling* 11(1): 1–19.
- 16 Fan, X., B. Thompson, and L. Wang. 1999. Effects of Sample Size, Estimation Methods, and Model Specification on Structural Equation Modeling Fit Indexes. *Structural Equation Modeling* 6: 56–83.
- 17 Feinian, C., P. J. Curran, K. A. Bollen, J. Kirby, and P. Paxton (2008), "An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models," *Sociological Methods & Research*, 36 (4), 462–94.
- 18 Graham, J. W. (2003), "Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models," *Structural Equation Modeling* 10 (1), 80–100.
- 19 Habelmo, T. 1943. The Statistical Implications of a System of Simultaneous Equations. *Econometrica* 11: 1–12.
- 20 Hair, J. F., B. J. Babin, and N. Krey (2017). Covariance-Based Structural Equation Modeling in the Journal of Advertising: Review and Recommendations. *Journal of Advertising*, 46 (1): 163–177.
- 21 Hayduk, L. A. 1996. *LISREL Issues, Debates and Strategies*. Baltimore: Johns Hopkins University Press.
- 22 Hayduck, L., G. Cummings, K. Boadu, H. P. Robinson, and S. Boulianne (2007), "Testing! Testing! One, Two, Three—Testing Theory in Structural Equation Models!" *Personality and Individual Differences*, 42: 841–50.
- 23 Hershberger, S. L. 2003. The Growth of Structural Equation Modeling: 1994–2001. *Structural Equation Modeling* 10(1): 35–46.
- 24 Hoelter, J. W. 1983. The Analysis of Covariance Structures: Goodness-of-Fit Indices. *Sociological Methods and Research* 11: 324–44.
- 25 Hu, L., and P. M. Bentler. 1999. Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equations Modeling* 6(1): 1–55.
- 26 Hunt, S. D. 2002. *Foundations of Marketing Theory: Toward a General Theory of Marketing*. Armonk, NY: M.E. Sharpe.
- 27 Jöreskog, K. G. 1970. A General Method for Analysis of Covariance Structures. *Biometrika* 57: 239–51.
- 28 Jöreskog, K. G. 1981. Basic Issues in the Application of LISREL. *Data* 1: 1–6.
- 29 Jöreskog, K. G., and D. Sörbom. 1976. *LISREL III: Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods*. Chicago: National Educational Resources, Inc.
- 30 Jöreskog, K. G., and D. Sörbom. (2015). *LISREL 9.20 for Windows*. Skokie, IL: Scientific Software International, Inc.
- 31 Kenny, D. A., and D. B. McCoach. 2003. Effect of the Number of Variables on Measures of Fit in Structural Equations Modeling. *Structural Equations Modeling* 10(3): 333–51.
- 32 Kline, R. B. (1998), Software Programs for Structural Equation Modeling: AMOS, EQS, and LISREL." *Journal of Psychoeducational Assessment*, 16: 302–23.
- 33 Little, T. D., W. A. Cunningham, G. Shahar, and K. F. Widaman. 2002. To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling* 9: 151–73.

- 34 MacCallum, R. C. 2003. Working with Imperfect Models. *Multivariate Behavioral Research* 38(1): 113–39.
- 35 MacCallum, R. C., K. F. Widaman, K. J. Preacher, and S. Hong. 2001. Sample Size in Factor Analysis: The Role of Model Error. *Multivariate Behavioral Research* 36(4): 611–37.
- 36 Maiti, S. S., and B. N. Mukherjee. 1991. Two New Goodness-of-Fit Indices for Covariance Matrices with Linear Structure. *British Journal of Mathematical and Statistical Psychology* 44: 153–80.
- 37 Marsh, H. W., and J. Balla. 1994. Goodness-of-Fit in CFA: The Effects of Sample Size and Model Parsimony. *Quality & Quantity* 28 (May): 185–217.
- 38 Marsh, H. W., K. T. Hau, and Z. Wen. 2004. In Search of Golden Rules: Comment on Hypothesis Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling* 11(3): 320–41.
- 39 Marsh, H. W., K. T. Hau, J. R. Balla, and D. Grayson. 1988. Is More Ever Too Much? The Number of Indicators per Factors in Confirmatory Factor Analysis. *Multivariate Behavioral Research* 33: 181–222.
- 40 Mulaik, S. A., L. R. James, J. Val Alstine, N. Bennett, S. Lind, and C. D. Stilwell. 1989. Evaluation of Goodness-of-Fit Indices for Structural Equations Models. *Psychological Bulletin* 105 (March): 430–45.
- 41 Muthén, L. K., and B. O. Muthén. 2017. *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén.
- 42 Netemeyer, R. G., W. O. Bearden, and S. Sharma. 2003. *Scaling Procedures: Issues and Applications*. Thousand Oaks, CA: Sage.
- 43 Olsson, U. H., T. Foss, and E. Breivik. 2004. Two Equivalent Discrepancy Functions for Maximum Likelihood Estimation: Do Their Test Statistics Follow a Noncentral CM-square Distribution Under Model Misspecification? *Sociological Methods & Research* 32 (May): 453–510.
- 44 Olsson, U. H., T. Foss, S. V. Troye, and R. D. Howell. 2000. The Performance of ML, GLS and WLS Estimation in Structural Equation Modeling Under Conditions of Misspecification and Nonnormality. *Structural Equation Modeling* 7: 557–95.
- 45 Pearl, J. 1998. Graphs, Causality and Structural Equation Models. *Sociological Methods & Research* 27 (November): 226–84.
- 46 Raykov, T., and G. A. Marcoulides. 2001. Can There Be Infinitely Many Models Equivalent to a Given Covariance Structure Model? *Structural Equation Modeling* 8(1): 142–49.
- 47 Rigdon, E. E. 1996. CFI Versus RMSEA: A Comparison of Two Fit Indices for Structural Equation Modeling. *Structural Equation Modeling* 3(4): 369–79.
- 48 Rubin, D. B. 1976. Inference and Missing Data. *Psychometrika* 63: 581–92.
- 49 Savalei, V. (2008). Is the ML Chi-Square Ever Robust to Nonnormality? A Cautionary Note with Missing Data, *Structural Equation Modeling*, 15 (1): 1–22.
- 50 Sharma, S., S. S. Mukherjee, A. Kumar, and W. R. Dillon. 2005. A Simulation Study to Investigate the Use of Cutoff Values for Assessing Model Fit in Covariance Structure Models. *Journal of Business Research* 58 (July): 935–43.
- 51 Sobel, M. E. 1998. Causal Inferences in Statistical Models of the Process of Socioeconomic Achievement. *Sociological Methods & Research* 27 (November): 318–48.
- 52 Tanaka, J. 1993. Multifaceted Conceptions of Fit in Structural Equation Models. In K. A. Bollen and J. S. Long (eds.), *Testing Structural Equation Models*. Newbury Park, CA: Sage.
- 53 Tanaka, J. S., and G. J. Huba. 1985. A Fit-Index for Covariance Structure Models Under Arbitrary GLS Estimation. *British Journal of Mathematics and Statistics* 42: 233–39.
- 54 Wang, L. L., X. Fan, and V. L. Wilson. 1996. Effects of Nonnormal Data on Parameter Estimates for a Model with Latent and Manifest Variables: An Empirical Study. *Structural Equation Modeling* 3(3): 228–47.
- 55 Wright, S. 1921. Correlation and Causation. *Journal of Agricultural Research* 20: 557–85.

# 10 SEM: Confirmatory Factor Analysis

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

Distinguish between exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

Understand the basic principles of statistical identification and know some of the primary causes of CFA identification problems.

Know how to represent a measurement model using a path diagram.

Understand the concept of model fit as it applies to measurement models and be able to assess the fit of a confirmatory factor analysis model.

Assess the construct validity of a measurement model.

Use CFA diagnostics to spot problems with a SEM model.

## Chapter Preview

The previous chapter introduced the basics of covariance-based structural equation modeling. It described the two basic parts to a conventional structural equation model. This chapter addresses the first part by demonstrating how confirmatory processes can test a proposed measurement theory. The measurement theory can be represented with a model that shows how measured variables come together to represent constructs. Confirmatory factor analysis (CFA) enables us to test how well the measured variables represent a set of theoretical latent constructs. CFA offers the key advantage of analytically testing a precise, conceptually grounded theory explaining how different measured variables represent important psychological, sociological, or business constructs. When CFA fit results are combined with construct validity tests, researchers know the quality of the theoretical measurement model.

The importance of assessing the quality of measures in a behavioral model cannot be overstated. No valid conclusions exist without valid measurement. **The procedures described in this chapter demonstrate how the validity of a measurement model can be tested using CFA.**

## Key Terms

Before beginning this chapter, review the key terms to develop an understanding of the concepts and terminology used. Throughout the chapter the key terms appear in **boldface**. Other points of emphasis in the chapter and key term cross-references are *italicized*. Illustrative examples are in shaded text.

**Average variance extracted (AVE)** A summary measure of convergence among a set of items representing a reflectively measured latent construct. It is the average percentage of variation explained (*variance extracted*) among the items of a construct.

**Between-construct error covariance** Covariance between two error terms of measured variables indicating different constructs.

**Communality** See *variance extracted*.

**Congeneric measurement model** Measurement model consisting of several *unidimensional* constructs with all cross-loadings assumed to be zero. Represented in CFA by all possible cross-loadings and all *between-* and *within-construct error covariances* being fixed at zero.

**Constraints** Fixing a potential relationship in a SEM model to some specified value (even if fixed to zero) rather than allowing the value to be estimated (free).

**Construct (composite) reliability (CR)** Measure of reliability and internal consistency of the measured variables representing a latent construct. Must be established before *construct validity* can be assessed.

**Construct validity** Extent to which a set of measured variables actually represents the theoretical latent construct those variables are designed to measure.

**Convergent validity** Extent to which indicators of a specific construct converge or share a high proportion of variance in common.

**Discriminant validity** Extent to which a construct is truly distinct from other constructs. The construct correlations can be useful in this assessment.

**Face validity** Extent to which the content of the items is consistent with the construct definition, based solely on the researcher's judgment.

**Formative measurement** Implies that (1) a set of measured variables forms a concept, and (2) the error in measurement is an inability to fully predict the concept. Here, a construct is not considered a latent factor. See also *reflective measurement theory*.

**Heywood case** a solution producing an error variance estimate of less than zero (a negative error variance—or alternatively, more than 100% prediction) leading to an improper solution.

**Identification** Whether enough information exists to identify (compute) a solution for a set of structural equations using covariance-based SEM. An identification problem leads to an inability of the proposed model to generate reliable estimates and makes a solution mathematically impossible. The three possible conditions of identification are *overidentified*, *just-identified*, and *underidentified*.

**Just-identified** SEM model containing just enough degrees of freedom to estimate all free parameters. Just-identified models have perfect fit by definition, meaning that a fit assessment is not meaningful.

**Measurement model** Specification of the *measurement theory* that shows how constructs are *operationalized* by sets of measured variables. The measurement theory, not the factor routine, dictates the number of factors, which items load on each factor, as well as any constraints representing variables or latent factors that are theoretically unrelated.

**Measurement theory** Series of relationships that suggest how measured variables represent a construct not measured directly (latent). A measurement theory can be represented by a series of regression-like equations mathematically relating a factor (construct) to the measured variables.

**Modification index** Amount the overall model  $\chi^2$  value would be reduced by freeing (estimating) any single particular path that is not currently estimated.

**Nomological validity** Test of validity that examines whether the correlations between the constructs in the *measurement theory* make sense. Construct correlations can be useful in this assessment.

**Operationalization** Manner in which a construct is represented. With CFA, a set of measured variables represents a construct.

**Order condition** Requirement that the degrees of freedom for a model be greater than zero; that is, the number of unique covariance and variance terms less the number of free parameter estimates must be positive.

**Overidentified model** Model that has more unique covariance and variance terms than parameters to be estimated, yielding positive net degrees of freedom. This is the preferred type of identification for a SEM model.

**Parameter** Numerical representation of some characteristic of a population. Parameters are numerical characteristics of the SEM relationships, comparable to regression coefficients in multiple regression.

**Psychometrics** Branch of psychology that addresses how to properly go about the quantitative measurement of psychological concepts such as personal traits, attitudes, affect, opinion, etc.

**Rank condition** Requirement that each individual parameter estimated be uniquely, algebraically defined. If you think of a set of equations that could define any dependent variable, the rank condition is violated if any two equations are mathematical duplicates.

**Reflective measurement** Based on the assumptions that (1) latent constructs cause the measured variables (indicators), and (2) measurement error results in an inability to fully predict the measured variables. It is the typical representation for a latent construct. See also *formative measurement*.

**Residuals** Individual differences between observed covariance terms and the estimated covariance terms.

**Specification search** Empirical trial-and-error approach that may lead to sequential changes in the model based on key model diagnostics.

**Squared multiple correlations** Values representing the extent to which a variable's variance is explained by a latent factor(s); like communality from EFA.

**Standardized residuals** Residuals divided by the standard error of the residual. Used as a diagnostic measure of model fit.

**Tau equivalence** Assumption that a *measurement model* is *congeneric* and that all factor loadings on a given factor are equal.

**Three-indicator rule** Refers to *congeneric measurement models* in which all constructs having at least three indicators are *identified*.

**Underidentified model** Model with more parameters to be estimated than there are item variance and covariances. The term *unidentified* is used in the same way as underidentified.

**Unidentified model** See *underidentified model*.

**Unidimensional measures** Set of measured variables (indicators) with only one underlying latent construct. That is, the indicator variables load on only one construct.

**Variance extracted** Total amount of variance a measured variable has in common with the constructs upon which it loads. Thus, it can be thought of as the variance explained in a measured variable by the construct. Also referred to as a *communality* and often as Average Variance Extracted (AVE).

**Within-construct error covariance** Covariance between two error terms of measured variables that are indicators of the same construct.

## What Is Confirmatory Factor Analysis?

This chapter begins by providing a description of confirmatory factor analysis (CFA). CFA is a way of testing how well a prespecified measurement theory composed of measured variables and factors fits reality as captured by data. The chapter illustrates this process by showing how CFA is similar to other multivariate techniques. Then, a simple example is provided. A few key aspects of CFA are discussed prior to describing the CFA stages in more detail and demonstrating CFA with an extended illustration.

### CFA AND EXPLORATORY FACTOR ANALYSIS

Chapter 3 described procedures for conducting exploratory factor analysis (EFA). EFA explores data and provides information suggesting empirically how many factors are needed to represent that data. With EFA, all measured variables load on *every* factor producing a factor loading estimate for each variable on all factors. Simple structure results when each measured variable loads highly on only one factor and has smaller loadings on all other factors (e.g., loadings <.4). EFA becomes useful in identifying variables with little communality with others being considered.

In contrast to CFA, EFA factors are derived from statistical results, not from theory. This means that the researcher runs the software and lets the underlying pattern of the data determine the factor structure. Thus, EFA, and Principal Components (PCA), are conducted without knowing how many factors really exist (if any), or which variables belong with which constructs. When either is applied, the researcher uses established guidelines to determine which variables load best on a factor and how many factors to keep. The researcher names the factors *after* the exploratory analysis is performed. In this respect, CFA and EFA are not the same. Note that in this chapter the terms *factor* and *construct* are used interchangeably.

Chapter 3 conducted EFA on 13 variables from the HBAT dataset. Based on the eigenvalues and the pattern of loadings, a four-factor solution was deemed most appropriate. The four factors were named based on the variables loading highly on each factor. Using this process, the factors were named (1) Customer Service, (2) Marketing, (3) Technical Support, and (4) Product Value (see Chapter 3 for more details).

In contrast to EFA, with CFA the researcher must specify both the number of factors that exist for a set of variables and which factor each variable will load on *before* results can be computed. Thus, the statistical technique does not assign variables to factors. Instead, the researcher makes this assignment based on the theory being tested before any results can be obtained. Moreover, if good measurement principals are employed in the theory, a variable can load on only a single factor and cross-loadings (loading on more than a single factor) are not permitted. Just as importantly, the measurement theory should specify independence for all error variance terms. Thus, cross-loadings and error variance correlations are constrained to zero. As a result, a factor uniquely determines each of its indicator variables. CFA then tests the extent to which a researcher's *a priori, theoretical* pattern of factor loadings on prespecified constructs (variables loading on specific constructs) represents the actual data. Thus, instead of

allowing the statistical method to determine the number of factors and loadings as in EFA, CFA statistics tell us how well our theoretical specification of the factors matches reality (the actual data). In a sense, CFA reveals the degree of confirmation for our preconceived measurement theory.

## MEASUREMENT THEORY AND PSYCHOMETRICS

The paragraph above describes how CFA is used to test a measurement theory. A **measurement theory** specifies precisely how measured variables logically and systematically represent constructs involved in a theoretical model. In other words, measurement theory specifies a series of relationships and constraints that suggest how measured variables represent a latent factor. The measurement theory may then be combined with a structural theory to fully specify a SEM model.

In fact, an entire branch of psychology, known as **psychometrics**, addresses how to properly go about the quantitative measurement of latent, psychological constructs. Psychometric theory provides the framework by which individual researchers can create a valid measurement theory. Factor analysis, in particular CFA, is the fundamental, multivariate statistical tool for psychometricians.

Measurement theory requires that a construct first be defined. Therefore, unlike EFA, with CFA a researcher uses theory to specify *a priori* the number of factors, as well as which variables load on those factors. This specification is often referred to as the way the conceptual constructs in a measurement model are operationalized. CFA cannot be conducted without a measurement theory. One can perform EFA without any prespecified theory or knowledge of factors that may exist. As such, the E (exploratory) in EFA is emphasized.

## A SIMPLE EXAMPLE OF CFA AND SEM

We are now going to illustrate a simple CFA utilizing two constructs from the example first introduced in Chapter 9. We will explain how the measurement theory is represented in a path diagram.

Consider a situation where a researcher is interested in studying factors that impact employee job satisfaction. After reviewing the relevant theory, the researcher concludes that two factors have the largest impact—Supervision and Work Environment. The measured variables for both factors are evaluated using a 7-point, agree-disagree Likert scale.

The construct Supervision can be defined as what workers think about the management capabilities of their immediate supervisor. The supervision construct can be represented by the following four items:

- My supervisor recognizes my potential.
- My supervisor helps me resolve problems at work.
- My supervisor understands the challenges of balancing work and home demands.
- My supervisor supports me when I have a problem.

The construct Work Environment can be defined as the aspects of the environment where people work that impact their productivity. The work environment construct can be represented by the following four measured variables:

- Supervisors and workers have similar values and ideas about what this organization should be doing.
- My organization provides the equipment needed to perform my job well.
- The temperature of my office and other working areas is comfortable.
- The physical arrangement of work areas at my organization helps me to manage my time on the job well.

## A VISUAL DIAGRAM

Measurement theories, like structural equation models in general, can be depicted using visual representations, which are called *path diagrams*. A path diagram depicts the theoretical pattern of linkages and constraints between specific

measured variables and their associated constructs and the relationships among constructs. “Paths” from the latent construct to the measured items (loadings) are based on a measurement theory formed using good psychometric principles. Psychometric principles include:

- 1 Only the loadings theoretically linking a measured item to its corresponding latent factor are freely estimated and thus shown in the path diagram.
- 2 All other possible loadings are assumed to be equal to, and therefore *constrained* to, zero. The absence of connections represents no association. The effect is the same as drawing a connection and specifying that it is 0.
- 3 No covariance exists among the residuals, which are represented by error variance terms in a model. Thus, no connections exist among the unobserved concepts representing error variances, meaning that a constraint of 0 (independence) exists for each possible connection among error variance terms.

Point two highlights a particularly key difference between EFA and CFA in that EFA produces a loading for every variable on every factor, and the researcher has no control over this fact.

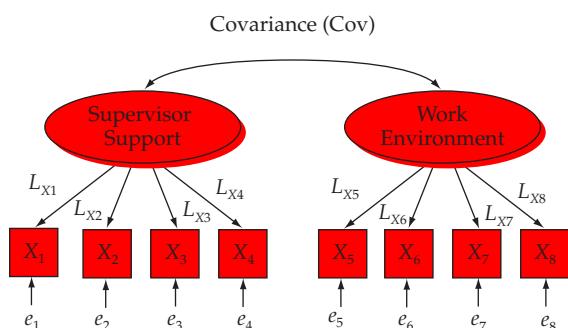
When conducting a CFA, we specify five elements to develop a measurement model:

- 1 the latent constructs,
- 2 the measured variables (indicators),
- 3 the pattern of item loadings on specific constructs,
- 4 the relationships among constructs,
- 5 the error variance and covariance terms for each indicator.

In CFA, one specifies all latent concepts as ellipses (circles or ovals) and the measured variables as rectangles. Correlational (or covariance) relationships are depicted by two-headed curved arrows, and the indicator variables typically are denoted by  $X$  (e.g.,  $X_1, X_2, \dots$ ). The relationships between the latent constructs and the respective measured variables (called *factor loadings*, as in EFA) are represented by single-headed arrows from the construct to the measured variable. Finally, each measured indicator variable has an error variance (shown as an  $e$  in our diagram but typically depicted with an ellipse), which is the extent to which the latent factor does not account for the measured variable’s variance. Although published papers often exclude the error-variance terms from the diagrams that depict models, they are vital to understanding CFA.

Figure 10.1 provides a complete specification of a CFA model. The two latent constructs are Supervision and Work Environment. The symbols  $x_1-x_8$  represent the measured indicator variables and  $L_{x1} - L_{x8}$  are the relationships between the latent constructs and the respective measured items (i.e., factor loadings). As you can see, the four items measuring Supervision are linked to that latent construct, as are the other four items to the Work Environment construct. The curved arrow between the two constructs denotes a correlational relationship between them. The symbols  $e_1-e_8$  represent the error variances associated with each measured item. All other possible connectors are omitted. We return to this point shortly.

Each SEM software uses a somewhat unique notation to denote the various elements of a SEM model. Some programs, such as lavaan and Mplus, do not provide for a graphical interface to execute the program. Thus, they



**Figure 10.1**  
Visual Representation (Path Diagram) of a Measurement Theory

## Construct Validity

Psychometric measurement theory means measurement models should be specified as congeneric.

Standardized loading estimates should be .5 or higher, and ideally, .7 or higher, to indicate convergent validity.

AVE should be .5 or greater to suggest adequate convergent validity.

AVE estimates for two factors also should be greater than the square of the correlation between the two factors to provide evidence of discriminant validity.

Construct reliability should be .7 or higher to indicate adequate convergence or internal consistency.

require knowledge of and capability with programming. The coding bears some similarity to SPSS or SAS syntax. Long-time users of LISREL rely on its program coding where each matrix or vector involved in the mathematics is represented by Greek characters (e.g., lambda, gamma, phi, etc.). In this text, we have developed a simplified notation. Chapter 9 describes the notation in greater detail. AMOS and LISREL both include a graphical interface that enables a user to avoid any program coding.

## SEM Stages for Testing Measurement Theory Validation with CFA

A six-stage SEM process was introduced in the previous chapter. Stages 1–4 will be discussed in more detail here because they involve examining measurement theory. Stages 5 and 6, which address the structural theory theoretically linking constructs to each other, will be discussed in the next chapter. Recognizing that valid measurement models from the first stages are absolutely necessary for a valid test of the structural theory, some researchers refer to the measurement model as an *auxiliary theory* [29].

### Stage 1: Defining Individual Constructs

The process begins by defining all constructs that will comprise the measurement model. If the researcher has experience with measuring one of these constructs, then perhaps some scale that was previously used can be applied again. If not, numerous compilations of validated scales are available for a wide range of constructs [6, 28]. When a previously applied scale is not available, the researcher may use psychometric principles and the steps of scale development to produce a measure. The process of designing a new construct measure involves several steps through which the researcher translates the theoretical definition of the construct into a set of specific measured variables that express the meaning of the construct. As such, it is essential that researchers consider not only the operational requirements (e.g., number of items, dimensionality), but also establish the construct validity of the newly designed scale [26]. The process of scale development is simple to follow but implementation can be frustrating and time consuming.

### Stage 2: Developing the Overall Measurement Model

In this step, the researcher must carefully consider how all the individual constructs will come together to form an overall measurement model. Several key issues should be highlighted at this point.

## Defining Individual Constructs

Factors must display adequate content validity, whether they are new scales or scales taken from previous research; even previously established scales should be carefully checked for content validity

Content validity should be of primary importance and judged both qualitatively (e.g., experts opinion) and empirically (e.g., unidimensionality and convergent validity).

Pilot test(s) should be used to purify measures prior to confirmatory testing

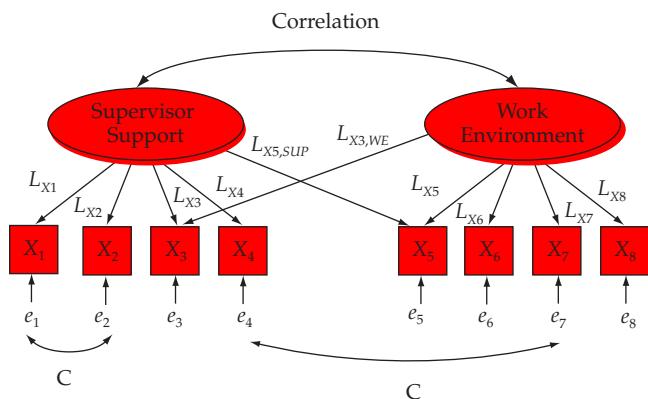
## UNIDIMENSIONALITY

We introduced unidimensionality in Chapter 3. **Unidimensional measures** mean that a set of measured variables (indicators) can be explained by only one underlying construct. Unidimensionality becomes critically important when more than two constructs are involved. In such a situation, each measured variable is hypothesized to relate to only a single construct. All cross-loadings and error variance covariance terms are hypothesized to be zero when unidimensional constructs exist.

Figure 10.1 hypothesizes two unidimensional constructs, because no measured item is determined by more than one construct (has more than one arrow from a latent construct to it). In other words, all cross-loadings are fixed at zero. In addition, no connections are drawn among any error variance terms.

One type of relationship among variables that impacts unidimensionality is when researchers allow a single measured variable to be caused by more than one construct. This situation is represented in the path model by arrows from a single construct pointing toward indicator variables associated with separate constructs. Remember that the researcher is seeking a model that produces a good fit. When one frees another path in a model to be estimated, the additional estimated path cannot reduce fit. That is, the difference between the estimated and observed covariance matrices ( $\Sigma_k - S$ ) is reduced unless the two variables are completely uncorrelated. Therefore, the  $\chi^2$  statistic will almost always be reduced by freeing additional paths.

Figure 10.2 is like the original model with the exception that several additional relationships are hypothesized. In contrast to the original measurement model, this one is not hypothesized to be unidimensional. Additional relationships are hypothesized between  $X_3$ , a measured variable, and the latent construct Work Environment and between  $X_5$  and the latent construct Supervision. These relationships are represented by  $L_{X3,WE}$  and  $L_{X5,SUP}$ , respectively. This means that Supervision indicator variable  $X_3$  and Work Environment indicator variable  $X_5$  are each hypothesized as loading on both of the latent constructs.



**Figure 10.2**  
Measurement Model with Hypothesized Cross-Loadings and Correlated Error Variance

As a rule, even if the addition of these paths leads to a significantly better fit, the researcher should not free (hypothesize) cross-loadings. Why? Because the existence of significant cross-loadings is evidence of a lack of construct validity. When a significant cross-loading is found to exist, any potential improvement in fit is artificial in the sense that it is obtained with the admission of a corresponding lack of construct validity.

Another form of relationships between variables is the covariance among error variance terms of two measured variables. Two types of covariance between error variance terms exist: (1) covariance among error terms of items indicating the same construct, referred to as **within-construct error covariance**, and (2) the covariance between two error terms of items indicating different constructs, referred to as **between-construct error covariance**.

Figure 10.2 also shows covariance (correlation) among some of the error terms. For example, the diagram shows covariance between the error variance terms of measured variables  $X_1$  and  $X_2$  (i.e., within-construct error covariance). It also indicates covariance between two error variance terms of indicator variables loading on different constructs— $e_{X4}$  and  $e_{X7}$ . Here the covariance between  $e_4$  and  $e_7$  is an example of between-construct error covariance between measured indicator variables.

In conventional measurement validation, testing CFA models that include covariances between error terms or cross-loadings violates basic psychometric principles of good measurement. Allowing these paths to be estimated (freeing them) will reduce the  $\chi^2$ , but at the same time seriously question the construct validity of the construct.

Although more focus is typically on the issues raised with between-construct correlations and their impact on the structural model, within-construct error covariance terms also threaten construct validity [15]. Significant between-construct error covariances suggest that the two items associated with these error terms are more highly related to each other than the original measurement model predicts. High within-construct error variance covariance suggests some other factor, not currently measured, likely exists that explains the relationship. Evidence that a significant cross-loading exists also shows a lack of discriminant validity. So again, although these paths can be freed (covariance permitted) and improve the model fit, doing so violates the assumptions of good measurement.

Therefore, we recommend that in standard CFA for psychometric measurement model validation do not free either type of error-covariance path. Relatively rare and specific situations exist where researchers may free these paths as a way of explaining some specific measurement issue not represented by standard factor loadings. For more information on this topic, the reader is referred to other sources [3].

## CONGENERIC MEASUREMENT MODEL

SEM terminology often states that a measurement model is *constrained* by the model hypotheses. The **constraints** refer specifically to the set of fixed **parameter** estimates. One type of common constraint is a measurement model hypothesized to consist of several unidimensional constructs with all cross-loadings constrained to zero. In addition, when a measurement model also hypothesizes no covariance between or within construct error variances, meaning they are all fixed at zero, the measurement model is said to be *congeneric*. **Congeneric measurement models** are considered to be sufficiently constrained to represent good psychometric properties [11]. A congeneric measurement model depicts construct validity and good CFA fit for the model provides evidence of its validity.

## ITEMS PER CONSTRUCT

Researchers are faced with somewhat of a dilemma in deciding how many indicators are needed per construct. On the one hand, researchers prefer many indicators to fully represent a construct and maximize reliability. On the other hand, parsimony encourages researchers to use a smallest number of indicators to adequately represent a construct.

More items (measured variables or indicators) are not necessarily better. Even though more items do produce higher reliability estimates and generalizability [4], more items also require larger sample sizes and can make it difficult to validate unidimensional factors. As researchers increase the number of scale items (indicators) representing a single construct (factor), they may include a subset of items that inadvertently focuses on some specific aspect of a problem and create a subfactor. This problem becomes particularly prevalent when the content of the items has not been carefully screened ahead of time.

In practice, you may find a CFA conducted with only a single item representing some factors. However, good practice dictates a minimum of three items per factor, preferably four, not only to provide minimum coverage of a construct's theoretical domain, but also to provide adequate identification for the construct, as discussed next.

**Items per Construct and Identification** A brief introduction to the concept of statistical identification is provided here to clarify why at least three or four items per construct are recommended. We discuss the issue of statistical identification within an overall SEM model in more detail later. In general, **identification** deals with whether enough information exists to *identify* a solution to a set of structural equations. As we saw earlier in our example, information is provided by the sample covariance matrix. In a CFA or SEM model, one parameter can be estimated for each unique variance or covariance in the observed covariance matrix (i.e., the number of unique terms in a covariance matrix). Thus, the covariance matrix provides the degrees of freedom used to estimate parameters, just as the number of respondents provided degrees of freedom in regression. In fact, the covariance matrix contains full information about the data.

If there are  $p$  measured items, then we can calculate the number of unique variances/covariances as:

$$1/2[p(p + 1)]$$

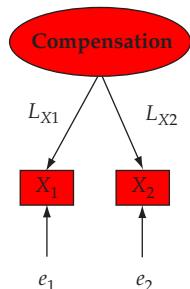
where  $p$  = the number of measured variables.

For example, if there are six items, then there are 21 unique variances/covariances ( $1/2(6 \times 7) = 21$ ). One degree of freedom is then used up for each parameter estimated. Another way to think about identification is that it costs one degree of freedom to free a parameter in any SEM.

Models, and even constructs, can be characterized by their degree of identification, which is defined by the degrees of freedom of a model after all the parameters to be estimated are specified. We will discuss the three levels of identification in terms of construct identification now and then discuss overall model identification at a later point.

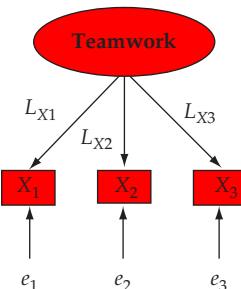
**UNDERIDENTIFIED** An **underidentified** model (also termed **unidentified**) has more parameters to be estimated than unique indicator variable variances and covariances in the observed variance/covariance matrix. For instance, the measurement model for a single construct with only two measured items as shown in Figure 10.3 is *underidentified*. The covariance matrix would be 2 by 2, consisting of one unique covariance and the variances of the two variables. Thus, there are three unique values. A measurement model of this construct would require, however, that two factor loadings ( $L_{X1}$  and  $L_{X2}$ ) and two error variances ( $e_1$  and  $e_2$ ) be estimated (presuming the factor variance is fixed to a value, usually 1). Thus, a unique solution cannot be found.

Underidentified



Four parameters to estimate  
( $L_{X1}, L_{X2}, e_{11}, e_{22}$ )

Just Identified



Six parameters to estimate

Figure 10.3

Underidentified and Just-Identified CFA Models

S	X <sub>1</sub>	X <sub>2</sub>
X <sub>1</sub>	<b>var(1)</b>	cov(1,2)
X <sub>2</sub>	cov(1,2)	<b>var(2)</b>

Three unique terms (shown in bold)

S	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>
X <sub>1</sub>	<b>var(1)</b>	cov(1,2)	cov(1,3)
X <sub>2</sub>	cov(1,2)	<b>var(2)</b>	cov(2,3)
X <sub>3</sub>	cov(1,3)	cov(2,3)	<b>var(3)</b>

Six unique terms (shown in bold)

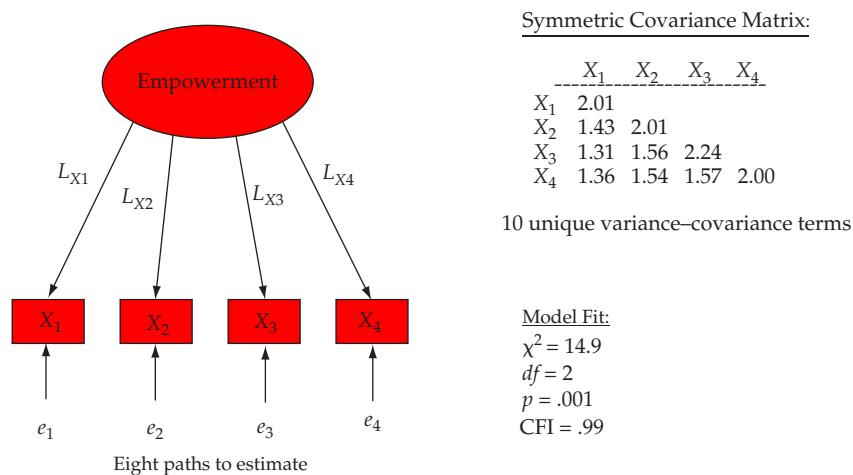
**JUST-IDENTIFIED** Using the same logic, the three-item indicator model in Figure 10.3 is **just-identified**. This means that there are just enough degrees of freedom to estimate all free parameters. All the information is used, which means the CFA analysis will reproduce the sample covariance matrix identically. Because of this, just-identified models have perfect fit by definition. To help understand this, you can use the equation for degrees of freedom given in Chapter 9, and you will see that the resulting degrees of freedom for a three-item factor would be zero:

$$[3(3 + 1)/2] - 6 = 0$$

In SEM terminology, a model with zero degrees of freedom is referred to as *saturated*. The resulting  $\chi^2$  goodness of fit statistic also is zero. Just-identified models do not test a theory, because their fit is determined by the circumstance. As a result, they are not of interest to researchers testing theories. Figure 10.3 illustrates the logic of both underidentified and just-identified single construct models. As noted in the figure, the number of unique variances/covariances is either exceeded by the number of estimated parameters (i.e., underidentified model) or equal to the number of estimated parameters (i.e., just-identified model).

**OVERIDENTIFIED** **Overidentified** models have more unique covariance and variance terms than parameters to be estimated. Thus, for any overidentified measurement model, a solution can be found with positive degrees of freedom and a useful chi-square goodness-of-fit value. A four-item, unidimensional measurement model produces an overidentified construct for which a fit value can be computed [19]. Increasing the number of measured items only strengthens the overidentified condition. Thus, the ideal objective when applying CFA and SEM is to model overidentified constructs within an overidentified measurement model.

Figure 10.4 illustrates an overidentification situation depicting a CFA testing a unidimensional Empowerment construct measured using the following four items: (1) This organization allows me to do things I find personally satisfying, (2) This organization provides an opportunity for me to excel in my job, (3) I am encouraged to make suggestions about how this organization can be more effective, and (4) This organization encourages people to solve problems by themselves. The items ( $X_1-X_4$ ) indicate how much employees perceive they are empowered on their job.



### Measured Items

- $X_1$  = This organization allows me to do things I find personally satisfying.
- $X_2$  = This organization provides an opportunity for me to excel in my job.
- $X_3$  = I am encouraged to make suggestions about how this organization can be more effective.
- $X_4$  = This organization encourages people to solve problems by themselves.

Loading Estimates	Error Variance Estimates
$L_{X1} = 0.78$	$e_1 = 0.39$
$L_{X2} = 0.89$	$e_2 = 0.21$
$L_{X3} = 0.83$	$e_3 = 0.31$
$L_{X4} = 0.87$	$e_4 = 0.24$

**Figure 10.4**  
**Four-Item Single Construct Model Is Overidentified**

A SEM program was used to calculate the results. The sample size was 800 respondents. By counting the number of items in the covariance matrix, we can see that there are a total of 10 unique covariance matrix values ( $\frac{1}{2}(4 \times 5) = 10$ ). We can also count the number of measurement parameters that are free to be estimated. There are four loading estimates ( $L_{x1}, L_{x2}, L_{x3}, L_{x4}$ ) and four error variances ( $e_1, e_2, e_3, e_4$ ), for a total of eight. Thus, the resulting model has two degrees of freedom ( $10 - 8$ ). The overidentified model produces a  $\chi^2 = 14.9$ , with two degrees of freedom.

But consider what would happen if only the first three items were used to measure Empowerment. There would be only six items in the covariance matrix, and exactly six item parameters to be estimated (three loadings and three error variances). The model would be just-identified. Finally, if only two items— $X_1$  and  $X_2$ —were used (as in Figure 10.3), four parameter estimates (two loadings and two error variances) would be needed, but there would be only three items in the covariance matrix. Therefore, the construct would be underidentified.

Note that even though a unidimensional two-item construct CFA is underidentified on its own, if it is integrated into a CFA model with other constructs the overall model may be overidentified. The same identification rules still apply to the overall model as described earlier. However, the extra degrees of freedom from some of the other constructs can provide the necessary degrees of freedom to identify the overall model.

Does this mean the researcher should not worry about the identification of individual constructs, only the overall model? The answer is NO! Even though the overall model is identified, this does not mean the underlying problems with two-item and single-item measures disappear completely when we integrate them into a larger model. Strictly speaking, unidimensionality of constructs with fewer than four item indicators cannot be determined separately [1]. The dimensionality of any construct with only one or two items can only be established relative to other constructs. Constructs measured with one or two items also increase the likelihood of problems with interpretational confounding [10]. One and two-item measures are more likely to lead to statistical problems, including problems with convergence (identifying an appropriate mathematical solution).

In summary, when specifying the number of indicators per latent construct, the following is recommended:

- Use at least four indicators whenever possible.
- Having three indicators per construct is acceptable, particularly when other constructs have more than three.
- Constructs with fewer than three indicators should be avoided or if used, constrained to identify.

**Single-Item Constructs** The exception to using multiple items to represent a construct comes when concepts can be adequately represented with a single item. Some concepts are very simple and lack the nuance and complexity that accompanies most psychological constructs. In other words, if there is little argument over the meaning of a term and that term is distinct and very easily understood, a single item can be sufficient. In marketing, some behavioral outcomes, such as sales, can be captured with a single item. The amount of sales is not really a latent concept. Sales is an observed variable. Other behaviors are directly observable (purchase/no purchase, fail/succeed, etc.). Some would argue that the concept of liking (How much do you like this store?) is a very simple and easily understood concept that does not require multiple items [7]. Although single items can adequately represent some phenomena, operationally, they are difficult to validate. When in doubt, and when multiple items are truly available, using multiple items is the safest approach. We will also address how to operationalize a single item measure in a SEM model in Chapter 11.

## REFLECTIVE VERSUS FORMATIVE MEASUREMENT

The issue of causality is an important consideration in measurement theory. Behavioral researchers typically study latent factors, in which the causal direction is from the factor to the measured variables. At times, however, the causality may be reversed. The contrasting direction of causality leads to different measurement approaches—reflective versus formative measurement models. Until now, our discussion of CFA assumed a reflective measurement theory. A reflective measurement theory is based on the idea that latent constructs cause the measured variables and that the error results in an inability to fully explain these measured variables. Thus, the arrows are

drawn from latent constructs to measured variables. As such, reflective measures are consistent with classical test theory, the psychometric basis for measurement [24].

In our earlier example of employee job satisfaction, the construct Job Search is believed to cause specific measured indicators, such as how often you search for another job, telling friends about looking for another job, and your frequency of looking at job listings on the web.

In contrast, a **formative measurement theory** assumes that measured variables cause, or actually form, a scale. The error in formative measurement models, therefore, is an inability of the measured variables to fully capture the variance in a scale. **A key assumption is that formative factors are not latent.** Instead, they are better viewed as indices where each indicator is a cause of the index score. A typical example would be a social class index [12]. Social class often is viewed as a composite of one's educational level, occupational prestige, and income (or sometimes wealth). Social class does not cause these indicators, as in the reflective case. Rather, each formative indicator is considered a cause of the index. With the measures, the researcher forms a composite score for an observation's social class. In contrast, one could measure social class with a series of items asking a respondent how he/she perceives her social class (e.g., "I fit well into upper-middle class society," etc.). Such an approach would represent reflective measurement. In fact, one could think of ways to represent some concepts like social class with either reflective or formative measurement.

Formative measures have become trendy in recent years, but they should be approached with a degree of caution. Besides issues related to specification in a SEM model (which are quite different than reflective constructs), formative measures have unique qualities in terms of their conceptual and practical meaning. For example, construct validity must be measured differently with formative measures. Thus, the traditional psychometric methods for measurement validation do not apply to formative measures [29]. Formative measurement is distinctly different from psychometric measurement [20]:

- 1 For formative measures, correlation among the indicators is not desirable since each indicator should have an independent effect.
- 2 High reliability or high AVE would provide evidence that a formative scale lacks validity.
- 3 For formative measures, one needs the complete set of variables that may form a factor. But in reflective measurement, the measured items are correlated so each item is viewed as interchangeable and thus representative of the factor. Omission of a variable in formative measurement means the index is improperly formed.

## Developing the Overall Measurement Model

In standard CFA applications testing a measurement theory, within and between error covariance terms should be fixed at zero and not estimated.

In standard CFA applications testing a measurement theory, all measured variables should be free to load only on one construct.

Latent constructs should be indicated by at least three measured variables, and preferably four or more; in other words, latent factors should be statistically identified.

Formative scales are not latent and are not validated in the same manner as conventional reflective scales.

They present greater difficulties with statistical identification but can be included in identified models provided other multiple-item, reflective concepts are included.

## Stage 3: Designing a Study to Produce Empirical Results

The third stage involves designing a study that will produce confirmatory results. In other words, the researcher's measurement theory will be tested. Here, all the standard rules and procedures that produce valid descriptive research apply [17]. If all goes well with the measurement model (CFA), the same sample is used to test the structural model (SEM). We should note that the initial data analysis procedures described in Chapter 2 should first be performed to identify any problems in the data including respondent error. After conducting these preliminary analyses, the researcher must make some key decisions on designing the CFA model.

### MEASUREMENT SCALES IN CFA

CFA models typically contain reflective indicators measured with an ordinal or better measurement scale. Indicators with ordinal responses of at least four categories can be treated as interval, or at least as if the variables are continuous, but should be interpreted accordingly. All the indicators for a construct are not required to be of the same scale type, and different scale values do not have to be normalized (mathematically transformed to a common scale range) prior to using SEM. Normalization, including recoding of oppositely worded items, can make interpreting coefficients and response values easier, so it is generally done prior to estimating the model. Thus, there are few restrictions on the type of data that can be used in SEM analyses, making typical survey research data suitable for a CFA model using SEM.

To illustrate the possibilities of using different scales when applying CFA, let us consider the job satisfaction example introduced earlier. The Job Satisfaction measure consists of four indicator variables. Each indicator variable could be measured with a different number of potential scale values (7 points, 10 points, 100 points, etc.). Although the researcher could transform the number of scale points to a common scale before estimating the model (e.g., all 7 points), it is not necessary to do so. CFA is capable of analyzing multiple indicator variables using a different number of scale points.

### SEM AND SAMPLING

Issues related to sample size and SEM in general were addressed in Chapter 9. But many times CFA requires the use of multiple samples, particularly when the measurement model includes new scales. An initial sample can be examined with EFA and the results used for further purification. Then an additional sample(s) should be drawn to perform the CFA. Even after CFA results are obtained, however, evidence of model stability and generalizability can only come from performing the analysis on additional samples and contexts.

### SPECIFYING THE MODEL

CFA, not EFA, should be used to test the measurement model. As noted earlier, a critical distinction between CFA and EFA is the ability of the researcher to use CFA to perform a test of the measurement theory by specifying theoretically supported correspondence rules between indicators and constructs. EFA provides insight into the structure of the items and may be helpful in proposing the measurement model, but it does not test a theory. As we have discussed, in CFA the researcher specifies (fixes or frees for estimation) the indicators associated with each construct and the correlations between constructs. In addition, the researcher does not specify cross loadings, which fixes the loadings at zero.

One unique feature in specifying the indicators for each construct is the process of "setting the scale" of a latent factor. Because it is unobserved, a latent construct has no metric scale, meaning no range of values. This must be done for both exogenous and endogenous constructs in one of two ways:

- Fix one of the factor loadings on each construct to a specific value (1 is typically used although any value could be inserted), or
- Fix the value of the variance of the construct (again 1 is a good value to use).

Most SEM software today will insert "1s" in appropriate places to set a scale for the latent constructs. For example, the AMOS software automatically fixes, or constrains, one of the factor loading estimates to 1. However, the researcher should get in the habit of making sure that the appropriate values are fixed. As one works with models, variables with fixed values can get changed or deleted, leading to constructs without the proper constraints needed

to set a scale. Likewise, the researcher should check that multiple values are not constrained to 1, as the effect would be to constrain the two to be equal in addition to setting a scale.

“Setting the scale” of a construct by fixing a loading to “1” does not imply perfect association. In earlier examples, we had discussed that there were two estimated parameters for each item—the loading and the error term. If you fix one of the factor loadings on a construct to 1, the error variance can still be estimated by setting it to free rather than fixed. A non-zero estimate for the error variance suggests the relationship is not perfect. If you fix the variance of a construct instead, then a loading and error variance is estimated for each item. So, in total, you still have free parameters equal to two parameters per item.

The researcher may also wish to place additional constraints on a CFA model. For instance, it is sometimes useful to set two or more parameters as equal or to set a specific parameter to a specific value. Information about imposing additional constraints can be found in the documentation for the SEM program of choice. We will discuss some procedures (e.g., testing for measurement invariance in Chapter 12) that employ these types of constraints on the estimated parameters.

## ISSUES IN IDENTIFICATION

Once the measurement model is specified, the researcher must revisit the issues relating to identification. Although in the earlier discussion we were concerned about the identification of constructs, here we are concerned about the overall model. As before, overidentification is the desired state for CFA and SEM models in general. Even though comparing the degrees of freedom to the number of parameters to be estimated seems simple, in practice, establishing the identification of a model can sometimes be frustrating. This complexity is partly due to the fact that a wide variety of problems and data idiosyncrasies can manifest themselves in a lack of convergence or a lack of identification.

During the estimation process, the most likely cause of a model not converging (or completing the estimation process) or producing meaningless results is a problem with statistical identification. As SEM models become more complex, however, ensuring that a model is identified can be problematic [9]. Once the problem is diagnosed, remedies must still be applied.

**Avoiding Identification Problems** Several guidelines help one determine the identification status of a SEM model [26] and assist the researcher in avoiding identification problems. The order and rank conditions for identification are the two most basic rules [8]. Mathematical identification problems can be avoided by following some basic rules in construct specification.

**MEETING THE ORDER AND RANK CONDITIONS** The order and rank conditions are the required mathematical properties for identification. The **order condition** refers to the requirement discussed earlier that the degrees of freedom for a model be greater than zero. That is, the number of unique covariance and variance terms less the number of free parameter estimates must be positive. The degrees of freedom for the overall model are always provided in the program output.

In contrast, the **rank condition** can be difficult to verify, and a detailed discussion would require a working knowledge of linear algebra. In general terms, it is the requirement that each parameter be estimated by a unique relationship (equation). As such, diagnosing a violation of the rank condition can be quite difficult. This is a problem encountered more often in the structural model relationships, particularly when non-recursive, or “feedback,” relationships are specified. But in CFA models, it can still occur in the presence of cross-loading of items and/or in correlated error terms. Although we discouraged the use of either cross-loadings or correlated errors on the basis of construct validity concerns, if they are used the researcher should be aware that identification problems may result.

**Three-Indicator Rule** Given the difficulty in establishing the rank condition, researchers turn to more general guidelines. These guidelines include the **three-indicator rule**. It is satisfied when all factors in a congeneric model have at least three indicators. A two-indicator rule also states that a congeneric factor model with two items per factor can be identified by adding constraints such as specifying both loadings to be equal. Single-item factors

also cause problems with identification. One-indicator constructs should be constrained rather than estimated to minimize problems with a lack of identification. That means fixing the loading and the error-variance term. Any under-identified factor can introduce statistical problems.

**Recognizing Identification Problems** Although identification issues underlie many estimation problems faced in SEM modeling, there are few certain indicators of the existence and source of identification problems. At times, the software programs provide an incomplete, and a potentially unreliable, solution even in the presence of identification issues. Therefore, researchers must consider a wide range of symptoms to assist in recognizing identification problems. Error warnings or messages sometimes suggest that a single parameter is not identified. Although the researcher can attempt to solve the problem by deleting the offending variable, many times this does not address the underlying mathematical cause, and the problem persists. SEM programs provide minimal diagnostic measures for identification problems. Thus, researchers must typically rely on other means of recognizing identification problems by the symptoms described in the following list:

- An inability of the program to invert the information matrix (no solution can be found).
- Wildly unreasonable or impossible estimates, such as negative error variances or very large parameter estimates, including standardized factor loadings and correlations among constructs outside the range of +1.0 to -1.0.
- Models that result in unstable parameter estimates in the presence of small changes suggest problematic data, potentially arising from identification issues. When questions about the identification of any single parameter occur, a second test can be performed. You first estimate a CFA model and obtain the parameter estimate. Next, fix the coefficient to its estimated value and rerun the model. If the overall fit results, including other parameter estimates, vary substantially, identification problems could exist.

As you can see, identification problems can be manifested in SEM results in many different ways. The researcher should never rely only on the software to recognize identification problems, but must also diligently examine the results to ensure that no data problems exist.

**Sources and Remedies of Identification Problems** Does the presence of identification problems mean your model is invalid? Although some models may need respecification, many times identification issues arise from common mistakes in specifying the model and the input data. In the discussion that follows, we will not only discuss the typical types of sources for identification problems, but also offer suggestions for dealing with the problems where possible. Some of the most common issues leading to problems with identification include the following.

**INCORRECT INDICATOR SPECIFICATION** Common mistakes often occur in specifying the measurement model relationships that cause to problems in estimating a valid solution. For instance, the researcher can make mistakes such as (1) not linking an item to any construct, (2) linking an indicator to two or more constructs, (3) assigning an indicator twice in the same model, or (4) not creating and linking an error term for each indicator. Although these mistakes seem obvious, as models get complicated, even experienced users may not avoid such problems on initial runs. Something as simple as listing a variable twice in the SELECT command in LISREL syntax creates linear dependence and makes a valid mathematical solution impossible. Even in programs such as AMOS, overlooking a loading between indicator and construct or an error term is quite easy in a complicated model. We encourage the researcher to carefully examine the model specification if identification problems are indicated and to be aware of issues such as how many degrees of freedom a model should produce as signs that indicate a specification that is true and free of setup errors.

**"SETTING THE SCALE" OF A CONSTRUCT** A second common mistake that creates identification problems is not "setting the scale" of each construct. As discussed earlier, each construct must have one value specified (either a loading of an indicator or the construct variance). Failure to do this for any construct will create a problem, and the model will not estimate. This type of problem occurs in initially specifying the model, but also in model respecification, when indicators may be eliminated from the model. If an indicator with the fixed loading is deleted for some reason, then

another indicator must be fixed. Even in AMOS, which will automatically fix one loading to 1.0 in the initial specification, it does not automatically fix another if that indicator is eliminated from the model at a later stage.

**TOO FEW DEGREES OF FREEDOM** This problem is likely accompanied by a violation of the three-indicator rule. Small sample size increases the likelihood of problems in this situation. The simplest solution is to avoid this situation by including enough measures to avoid violating these rules. If this is not possible, the researcher can try to add some constraints that will free up degrees of freedom [18]. One possible solution is imposing **tau equivalence** assumptions, which require all factor loadings on a particular factor to be equal. Tau equivalence can be done for one or more factors. A second solution is to fix the error variances to a known or specified value. Third, the correlations between constructs can be fixed if some theoretical value can be assigned. The researcher should remember that each of these solutions, however, has implications for the construct validity of the constructs involved and should be undertaken with great care and an understanding of their impact on the constructs.

Identification problems must be solved before the results can be trusted. Although careful model specification using the guidelines discussed earlier can help avoid many of these problems, researchers must always be vigilant in scrutinizing the results to recognize identification problems wherever they occur.

## PROBLEMS IN ESTIMATION

Like other multivariate techniques, the SEM user needs to be wary of results that seem implausible. Just like in regression analysis, for example, one might encounter a standardized coefficient suggesting greater than 100 percent explained variance. We will discuss the two most common types of estimation problems as well as potential causes and solutions.

**Illogical Standardized Parameters** The most basic estimation problem with SEM results is when correlation estimates (i.e., standardized estimates) between constructs exceed  $|1.0|$  or standardized path coefficients exceed  $|1.0|$ . These estimates are theoretically impossible, and many times identification problems are the cause. But they also may occur from data issues (e.g., highly correlated constructs) or even poorly specified constructs (e.g., extremely low reliability or other issues in construct validity).

**Heywood Cases** A SEM solution that produces an error variance estimate of less than zero (a negative error variance) is termed a **Heywood case**. Such a result is logically impossible because it implies a less than 0 percent error variance in an item, and by inference it implies that more than 100 percent of the variance in an item or a construct is explained. Heywood cases are particularly problematic in CFA models with small samples or when the at least three-indicator rule is not followed [22]. Models with sample size of at least 100 that adhere to the three-indicator rule rarely produce Heywood cases. Even when a Heywood case(s) is present, the SEM program may produce a solution; that solution may be one in which the model did not fully converge. This is usually accompanied by a warning or error message indicating that an error variance estimate is not identified and cautioning that the solution may not be reliable.

Several options are possible when Heywood cases arise. The first solution should be to ensure construct validity. This may involve the elimination of an offending item, but the researcher may be limited if this creates a violation of the three-indicator rule. An alternative is to try and add more items if possible or assume tau equivalence (all loadings in that construct are equal). Each of these is preferable to the “last resort” solution, which is to fix the offending estimate to a very small value, such as .005 [13]. Although this value may identify the parameter, it can lead to lower fit, because the value is not likely to be the true sample value. It also means that the underlying cause is not remedied in the model specification, but must be addressed in an “ad hoc” fashion.

## Stage 4: Assessing Measurement Model Validity

Once the measurement model is correctly specified, a corresponding SEM model is estimated to provide empirical assessment of the accuracy of the measurement theory. The results enable us to compare the theory against reality as represented by the sample data. In other words, we see how well the measurement theory fits the data.

## Designing a Study to Provide Empirical Results

The scale of a latent construct can be set by either:

Fixing one loading and setting its value to 1, or

Fixing the construct variance and setting its value to 1.

Congeneric, reflective measurement models, in which all constructs have at least three item indicators, are statistically identified both within constructs and for the model overall.

The researcher should check for errors in the specification of the measurement model when identification problems are indicated.

Models with sufficient sample size that adhere to the three indicator rule generally do not produce Heywood cases.

When factors are included with less than three indicators, the values of loadings may need to be constrained to produce a mathematically identified and stable solution.

## ASSESSING FIT

Fit was discussed in detail in Chapter 9. Recall that reality is represented by a covariance matrix of measured items ( $S$ ), and the theory is represented by the proposed measurement model structure. Equations are implied by the model, as discussed earlier in this chapter and in Chapter 9. The equations enable us to estimate reality by computing an estimated covariance matrix based on our theory ( $\Sigma_k$ ). Fit compares the two covariance matrices.

Guidelines for good fit provided in Chapter 9 apply. Here the researcher attempts to examine all aspects of construct validity through various empirical measures. The result is that CFA enables us to test the validity of a theoretical measurement model. CFA is quite different from EFA, which explores data to identify potential constructs. Many researchers conduct EFA on one or more separate samples before reaching the point of trying to confirm a model. EFA is an appropriate tool for identifying factors among multiple variables. As such, EFA results can be useful in helping to develop theory that will lead to a proposed measurement model. After a researcher builds confidence in a measurement model, CFA enters the picture.

## PATH ESTIMATES

One of the most fundamental assessments of construct validity involves the measurement relationships between items and constructs (i.e., the *path estimates* linking constructs to indicator variables). When testing a measurement model, the researcher should expect to find relatively high loadings. After all, once CFA is used a good conceptual understanding of the constructs and its items should exist. This knowledge, along with preliminary empirical results from exploratory studies, should provide these expectations.

**Size of Path Estimates and Statistical Significance** In Chapter 9, we provided rules of thumb suggesting that standardized indicator loadings should be at least .5 and ideally .7 or higher. Loadings of this size or larger confirm that the indicators are strongly related to their associated constructs and are one indication of construct validity. Note that these guidelines apply to the standardized loadings estimates, which remove effects due to the scale of the measures, much like the differences between correlation and covariance. Thus, the researcher must be certain that they are included in the output. The default output often displays the unstandardized maximum likelihood estimates, which are more difficult to interpret with respect to these guidelines.

Researchers should also assess the statistical significance of each estimated (free) coefficient. Nonsignificant estimates suggest that an item should be dropped. Conversely, a significant loading alone does not indicate an item

is performing adequately. A loading can be significant at impressive levels (i.e.,  $p < .01$ ) but still be considerably below  $.5$ . Low loadings suggest that a variable is a candidate for deletion from the model.

SEM output typically displays the **squared multiple correlations** for each measured variable. In a CFA model, this value represents the extent to which a measured variable's variance is explained by a latent factor. From a measurement perspective, it represents how well an item measures a construct. Squared multiple correlations are sometimes referred to as *item reliability, communality, or variance extracted* (more discussion in following section on construct validity). We do not provide specific rules for interpreting these values here, because in a congeneric measurement model they are a function of the loading estimates. Recall that a congeneric model is one in which no measured variable loads on more than one construct.

**Identifying Problems** Loadings also should be examined for offending estimates as indications of overall problems. One often overlooked task is to make sure the loadings make sense. For instance, items with the same valence (e.g., positive or negative wording) should produce the same sign. If an attitude scale consists of responses to four items—good, likeable, unfavorable, bad—then two items should carry positive loadings and two should carry negative loadings (unless they have previously been recoded). If the signs of the loadings are not opposite, the researcher should not have confidence in the results.

As discussed earlier, standardized loadings above 1.0 or below  $-1.0$  are out of the feasible range and are an important indicator of a problem with the model. The reader can refer to the discussion of problems in parameter estimation to examine what this situation may mean for the model overall. It is important to point out that the problem may not reside solely in the variable with the out-of-range loading. So simply dropping this item may not provide the best solution. To summarize, the loading estimates can suggest either dropping an individual item or that some offending estimate indicates a larger overall problem.

## CONSTRUCT VALIDITY

Recall that in Chapter 3 *validity* was defined as the extent to which research is accurate, and the discussion centered on validating summated scales. CFA eliminates the need to summate scales, because the SEM programs compute latent construct scores for each respondent. This process enables relationships between constructs to be corrected for the amount of error variance that exists in the construct measures.

One of the primary objectives of CFA/SEM is to assess the construct validity of a proposed measurement theory. Construct validity is the extent to which a sets of measured items accurately reflect the theoretical latent constructs they are designed to measure. Thus, construct validity deals with the accuracy of measurement. Evidence of construct validity provides confidence that item measures taken from a sample represent the actual true score that exists in the population. Poor fit to a measurement model with congeneric constraints would be *prima facie* evidence of a lack of construct validity.

Once good fit is established (fit validity), four additional components of construct validity are evaluated. Each component was introduced in Chapter 3. Here, we expand on those ideas and discuss them in terms more appropriate for CFA. Note that the comments in the following sections are associated with reflective measurement models, and that formatively measured constructs are evaluated using different criteria.

**Convergent Validity** The items that are indicators of a specific construct should converge or share a high proportion of variance in common, known as **convergent validity**. Several ways are available to estimate the relative amount of convergent validity among item measures.

**Factor Loadings** The size of the factor loading is one important consideration. In the case of high convergent validity, high loadings on a factor would indicate that they converge on a common point, the latent construct. At a minimum, all factor loadings should be statistically significant [1]. Because a statistically significant loading could still be very weak in strength, particularly with large samples, a good rule of thumb is that standardized loading

estimates should be .5 or higher, and ideally .7 or higher. In most cases, researchers should interpret standardized parameter estimates, because they are constrained to range between -1.0 and +1.0.

The rationale behind the rule of thumb for loading size can be understood in the context of an item's **communality** (see Chapter 3). The square of a standardized factor loading represents how much variation in an item is explained by the latent factor and is termed the **variance extracted** of the item. Thus, a loading of .71 squared yields a communality of .5. In short, the factor is explaining half the variation in the item with the other half being error variance. As loadings fall below .7, they still are usually statistically significant, but more of the variance in the measure is error variance than explained variance. Thus, factor loadings illustrate another case where statistical significance is not particularly meaningful.

**AVERAGE VARIANCE EXTRACTED** With CFA, the **average variance extracted (AVE)** is calculated as the mean variance extracted for the items loading on a construct and is a summary indicator of convergence [14]. This value can be calculated using standardized loadings:

$$\text{AVE} = \frac{\sum_{i=1}^n L_i^2}{n}$$

$L_i$  represents the completely standardized factor loading for the  $i$ th measured variable and  $n$  is the number of item indicators for a construct. In words, AVE is computed as the total of all squared standardized factor loadings (squared multiple correlations) divided by the number of items. (SEM programs offer several different types of standardization. Where we use the term *standardized*, we refer to completely standardized estimates unless otherwise noted.) In other words, it is the average squared factor loading or average communality. Using this same logic, an AVE of .5 or higher is a good rule of thumb suggesting adequate convergence. An AVE of less than .5 indicates that, on average, more error remains in the items than variance held in common with the latent factor upon which they load. An AVE measure should be computed for each latent construct in a measurement model. In Figure 10.1, an AVE estimate is needed for both the Supervision and Work Environment constructs.

**RELIABILITY** Reliability is also an indicator of convergent validity. Considerable debate centers around which of several alternative reliability estimates is best [4]. Coefficient alpha remains a commonly applied estimate, although it may underestimate reliability. Different reliability coefficients do not produce dramatically different reliability estimates, but a slightly different **construct reliability (CR)** value is often used in conjunction with SEM models. It is computed from the squared sum of factor loadings ( $L_i$ ) for each construct and the sum of the error variance terms for a construct ( $e_i$ ) as:

$$\text{CR} = \frac{\left( \sum_{i=1}^n L_i \right)^2}{\left( \sum_{i=1}^n L_i \right)^2 + \left( \sum_{i=1}^n e_i \right)}$$

High construct reliability indicates that internal consistency exists ( $\geq .7$ ), meaning that the measures all consistently represent the same latent construct.

**Discriminant Validity** Discriminant validity is the extent to which a construct or variable is truly distinct from other constructs or variables. Thus, high discriminant validity provides evidence that a construct is unique and captures some phenomena other measures do not. CFA provides two common ways of assessing discriminant validity.

First, the correlation between any two constructs can be specified (fixed) as equal to one. In essence, unity correlation means the items making up two constructs could just as well make up only one construct. If the fit of the two-construct model is significantly different from that of the one-construct model, then discriminant validity is supported [1, 5]. The researcher could then test a model with the specification of all items for both constructs on a single factor and compare its fit to the fit of the original two-factor model. If the model fits were significantly

different, this would suggest that the eight items better represent two separate constructs. In practice, however, this test does not provide strict evidence of discriminant validity, because high correlations, sometimes as high as .9, can still produce significant differences in fit between the two models.

A more rigorous test is to compare the average variance-extracted values for any two constructs with the square of the correlation estimate between these two constructs [14]. The variance-extracted estimates should be greater than the squared correlation estimate. The logic here is based on the idea that a latent construct should explain more of the variance in its item measures than it shares with another construct. Passing this test provides good evidence of discriminant validity.

In addition to distinctiveness between constructs, discriminant validity also means that individual measured items should represent only one latent construct. The presence of cross-loadings indicates a discriminant validity problem. If high cross-loadings do indeed exist, and they are not represented by the measurement model, the CFA fit should not be good.

**Nomological Validity and Face Validity** Constructs also should have face validity and nomological validity. The processes for testing these properties are the same whether using CFA or EFA, so the reader is referred to Chapter 3 for more detailed clarification. **Face validity** must be established *prior* to any theoretical testing when using CFA. Without an understanding of every item's content or meaning, it is impossible to express and correctly specify a measurement theory. Thus, in a very real way, face validity is the most important validity test. **Nomological validity** is then tested by examining whether the correlations among the constructs in a measurement theory make sense. The matrix of construct correlations can be useful in this assessment.

Researchers often test a measurement theory using constructs measured by multi-item scales developed in previous research. For instance, if HBAT wished to measure customer satisfaction with their services, it could do so by evaluating and selecting one of several customer satisfaction scales in the marketing literature. Handbooks exist in many social science disciplines that catalog multi-item scales [6, 28]. Similarly, if HBAT wanted to examine the relationship between cognitive dissonance and customer satisfaction, a previously applied cognitive dissonance scale could be used.

Any time previously used scales are in the same model, even if they have been applied successfully with adequate reliability and validity in other research, the psychometrician should pay careful attention that the item content of the scales does not overlap. In other words, when using borrowed scales, the researcher should still check for face validity. It is quite possible that when two borrowed scales are used together in a single measurement model, face validity issues become apparent that were not seen when the scales were used individually.

## MODEL DIAGNOSTICS

CFA's goal is to obtain an answer as to whether a given theoretical measurement model is valid. But the process of testing using CFA provides additional diagnostic information that may suggest modifications for either addressing unresolved problems or improving the model's validity.

When the researcher respecifies a model after the initial test, he/she is admitting to flaws in the original model. If the modifications are minor, then the theoretical integrity of a measurement model may not be severely damaged, and the research can proceed using the prescribed model and data after making suggested changes. If the modifications are more than minor, then the researcher must be willing to modify the measurement theory, which will result in a new measurement model and potentially require a new sample. Given the strong theoretical basis for CFA, the researcher should avoid making changes based solely on empirical criteria, such as the diagnostics provided by CFA. Moreover, other concerns should be considered before making any change, including the theoretical integrity of the individual constructs and overall measurement model and the assumptions and guidelines that go along with good practice, much of which have already been discussed.

What diagnostic cues are provided when using CFA? They include fit indices such as those discussed and analyses of residuals as well as some specific diagnostic information provided in most CFA output. Many diagnostic cues are provided, and we focus here on those that are both useful and easy to apply. Some areas that can be used to identify problems with measures are standardized residuals, modification indices, and specification search.

**Standardized Residuals** The standard output produced by most SEM programs includes residuals. **Residuals** are the individual differences between observed covariance terms and the fitted (estimated) covariance terms. The better the fit, the smaller are the residuals. Thus, a residual term is associated with every unique value in the observed covariance matrix. The **standardized residuals** are simply the raw residuals divided by the standard error of the residual. They are not dependent on the actual measurement scale range, which makes them useful in diagnosing problems with a measurement model.

Residuals can be either positive or negative, depending on whether the estimated covariance is under or over the corresponding observed covariance. Researchers can use these values to identify item pairs for which the specified measurement model does not accurately predict the observed covariance between those two items. Typically, standardized residuals less than |2.5| do not suggest a problem in a model of moderate or high complexity. **Conversely, residuals greater than |4.0| can raise a red flag and indicate a potentially unacceptable degree of error. We may accept and even expect one or two large residuals in complex models (>30 indicators).** More than any individual residual though, what is of concern is a consistent pattern of large standardized residuals associated either with a single variable and a number of other variables or residuals for several of the variables within a construct. Either occurrence suggests problems. The most likely, but not automatic, response is dropping one of the items associated with a residual greater than |4.0|. Standardized residuals between |2.5| and |4.0| deserve some attention, but may not suggest any changes to the model if no other problems are associated with those two items.

**Modification Indices** Typical SEM output also provides modification indices. A **modification index** is calculated for every possible relationship that is *not* estimated in a model. For example, in Figure 10.1 variable  $X_1$  has a loading on the Supervision construct, but not on the Work Environment construct. That is, the loading of  $X_1$  on the Work Environment construct is fixed at zero. There would then be a modification index value for the possible loading of  $X_1$  on the other construct. The modification index value would show how much the overall model  $\chi^2$  value would be reduced by also estimating a loading for  $X_1$  to the Work Environment construct. Likewise, there would be modification indices calculated for the remainder of the items that loaded on Supervision and not Work Environment, as well as vice versa (those items that loaded on Work Environment and not on Supervision).

**Modification indices of approximately 4.0 or greater** suggest that the fit could be improved significantly by freeing the corresponding path to be estimated. But making model changes based solely on modification indices is not recommended. Doing so would be inconsistent with the theoretical basis of CFA and SEM in general. Modifications do provide important diagnostic information about the potential cross-loadings that could exist if specified. As such, they assist the researcher in assessing the extent of model misspecification without estimating a large number of new models. This is an important tool for identifying problematic indicator variables if they exhibit the potential for cross-loadings. Modification indices are estimated for all non-estimated parameters, so they are also generally provided for diagnosing error term correlations and also correlational relationships between constructs that may not be initially specified in the CFA model. Researchers should consult other residual diagnostics for a change suggested by a modification index and then take appropriate action, if justified by theory.

**Specification Searches** A **specification search** is an empirical trial-and-error approach that uses model diagnostics to suggest changes in the model. In fact, when we make changes based on any diagnostic indicator, we are performing a type of specification search [27]. SEM programs such as AMOS and LISREL can perform specification searches automatically. These searches identify the set of “new” relationships that best improve the overall model fit. This process is based on freeing fixed (non-estimated) relationships with the largest modification index. Specification searches are fairly easy to implement.

Although it may be tempting to rely largely on specification searches as a way of finding a model with a good fit, this approach is not recommended [21]. The biggest problem is its inconsistency with the intended purpose and use of procedures such as CFA. Namely, CFA tests theory and is generally applied in a confirmatory approach, not

as an exploratory tool. Second, the results for one parameter depend on the results of estimating other parameters, which makes it difficult to be certain that the true problem with a model is isolated in the variables suggested by a modification index. Third, empirical research using simulated data has shown that mechanical specification searches are unreliable in identifying a true model and thus can provide misleading results. Therefore, CFA specification searches should involve identifying only a small number of major problems. A researcher in exploratory mode can use specification searches to identify a plausible measurement theory. But new construct structures suggested by specification searches must be confirmed using a new dataset.

**Caveats in Model Respecification** What types of modifications are more than minor? The answer to this question is not simple or clear-cut. If model diagnostics indicate the existence of some new factor not suggested by the original measurement theory, verifying such a change would require a new dataset. When more than 20 percent of the measured variables are dropped or changed with respect to the factor they indicate, then a new dataset should be used for further verification. In contrast, dropping one or two items from a large battery of items is less consequential, and the confirmatory test may not be jeopardized.

Because CFA tests a measurement theory, changes to the model should be made only after careful consideration. The most common change would be the deletion of an item that does not perform well with respect to model integrity, model fit, or construct validity. At times, however, an item may be retained even if diagnostic information suggests that it is problematic. For instance, consider an item with high content validity (e.g., “I was very satisfied,” in a satisfaction scale) within an overall CFA model with good overall fit and strong evidence for construct validity. Dropping may improve empirical performance at the expense of potentially changing the meaning of the scale. In sum, a poorly performing item may be retained at times to satisfy statistical identification requirements, to meet the minimal number of items per factor, or based on face validity considerations. In the end, however, theory should always be prominently considered in making model modifications.

## Assessing the Measurement Model Validity

Loading estimates can be statistically significant but still be too low to qualify as a good item (standardized loadings below  $|.5|$ ); in CFA, items with low loadings become candidates for deletion.

Completely standardized loadings above 1.0 or below  $-1.0$  are out of the possible range and can be an important indicator of some problem with the data.

Typically, standardized residuals less than  $|2.5|$  do not suggest a problem.

Standardized residuals greater than  $|4.0|$  indicate a potentially unacceptable degree of error that may call for the deletion of an offending item.

Standardized residuals between  $|2.5|$  and  $|4.0|$  deserve some attention, but may not suggest any changes to the model if no other problems are associated with those two items.

Patterns of relatively high residuals suggest that something is missing from the model, such as another factor.

The researcher should use the modification indices as a guideline for model improvements of those relationships that can theoretically be justified.

Specification searches based on purely empirical grounds are discouraged because they are inconsistent with the theoretical basis of CFA and SEM.

CFA is not an exploratory tool.

CFA results suggesting more than minor modification should be re-evaluated with a new dataset (e.g., if more than 20 percent of the measured variables are deleted, then the modifications cannot be considered minor).

**Table 10.1**

Model Fit Measures, Loadings, Standardized Residuals, and Modification Indices in CFA

Overall Model Fit Measures				Fit Indices	
$\chi^2 = 68.0$ with 26 degrees of freedom ( $p = .000013$ )					
CFI = .99					
RMSEA = .04					
Standardized Loadings (AMOS = Regression Weights)					
		EMPOWERMENT		JOB SATISFACTION	
$X_1$	0.78			—	New Construct Job Satisfaction has five indicator variables.
$X_2$	0.89			—	
$X_3$	0.83			—	
$X_4$	0.87			—	
$X_5$	—			0.58	
$X_6$	—			0.71	
$X_7$	—			0.69	
$X_8$	—			0.52	
$X_9$	—			0.46	
Largest Negative Standardized Residuals			Largest Positive Standardized Residuals		
RESIDUAL	FOR	$X_3$	AND	$X_1$	-3.12
RESIDUAL	FOR	$X_4$	AND	$X_2$	-3.04
RESIDUAL	FOR	$X_6$	AND	$X_4$	-2.70
RESIDUAL	FOR	$X_9$	AND	$X_5$	-3.76
Modification Indices for Cross-Loading Estimates					
		EMPOWERMENT		JOB SATISFACTION	
$X_1$	—			0.00	Modification indices for each cross-loading not estimated above.
$X_2$	—			5.04	
$X_3$	—			0.01	
$X_4$	—			5.29	
$X_5$	4.09			—	
$X_6$	2.72			—	
$X_7$	0.04			—	
$X_8$	2.30			—	
$X_9$	2.06			—	
Modification Indices for Error Term Estimates					
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
$X_1$	—				
$X_2$	9.30	—			
$X_3$	9.72	0.90	—		
$X_4$	0.01	9.26	15.17	—	
$X_5$	10.04	2.40	2.62	1.86	—
$X_6$	0.28	0.00	1.40	2.73	0.86
$X_7$	2.04	0.09	0.17	0.28	0.16
$X_8$	0.00	0.84	3.82	0.06	6.62
$X_9$	0.78	0.08	2.14	0.00	14.15
					4.98

## SUMMARY EXAMPLE

We will now illustrate not only how to assess the overall model fit of a CFA model, but also the use of several diagnostic measures. The measures will include standardized loadings, standardized residuals and modification indices. Table 10.1 shows selected output from testing a CFA model that extends the model shown in Figure 10.4. Another construct, Job Satisfaction (JS), has been added to the model. The two constructs represent employee Empowerment and Job Satisfaction. The model fit as indicated by the CFI (.99) and RMSEA (.04) appears good. The model  $\chi^2$  is 68 with 26 df, which is significant as to be expected given the large sample ( $N = 800$ ). The fit compares favorably to the guidelines in the previous chapter.

All loadings estimates are statistically significant. Given the sample size of 800, we can expect that all factor loading estimates are statistically significant. But, statistical significance does not make the case for convergent validity. Three Job Satisfaction loading estimates fall below the .7 cut-off and one falls below the less conservative .5 cut-off ( $X_9$ ). Thus,  $X_9$  becomes a prime candidate for deletion. The loadings for  $X_5$  and  $X_8$  are lower than preferred, but unless some other evidence suggests they are problematic, they will likely be retained to support content validity. For all practical purposes,  $X_7$ 's loading is adequate given it is only .01 below .70. Rounded to one significant digit, it meets the .7 cut-off.

The next step is to calculate the construct reliabilities of both constructs. The reliability for Empowerment is .91, whereas the reliability of Job Satisfaction is .73. Both exceed the suggested threshold of .70. In terms of discriminant validity, we need to compare the AVEs for each construct with the square of the estimated correlation between these constructs. The AVEs are .71 and .36 for Empowerment and Job Satisfaction, respectively. Note that Job Satisfaction's AVE falls below the suggested level of .50, another indicator of perhaps improvement of the construct by eliminating an item. The correlation between constructs is .48, and its squared value is .23. Thus, discriminant validity of the two constructs is supported because the AVE of both constructs is greater than the squared correlation between them.

In terms of other diagnostic measures, we next examine the standardized residuals. In the table, all standardized residuals greater than |2.5| are shown. Two residuals approach but do not exceed 4.0. The largest, between  $X_3$  and  $X_4$  (3.90), suggests that the covariance estimate between these indicator variables could be more accurate. In this case, no change will be made based on the residual between  $X_3$  and  $X_4$ , because the fit remains good despite the high residual. Deleting either variable would leave fewer than four indicator variables for this construct. Also, freeing the parameter representing the error covariance between these two would be inconsistent with the congeneric properties of the measurement model. Thus, it appears "we can live with" this somewhat high residual for now. The second highest residual is between  $X_5$  and  $X_9$  (-3.76). It provides further evidence (in addition to the low standardized loading) that  $X_9$  may need to be dropped.

Finally, we examine the modification index associated with each of the loadings of the indicators. As we see, the values represent the cross-loadings of items if they were estimated. Here the information is consistent with that obtained from the residuals, leading to much the same conclusion. First, none of the values among loadings are high enough to indicate that a cross-loading is required. In looking at the modification indices for the error terms, we see that the value for the covariance between  $X_3$  and  $X_4$  error terms is 15.17. Although we do not recommend adding this relationship to the model, it does indicate a high degree of covariance between these two items that is not captured by the construct. But given the high loading estimates for each, no change is made. Looking further,  $X_9$  and  $X_5$  yields a high value of 14.15, just one more indication of a poorly performing item.

Given that there a high standardized residual associated with  $X_9$  (-3.76), there is a high modification index between  $X_9$  and  $X_5$  (14.15), and its loading is below .5,  $X_9$  thus becomes a candidate for deletion. The final decision should be made based not only on model fit improvement, but the extent to which deleting  $X_9$  would diminish the content validity of the construct. Overall, the results are not supportive of the Job Satisfaction scale.

## CFA Illustration

We now illustrate CFA using HBAT as an example. In this section, we apply the first four stages of the six-stage process to a problem faced by management. We begin by briefly introducing the context for this new HBAT study.

HBAT employs thousands of workers in different operations around the world. Like many firms, one of its biggest management problems is attracting and keeping productive employees. The cost to replace and retrain employees is high. Yet the average new person hired works for HBAT less than three years. In most jobs, the first year is not productive, meaning the employee is not contributing as much as the costs associated with employing him/her. After the first year, most employees become productive. HBAT management would like to understand the factors that contribute to employee retention. A better understanding can be gained by learning how to measure the key constructs. Thus, HBAT is interested in developing and testing a measurement model made up of constructs that affect employees' attitudes and behaviors about remaining with HBAT.

## STAGE 1: DEFINING INDIVIDUAL CONSTRUCTS

With the general research question defined, the researcher now selects the specific constructs that represent the theoretical framework to be tested and will be included in the analysis. The indicators used to operationalize the constructs may come from prior research or be developed specifically for this project.

HBAT initiated a research project to study the employee turnover problem. Preliminary research discovered that a large number of employees are exploring job options with the intention of leaving HBAT should an acceptable offer be obtained from another firm. To conduct the study, HBAT hired consultants with a working knowledge of the organizational behavior theory dealing with employee retention. Based on published literature and some preliminary interviews with employees, a study was designed focusing on five key constructs. The consulting team and HBAT management also agreed on construct definitions based on how they have been used in the past. The five constructs along with a working definition are as follows:

- *Job Satisfaction (JS)*. Reactions resulting from an appraisal of one's job situation.
- *Organizational Commitment (OC)*. The extent to which an employee identifies and feels part of HBAT.
- *Staying Intentions (SI)*. The extent to which an employee intends to continue working for HBAT and is not participating in activities that make quitting more likely.
- *Environmental Perceptions (EP)*. Beliefs an employee has about day-to-day, physical working conditions.
- *Attitudes Toward Coworkers (AC)*. Attitudes an employee has toward the coworkers he/she interacts with on a regular basis.

The consultants proposed a set of multiple-item reflective scales to measure each construct. Face validity appears evident, and the conceptual definitions match well with the item wordings. Additionally, a pretest was performed in which three independent judges matched items with the construct names. No judge had difficulty matching items to constructs, providing further confidence the scales contain face validity. Having established face validity, HBAT proceeded to finalize the scales. Scale purification based on item-total correlations and EFA results (as in Chapter 3) from a pretest involving 100 HBAT employees resulted in the measures shown in Table 10.2. The job satisfaction scale contains multiple measures each assessing the degree of satisfaction felt by respondents with a different type of scale. The complete questionnaire is available online.

## STAGE 2: DEVELOPING THE OVERALL MEASUREMENT MODEL

With the constructs specified, the researcher next must specify the measurement model to be tested. In doing so, not only are relationships among constructs defined, but also the nature of each construct (reflective versus formative) is specified.

A visual diagram depicting the measurement model is shown in Figure 10.5. The model displays 21 measured indicator variables and five latent constructs. Without a reason to think the constructs are independent, all constructs can correlate with all other constructs. All measured items can load on only one construct each. Moreover, the error variance terms are not shown in the illustration, but the reader needs to be fully aware that they exist and are not allowed to relate to any other measured variable or error variance. In this way, the proposed measurement model is congeneric. Four constructs are indicated by four measured

**Table 10.2** Observed Indicators Used in HBAT CFA of Employee Behavior

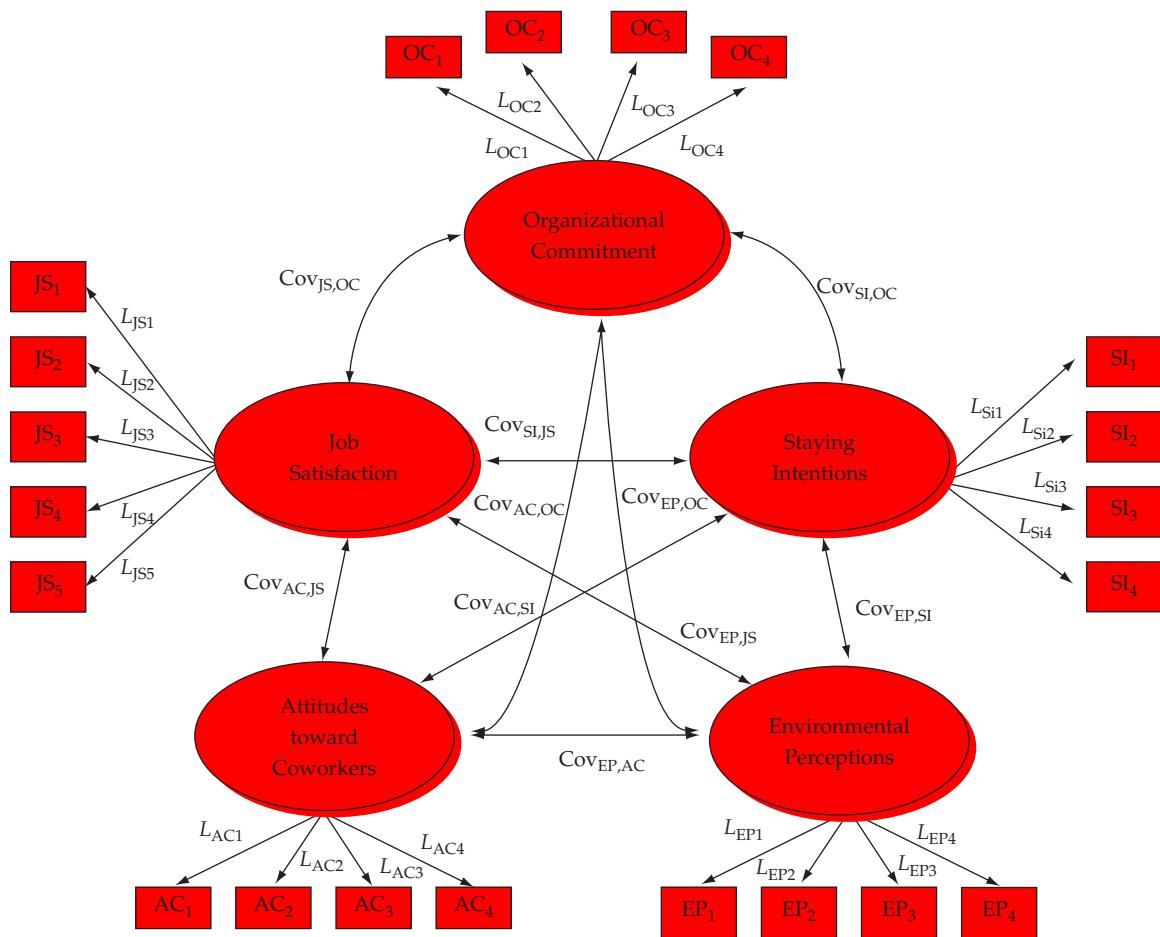
Item	Scale Type	Description	Construct
JS <sub>1</sub>	0–10 Likert Disagree–Agree	All things considered, I feel very satisfied when I think about my job.	JS
OC <sub>1</sub>	0–10 Likert Disagree–Agree	My work at HBAT gives me a sense of accomplishment.	OC
OC <sub>2</sub>	0–10 Likert Disagree–Agree	I am willing to put in a great deal of effort beyond that normally expected to help HBAT be successful.	OC
EP <sub>1</sub>	0–10 Likert Disagree–Agree	I am comfortable with my physical work environment at HBAT.	EP
OC <sub>3</sub>	0–10 Likert Disagree–Agree	I have a sense of loyalty to HBAT.	OC
OC <sub>4</sub>	0–10 Likert Disagree–Agree	I am proud to tell others that I work for HBAT.	OC
EP <sub>2</sub>	0–10 Likert Disagree–Agree	The place I work in is designed to help me do my job better.	EP
EP <sub>3</sub>	0–10 Likert Disagree–Agree	There are few obstacles to make me less productive in my workplace.	EP
AC <sub>1</sub>	5-point Likert	How happy are you with the work of your coworkers? ____ Not happy ____ Somewhat happy ____ Happy ____ Very happy ____ Extremely happy	AC
EP <sub>4</sub>	7-point Semantic Differential	What term best describes your work environment at HBAT? Too hectic _____ Very soothing	EP
JS <sub>2</sub>	7-point Semantic Differential	When you think of your job, how satisfied do you feel? Not at all satisfied _____ Very much satisfied	JS
JS <sub>3</sub>	7-point Semantic Differential	How satisfied are you with your current job at HBAT? Very unsatisfied _____ Very satisfied	JS
AC <sub>2</sub>	7-point Semantic Differential	How do you feel about your coworkers? Very unfavorable _____ Very favorable	AC
SI <sub>1</sub>	5-point Likert Disagree–Agree	I am not actively searching for another job. Strongly disagree _____ Strongly agree	SI
JS <sub>4</sub>	5-point Likert	How satisfied are you with HBAT as an employer? ____ Not at all ____ Little ____ Average ____ A lot ____ Very much	JS
SI <sub>2</sub>	5-point Likert Disagree–Agree	I seldom look at the job listings on monster.com. Strongly disagree _____ Strongly agree	SI
JS <sub>5</sub>	Percent Satisfaction	Indicate your satisfaction with your current job at HBAT by placing a percentage in the blank, with 0% = Not satisfied at all, and 100% = Highly satisfied. _____	JS
AC <sub>3</sub>	5-point Likert	How often do you do things with your coworkers on your days off? Never ____ Rarely ____ Occasionally ____ Often ____ Very often	AC
SI <sub>3</sub>	5-point Likert Disagree–Agree	I have no interest in searching for a job in the next year. Strongly disagree _____ Strongly agree	SI
AC <sub>4</sub>	6-point Semantic Differential	Generally, how similar are your coworkers to you? Very different _____ Very similar	AC
SI <sub>4</sub>	5-point Likert	How likely is it that you will be working at HBAT one year from today? ____ Very unlikely ____ Unlikely ____ Somewhat likely ____ Likely ____ Very likely	SI

items, and one (JS) is indicated by five measured items. Every individual construct is identified. The overall model has more degrees of freedom than paths to be estimated. Therefore, in a manner consistent with the rule of thumb recommending a minimum of three indicators per construct but encouraging at least four, the order condition is satisfied. In other words, the model is overidentified. Given the number of indicators and a sufficient sample size, no problems with the rank condition are expected either. Any such problems should emerge during the analysis.

In the proposed model, all of the measures are hypothesized as reflective. That is, the direction of causality is from the latent construct to the measured items. For instance, an employee's desire to quit would tend to cause low

**Figure 10.5**

**Measurement Theory Model (CFA) for HBAT Employees**



scores on each of four indicators loading on the Staying Intentions (SI) construct. Each construct also has a series of indicators that share a similar conceptual basis, and empirically they would tend to move together. That is, we would expect that when one changes, systematic change will occur in the other.

### STAGE 3: DESIGNING A STUDY TO PRODUCE EMPIRICAL RESULTS

The next step requires that the study be designed and executed to collect data for testing the measurement model. The researcher must consider issues such as sample size and model specification, particularly in establishing the identification of the model.

HBAT next designed a study to test the measurement model. HBAT's interest was among its hourly employees and not its management team. Therefore, the HBAT personnel department supplied a random sample of 500 employees. The 500 represent employees from each of HBAT's divisions, including their operations in the United States, Europe, Asia, and Australia. Employees could respond to the questionnaires while they were at work and return them anonymously. Four hundred completed responses were obtained on the scale items described in Table 10.2. Several classification variables also were collected with the questionnaire. Initial screening showed no problems with missing data. Only two responses included any missing data. In one case, an out-of-range response was given, which is treated as a missing response. Using our rule of thumb from the previous chapter, the effective

sample size using pairwise deletion (otherwise known as the *all-available treatment*) is 399, because it is the minimum number of observations for any observed covariance.

**Specifying the Model** Depending on the software you use, different approaches are required at this point. Two of the most popular software packages will be discussed, although many other software packages can be used to obtain like results.

If you choose to use AMOS, then you begin by using the graphical interface to draw the model depicted in Figure 10.5. Once the model is drawn, you can drag the measured variables into the model and run the software. In contrast, if you choose to use LISREL, you can either use the drop-down menus to generate the syntax that matches the measurement model, draw the measurement model using a path diagram, or write the appropriate code into a syntax window. If either of the first two alternatives is chosen, LISREL can generate the program syntax automatically.

**Identification** Once the measurement model is specified, the researcher is ready to estimate the model. The SEM software will provide a solution for the specified model if everything is properly specified. The default estimation procedure is maximum likelihood, which will be used in this case because preliminary analysis with the data leads HBAT to believe that the distributional properties of the data are acceptable for this approach. The researcher must now choose the remaining options that are needed to properly analyze the results. A more complete discussion of options is included online.

Table 10.3 shows an initial portion of an output from the CFA results for this model. The output provides an easy way to quickly identify the parameters to be estimated and understand the degrees of freedom for the model. In this case, 52 parameters are to be estimated. Of the 52 free parameters, 16 are factor loadings (a single loading estimate per construct is fixed to 1 to set the scale), 15 represent factor variance and covariance terms, and 21 represent error variance terms. The total number of unique variance and covariance terms equals the initial (premodel) degrees of freedom ( $df_{initial}$ ):

$$df_{initial} = \frac{p(p + 1)}{2} = \frac{21(21 + 1)}{2} = 231$$

Because 231 is greater than 52, the model is identified with respect to the order condition. It includes more degrees of freedom than free parameters. Every free parameter costs one df, so the net df for the model are:

$$df = df_{initial} - \# \text{ free parameters} = 231 - 52 = 179$$

No problems emerge with the rank condition for identification, because we have at least four indicators for each construct. Furthermore, our sample size is sufficient, so we believe the model will converge and produce reliable results. SEM users need to understand these basic principles to help avoid or spot potential identification problems.

## STAGE 4: ASSESSING MEASUREMENT MODEL VALIDITY

We now examine the results of testing this measurement theory by comparing the theoretical measurement model against reality, as represented by the sample covariance matrix. Both the overall model fit and the criteria for construct validity must be examined. Therefore, we will review key fit statistics and the parameter estimates.

**Overall Fit** CFA output includes many fit indices. We did not present all possible fit indices. Rather, we will focus on the key GOF values using our rules of thumb to provide some assessment of fit. Each SEM program (AMOS, LISREL, EQS, etc.) includes a slightly different set, but most contain all key values, particularly the  $\chi^2$  statistic, as well as the CFI and RMSEA.

Table 10.4 includes selected fit statistics from the CFA output. The overall model  $\chi^2$  is 236.62 with 179 degrees of freedom. The  $p$ -value associated with this result is .0061. This  $p$ -value is significant using a type I error rate of .05.

**Table 10.3** Parameters to be Estimated in the HBAT CFA Model

Indicator Variable Loadings					
	JS	OC	SI	EP	AC
JS <sub>1</sub>	0	0	0	0	0
JS <sub>2</sub>	1	0	0	0	0
JS <sub>3</sub>	2	0	0	0	0
JS <sub>4</sub>	3	0	0	0	0
JS <sub>5</sub>	4	0	0	0	0
OC <sub>1</sub>	0	0	0	0	0
OC <sub>2</sub>	0	5	0	0	0
OC <sub>3</sub>	0	6	0	0	0
OC <sub>4</sub>	0	7	0	0	0
SI <sub>1</sub>	0	0	0	0	0
SI <sub>2</sub>	0	0	8	0	0
SI <sub>3</sub>	0	0	9	0	0
SI <sub>4</sub>	0	0	10	0	0
EP <sub>1</sub>	0	0	0	0	0
EP <sub>2</sub>	0	0	0	11	0
EP <sub>3</sub>	0	0	0	12	0
EP <sub>4</sub>	0	0	0	13	0
AC <sub>1</sub>	0	0	0	0	0
AC <sub>2</sub>	0	0	0	0	14
AC <sub>3</sub>	0	0	0	0	15
AC <sub>4</sub>	0	0	0	0	16

Construct Variances and Covariances					
	JS	OC	SI	EP	AC
JS	17				
OC	18	19			
SI	20	21	22		
EP	23	24	25	26	
AC	27	28	29	30	31

**Error Terms for Indicators (one per indicator) = 21**  
**Total number of estimated parameters: 16 + 15 + 21 = 52**

Note that one loading per construct is not estimated as it will be set to a value of 1.0 to "set the scale." Thus, 16 parameters are estimated for loadings.

Fifteen estimated parameters—one variance for each construct plus the 10 unique covariances among constructs

Thus, the  $\chi^2$  goodness of fit statistic does not indicate that the observed covariance matrix matches the estimated covariance matrix within sampling variance. However, given the problems associated with statistical power, and the effective sample size of 399, we examine other fit statistics closely as well.

Next we look at several other fit indices. Our rule of thumb suggests that we rely on at least one absolute fit index and one incremental fit index, in addition to the  $\chi^2$  results. The value for RMSEA, an absolute fit index, is 0.027. This value appears quite low and is below the .08 guideline for a model with 21 measured variables and a sample size of 400. Using the 90 percent confidence interval for this RMSEA, we conclude the true value of RMSEA is between 0.015 and 0.036. Thus, even the upper bound of RMSEA is low in this case. The RMSEA therefore provides additional support for model fit. Next we see the standardized root mean square residual (SRMR) with a value of .035, below even the conservative cut-off value of .05. The third absolute fit statistic is the normed  $\chi^2$ , which is 1.32. This measure is the chi-square value divided by the degrees of freedom ( $236.62/179 = 1.32$ ). The normed  $\chi^2$  is a rather crude statistic and should not be considered a replacement for the actual  $\chi^2$  value. However, some consider a value of 2.0 or less to be good, and between 2.0 and as high as 5.0 acceptable in very complex models with large samples.

**Table 10.4** The HBAT CFA Goodness-of-Fit Statistics

Chi-square ( $\chi^2$ )
Chi-square = 236.62 ( $p = 0.0061$ )
Degrees of freedom = 179
Absolute Fit Measures
Goodness-of-fit index (GFI) = 0.95
Root mean square error of approximation (RMSEA) = 0.027
90 percent confidence interval for RMSEA = (0.015; 0.036)
Root mean square residual (RMR) = 0.086
Standardized root mean residual (SRMR) = 0.035
Normed chi-square = 1.32
Incremental Fit Indices
Normed fit index (NFI) = 0.97
Non-normed fit index (NNFI) = 0.99
Comparative fit index (CFI) = 0.99
Relative fit index (RFI) = 0.97
Parsimony Fit Indices
Adjusted goodness-of-fit index (AGFI) = 0.93
Parsimony normed fit index (PNFI) = 0.83

Moving to the incremental fit indices, the CFI is the most widely used index. In our HBAT CFA model the CFI has a value of 0.99, which, like the RMSEA, exceeds the CFI guidelines of greater than .94 for a model of this complexity and sample size (see Table 9.4). The other incremental fit indices also exceed suggested cut-off values. Although this model is not compared to other models, the parsimony normed fit index (PNFI) is 0.83, which may become useful if the user considers alternative models.

The CFA results suggest the HBAT measurement model provides a reasonably good fit, and thus it is suitable to proceed to further examination of the model results. Issues related to construct validity will be examined next and then the focus shifts to model diagnostics aimed at improving the specified model.

**Construct Validity** To assess construct validity, we examine convergent, discriminant, and nomological validity. Face validity, as noted earlier, was established based on the content of the corresponding items.

**CONVERGENT VALIDITY** CFA provides a range of information used in evaluating convergent validity. Even though maximum likelihood factor loading estimates are not associated with a specified range of acceptable or unacceptable values, their magnitude, direction, and statistical significance should be evaluated.

We begin by examining the unstandardized factor loading estimates in Table 10.5. In LISREL, these are termed lambda values, whereas in AMOS the loading estimates are not distinguished from regression weights and are shown under the “Estimates” portion of the output. Loading estimates that are statistically significant provide a useful start in assessing the convergent validity of the measurement model. The results confirm that all loadings in the HBAT model are highly significant as required for convergent validity.

Maximum likelihood estimates are the typical default estimation option for SEM programs. The model provides unstandardized loading estimates, but they offer little diagnostic information other than directionality and statistical significance. Instead, we examine standardized loadings because they are needed to calculate discriminant validity and reliability estimates. For construct validity, our guidelines are that individual standardized factor loadings (regression weights) should be at least .5, and preferably .7. Moreover, variance-extracted measures should equal or exceed 50 percent, and .7 is considered the minimum threshold for construct reliability, except when conducting exploratory research. Realize the loading estimates must average .7 or better to reach the AVE $\geq$ .5 cutoff.

**Table 10.5** HBAT CFA Factor Loading Estimates and t-values

Indicator	Construct	Estimated Loading	Standard Error	t-value
JS <sub>1</sub>	JS	1.00	— <sup>a</sup>	— <sup>a</sup>
JS <sub>2</sub>	JS	1.03	0.08	13.65
JS <sub>3</sub>	JS	0.90	0.07	12.49
JS <sub>4</sub>	JS	0.91	0.07	12.93
JS <sub>5</sub>	JS	1.14	0.09	13.38
OC <sub>1</sub>	OC	1.00	— <sup>a</sup>	— <sup>a</sup>
OC <sub>2</sub>	OC	1.31	0.11	12.17
OC <sub>3</sub>	OC	0.78	0.08	10.30
OC <sub>4</sub>	OC	1.17	0.10	11.94
SI <sub>1</sub>	SI	1.00	— <sup>a</sup>	— <sup>a</sup>
SI <sub>2</sub>	SI	1.07	0.07	16.01
SI <sub>3</sub>	SI	1.06	0.07	16.01
SI <sub>4</sub>	SI	1.17	0.06	19.18
EP <sub>1</sub>	EP	1.00	— <sup>a</sup>	— <sup>a</sup>
EP <sub>2</sub>	EP	1.03	0.07	14.31
EP <sub>3</sub>	EP	0.80	0.06	13.68
EP <sub>4</sub>	EP	0.90	0.06	14.48
AC <sub>1</sub>	AC	1.00	— <sup>a</sup>	— <sup>a</sup>
AC <sub>2</sub>	AC	1.24	0.06	18.36
AC <sub>3</sub>	AC	1.04	0.06	18.82
AC <sub>4</sub>	AC	1.15	0.06	18.23

<sup>a</sup>Not estimated when loading set to fixed value (i.e., 1.0).

Table 10.6 displays standardized loadings (standardized regression weights using AMOS terminology). The lowest loading obtained is .58, linking organizational commitment (OC) to item OC1. Two other loadings estimates fall just below the .7 standard. The average variance extracted estimates and the construct reliabilities are shown at the bottom of Table 10.6. The AVE estimates range from 51.9 percent for JS to 68.1 percent for AC. All exceed the 50 percent rule of thumb. Construct reliabilities range from .83 for the OC construct to .89 for both SI and AC. Once again, these exceed .7, suggesting adequate reliability. These values were computed using the formulas shown earlier in the chapter when convergent validity was discussed. As of this date, SEM programs do not routinely provide these values.

Taken together, the evidence supports the convergent validity of the measurement model. Although three loading estimates are below .7, two of these are just below the .7, and the other does not appear to be significantly harming model fit or internal consistency. The average variance extracted estimates all exceed .5, and the reliability estimates all exceed .7. In addition, the model fits relatively well. Therefore, all the items are retained at this point and adequate evidence of convergent validity is provided.

**DISCRIMINANT VALIDITY** We now turn to discriminant validity. First, we examine the interconstruct covariances. After standardization, the covariances are expressed as correlations. All SEM programs provide the construct correlations whenever standardized results are requested. Some (LISREL) will have a default text output that prints them as an actual correlation matrix. Others (i.e., AMOS) may simply list them in text output. The information is the same.

The conservative approach for establishing discriminant validity compares the AVE estimates for each factor with the squared interconstruct correlations associated with that factor. All AVE estimates from Table 10.6 are greater than the corresponding interconstruct squared correlation estimates in Table 10.7 (above the diagonal). Therefore, this test indicates there are no problems with discriminant validity for the HBAT CFA model.

**Table 10.6** HBAT Standardized Factor Loadings, Average Variance Extracted, and Reliability Estimates

	JS	OC	SI	EP	AC
JS <sub>1</sub>	0.74				
JS <sub>2</sub>	0.75				
JS <sub>3</sub>	0.68				
JS <sub>4</sub>	0.70				
JS <sub>5</sub>	0.73				
OC <sub>1</sub>		0.58			
OC <sub>2</sub>		0.58			
OC <sub>3</sub>		0.66			
OC <sub>4</sub>		0.84			
SI <sub>1</sub>			0.81		
SI <sub>2</sub>			0.86		
SI <sub>3</sub>			0.74		
SI <sub>4</sub>			0.85		
EP <sub>1</sub>				0.70	
EP <sub>2</sub>				0.81	
EP <sub>3</sub>				0.77	
EP <sub>4</sub>				0.82	
AC <sub>1</sub>					0.82
AC <sub>2</sub>					0.82
AC <sub>3</sub>					0.84
AC <sub>4</sub>					0.82
Average					
Variance					
Extracted	51.9%	56.3%	66.7%	60.3%	68.1%
Construct					
Reliability	0.84	0.83	0.89	0.86	0.89

Computed using the formula above as the average squared factor loading (squared multiple correlation).

Computed using the formula from above and the squared sum of the factor loadings.

**Table 10.7** HBAT Construct Correlation Matrix (Standardized)

	JS	OC	SI	EP	AC
JS	1.00	.04	.05	.06	.00
OC	0.21***	1.00	.30	.25	.09
SI	0.23***	0.55***	1.00	.31	.10
EP	0.24***	0.50***	0.56***	1.00	.06
AC	0.05	0.30***	0.31***	0.25***	1.00

Significance Level: \* = .05, \*\* = .01, \*\*\* = .001

Note: Values below the diagonal are correlation estimates among constructs, diagonal elements are construct variances, and values above the diagonal are squared correlations.

The congeneric measurement model also models discriminant validity, because it does not contain any cross-loadings among either the measured variables or the error terms. This congeneric measurement model provides a good fit and shows little evidence of substantial cross-loadings. Taken together, these results support the discriminant validity of the HBAT measurement model.

**NOMOLOGICAL VALIDITY** Assessment of nomological validity is based on the approach outlined in Chapter 3 for EFA. The correlation matrix provides a useful start in this effort to the extent that the constructs are expected to relate to one another. Previous organizational behavior research suggests that more favorable evaluations of all constructs

are generally expected to produce positive employee outcomes. For example, these constructs are expected to be positively related to whether an employee wishes to stay at HBAT. Moreover, satisfied employees are more likely to continue working for the same company. Most important, this relationship simply makes sense.

Correlations between the factor scores for each construct are shown in Table 10.7. The results support the prediction that these constructs are positively related to one another. Specifically, satisfaction, organizational commitment, environmental perceptions, and attitudes toward coworkers all have significant positive correlations with staying intentions. In fact, only one correlation is inconsistent with this prediction. The correlation estimate between AC and JS is positive, but not significant ( $p >= 0.87$ ). Because the other correlations are consistent, this one exception is not a major concern.

Nomological validity can also be supported by demonstrating that the constructs are related to other constructs not included in the model in a manner that supports the theoretical framework. Here the researcher must select additional constructs that depict key relationships in the theoretical framework being studied. In addition to the measured variables used as indicators for the constructs, several classification variables, such as employee age, gender, and years of experience, were also collected. Moreover, the performance of each employee was evaluated by management on a 5-point scale ranging from 1 = “Ver Low Performance” to 5 = “Ver High Performance.” Management provided this information to the consultants who then entered it into the database.

These other measures are helpful in establishing nomological validity. Previous research suggests that job performance is determined by an employee’s working conditions [2, 25]. The job performance–job satisfaction relationship is generally positive, but typically not a strong relationship. A positive organizational commitment–job performance relationship also is expected. In contrast, the relationship between job performance and staying is not as clear. Better-performing employees tend to have more job opportunities that can cancel out the effects of “employees who perform better are more comfortable on the job.” A positive environmental perceptions–job performance relationship is expected, because one’s working conditions directly contribute to how one performs a job. We also expect that experience will be associated with staying intentions. Thus, when intentions to stay are higher, an employee is more likely to remain with an organization. Age and staying intentions are not likely to be highly related. Employees approaching retirement are relatively older and could possibly report lower intentions to stay. This result would interfere with a positive age–staying intentions relationship that might exist otherwise.

Correlations between these three items and the factor scores for each measurement model construct are shown in Table 10.8. The predictions made in the previous paragraph can be compared with the results. This comparison shows that the correlations are consistent with the theoretical expectations as described. Therefore, the analysis of the correlations among the measurement model constructs and the analysis of correlations between these constructs and other variables both support the nomological validity of the model.

**Modifying the Measurement Model** In addition to evaluating goodness-of-fit statistics, the researcher must also check a number of model diagnostics. They may suggest some way to further improve the model or perhaps some specific problem area not revealed to this point. The following diagnostic measures from CFA should be checked: path estimates, standardized residuals, and modification indices.

**PATH ESTIMATES** Evaluation of the loadings of each indicator on a construct provides the researcher with evidence of the indicators that may be candidates for elimination. Loadings below the suggested cut-off values should be evaluated for deletion, but the decision is not made based just on the loadings, but on the other diagnostic measures as well.

**Table 10.8 Correlations Between Constructs and Age, Experience, and Job Performance**

	JS	OC	SI	EP	AC
Job Performance (JP)	.15 (.003)	.27 (.000)	.10 (.041)	.29 (.000)	.06 (.216)
Age	.14 (.005)	.12 (.021)	.06 (.233)	-.01 (.861)	.15 (.003)
Experience (EXP)	.08 (.110)	.07 (.159)	.15 (.004)	.01 (.843)	.12 (.018)

Note:  $p$ -values shown in parentheses.

Results are positive to this point. Even with good fit statistics, however, HBAT should check the model diagnostics. The path estimates were examined earlier. One loading estimate—the .58 associated with OC<sub>1</sub>—was noted because it fell below the ideal loading cut-off of .7. It did not appear to be causing problems, however, because the fit remained high. If other diagnostic information suggests a problem with this variable, action may be needed.

**STANDARDIZED RESIDUALS** The next diagnostic measures are the standardized residuals. The LISREL output shows the highest standardized residuals (e.g., greater than |2.5|), which prevents the researcher from having to search through all of the residuals. This can be a substantial task, because a residual term is computed for every covariance and variance term in the observed covariance matrix.

The HBAT CFA model has 231 residuals (remember, this was the number of unique elements in the observed covariance matrix). We will not display them all here. In Table 10.9, we show all standardized residuals greater than |2.5|. No standardized residuals exceed |4.0|, the benchmark value that may indicate a problem with one of

**Table 10.9** Model Diagnostics for the HBAT CFA Model

Standardized Residuals (all residuals greater than  2.5 )					
Negative Standardized Residuals					
SI <sub>3</sub>	and	OC <sub>1</sub>	-2.68		
SI <sub>4</sub>	and	OC <sub>1</sub>	-2.74		
EP <sub>3</sub>	and	OC <sub>1</sub>	-2.59		
Positive Standardized Residuals					
SI <sub>2</sub>	and	SI <sub>1</sub>	3.80		
SI <sub>4</sub>	and	SI <sub>3</sub>	3.07		
EP <sub>2</sub>	and	OC <sub>3</sub>	2.98		
EP <sub>4</sub>	and	OC <sub>3</sub>	2.88		
EP <sub>4</sub>	and	EP <sub>3</sub>	3.28		
Modification Indices for factor loadings					
	JS	OC	SI	EP	AC
JS <sub>1</sub>	–	0.19	1.44	2.71	0.69
JS <sub>2</sub>	–	2.11	0.32	0.53	2.55
JS <sub>3</sub>	–	0.00	0.29	0.16	0.00
JS <sub>4</sub>	–	0.59	0.09	0.40	0.10
JS <sub>5</sub>	–	3.20	2.59	1.38	4.96
OC <sub>1</sub>	0.64	–	10.86	3.02	2.75
OC <sub>2</sub>	0.07	–	10.84	0.51	7.14
OC <sub>3</sub>	1.01	–	3.15	7.59	1.86
OC <sub>4</sub>	0.00	–	0.07	0.02	1.02
SI <sub>1</sub>	0.00	0.00	–	0.29	0.02
SI <sub>2</sub>	1.89	0.08	–	1.66	0.59
SI <sub>3</sub>	0.15	1.85	–	0.10	0.00
SI <sub>4</sub>	2.78	0.55	–	2.46	0.37
EP <sub>1</sub>	0.10	1.85	1.74	–	0.05
EP <sub>2</sub>	0.11	3.48	0.78	–	0.53
EP <sub>3</sub>	0.31	0.17	3.00	–	0.00
EP <sub>4</sub>	0.17	0.17	0.11	–	0.85
AC <sub>1</sub>	0.70	0.38	0.02	0.07	–
AC <sub>2</sub>	0.43	2.45	0.84	0.22	–
AC <sub>3</sub>	1.59	0.07	0.02	0.89	–
AC <sub>4</sub>	1.29	3.70	0.89	3.01	–

the measures. Those between |2.5| and |4.0| also may deserve attention if the other diagnostics indicate a problem. The largest residual is 3.80 for the covariance between  $SI_2$  and  $SI_1$ . Both of these variables have a loading estimate greater than .8 on the SI construct. This residual may be explained by the content of the items. In this case,  $SI_2$  and  $SI_1$  may have slightly more in common with each other contentwise than they do with  $SI_3$  and  $SI_4$ , the other two items representing SI.

The HBAT analyst decides not to take action in this case given the high reliability and high variance extracted for the construct. In addition, the model fit does not suggest a great need for improvement. Three of the highest negative residuals are associated with variable  $OC_1$ , which also is the variable with the lowest loading estimate (.58). Again, no action is taken at this point given the overall positive results. If a residual associated with  $OC_1$  exceeded |4.0|, however, or if the model fit was marginal,  $OC_1$  would be a prime candidate for being dropped from the model. In this case, the congeneric representation, which meets the standards of good measurement practice, appears to hold quite well.

**MODIFICATION INDICES** Modification indices (MI) are calculated for every fixed parameter (i.e., all of the possible parameters that were not estimated in the model). The two sets of MIs most useful in a CFA are for the factor loadings and the error terms between items. Note that there are generally not any MIs for the relationships between constructs, because each construct has an estimated path to every other construct. As you would expect, a full listing of all modification indices is quite extensive and will not be provided here. Instead, we will identify the largest MI and also examine the MIs for the factor loadings.

First, the largest modification index is 14.44 for the covariance of the error terms of  $SI_1$  and  $SI_2$  (the full output can be found at the text's online resources). Although the modification indices for the error term correlations are somewhat useful in diagnosing problems with specific items, the researcher should avoid making model respecifications that involve correlated error terms.

The second type of MI that is quite useful in a CFA is for the factor loadings. As you can see in Table 10.9, each item has a modification index for all the constructs except the one it is hypothesized to relate to. This provides the researcher with an empirical estimate of how strongly each item is associated with other constructs (i.e., the potential for cross-loadings). As you can see, most of the values above 4.0 are associated with the items in the OC construct, which have fairly large values for the SI, EP and AC constructs.  $OC_2$  may be most problematic in that it has high values for both SI and AC, although  $OC_1$  and  $OC_3$  also have high values with at least one other construct. This may indicate some lack of unidimensionality for these two items, and this is reinforced by the fact that they have the two lowest standardized loadings across all items. But elimination of both items would violate the three-indicator rule, so they will be retained at this time.

A further specification search is not needed because the model has a solid theoretical foundation and because the CFA is testing rather than developing a model. If the fit were poor, however, a specification search could take place as described earlier in the chapter. Such an effort would rely heavily on the combined diagnostics provided by the factor loading estimates, the standardized residuals, and the modification indices. At this point, HBAT can proceed with confidence that the questionnaire measures these key constructs well.

## HBAT CFA SUMMARY

Four SEM stages are complete. The CFA results generally support the measurement model. The  $\chi^2$  statistic is significant above the .01 level, which is not unusual given a total sample size of 400 (with an effective sample size of 399 using the all-available [PD] approach). Both the CFI and RMSEA appear quite good. Overall, the fit statistics suggest that the estimated model reproduces the sample covariance matrix reasonably well. Further, evidence of construct validity is present in terms of convergent, discriminant, and nomological validity. Thus, HBAT can be fairly confident at this point that the measures behave as they should in terms of the unidimensionality of the five measures and in the way the constructs relate to other measures. Remember, however, that even a good fit is no guarantee that some other combination of the 21 measured variables would not provide an equal or better fit. The fact that the results are conceptually consistent is of even greater importance than are fit results alone.

## CFA RESULTS DETECT PROBLEMS

Textbook examples serve first to illustrate how tools like CFA work. When students use real data, however, things don't always go as smoothly as they do in the textbook example. Here, we illustrate a situation typical to researchers as they go about trying to confirm a proposed measurement model. In this case, we consider a hypothetical situation where the HBAT researchers wish to incorporate a construct to represent Supervisor Support as an additional hygiene factor (in addition to the attitude toward coworkers and environmental perceptions constructs). For this data, we use the HBATSEM6CON data. The data includes the 400 observations from the previous analysis with the addition of indicators for a sixth construct as described below.

**A Sixth Construct** The HBAT researcher goes to a scales book and quickly finds a six-item Supervisor Support scale. The scale consists of the following items:

- 1 SP1.** A 100-point slider scale anchored by "I have no problems with my supervisor" (scored 0) on the left and "My supervisor and I constantly have problems" (scored 100).
- 2 SP2.** A 9-point semantic differential scale anchored by 1 = "I personally like my supervisor" to 9 = "I personally dislike my supervisor."
- 3 SP3.** A 5-point Likert scale expressing agreement with the statement "My company often leaves me without all the resources I need to do my job."
- 4 SP4.** An 11-point scale asking agreement with "My supervisor helps me resolve all my problems at work."
- 5 SP5.** A 9-point scale asking how often "My supervisor gives me a pat on the back" ranging from 1 = never to 9 = whenever I deserve one.
- 6 SP6.** A 5-point Likert scale asking agreement with the statement "Management supports me when I have a problem."

The researcher is anxious to know if the Supervisor Support construct will work out and so checks the Cronbach alpha as an indicator of reliability and examines loadings from a single-component Principal Components Analysis. The coefficient alpha for the six items (with all items scored in the same direction) is 0.68. Plus, the SPSS scale reliability analysis does not suggest that dropping any item would result in more than a trivial improvement in the scale (.01). The loadings from Principal Components Analysis range from .45 to .81. Further, the scale correlates with the Job Performance variable at  $r = .46$ , suggesting nomological validity ( $p < .001$ ). None of these findings are particularly out of line with the rules of thumb for scales. Thus, the researcher is encouraged and moves forward with the CFA.

**6 Construct CFA Results** The researcher estimates the 6-construct, congeneric, measurement model using CFA. The overall model  $\chi^2$  is 623.6 with 309 degrees of freedom. As expected with a sample size of 400 and a relatively large number of variables (27), the  $\chi^2$  is statistically significant ( $p < 0.001$ ). Further, the researcher confirms that the model df are correct, using the formula discussed earlier in the chapter, as an indicator that the model is correctly specified. The model CFI is 0.94 and the RMSEA is 0.051. Using the guidelines from the previous chapter (see Table 9.4), the CFI is borderline, although the RMSEA falls within the range of good fit. Thus, the fit results leave some questions about the model but the researcher moves on to check other elements of construct validity.

Table 10.10 displays the standardized factor loadings, AVEs, construct reliabilities, inter-construct correlations and their squared values. As before, all are helpful in assessing the construct validity further, particularly given the questionable fit. First, we look at the loading estimates. Like before, all loading estimates are statistically significant, but statistical significance is of little value in determining measurement model validity. The loading estimates for the original five constructs are almost identical to those seen earlier (see Table 10.6), which provides evidence that the five-construct data are free of characteristics (i.e., distributional issues) that lead to interpretational confounding. However, four of the six Supervisor Support loading estimates fall below the ROT of |0.7|, raising some questions

**Table 10.10 HBAT 6 Construct CFA Results**

	JS	OC	SI	EP	AC	SUP
JS <sub>1</sub>	0.74					
JS <sub>2</sub>	0.75					
JS <sub>3</sub>	0.68					
JS <sub>4</sub>	0.71					
JS <sub>5</sub>	0.73					
OC <sub>1</sub>		0.58				
OC <sub>2</sub>		0.88				
OC <sub>3</sub>		0.66				
OC <sub>4</sub>		0.84				
SI <sub>1</sub>			0.81			
SI <sub>2</sub>			0.86			
SI <sub>3</sub>			0.74			
SI <sub>4</sub>			0.85			
EP <sub>1</sub>				0.69		
EP <sub>2</sub>				0.81		
EP <sub>3</sub>				0.74		
EP <sub>4</sub>				0.85		
AC <sub>1</sub>					0.82	
AC <sub>2</sub>					0.82	
AC <sub>3</sub>					0.84	
AC <sub>4</sub>					0.82	
SP <sub>1</sub>						-0.83
SP <sub>2</sub>						-0.75
SP <sub>3</sub>						-0.41
SP <sub>4</sub>						0.33
SP <sub>5</sub>						0.49
SP <sub>6</sub>						0.29
Variance Extracted	52.0%	56.3%	67.0%	60.0%	68.2%	30.9%
Construct Reliability	0.84	0.83	0.89	0.86	0.90	0.70
<b>Interconstruct Correlations (<math>\phi</math> matrix):</b>						
JS	1.00					
OC	0.21	1.00				
SI	0.23	0.55	1.00			
EP	0.24	0.50	0.54	1.00		
AC	0.05	0.30	0.31	0.25	1.00	
SUP	0.01	0.17	0.03	0.16	0.00	1
<b>Squared Interconstruct Correlations:</b>						
JS	1.00					
OC	0.04	1.00				
SI	0.05	0.30	1.00			
EP	0.06	0.25	0.29	1.00		
AC	0.00	0.09	0.10	0.06	1.00	
SUP	0.00	0.03	0.00	0.02	0.00	1

about the measurement model. Second, examining convergent validity further, the construct reliabilities for the first five constructs remain nearly unchanged. The construct reliability for Supervisor Support is 0.70, meeting our ROT of 0.7. (Alpha is 0.68 and CR is 0.70. Thus, the values are consistent and the small difference illustrates how coefficient alpha's formula tends to underestimate reliability.) The AVE for the new construct, however, falls below the .5 ROT at .309, providing another sign of problems. Third, we examine discriminant validity. Once again, the interconstruct correlation estimates exceed all their squared values, providing evidence of discriminant validity.

Taken together, the results cast doubt on the validity of the HBAT six-construct theoretical measurement model. Both the fit validity and convergent validity raise questions. Although the low loading estimates provide some indication of where the problems originate and what, if any, items might be dropped, the residuals provide the most useful diagnostics. In this case, several high standardized residuals result. In particular, they indicate that the model does a poor job of fitting the covariances between (Residual values obtained using AMOS.):

- SP5–SP6 = 8.03
- SP3–JS3 = 7.18
- SP3–JS5 = 7.14
- SP3–JS2 = 6.96
- SP3–JS1 = 6.71
- SP3–JS4 = 6.47

The largest single residual is for SP5 and SP6, SP3 provides several high residuals. Both SP3 and SP6 demonstrate a pattern of relatively high standardized residuals with many variables. The largest modification index is for the covariance between the error variances for SP5 and SP6, consistent with the highest residual. Interestingly, the second highest modification index is for the loading between Job Satisfaction and SP3, suggesting a cross-loading. Removing the constraints indicating no relationship for either the error-variance covariance or the cross-loading would create a model proposing that the psychometric properties needed to establish valid measurement do not hold. Thus, neither is an acceptable solution.

After some consideration of how to proceed, the HBAT research team decides to drop both SP3 and SP6. The model with 25 indicators yields a  $\chi^2$  of 328.8 with 260 degrees of freedom ( $p = 0.002$ ), a CFI of 0.985 and a RMSEA of 0.026. Thus, the model displays substantially more validity in terms of fit. The  $\chi^2$  difference between the two models is 197.8 with 49 degrees of freedom ( $p < 0.001$ ), indicating that the revised model fits significantly better than the original six-construct CFA. The loading estimates for the remaining four Supervisor Support items remain practically the same. As a consequence, the resulting construct reliability is 0.71 and the AVE is 0.40. The AVE, while improved, remains well below the .5 ROT. Thus, although the model now fits well, the convergent validity of Supervisor Support remains a question.

Several possibilities exist. One or two more Supervisor Support items could be dropped. Clearly, with only items SP1 and SP2, the convergent validity ROT would be met. However, dropping half or more of the proposed scale items would question whether the scale was really being confirmed versus refined. Further, potential identification issues with the employment of a two-item scale could surface. Another possibility is that the scale is not unidimensional, as theoretically proposed. Sometimes, a scale that contains both negatively and positively balanced items, as is the case with Supervisor Support's six items, will tend to produce separate positive and negative dimensions. In conclusion though, the researchers decide to move forward without the Supervisor Support construct. After looking at the content of the items, the researchers believe that perhaps only the first two items truly capture perceptions of a supportive supervisor. The other items seem to indicate something to do with the company or factors beyond the supervisor's control. Therefore, any theory involving implications of perceptions of a supportive supervisor will have to wait for another data collection to gather a new sample. In that effort, the researchers plan to take more time to identify a Supervisor Support scale with better face validity.

The widespread use of confirmatory factor analysis (CFA) has greatly improved psychometric measurement across all social sciences. Researchers now have a tool that provides a strong test of one's measurement theory. The key advantage is that the researcher can analytically test a conceptually grounded theory explaining how several different measured items represent important psychological, sociological, or business measures. CFA results and the corresponding additional construct validity indicators provide a thorough test and allow users a good understanding of the quality of their measures. Therefore, as we move from exploratory multivariate procedures toward more specific empirical testing of conceptual ideas, CFA becomes an essential multivariate tool.

It is difficult to highlight in a paragraph or two the key points about CFA. However, some important points that will help in understanding and using CFA include those corresponding to the objectives of the chapter:

**Distinguish between exploratory factor analysis and confirmatory factor analysis.** CFA cannot be conducted appropriately unless the researcher can specify both the number of constructs that exist within data to be analyzed and which specific measures should be assigned to each of these constructs. In contrast, EFA is conducted without knowledge of either of these things. EFA does not provide an assessment of fit. CFA provides an assessment of fit.

**Understand the basic principles of statistical identification and know some of the primary causes of SEM identification problems.** Statistical identification is extremely important in understanding CFA and obtaining useful CFA results. Underidentified models cannot produce statistical results. Any preliminary results from an underidentified model are likely not correct. Overidentified models with an excess number of degrees of freedom are required for statistical identification. In addition, each estimated parameter should be statistically identified. Many naïve users' frustrations with using SEM arise from not understanding identification deficits within constructs or within their overall model. Identification issues can be minimized by using construct scales that include at least three indicators at a minimum, and preferably at least four indicators.

**Know how to represent a measurement model using a path diagram.** Visual diagrams, or path diagrams, are useful tools in helping to translate a measurement theory into something that can be tested using standard CFA procedures. SEM programs make use of these path diagrams to show how constructs are related to measured variables. Good measurement practice suggests that a measurement model should be *cogeneric*, meaning that each measured variable loads on only one construct and no relationships among any error variance terms exist. Unless some strong theoretical reason indicates doing otherwise, all constructs should be linked with a two-headed, curved arrow in the path diagram indicating that the correlation between constructs will be estimated.

**Understand the concept of fit as it applies to measurement models and be able to assess the fit of a confirmatory factor analysis model.** CFA is a multivariate tool that computes a theory-consistent estimated covariance matrix using equations corresponding to the theoretical structure. The estimated covariance matrix is then compared to the actual, data-derived, or observed, covariance matrix. Models fit relatively well as these two matrices become more similar. Multiple fit statistics should be reported to help understand how well a model truly fits. They include the  $\chi^2$  goodness-of-fit statistic and degrees of freedom, one absolute fit index (such as the GFI or RMSEA), and one incremental fit index (such as the TLI or CFI). One of these indices should also be a badness-of-fit indicator, such as the SRMR or RMSEA. No absolute value for the various fit indices suggests a good fit (with the exception of the  $\chi^2$  goodness-of-fit statistic), only suggested guidelines are available for this task. The values associated with acceptable models vary from situation to situation and depend considerably on the sample size, number of measured variables, and the communalities of the factors.

**Assess the construct validity of a measurement model.** Construct validity is essential in confirming a measurement model. Multiple components of construct validity include convergent validity, discriminant validity, face validity, and nomological validity. Construct reliabilities and variance-extracted estimates are useful in establishing convergent validity. Discriminant validity is supported when the average variance extracted for a construct is greater than the shared variance between constructs. Good fit for a *cogeneric* model representation also provides evidence of construct validity. Face validity is established when the measured items are conceptually consistent with

How does CFA differ from EFA?

Looking back at some of the basic data issues described in Chapter 2, do any of the variables SP1-SP6 display properties that suggest it (they) should not be included in the CFA to begin with?

- 7 Berkovitz, L. R., and J. R. Rossiter. 2007. The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs. *Journal of Marketing Research* 44: 175–84.
- 8 Blalock, H. M. 1964. *Causal Inferences in Nonexperimental Research*. Chapel Hill: University of North Carolina Press.
- 9 Bollen, K. A., and K. G. Jöreskog. 1985. Uniqueness Does Not Imply Identification. *Sociological Methods and Research* 14: 155–63.
- 10 Burt, R. S. 1976. Interpretational Confounding of Unobserved Variables in Structural Equations Models. *Sociological Methods Research* 5: 3–52.
- 11 Carmines, E. G., and J. P. McIver. 1981. Analyzing Models with Unobserved Variables: Analysis of Covariance Structures. In G. W. Bohrnstedt and E. F. Borgatta (eds.), *Social Measurement: Current Issues*. Beverly Hills, CA: Sage, pp. 65–115.
- 12 Diamantopoulos, Adamantios, and Heidi M. Winklhofer. 2001. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research* 38: 269–77.
- 13 Dillon, W., A. Kumar, and N. Mulani. 1987. Offending Estimates in Covariance Structure Analysis—Comments on the Causes and Solutions to Heywood Cases. *Psychological Bulletin* 101: 126–35.
- 14 Fornell, C., and D. F. Larcker. 1981. Evaluating Structural Equations Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research* 18: 39–50.
- 15 Gerbing, D. W., and J. C. Anderson. 1984. On the Meaning of Within-Factor Correlated Measurement Errors. *Journal of Consumer Research* 11: 572–80.
- 16 Griffin, M., B. J. Babin, and D. Modianos. 2000. Shopping Values of Russian Consumers: The Impact of Habituation in a Developing Economy. *Journal of Retailing* 76: 33–52.
- 17 Hair, J. F., B. J. Babin, A. Money, and P. Samouel. 2003. *Essentials of Business Research*. Indianapolis, IN: Wiley.
- 18 Hayduk, L. A. 1987. *Structural Equation Modeling with LISREL*. Baltimore, MD: Johns Hopkins University Press.
- 19 Herting, J. R., and H. L. Costner. 1985. Respecification in Multiple Response Indicator Models. In *Causal Models in the Social Sciences*, 2nd edn. New York: Aldine, pp. 321–93.
- 20 MacCallum, R. C. 2003. Working with Imperfect Models. *Multivariate Behavioral Research* 38(1): 113–39.
- 21 MacCallum, R. C., M. Roznowski, and L. B. Necowitz. 1992. Model Modification in Covariance Structure Analysis: The Problem of Capitalization on Chance. *Psychological Bulletin* 111: 490–504.
- 22 Nasser, F., and J. Wisenbaker. 2003. A Monte Carlo Study Investigating the Impact of Item Parceling on Measures of Fit in Confirmatory Factor Analysis. *Educational and Psychological Measurement* 63: 729–57.
- 23 Netemeyer, R.G., W. O. Bearden, and S. Sharma. 2003. *Scaling Procedures: Issues and Applications*. Thousand Oaks, CA: Sage.
- 24 Nunnally, J. C. 1978. *Psychometric Theory*. New York: McGraw-Hill.
- 25 Quinones, Miguel A., and Kevin J. Ford. 1995. The Relationship Between Work Experience and Job Performance: A Conceptual and Meta-Analytic Review. *Personnel Psychology* 48: 887–910.
- 26 Rigdon, E. E. 1995. A Necessary and Sufficient Identification Rule for Structural Models Estimated in Practice. *Multivariate Behavior Research* 30: 359–83.
- 27 Silvia, E., M. Suyapa, and R. C. MacCallum. 1988. Some Factors Affecting the Success of Specification Searches in Covariance Structure Modeling. *Multivariate Behavioral Research* 23: 297–326.
- 28 Smitherman, H. O., and S. L. Brodsky. 1983. *Handbook of Scales for Research in Crime and Delinquency*. New York: Plenum Press.
- 29 Sajtos, L., and B. Magyar. 2016. Auxiliary Theories as Translation Mechanisms for Measurement Model Specification. *Journal of Business Research* 69, 3186–91.

# 11 Testing Structural Equation Models

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

- Distinguish a theoretical measurement model from a theoretical structural model.
- Describe the similarities between SEM and other multivariate techniques.
- Depict a model with dependence relationships using a path diagram.
- Test a structural model using SEM.
- Diagnose problems with the SEM results.

## Chapter Preview

The process of testing structural equations models (SEM) was introduced in Chapter 9 as involving both theoretical measurement and theoretical structural models. Chapter 10 provides an overview of developing a measurement model based on theory and then testing it with confirmatory factor analysis (CFA). CFA tests measurement theory based on the covariance between all measured items. As such, the CFA model provides the foundation for all further theory testing.

This chapter focuses on the second model—testing the theoretical structural model, where the primary focus shifts to a network of relationships among latent constructs. With SEM, we focus on how well the proposed structure fits, or reproduces, the observed covariance matrix. After assessing fit, we examine piecemeal relationships among constructs much as we examined relationships between independent and dependent variables in multiple regression analysis (Chapter 4). Even though we saw that summated factors representing theoretical constructs could be entered as variables in OLS regression models, those regression models treated variables and constructs identically. That is, multiple regression did not consider any of the measurement properties that go along with forming a multiple-item construct when estimating the relationship. SEM provides a better way of empirically examining a theoretical model by involving both the measurement model and the structural model in one analysis. In other words, it takes information about measurement into account in testing the structural model.

The chapter begins by describing some terminology associated with the testing of the structural model with SEM. In addition, we discuss the similarities and differences between SEM and other multivariate techniques. We then describe the last two stages (Stages 5 and 6) in the six-stage process for testing theoretical models and provide an illustration using the HBAT\_SEM dataset.

## Key Terms

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

**Causal model** Structural model that infers that relationships have a sequential ordering in which a change in one brings about a change in another.

**Feedback loop** Relationship when a construct serves as both a predictor and an outcome of another single construct. Feedback loops can involve either direct or indirect relationships. Also called a *non-recursive relationship*.

**Interpretational confounding** Measurement estimates for one construct are significantly affected by relationships other than those among the specific measures. It is indicated when loading estimates vary substantially from one model to another model that is the same except for the change in specification of one or more relationships.

**Non-recursive model** Structural model containing *feedback loops*.

**Path estimate** See *structural parameter estimate*.

**Post hoc analysis** After-the-fact tests of relationships for which no hypothesis was theorized. In other words, a path is tested where the original theory did not indicate a path.

**Recursive models** Structural models in which paths between constructs all proceed only from the antecedent construct to the consequences (outcome construct). No construct is both a cause and an effect of any other single construct.

**Saturated structural model** Recursive SEM model specifying the same number of direct *structural relationships* as the number of possible construct correlations in the CFA. The fit statistics for a saturated theoretical model should be the same as those obtained for the CFA model.

**Structural model** A process model specifying how constructs affect each other (i.e., the *structural theory*).

**Structural relationship** Dependence relationship (regression type) specified between any two latent constructs. Structural relationships are represented with a single-headed arrow and suggest that one construct is dependent on another. Exogenous constructs cannot be dependent on another construct. Endogenous constructs can be dependent on either exogenous or endogenous constructs.

**Structural parameter estimate** SEM equivalent of a regression coefficient that measures the linear relationship between a predictor construct and an outcome construct. Also called a *path estimate*.

**Structural theory** Conceptual representation of the relationships between constructs.

**Two-step SEM process** Approach to SEM in which the measurement model fit and construct validity are first assessed using CFA and then the *structural model* is tested, including an assessment of the significance of relationships. The structural model is tested only after adequate measurement and construct validity are established.

**Unit of analysis** Unit or level to which results apply. In business research, it often deals with the question of testing relationships between individuals' (people) perceptions versus between organizations. Analyses that involve both are known as multi-level.

## What Is a Structural Model?

In the previous chapter, we learned that the goal of psychometric measurement theory is to produce ways of measuring concepts in a reliable and valid manner. Measurement theory tests place the focus on patterns of potential relationships between measured variables and latent constructs. CFA tests a measurement theory by providing evidence of the validity of individual measures based on the model's overall fit (model's ability to reproduce observed covariance matrix) and other evidence of construct validity. CFA typically presumes all constructs are related to one another. The consequence is that the model is theoretically saturated at the construct level. A CFA has no over-identifying assumptions at the construct level. The results is that construct relationships are simple correlations, which are error disattenuated, meaning corrected for measurement error.

A **structural theory** is a conceptual representation of the **structural relationships** among constructs. The **structural model** component represents proposed theory with a set of structural equations specifying what things are related or not related to each other. A researcher can depict the equations with a visual diagram displaying the pattern of connections depicting proposed relationships and the non-connections depicting construct independence. Thus, in the CFA, every possible relationship is allowed, creating a saturated model, and the research can add the structure by taking away relationships not proposed to exist and replacing correlational relationships with implied causal links (single-headed arrows) for those relationships that theoretically should exist. The structural relationship

between any two constructs is represented empirically by the **structural parameter estimate**, also known as a **path estimate**. Structural models are referred to by several terms, including a theoretical model or **causal model**. A causal model infers that the relationships meet the conditions necessary for causation. The conditions for causality were discussed in Chapter 9 and the researcher should be careful not to depict the model as having causal inferences unless a theoretically sound explanation can be offered.

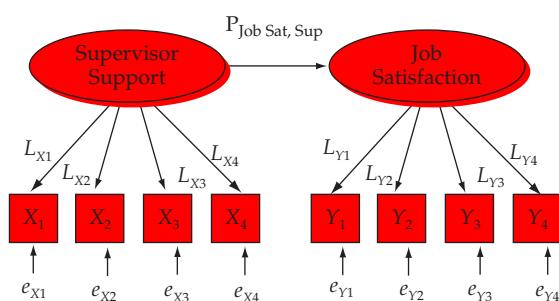
## A Simple Example of a Structural Model

The transition from a measurement model to a structural model is strictly the application of the structural theory in terms of relationships among constructs. Recall from Chapter 10 that a measurement model typically represents all construct relationships with proposed covariance/correlation links. We will revisit our simple measurement model example from Chapter 10 to illustrate the process of constraining a CFA model into a recursive structural model.

Figure 11.1 shows a two-construct structural model. The assumption now is that the first construct—Supervisor Support—is related to Job Satisfaction in a way that the relationship can be expressed as a regression coefficient. In the figure this relationship is shown as a structural relationship and labeled with a  $P =$  path estimate. In a causal theory, the model would imply that Supervisor Support causes or helps bring about Job Satisfaction.

The diagram in Figure 11.1 is similar to the CFA model. But when we move from the measurement model (CFA) to the structural model (SEM), there are some changes in abbreviations, terminology, and notation. No changes are made to the left side of the diagram representing the Supervisor Support construct. But there are changes in other areas, including the following:

- The relationship between the Supervisor Support and Job Satisfaction constructs in a CFA would be represented by a two-headed curved arrow (correlational relationship). This relationship changes to a dependence relationship and is now represented in Figure 11.1 by a single-headed arrow ( $P =$  path estimate). This arrow can be thought of as a relationship that is represented in a multiple regression model and estimated by a regression coefficient. This path shows the direction of the relationship in a structural model and represents the structural relationship that will be estimated to depict the strength of relationship between the two constructs.
- The constructs are now identified differently. In CFA exogenous and endogenous constructs are not distinguished, but with a structural model we must distinguish between exogenous and endogenous constructs. The traditional independent variables are now labeled exogenous and are still connected by correlations (double-headed curved arrows) unless a rationale exists for their independence. Traditional dependent variables are now labeled endogenous. Theory is tested by examining the effect of exogenous constructs (predictors) on endogenous constructs (outcomes). Also, with most SEM models there is more than one endogenous construct, so you have one endogenous construct predicting another. In such cases, one or more of the endogenous constructs operates as both a predictor and an outcome variable. This situation is not represented in Figure 11.1, but it will be shown in later examples.



**Figure 11.1**  
Visual Representation (Path Diagram) of a Simple Structural Theory

- The measured indicator variables (items) are no longer all represented by the letter X. Only the indicator variables for the exogenous construct are still represented by X. In contrast, the indicator variables for the endogenous construct are now represented by the letter Y. This is a typical distinction in structural models and is consistent with the approach used in other multivariate procedures (X associated with predictors, and Y associated with outcomes).
- The error variance terms now also have a notation that matches the exogenous–endogenous distinction. Error terms for all the variables are now labeled by the appropriate item (i.e., X variables or Y variables and the item number). Although not shown in the exhibit, each endogenous construct also includes an error variance term.
- The loading estimates are also changed to indicate exogenous or endogenous constructs. Variable loading estimates for exogenous constructs are represented by X and the item number, whereas variable loading estimates for endogenous constructs are represented by Y and the number.

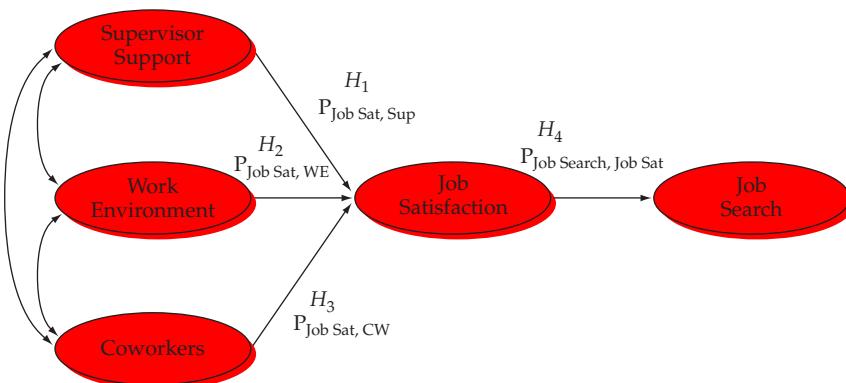
Structural models differ from measurement models in that the emphasis moves from the relationship between latent constructs and measured variables to the nature and magnitude of the relationships between constructs. Measurement models are tested using CFA. The CFA model is then altered based on the nature of relationships among constructs. The result is a structural model specification that is used to test the hypothesized theoretical model.

With these theoretical distinctions between CFA and SEM represented in the path diagram, we now move on to estimate the structural model using SEM procedures. We should note that in this type of situation, the observed covariance model does not change between models. Differences in model fit are based solely on the different relationships represented in the structural model.

## An Overview of Theory Testing with SEM

Given that the measurement model has already been examined and validated in a CFA analysis, the focus in a SEM analysis is testing the proposed structural theory by examining two issues: (1) overall and relative model fit as a measure of acceptance of the proposed theory and (2) structural parameter estimates representing direct and indirect relationships with one-headed arrows within a path diagram.

The validity of the theoretical model shown in Figure 11.2 is evaluated based on how well it reproduces the observed covariance matrix and on the size and direction of the hypothesized paths. If a relationship truly exists where the theoretical model proposed that none exists, or vice-versa, the fit will suffer. Note that in this figure we identify each hypothesized direct relationship in the path diagram ( $H_{\text{number}}$ ). If the model shows good fit, then the model is supported. The proposed relationships are further examined by considering the parameter estimate magnitudes and direction. But good fit does not mean that some alternative model might not fit better or be more accurate. Thus, the



**Figure 11.2**  
Expanded “Theoretical Model”  
of Job Search

model diagnostics should be examined to see if some alternative and theoretically sensible model may exist. Most typically, the diagnostics may suggest that some constraints may be removed (additional paths freed) to improve fit. However, like with CFA, any significant change to the model would need to be reconfirmed with new data.

The estimation of the structural parameter estimates is the same process used in CFA models. The primary distinction is the structural model does not have all the constructs related to each other, as in CFA. Thus, the structural model replaces the correlational relationships with dependence relationships. In doing so, we introduce the concept of direct and indirect effects. The derivation of the path estimates and the identification of direct and indirect effects are described in the Basic Stats appendix and materials available online. The reader may find this information helpful in understanding the full impact of any dependence relationship.

## Stages in Testing Structural Theory

Theory testing with SEM closely follows the way measurement theory is tested using CFA. The process is similar conceptually in that a theory is proposed and then tested based on how well it fits the data. As we deal with the theoretical relationships between constructs, greater attention is focused on the different types of processes that may exist.

### ONE-STEP VERSUS TWO-STEP APPROACHES

Even though SEM has the advantage of simultaneously estimating the measurement model and the structural model, our six-stage overall process is consistent within a **two-step SEM process** [1]. By two steps, we mean that in the first step we test the measurement model fit and other elements of construct validity with CFA. Once a satisfactory measurement model is obtained, the second step is to test the structural theory. Thus, two key tests—one measurement and one structural—totally assess fit and validity. Thus, the measurement model fit provides a relative basis for assessing the validity of the structural theory; the *ATFI* introduced in Chapter 9 is useful for describing the fit of the structural model relative to the CFA [4].

Early SEM models practiced a one-step approach, in which the overall fit of a model is tested without regard to separate measurement and structural models [3]. Yet a one-step model provides only one key test of fit and validity that combines measurement and structural aspects. With experience, researchers came to adopt the two-step approach for most conventional SEM applications.

The authors recommend separate testing of the measurement model via a two-step approach as essential because valid structural theory tests cannot be conducted with bad measures. A valid measurement model is essential because with poor measures we would not know what the constructs truly mean. Therefore, if a measurement model cannot be validated, researchers should first refine their measures and collect new data. If the revised measurement model can be validated, then and only then do we advise proceeding with a test of the full structural model. A more detailed discussion of this issue is presented later in the chapter in the section on interpretational confounding.

The six SEM stages now continue. Stages 1–4 covered the CFA process from identifying model constructs to assessing the measurement model validity (see Chapter 10). If the measurement is deemed sufficiently valid, then the researcher can test a structural model composed of these measures, bringing us to Stages 5 and 6 of the SEM process. Stage 5 involves specifying the structural model and Stage 6 involves assessing its validity.

## Stage 5: Specifying the Structural Model

We turn now to the task of specifying the structural model. This process involves determining the appropriate unit of analysis, representing the theory visually using a path diagram, clarifying which constructs are exogenous and endogenous, and several related issues, such as sample size and identification.

## UNIT OF ANALYSIS

One issue not visible in a model is the **unit of analysis**. The researcher must ensure that the model's measures capture the appropriate unit of analysis. For instance, organizational researchers often face the choice of testing relationships representing individual perceptions versus the organization or business unit. Marketing researchers also study organizations, but they sometimes look at an exchange dyad (buyer and seller), retail store, or advertisement as the unit of analysis. Individual perceptions represent each person's opinions or feelings. Organizational factors represent characteristics that describe an individual organization. A construct such as employee *esprit de corps* may well exist at both the individual and organizational levels. *Esprit de corps* can be thought of as how much enthusiasm an employee has for the work and the firm. In this way, one employee can be compared to another. But it also can be thought of as a characteristic of the firm overall. In this way, one firm can be compared to another and the sample size is now determined by how many firms are measured rather than by the number of individual respondents. The choice of unit of analysis determines how a scale is treated.

For example, a multiple-item scale could be used to assess the *esprit de corps* construct. If the desired unit of analysis is at the individual level and we want to understand relationships that exist among individuals, the research can proceed with individual responses. However, if the unit of analysis is the organization, or any other group, responses must be aggregated over all individuals responding for that group. Thus, organizational-level studies require considerably more data because multiple responses must be aggregated into one group. Organizations also possess many nonlatent characteristics such as size, location and organizational structure.

Once the unit of analysis is decided and data are collected, the researcher must aggregate the data if group-level responses are used to set up the appropriate SEM. If the unit of analysis is the individual, the researcher can proceed as before. Sometimes, multiple units of analysis are included in the same model. For instance, organizational culture may cause an individual's job satisfaction. The term multilevel model refers to these analyses as discussed in a previous chapter.

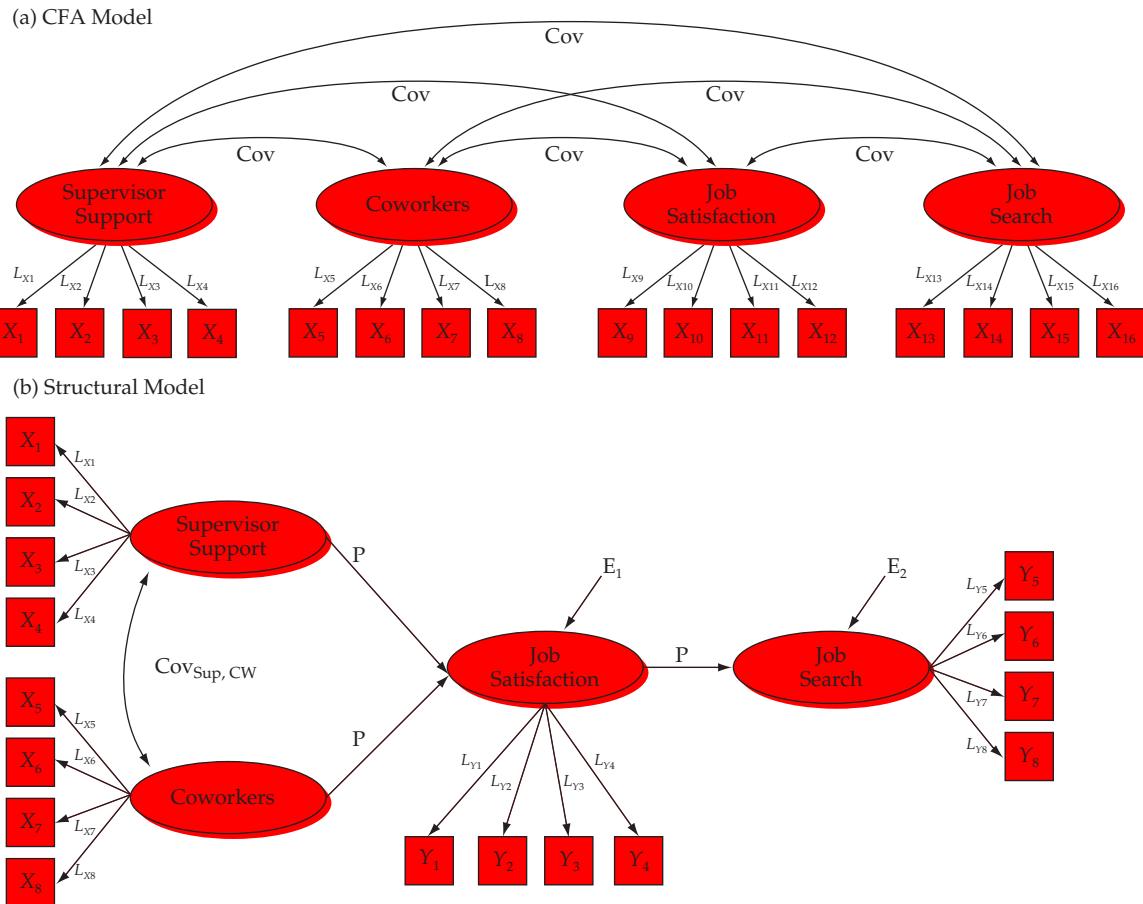
## MODEL SPECIFICATION USING A PATH DIAGRAM

We now consider how a theory is represented by visual diagrams. Paths indicate relationships and a lack of a connection indicates an independence constraint. Constraints provide over-identification and are represented by *fixed parameters*, meaning relationships not estimated by the SEM routine. These typically are assumed to be (set at) zero and are typically not shown on a visual diagram. *Free parameters* refer to a relationship that will be estimated. These parameters are generally depicted by an arrow in a visual diagram.

Figure 11.2 included both fixed and free parameters. For instance, there is no relationship specified between Supervisor Support and Job Search. Therefore, no arrow is shown here, and the theory assumes this direct path is equal to zero (fixed at zero). But there is a path between Job Satisfaction and Job Search that represents an implied causal relationship between the two constructs and for which a parameter will be estimated. When the Job Satisfaction to Job Search path is combined with the Supervisor Support to Job Satisfaction path, the result is a theoretically proposed indirect relationship between Supervisor Support and Job Search. Thus in this case, the constraint between supervisor support and job search (no direct path) does not imply that the two are unrelated.

The parameters representing structural relationships between constructs are now our focus. These are in many ways the equivalent of regression coefficients and can be interpreted in a similar way. With SEM these parameters are divided into two types: (1) relationships between exogenous and endogenous constructs and (2) relationships between two endogenous constructs. Some software programs make a distinction between these two types (e.g., LISREL), whereas others (e.g., AMOS) consider all structural relationships similarly. For more detail on LISREL notation, refer to materials on the text's online resources.

**Starting with a Measurement Model** Figure 11.3 shows a CFA model that is converted into a subsequent structural model. The constructs are based on the previous employee Job Satisfaction example. Figure 11.3a shows a CFA that tests the measurement model. Each construct is measured by four indicator items. Thus, four latent constructs are measured by 16 measured indicator variables ( $X_1-X_{16}$ ). The error variance terms are not shown in the exhibit, but

**Figure 11.3** Changing a CFA Model to a Structural Model

each of the 16 indicator items also has a corresponding, free, error variance term. Relationships between constructs are estimated by correlational relationships (Cov). In this case, there are six covariance/correlation terms between constructs.

**Transforming to a Structural Model** The primary objective is to specify the structural model relationships as replacements for the correlational relationships found in the CFA model. This process, however, also involves a series of other changes, some just in terms of notation and others of more substantive issues (e.g., changing from exogenous to endogenous constructs). The following section describes the theoretical and notational changes involved in this process.

**THEORETICAL CHANGES** Specifying the structural model based on the measurement model necessitates the use of structural relationships (single-headed arrows for the hypothesized causal relationships) in place of the correlational relationships among constructs used in CFA. The structural theory is specified by using free parameters (ones to be estimated) and fixed parameters (ones fixed/constrained at a value, usually zero) to represent the proposed process. But in specifying the structural relationships, two other significant changes occur. First, there must now be the distinction between endogenous and exogenous constructs. Recall that in the CFA model this distinction is not made. But now those constructs that act as outcomes (i.e., structural relationships predict them) must be specified as endogenous. Endogenous constructs are easily recognized in the path diagram because they have one or more arrows depicting structural relationships pointing toward them. A second change is that endogenous constructs are

not fully explained and so each is associated with an error variance term ( $E$ ). The user should be careful to insert an error variance term for endogenous constructs when using AMOS. LISREL includes the error term for endogenous constructs automatically.

Note that although it may seem that changing a relationship from a correlational to a dependence relationship only involves changing a double-headed arrow to a single-headed arrow, the implications for the estimation of parameters in the model may be substantial. As described in the Basic Stats appendix on the text's online resources, effects between constructs include estimates of both the direct and indirect effect presumed by a model. Specifying a relationship as a dependence relationship specifies constraints on which effects are used in estimating the path estimate.

**NOTATIONAL CHANGES** The second type of change is one that is primarily notational to reflect either the change in the type of relationship (correlational to structural) or the type of construct (exogenous versus endogenous). As we have discussed, the underlying indicators do not change, but their notation may change. In a CFA model, all indicators used the designation of  $X$ . But indicators of endogenous constructs are distinguished by using the designation of  $Y$ . This impacts not only the item labeling but also the notation used for factor loadings and error terms. Remember, the underlying observed measures do not change, just their notation in the model.

We will now demonstrate the changes that occur in translating the CFA model described in Chapter 10 into a SEM model with hypothesized structural relationships. Let us assume the employee job satisfaction theory hypothesized that Supervisor Support and Coworkers are related to Job Satisfaction. This implies a single structural relationship with Job Satisfaction as a function of Supervisor Support and Coworkers. When Job Search is included in the theoretical model, it is viewed as an outcome of Job Satisfaction. Supervisor Support and Coworkers are not hypothesized as being directly related to Job Search.

Figure 11.3b corresponds to this structural theory. Several changes can be seen in transforming the CFA measurement model into the SEM structural model.

### Theoretical Changes.

- 1 Based on the proposed theory, there are two exogenous constructs and two endogenous constructs. Supervisor Support and Coworkers are exogenous based on our theory, and the model therefore has no arrows pointing at them. Job Satisfaction is a function of Supervisor Support and Coworkers, and is therefore endogenous (arrows pointing at it). Job Search is a function of Job Satisfaction, and therefore is also endogenous. Thus, the representation of Supervisor Support and Coworkers is not changed.
- 2 The hypothesized relationships between Supervisor Support and Job Satisfaction and Coworkers and Job Satisfaction, as well as the relationship between Job Satisfaction and Job Search, are all represented by a  $P$  (path coefficient).
- 3 No direct relationships are shown between Supervisor Support and Job Search or Coworkers and Job Search because they are fixed at zero based on our theory. That is, the theory does not hypothesize a direct relationship between either Supervisor Support and Job Search or Coworkers and Job Search.
- 4 The hypothesized relationship between Supervisor Support and Coworkers remains a correlational relationship and is still a two-headed arrow and represented by Cov.

### Notational Changes.

- 1 The measured indicator variables for the exogenous constructs are still identified as  $X_1$  to  $X_8$ . But the measured indicator variables for the endogenous constructs are now identified as  $Y_1$  to  $Y_8$ .
- 2 The parameter coefficients representing the loading paths for exogenous constructs are still identified as  $L_{X1}$  to  $L_{X8}$ . In contrast, the parameter coefficients representing the loading paths for endogenous constructs are now identified as  $L_{Y1}$  to  $L_{Y8}$ .
- 3 Two other new terms appear— $E_1$  and  $E_2$ . They represent the error variance of prediction for the two endogenous constructs. After these error variances are standardized, they can be thought of as the opposite of an  $R^2$ . That is, they are similar to the residual in regression analysis.

**DEGREES OF FREEDOM** Computation of the degrees of freedom in a structural model proceeds in the same fashion as the CFA model, except that the number of estimated parameters is generally smaller. First, because the number of indicators does not change, neither does the initial number of degrees of freedom available. What primarily changes in most situations is a reduction in the number of structural relationships between constructs relative to the CFA model, in which all possible relationships between constructs are estimated.

In our example, there are 16 indicators resulting in a total of 136 unique values in the covariance matrix:

$$[p \times (p + 1)]/2 = (16 \times 17)/2 = 136$$

In Chapter 10 we saw that the CFA for this model requires 38 estimated parameters, leaving 98 degrees of freedom for the CFA model.

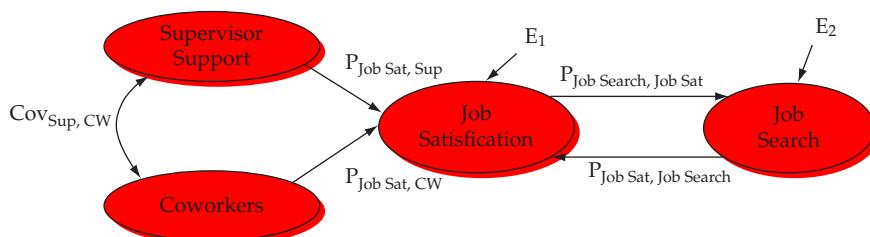
For the structural model, we still have the same 136 unique values in the covariance matrix, but the degrees of freedom now differ in the relationships between constructs. We still have 32 estimated parameters for the indicators (a loading and error term for each of the 16 indicators – presuming the variance for). Now, instead of the six correlational relationships between the four constructs, we have one correlational relationship (Supervisor Support ↔ Coworkers) and three structural relationships: Supervisor Support → Job Satisfaction, Coworkers → Job Satisfaction, Job Satisfaction → Job Search. This gives a total of 36 free parameters or a total of 100 degrees of freedom ( $136 - 36 = 100$ ). The two additional degrees of freedom come from not specifying (i.e., constraining to 0) the direct relationships from Supervisor Support and Coworkers to Job Search. Rather, they both are directly related to Job Satisfaction, which in turn directly predicts Job Search.

**Recursive Versus Non-recursive Models** One final distinction that must be made when specifying the structural model is if it is to be a recursive or non-recursive model. A model is considered **recursive** if the paths between constructs all proceed only from the predictor (antecedent) construct to the dependent or outcome construct (consequences). In other words, a recursive model does not contain any constructs that are both determined by some antecedent and help determine that antecedent (i.e., no pair of constructs has arrows going both ways between them). Recursive structural models will never have fewer degrees of freedom than a CFA model involving the same constructs and variables because they contain more constraints (i.e., proposed connections have been removed).

In contrast, a **non-recursive model** contains feedback loops. A **feedback loop** exists when a construct is seen as both a predictor and an outcome of another single construct. The feedback loop can involve direct or indirect relationships. In the indirect relationship the feedback occurs through a series of paths or even through correlated error terms.

Figure 11.4 illustrates a structural model that is non-recursive. Notice that the construct Job Search is both determined by and determines Job Satisfaction. The parameters for the model include the path coefficients corresponding to both of these paths ( $P$ ). If the model included a path from Job Search back to Coworkers, the model would also be non-recursive. This is because Job Search would be determined indirectly by Coworkers through Job Satisfaction, and Coworkers would be directly determined by Job Search with the new path.

A theoretical interpretation of a non-recursive relationship between two constructs is that one construct is both a cause and effect of the other. Although this situation is unlikely with cross-sectional data, it becomes more



**Figure 11.4**  
A Non-recursive SEM Model

plausible with longitudinal data. It is difficult to produce a set of conditions that support a reciprocal relationship with cross-sectional data.

For instance, both intelligence and success in school can be thought of as latent constructs measured by multiple items. Does intelligence cause success in school or does success in school cause intelligence? Could it be that both are causal influences on each other? Longitudinal data may help sort out this issue because the time sequence of events can be taken into account.

Non-recursive models many times have problems with statistical identification. By including additional constructs and/or measured variables, we can help ensure that the order condition is met. The rank condition for identification could remain problematic, however, because a unique estimate for a single parameter may no longer exist (see Chapter 10). Therefore, we recommend avoiding non-recursive models, particularly with cross-sectional data.

## DESIGNING THE STUDY

Chapter 10 covered conditions for mathematical identification. If a CFA model is identified, then recursive structural models derived from it are likely identified too. A recursive structural model is nested within a CFA model and is more parsimonious because it contains more constraints, meaning fewer coefficients to estimate. Therefore, if the CFA model is identified, the structural model also should be identified—as long as the model is recursive, no interaction terms are included, the sample size is adequate, and a minimum of three measured items per construct is used. We now turn to some other issues that may occur in transitioning from a measurement model to a structural model.

**Single-Item Measures** Occasionally a structural model will involve single-item measures. A structural model does not require multi-item measurement. Single-items can be incorporated into a structural model in different ways:

- 1 The single items can be entered in the model directly as measured variables. In this case, the variables do not represent latent constructs but rather are directly observed variables.
- 2 The single items represent a latent construct or composite and is entered in the model as a construct represented by a single measured variable, which also includes an error variance term.

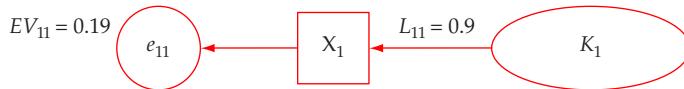
In the first case, the research may involve only non-psychological constructs, all of which lend themselves to direct measurement. For example, a structural model may contain statistics for demographic and economic variables taken from the census bureau. The US Census Bureau's website ([www.census.gov](http://www.census.gov)) provides nearly countless facts about the people and commerce within states, urban areas, zip codes, etc. More and more, the internet provides access to an abundance of similar archival (i.e., secondary) data with facts and figures on industries, employees, companies, and consumers. In addition, a researcher may wish to incorporate control variables into a larger SEM model or use psychological constructs to explain outcome variables like sales, stock-values, or firm performance. The result would be a structural model with a blend of multi-item constructs (ovals) and individual single-item variables (rectangles)—for example, see [5, 6]. In this way, the model proposes structural relationships among single variables and latent constructs. When variables are added directly to the model, they are assumed to be measured without error.

In the second approach, the single-item structurally plays the same role as a multi-item latent construct. The problem is that single-item measures, beyond the lack of statistical identification, do not lend themselves to psychometric validation through the computation of reliability and convergent validity. The question then becomes how can a single-item measure be represented within a SEM framework? How is it specified? Because its measurement characteristics are unknown, it requires the researcher's best judgment to fix the measurement parameters associated with the single item.

The difference between the variable representing a single-item construct and all other variables forming a covariance matrix is that it will be the only item associated with its construct. Thus, one measurement path links them

together with a factor loading. The measured variable, to be fully represented, also contains error variance. Clearly then, as noted earlier, a single-item construct would not be identified if a research tried to estimate its parameters (loading and error-variance). So, the researcher must specify both values. Factor loadings and error-variance terms for single-item constructs should be set based on the best knowledge available. The construct can be thought of as the “true” value and the variable as the observed value. If the researcher feels there is very little measurement error (i.e., the “true” score is very close to the observed score), then a high loading value and corresponding low error term will be specified. For example, if the researcher felt that there was no error in the observed value, then the loading would be set to 1.0 and the error term to zero. The end result would be the same as including links directly from a measured variable as in the case above.

But the assumption of no error for psychological measurement is questionable. Instead, the researcher may use past research providing an estimate of similar multi-item measures’ reliability or expert judgment to make an educated guess. The factor loading is then set (fixed) to the square root of the best-guess reliability. The corresponding error variance term is set to 1.0 minus the best-guess reliability. Thus, both parameters for the single-item construct are constrained to non-zero values. The illustration below depicts the process visually using a construct with a best-guess reliability of 0.81:



**Modeling the Construct Loadings When Testing Structural Theory** The CFA model in Figure 11.3a is modified to test the structural model shown in the bottom portion. The measurement portion of the structural model consists of the loading estimates and the corresponding error-variances (not shown in the figure) for the measured items and each of the covariance estimates between pairs of latent constructs. A good way to think about the proposed structural model is that the proposed theoretical structure replaces all the between-construct covariance estimates. Visually, one can think of first deleting all the two-headed arrows depicting construct covariation (correlation if standardized) and inserting single-headed arrows only in those places where causality is implied by theory. Having completed the deletion of two-headed arrows and insertion of single-headed arrows, the graphical diagram conversion from CFA to structural model is complete.

However, it's worth thinking about what happens with the factor loading estimates in the process. One argument suggests that, with the CFA model already estimated at this point, the factor loading estimates are known. Therefore, their values should be fixed and specified to the loading estimates obtained from the CFA model. In other words, they should no longer be free parameter estimates. Similarly, because the error variance terms are provided from the CFA, their values can also be fixed rather than estimated.

The rationale for fixing these values is that they are “known” and should not be subject to change because of relationships specified in the structural model. If they would change, this would be evidence of **interpretational confounding**, which means the measurement estimates for one construct are being significantly affected by the pattern of relationships between constructs. In other words, the loading estimates for any construct should not change noticeably just because a change is made to the structural model. An advantage to this approach is that the structural model is easier to estimate because so many more parameters are constrained. A complication of the process is that the CFA fit, unless re-estimated with all fixed parameters, is lost as a basis of assessing the structural model fit.

Another approach is to use the CFA factor pattern and allow the coefficients for the loadings and the error variance terms to be estimated along with the structural model coefficients, simplifying the transition from CFA to structural testing and eliminating the need to go through the process of fixing values for all the construct loading estimates and error variance terms. The process also can reveal any interpretational confounding by comparing the CFA loading estimates with those obtained from the structural model. If the standardized loading estimates vary substantially, then evidence of interpretational confounding exists. Small fluctuations are expected (.05 or less).

## Specifying the Structural Model

A structural model should be tested after CFA has validated the measurement model.

The structural relationships between constructs can be created by replacing the two-headed arrows from CFA with single-headed arrows representing implied cause-and-effect relationships and processes.

Recursive SEM models cannot produce fewer degrees of freedom than a CFA model involving the same constructs and variables.

Non-recursive models involving cross-sectional data should be avoided in most instances.

When a structural model is being specified, it should use the CFA factor pattern corresponding to the measurement theory and allow the coefficients for the loadings and the error variance terms to be estimated along with the structural model coefficients.

Measurement paths and error variance terms for constructs measured by only a single item (single measured variables or summated construct scores) should be based on the best knowledge available.

When measurement error for a single item is modeled:

The loading estimate between the variable and the latent construct is set (fixed) to the square root of the best estimate of its reliability.

The corresponding error term is set (fixed) to 1.0 minus the reliability estimate.

Given the issues with estimation and validation of single-item measures, they are typically not included in CFA.

The exception may be to specifically examine discriminant validity of the single-item with other variables or factors.

As inconsistencies increase in size and number, however, the researcher should examine the measures more closely. Evidence of interpretational confounding could suggest problems like endogeneity (see Chapter 9), model misspecification, or multicollinearity due to highly related constructs. Another advantage of this approach is that the original CFA model fit becomes a convenient basis of comparison in assessing the fit for the structural model (as illustrated in Chapter 9). This approach is used most often in practice, and it is the one recommended here.

## Stage 6: Assessing the Structural Model Validity

The final stage of the decision process evaluates the validity of the structural model based on the a comparison of the structural model fit compared to the CFA model as well as an examination of model diagnostics. The comparison of structural model fit to the CFA model assesses the degree to which the structural model decreases model fit due to its specified relationships. Here the researcher also determines the degree to which each specified relationship is supported by the estimated model (i.e., the statistical significance of each hypothesized path). Finally, model diagnostics are used to determine if any model respecification is indicated.

### UNDERSTANDING STRUCTURAL MODEL FIT FROM CFA FIT

This stage assesses the structural model's validity. The observed data are still represented by the observed sample covariance matrix, which will be compared to the estimated covariance matrix. In CFA, the estimated covariance matrix is computed-based on the restrictions (pattern of free and fixed parameter estimates) corresponding to the measurement theory. If the structural theory is recursive, then it cannot include more free relationships between

constructs than does the CFA model from which it is developed. Consequently, a recursive structural model cannot have a lower  $\chi^2$  value than that obtained in the CFA. Researchers are mistaken if they hope problems with a poorly performing CFA model will disappear in the structural model. If adequate fit is not found in the CFA model, the focus should be on finding a better theoretical measurement model or identifying better construct measures.

**Saturated Theoretical Models** If the SEM model specifies the same number of structural relationships as are possible construct correlations in the CFA, the model is considered a **saturated structural model**. Saturated theoretical models are not generally interesting because they usually cannot reveal any more insight than the CFA model. The fit statistics for a saturated theoretical model should be the same as those obtained for the CFA model, which is a useful point to know, and provides the basis for the Theoretical Fit Index (TFI), introduced in Chapter 9. One way researchers can check to see whether the transition from a CFA model setup to a structural model setup is correct is to test a saturated structural model. If its fit does not equal the CFA model fit, a mistake has been made.

**Assessing Overall Structural Model Fit** The structural model fit is assessed as was the CFA model fit. Therefore, good practice dictates that more than one fit statistic should be used. Recall from Chapter 12 that we recommended one absolute index, one incremental index, and the model  $\chi^2$  be used at a minimum. Once again, no magic set of numbers suggests good fit in all situations. Even a CFI equal to 1.0 and an insignificant  $\chi^2$  may not have a great deal of practical meaning in a simple model. Therefore, only general guidelines are given for different situations. Those guidelines remain the same for evaluating the fit of a structural model.

**Comparing the CFA Fit and Structural Model Fit** The CFA fit provides a useful baseline to assess the structural or theoretical fit. A recursive structural model cannot fit any better (have a lower  $\chi^2$ ) than the overall CFA. Therefore, one can conclude that the structural theory lacks validity if the structural model fit is substantially worse than the CFA model fit [2, 4]. A structural theory seeks to explain all of the relationships between constructs as simply as possible. The standard CFA model assumes a relationship exists between each pair of constructs. Only a saturated structural model would make this assumption. So, SEM models attempt to explain interconstruct relationships more simply and precisely than does CFA. When they fail to do so, the failure is reflected with relatively poor fit statistics. Conversely, a structural model demonstrating an insignificant  $\Delta\chi^2$  value with its CFA model is strongly suggestive of adequate structural fit.

**Examining Hypothesized Dependence Relationships** Recall that assessment of CFA model validity was based not only on fit, but also on construct validity. Likewise, the interpretation of SEM results may begin with, but does not end with, the assessment of fit. The researcher also examines the structural parameter estimates against the corresponding hypotheses.

Theoretical validity increases to the extent that the parameter estimates are *statistically and practically significant in the predicted direction*. That is, the estimated effect size is non-trivial and greater than 0 if hypothesized as positive and less than 0 if hypothesized as negative. The same standard applies to direct or indirect effects. Realize that with large sample sizes, even very small path estimates will be statistically significant. Only in rare instances would a standardized effect size less than  $|.1|$  be practically meaningful, even if statistically significant.

The researcher also can examine the squared multiple correlations for the endogenous constructs, which essentially provides an analysis of the  $R^2$ . The same general guidelines apply for these values as applied with multiple regression.

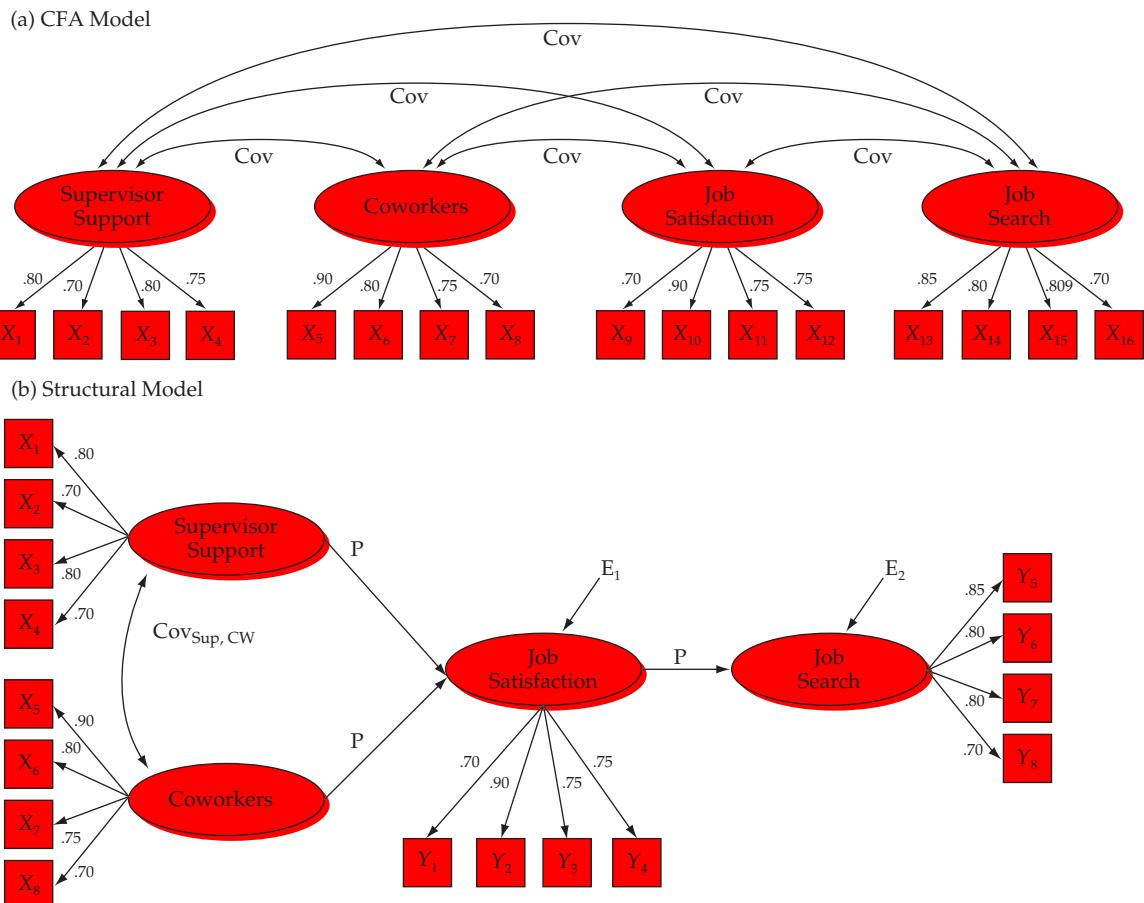
Also, we must remember, particularly in structural models, that good fit does not guarantee that the SEM model is the single best representation of the data. Like CFA models, alternative models can often produce the equally good or better empirical results. The concept of equivalent models takes on much more meaning when examining a structural model in the sense that an equivalent model is an “alternative theory” that has identical fit to the model being tested. Once again, theory becomes essential in assessing the validity of a structural model.

## EXAMINE THE MODEL DIAGNOSTICS

The same model diagnostics are provided for SEM as for the CFA model. For example, the pattern and size of standardized residuals can be used to identify problems in fit. We can assume the CFA model has sufficient validity if we have reached this stage, so the focus is on the diagnostic information about relationships between constructs. Particular attention is paid to path estimates, standardized residuals, and modification indices associated with the possible relationships between constructs in any of three possible forms (exogenous  $\rightarrow$  endogenous constructs, endogenous  $\rightarrow$  endogenous constructs, and error covariance among endogenous constructs). For instance, if a problem with model fit is due to a currently fixed relationship between an exogenous construct and an endogenous construct, it likely will be revealed through a standardized residual or a high modification index.

Consider the structural model in Figure 11.5. The model does not include a path linking Supervisor Support and Job Search. If the model were tested and a non-trivial relationship between these two really exists, a high standardized residual would likely be found between indicator variables that make up these two constructs ( $X_1-X_4$  and  $Y_5-Y_8$  in this case). The residuals would be telling us that the covariance between these sets of items has not been accurately reproduced by our initial theory. In addition, a high modification index might exist for the path that would be labeled  $P_{Job\ Search,\ Supervisor\ Support}$  (the causal relationship from Supervisor Support to Job Search). Modification indices, one for each constrained parameter, are shown in the standard SEM output. The indices can also be requested to be shown on a path diagram by using the appropriate drop-down menus. Model diagnostics are examined in the same manner as they are for CFA models.

**Figure 11.5**  
CFA Loading Estimates in a Structural Model



Should a model be respecified based on this diagnostic information? Researchers commonly conduct **post hoc analyses** following the theory test. Post hoc analyses are after-the-fact explorations of relationships not estimated in the structural theory test. In other words, a path is freed where the original theory constrained it to zero, meaning the original model did not contain that path. Recall that SEM provides an excellent tool for *testing theory*. Therefore, any relationship revealed in a post hoc analysis provides only empirical evidence, not theoretical support. For this reason, relationships identified post hoc should not be relied upon in the same way as the original theoretical relationships. Only when all the caveats of a model respecification strategy are noted should these changes be considered. Post hoc structural analyses are useful only in specifying potential model improvements that *must* make both theoretical sense and be cross-validated by testing the model with new data drawn from the same population. Thus, post hoc analyses are not useful in theory testing, and attempts by a researcher to represent posthoc findings as deductively derived theory are unprofessional and scientifically misleading.

## SEM Illustration

The CFA illustrations in the previous chapter began by testing a measurement theory. The result was validation of a set of construct indicators that enable HBAT to study relationships among five important constructs. HBAT would like to understand why some employees stay on the job longer than others. They know they can improve service quality and profitability when employees stay with the company longer. The six-stage SEM process begins with this goal in mind. For this illustration, we use the HBAT\_SEM dataset, available on the text's online resources.

The full measurement model was tested in the CFA chapter and was shown to have adequate fit and construct validity. Recall that the CFA fit statistics for this model were:

- $\chi^2$  is 236.6 with 179 degrees of freedom ( $p < .05$ )
- CFI = .99
- RMSEA = 0.027

To refresh your memory, the five constructs are defined here:

- *Job Satisfaction (JS)*. Reactions resulting from an appraisal of one's job situation.
- *Organizational Commitment (OC)*. The extent to which an employee identifies and feels part of HBAT.
- *Staying Intentions (SI)*. The extent to which an employee intends to continue working for HBAT and is not participating in activities that make quitting more likely.
- *Environmental Perceptions (EP)*. Beliefs an employee has about day-to-day, physical working conditions.
- *Attitudes Toward Coworkers (AC)*. Attitudes an employee has toward the coworkers he/she interacts with on a regular basis.

The analysis will be conducted at the individual level. HBAT is now ready to test the structural model using SEM.

## STAGE 5: SPECIFYING THE STRUCTURAL MODEL

With the construct measures in place, researchers now must establish the structural relationships among the constructs and translate them into a form suitable for SEM analysis. The following sections detail the structural theory underlying the analysis and the path diagram used for estimation of the relationships.

**Defining a Structural Theory** The HBAT research team proposes a theory based on the organizational literature and the collective experience of key HBAT management employees. They agree it is impossible to include all the constructs that might possibly relate to employee retention (staying intentions). It would be too costly and too demanding on the respondents based on the large number of survey items to be completed. Thus, the study is conducted with the five constructs listed previously.

The theory leads HBAT to expect that EP, AC, JS, and OC are all related to SI, but in different ways. For example, a high EP score means that employees believe their work environment is comfortable and allows them to freely conduct their work. This environment is likely to create high job satisfaction, which in turn will facilitate indirect links between EP and SI and between AC and SI. Thus, the theory suggests a process of effects. Because it would require an extensive presentation of key organizational concepts and findings, we will not develop the theory in detail here.

The theory developed by HBAT implies the following piecemeal hypotheses:

- $H_1$ : Environmental perceptions are positively related to job satisfaction.
- $H_2$ : Environmental perceptions are positively related to organizational commitment.
- $H_3$ : Attitudes toward coworkers are positively related to job satisfaction.
- $H_4$ : Attitudes toward coworkers are positively related to organizational commitment.
- $H_5$ : Job satisfaction is related positively to organizational commitment.
- $H_6$ : Job satisfaction is related positively to staying intentions.
- $H_7$ : Organizational commitment is related positively to staying intentions.

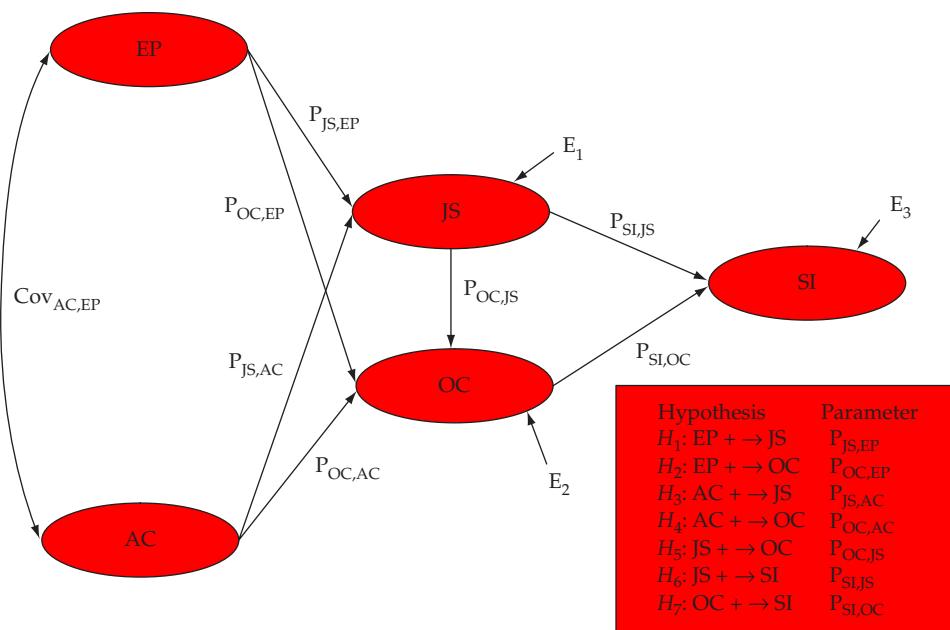
**Visual Diagram** The theory can be expressed visually. Figure 11.6 shows the diagram corresponding to this theory. For simplicity, the measured indicator variables and their corresponding paths and errors have been left off the diagram. If a graphical interface is used with a SEM program, then all measured variables and error variance terms are shown on the path diagram.

**EXOGENOUS CONSTRUCTS** EP and AC are exogenous constructs in this model. The exogenous constructs are considered to be determined by things outside of this model. In practical terms, this means that no hypothesis predicts either of these constructs. Like independent variables in regression, they are used only to predict other constructs.

The two exogenous constructs—EP and AC—are drawn at the far left. No single-headed arrows enter the exogenous constructs. A curved two-headed arrow is included to capture any covariance between these two constructs ( $\text{Cov}_{\text{AC},\text{EP}}$ ). We maintain the CFA convention of assuming exogenous constructs are related unless we have a reason not to think so.

**ENDOGENOUS CONSTRUCTS** JS, OC, and SI are endogenous constructs in this model. Each is determined by constructs included in the model. Notice that both JS and OC serve both as outcomes and as predictors. This is perfectly

**Figure 11.6**  
The HBAT Employee Retention Model



acceptable in SEM, and all hypothesized effects are produced with one structural model test. This would not be possible with a single OLS regression model because we would be limited to a single dependent variable.

The structural path model begins to develop from the exogenous constructs. A path should connect any two constructs linked theoretically by a hypothesis. Therefore, after drawing the three endogenous constructs (JS, OC, and SI), single-headed arrows are placed connecting the predictor (exogenous) constructs with their respective outcomes based on the hypotheses. The legend in the bottom right of Figure 11.6 lists each individual hypothesized relationship and the path to which it belongs. Each single-headed arrow represents a direct path and is labeled with the appropriate parameter estimate. For example,  $H_2$  proposes a positive EP–OC relationship. A parameter estimate linking an exogenous construct to an endogenous construct is designated by the symbol P. The convention is that the subscript first lists the number (or abbreviation) of the construct to which the path points and then the subscript for the construct from which the path begins. So,  $H_1$  is represented by  $P_{JS,EP}$ . Similarly then,  $H_7$ , linking SI with OC, is represented by  $P_{SI,OC}$ .

As discussed earlier, the CFA model must be transformed into a structural model within the software programs. Although issues are specific to each program, the user must essentially redefine construct types (exogenous to endogenous), replace the correlational paths with the structural relationships, and change the notation associated with these changes. Materials on the text's online resources describe this process in more detail.

## STAGE 6: ASSESSING THE STRUCTURAL MODEL VALIDITY

The structural model shown in the path diagram in Figure 11.6 can now be estimated and assessed. To do so, the emphasis first will be on SEM model fit and then whether the structural relationships are consistent with theoretical expectations.

The information in Table 11.1 shows the overall fit statistics from testing the Employee Retention model. The  $\chi^2$  is 283.4 with 181 degrees of freedom ( $p < .05$ ), which means the normed chi-square is 1.57. The model CFI is .99 with a RMSEA of .036, which corresponds to a 90 percent confidence interval of .027 to .045. Although 0 is not in the confidence interval, the RMSEA value is relatively low. All measures are within a range that would be associated with

**Table 11.1** Comparison of Goodness-of-Fit Measures Between HBAT Employee Retention and CFA Models

GOF Index	Employee Retention Model	CFA Model
<b>Absolute Measures</b>		
$\chi^2$ (chi-square)	283.43	236.62
Degrees of freedom	181	179
Probability	0.00	0.00
GFI	.94	.95
RMSEA	.036	.027
Confidence interval of RMSEA	.027–.045	.015–.036
RMR	.110	.085
SRMR	.060	.035
Normed chi-square	1.57	1.32
<b>Incremental Fit Measures</b>		
NFI	.96	.97
NNFI	.98	.99
CFI	.99	.99
RFI	.96	.97
<b>Parsimony Measures</b>		
AGFI	.92	.93
PNFI	.83	.83

good fit (see Chapter 9 for a review of fit guidelines). We can also see that overall model fit changed a little from the CFA model (Table 11.1). The only mark on an otherwise good fit is a chi-square increase of 46.8 and a difference of two degrees of freedom ( $p < .05$ ) in moving from the CFA to the structural model. While the researchers conclude the model has validity, the difference in chi-squares are kept in mind as they move forward. The standardized path coefficients are shown in Figure 11.7.<sup>1</sup>

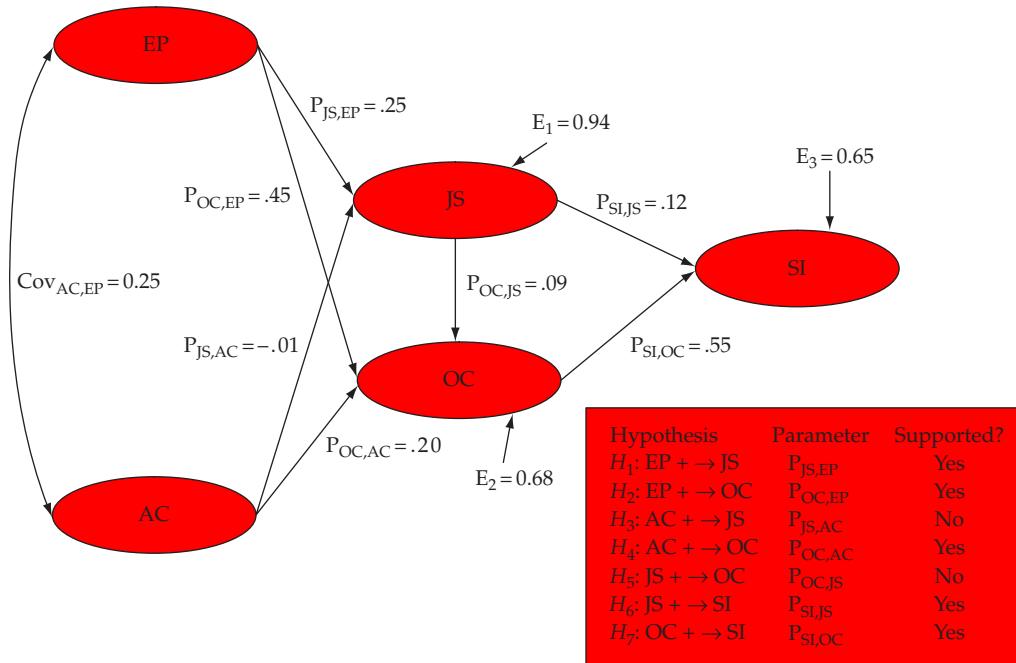
We next examine the loading estimates to make sure they have not changed substantially from the CFA model (see Table 11.2). The loading estimates are virtually unchanged from the CFA results. Only three estimated standardized loadings change and the maximum change is .01. Thus, if parameter stability had not already been tested in the CFA stage, there is now evidence of stability among the measured indicator variables. In technical terms, this indicates no problem is evident due to interpretational confounding and further supports the measurement model's validity. As we would expect with so little change in loadings, the construct reliabilities are identical as well.

Validation of the model is not complete without examining the individual parameter estimates. Are they statistically significant and practically meaningful? These questions are addressed along with assessing model fit.

Table 11.3 shows the estimated unstandardized and standardized structural path estimates. All but two structural path estimates are significant and in the expected direction. The exceptions are the estimates between AC and JS and between JS and OC. Both estimates have significance below the critical  $t$ -value for a Type I error of .05. Therefore, although the estimate is in the hypothesized direction, it is not supported. Overall, however, given that five of seven estimates are consistent with individual hypotheses, these results support the theoretical model, with a caveat for the two insignificant paths.

One final comparison between the employee retention model and the CFA model is in terms of the structural model estimates. Table 11.4 contains the standardized parameter estimates for all seven of the structural relationships as well as the correlational relationship among EP and AC. As noted earlier, five of these seven relationships were supported with significant path estimates. But what about the relationships not in the hypothesized model? We will

**Figure 11.7**  
Standardized Path Estimates for the HBAT Structural Model



<sup>1</sup> Recently, the American Psychological Association helped develop a set of standards for presenting quantitative research results, including results from techniques like multiple regression and SEM. See: Appelbaum, M, R.B. Kline, A. Nezu, H. Cooper, E. Mayo-Wilson, and S.M. Rao (2018), "Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report," *American Psychologist*, 73, 3–25.

**Table 11.2** Comparison of Standardized Factor Loadings and Construct Reliabilities for HBAT Employee Retention and CFA Models

Indicator	Construct	Employee Retention Model	CFA MODEL
<b>Standardized Factor Loading</b>			
JS <sub>1</sub>	JS	0.74	0.74
JS <sub>2</sub>	JS	0.75	0.75
JS <sub>3</sub>	JS	0.68	0.68
JS <sub>4</sub>	JS	0.70	0.70
JS <sub>5</sub>	JS	0.73	0.73
OC <sub>1</sub>	OC	0.58	0.58
OC <sub>2</sub>	OC	0.88	0.88
OC <sub>3</sub>	OC	0.66	0.66
OC <sub>4</sub>	OC	0.83	0.84
SI <sub>1</sub>	SI	0.81	0.81
SI <sub>2</sub>	SI	0.87	0.86
SI <sub>3</sub>	SI	0.74	0.74
SI <sub>4</sub>	SI	0.85	0.85
EP <sub>1</sub>	EP	0.69	0.70
EP <sub>2</sub>	EP	0.81	0.81
EP <sub>3</sub>	EP	0.77	0.77
EP <sub>4</sub>	EP	0.82	0.82
AC <sub>1</sub>	AC	0.82	0.82
AC <sub>2</sub>	AC	0.82	0.82
AC <sub>3</sub>	AC	0.84	0.84
AC <sub>4</sub>	AC	0.82	0.82
<b>Construct Reliabilities</b>			
	JS	0.84	0.84
	OC	0.83	0.83
	SI	0.89	0.89
	EP	0.86	0.86
	AC	0.89	0.89

**Table 11.3** Structural Parameter Estimates for HBAT Employee Retention Model

Structural Relationship	Unstandardized Parameter Estimate	Standard Error	t-value	Standardized Parameter Estimate
H <sub>1</sub> : EP → JS	0.20	0.05	4.02	0.25
H <sub>2</sub> : EP → OC	0.52	0.08	6.65	0.45
H <sub>3</sub> : AC → JS	-0.01	0.05	-0.17	-0.01
H <sub>4</sub> : AC → OC	0.26	0.07	3.76	0.20
H <sub>5</sub> : JS → OC	0.13	0.08	1.60	0.09
H <sub>6</sub> : JS → SI	0.09	0.04	2.38	0.12
H <sub>7</sub> : OC → SI	0.27	0.03	8.26	0.55
EP correlated AC	0.37	0.09	4.19	0.25

**Table 11.4** Comparison of Structural Relationships with CFA Correlational Relationships

HBAT Employee Retention Model		HBAT CFA Model	
Structural Relationship	Standardized Parameter Estimate	Comparable Correlational Relationship	Standardized Correlation Estimate
$H_1: EP \rightarrow JS$	0.25	EP correlated JS	0.24
$H_2: EP \rightarrow OC$	0.45	EP correlated OC	0.50
$H_3: AC \rightarrow JS$	-0.01	AC correlated JS	0.05
$H_4: AC \rightarrow OC$	0.20	AC correlated OC	0.30
$H_5: JS \rightarrow OC$	0.09	JS correlated OC	0.21
$H_6: JS \rightarrow SI$	0.12	JS correlated SI	0.23
$H_7: OC \rightarrow SI$	0.55	OC correlated SI	0.55
EP correlated AC	0.25	EP correlated AC	0.25
Not estimated	—	EP correlated SI	0.56
Not estimated	—	AC correlated SI	0.31

examine model diagnostics in the next section, but we can also compare the correlational relationships from the CFA model with the structural relationships. As we see in Table 11.4, the estimated parameters are quite comparable between the two models. But we also see that the two excluded structural relationships ( $EP \rightarrow SI$  and  $AC \rightarrow SI$ ) correspond to significant relationships in the CFA model. This might suggest that model performance would improve with the addition of one or more of these relationships.

**Examining Model Diagnostics** As discussed earlier, several diagnostic measures are available for researchers to evaluate SEM models. They range from fit indices to standardized residuals and modification indices. Each of these will be examined in the following discussion to determine if model respecification should be considered.

The first comparison in fit statistics is the chi-square difference between the hypothesized model and the measurement model where we see a  $\Delta\chi^2$  of 46.8 with two degrees of freedom ( $p < .001$ ). The difference in degrees of freedom is two, which is due to the fact that the model includes all but two of the possible structural relationships. The possibility of another meaningful structural path should be explored, particularly if other diagnostic information points in that direction. All the other fit statistics are also supportive of the model, and there were no substantive changes in the other fit indices between the CFA and structural model.

The next step is to examine the standardized residuals and modification indices for the structural model. As before, patterns of large standardized residuals and/or large modification indices indicate changes in the structural model that may lead to model improvement. Table 11.5 contains the standardized residuals greater than  $|2.5|$ . In looking for patterns of residuals for a variable or set of variables, one pattern is obvious: each item of the EP construct ( $EP_1$  to  $EP_4$ ) has significant standardized residual with at least three of the four items in the SI construct. This indicates that there may be a substantial relationship omitted between these two constructs. At the moment, there is no direct relationship between these two constructs, only indirect relationships ( $EP \rightarrow JS \rightarrow SI$  and  $EP \rightarrow JS \rightarrow OC \rightarrow SI$ ).

Examination of the two modification indices for the direct paths of  $EP \rightarrow SI$  and  $AC \rightarrow SI$  shows that both have values over 4.0, although the  $EP \rightarrow SI$  value is much higher (40.12 versus 8.98). This strongly supports the addition of the  $EP \rightarrow SI$  relationship if it can be supported theoretically. This also corresponds to the pattern of residuals described between the indicators of these two constructs. It also casts doubt on the premise that JS mediates the relationship between EP and SI.

Note that modification indices can be estimated for the “second half” of the recursive relationships between  $SI \rightarrow JS$  and  $SI \rightarrow OC$ . As we can see, both indicate substantial improvement in model fit. But more important,

**Table 11.5** Model Diagnostics for HBAT Employee Retention Model

Standardized Residuals (All Residuals Greater Than  2.5 )			
Largest Negative Standardized Residuals			
SI <sub>2</sub>	and	OC <sub>1</sub>	-2.90
SI <sub>3</sub>	and	OC <sub>1</sub>	-2.88
SI <sub>4</sub>	and	OC <sub>1</sub>	-2.99
EP <sub>3</sub>	and	OC <sub>1</sub>	-2.90
Largest Positive Standardized Residuals			
SI <sub>2</sub>	and	SI <sub>1</sub>	3.45
SI <sub>4</sub>	and	SI <sub>3</sub>	3.47
EP <sub>1</sub>	and	SI <sub>1</sub>	3.78
EP <sub>1</sub>	and	SI <sub>2</sub>	4.27
EP <sub>1</sub>	and	SI <sub>3</sub>	3.78
EP <sub>1</sub>	and	SI <sub>4</sub>	4.50
EP <sub>2</sub>	and	OC <sub>3</sub>	2.63
EP <sub>2</sub>	and	SI <sub>1</sub>	3.41
EP <sub>2</sub>	and	SI <sub>2</sub>	4.20
EP <sub>2</sub>	and	SI <sub>3</sub>	3.70
EP <sub>2</sub>	and	SI <sub>4</sub>	5.84
EP <sub>3</sub>	and	SI <sub>2</sub>	2.69
EP <sub>3</sub>	and	SI <sub>3</sub>	2.73
EP <sub>3</sub>	and	SI <sub>4</sub>	3.76
EP <sub>4</sub>	and	SI <sub>1</sub>	4.52
EP <sub>4</sub>	and	SI <sub>2</sub>	3.60
EP <sub>4</sub>	and	SI <sub>3</sub>	3.97
EP <sub>4</sub>	and	SI <sub>4</sub>	4.27
EP <sub>4</sub>	and	EP <sub>3</sub>	3.01
AC <sub>3</sub>	and	SI <sub>4</sub>	2.73
AC <sub>4</sub>	and	SI <sub>4</sub>	2.95
Modification Indices for Structural Relationships			
Structural Relationship (Not Estimated)		Modification Index	
EP → SI		40.12	
AC → SI		8.98	
SI → JS		38.66	
SI → OC		45.12	

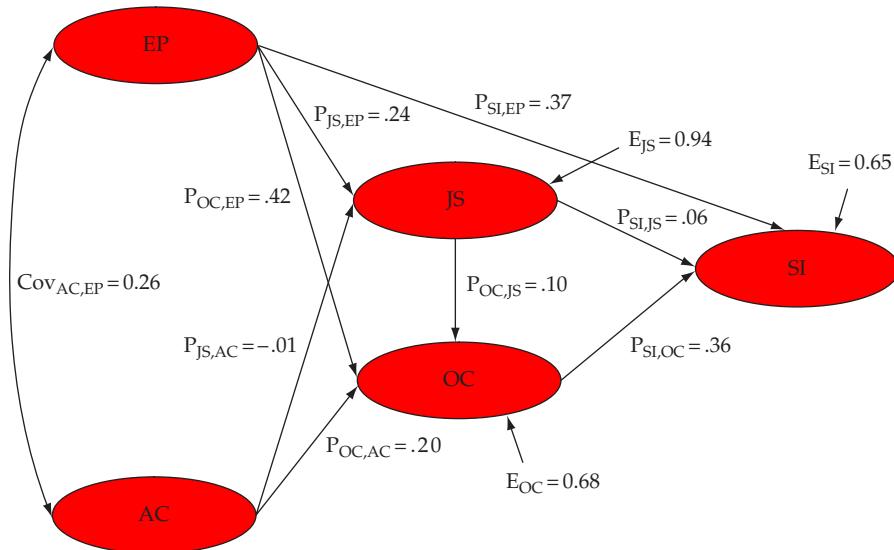
because they have no theoretical basis for inclusion in the model, they highlight the potential dangers of making model respecifications based solely on improvement of model fit without regard to a theoretical basis.

The researcher must evaluate not only the information provided by the model fit measures and other diagnostics but determine (a) the level of theoretical support provided by the results and (2) any potential model respecifications that would provide improvement in the model while also having theoretical support.

**Model Respecification** Many times the diagnostic measures in SEM indicate model respecification should be considered. Any respecification must have strong theoretical as well as empirical support of the nature that would allow an *a priori* alternative model. Model respecification should not be the result of searching for relationships,

but rather for improving model fit that is theoretically justified. Based on the residuals and modification indices information from the initial SEM model, we examine a respecification of our HBAT example.

To further assess the SEM model, the HBAT research team conducts a post hoc analysis adding a direct relationship between EP and SI. The SEM program is instructed to free this path, and the model is re-estimated. Figure 11.8 shows the model including a free path corresponding to EP → SI ( $P_{SI,EP}$ ). Table 11.6 compares the



**Figure 11.8**  
Standardized Path  
Estimates for the Revised  
HBAT Structural Model

**TABLE 11.6** Comparison of Goodness-of-Fit Measures Between HBAT Employee Retention and Revised Employee Retention Models

GOF Index	Revised Employee Retention Model	Employee Retention Model
<b>Absolute Measures</b>		
$\chi^2$ (chi-square)	242.23	283.43
Degrees of freedom	180	181
Probability	0.00	0.00
GFI	.95	.94
RMSEA	.029	.036
Confidence interval of RMSEA	.018–.038	.027–.045
RMR	.090	.110
SRMR	.040	.060
Normed chi-square	1.346	1.57
<b>Incremental Fit Measures</b>		
NFI	.97	.96
NNFI	.99	.98
CFI	.99	.99
RFI	.96	.96
<b>Parsimony Measures</b>		
AGFI	.93	.92
PNFI	.83	.83

**TABLE 11.7** Comparison of Structural Relationships for the Original and Revised HBAT Employee Retention Models

HBAT Employee Retention Model		Revised HBAT Employee Retention Model	
Structural Relationship	Standardized Parameter Estimate	Structural Relationship	Standardized Parameter Estimate
$H_1: EP \rightarrow JS$	0.25*	$H_1: EP \rightarrow JS$	0.24*
$H_2: EP \rightarrow OC$	0.45*	$H_2: EP \rightarrow OC$	0.42*
$H_3: AC \rightarrow JS$	-0.01	$H_3: AC \rightarrow JS$	-0.01
$H_4: AC \rightarrow OC$	0.20*	$H_4: AC \rightarrow OC$	0.20*
$H_5: JS \rightarrow OC$	0.09	$H_5: JS \rightarrow OC$	0.10
$H_6: JS \rightarrow SI$	0.12*	$H_6: JS \rightarrow SI$	0.06
$H_7: OC \rightarrow SI$	0.55*	$H_7: OC \rightarrow SI$	0.36*
EP correlated AC	0.25*	EP correlated AC	0.26*
		EP $\rightarrow$ SI	0.37*

\*Statistically significant at .05 level.

GOF measures for the “original” and revised models. The resulting standardized parameter estimate for  $P_{SI,EP}$  is 0.37 ( $p < .001$ ). In addition, the overall fit reveals a  $\chi^2$  value of 242.2 with 180 degrees of freedom and a normed  $\chi^2$  value of 1.346. The CFI remains .99, and the RMSEA is .029, which is practically the same as the value for the CFA model. This is a better fit than the original structural model because the  $\Delta\chi^2$  is 41.2 with one degree of freedom, which is significant ( $p < .001$ ). In this case, the respecification casts doubt on the role of JS and OC in mediating the effects of EP.

Several of the path estimates from the original model have changed slightly, as would be expected (see Table 11.7). Most notably, the JS–SI relationship ( $P_{SI,JS} = .06$ ) is no longer significant, and the SI–OC relationship ( $P_{SI,OC} = .36$ ) remains significant but is substantially smaller than before.

The  $\Delta\chi^2$  value between the revised SEM and CFA models is 5.61 with one degree of freedom, which is significant at a type I error rate of .05. The squared multiple correlation (i.e.,  $R^2$ ) for SI also improves from .35 to .45 with the addition of this relationship. These findings suggest that the structural model does a good, but not perfect, job in explaining the observed covariance matrix. Thus, we can proceed to interpret the precise nature of the relationships with a fair degree of confidence.

At this point, HBAT has tested its original structural model. The results showed reasonably good overall model fit and the hypothesized relationships were generally supported. However, the large difference in fit between the structural model and CFA model and several key diagnostics, including the standardized residuals, suggested one improvement to the model. This change improved the model fit. Now, HBAT must consider testing this model with new data to examine its generalizability.

A complete SEM analysis involves both the test of a measurement theory and the structural theory that links constructs together in a logically meaningful way. In this chapter, we learned how to complete the analysis by extending our CFA model in a way that enabled a test of the overall structural model, which includes the set of relationships showing how constructs are related to one another. SEM is not just another multivariate statistical procedure. It is a way of testing theory. Much easier and more appropriate statistical tools are available for exploring relationships. However, when a researcher becomes knowledgeable enough about a subject matter to specify a set of relationships between constructs, in addition to the way these constructs are measured, SEM is an appropriate and powerful tool. This chapter highlights several key points associated with SEM, including the following:

**Distinguish a measurement model from a structural model.** The key difference between a measurement model and a structural model is the way the relationships between constructs are treated. In CFA, a measurement model is tested that usually assumes each construct is related to each other construct. No distinction is made between exogenous and endogenous constructs, and the relationships are represented as simple correlations with a two-headed curved arrow. In the structural model, endogenous constructs are distinguished from exogenous constructs. Exogenous constructs have no arrows entering them. Endogenous constructs are determined by other constructs in the model as indicated visually by the pattern of single-headed arrows that point to endogenous constructs.

**Describe the similarities between SEM and other multivariate techniques.** Although CFA has much in common with EFA, the structural portion of SEM is similar to multiple regression. The key differences lie in the fact that the focus is generally on how constructs relate to one another instead of how variables relate to one another. Also, it is quite possible for one endogenous construct to be used as a predictor of another endogenous construct within the SEM model.

**Depict a theoretical model with dependence relationships using a path diagram.** The chapter described procedures for converting a CFA path diagram into a structural path diagram. In a path diagram, the relationships between constructs are represented with single-headed arrows. Also, the common abbreviations change. Measured indicator items for endogenous constructs are generally referred to with a *Y*, whereas the exogenous construct indicators are referred to with an *X*.

**Test a structural model using SEM.** The CFA setup can be modified and the structural model tested using the same SEM program. Models are supported to a greater extent as the fit statistics suggest that the observed covariances are reproduced adequately by the model. The same guidelines that apply to CFA models apply to the structural model fit. Also, the closer the structural model fit is to the CFA model fit, then the more confidence the researcher can have in the model. Finally, the researcher also must examine the statistical significance and direction of the relationships. The model is supported to the extent that the parameter estimates are consistent with the hypotheses that represented them prior to testing.

**Diagnose problems with the SEM results.** The same diagnostic information can be used for structural model fit as for CFA model fit. The statistical significance, or lack thereof, of key relationships, the standardized residuals, and modification indices all can be used to identify problems with a SEM model.

In what ways is a measurement theory different from a structural theory? What implications do these differences have for the way a SEM model is tested? How does the visual diagram for a measurement model differ from that of a SEM model?

How can a measured variable represented with a single item be incorporated into a SEM model?

What is the distinguishing characteristic of a non-recursive SEM model?

How is the validity of a SEM model estimated?

Why is it important to examine the results of a measurement model before proceeding to test a structural model?

Comment on the difference in orientation between testing “a theory” versus testing “hypotheses.”

Can single items be used in SEM? Explain your answer.

Using the HBAT6SEM data set, test a structural model that adds the single item SP1 as an additional work environment variable to represent supervisor characteristics (in addition to AC and EP) in trying to explain why employees continue to work at HBAT.

A list of suggested readings and other materials relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com)).

Chapter 11 introduced the possibility of single item measures within a SEM. Although clearly most SEM applications in behavioral sciences involve multiple-item assessments of latent constructs, the user may benefit from an illustration involving a mixture of both single item and multiple item constructs. The illustration provided here uses the HBATSEM6CON.sav data.

In this case, researchers are interested in explaining the job performance of HBAT sales personnel as a function of job experience. In this case, a single job performance variable is included in the data and represents the employee's previous annual performance rating on a 1 to 5 scale. The data include both the age of the employee and the number of years of experience, as well as the employee's gender (dummy coded as 0=male and 1=female). All of those variables are taken from HR records and are presumed to have no error. The researcher is impressed by a management theory suggesting that job satisfaction facilitates the relationship between experience and job performance. The theory also suggests that older employees tend to be more satisfied. Thus, the hypothesized model can be depicted as:

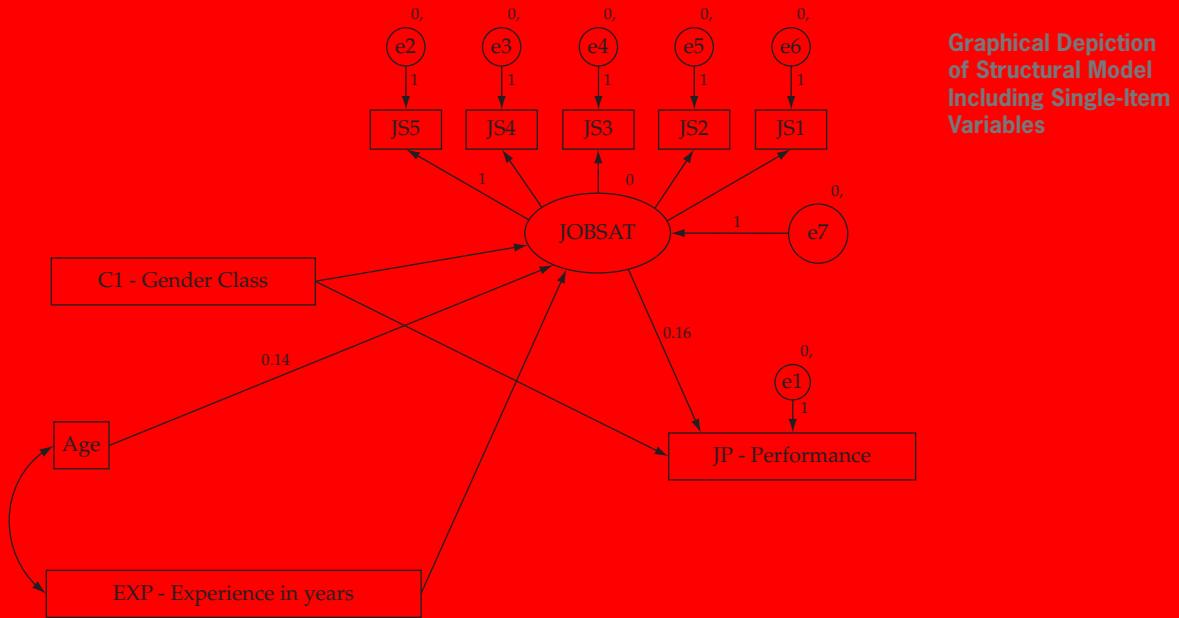


In addition, the researcher includes employee gender as a control variable. Typically, control variables are not shown in formal diagrams of the theory but they must be accounted for in the actual structural equation model.

The researcher decides to use an SEM program's graphical interface to test the model. In addition to the structural paths, age and experience should logically be correlated, and thus a two-headed arrow is included between them. In contrast, the researcher has no reason to suspect that the control variable, gender, is related to either age or experience. Appendix Figure A11.1 depicts the way the model looks in the SEM program (in this case, AMOS).

The researcher estimates the model. The model yields a  $\chi^2$  of 25.9 with 25 degrees of freedom ( $p = 0.411$ ). Thus, the model reproduces the sample covariance matrix within sampling error. Not surprisingly then, the CFI is very high at 0.999 and the RMSEA is very low .010. The 90 percent confidence interval for RMSEA includes 0, which suggests the possibility of 0 error cannot be rejected. All indicators clearly suggest good fit for the model.

Gender, the control variable, does not produce a significant path coefficient with either job satisfaction or job performance. The two significant structural path estimates are for age → job satisfaction (0.14) and job satisfaction



to job performance (0.16). Interestingly, or perhaps not so interestingly, experience has no significant effect on job satisfaction (0.05) and thus does not indirectly influence job performance. No residuals suggest any improvement in the model, thus the research concludes the results are valid. After talking it over with colleagues, the researcher learns that sometimes, experience does not affect performance or satisfaction because employees sometimes become stuck in a job, not performing well enough to advance and not having a lot of alternatives. Thus, HBAT researchers are back to work theorizing a new model that includes the notion of being stuck in a job.

The simple example illustrates that including single-item variables, and in particular including control variables, is a straightforward process. If the control variables turn out not to influence any of the other relationships in a model, they can be deleted from the final theoretical model.

- 1 Anderson, J. C., and D. W. Gerbing. 1988. Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin* 103: 411–23.
- 2 Anderson, J. C., and D. W. Gerbing. 1992. Assumptions and Comparative Strengths of the Two-Step Approach. *Sociological Methods and Research* 20: 321–33.
- 3 Fornell, C., and Y. Yi. 1992. Assumptions of the Two-Step Approach to Latent Variable Modeling. *Sociological Methods and Research* 20: 291–320.
- 4 Hair, J. F., B. J. Babin, and N. Krey. 2017. Covariance-Based Structural Equation Modeling in the Journal of Advertising: Review and Recommendations. *Journal of Advertising* 46: 163–77.
- 5 Peterson, S. J., F. O. Walumbwa, K. Byron, and J. Myrowitz. 2009. CEO Positive Psychological Traits, Transformational Leadership, and Firm Performance in High-Technology Start-Up and Established Firms. *Journal of Management* 35: 348–68.
- 6 Zattoni, A., L. Gnan, and M. Huse. 2015. Does Family Involvement Influence Firm Performance? Exploring the Mediating Effects of Board Processes and Tasks. *Journal of Management* 41: 1214–43.

# 12 Advanced SEM Topics

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

- Understand the differences between reflective and formative measurement.
- Know how to specify formative scales in SEM models.
- Identify when higher-order factor analysis models may be appropriate.
- Use multigroup methods to perform an invariance measurement analysis.
- Understand the concepts of statistical mediation and moderation.
- Be aware of current developments in SEM.

## Chapter Preview

Chapters 9, 10, and 11 introduced the basic fundamentals of covariance-based structural equation modeling and its two most basic applications—confirmatory factor analysis and estimation of a structural model. This chapter extends the discussion to several more advanced topics faced by many researchers today. We first examine the use of formative rather than reflective measurement theory. As will be discussed, not only are estimation issues involved in the use of formative scales, but there are also questions involving their appropriateness in structural equation modeling. We then discuss the applicability and use of higher-order factor models. In these instances, the latent constructs we estimate with measured variables now act as “indicators” of a higher-order latent construct. In a sense, we represent a latent construct with other latent constructs. We identify the situations in which this approach is most applicable and discuss the issues of estimation and interpretation. We then focus on multigroup models, a form of SEM analysis comparing the same model across multiple samples of respondents. A specific type of multigroup analysis, measurement invariance testing, is discussed in detail to clarify its objective of assessing the measurement model’s equivalence across groups. This is the foundation for making between-group comparisons, which are particularly relevant in cross-cultural research.

We then shift focus to the structural model, presenting two common types of relationships—mediation and moderation. Each is discussed in terms of how it is estimated in a SEM model. We also present the underlying logic of each type of relationship so the researcher can select the appropriate type. The final section discusses alternative methods of estimating structural models including Bayesian SEM.

## Key Terms

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

**Bayesian statistics** An alternative to traditional frequentist statistical approach built around the idea of conditional probabilities such that parameters are not sample dependent.

**Between-group constraints** Fixing a relationship to be equivalent across two or more group models. A single estimate is made for all groups rather than a unique estimate in each group.

**Bootstrapping** Produces  $t$  values for parameter estimates through a non-parametric process involving reestimation of the model hundreds or thousands of times by sampling with replacement from the original sample to produce each iteration's input.

**Chi-square difference ( $\Delta\chi^2$ )** Measure for assessing the statistical significance of the difference in overall model fit between two models based on the relative chi-square values of each model. A nonsignificant value indicates that the two models provide the same level of model fit and can be considered equivalent in terms of explanation. The degrees of freedom for the chi-square represent the difference in the number of estimated parameters in the two models.

**Complete mediation** See *full mediation*.

**Configural invariance** Exists when an acceptable fit is obtained from a *multisample CFA* model that simultaneously estimates a factor solution for all groups with each group configured with the same structure (same pattern of free and fixed parameters). Also see *totally free multiple group model*.

**Common methods bias (CMB)** Relationships among variables and/or constructs are influenced by the data collection method (e.g., same collection method, questionnaire format, or even scale type)—a result of common methods variance (CMV) and sometimes referred to as constant methods bias or monomethods bias.

**Competence interval** The Bayesian cousin to confidence intervals that reveal, typically with 95 percent probability, the lower and upper bounds for a mean parameter estimate value given a sample. In other words, the expected value distribution for a parameter given the data.

**Cross-validation** Attempt to reproduce the results found in one sample using data from a different sample, usually drawn from the same population.

**Direct effect** Relationship linking two constructs with a single arrow between the two.

**Error term invariance** Occurs when the error variance terms for each measured variable are equal across the groups being studied. Achieving error term invariance denotes equal reliabilities for the constructs across groups.

**Factor covariance invariance** An advanced stage of the *measurement invariance* process where the covariances between constructs are the same across groups.

**First-order factor model** Covariances between measured variables explained with a single latent factor layer. See also *second-order factor model*, which has two layers of latent factors.

**Formative measurement theory** Theory presenting formative scales as operational factors (see Chapter 9).

**Full invariance** Achieved when the *chi-square difference* test is nonsignificant for the complete set of constraints when testing for *measurement invariance* in MCFA.

**Full mediation** A sequence relationship proposing that an independent variable causes an intermediate variable, which in turn causes an outcome variable ( $X \rightarrow M \rightarrow Y$ ). In other words, the relationship between a predictor and outcome depends on an intermediate variable.

**Indirect effect** Sequence of relationships with at least one intervening construct involved. That is, a sequence of two or more sequential (implied causal) *direct effects*.

**Measurement equivalence** See *measurement invariance*.

**Measurement invariance** *Measurement theory* condition in which the measures forming a measurement model have the same meaning and are used in the same way by different groups of respondents. Tested through a series of increasingly rigorous MCFA models where *between-group constraints* restrict different elements of the measurement model (e.g., *metric invariance* tests for equivalence of the factor loading estimates).

**Measurement theory** Series of relationships that suggest how measured variables represent a construct not measured directly (latent). A measurement theory can be represented by a series of regression-like equations mathematically relating a factor (construct) to the measured variables.

**Mediating effect** Effect of a third variable/construct intervening between two other related constructs.

**Metric invariance** An important stage in the *measurement invariance* process that assesses the extent to which factor loading estimates are equivalent across groups. Metric invariance provides support that respondents use the numeric rating scales similarly across groups so the differences in relationships between groups can be compared.

**Moderating effect** Effect of a third variable or construct changing the relationship between two related variables/constructs. That is, the relationship between two variables changes based on the level/amount of a moderator. For example, if a relationship changes significantly when measured for males versus females, then sex moderates the relationship.

**Multiple group analysis** A form of SEM analysis where two or more samples of respondents are compared using similar models. *Between-group constraints* are used to assess the similarities between groups on any model parameter(s).

**Multisample confirmatory factor analysis (MCFA)** A form of *multiple group analysis* where multiple CFA models, one for each group of respondents, are estimated and then measures of fit calculated for all the models collectively.

**Nuisance factor** An external effect to the SEM model that may impact the results in some fashion and thus needs to be accounted for. Examples can be types of questions used in the questionnaire or differing conditions at various times of data collection. For an example, see *common methods bias*.

**Partial invariance** When the *chi-square difference* indicates that only a subset of possible *between-groups constraints* (at least two per construct) are nonsignificant when testing *measurement invariance* in MCFA.

**Partial mediation** Effect when a direct relationship between a predictor and an outcome is reduced but remains significant when a mediator is also entered as an additional predictor.

**Reflective measurement theory** Theory presenting reflecting scales as factors (see Chapter 10). It is the typical psychometric representation for a psychological latent construct.

**Scalar invariance** A stage in the *measurement invariance* process following *metric invariance*, which assesses the extent to which intercept terms for the equations explaining measured variables are equivalent across groups in a MCFA. Scalar invariance supports valid comparison of the latent construct means between groups.

**Second-order factor model** *Measurement theory* involving two “layers” of latent constructs. These models introduce a second-order latent factor(s) that causes multiple *first-order latent factors*, which, in turn, cause the measured variables ( $x$ ).

**TF** See *totally free multiple group model*.

**Totally free multiple group model (TF)** Model that uses the same structure (pattern of fixed and free parameters) on all groups but allows parameters to take on different values (no equality constraints between groups) in a *multiple group analysis*.

## Reflective Versus Formative Scales

The issue of causality (i.e., correlational versus dependence relationships) has played a key role in our specification of the structural model. Traditional psychometric theory relies on reflective measurement where the **measurement theory** presumes latent factors (constructs) cause measured variables. Can the direction of causality be reversed? Can a measured variable cause a factor? This contrasting direction of causality leads to a different measurement approach known as *formative measurement models*.

### REFLECTIVE VERSUS FORMATIVE MEASUREMENT THEORY

A **reflective measurement theory** is based on the idea that latent constructs cause the measured variables and that the error in measurement results in an inability of the construct to fully explain individual measured variables. Thus, the direction of the arrows is from latent constructs to measured variables, and error terms are associated with each measured variable. As such, reflective measures are consistent with psychometrics and classical test theory [45]. Construct validity of a reflective latent construct ensures that the “meaning” of the factor will remain consistent given the measures used and it should not vary when associated with other constructs.

Because a reflective measure presumes that all indicator items are caused by the same latent construct, items within a construct should be highly correlated with each other. Individual items should be interchangeable and any single item can be left out without changing the construct’s meaning. The scale measurement must demonstrate sufficient construct validity and at least three items must be specified to avoid identification problems [9]. Reflective indicators can be viewed as a sample of all the possible items available within the conceptual domain of the construct [18]. As a consequence, reflective indicators of a given construct are expected to move together, meaning that changes in one indicator are associated with proportional changes in the other indicators.

Reflective measurement models are the predominant measurement theory used in the social sciences [8]. Typical social science constructs such as personality, other individual traits, work-related traits, and behavioral intentions, to name a few, represent reflective measurement [7, 9]. Likewise, a study of symptoms typically would be reflective. For example, symptoms such as shortness of breath, tiring easily, wheezing, and reduced lung functioning would be considered indicators that would reflect the latent factor of emphysema. The symptoms *do not cause* the disease. Rather, the disease causes the symptoms.

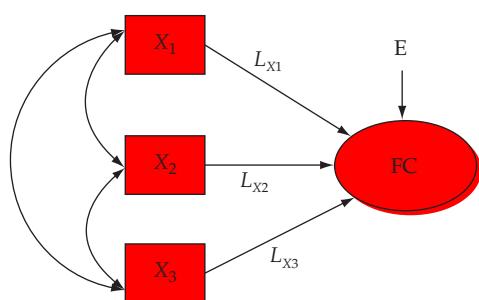
In contrast, **formative measurement** implies that measured variables can cause the factor. A typical example would be a social class index [21]. Social class often is viewed as a composite of one's educational level, occupational prestige, and income (or sometimes wealth). Social class does not cause these indicators as in the reflective case. Rather, each formative indicator is considered a partial cause of the index. In a business setting, investors often compute a bankruptcy index indicating how close an individual or company is to financial bankruptcy. Key financial measures (e.g., retained earnings, working capital, equity, and sales to assets, among others) could be thought of as causing bankruptcy, and thus they would be appropriate as formative indicators. Finally, using a health-related example, a formative emphysema index might specify indicators such as cigarette consumption, exposure to toxins, chronic bronchitis, and others. These indicators would *form*, rather than reflect, the probability of an individual having emphysema. The fact that one smokes cigarettes has little connection to the other indicators. Under formative measurement theory, the measures can not share commonality.

Although a formative scale may seem quite simple—just reverse the arrows—it also reverses the way we think about scales. A key assumption is that formative factors are not considered latent, and thus the indicators need not have a consistent inherent meaning [32]. Formative factors are better viewed as indices or composites where each indicator is a potential contributing cause. One issue with formative measurement is that factor loading estimates are determined by other relationships in the model. This is because when a formative measure requires at least two separate reflective items or other endogenous constructs to act as “outcome” measures to be identified and estimated (see later discussion) [28]. The result is that a formative factor changes meanings depending on the other variables or factors used as outcomes. In some sense selecting the outcome measures is as important to a formative scale as the construct indicators themselves [19]. The problem epitomizes the concept of interpretational confounding discussed in Chapter 11 [13].

## OPERATIONALIZING A FORMATIVE SCALE

Figure 12.1 illustrates a formative measurement model. Each indicator ( $X$ ) is an index item that causes the composite construct. A correlation is shown among the index items ( $X_1 - X_3$ ), and  $E$  is a parameter (error term) indicating the amount of error variance in the construct. Notice that the error is now associated with the construct, and not with the measured items. Similarly, because the causality is from the items to the factor, and not the reverse, the construct does not explain the item intercorrelations. These differences lead to some changes in scale assessment and usage as discussed in the next sections.

The implications of collinearity and dropping indicator items are different in reflective and formative models. Reflective items are presumed to each be representative of the same conceptual domain. Therefore, dropping reflective items does not change the meaning of the latent construct. Items with low factor loadings can be dropped from



**Figure 12.1**  
Example of Formative Measurement Theory

reflective models without serious consequences, if a construct retains a sufficient number of indicators. Moreover, collinearity is expected in reflective measures as an indication of convergent validity. But in formative models, the items define the factor, so dropping or adding an item can have profound changes in its meaning. Conceptually, a formative factor should be represented by the entire population of indicators that form it [32]. Also, because there is no “common cause” for the items comprising a formative scale, there is no requirement that the items be correlated. As a matter of fact, collinearity among formative indicators presents significant problems because the loadings of the formative indicators to the construct can become unreliable in estimation (similar to the impact of multicollinearity in multiple regression as discussed in Chapter 5). If these parameters are unreliable, then it becomes impossible to validate scale items. Thus, the researcher faces a dilemma: Dropping an item may make the index incomplete, but keeping it may make an estimate(s) invalid. These issues associated with formative measurement models have yet to be resolved [20, 21].

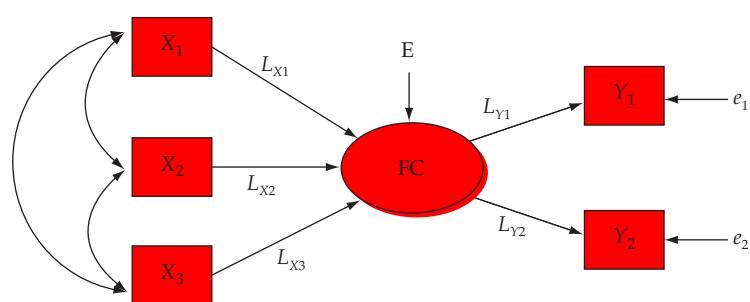
Formative measurement models require a different validation process. Because formative indicators need not be correlated, internal consistency, reliability (e.g., Cronbach Alpha), and average variance extracted (AVE) are not appropriate validation criteria. Indeed, formative items ideally are mutually exclusive [32]. Because the error is in the factor, the most important validation criteria relate to criterion or predictive validity. As noted earlier, the “validity” of a formative scale is contingent on the other model constructs or variables. Psychometric guidelines applied by researchers using reflective measurement theory (see Chapter 9) do not apply in to formative measurement theory [21, 52].

As previously noted, the mathematical identification of a formative scale is problematic (various approaches exist to cope with under-identification, one of which we will see in the next chapter). Figure 12.2 depicts three methods that allow mathematical identification, which is made particularly salient in CB-SEM. In Figure 12.2a we see that two reflective measured indicators (rectangles) have been added to the formative measure similar to the MIMIC model [33, 35]. In Figure 12.2b, the formative measure is related to two other multiple item reflective constructs. Finally, in Figure 12.2c a combination of one reflective construct and one reflective measured indicator (rectangle) provides identification for the formative scale. Each of these approaches has advantages and disadvantages that are the topic of continued discussion and research [19].

## DIFFERENCES BETWEEN REFLECTIVE AND FORMATIVE SCALES

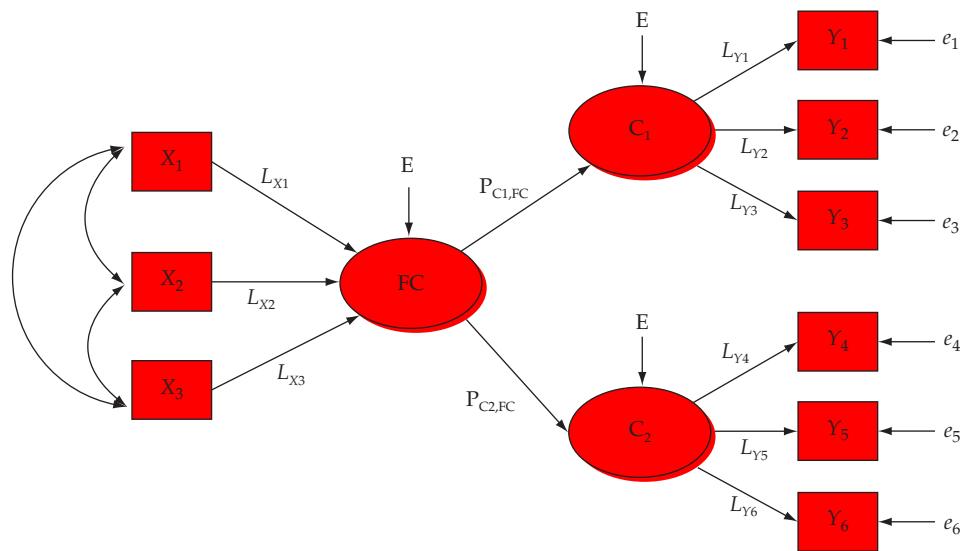
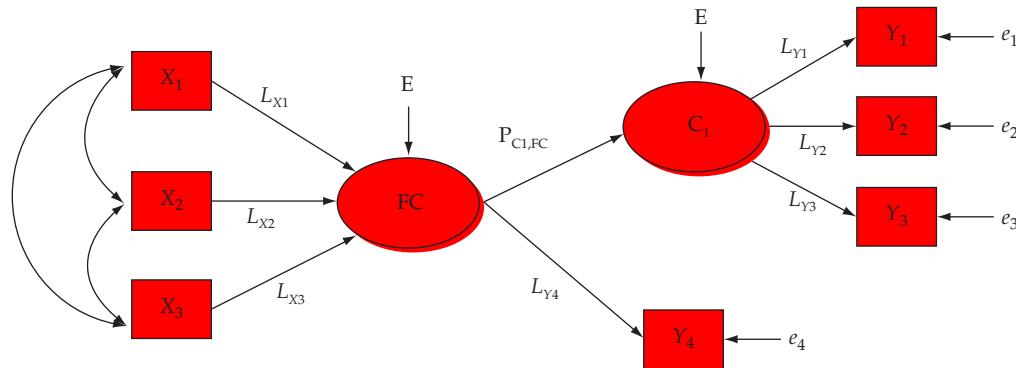
Meaningful differences separate reflective and formative measurement models, but conceptually differentiating between the two is not always easy. Reflective models are generally easier to work with, have traditionally been more commonly employed in the social sciences, and are thought to best represent many individual difference characteristics and perceptual measures. Formative measurement models may have a place when representing less psychological types of concepts [9, 32]. Additionally, the type of measurement does not affect the true nature of a concept. In fact, many scales, like perceived coercion in a B2B context, could be considered either reflective or formative depending on the context of the survey [30, 52]. One of the distinctions between reflective and formative measures also lies

a. Inclusion of Two Reflective Indicators



**Figure 12.2**  
Three Approaches to Identification of  
Formative Scales

(Continued)

**b. Inclusion of Two Reflective Constructs****Figure 12.2 (Continued)****c. Inclusion of One Reflective Construct and One Reflective Indicator**

around scientific objectivity. Changing the items for a formative scale or changing the constructs it predicts changes the meaning of the concept. Thus, the researcher determines the meaning through the choice of measured variables.

Table 12.1 presents a series of characteristics that may assist in selecting the appropriate measurement model form. As discussed earlier, the reflective and formative measurement theories are directly opposite approaches. A reflective theory presumes the construct causes the scale item indicators and are modeled as “outcomes” of the latent construct. All the items should have some conceptual linkage and should covary together. The items in a reflective scale need only to be a representative sample, and items can be added or dropped if the set of items provides coverage of the domain and the construct is identified. Reflective scales are required to exhibit internal consistency as a requirement for validity, whereas validity must be established internally (convergent validity) as well as externally (discriminant and predictive/criterion validity).

Formative scales are best characterized as indices rather than latent constructs because there is nothing “unobservable” when items define the construct. Specification of the complete domain of the concept through an exhaustive set of possible items is required, thus raising the possibility of compromising content validity if essential items are omitted or dropped. Because items are not required to be conceptually linked, except in their relationship to other constructs, there is also no requirement for collinearity among the items, and thus no level of internal consistency. In terms of construct validity, the lack of any internal validity measures requires that validity only be established through criterion or predictive validity and is contingent on the constructs used in the validation process.

## WHICH TO USE—REFLECTIVE OR FORMATIVE?

Formative scales have received considerable attention in recent years [19], particularly given a few published articles that claim previous research often presents reflective measurement that should be formative in the marketing, management, and strategic management literatures [32, 47]. However, the findings of these studies are questionable since the researchers often did not have access to the original questions or instructions used in the studies. Research that misspecifies measurement runs the risk of presenting inaccurate structural parameter relationships [37]. CB-SEM procedures can be used to model formative scales as long as the models are statistically identified, such as the models shown in Figure 12.2.

The trend toward more widespread use of formative scales has not been without concerns. Perhaps the most widespread concern has been the inherent lack of internal validity in formative measures and their potential for interpretation confounding based on the constructs and/or approach selected for identification and estimation purposes (see [52] for a comprehensive review). Additional issues of both a conceptual and empirical nature have been raised to the extent that some researchers call into question any use of formative measures [10, 11, 46, 52].

At this point, the question—"Which one do I use?"—does not have a definite answer. Research is continuing to address the questions raised by both proponents and opponents. As mentioned earlier, one could think of ways to measure many concepts either way [10], and in the end concepts exist without any concern over being formative or reflective. Table 12.1 provides guidance for the researcher in trying to decide.

## Higher-Order Factor Models

The CFA model described in Chapter 10 is a **first-order factor model**. A first-order factor model means the covariances between measured items are explained with a single latent factor layer. For now, think of a layer as one level of latent constructs.

Researchers increasingly are employing higher-order factor analyses although this aspect of measurement theory is not new. Higher-order CFAs most often test a **second-order factor** structure that contains two layers of latent

**Table 12.1** Distinguishing Between Reflective and Formative Measures

Characteristic	Indicative of:	
	Reflective	Formative
Causality of construct	Measured indicators assumed caused by factor.	Assumed causally formed from item indicators.
Conceptual relationship among items	All items are conceptually related because they have a common cause.	No requirement of conceptual linkage to other items.
Domain of items	Representative sample of potential items. All items have the same meaning.	Exhaustive inventory of all possible items. Items each have different meanings.
Covariance among items	Collinearity among items expected.	No expectation of collinearity. High collinearity among formative items questions the approach.
Internal consistency	Required.	Not required and in fact problematic.
Forms of construct validity	Standard psychometric assessment (as in Chapter 10).	Only external construct validity. Does the index predict other things?
Interpretational Confounding	Rare if factor is identified and meets ROT for convergence.	Inherent as formative loadings depend on outcomes.

## Formative Measurement

Formative scales do not represent latent factors and are not validated using the same methods as with conventional reflective factors.

The variables that make up a formative scale should explain the largest portion of variation and should highly correlate with other outcomes that are conceptually related (minimum correlation of .5).

Formative factors present greater difficulties with statistical identification.

At least two additional reflective variables or constructs with reflective variables must be included in the theoretical model along with a formative measure to achieve an overidentified model.

A formative scale should be represented by the entire population of items that form it. Therefore, items should never be dropped.

CB-SEM is appropriate for analyzing theoretical models with formative factors so long as statistical identification is possible.

constructs. One type of second-order latent factor(s), referred to as formative-formative, is specified so the second order factor causes multiple first-order latent factors, which in turn cause the measured variables ( $x$ ). But there are other approaches to model higher order factors (see Chapter 13). Theoretically this process can be extended to any number of multiple layers. Thus, the term *higher-order factor analysis*. Researchers seldom examine theories beyond a second-order model. Figure 12.3 contrasts path diagrams between a conventional first-order factor model with one layer in part (a) and a second-order factor model (b) with two layers in part (b).

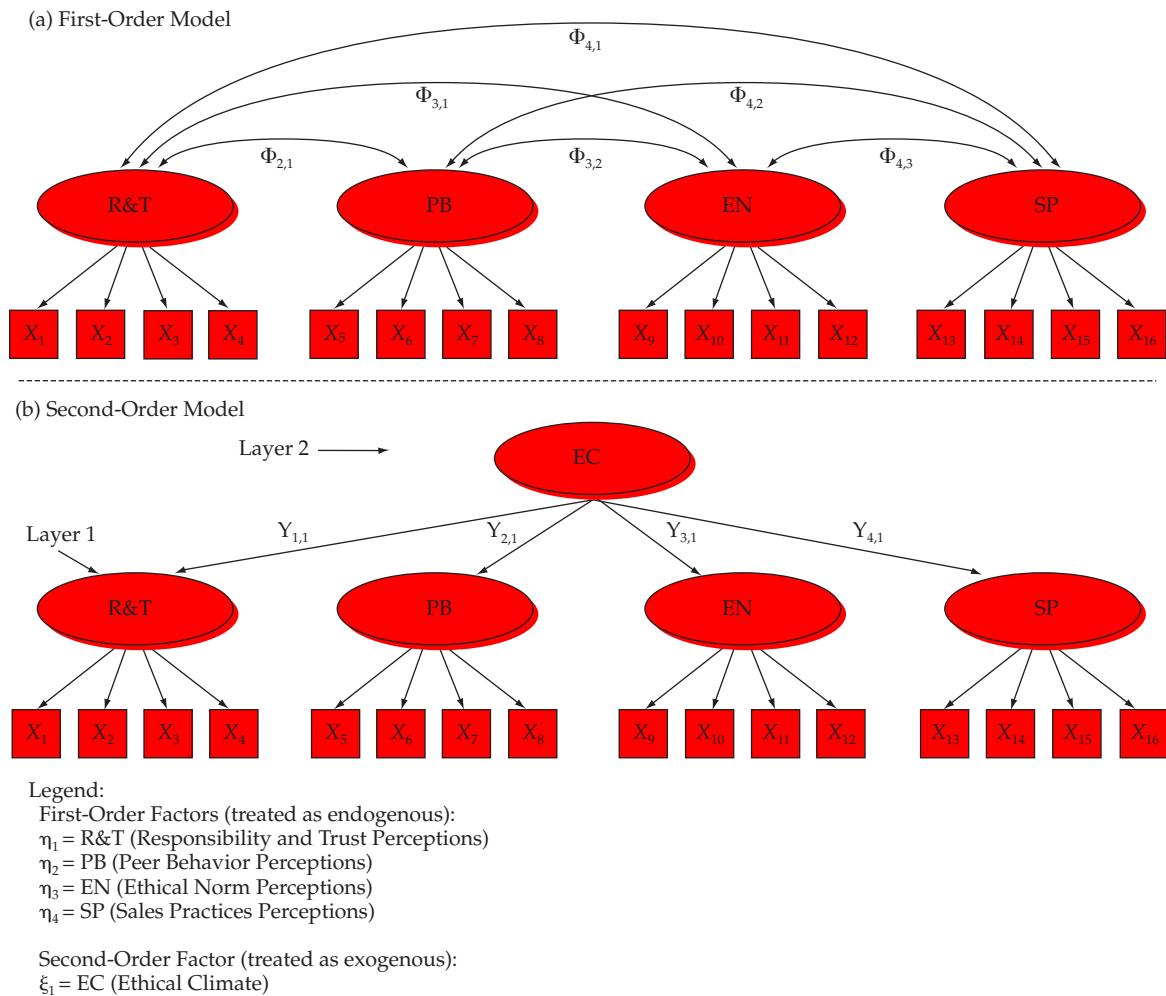
## EMPIRICAL CONCERNs

Both theoretical and empirical considerations are associated with higher-order CFA. All CFA models must account for the relationships among constructs. In a first-order CFA model, these covariance terms are typically free (estimated) unless the researcher has a strong theoretical reason to hypothesize independent dimensions. Recall from Chapter 9 that double-headed covariance terms could suggest the existence of an unmeasured latent factor. Higher-order factors can be thought of as explicitly representing the correlations among first-order factors. Thus, another way to view a higher-order factor is that it accounts for covariance between constructs just as first-order factors account for covariation between observed variables [2]. In other words, the first-order factors now act as indicators of the second-order factor. All the considerations and rules of thumb (items per factor, identification, scale, etc.) apply to second-order factors just as they do to first-order factors. The difference is the researcher must consider the first-order constructs as indicators of the second-order construct.

Figure 12.3a shows a conventional factor model with six unique covariances between four latent factors. Figure 12.3b depicts a CFA model where a second-order factor (Ethical Climate, EC) is introduced as the cause of the four first-order factors (R&T, PB, EN, and SP), each measured by four reflective items. Note that the introduction of a second-order factor changes the designation of the constructs. First, the first-order factors from the CFA model (which were originally exogenous constructs) now become endogenous constructs (note the arrows point from the higher-order construct toward the first-order constructs). The second-order factor is now the specified cause of the four first-order constructs versus using the correlational relationships among constructs to represent an unspecified common cause as was done in the CFA. Second, the higher-order construct is now the exogenous construct and it has no measured variables as indicators. Because it represents a relationship among constructs, the first-order factors act as its “indicators” through the structural model relationships. Finally, just as was required in specifying each first-order construct, the scale must be set for the second-order construct as well. The same two approaches are

**Figure 12.3**

Contrasting Path Diagrams for a First- and Second-Order Measurement Theory



available that we discussed when using measured variables. First, one structural path from the second-order factor to a first-order factor can be fixed at 1.0 to set the scale. Alternatively, all four loading estimates can be estimated if the variance of the second-order factor is fixed at 1.0.

## THEORETICAL CONCERNs

Theoretically, constructs sometimes can be operationalized at different levels of abstraction. Each layer in Figure 12-3b refers to a different level of abstraction. We will discuss two examples that illustrate the role of second-order factors.

Psychological constructs are often defined at different levels of abstraction. Personality can be represented by numerous related first-order factors. Each can be measured using dozens of multiple-item scales tapping a specific personality dimension. Psychological constructs representing first-order factors include scales for anxiety, pessimism, creativity, imaginativeness, and self-esteem, among many others. Alternatively, the first-order factors can be viewed as indicators of a smaller set of more abstract higher-order factors that reflect broader, more abstract personality orientations such as extraversion, neuroticism, conscientiousness, agreeableness,

and intellect [7, 48]. These more abstract personality constructs sometimes are referred to as the “Big Five” personality factors.

Similarly, one can imagine that many different factors might indicate how well one would do in graduate school. Multiple indicators from a standardized test could be used to represent verbal performance and quantitative performance, among other exam characteristics. Multiple items also could be used to assess how well a candidate performs in school, including GPAs in college, GPAs in high school, and perhaps several other grade-related scores. We also could use multiple-item scales to assess how motivated one is to succeed in graduate school. Once we have identified all the factors related to performance in graduate school, we may end up with a few dozen indicator variables for several factors, such as reading comprehension, quantitative ability, problem solving, school performance, and motivation. Each of these aspects is in itself a factor. However, they may all be driven by a higher-order factor that we could label “Likelihood of Success.” It may be difficult to look at one’s credentials and directly assess likelihood of success. However, it may be indicated quite well by more tangible factors such as problem-solving ability. In the end, key decisions may be made based on the more abstract success factor, and hopefully these decisions are better than relying on the individual more specific factors. Thus, the individual factors are first-order factors and Likelihood of Success could be thought of as a second-order factor. This type of situation calls for the testing of a second-order CFA model.

It cannot be emphasized enough that the ultimate criterion in deciding to form a second-order measurement model is theory. Does it make theoretical sense? What logical reason leads us to expect layers of constructs? The well-known Big Five Personality theory represents the notion of higher order factors well. Personality is a higher order factor to the big five, extraversion, agreeableness, openness to experience, conscientiousness, and neuroticism. Each of the five factors has numerous measured indicator items. The increasing number of second-order factor models seen in the literature is partially the result of more researchers learning how to use SEM to represent and test a higher-order factor structure. The ability to conduct a second-order test does not justify doing so. The need for theory is particularly true when trying to decide between a first- and second-order factor configuration for a given measurement theory.

## USING SECOND-ORDER MEASUREMENT THEORIES

The specification of a second-order CFA model is actually quite similar to a first-order model if we view the first-order constructs as indicators. Considering Figure 12.3a, the first-order model estimates a relationship (two-headed path in this case) for each potential covariance. The higher-order model in Figure 12.3b accounts for these six relationships with four factor loadings. Although the comparison between a first- and second-order measurement model is generally nested, the empirical comparison using a  $\Delta\chi^2$  statistic is not as useful as it is when comparing competing measurement models of the same order [38]. The first-order model should fit better in absolute terms because it uses more paths to capture the same amount of covariance.

In contrast, the higher-order model is more parsimonious (it consumes fewer degrees of freedom). Thus, it should perform better on indices that reflect parsimony (PNFI, RMSEA, etc.). Note, however, that even though a higher-order model is more parsimonious from the standpoint of degrees of freedom, one may not view it as “simpler” because it involves multiple levels of abstraction. This complicates empirical comparisons, thus placing more weight on theoretical and pragmatic concerns.

Higher-order measurement models are also still subject to construct validity standards. In particular, second-order factors should be rigorously examined for nomological validity, because it is possible that various confounding explanations may exist for a higher-order factor. For example, if all item measures use the same type of rating scale, there could be a common methods factor influencing all first-order constructs. The second-order factor could be interpreted as common measurement in this case. When a second-order factor reacts to other theoretical constructs as expected, the chance of it being an artifact of the research design is lower. More specifically, if the higher-order factor explains theoretically related outcomes such as organizational commitment and job satisfaction as well as or better than the combined set of first-order factors, then evidence in favor of the higher-order representation is

provided [38]. Thus, a primary validation criterion becomes how well a higher-order factor explains theoretically related constructs. When comparing measurement models of different orders, a second-order model is supported to the extent that it shows greater nomological validity than a first-order model.

## WHEN TO USE HIGHER-ORDER FACTOR ANALYSIS

Although higher-order measurement models might seem to have many advantages, we must also consider the disadvantages. In general, they are conceptually more complicated. A construct can become so abstract that it is difficult to adequately describe its meaning. The added complexity also can diminish the diagnostic value of a construct as it becomes further removed from the tangible measured items. Higher-order CFA models also create more potential for unidentified or improper CFA solutions. For instance, researchers may have one or more higher-order factors with fewer than three indicators. Just as we saw in Chapter 9 with first order factors, a two-item higher order factor can exist in an identified CFA but the factor itself is mathematically under-identified without added constraints.

With a reflective second-order or higher factor model, all first-order factors, which are now indicators of the second-order factor, are expected to move together (covary), just as with the measured items indicating first-order factors. When multiple first-order factors are used as indicators of a second-order factor, the researcher gives up the ability to test for relationships between these first-order factors and other key constructs. Thus, a drawback of the measurement model shown in Figure 12.3b is that we cannot investigate, for example, direct relationships between peer behavior (PB) and other key job outcomes such as turnover. Thus, the presumption is that all four first-order indicators would influence any other construct (e.g., turnover) the same way. If a conceptual case can be made that anyone of these first-order factors would affect another key construct differently, then perhaps a second-order measurement theory should not be used. This case is typified when one of a set of related first-order constructs would be expected to affect some other construct positively whereas other first-order constructs would affect it negatively.

Some questions that can help determine whether a higher-order measurement model is appropriate are listed here:

- 1 Is there a theoretical reason to expect that multiple conceptual layers of a construct exist?
- 2 Are all the first-order factors expected to influence other nomologically related constructs in the same way?
- 3 Are the higher-order factors going to be used to predict other constructs of the same general level of abstraction (i.e., global personality-global attitudes)?
- 4 Are the minimum conditions for identification and good measurement practice present in both the first-order and higher-order layers of the measurement theory?

If the answer to each of these questions is yes, then a higher-order measurement model becomes applicable. After empirically testing higher-order models, the following questions should be addressed:

- 1 Does the higher-order factor model exhibit adequate fit?
- 2 Do the higher-order factors predict other conceptually related constructs adequately and as expected?
- 3 When comparing to a lower-order factor model, does the higher-order model exhibit equal or better predictive validity (higher order factor does not mask findings from individual lower order factors)?

Once again, if the answer to these questions is yes, then a higher-order measurement theory would be supported.

## Multiple Groups Analysis

Numerous SEM applications involve analyzing groups of respondents. Groups sometimes represent dividing a sample by a meaningful characteristic such as a respondent's gender or home country. For example, we may expect that men and women do not respond similarly across a wide range of social psychological issues. Many times, different populations are sampled with the ultimate aim of testing for similarities and differences between those populations. For example, the populations may involve people from different cultures. Alternatively, a large sample may be broken randomly into two subsamples so that model cross-validation can take place.

## Higher-Order Factor Models

Higher-order factors must have a theoretical justification and should be used to predict other constructs of the same general level of abstraction

All first-order factors should influence other related constructs in the same way

All rules that apply for identification at the first order apply at higher orders

**Multiple group analysis** is a SEM framework for testing any number or type of differences between models estimated for different groups. The general objective is to test for potential differences between individual group models. Although very specific tests of differences can be performed for unique research questions, a general framework has emerged for comparing the measurement models and then structural models across groups. We will first discuss measurement model invariance, because it often is a prerequisite for making comparisons at the structural model level, particularly in cross-cultural applications.

## MEASUREMENT MODEL COMPARISONS

A key benefit of achieving construct validity is that a construct will meet all of the requirements of reliability and validity, not only in one situation, but hopefully across all potential situations in which it can be applied. Although rigorous testing in the development stage may support construct validity, researchers have long been aware of the need to reassess a construct when scales are applied in different contexts [22, 29, 30, 31, 43]. Increased applications of SEM to cross-cultural studies, longitudinal studies, assessment of differences based on personal differences (e.g., gender), and even environments (e.g., type of workplace setting), have brought to our attention the need for a formalized process of making group comparisons [2, 15, 16, 49, 51].

We discuss measurement model comparisons in two related areas. The first is broadly known as **measurement invariance** (or **measurement equivalence**). The primary objective is to ensure that measurement models conducted among different populations yield equivalent representations of the same construct. As we have discussed, the types of situations using this approach have become quite extensive. Yet even with the increased awareness and availability of SEM programs, in many areas of research that require measurement invariance (e.g., cross-cultural research), it is still not always used [53]. A more specific instance of measurement model comparisons is **cross-validation**, the attempt to reproduce the results found in one sample using data from a different sample. Generally, cross-validation uses two samples drawn from the same population. In other words, the sampling units in each group would have the same characteristics. Perhaps the most basic application is providing a second confirmation of a measurement theory that survived initial testing. One way of accomplishing this task is to split a large sample randomly into two groups so that each sample meets the minimum size requirements discussed earlier.

**A Six-Step Process of Group Comparisons** What both of these areas have in common is their joint use of what has become known as **multisample confirmatory factors analysis** (MCFA). CFA, as discussed in Chapter 9, provides the basis for establishing construct validity through the measurement model. MCFA now extends CFA into a multigroup situation where separate samples are collected for each group and then comparisons made to determine their invariance (or equivalence). As might be expected, numerous aspects of a CFA model are available for comparison. A systematic framework has been proposed for evaluating these aspects in a progressively more rigorous set of comparisons that addresses the most elemental aspects in the earlier stages [40, 41].

The foundation of the process is a series of empirical comparisons of models with increasingly restrictive constraints. The fundamental measure of difference used is the chi-square difference ( $\Delta\chi^2$ ). This measure allows for an overall comparison between two model specifications (e.g., one with fewer constraints than the other). The basic logic remains that if a more constrained model fits (as measured by chi-square) as well as a less constrained model, then results support the more constrained model [6, 26, 36, 41].

General practice is to start with a relatively unconstrained model. For example, the initial multi-group CFA would allow separate and unique loadings for each group. Then **between-group constraints** are added to reflect specific measurement model comparisons. A typical between-group constraint restricts parameter estimates to be equal in each group. Rather than freeing a factor loading in each group, in essence the constraint frees the loading in one group and forces the corresponding loading to the same value in the other group. In this way, one can test a hypothesis that a relationship is invariant (equal) across the groups. If the model with additional constraints does not fit significantly worse than the lesser constrained model, then the results support invariance.

One feature of the MCFA approach is that although basic measures of fit exist for each group model, model fit measures are provided for the collective set of group models. In simple terms the chi-square for each group model added together equals the overall chi-square. Measures such as CFI, RMSEA, and other fit indices, are calculated for the entire model set. In this way, comparisons can be made on model fit measures (e.g.,  $\Delta\chi^2$ ) across MCFA models with differing sets of constraints.

Although all model fit indices are available for the set of group models, the primary measure used for comparison remains the **chi-square difference ( $\Delta\chi^2$ )**, which can be assessed with a statistical significance level. The degrees of freedom for any  $\Delta\chi^2$  model comparison equals the number of additional constraints introduced from one stage to the more restrictive stage. For example, assume that we have three groups. We could model separate models for each group, creating three unique  $\chi^2$  values and separate loading estimates for each variable in each group. Next, we assume that the loading estimates are equal. Instead of three estimates for a given loading, we would have only one estimate that is the best single loading estimate for the three groups combined. Now, we estimate one value instead of three, saving two degrees of freedom. So, in general the number of degrees of freedom for the  $\Delta\chi^2$  test is equal to the number of added equality constraints multiplied by one less than the number of groups.

We will first describe the six stages in terms of both the measurement model issues they address as well as the nature of the constraints used. Each stage introduces a new set of constraints to those in the previous, lesser-constrained, model. For example, the model at Stage 3 will have all constraints imposed in Stage 2 plus those added in Stage 3. So, the  $\Delta\chi^2$  test can be made between models at each stage rather than only with the initial or baseline model.

**STAGE 1: CONFIGURAL INVARIANCE** The first stage confirms **configural invariance**—that the same basic factor structure exists in all groups. Researchers should confirm that each group CFA model has the same number of constructs and items associated with each construct. Moreover, it must be shown that each group model meets appropriate levels of model fit and construct validity. In measurement theory terms, we are now ensuring that the constructs are congeneric across groups. This model is sometimes referred to as the **totally free multiple group model (TF)** because it estimates all free parameters separately meaning each parameter takes on its own value in each group. The TF (configural invariance) model becomes the baseline model for comparison.

**STAGE 2: METRIC INVARIANCE** The second stage provides the first empirical comparison between MCFA models groups and involves the equivalence of factor loadings. **Metric invariance** represents equivalence in the relationships between measured variables and constructs and is a critical test of invariance. The degree to which metric invariance exists determines cross-group equivalence beyond the basic factor structure. Constraints are set so that the factor loadings are equal across groups (e.g.,  $L_{X1,\text{Group}1} = L_{X1,\text{Group}2}$ ,  $L_{X2,\text{Group}1} = L_{X2,\text{Group}2}$ , ...). The  $\Delta\chi^2$  is computed between this model and the TF (configural invariance) model with degrees of freedom equaling the number of loading estimates constrained equal across the groups.

**STAGE 3: SCALAR INVARIANCE** The third stage is **scalar invariance**, which tests for the equality of the intercept terms in the equations explaining the measured variables. Support for scalar invariance is required if any comparisons of relative construct level (e.g., mean scores) are made across groups. That is, scalar invariance allows the relative amounts of latent constructs to be compared between groups.

**STAGE 4: FACTOR COVARIANCE INVARIANCE** In a fourth stage, the covariances between constructs can be constrained. **Factor covariance invariance** tests if constructs are related to each other in a similar fashion across groups. The difference in degrees of freedom represents the number of equality constraints for factor covariances. If the  $\Delta\chi^2$  is not significant, the researcher concludes that the covariances among factors are the same in each group.

**STAGE 5: FACTOR VARIANCE INVARIANCE** Now the test is for **factor variance invariance**, which assesses the equality of the variances of the constructs across the groups. If factor variances and covariances are equivalent across groups, then the latent construct correlations also are equal.

**STAGE 6: ERROR VARIANCE INVARIANCE** The final stage tests for the **error term invariance** for each measured variable across the groups. This test is whether the measurement error variance in the indicators is equivalent among groups. Generally, the most important comparisons are through Stage 3.

**Full Versus Partial Invariance** The  $\Delta\chi^2$  test for metric invariance or scalar invariance tests **full invariance**, meaning that constraining all the parameters relative to that type of invariance to be the same in each group does not significantly worsen fit. In the case of metric invariance, this would mean constraining each of the corresponding loadings to be the same in each group. Yet, full invariance becomes more difficult to achieve as models become complex and the tests progress to later stages [29]. **Partial invariance** is a less conservative standard involving more than one estimate per construct to be equivalent across groups [14]. Although a weaker test of invariance, if two parameters per construct (e.g., loadings in metric invariance, intercepts in scalar invariance, or even error terms in error variance invariance) are found to be invariant, then partial invariance is found, and the process can extend to the next stage.

If full invariance is not supported, the researcher can systematically “free” the individual equality constraints for variables showing the greatest differences in unstandardized loadings in the hope that the  $\Delta\chi^2$  will become nonsignificant with at least two constraints per construct. One approach to choosing the constraints that should be eliminated first is to examine the modification indices for the fully constrained model. Remember that modification indices suggest the change in  $\chi^2$  associated with estimating a relationship. Thus, equality constraints with the largest modification indices should be freed first. An approach using the specification search feature of SEM programs to identify and free the most restrictive constraints has also been proposed [54]. No matter what approach is taken, the objective is to free as few constraints as possible to achieve invariance.

**Alternative to Partial Invariance** The concept of partial invariance can be tedious to test and theoretically unsatisfying. Why are two equal indicators sufficient in a scale that may contain many more items? An alternative approach to establishing invariance mirrors that of validating a CFA whose diagnostics suggest problems. Perhaps the problems with invariance reside in only a single variable. If so, a much simpler solution to partial metric invariance is to delete the offending item, particularly if the deletion does not cause problems with measurement model overall [3]. An item can be dropped and the test of metric invariance (or scalar invariance as the case may be) can be examined. If the  $\Delta\chi^2$  is insignificant after dropping the item, full invariance is established on the reduced model and the researcher can move forward with the relevant comparisons [3]. Following the same guidelines as for establishing a valid CFA model (see Chapter 10), the researcher may be able to drop a small number of items and establish invariance. If the number of items dropped is fewer than the number of constructs, and the deletions do not cause problems with identification of the model, then the approach is preferable to trying to establish partial invariance. The analysis moves forward with full invariance.

**What Level of Invariance is Needed?** With six stages of invariance now defined, one might ask: What level of invariance is needed? Is it necessary to achieve invariance for all six stages? The answer to that question depends on the type of research question being addressed. A research question that involves comparing construct means across groups requires scalar invariance. A research question that involves comparing relationships across groups, such as the case with a moderation hypothesis, requires metric invariance. Moreover, error variance invariance is rarely examined and demonstrates equal construct reliability across the groups. Error variance invariance also would mean that both unstandardized and standardized relationships can be compared across groups. Typically though, the examination of unstandardized relationships is sufficient to establish differences between groups.

Table 12.2 provides guidelines for the *minimum* level of invariance needed for different types of research questions [16, 49, 51]. Focusing on the measurement model issues, we can see that the most common issue, equivalence of the basic structure of the construct, requires configural invariance only. Comparisons of relationships between groups requires full or partial metric invariance. When comparisons of the latent construct means are made, scalar invariance is required [43]. Again though, the deletion of an offending item offers an often preferable alternative to partial invariance.

**HBAT Invariance Analysis** During the course of interviews between HBAT management and the consultants, numerous issues arose suggesting a need to compare full-time versus part-time employees. The concern was that just taking an overall perspective on employees might overlook noticeable differences between these two groups. It was felt that the first step should be to make sure that these two groups had common perceptions about the workplace attitudes that HBAT considered important in their employee retention efforts based on the assumption that part-time and full-time employees represent two distinct populations, similar to that of comparing employees from two different countries. This led to a call for an empirical comparison of the two groups on the five constructs in the Employee Retention model (see Chapters 10 and 11 for a more complete description of the constructs and model).

Groups were formed from the respondents to the HBAT employee survey based on work status—full-time or part-time. In this case, the groups were of almost equal size (191 part-time employees, 209 full-time employees). With the groups and their responses defined, the invariance testing process can begin.

**SIX-STAGE INVARIANCE TESTING PROCESS** Group models were specified based on the six-stage process and then estimated. Table 12.3 contains the model fit statistics for each model and the chi-square difference test for each model comparison. In the first stage of configural invariance, the separate models for full-time and part-time employees both exhibit acceptable levels of model fit, as does the combined MCFA model ( $\chi^2 = 438.1$ ,  $df = 358$ ,  $p = .002$ , RMSEA = .024, and CFI = .98). Note that the MCFA  $\chi^2$  is equal to the sum of the two employee group models and the other fit

**Table 12.2** Suggested Minimum Levels of Invariance by Type of Research Question

	Measurement Model Comparisons <sup>a</sup>		Structural Model Comparisons <sup>a</sup>	
	Basic Structure: Is the construct perceived and used in a similar manner?	Mean Levels: Do the groups have equal amounts of latent constructs?	Theoretical Relationship Equivalence: Is the relationship between constructs the same across groups?	Theoretical Relationship Equivalence: Are correlations or standardized loadings the same across groups?
Configural	Full	Full	Full	Full
Metric	Partial	Partial	Partial	Partial
Scalar		Partial		Partial
Factor covariance				Partial
Factor variance				
Error term				

<sup>a</sup>Minimum levels of invariance required.

**Table 12.3** Measurement Invariance Tests for Full-Time Versus Part-Time Employees

Model Tested	Model Fit Measures					Model Differences		
	$\chi^2$	df	p	RMSEA	CFI	$\Delta\chi^2$	$\Delta df$	p
Separate groups								
Part-time employees	259.8	179	.000	.049	.96			
Full-time employees	178.3	179	.500	.000	1.00			
Configural invariance	438.1	358	.002	.024	.98			
Metric invariance	450.8	374	.004	.023	.98	12.7	16	.69
Scalar invariance	521.8	395	.000	.028	.97	71.0	21	.00
Factor covariance invariance	535.6	405	.000	.028	.97	13.8	10	.19
Factor variance invariance	536.2	410	.000	.028	.97	.6	5	.98
Error variance invariance	587.7	431	.000	.030	.96	51.5	21	.00
Partial scalar invariance	457.4	384	.006	.022	.98	6.6	10	.77

measures signify acceptable fit across the two groups—indicating configural invariance. The next test is for metric invariance and involves constraining each matching loading to be equal across the groups. We can see that the  $\Delta\chi^2$  is only 12.7 with 16 degrees of freedom, which indicates a nonsignificant difference. The 16 degrees of freedom represent the 16 free factor loadings that were constrained to be equal to the other group (remember that one parameter was already constrained to 1.0 to set the scale on each construct, thus leaving 16 free parameters across the measured variables). Thus, the two models exhibit full metric invariance.

The next stage is to test for scalar invariance. Here the  $\Delta\chi^2$  is 71.0 with 21 degrees of freedom. This difference is statistically significant, indicating that full scalar variance is not supported. For illustrative purposes we will continue to the next stage of the process and then return to assess if at least partial scalar invariance can be achieved. The next stage tests for factor covariance invariance, and the  $\Delta\chi^2$  of 13.8 with 10 degrees of freedom is nonsignificant. From this result we can support invariance in the covariances among matching constructs. The next test for factor variance invariance shows very little difference for this constraint ( $\Delta\chi^2 = .6$ ,  $df = 5$ ), indicating that factor variances are almost identical between the two models. The final invariance test is for equivalence of the error terms of the indicators. As expected this test had a significant chi-square difference ( $\Delta\chi^2 = 51.5$ ,  $df = 21$ ). If equivalence of the error terms of the indicators would have been supported, then equal reliabilities would have been found for constructs in each group.

We now return to see if we can achieve at least partial scalar invariance. Modification indices for the scalar invariance model were examined to identify the constraints of item intercepts that could be freed to most reduce the chi-square difference. As a result, 10 items (two per construct) were identified to retain their constraints in a test of partial scalar invariance. The items were JS<sub>2</sub>, JS<sub>4</sub>, OC<sub>3</sub>, OC<sub>4</sub>, SI<sub>1</sub>, SI<sub>3</sub>, AC<sub>1</sub>, AC<sub>4</sub>, EP<sub>2</sub>, and EP<sub>4</sub>. A model constraining each of these parameters to be equal to one another in each group produced a  $\Delta\chi^2$  of only 6.6 ( $df = 10$ ) from the metric invariance model, which was nonsignificant. Thus, partial scalar invariance can be supported as well. Thus, comparisons between construct means are possible.

**MEASUREMENT INVARIANCE CONCLUSIONS** The measurement invariance testing process demonstrated that the five constructs used in the Employee Retention model met the criteria for configural invariance, full metric invariance and partial scalar invariance. As a result, almost any form of group comparison can be made without concern that the differences are due to differing measurement properties between the two groups.

## STRUCTURAL MODEL COMPARISONS

The process of group comparisons for structural model parameters first builds upon the measurement model process and then performs similar types of comparisons to assess the differences in the structural model. Any type of structural model comparison first requires at metric invariance (partial if full cannot be established) of the measurement

model to ensure that the constructs are comparable. If metric invariance is not achieved, then the researcher cannot be sure whether the differences seen in a structural model parameter are due to a group idiosyncrasy, or if they truly represent a differing structural relationship.

Structural model comparisons provide a specific test for addressing any number of research hypotheses, but the most common use is the test of moderation. Discussed in more detail in a following section of this chapter, moderation assesses the differences in structural relationships between groups formed on a third variable. Group model comparisons can identify the extent of the differences either for an entire model or a specific relationship. We provide a complete discussion of moderation and an example using the HBAT Employee Retention model in a later section.

The other research question concerning the structural model is to compare the means of latent constructs. In making these comparisons, SEM programs compare means only in a relative sense. In other words, they can tell you whether the mean is higher or lower relative to another group [43]. The interpretation of the difference in means assumes the mean is zero in a reference group, and the mean values estimated for other groups shows how much more or less the means are than in the reference group. Positive estimates indicate a higher mean than in the reference group and negative means indicate a lower average.

## Measurement Type Bias

Researchers sometimes become concerned that survey responses are biased based on how questions are asked. For instance, it could be argued that the order in which questions are asked could be responsible for the covariance among items that are grouped close together. Similarly, researchers often are faced with resolving the question of potential common methods bias. **Common (sometimes called constant) methods bias (CMB)** would imply that the covariance among measured items is influenced by the fact that some or all of the responses are collected with the same type of scale. Thus, the covariance could be explained by the way respondents use a certain scale type in addition to or instead of the content of the scale items.

CMB is an example of what is known as a nuisance factor. A **nuisance factor** is something that may affect the responses but is not of primary interest to the research question. For many effects of this type it is assumed that they are just represented in the error terms. But if the impact of a nuisance factor is substantial or systematic enough to impact the results, then it should be included in the model. We first discuss the model specification issues involved in assessing the impact of a nuisance factor and then examine model estimates to assess the extent of the effect.

## MODEL SPECIFICATION

The concept of a nuisance factor is widely used in experimental designs, where factors in administration of the experiment (time of day, room conditions or even administrator characteristics) may be thought to have some impact and thus need to be “controlled for” in the design. The most common approach is to introduce blocking factors or covariates to try to control for potential confounding effects. In structural equations modeling, we follow a similar approach. To do so, we create a latent construct to represent the nuisance factor and then account for it in the model. With CMB, the focus is on effects related to the questionnaire design and administration, but the approach could be extended to any type of nuisance effect.

In survey research, CMB may arise because a common type of scale causes a response style that is independent of the content of the item. In SEM terms, this means that an external effect (e.g., the scale type) impacts the measured variables. We can represent this external effect in SEM by creating an additional latent construct for each scale type and relate it to the measured variables collected by that type of scale. Because the “cause” impacts the measured variables, the construct is reflective (i.e., arrows go from construct to measured variables) and operates as any other construct of this type. Note that a construct of this type violates the principles of good measurement theory

## Multiple Group Models and Measurement Invariance Testing

Multigroup models provide a comprehensive framework for comparing any model parameter between two or more samples of respondents.

Multisample confirmatory factor analysis (MCFA), a form of multigroup analysis, is the framework for assessing measurement invariance.

The chi-square difference test is the empirical means of assessing if a between-group constraint is statistically significant.

If a between-group constraint is nonsignificant, this means that the parameter being evaluated does not vary between groups.

Metric invariance, involving the equivalence of factor loadings, is needed for making model comparisons of relationships between constructs.

Scalar invariance, the equality of indicator intercepts, allows for comparing latent construct means across groups.

If full invariance cannot be achieved, partial invariance is acceptable if two indicators per construct are found to be invariant. However, a preferable approach would be to drop offending items so long as they are only a very small number of variables responsible for the invariance, and move forward with full invariance.

because it creates cross-loadings for the measured variables involved and thus impacts the unidimensionality of the construct. In addition, because the scale type should exert a constant effect on variables, the loadings from the method (nuisance) factor can be constrained to be equal to each other.

Let's use the HBAT employee questionnaire as an example. In gathering responses, several different types of rating scales were used. Although it could be argued that respondents prefer a single format on any questionnaire, several advantages come with using a small number of different formats. One advantage is that the researcher can assess the extent to which any particular scale type may be biasing the results. Also, by including different scale types, the researcher attempts to address concerns about CMB in the design.

In this case, HBAT is concerned that the semantic differential items are accounting for some common variance. The analyst suggests that respondents may have consistent patterns of responses to semantic differential scales no matter what the subject of the item is. Therefore, a semantic differential factor may help explain results. A CFA model can be used to test this proposition. To do so, an additional construct is created to represent the effect causing the semantic differential items. In this case, items EP<sub>4</sub>, JS<sub>2</sub>, JS<sub>3</sub>, AC<sub>2</sub>, and SI<sub>4</sub> are all measured with semantic differential scales. Thus, the model needs to estimate paths between this new construct and these measured items. Remember that factors of this type that are added will not have congeneric measurement properties.

We will modify the original HBAT CFA model (see Figure 11.5 in Chapter 11) by adding a sixth construct to represent the constant methods bias with dependence paths (loadings) to the five measured variables (EP<sub>4</sub>, JS<sub>2</sub>, JS<sub>3</sub>, AC<sub>2</sub>, and SI<sub>4</sub>). These five variables will now have two factor loadings, one to their original latent construct (e.g., EP<sub>4</sub> loading on the EP construct) as well as a loading to the constant methods bias construct. Thus, the measurement model no longer exhibits simple structure because each of the five measured variables is now determined both by its conceptual factor and by the new construct. We provide instructions on the text's online resources on how to do this using either LISREL or AMOS. When all items are measured with the same scale, the equivalent addition would be a single factor causing all measured items.

## MODEL INTERPRETATION

Assessing the potential impact of the common methods (or any other nuisance factor) follows procedures for comparing any other set of competing models. First, overall model fits are compared to see if the additional factor has a significant impact. This is done through a chi-square difference ( $\Delta\chi^2$ ) test and examination of the model fit indices. If the model fit significantly improves with the addition of the nuisance factor construct, then the researcher would know that the effect was substantial and should proceed to understanding the nature of the effect. Examination of the factor loadings provides a more precise estimate of the extent and magnitude of the effects on the individual items. Substantial loadings indicate the presence of another cause for the measured variables in addition to the original latent construct it represents.

Researchers should always be cautious in interpreting these effects because they are based on the assumption that the nuisance factor actually represents that effect (e.g., common methods) and does not, in actuality, represent some other factor. Because the nuisance factor is many times not directly observed, the researcher must be careful to not allow spurious or other non-specified effects to confound the nuisance factor.

To test for a constant methods effect due to the semantic differential scale, the CFA model is estimated with the addition of the sixth construct. The following fit statistics are obtained:  $\chi^2 = 232.6$  with 174 degrees of freedom and the RMSEA, PNFI, and CFI are .028, .80, and .99, respectively. The  $\Delta\chi^2$  of 4.0 ( $236.6 - 232.6 = 4.0$ ) with 5 ( $179 - 174 = 5$ ) degrees of freedom is insignificant. In addition, the estimates associated with the nuisance construct are nonsignificant. Also, the values for the original parameter estimates remain virtually unchanged as well. Although not the case here, the researcher would be justified to force an equality constraint on all the method “scale” factor loadings based on the assumption that common scaling should affect all item responses equally.

Based on the model fit comparisons, the insignificant parameter estimates, and the parameter stability, no evidence supports the proposition that responses to semantic differential items are biasing results, and in this case the responses are not subject to measurement bias. Another factor could be added to act as a potential nuisance cause for the items representing another scale type, such as all Likert items. The test would proceed in much the same way.

Another approach for examining for the possibility of CMB involves using the first eigenvalue obtained from principal components analysis of the set of measured items used in the model. The test is sometimes referred to as the Harmon's one factor test. The idea is that if a single component (or factor) accounts for less than half the common variance among the items, then CMB is not likely present. Although the test is criticized as not being conservative, a recent simulation study suggests that common methods variance has to be quite high before CMB becomes problematic [24]. The results suggest that as long as scale AVEs and reliabilities meet the established guidelines, Harmon's one factor test actually performs quite well in detecting problematic levels of common methods variance. To be even more conservative, one could drop the percent of variance accounted for by the first factor down to 40 percent. If the first component (factor) accounts for 40 percent of the total variance or less, CMB is not likely to be problematic.

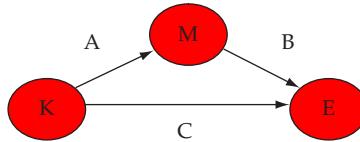
Some have previously emphasized that CMB is a major problem in empirical survey research [47]. Recent research indicates, however, that CMB is less of a problem than previously thought if the proper design is applied in developing the questionnaire. Moreover, the so-called Harmon Single Factor test was previously criticized as a weak test of CMV, but it can be viewed as a simple and useful method of identifying the likelihood of CMB [4, 24].

## Relationship Types: Mediation and Moderation

In Chapter 12 we introduced two basic types of relationships: correlational and dependence. As we have seen in the models using these relationship types, they are “building blocks” of our structural models. We now discuss two variations of these basic relationships: mediation and moderation. For each new relationship type the discussion will first focus on the theoretical nature of the relationship and how it can be incorporated into our SEM models. Then we discuss the process whereby we incorporate these relationships into our existing models.

## MEDIATION

In simple terms, a **mediating effect** is created when a third variable/construct intervenes between two other related constructs [5]. Mediating effects highlight the distinction between direct and indirect effects. **Direct effects** are the relationship linking two constructs with a single arrow (connection). **Indirect effects** are those relationships that involve a sequence of relationships with at least one intervening construct involved. Thus, an indirect effect is a sequence of two or more direct effects (compound path) and is represented visually by multiple arrows. The following diagram shows both a direct effect ( $K \rightarrow E$ ) and an indirect effect of  $K$  on  $E$  in the form of a  $K \rightarrow M \rightarrow E$  sequence:



The indirect effect ( $K \rightarrow M \rightarrow E$ ) represents the mediating effect of construct  $M$  on the relationship between  $K$  and  $E$ .

**Conceptual Basis for Mediation** From a theoretical perspective, the most common application of mediation is to “explain” why a relationship between two constructs exists. We may observe a relationship between constructs (e.g.,  $K$  and  $E$ ), but not know “why” it exists. We can then posit some explanation in terms of an intervening or facilitating variable, which operates to takes the “inputs” from  $K$  and translate them into the “output,”  $E$ . As such, the mediator ( $M$ ) facilitates and explains why the relationship between the two original constructs exists. A simple example illustrates these points.

The construct  $K$  could be a student’s intelligence and  $E$  could be classroom performance. What is interesting is not just that the relationship exists between intelligence and classroom performances, but how it works. Can we explain how students translate their intelligence into performance? We may find that sometimes a student exhibits high intelligence, but does not always perform well. Moreover, some students with lower intelligence scores perform extremely well. So, is there some other process going on that translates student intelligence into actual classroom performance?

The intervening process is the mediating effect. In this case, we could propose a construct termed “study effectiveness.” Here we refer to such characteristics as the ability of students to focus their efforts on their class work, organize their class-related and other activities to provide them with sufficient time to complete their homework, and other “good” study habits. If a student is intelligent, this quality may encourage the student to study longer and better, which could result in higher classroom performance. In such a case, the significant correlation between  $K$  and  $E$  would be explained by the  $K \rightarrow M \rightarrow E$  sequence of relationships. We could then say that study effectiveness mediates the relationship between student’s intelligence and classroom performance.

Indirect relationships, and thus mediation, commonly appear in structural models. If we view exogenous constructs as the “inputs” to our model explaining some final “outcome” represented by an endogenous construct, then any constructs going between these correspond to a definition of mediation in some way. A model proposing mediation that exhibits good fit provides evidence that the mediation exists.

**Testing for Mediation** Mediation requires significant correlations among all three constructs. Theoretically, a mediating construct facilitates the relationship between the other two constructs involved. If the mediating construct completely explains the relationship between the two original constructs (e.g.,  $K$  and  $E$ ), then we term this **complete mediation**. But if we find that there is still some of the relationship between  $K$  and  $E$  that is not explained away by the mediator, then we denote this as **partial mediation**.

A researcher can determine if mediation exists, and whether it is complete or partial, in several ways. First, if the path labeled  $C$  is expected to be zero due to mediation (representing complete mediation), a SEM model can represent mediation by including only the paths  $A$  and  $B$  in the model. It would not include a path directly from  $K$  to  $E$ . If the estimated model suggests the sequence  $K \rightarrow M \rightarrow E$  provides a good fit, it would support a mediating role for  $M$ . In addition, the fit of this model could be compared with the SEM results of a model including the  $K \rightarrow E$  path ( $C$ ). If the addition of path  $C$  improves fit significantly, as indicated by the  $\Delta\chi^2$ , then complete mediation is not

supported. If the two models produce similar  $\chi^2$ , and the  $\Delta\chi^2$  is relatively small and perhaps even not significant, then mediation is supported. The researcher also can compare the relative fit indices such as the CFI and RMSEA in determining which of the alternative models provides the best fit.

Because the conclusion is not always clear, a series of steps can be followed to evaluate mediation. These steps apply whether using SEM or any other general linear model (GLM) approach, including multiple regression analysis. Using the previous mediation diagram, the steps are [17]:

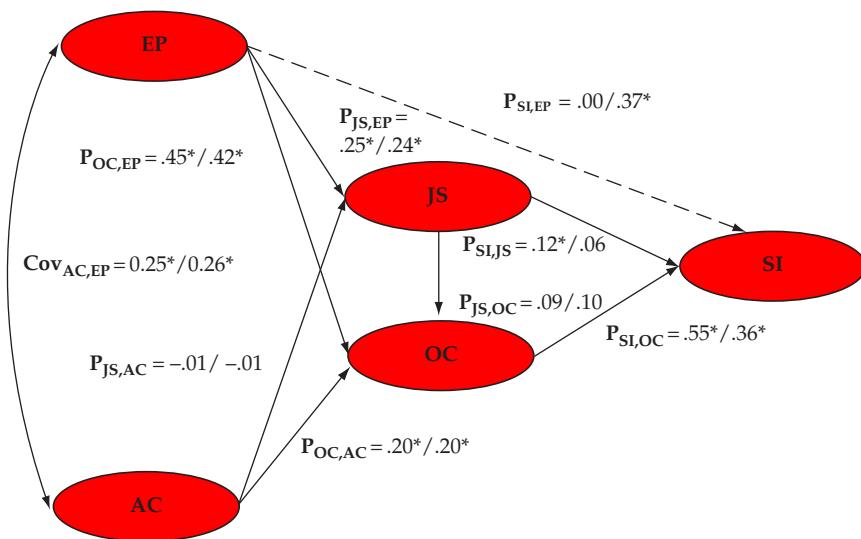
- 1** Establish that the individual relationships have statistically significant relationships:
  - a** *K is related to E:* Here we are establishing whether a relationship exists. The test of mediation proceeds whether this relationship is significant or not [55].
  - b** *K is related to M:* Here we establish that the mediator is related to the “input” construct.
  - c** *M is related to E:* Here we establish that the mediator does have a relationship with the outcome construct.
- 2** Estimate an initial model with only the direct effect (C) between K and E. Then estimate a second model adding in the mediating variable (M) and the two additional path estimates (A and B). Then assess the extent of mediation as follows:
  - a** If the relationship between K and E (C) remains *significant and largely unchanged* once M is included in the model as an additional predictor (K and M are both modeled as predicting E), then mediation is not supported.
  - b** If C is *reduced but remains significant* when M is included as an additional predictor, then partial mediation is supported.
  - c** If C is reduced to a point where it is *not statistically significantly* after M is included as a mediating construct, then **full mediation** is supported.

However, keep in mind that the test of fit based on model chi-square for the mediation model, along with a significant and non-trivial indirect relationship (indirect relationship = A  $\times$  B), provides the strongest evidence for mediation or the lack thereof.

**Other Causes** Recall the issue of endogeneity introduced in Chapter 9. The mediation process places attention on the fact that if a model is underspecified in a way that some unmeasured variable or construct exists that affects the constructs involved in the mediation sequence, then the parameter estimates’ corresponding standard errors and *t* values may be biased [12]. One check for such bias is to examine the residuals from a model. If the residuals are random, meaning no pattern is evidence, then one need not be concerned about bias. However, if evidence suggests under-specification through non-random residuals, then bootstrapping provides an alternative to conventional *t*-tests for the significance of parameter estimates. **Bootstrapping** produces *t* values through a non-parametric process involving re-estimation of the model hundreds or thousands of times by sampling with replacement from the original sample to produce each iteration’s input. Presuming the bootstrapping process involves 500 re-estimations of the model, each parameter coefficient is estimated 500 times. Bootstrapping uses a *t*-test to examine whether the average of those estimates is significantly different from 0 to draw conclusions about the strength of a relationship.

Most SEM software provides an option for bootstrapped *t* values. In addition, the PROCESS macro facilitates bootstrapping and computes indirect effects across a wide range of multiple regression mediation models for standard GLM applications in SPSS or SAS, but does not apply to SEM [27]. When the researcher can identify the potential for endogeneity through the residuals, the effect of unmeasured causes can be modeled in SEM software by freeing covariances among the corresponding construct error-variance (or variance for exogenous constructs) terms. In effect, the free covariance terms serve the role of a control function and address the concern of bias due to an unmeasured cause, providing an alternative to bootstrapping. Also, if the addition of the covariance term(s) fails to improve model fit, then the concern for problematic endogeneity is reduced.

**HBAT Illustration of Mediation** The HBAT model shown in Figure 12.4 hypothesizes several mediating effects. The relationships of both exogenous constructs (Environmental Perceptions [EP] and Attitudes Toward Coworkers [AC]) with Staying Intentions (SI) are hypothesized to be fully mediated by two endogenous constructs—Job Satisfaction (JS) and Organizational Commitment (OC). The model hypothesizes that the effects of EP and AC can be fully



**Figure 12.4**  
Adding Direct Effects For  
Testing Mediation in the HBAT  
Employee Retention Model

Values represent parameter estimates for initial/respecified (direct effect added) models.

\*Statistically significant at .05 level.

explained through these two mediating constructs. In Chapter 11, we examined model fit as an explanation of the overall set of relationships and confirmed that model. But we can also explicitly test for mediation, as described above.

Let us examine the possibility that the relationship between EP and SI is mediated by the two constructs of JS and OC. To do so, we'll follow the two-step process just described.

**STEP 1: ESTABLISH SIGNIFICANT RELATIONSHIPS BETWEEN THE CONSTRUCTS** In this situation, we can refer back to Table 10.7, which provided the correlations between constructs in the CFA model. From that analysis, we can first see that EP is significantly related to SI (.56), ensuring that the direct, unmediated relationship was significant. We also find that EP was significantly related to both JS (.24) and OC (.50), establishing a relationship with both potential mediators. Finally, SI is significantly related to both JS (.23) and OC (.55), thus supporting relationships between the mediators and the outcome variable.

**STEP 2: ESTIMATE THE MEDIATED MODEL AND ASSESS LEVEL OF MEDIATION AND INDIRECT EFFECTS** In this analysis, we will first use the original Employee Retention model, which did not estimate the direct effect from EP to SI. To assess if adding the direct effect would substantially change the model fit, we can estimate a revised HBAT model (dotted path in Figure 12.4), which adds the direct path between EP and SI.

The revised model with the direct relationship yields a significant decrease in chi-square ( $\Delta\chi^2 = 41.2, df = 1, p = .00$ ), a substantive improvement in model fit, and a significant path estimate for the EP → SI relationship (see Table 12.4). These results suggest that there is not complete mediation. But is there partial mediation? To establish partial mediation, we will need to identify a significant indirect effect leading from EP to SI through the mediating variables. A number of potential compound paths are possible, but only three reflect indirect causal mediated effects: (1) EP → OC → SI; (2) EP → JS → OC → SI; and (3) EP → JS → SI. If one or more of these indirect effects contains paths that are significant, then the model supports partial mediation. By this we mean that there is a significant direct relationship between EP → SI, but there is also a significant indirect effect through the mediator.

In the revised model the path estimates between EP and both mediators (JS and OC) are still significant (Table 12.4). Although one mediator still has a significant relationship with SI (OC → SI is significant), the other relationship from a mediator to SI (JS → SI) becomes nonsignificant. Does this mean that there is not partial mediation? No, it just means that while JS does not act as a mediator, OC can still have a mediating effect. Note that in the three indirect mediated effects, the nonsignificant relationship (JS → SI) is only a part of the third effect. The other two indirect mediating effects still have all of the individual paths statistically significant. Thus, the model supports the finding that OC provides partial mediation of the relationship between EP and SI.

**Table 12.4** Testing for Mediation in the HBAT Employee Retention Model

Model Element	HBAT Employee Retention Model	Revised Model with Direct Effect
Model fit		
$\chi^2$ (chi-square)	283.4	242.2
Degrees of freedom	181	180
Probability	0.00	0.00
RMSEA	.036	.029
CFI	.99	.99
Standardized parameter estimates		
EP → JS	0.25*	0.24*
EP → OC	0.45*	0.42*
JS → SI	0.12*	0.06
JS → OC	0.09	0.10
OC → SI	0.55*	0.36*
EP → SI	Not estimated	0.26*

\*Statistically significant at .05 level.

The magnitude of the mediating effects is demonstrated by breaking down the total effects into direct and indirect effects. Most software programs provide this decomposition of effects. However, the researcher can also calculate the effects. The interested reader is referred to the Basic Stats appendix on the text's online resources where this process is described in more detail. Table 12.5 provides a breakdown of the effects of EP → SI both in the original HBAT model (no direct effects from EP → SI) and the revised HBAT model (direct effect added for EP → SI). As we can see in the original model, there are non-trivial indirect effects, thus supporting the presence of mediating effects of JS and OC. The remaining question is: Do those effects remain when the direct path is added in the revised model? As we can see, although the indirect effects do decrease, they are still significant and represent a substantial portion of the total effects. Also note that the direct effect is significant, adds considerably to the total effects, and constitutes the majority of the total effects, making this a partial mediation situation. Further, in this case, the bootstrapped *t* values are consistent with the conventional values and would lead to no differences in interpretation.

The magnitude of any individual mediating effect can be calculated using the process for calculating indirect effects. If this is done, you will find that most of the mediating effect comes from the EP → OC → SI mediating relationship. Given this post hoc theoretical analysis, HBAT needs to cross-validate this result with new data before considering it reliable. However, managerial implications for each of the supported hypotheses can be developed based on the overall positive results.

## MODERATION

A **moderating effect** occurs when a third variable or construct changes the relationship between two related variables/constructs [2]. For example, we would say that a relationship is moderated by gender if we found that the relationship between two variables differed significantly between males and females. For example, the relationship

**Table 12.5** Assessing Direct and Indirect Effects in a Mediated Model

Effects <sup>a</sup> of EP → SI	Original HBAT Model (Only Indirect Effects)	Revised HBAT Model (Indirect and Direct Effects)
Total effects	.29	.55
Direct effects	.00	.37
Indirect effects	.29	.18

<sup>a</sup>Values in the table represent standardized effects.

between two variables may be negative for males and positive for females or significant in one group and not the other. In this type of situation, we would need to know whether the respondents were males or females before we could accurately estimate and interpret the relationship.

We have discussed moderators in the other multivariate techniques. In multiple regression, for example, there were interaction terms where the regression coefficient changed based on the value of a second variable. In ANOVA/MANOVA, interaction effects were used to assess whether the differences between groups were constant across the values of another variable. In both examples, the interaction effects play the role of a moderator.

Moderating variables must be chosen with strong theoretical support. Problems in examining moderation occur as the moderator becomes correlated with either of the variables in the relationship. Therefore, analysis of moderators is easiest when the moderator has no significant linear relationship with either exogenous or endogenous constructs [2, 17, 25]. The lack of a relationship between the moderator and the other constructs helps distinguish moderators from mediators (remember that the mediator must be related to both constructs in the relationship being mediated). In non-experimental relationship, multicollinearity problems present themselves in moderated regression analysis as the interaction term is a mathematical function of two independent variables. Thus, they are related by mathematical definition. As a consequence, in multiple regression moderation analysis relies on hierarchical regression analysis using standardized variables where the interpretation depends on the change in model F and  $R^2$  with the addition of the moderating variable rather than on the parameter coefficients, which may well be biased due to multicollinearity.

**Nonmetric Moderators** A moderator variable can be metric or nonmetric. Nonmetric, categorical variables often are hypothesized to be moderators. These moderators typically are classification variables of some type. One common type of moderator is respondent characteristics, such as gender, age, or other characteristic. Differing situations or contexts are another type of categorical moderator. A common example would be cross-cultural studies where country-of-origin becomes a moderating variable. Similarly, dividing respondents into current customers versus non-customers would be using customer status as a moderating variable.

As noted earlier, theory is important in evaluating a moderator because a researcher should find some reason to expect that the moderator changes a relationship. Researchers have any number of ways for dividing the sample into groups, but the selection of a moderator should not be based on whether it demonstrates significant moderating effects, but rather on its theoretical foundation.

Once the moderating variable is selected, groups of respondents can be defined and multigroup analysis applied. The basic procedure described earlier in the discussion of invariance testing provides results, with the primary difference being that now the focus is on structural model fits and structural path estimates rather than on measurement model fits.

**Metric Moderators** A moderator can also be a continuous/metric variable but treated as categorical if it can be divided into several sub-groups. If the continuous variable can be categorized in a way that makes sense (i.e., is based on theory or logic), then groups can be created and the same procedures used for nonmetric moderators can be applied. For instance, if the continuous variable shows bimodality (i.e., the frequency distribution shows two clear peaks rather than one), then logical groups could be created around each mode. As an example, job satisfaction could be measured metrically, but if there is a highly satisfied group versus a not satisfied group, then the two groups defined by their feelings about satisfaction would be used as a moderating variable. Cluster analysis also might be useful to form groups. However, if the moderator variable displays a clear unimodal distribution (one peak), then grouping is not justified. It is possible that some fraction (i.e., one-fourth to one-third) of the observations around the median value could be deleted and the remaining observations (which are likely now bimodal) put into groups. Treating normally distributed variables as categorical can be criticized for loss of information. Another drawback is the need for a relatively large sample given some is thrown out. The advantage is that multigroup analysis can be used as an intuitive way of testing and demonstrating moderation.

Researchers also can model a metric moderator by creating interaction terms similar to what is required with a regression approach. Using regression terminology, the independent variable can be multiplied by the moderator to create an interaction term. In standard SEM, the measured item indicators of the exogenous construct(s) and the moderator would be multiplied to each other. These product terms can be used to indicate the interaction. However, taking this approach with multiple-item constructs is complicated by numerous factors, not the least of which is the violation of the assumption that the indicators of the interaction and the exogenous construct and the moderator are unrelated. Standardization sometimes helps reduce, but often does not eliminate, multicollinearity. Although many approaches exist to implement continuous interactions in regression models, we advise use of the nonmetric multigroup approach unless it simply cannot be justified. Multigroup moderation allows an easily understood presentation of results.

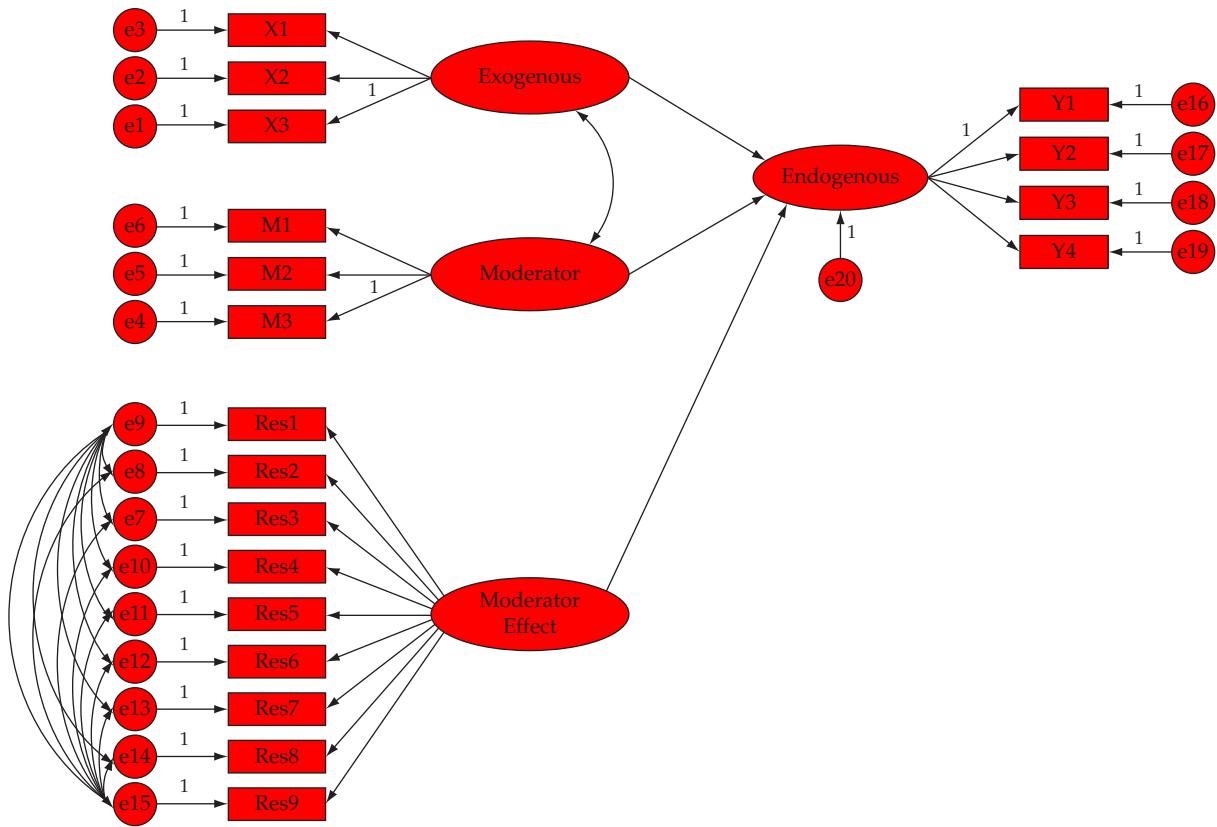
If the nature of the data simply does not allow for grouping, then some alternative approaches exist for operationalizing a moderator within a single SEM model. One approach involves using a method that forces the construct representing the moderator to be independent of the exogenous construct [34]. The orthogonal moderator approach involves these steps, presuming one exogenous construct, one moderator construct, and one endogenous construct [50]:

- 1 Create intermediate interaction indicators by multiplying each measured variable indicator for the exogenous construct by each measured variable indicator of the moderator. If each construct has three indicators, nine products result.
- 2 Run a multiple regression model using each product above as a dependent variable and each exogenous construct and moderator construct indicator as independent variables.
- 3 Save the residuals from each regression as new variables in the dataset along with the other construct indicators. The residuals retain the variation of the product term.
- 4 Include the residuals for each regression equation in the covariance matrix to be analyzed in the SEM model. The result will be a set of indicators that will represent the moderating effect.
- 5 Set up the structural model as before with both the exogenous construct and the moderator construct predicting the endogenous construct. This part is just like setting up a structural model as described in the previous chapter.
- 6 Use the residual variables as indicators of a new construct that represents the moderating effect. Given that the moderator effect construct is represented by residual terms, it is orthogonal (independent) to the exogenous construct and moderator construct. Thus, there is no need to correlate this moderator effect with the exogenous or moderator construct.
- 7 Given that the nature of the computations involved, the residual indicators themselves correlate with each other and thus the error-variance terms among the indicators of the moderator effect construct should be freed to correlate with each other. However, there is no need to remove the constraints of independence on the indicators of the exogenous construct of moderator construct.
- 8 Estimate the model and interpret as described in the previous chapter. Figure 12.5 depicts the process.

Although the approach can be useful when needed, as a model becomes more complex, properly depicting the model becomes more tedious. Thus, we suggest using the approach only when the multi-group method is not feasible. We describe that process here.

**Using Multigroup SEM to Test Moderation** Multigroup SEM is used to test moderating effects when the moderating variable is either nonmetric or a metric moderator has been transformed into a nonmetric variable (see previous description above). Moderation typically involves the testing of structural model estimates. Thus, the process becomes an extension of the multigroup analysis for testing measurement invariance. When the moderating variable is culture or some other variable representing populations that may respond to scales differently, metric invariance must be first established. On the other hand, if the different groups are from the same population, such as when they speak the same language, share characteristics in common etc., metric invariance is not needed because there is no reason to suspect different reactions to the scale items [2].

**Figure 12.5**  
Orthogonal Moderating Construct Approach



The multigroup structural model proceeds much like invariance testing in CFA. The first group model is estimated with path estimates calculated separately for each group. This is identical to the TF (totally free) model described earlier. Then a second group model is estimated where the path estimate of interest is constrained to be equal between the groups. Comparison of the differences between models with a chi-square difference test ( $\Delta\chi^2$ ) indicates if the model fit decreased significantly (i.e., an increase in chi-square) when the estimates were constrained to be equal. A statistically significant difference between models indicates that the path estimates were different (i.e., model fit was significantly better when separate path estimates were made) and that moderation does exist. If the models are not significantly different, then there is no support for moderation (because the path estimates were not different between groups). When testing for moderation, the researcher is looking for significant differences in the two models to support the hypothesis of differences in the path estimates. The researcher should also examine the path estimates in question to assess if the differences in both group models are theoretically consistent.

**HBAT Illustration of Moderation** The HBAT management team suspects that men and women may not exhibit the same relationships in Attitudes toward Coworkers to job satisfaction relationship. Theory suggests a gender difference in this relationship, whereby the effect would be greater among women relative to men. This was reinforced when the Employee Retention model discussed in Chapter 10 found a nonsignificant relationship between AC  $\rightarrow$  JS. Questions arose that perhaps the “common” path estimate from combining both groups was not reflecting the actual differences between groups. As a result, the HBAT research team decided to conduct a multigroup analysis using the gender classification variable.

**Table 12.6** Testing for Gender as a Moderator in the HBAT Employee Retention Model

Model Characteristic	Unconstrained Group Model (TF for Each Group)	Constrained Group Model (AC & JS Equal Across Groups)	Model Differences
Model fit			
Chi-square	401.1	412.2	11.1
df	360	361	1
CFI	0.99	0.99	—
RMSEA	0.024	0.027	—
Path estimate ( $P_{JS,AC}$ )	.24 (female)* -.17 (male)*	-.01 (combined)	

\*Significant at .05 level.

The multigroup CFA established metric invariance, as described earlier in this chapter. This was sufficient to now test for moderation in the relationships between constructs, specifically the AC → JS relationship.

Following the same steps used to specify the two-group CFA model testing for differences based on gender, a two-group structural model was set up. The TF structural model estimates an identical structural model in both groups simultaneously. The model fit statistics and path estimates for the AC → JS relationship are shown in Table 12.6. Then a second group model is estimated, the only difference being that the AC → JS path estimate is constrained to be equal in both groups. The fit results and path estimates are also shown in Table 12.6.

Both models show acceptable fit indices (CFI and RMSEA) indicating their overall acceptability. The chi-square difference between models ( $\Delta\chi^2$ ) is 11.1 with one degree of freedom. This is significant ( $p < .001$ ), indicating that constraining the AC → JS path estimate to be equal between groups produces worse fit. Therefore, the unconstrained (TF) model in which the AC → JS relationship is freely estimated provides better fit. This result suggests that gender does moderate the relationship between AC and JS.

Looking at the standardized parameter estimates for the TF results, HBAT researchers find that the AC → JS relationship is significant in both groups. As predicted, the relationship is greater for women, with a completely standardized estimate of 0.24, as compared to a completely standardized estimate of -0.17 for men. Thus, it seems that attitudes toward coworkers is positively related to job satisfaction among women, but negatively related to job satisfaction among men. Moreover, the nonsignificant path estimate from the combined model (males and females together) may lead to the incorrect conclusion that AC is not related to JS. The moderated relationship “cancels” out the different effects of males and females when estimated together. The result is a clear case of moderation where the nature of a relationship (AC → JS in this case) changes based on a third variable (gender).

## Developments in Advanced SEM Approaches

### LONGITUDINAL DATA

SEM is increasingly applied to longitudinal data. Given the added insight from tracking changes in constructs and relationships over time, the increasing use of longitudinal data may be beneficial in many fields. Because many different types of longitudinal study designs lead to many different SEM applications, this section provides only a brief introduction to some of the key differences in dealing with longitudinal data. The interested reader is referred to other sources for a more detailed discussion and review [23].

### LATENT GROWTH MODELS

One of the key issues in modeling longitudinal data with SEM involves added sources of covariance associated with taking measures on the same units over time. For instance, consider a model hypothesizing that study habits relate to academic performance. If one has data for the same students over same years, the study habits and academic

## Mediation and Moderation

Mediation involves the comparison of a direct effect between two constructs while also including an indirect effect through a third construct

Good fit for a model implying mediation provides support for mediation

Full mediation is found when the direct effect is nonsignificant in the presence of the indirect effect, whereas partial mediation occurs when the direct effect is reduced, but still significant

Although individual indirect effects that are small (less than .1) are not of interest because they are likely trivial, the combined total of all indirect effects may be substantial

Moderation by a classification variable can be tested with multigroup SEM:

A multigroup SEM first allows all free (unconstrained) structural parameters to be estimated freely

Then, a second model is estimated in which the relationships that are thought to be moderated are constrained to be equal in all groups

If the second model fits as well as the first, moderation is not supported. If its fit is significantly worse, then moderation is evident

The multigroup model is convenient for testing moderation:

If a continuous moderating variable can be collapsed into groups in a way that makes sense, then groups can be created and the procedures described previously can be used to test for moderation

Cluster analysis may be used to identify groups for multigroup comparisons

Unimodal data should not be split into groups based on a simple median split

When a continuous variable moderator must be used, consider the orthogonal moderator effect approach.

performance from one year are almost assuredly related to the study habits and academic performance of the previous year. The extra source of covariance complicates matters and creates a situation much like that of the panel data models discussed in an earlier chapter.

Although multiple approaches exist to handle longitudinal data in SEM models, perhaps the most applied is something referred to as a **latent growth model**. The **latent growth model** specifically focuses on “growth” in concepts over time. More generally, how do levels of constructs change over time and why? A detailed discussion of latent growth models is beyond the scope of the text. Be aware however that major SEM software (LISREL, AMOS, Mplus, lavaan, etc.) incorporates applications facilitating latent growth models.

## BAYESIAN SEM

Researchers increasingly recognize that statistical meaningfulness need not always be defined by statistical significance and  $p$  values [4]. Although the bulk of this text takes what some refer to as frequentist statistics approach, with a focus on  $p$ -values that represent the probability of a given finding based on sample characteristics, an alternative approach lies in Bayesian statistics. Bayesian statistics rely on conditional probabilities and treat samples as constant and parameter estimates as variable. Thus, Bayesian results are not defined by traditional statistical significance tests and are not impacted by sample size. One can think of Bayesian results as information learned to adjust prior beliefs. The new information often is referred to as posterior, which comes after the analysis.

Bayesian SEM represents an alternative SEM approach in the Bayesian tradition. Should the research have concerns about small sample sizes, problematic measured variable distributions, Bayesian SEM is an alternative. Bayesian SEM also allows the user to incorporate prior information into the analysis. SEM software packages including Mplus, LISREL, AMOS, and a partner program to the R-package lavaan (namely, blavaan) provide a mechanism for conducting Bayesian SEM. A detailed treatment of Bayesian SEM requires a background in Bayesian statistics,

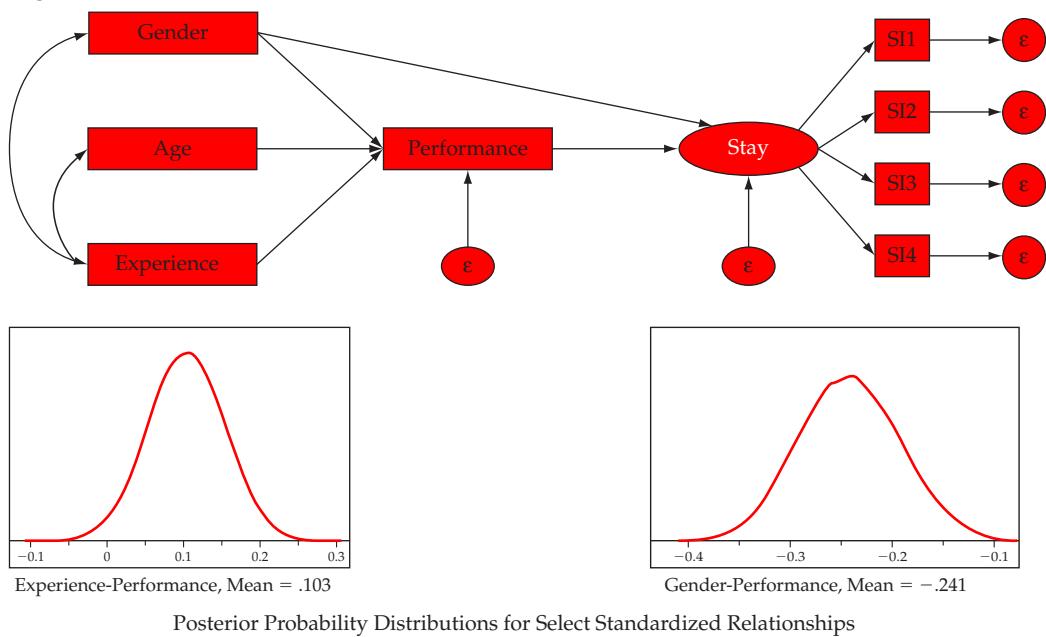
Markov chains, and Markov chain simulation. Thus, the reader is referred to other sources for a detailed description [42]. More advanced applications may also require specialized software such as Winbugs [44]. The process of Bayesian SEM is computationally intense and even with fast microprocessor speeds, model estimation takes much longer than the other statistical models discussed in the book.

However, the Bayesian software applications provide for a familiar mechanism compared to conventional SEM in conducting a model test. The key statistics in interpretation shift however. The Bayes factor, Deviance Information Criterion, and the Posterior predictive *p*-values help interpret model fit, facilitate alternative model fit comparisons, and assist in examining parameter estimates [1]. Parameter estimates are not interpreted in terms of statistical significance but in terms of how likely each would be to occur given the data. Parameter estimates can be interpreted with an alternative to confidence intervals known as a **competence interval**. The competence interval provides a 95 percent probability (the typical default) that a parameter values lies between an upper and lower value in a given population (represented by the sample). Often time, these competence intervals are depicted as posterior density graphs. Thus, although a Bayesian approach may seem quite different philosophically, the interpretation is not entirely different than traditional SEM analysis.

Figure 12.6 displays a structural model examining HBAT employees intention to continue working for the company as a function of job performance, serving in a mediating role, facilitating the relationship between experience and continuing on the job. The theory is that experience leads people to stay on the job, but only if the employees perform well. Because the researcher is concerned that job performance measures can be very skewed, a Bayesian analysis is included. Gender and age are included as control variables. First, consistent with good practice, the researcher conducts a conventional CB-SEM analysis. An initial model that did not include the gender-staying relationship provided poor fit and the residuals suggested adding a direct relationship from gender to stay.

The model fit results of this solution are:  $\chi^2 = 31.6$ , with 18 df ( $p = 0.025$ ), which yields a CFI = 0.986, with a RMSEA = 0.043. Thus, the fit appears good based on the guidelines from Chapter 9, but with such a simple model, one might expect an insignificant  $\chi^2$  test. Three structural relationships yield significant parameter estimates. The Performance-Stay standardized parameter estimate is 0.110, the Experience-Performance parameter estimate is

**Figure 12.6**  
Bayesian SEM Illustration



Posterior Probability Distributions for Select Standardized Relationships

0.104, and among the control relationships, the Gender-Stay relationship is  $-0.241$ . Thus, the results present modest support for the mediation theory because the relationships making up mediation are statistically significant, although weak. These results can be compared to the Bayesian results.

Bayesian fit is indicated by the Posterior Predictive P (PPP). Values closer to 0.5 suggest better fit than values further from 0.5. In this case, the PPP = 0.13 and the DIC = 83.9. While both of these become more obviously useful in comparing model fit, the PPP in particular may not be as supportive of the fit as was the conventional fit results. In this particular case, the standardized Bayesian parameter estimates are the same as the ML estimates to three decimal places. Figure 12.6 shows the posterior distributions for the Experience-Performance and Gender-Performance relationships. The corresponding competence intervals are [0.070:0.138] and [-.275: -.208], respectively. Thus, in neither case is 0 in the competence interval. The results support these two relationships. One interesting substantive finding depends on the coding of the gender variable. Men are coded with a value of 0 and women with a value of 1 in this particular data. Thus, the results suggest that women express lower intentions to continue working for HBAT. The illustration points also to the comparison of the conventional SEM and Bayesian SEM results. In this case, the results are very similar, which is comforting to the researcher. When the two produce drastically different results, the researcher must be certain to know why the results differ before drawing strong conclusions. Bayesian SEM becomes relatively preferable when the study execution departs substantially from the assumptions for conventional SEM analysis.

The widespread use of SEM models in almost every discipline has also heightened interest in using SEM methods for more advanced issues. Whether using higher-order factor models, testing mediation or moderation, or assessing construct invariance across groups, SEM researchers can utilize the flexibility of SEM models to address all these questions and more. Yet as these new applications arise, the researcher must be cautious and fully understand the theoretical foundations for the approach and the issues of estimation and interpretation that are involved. As SEM models become more accepted their use for these more specific research questions will become more widespread as well.

It is difficult to highlight in a paragraph or two all of the issues concerning these advanced topics. However, some important points that will help in understanding the issues and modeling benefits include those corresponding to the objectives of the chapter:

**Understand the differences between reflective and formative measurement approaches.** Reflective measurement remains consistent with psychometric theory and practice and assessments of measurement model validity include fit, convergent validity, discriminant validity, and nomological/predictive validity. Reflective scales represent latent constructs. Formative scales do not depend on the same validation processes and in particular, evidence of convergent validity does not support a formative scale. Formative scales do not have to represent a latent construct and are often referred to as indices. Formative scales present issues with statistical identification and interpretational confounding.

**Specify formative scales in SEM models.** A formative scale becomes statistically identified when (by) either including two reflective indicators or by including path coefficients to two reflectively scaled latent constructs, which serve as “outcomes” or dependent variables and parameterized the formative index. Thus, the loadings of the formative scale are determined based on the “outcome” measures used. As such, loadings can vary markedly if different outcomes are used.

**Identify when higher-order factor analysis models are appropriate.** The most common higher-order model is the second-order model. A second-order latent factor(s) causes multiple first-order latent factors, which in turn cause the measured variables ( $x$ ). In a simple sense, the first-order latent constructs become the “indicators” of the second-order latent construct. As one would expect, the higher-order construct is more abstract because it has only latent constructs as its indicators. When theoretical support can be found for a higher-order model, then any SEM program can estimate the higher-order model.

**Know how SEM can be used to compare results between groups.** Multigroup comparisons can be useful for examining both the measurement and structural models. They require that the researchers test their hypotheses of group differences with between-group constraints representing the various degrees of measurement model invariance or equality of structural relationships between the groups. The  $\Delta\chi^2$  is a primary statistic for testing invariance and for drawing conclusions about the differences between groups.

**Use multigroup methods to perform an invariance measurement analysis.** Invariance testing is performed through a six-stage process. At each stage, similar models are estimated for each group, and measures of model fit are calculated for the collective set of models. For each stage, a specific type of between-group constraint is specified, adding a new type of constraint to the previous model. So each model becomes more constrained. The configural invariance model only imposes the same structure on each group with each individual loading estimate totally free to take on its own value. The metric invariance model involves more constraints, with the added constraints fixing the corresponding loading parameters to be equal between groups. So, in the case of metric invariance, if the chi-square difference from the configural invariance model is not statistically significant, the results support full metric invariance indicating that the factor loadings are the same across the groups. This chi-square difference test is tested at each stage where a more restrictive model is specified. If the chi-square difference test is significant (meaning that full invariance is not supported), then the researcher can try and achieve partial invariance. A simpler alternative to searching for partial invariance is to drop an offending item. Scalar invariance can be tested by adding constraints that the intercept terms for the measured variable equations are equal between groups to the metric invariance model and compared in the same way.

**Understand the concepts of statistical mediation and moderation.** Several different types of relationships were discussed. In particular, the concepts of mediation and moderation were explained. Mediation involves a sequencing of relationships so that some construct intervenes in a sequence between two other constructs. Moderation involves changes in relationships based on the influence of some third variable or construct. Moderation was discussed in the context of multigroup SEM models and continuous variable interactions. Whenever possible, the multigroup approach is recommended but the orthogonal moderator approach may be an alternative when continuous variables must be used.

**Developments in SEM.** The chapter briefly introduced two SEM applications seeing increased usage. Latent growth analysis provides a way of analyzing longitudinal data using SEM. The latent growth analysis follows the procedures for analyzing panel data and allows one to model changes in construct values over time. Bayesian SEM provides an alternative to the frequentists statistical approach to conventional SEM. The Bayesian approach turns things around and asks how likely is the model given the data? Bayesian SEM also can be useful when the researcher is confronted with severely skewed or kurtotic data distributions or when the researcher has concerns that the sample size is too small. Parameter estimates are interpreted using posterior distributions and competence intervals rather than  $p$  values.

What conditions make a second-order factor model appropriate?

What conditions must be satisfied to draw valid conclusions about differences in relationships and differences in means between three different groups of respondents—one from Canada, one from Italy, and one from Japan? Explain.

An interviewer collects data on automobile satisfaction. Ten questions are collected via personal interview. Then, the respondent responds to another 20 items by marking the items using a pencil. How can CFA be used to test whether the question format has biased the results?

What is meant by a *formative scale*? Can programs like LISREL, AMOS, and Mplus accommodate formative scales?

How does formative theory differ from reflective measurement theory? Discuss both in specification/estimation and interpretation?

What conditions make a second-order factor model appropriate?

What conclusions can be drawn from measurement invariance testing?

What are the most common uses of multigroup testing?

Describe the process of multigroup testing in general terms.

What is a major concern when using SEM techniques with longitudinal data?

Draw a structural model hypothesizing that three exogenous constructs—X, Y, and Z—each affects a

mediating construct, M, which in turns determines two other outcomes, P and R. What is an indirect effect?

How can SEM test for a moderating effect? Name at least two ways.

Describe differences between conventional SEM and Bayesian SEM.

A list of suggested readings and other materials relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com)).

- 1 Assaf, A.G., M. Tsionas, and H. Oh. 2018. The Time Has Come: Bayesian SEM Estimation in Tourism Research. *Tourism Management* 64: 98–109.
- 2 Babin, B. J., J. B. Boles, and D. P. Robin. 2000. Representing the Perceived Ethical Work Climate Among Marketing Employees. *Journal of the Academy of Marketing Science* 28: 345–59.
- 3 Babin, B. J., A. Borges, and K. James. 2016. The Role of Retail Price Image in a Multi-country Context: France and the SA. *Journal of Business Research* 69: 1074–81.
- 4 Babin, B. J., M. Griffin, and J. F. Hair. 2016. Heresies and Sacred Cows in Scholarly Marketing Publications. *Journal of Business Research* 69: 3133–8.
- 5 Baron, R. M., and D. A. Kenny. 1986. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations. *Journal of Personality and Social Psychology* 51: 1173–82.
- 6 Bentler, P. M. 1980. Multivariate Analysis with Latent Variables: Causal Modeling. *Annual Review of Psychology* 31: 419–56.
- 7 Blaha, John, S. P. Merydith, F. H. Wallbrown, and T. E. Dowd. 2001. Bringing Another Perspective to Bear on the Factor Structure of the Minnesota Multiphasic Personality Inventory-2. *Measurement and Evaluation in Counseling and Development* 33: 234–43.
- 8 Bollen, K. A. 2002. Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology* 53: 605–34.
- 9 Bollen, K., and R. Lennox. 1991. Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin* 110: 305–14.
- 10 Bollen, K. A., and K. Ting. 2000. A Tetra Test for Causal Indicators. *Psychological Methods* 5: 3–32.
- 11 Borsboom, D., G. J. Mellenbergh, and J. van Heerden. 2003. The Theoretical Status of Latent Variables. *Psychological Review* 110: 203–19.
- 12 Bullock, J. G., D. P. Green, and S. E. Ha. 2010. Yes, but what's the Mechanism? (Don't Expect an Easy Answer). *Journal of Personality and Social Psychology* 98: 550–8.
- 13 Burt, R. S. 1976. Interpretational Confounding of Unobserved Variables in Structural Equations Models. *Sociological Methods Research* 5: 3–52.
- 14 Byrne, B. M., R. J. Shavelson, and B. Muthén. Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin* 105: 456–66.
- 15 Cattell, R. B. 1956. Validation and Intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology* 12: 205–14.
- 16 Cheung, G. W., and R. B. Rensvold. 2002. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling* 9: 233–55.
- 17 Cohen, J., and P. Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum Associates.
- 18 DeVellis, R. F. 1991. *Scale Development: Theory and Applications*. Newbury Park, CA: Sage.
- 19 Diamantopoulos, A., P. Riefler, and K. P. Roth. 2008. Advancing Formative Measurement Models. *Journal of Business Research* 61: 1203–18.
- 20 Diamantopoulos, A., and J. Siguaw. 2006. Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. *British Journal of Management* 17: 263–82.
- 21 Diamantopoulos, A., and H. M. Winklhofer. 2001. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research* 38: 269–77.
- 22 Drasgow, F. 1984. Scrutinizing Psychological Tests: Measurement Equivalence and Equivalent Relations with External Variables Are the Central Issues. *Psychological Bulletin* 95: 134–5.
- 23 Ferrer, E., F. Hamagami, and J. J. McArdle. 2004. Modeling Latent Growth Curves with Incomplete Data Using Different Types of SEM and Multi-Level Software. *Structural Equation Modeling* 11: 452–83.

- 24 Fuller, C. M., M. J. Simmering, G. Atinc, Y. Atinc, and B. J. Babin (2016). Common Methods Variance Detection in Business Research. *Journal of Business Research*, 69, 3192–8.
- 25 Gogineni, A., R. Alsup, and D. F. Gillespie. 1995. Mediation and Moderation in Social Work Research. *Social Work Research* 19: 57–63.
- 26 Griffin, M., B. J. Babin, and D. Modianos. 2000. Shopping Values of Russian Consumers: The Impact of Habituation in a Developing Economy. *Journal of Retailing* 76: 33–52.
- 27 Hayes, A. F. 2017. *Introduction to Moderation, Mediation, and Conditional Process Analysis*. New York: Guilford Press.
- 28 Heise, D. R. 1972. Employing Nominal Variables, Induced Variables, and Block Variables in Path Analysis. *Sociological Research Methods* 1: 147–73.
- 29 Horn, J. L. 1991. Comments on “Issues in Factorial Invariance”. In L. M. Collins and J. L. Horn (eds.), *Best Methods for the Analysis of Change*. Washington, DC: American Psychological Association, pp. 114–25.
- 30 Howell, R. D., E. Breivik, and J. B. Wilco. 2007. Reconsidering Formative Measurement. *Psychological Methods* 12: 205–18.
- 31 Hui, C. H., and H. C. Triandis. 1985. Measurement in Cross-Cultural Psychology: A Review and Comparison of Strategies. *Journal of Cross-Cultural Psychology* 16: 131–52.
- 32 Jarvis, C. B., S. B. Mackenzie, and P. M. Podsakoff. 2003. A Critical View of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. *Journal of Consumer Research* 30: 199–218.
- 33 Landis, R. S., D. J. Beal, and P. E. Tesluk. 2000. A Comparison of Approaches to Forming Composite Measures in Structural Equation Models. *Organizational Research Methods* 3: 186–207.
- 34 Little, D. C., S. R. Shah, S. D. St Peter, C. M. Calkins, S. E. Morrow, J. P. Murphy, R. J. Sharp, W. S. Andrews, G. W. Holcomb, D. J. Ostlie, and C. J. Snyder. 2006. Esophageal Foreign Bodies in the Pediatric Population: Our First 500 Cases. *Journal of Pediatric Surgery* 41: 914–8.
- 35 MacCallum, R. C., and M. W. Browne. 1993. The Use of Causal Indicators in Covariance Structure Models: Some Practical Issues. *Psychological Bulletin* 114(3): 533–41.
- 36 MacCallum, R., M. Rosnowski, C. Mar, and J. Reith. 1994. Alternative Strategies for Cross-validation of Covariance Structure Models. *Multivariate Behavioral Research* 29: 1–32.
- 37 MacKenzie, S. B., P. M. Podsakoff, and C. B. Jarvis. 2005. The Problem of Measurement Model Misspecification in Behavioral and Organizational Research and Some Recommended Solutions. *Journal of Applied Psychology* 90: 710–30.
- 38 Marsh, H. W., and S. Jackson. 1999. Flow Experience in Sport: Construct Validation of Multidimensional, Hierarchical State and Trait Responses. *Structural Equations Modeling* 6: 343–71.
- 39 McDonald, R. P. 1996. Path Analysis with Composite Variables. *Multivariate Behavioral Research* 31: 239–70.
- 40 Meade, A. W., and G. J. Lautenschlager. 2004. A Monte Carlo Study of Confirmatory Factor Analytic Tests of Measurement Equivalence/Invariance. *Structural Equation Modeling* 11: 60–72.
- 41 Meredith, W. 1993. Measurement Invariance, Factor Analysis, and Factorial Invariance. *Psychometrika* 58: 525–43.
- 42 Merkle, E. C., and Y. Rosseel. 2016. *Blavaan: Bayesian Structural Equation Models via Parameter Expansion*. Ithaca, NY: Cornell University Press.
- 43 Millsap, R. E., and H. Everson. 1991. Confirmatory Measurement Model Using Latent Means. *Multivariate Behavioral Research* 26: 479–97.
- 44 [www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/](http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/), accessed 30 October 2017.
- 45 Nunnally, J. C. 1978. *Psychometric Theory*. New York: McGraw-Hill.
- 46 Ping, R. A. 2004. On Assuring Valid Measurement for Theoretical Models Using Survey Data. *Journal of Business Research* 57: 125–41.
- 47 Podsakoff, N. P., W. Shen, and P. M. Podsakoff. 2006. The Role of Formative Measurement Models in Strategic Management Research: Review, Critique, and Implications for Future Research. *Research Methodology in Strategic Management* 3: 197–252.
- 48 Shafer, A. B. 1999. Relation of the Big Five and Factor V Subcomponents to Social Intelligence. *European Journal of Personality* 13: 225–40.
- 49 Steenkamp, J., and H. Baumgartner. 1998. Assessing Measurement Invariance in Cross-cultural Research. *Journal of Consumer Research* 25: 78–79.
- 50 Steinmetz, H., C. A. Vassiliadis, V. Belou, and A. Andronikidis (2011). Three Approaches to Estimate Latent Interaction Effects: Intention and Perceived Behavioral Control in the Theory of Planned Behavior. *Methodological Innovations Online* 6, 95–110.
- 51 Vandenberg, R. J., and C. E. Lance. 2000. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Record* 3: 4–69.
- 52 Wilcox, J. B., R. D. Howell, and E. Breivik. 2008. Questions About Formative Measurement, *Journal of Business Research*: 61: 1219–28.
- 53 Yi He, Y., M. D. Merz, and D. L. Alden. 2008. Diffusion of Measurement Invariance Assessment in Cross-National Empirical Marketing Research: Perspectives from the Literature and a Survey of Researchers. *Journal of International Marketing* 16: 64–83.
- 54 Yoon, M., and R. E. Millsap. 2007. Detecting Violations of Factorial Invariance Using Data-Based Specification Searches: A Monte Carlo Study. *Structural Equation Modeling* 14: 435–563.
- 55 Zhao, X., J. G. Lynch, Jr, and Q. Chen. (2010). Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research* 37: 197–206.

# 13 Partial Least Squares Structural Equation Modeling (PLS-SEM)

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

Understand the distinguishing characteristics of PLS-SEM.

Explain how the PLS-SEM algorithm is estimated.

Describe the stages of the PLS-SEM decision process.

Distinguish between common factor modeling and composite modeling.

Explain how to assess reflective and formative measurement models.

Interpret the results of the HBAT data used with PLS-SEM.

Describe when PLS-SEM and CB-SEM are the appropriate structural modeling method.

## Chapter Preview

Use of the multivariate statistical technique of partial least squares structural equation modeling (PLS-SEM) has increased substantially during the past decade in all fields of social sciences research, particularly business. As the number of structural relationships and constructs to be considered in multivariate modeling increases, so does the need for increased knowledge of the structure and interrelationships of the variables. This chapter focuses on PLS-SEM, a very useful technique for analyzing complex structural models increasingly encountered by researchers in this era of Big Data, particularly from secondary data sources. It defines and explains in broad, conceptual terms the fundamental aspects of PLS-SEM, and when it is the appropriate SEM method to use. The method can be utilized to examine the measurement models and structural relationships for complex models and obtain solutions that are not possible when applying CB-SEM. To further clarify the methodological concepts, basic guidelines for presenting and interpreting the results of applying PLS-SEM are also included.

## Key Terms

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

**Adjusted coefficient of determination ( $R^2$ )** A modified measure of the *coefficient of determination* that is adjusted based on the number of predictor constructs. It is used to compare models with different numbers of predictor constructs and/or different sample sizes, or both.

**AVE** See *Average variance extracted*.

**Average variance extracted** A measure of convergent validity. It is the degree to which a latent construct explains the variance of its indicators; see *Communality (construct)*.

**Bias-corrected and accelerated (BCa) bootstrap confidence intervals** It improves the *percentile method* by adjusting for biases and skewness in the bootstrap distribution. The method yields very low Type I errors but is limited in terms of statistical power.

**Blindfolding** A sample reuse technique for PLS-SEM that omits part of the data matrix and uses the SEM model estimates to predict the omitted part.

**Bootstrap cases** The number of observations drawn in the bootstrap samples. The number is set equal to the number of valid observations in the original dataset.

**Bootstrap confidence interval** An estimated range of values that is likely to include an unknown population parameter. It is based on the upper and lower bounds, which depend on a predefined probability of error and the standard error of the estimation for a given set of sample data. When zero does not fall in the confidence interval, an estimated parameter is considered to be significantly different from zero for the prespecified probability of error (e.g., 5%).

**Bootstrap samples** The number of samples drawn when the bootstrapping method is applied. Generally, a minimum of 1,000 samples is recommended, but some authors recommend 5,000.

**Bootstrapping** A resampling technique that randomly withdraws a large number of sub samples from the original data (with replacement) and estimates models for each subsample. It is used to determine the standard errors of coefficients to determine their statistical significance without applying distributional assumptions.

**CB-SEM** See *covariance-based structural equation modeling*.

**Common variance** Variance shared with other variables in the statistical analysis.

**Coefficient of determination ( $R^2$ )** A measure of the proportion of the variance in the endogenous construct(s) that is explained by the predictor constructs.

**Collinearity** Is present when two variables are highly correlated.

**Common factor model** Assumes that each indicator in a set of observed measures is a linear function of one or more common factors. Confirmatory factor analysis (CFA) and CB-SEM are the two types of analyses based on common factor models, but exploratory factor analysis (EFA) also has an option that can be run with only common variance.

**Communality (construct)** See *average variance extracted*.

**Communality (item)** See *indicator reliability*.

**Composite indicators** A type of indicator used in formative measurement models. A linear combination is used to combine the composite indicators into the latent construct (or composite). As a result, composite indicators are not necessarily conceptually united.

**Composite reliability** A measure of internal consistency reliability, which, in contrast to Cronbach's alpha, does not assume equally weighted indicator loadings. Composite reliability should be above 0.60 in exploratory research, and above 0.70 as a general guideline, but not above 0.95.

**Composite variable** A latent variable that is a linear combination of several variables (indicators).

**Construct scores** A linear combination of the data from all indicators used to measure a particular construct.

**Constructs** Indirectly measured concepts that are complex, abstract and can not be directly observed. Constructs are shown in path models as circles or ovals, to which directly measured indicators are attached, and are also referred to as *latent variables*.

**Continuous moderator variable** A variable that is measured continuously (not categorically) and may affect the direction and/or strength of the relationship between two latent variables in a structural model.

**Convergence** Reached when the results of the PLS-SEM algorithm change very little. Convergence occurs when the PLS-SEM algorithm stops based on a pre-specified stop criterion (i.e., a very small number such as 0.00001) that indicates only very small changes in PLS-SEM computations are possible. Thus, convergence is reached when the PLS-SEM algorithm stops because the prespecified stop criterion has been met but not necessarily the maximum number of iterations specified.

**Convergent validity** For formatively measured constructs, it is the extent to which the formative construct correlates positively with an alternative measure (reflective, single- or multi-item) of the same construct; see *redundancy analysis*. For reflectively measured constructs, it is the extent to which a latent construct explains the variance of its indicators; see *communality (construct)*.

**Covariance-based structural equation modeling (CB-SEM)** is based on common variance only, as measured by the variable covariances, and is used to confirm (or reject) theories. It does this by determining how well a theoretical model can estimate a second covariance matrix that is not significantly different from the original observed for a sample dataset.

**Critical  $t$  value** The criterion applied to determine the significance of a coefficient. If the *empirical t value* is larger than the critical  $t$  value, the null hypothesis of no effect is rejected. Typical critical  $t$  values are 2.57, 1.96, and 1.65 for a two-tailed significance level of 1%, 5%, and 10%, respectively.

**Critical value** See *significance testing*.

**Cronbach's alpha** A measure of internal consistency reliability that ranges from 0 to 1, and assumes equal (unweighted) indicator loadings. When SEM is used with reflectively measured constructs, composite reliability is considered a more suitable criterion of reliability. But Cronbach's alpha is still considered a conservative measure of internal consistency reliability.

**Degrees of freedom (df)** The number of values in the final calculation of the test statistic that are free to vary.

**Discriminant validity** The extent to which a construct is distinct from other constructs in a theoretical structural model. It is measured based on how much it correlates with other constructs in the theoretical model, compared to how much indicators represent only a single construct.

**Endogenous constructs** See *endogenous latent variables*.

**Endogenous latent variables** Serve as only dependent variables, or as both independent and dependent variables in a structural model.

**Error terms** Represent the unexplained variance in constructs and indicators when structural models are estimated.

**Evaluation criteria** Are recommended to use in evaluating the quality of the results for measurement models and the structural model in statistical analysis, particularly SEM. Examples for PLS-SEM include criteria for  $f^2$  effect size, bootstrapping and blindfolding.

**Exogenous constructs** See *exogenous latent variables*.

**Exogenous latent variables** Latent multi-item variables that serve only as independent variables in a structural model.

**Explained variance** See *coefficient of determination*.

**Exploratory** Describes research situations that focus on exploring data patterns and identifying relationships.

**$f^2$  effect size** A measure used to assess the relative impact of a predictor construct on an endogenous construct.

**Factor (score) indeterminacy** Means the statistical analysis can compute an infinite number of factor scores that match the specific requirements of a specified common factor model. In contrast to estimation of scores in PLS-SEM, which are determinate, the scores of common factors are indeterminate.

**Formative measurement** A type of measurement model in which the indicators completely form (see *Composite indicators*) or cause (see *Causal indicators*) the construct. In the path model the arrows point from the indicators to the construct.

**Formative measurement model** A measurement model in which the direction of the arrows is from the indicator variables to the construct, that assumes the indicator variables cause (or form) the measurement of the construct. See formative measurement.

**Formative measures** See *measurement models*.

**Fornell-Larcker criterion** A measure of discriminant validity that compares the shared variance within the constructs to the shared variance between the constructs. The shared variance within should be larger than the shared variance between.

**HCM** See *hierarchical component model*.

**Heterogeneity** When the responses for two or more groups are significantly different, and therefore produce different model parameters. Heterogeneity can be either observed, such as firm size or gender, or unobserved in which it is not known before a post hoc analysis.

**Heterotrait-heteromethod correlations** A procedure for assessing discriminant validity that estimates the true correlation between two constructs if they were perfectly measured (i.e., if they were perfectly reliable). The HTMT correlations are the average of correlations of the indicators within a construct measured across different constructs.

**Heterotrait-monotrait ratio (HTMT)** An alternative procedure for assessing discriminant validity, it estimates the true correlation between two constructs if they were perfectly measured (i.e., if they were perfectly reliable).

**Hierarchical component model (HCM)** A higher-order structure (HOC) of the relationships between constructs, most often second-order. The structure contains several layers of constructs and involves a higher level of abstraction.

**Higher-order component (HOC)** A general construct that represents all underlying LOCs in an HCM.

**Higher-order model** See *hierarchical component model (HCM)*.

**HOC** See *higher-order component*.

**HTMT** See *heterotrait-monotrait ratio*.

**Hypothesized relationships** Proposed relationships between constructs that define the paths in the structural model. PLS-SEM enables researchers to statistically test these hypotheses and thereby empirically evaluate whether the proposed path relationships are significant and meaningful.

**In-sample predictive power** See *coefficient of determination*.

**Indicator reliability** The square of a standardized indicator's outer loading. It represents the variation in an item explained by the construct, and is referred to as the variance extracted from the item; see *communality (item)*.

**Indicators** Directly measured observations (raw data), generally referred to as either *items* or *manifest variables*. In path models they are shown as rectangles.

**Indirect effect** Represents a relationship between two latent variables in which a third variable/construct (e.g., mediator) intervenes.

**Inner model** See *structural model*.

**Latent variables** The unobserved (not directly measured) concepts in the structural model, shown as circles or ovals.

**LOC** See *lower-order component*.

**Lower-order component (LOC)** A first-order construct related to the HOC in an HCM.

**Manifest variables** See *indicators*.

**Maximum number of iterations** The number of iterations needed to ensure that the algorithm converges. If convergence cannot be reached with the specified number of iterations, the number of iterations can be increased so convergence can be achieved.

**Measurement** The process of assigning numbers to a variable based on a set of rules.

**Measurement error** The difference between the true value of a variable and the value obtained by a measurement. Also considered inaccuracies in measuring the “true” variable values due to the fallibility of the measurement instrument (i.e., inappropriate response scales), data entry errors, or respondent errors.

**Measurement model** A component of a theoretical path model that contains the indicators and their relationships with the constructs; also called the *outer model* in PLS-SEM.

**Measurement model misspecification** Specifying a reflective measurement model as formative when it actually is reflective, or vice versa.

**Measurement scale** A method to obtain responses to a question that has a predetermined number of responses.

**Measurement theory** Specifies how constructs should be measured with several indicators. The theory specifies which indicators to use for construct measurement as well as the directional relationship between the construct and its indicators.

**Mediating effect** See *mediation*.

**Mediation** A situation in which one or more mediator variable(s) facilitate the explanation of the relationship(s) between two other variables/constructs.

**Mediation model** See *mediation*.

**Mediator variable** See *mediation*.

**Minimum sample size** The number of observations needed to represent the underlying population and to meet the requirements to obtain an acceptable solution. See *ten (10) times rule*.

**Mode A** An algorithm setting for determining measurement model estimates, typically used to estimate reflective measurement models.

**Mode B** An algorithm setting for determining measurement model estimates, typically used to estimate formative measurement models.

**Model complexity** A complex model is one that has many latent variables/constructs, numerous structural model relationships, and perhaps both reflective and/or formative measurement models. PLS-SEM has virtually no limits on model complexity.

**Moderating effect** See *moderation*.

**Moderation** Occurs when the effect of one latent variable on another latent variable depends on the value of a third variable, referred to as a moderator variable.

**Moderator effect** See *moderation*.

**Moderator variable** See *moderation*.

**Multicollinearity** Extent to which a variable in the measurement model for a construct is explained by the other variables in the same measurement model. Also refers to the extent to which two or more constructs in a structural model are correlated with each other.

**Multi-group analysis** A type of analysis where a categorical variable (e.g., two categories) potentially affects all relationships in the structural model. The method examines whether parameters differ significantly between the groups. The focus is primarily on comparison of path coefficients.

**Out-of-sample predictive power** See  *$Q^2$  value*.

**Outer loadings** Are the bivariate correlations between each variable and their associated constructs. They represent the absolute contribution of an indicator to its construct.

**Outer models** See *measurement model*.

**Outer weights** The indicators’ partial correlations with their associated construct. They indicate each indicator’s relative importance to formative measurement models.

**p value** The probability of incorrectly rejecting a true null hypothesis in a given statistical test. That is, concluding a path coefficient is not significantly different from zero when in fact it is actually not zero (i.e., significantly different).

**Parsimonious models** Models with as few parameters as possible to represent theoretical relationships.

**Partial least squares structural equation modeling (PLS-SEM)** A variance-based method to estimate structural equation models that uses total variance and focuses on maximizing the explained variance of the endogenous latent variable(s).

**Path coefficients** Estimates of the path relationships in the structural model (i.e., between the constructs in the model), which correspond to standardized betas in regression analysis.

**Path models** Diagrams that visually display the hypotheses and variable relationships that are examined with SEM.

**PLS path modeling** See *partial least squares structural equation modeling*.

**PLS regression** Explores the linear relationships between several multi-item independent variables and one dependent variable that may be a single item or a single multi-item construct. The procedure constructs composites from both the several multi-item independent variables and the dependent variable(s) by means of principal component analysis.

**PLS-SEM algorithm** The basic building block for the method. The algorithm uses the PLS path model and the available indicator data to estimate the scores of all latent variables in the model, which in turn are used to estimate all path model relationships.

**PLS-SEM bias** An issue regarding solutions obtained when comparing PLS-SEM and CB-SEM results. The different statistical objectives of the estimation of PLS-SEM and CB-SEM methods produce different results in the model estimates. PLS-SEM results are biased only if one considers CB-SEM results to be a more accurate estimation method. PLS-SEM scholars do not consider it a bias, but rather the differences between the statistical objectives of two different methods.

**Prediction** The primary objective of the PLS-SEM method. The higher the *coefficient of determination ( $R^2$  values)* of endogenous constructs (*latent variables*), the better their prediction by the PLS-SEM model.

**Prediction error** The difference between the prediction of the data points and their original value in the sample.

**Predictive relevance ( $Q^2$ )** A measure of a model's predictive power (i.e., predictive relevance). It assesses whether a model accurately predicts data not used in the estimation of model parameters.

**Principal components analysis** An approach to exploratory factor analysis (EFA) that is based on total variance, and not the common factor model.

**$Q^2$  value** A measure of a PLS-SEM model's predictive power. The computation of  $Q^2$  is based on the *blindfolding* technique, which uses a subset of the available data to estimate model parameters and then predicts the omitted data.

**$R^2$  value** See *coefficient of determination*.

**$R^2$  values** The amount of explained variance of endogenous *latent variables* in the *structural model*. The higher the  $R^2$  values, the better the construct is explained by the latent variables in the structural model. See *coefficient of determination*.

**Redundancy analysis** Measures a formative construct's *convergent validity* by determining whether a formatively measured construct is highly correlated with a reflective or single-item measure of the same construct.

**Reflective measure** See *reflective measurement*.

**Reflective measurement** A type of measurement model setup in which indicators represent the effects (or manifestations) of an underlying construct. Causality is from the construct to its indicators (items).

**Reflective measurement model** See *reflective measurement*

**Reliability** The consistency of a multi-item scale or construct. A scale is reliable when it produces consistent outcomes under similar or the same conditions. The most commonly used measure of reliability is the *internal consistency reliability*. Examples are Cronbach's Alpha and composite reliability.

**SEM** See *structural equation modeling*.

**Single-item construct** A construct that has only a single item/indicator measuring it.

**Specific variance** Variance of each variable unique to that variable and not explained or associated with other variables in the analysis.

**Standardized data** Have a mean value of 0 and a standard deviation of 1 (*z-standardization*). PLS-SEM results are based on standardized *raw data*.

**Statistical power** The probability of concluding that an effect as significant when in fact the effect is significant.

**Stop criterion** See *convergence*.

**Structural equation modeling (SEM)** Method to estimate a *structural model* containing a series of interrelated dependence relationships (equations) where the dependent construct in one relationship may be a predictor construct in another relationship. Also includes a *measurement model* where multiple indicators are used to define each construct/variable (known as a *latent construct*) used in the structural model.

**Structural model** The theoretical or conceptual components of the path model. The structural model (also called inner model in PLS-SEM) includes the latent variables/constructs and their path relationships.

**Structural theory** Specifies how the latent variables are related to each other. That is, it shows the constructs and the paths/relationships between them.

**Ten (10) times rule** A method of determining the minimum sample size specific to the PLS path model (i.e., 10 times the number of independent variables of the most complex OLS regression in the structural or formative measurement model). The 10 times rule should be viewed as a rough guideline for the minimum sample size.

**Theory** A set of systematically related hypotheses developed following the scientific method that can be used to explain and predict outcomes and can be tested empirically.

**Total effect** The sum of the direct effect and the indirect effect between an exogenous and an endogenous latent variable in the path model. It also could be the sum of the direct effect and the indirect effect between two or more endogenous constructs.

**Total indirect effect** The sum of all indirect effects in a multiple mediation model.

**Validity** The extent to which a construct's indicators jointly measure what they are supposed to measure.

**Variance-based SEM** See *partial least squares structural equation modeling*.

**Variance inflation factor (VIF)** A statistic used to evaluate the severity of collinearity among the indicators in a formative measurement model, or between the constructs in a structural model. The VIF is directly related to the tolerance value.

**Variate** Linear combination of variables formed by deriving empirical weights applied to a set of variables specified by the researcher. See also *composite variable*.

**VIF** See *variance inflation factor*.

## What is PLS-SEM?

Partial least squares structural equation modeling, often referred to as PLS-SEM, is a combination of interdependence and dependence techniques, as defined in Chapter 1. The method belongs to the family of statistical models that seek to explain the relationships among multiple variables simultaneously. As with covariance-based SEM (CB-SEM) discussed in Chapters 9 through 12, PLS-SEM consists of two models, the **measurement model** (representing how measured variables represent the constructs) and the **structural model** (showing how constructs are associated with each other). In PLS-SEM the measurement model is often referred to as the **outer model** and the structural model is termed the **inner model**. But these two models operate in fundamentally the same exact manner in both approaches. Let's first examine these two models briefly as to their basic purpose.

### STRUCTURAL MODEL

The **structural model** examines the *structure* of interrelationships expressed in a series of equations, similar to a series of multiple regression equations. These equations estimate a series of separate, but interdependent, multiple regression equations simultaneously. Just as with CB-SEM, the researcher draws upon theory, prior experience, and the research objectives to distinguish which independent variables predict each dependent variable. Dependent variables in one relationship can become independent variables in subsequent relationships, giving rise to the interdependent nature of the structural model. Moreover, many of the same variables affect each of the dependent variables, but with differing effects. The structural model expresses these **dependence relationships** among independent and dependent variables/constructs, even when a dependent variable becomes an independent variable in other relationships.

### MEASUREMENT MODEL

The second basic model is the **measurement model**, which defines the **latent constructs**. Also termed a **latent variable**, a latent construct is a hypothesized and unobserved concept that can be represented by observable or measurable variables. In the past, researchers have sometimes attempted to distinguish between latent variables/constructs that are factor-based and those that are composite-based [85]. But in most instances latent constructs are specified as representing conceptual variables in statistical models, and the latent constructs are proxies that facilitate empirical testing of hypotheses that represent relationships between conceptual variables in a structural equation model [107]. Thus, all measures of conceptual variables are approximations of or proxies for conceptual variables, irrespective of the construct definition or the theoretical foundation [145]. Any measurement model in a SEM, therefore, whether common factor- or composite-based, develops linear combinations that are proxies for latent variables [107, 124].

Latent constructs are measured indirectly by examining the relationships between multiple measured variables, sometimes referred to as **manifest variables**, or **indicators**, which are gathered through various data collection methods (e.g., surveys, tests, observational methods) as well as from secondary sources (e.g., social media websites, mobile phone call records, GPS signals, government sources, information from company data warehouses). As with covariance-based structural modeling, PLS-SEM provides metrics to evaluate the reliability, validity and measurement error associated with the constructs. As noted earlier, practical and theoretical perspectives indicate we cannot perfectly measure a concept. While we may be able to reduce error when measuring physical concepts such as time (e.g., measurement with atomic clocks), most theoretical or abstract concepts are necessarily always subject to some degree of **measurement error**. For example, when asking about something as straightforward as age, income or weight, we know from experience that some people will answer incorrectly, either overstating age and income or under-stating weight, or not knowing it precisely. The answers provided to these rather simple questions, as well as to more abstract questions such as level of commitment or trust, have some measurement error, and thus affect the estimate of the true measurement and structural coefficients. Similar to CB-SEM, PLS-SEM automatically applies a correction procedure that accounts for the measurement error.

## THEORY AND PATH MODELS IN PLS-SEM

Theory can be thought of as a systematic set of relationships providing a consistent and comprehensive explanation of phenomena. From this definition, we see that theory is not the exclusive domain of academia, but is often based on experience and practice obtained by observation of real-world behavior. It provides the foundation upon which both the structural and measurement models are specified.

A **path model** is a visual representation of a theory. SEM models consisting of both measurement and structural models can be quite complex. While all of the relationships can be expressed in mathematical path analysis notation, many researchers find it more convenient to portray a model in a visual form, known as a **path diagram**. This path diagram displays the hypotheses and variable relationships to be estimated in a structural equation modeling analysis [8]. Common convention is to represent constructs as circles or ovals and the manifest variables or indicators of the constructs as rectangles. Directional arrows represent the types of relationships in both the outer/measurement models as well as the inner/structural models. For more details on developing and portraying path models, see the discussion of this topic in Chapter 9.

## THE EMERGENCE OF SEM

Structural equation modeling (SEM) is a relatively new analytical tool, but its roots extend back to the first half of the twentieth century when the concepts underlying the structural model were developed [150]. It was not until the late 1960s and early 1970s that the work of Jöreskog [77, 78] led to the emergence of the measurement model and the development of common factor-based SEM (also called covariance-based SEM and CB-SEM). CB-SEM extended prior models by the simultaneous maximum likelihood estimation of the relationships between constructs and measured indicator variables (i.e., the measurement model), as well as the relationships among latent constructs (i.e., the structural model). In the late 1970s, another scholar [146, 147, 148, 149] began working on an alternative approach to structural modeling that emerged directly from ordinary least squares multiple regression and principal components factor analysis [90] and also involved simultaneous estimation of relationships (paths) among constructs (structural model) as well as between constructs and measured variables (measurement model). This alternative was partial least squares structural equation modeling (PLS-SEM) [55]. The CB-SEM method was referred to by Wold as “hard modeling” because of the more restrictive assumptions of the method, such as normally distributed data and larger sample sizes. In contrast, he referred to PLS-SEM as “soft modeling” because it does not require normally distributed data and performs well even when the data are highly skewed [54] as is typical of much social sciences data. As Lohmöller [84, p. 64] notes, “it is not the concepts nor the models nor the estimation techniques which are ‘soft,’ only the distributional assumptions.”

CB-SEM became the dominant multivariate analysis technique in the 1990’s and remained so well after 2000. This was in large part due to the early availability of user-friendly software as well as the substantial benefits of SEM in general, which provides a conceptually appealing way to develop as well as test theory. The first user-friendly PLS-SEM software was PLSGraph [15], and a limited number of PLS applications began appearing in journals. Another user-friendly PLS-SEM software became available in 2005 (SmartPLS 2) [118], and was then updated and extended in 2015 [119]. The later version of SmartPLS automated many analytical functions and the applications of PLS-SEM in journals expanded substantially. Increasingly, PLS-SEM is being applied in a broad range of disciplines including marketing [59], strategic management [57], group and organization management [135], international management [106], operations management [101], management information systems [48, 117], supply chain management [80], accounting [82], and tourism [31]. Contributions in terms of books, edited volumes, and journal articles applying PLS-SEM or proposing methodological extensions are also appearing (see for example, [49, 58, 92, 129, 130]).

PLS-SEM is similar to many other multivariate techniques in that it is a variance analysis technique rather than a covariance structure analysis technique, as is CB-SEM. As a result, PLS-SEM focuses on explaining the variance in the dependent (endogeneous) construct(s), and does not include the concept of “fit” based on a covariance matrix.

## ROLE OF PLS-SEM VERSUS CB-SEM

While we have seen that CB-SEM and PLS-SEM are comparable in their basic elements, there are some distinct differences we will detail in this chapter. We introduce partial least squares structural equation modeling as our second approach to structural modeling because the primary statistical objective is fundamentally different than covariance-based structural modeling, and can obtain solutions that are not possible with CB-SEM. The basic distinction between these two approaches is:

- CB-SEM: While prediction is possible with covariance-based structural modeling, the primary statistical objective of CB-SEM is confirming theory by estimating a new covariance matrix that is not significantly different from the original observed covariance matrix.
- PLS-SEM: In contrast, the primary statistical objective of PLS-SEM is prediction that maximizes the explained variance in the dependent variable(s).

This chapter will review the foundational elements of PLS-SEM and provide a comparative perspective with CB-SEM. Readers are encouraged to also review Chapters 9 through 12, which discuss many of the common topics between these approaches in more detail. We will use the more generic term SEM to refer to concepts common to both CB-SEM and PLS-SEM.

## Estimation of Path Models with PLS-SEM

In contrast to CB-SEM, which uses a common-factor approach to developing proxies for constructs, PLS-SEM calculates composites for the latent variables. The composites are estimated as exact linear combinations of their empirical indicators [38] that maximize the variance in the exogenous construct indicators that is useful for predicting the endogenous construct(s) indicators [93]. PLS-SEM treats these composites as proxies for the conceptual variables under investigation [67, 107], similar to the way CB-SEM treats factor-based constructs as proxies [124].

While PLS-SEM's composite modeling method has traditionally been considered more consistent with formative measurement models than reflective ones [43], the method easily accommodates both measurement model types without encountering identification problems [55]. PLS path model identification only requires that each of the constructs be linked to the nomological net of constructs [67]. This requirement also applies to structural models where endogenous constructs are specified as formative, since PLS-SEM involves a multistage estimation process that separates measurement model estimation from structural model estimation [110].

## MEASUREMENT MODEL ESTIMATION

Whether the measurement models are specified as reflective or formative, PLS-SEM always uses linear combinations of sets of indicators to represent the latent constructs when estimating model parameters. PLS-SEM has been criticized previously for underestimating structural model parameters and overestimating measurement model parameters, many times referred to as PLS-SEM bias [17]. But research has shown that the size of PLS-SEM bias is small in absolute terms [105] and decreases when the number of indicators per construct and sample size increase [76]. Furthermore, recent research has shown that the composite model approach of PLS-SEM produces more consistent parameter estimates [124] and higher power with smaller sample sizes [41, 105]. Moreover, the differences previously attributed to the PLS-SEM bias are a result of comparing PLS results to those from common factor modeling (CB-SEM), and thus assumes the CB-SEM results are more accurate. Recent research indicates, however, that this assumption is not true, and that both approaches are proxy measures (estimates) of construct scores [65, 121].

The confirmation of theoretical measurement models in PLS-SEM tests the hypothesis that theoretical relationships actually exist between the observed indicator variables and their underlying latent constructs. To do so, a variate

of the indicators is derived to represent the constructs, similar to the variate that is the building block in all other multivariate methods. With PLS-SEM, to confirm reflective measurement model hypotheses the metrics applied are the size and significance of the loadings and/or coefficients, reliability, convergent validity, and discriminant validity [49, 58]. Note that when confirming measurement models with PLS-SEM the process is referred to as confirmatory composite analysis (CCA) [65], while with CB-SEM it is called confirmatory factor analysis (CFA). When examining the structural model relationships, for both methods the size and significance of the path coefficients are assessed. In addition, when using CB-SEM researchers also assess goodness of fit, while with PLS-SEM researchers assess  $R^2$ ,  $f^2$  effect size, and  $Q^2$ .

## STRUCTURAL MODEL ESTIMATION

The use of linear composites with PLS-SEM has implications not only for the method's measurement philosophy, but also for its application in the estimation of the structural model. Once the algorithm derives the weights, PLS-SEM always produces a single specific (i.e., determinant) score for each observation on all composites. These scores are proxies of the concepts being measured, just as common factors are proxies of the conceptual variables in CB-SEM [5]. Using these proxies as input, PLS-SEM applies ordinary least squares regression based on the statistical objective of minimizing the unexplained variance in the dependent variables (i.e., the residual variance), which achieves the prediction objective. In short, the PLS-SEM algorithm estimates path coefficients that maximize the explained variance ( $R^2$ ) in the endogenous constructs. Indeed, simulation studies [3, 36] support the superior predictive capabilities of PLS-SEM compared to CB-SEM.

## ESTIMATING THE PATH MODEL USING THE PLS-SEM ALGORITHM

PLS-SEM derives solutions for the path model, including both the measurement and structural models, using a three-stage approach: initial estimates of the measurement model, initial estimates of the structural model and final estimates of both models. Each stage will be discussed below.

**Stage 1: Initial Measurement Model Estimates** In the first stage the PLS-SEM algorithm iteratively determines the inner weights for the structural relationships and the construct scores using a four-step procedure. In Step 1 preliminary latent variable scores are calculated typically by using unit weights (i.e., 1) for all indicators of each construct [49]. These construct scores are used to determine the inner weights between the adjacent constructs in the structural model. Then proxies for all latent variable constructs are computed. Finally, new outer weights are computed for all the indicators indicating the strength of the relationships between each latent construct and its corresponding indicators.

The PLS-SEM algorithm uses two different estimation modes to compute the outer weights. Mode A uses the bivariate correlation between each indicator and the construct to determine the outer weights. In contrast, Mode B uses the OLS regression results to obtain the outer weights. In this regression model, the construct is the dependent variable and the indicators the independent variables. As a result, the regression weights consider not only the correlations between the construct and each indicator, but also the correlations between the indicators. Mode A is used in PLS-SEM to obtain solutions for reflectively specified constructs and Mode B for formatively specified constructs.

**Stages 2 and 3: Initial Structural Model and Final Model Estimates** The second and third stages use the final latent variable scores from Stage 1 as input for a series of ordinary least squares regressions. These OLS regressions compute the final outer loadings for reflective measurement models, the final outer weights for formative measurement models, and the final path coefficients as well as related elements. The related elements include the indirect and total effects,  $R^2$  values of the endogenous latent constructs, and the indicator and latent variable correlations [84] as well as the  $f^2$  effect size for each predictor construct, and an additional measure of the model's predictive power termed  $Q^2$ . All of these elements will be discussed in more detail in a later section.

## PLS-SEM Decision Process

We focus the discussion of the PLS-SEM decision process on the model-building paradigm introduced in Chapter 1. This process varies slightly from that introduced in previous chapters to reflect the unique terminology and procedures of PLS-SEM. The six stages are as follows:

- Stage 1:** Defining research objectives and selecting constructs
- Stage 2:** Designing a study to produce empirical results
- Stage 3:** Specifying the measurement and structural models
- Stage 4:** Assessing measurement model validity
- Stage 5:** Assessing the structural model
- Stage 6:** Advanced analyses with PLS-SEM

The remainder of this chapter provides an overview and introduction of these six stages. We also include an application of PLS-SEM to the HBAT example data previously introduced in the CB-SEM chapters (9 through 12). This application will facilitate a direct comparison of the results of PLS-SEM to those of CB-SEM.

### Stage 1: Defining Research Objectives and Selecting Constructs

A critical element in common to both PLS-SEM and CB-SEM is the importance of the measurement model. Testing hypotheses involving structural relationships between constructs is only as reliable and valid as the measurement models are in justifying the constructs as proxy measures of multi-item latent variables. Researchers typically rely on established scales to specify the measurement models, or at minimum they begin their search there. But sometimes they must develop a new scale or substantially modify an existing scale to the new context. Selecting the indicators to measure each construct sets the foundation for the remainder of the analysis, whether applying PLS-SEM or CB-SEM. Significant time and effort must be devoted early in the research process to make sure the quality of the measurement will support valid conclusions. Finally, when measures are either taken from various sources or developed for a study, some type of pretest should always be performed.

In terms of the structural model, CB-SEM and PLS-SEM are both used to confirm theoretical measurement models and structural model relationships. In evaluating the results to confirm theoretical measurement and structural models, the essential differences between the two methods are CB-SEM uses only common variance derived from a covariance matrix, and the results are assessed first on the basis of goodness of fit (GOF), and then on reliability, convergent validity, and discriminant validity. In contrast, PLS-SEM derives solutions from total variance, not covariances, and thus does not have a similar fit measure. In addition, with PLS-SEM, the results are first assessed on the reliability, convergent validity, and discriminant validity of the measurement models, and then on the predictive ability of the structural model as measured by  $R^2$ ,  $f^2$ , and  $Q^2$ .

But while CB-SEM is exclusively used to test well-developed theory, PLS-SEM has the additional capability to perform exploratory as well as confirmatory research [49, 58]. Exploratory or descriptive modeling involves no underlying causal theory, except perhaps to a limited extent in a few situations [134]. In exploratory research researchers are exploring possible relationships not based on well-developed theoretical or causal justification, but instead potential associations that may lead to further theory development. CB-SEM is not appropriate for exploratory/descriptive modeling because the exploratory nature of the modeling process is not based on well-developed theory, and specific hypotheses are not necessarily prespecified. In contrast, PLS-SEM is appropriate for exploratory modeling, measurement and structural models can be changed based on the data, and hypotheses do not have to be prespecified to be tested. The distinction of exploratory research is important and involves at least three situations:

- *Less well-defined research problems.* As noted earlier, when research problems have not been clearly defined exploratory research is the appropriate method. While sufficient information may not be available to make conceptual distinctions or to propose explanatory relationships, researchers can let the method and the data define the nature of the relationships.

- *Generating hypotheses.* Exploratory research can be used to generate hypotheses from data obtained from qualitative methods, and also to test hypotheses using quantitative research. For example, if hypotheses are developed from research in one context, such as in the US, the focus might be on testing the same or similar hypotheses in another country.
- *Types of research questions.* Exploratory research can examine all types of research questions, including what, when, why and how [134]. In contrast, confirmatory research examines previously specified hypotheses that predict specific outcomes based on underlying theory, and the hypotheses usually are derived from established theories or previous studies conducted within the same context.

Another difference between the two approaches is that PLS-SEM focuses on predictive modeling, which is the process of applying a statistical model to examine data representing both independent and dependent variables, with the objective of predicting existing or new observations [134]. This type of statistical modeling provides an opportunity to develop conclusions based on statistical significance of relationships, relative influence of antecedents, explained variance, effect sizes, and prediction as they relate to proposed hypotheses. In social sciences research, particularly business, the statistical models used for testing hypotheses are often based on correlations derived from observational or survey data, and increasingly secondary data. Thus, predictive modeling is more than just statistical modeling. It is also explanatory modeling, or the application of statistical modeling to test and explain relationships between variables, accept/reject hypotheses, and predict outcome variables.

PLS-SEM and CB-SEM share some fundamental objectives common to all SEM applications. But as we have briefly discussed and will illustrate in the following sections, the two approaches take quite different paths to their final results. Each approach has both advantages and limitations and the researcher, rather than selecting just one method for all situations, should evaluate each research problem for its unique characteristics and then select the approach that is most applicable.

## Stage 2: Designing a Study to Produce Empirical Results

With the basic theoretical models specified, the researcher must consider issues involved with research design and estimation. Our discussion will focus on issues related to both research design and model estimation, especially as they relate to PLS-SEM. In the area of research design, we will discuss (1) the type of data to be analyzed; (2) the impact and remedies for missing data; (3) statistical power; and (4) the impact of sample size.

### METRIC VERSUS NONMETRIC DATA AND MULTIVARIATE NORMALITY

The observed or measured variables in many multivariate analyses have traditionally been restricted to metric data (ratio or interval), since this type of data is more intuitive to interpret and allows for greater flexibility in the analysis (e.g., in terms of measures of centrality and dispersion). PLS-SEM, however, is a non-parametric method, allowing for the use of nonmetric data types (nominal or ordinal) in addition to metric data. Therefore, PLS-SEM is more flexible regarding the type of data that can be used.

With previous multivariate methods, researchers had to be concerned about the assumption of multivariate normality, which is required for parameter estimation and statistical inference (i.e., statistical significance levels) of the estimated parameters. But PLS-SEM is a non-parametric statistical method and obtains good solutions with non-normal data, which is typical of the data used in most social sciences studies. Many previous studies cited a lack of assumptions about the normality of data distributions as their main reason for choosing PLS-SEM [94, 95, 31]. But scholars have noted that this justification for PLS-SEM use is insufficient by itself since maximum likelihood estimation in CB-SEM is fairly robust against violations of normality [18, 99], and procedures are available for parameter estimation with non-normal distributions [83]. Thus, it is not sufficient to justify the use of PLS-SEM solely on the grounds of the data distribution.

## MISSING DATA

As with other multivariate procedures, researchers must be concerned about missing data. To deal with this issue, two questions must be addressed (1) Is the missing data sufficiently large and non-random, to the extent that it causes problems in estimation or interpretation?, and (2) If missing data must be remedied, what is the best approach? In general, missing data complicate the testing of all SEM models using either PLS-SEM or CB-SEM since the remedies for missing data in most cases reduce the sample size to some extent from the original number of cases or require extensive modeling for the imputation of the missing data. Depending on the missing data approach taken and the extent of missing data anticipated, where possible the researcher should plan to obtain a larger sample size than required to offset any problems of missing data. The reader is referred back to Chapters 2 and 9, where a more complete discussion is provided on the methods of assessing the extent and pattern of missing data and the approaches to remedy missing data, if needed.

## STATISTICAL POWER

**Statistical power** is the probability of making the correct decision if the alternative hypothesis is true and the null hypothesis should be rejected. That is, power is the probability of concluding there is a significant difference when in fact there actually is a difference. PLS-SEM has greater statistical power compared to CB-SEM [105, 41] and researchers benefit from this higher statistical power. The higher statistical power of PLS-SEM means that a specific relationship is more likely to be found significant when it is present in the population. The higher statistical power of PLS-SEM makes the method particularly suitable for exploratory research where theory is less developed and sample sizes are relatively smaller. As Wold [146] notes, model development is an evolutionary process and the structural relationships are often tentative. Thus, as the empirical content of the PLS-SEM model is extracted from the data, the model is improved through the estimation procedure by interactions between the model, the data and the researcher's responses.

## MODEL COMPLEXITY AND SAMPLE SIZE

PLS-SEM provides a unique capacity in SEM applications to be applied with both large and small sample sizes, and many times it is the small sample capability that distinguishes it from other methods [38, 143]. The critical question in discussing sample sizes for SEM applications involves how large a sample is needed to produce trustworthy results, and this decision involves three aspects of model complexity.

- 1 *Number of constructs.* Prior reviews indicate the average number of constructs per model is higher in PLS-SEM (approximately eight constructs [80, 117]) compared to factor-based CB-SEM (approximately five constructs [2, 135]).
- 2 *Number of indicators per construct.* At the same time, the number of indicators per construct is typically higher in PLS-SEM compared to CB-SEM. In contrast to CB-SEM, the PLS-SEM algorithm does not simultaneously compute all the model relationships, but instead uses separate ordinary least squares regressions to estimate the partial regression relationships.
- 3 *Number of observations per estimated parameter.* Finally, the number of model parameters can be very high relative to the sample size, as long as each partial regression relationship has a sufficient number of observations (more than the number of indicators/estimated parameters). There is a basic rule of thumb for sample size, however, and it is 10 times the number of arrows pointing at a construct, whether as a formative indicator to a construct or a structural path to an endogenous construct. Reinartz et al. [105], Henseler et al. [65], and Sarstedt et al. [124] show that the PLS-SEM algorithm obtains solutions when other methods do not converge or develop inadmissible solutions.

As Hair et al. ([56], p. 2) note, “some researchers abuse this advantage by relying on extremely small samples relative to the underlying population” and PLS-SEM scholars have been criticized for this [87]. The general guideline

is when it is possible to obtain more observations, such as with most consumer research projects, then researchers should do so. But in some situations, such as business-to-business research, the population size is quite small (e.g., < 100) and PLS-SEM is the only structural modeling approach that can obtain meaningful solutions with such small sample sizes.

While PLS-SEM can be applied with smaller samples that would be unacceptable with other methods, the acceptability of this practice depends on several considerations. The first involves population heterogeneity, which is a consideration for all statistical techniques. A more heterogeneous population requires a larger sample to arrive at an acceptable level of sampling error, all other considerations being equal [19]. When these fundamentals of sampling theory are ignored, the estimation results are highly questionable, no matter which SEM method is used. Also to be considered are expected effect sizes and desired levels of statistical significance [86], which impact the expected statistical power for a range of path model designs. Hair et al. [49] include a power table that provides guidelines on the minimum sample size when applying PLS-SEM.

To this point considerable attention has been devoted to PLS-SEM's small sample size capabilities [41]. These discussions often overlook, however, the suitability of PLS-SEM for analyzing much larger samples, such as those produced in social media research. Social media analyses typically focus on prediction, often lack a comprehensive substantiation on the grounds of measurement theory [108], and rely on complex models with limited theoretical foundation [138]. The non-parametric nature of PLS-SEM, the ability to handle complex models with many constructs (e.g., 10 or more) as well as a large number of indicators (e.g., 70 or more), and the characteristic of higher statistical power, make the method a valuable tool for social media and digital analytics, as well as the analysis of other large-scale datasets.

## Stage 3: Specifying the Measurement and Structural Models

After the constructs and their respective indicators are identified, the researcher specifies the measurement and structural models. In this stage, each latent construct in the model is identified as either exogenous or endogenous, and the measured variables (manifest variables or indicators) are assigned to latent constructs. When specifying a path model in PLS-SEM, you do not start by first specifying and testing the measurement model (CFA), as you

### General Guidelines on Using PLS-SEM

PLS-SEM is the preferred method when prediction is the statistical objective of your research.

Missing data and outliers must be dealt with before running a PLS model.

PLS-SEM produces good results with non-normal data and when heteroscedasticity is present.

PLS-SEM works well with both metric and non metric data.

PLS-SEM is suitable for survey data and for secondary (archival) data.

PLS-SEM has higher statistical power than CB-SEM, which means it is particularly suitable for exploratory research where theory is less developed, and sample sizes are relatively smaller.

The recommended sample size should be based on statistical power considerations and the context of the research.

In general larger sample sizes (> 100) are preferable, but smaller sample sizes (< 100) are acceptable depending on the context of the research; e.g., B-to-B research that is restricted to smaller populations will necessarily involve smaller sample sizes.

PLS-SEM works well with both reflective and formative measurement models.

Goodness of fit measures like  $\chi^2$ , GFI, CFI, RMSEA, and similar heuristics that are used with CB-SEM do not apply with PLS-SEM.

do in CB-SEM. Instead, the structural model is specified at the same time as the measurement models. Moreover, with PLS-SEM correlational relationships are not specified in the path model between the exogenous constructs. Thus, to specify the PLS path model researchers must rely on their knowledge of both measurement theory and structural theory.

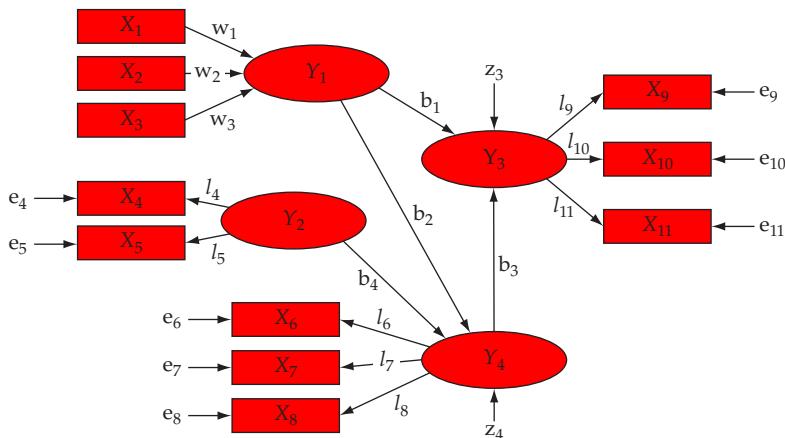
Recall that with CB-SEM different terminology is used to refer to the independent and dependent variables than with previous dependence techniques, such as multiple regression and MANOVA. This same terminology applies to PLS-SEM applications. The independent variables/constructs are referred to as **exogenous constructs**, and are measured by the latent, multi-item equivalent of independent variables. The exogenous constructs are determined by factors outside of the model (i.e., they are not explained by any other construct or variable in the model), and are thus identified as *independent*. Since PLS-SEM path models are also depicted by a visual diagram, it is useful to know how to identify the exogenous constructs. Given that exogenous variables are independent of any other construct in the model, an exogenous construct does not have any paths (single-headed arrows) going into it from any other construct or variable.

PLS-SEM also refers to the dependent constructs as endogenous. The **endogenous constructs** are the latent, multi-item equivalent to dependent variables (i.e., a variate of individual indicator variables). These constructs are theoretically determined by other constructs within the model. Thus, they are dependent on other constructs/variables, and this dependence is represented visually by a path to an endogenous construct from an exogenous construct (or from another endogenous construct). The measurement and path models can be represented by equations, but it is simpler to represent this process with a diagram.

Figure 13.1 shows a path model with four constructs—three of the constructs have three indicators and one has two indicators. In addition, there are two exogenous constructs ( $Y_1$  and  $Y_2$ ) and two endogenous constructs ( $Y_3$  and  $Y_4$ ), and one of the endogenous constructs ( $Y_4$ ) is a mediator. Note that while a minimum of three indicators per construct is required for construct identification in CB-SEM, with PLS-SEM this is not a requirement. But, while possible it does raise the question of whether a concept can be adequately measured with only two indicators.

Constructs in PLS path models are the same as conceptual variables. The constructs are drawn as ovals or circles ( $Y_1$  to  $Y_4$ ) in path models. The indicators attached to the constructs are the measured variables, and they are represented by rectangles ( $X_1$  to  $X_{11}$ ). The relationships between constructs, and the relationships between the indicators and their respective constructs, are shown as arrows. In PLS-SEM, the arrows are always single-headed (there are no double-headed arrows between constructs). The single-headed arrows are directional and considered to be predictive relationships.

**Figure 13.1**  
An example of a PLS-SEM path model



## MEASUREMENT THEORY AND MODELS

Researchers rely on measurement theory to help them specify how to measure latent constructs. To do so, they typically choose between two types of measurement models [21]: reflective measurement models and formative measurement models. Reflective measurement models have direct relationships (arrows) from the construct to the indicators and treat the indicators as error-prone representations of the underlying construct [7]. The following equation illustrates the relationship between a latent variable and its observed indicators:

$$X = l \cdot Y + e$$

where  $X$  is the observed (measured) indicator variable,  $Y$  is the latent construct, the loading ( $l$ ) is a standardized regression coefficient representing the strength of the relationship between  $X$  and  $Y$ , and  $e$  is the random measurement error. The latent variables  $Y_2$ ,  $Y_3$ , and  $Y_4$  in the path model have reflective measurement models, two with three indicators each and one with only two indicators. When developing reflective measurement models, the indicators should be a representative sample of all items of the construct's conceptual domain [98]. If the items are specified from the same domain, together they will represent the same concept and are expected to be highly correlated [34]. Endogenous latent variables always have error terms associated with them. In Figure 13.1, the endogenous latent variables  $Y_3$  and  $Y_4$  have one error term each ( $Z_3$  and  $Z_4$ ), which reflect the sources of variance not predicted by the antecedent construct(s) in the structural model.

Formative measurement models differ from reflective measurement models. They are a linear combination of a set of indicators that form the construct (i.e., the relationship/arrow is from the indicators to the construct). As a result, variation in the indicators occurs before variation in the latent construct [13]. For example, when the respondent's evaluation of a formative indicator changes so does the construct value. Indicators of reflectively measured constructs are typically highly correlated, but formatively measured indicators are often not very correlated. High correlations can occur, however, among indicators with formative measurement models, but this does not necessarily imply the measurement model is reflective [95].

Formative measurement models can be modeled with two types of indicators: causal indicators and composite indicators [9, 10]. Constructs measured with causal indicators have an error term that implies the construct has not been perfectly measured by its indicators [10]. Causal indicators also exhibit conceptual unity since they are selected based on the definition of the concept [11]. A limitation, however, is that researchers are almost never able to identify all relevant indicators constituting the construct's domain [12]. Thus, the error term captures all the other "causes" or explanations of the construct not represented by the set of causal indicators [22]. As Grace and Bollen [43] note, the error term in causal indicator measurement models implies that the construct is equivalent to the conceptual variable, and the error term accounts for the other "causes."

Composite indicators are the second type of formative measurement model indicators. Composite indicators are similar to causal indicators, except the error term of a construct measured with composite indicators is set to zero [23]. The indicators, therefore, completely form the composite representing the construct based on a linear combination, and the composite indicators in combination are assumed to form a new entity. That is, composite indicators are a way to model conceptual variables that define the empirical meaning of the construct. For example, Aaker's [1] conceptualization of brand equity, as defined by Henseler [63], is a typical conceptual variable with composite indicators (artifacts). Thus, when combined the indicators represent the elements of brand awareness, brand associations, brand quality, brand loyalty, as well as other proprietary assets.

PLS-SEM only computes measurement models with composite indicators and not causal indicators. The exogenous construct  $Y_1$  in Figure 13.1 is modeled as a composite formative construct, so the error term  $Z_1$  (not shown in this path model) is set (fixed) as zero (0). Thus, the composite representing the construct is formed in full by a linear combination of the indicators. Note that for both formative and reflectively measured constructs, the error term is typically not shown on PLS path models.

The concept of artifacts is especially relevant when examining theoretical models based on secondary/archival data, which are often not based on established measurement theory [74, 109]. For example, secondary data could be

collected from the internet or a company data warehouse to form an index of information search activities. The index would represent the sum of the activities customers engage in when seeking information from vendors, promotional materials, and other sources [126]. But it is also more likely that measurements based on secondary data will include more artifacts than survey data.

Measurement models with composite indicators can also be used for dimension reduction. When dimension reduction is the focus, the objective is to condense (combine) the measures so they adequately cover the salient features of a conceptual variable [27]. For example, a researcher may be interested in representing the salient aspects of customer loyalty by including three (composite) indicators that measure relevant elements (e.g., satisfaction, recommend to others, and future purchase likelihood). Sarstedt et al. [124] note that composite indicators can be used to measure any concept including attitudes, perceptions, and behavioral intentions, as long as researchers specify an appropriate construct definition and include items that closely match the construct definition. In sum, construct measurement based on composite indicator models is a method to develop proxies for conceptual variables that acknowledges the practical problems of measuring unobservable conceptual variables.

## STRUCTURAL THEORY AND PATH MODELS

Structural theory specifies the latent constructs in the theoretical SEM model and their relationships. The sequence and position of the latent constructs are based on theory and the accumulated knowledge and experience of the researcher [37]. Structural path models are specified with the latent variables/constructs on the left side of the path model as independent variables (antecedents), and the latent variables on the right side are dependent variables (outcome variables). In addition, latent constructs in SEM can also serve as both independent and dependent variables in the structural model [46].

The relationships (structural paths) between the constructs (inner model) connect the exogenous and endogenous constructs. As with multiple regression and CB-SEM, the strength of the relationships between the latent variables/constructs is represented by path coefficients, and the coefficients are the result of regressions of each endogenous latent variable on their direct predecessor exogenous constructs. In Figure 13.1, these path coefficients are labeled as  $b_1$ ,  $b_2$ ,  $b_3$  and  $b_4$ . Finally, recall that directly measured variables are typically represented in visual diagrams by rectangles, and the indirectly measured latent variables are represented as circles or ovals.

## Stage 4: Assessing Measurement Model Validity

Evaluation of PLS-SEM results is a two-step process, similar to that with CB-SEM. The focus of Step 1 (discussed in this section as Stage 4) is on evaluating the measurement models. Evaluation of the measurement models differs depending on whether the measurement is reflective or formative. For example, if the measurement theory is reflective, then Step 1 examines the size and significance of the loadings, reliability, and convergent and discriminant validity. But if the measurement theory is formative, then Step 1 examines convergent validity, multicollinearity, and the size and the significance of the indicator weights.

If the evaluation in Step 1 supports measurement quality, the researcher continues with the structural model evaluation in Step 2 [49]. Discussed in the next section as Stage 5, the second step evaluates the structural theory that tests the proposed hypotheses. Step 2 involves determining if the structural relationships are statistically significant and meaningful, and if the predictive ability of the theoretical model is acceptable, as measured by the  $R^2$ ,  $f^2$  effect size, and  $Q^2$ .

As with CB-SEM and other multivariate methods, rules of thumb are used in evaluating the results of the measurement and structural model estimations [15, 16, 42, 49, 58, 71, 120, 141]. Since rules of thumb are broadly applied guidelines for decision-making, they should not be interpreted the same for all aspects of a model or research contexts. For example, the rules of thumb differ for reflective or formative measurement models based on the assumption of each type of model. Moreover, the threshold for a rule of thumb may differ depending on the research

context. For example, most social science research applies  $<0.05$  as the level for statistical significance. But since statistical significance is sensitive to sample size, and PLS-SEM can be used with smaller sample sizes in the appropriate context, some scholars apply  $<0.10$  as the level for statistical significance with this method. The researcher is encouraged to apply rules of thumb whenever possible, but also remember that they must be conditioned on the research context.

## ASSESSING REFLECTIVE MEASUREMENT MODELS

Reflective measurement model assessment, whether with PLS-SEM or CB-SEM, involves four aspects of each model construct: size and significance of indicator loadings, construct reliability, convergent validity and discriminant validity. These issues were discussed in Chapters 9 and 10 for CB-SEM. But there are some differences between PLS-SEM and CB-SEM even in these basic aspects, so the following discussion will address each aspect briefly and focus on any differences between PLS-SEM and CB-SEM.

**Indicator Loadings** The assessment process begins by examining the indicator loadings. Loadings above 0.708 indicate the construct explains more than 50 percent of the indicator's variance. This confirms the indicator exhibits acceptable item reliability. While indicator loadings have the same interpretation in both PLS-SEM and CB-SEM, we should note that PLS-SEM typically has somewhat higher indicator loadings than CB-SEM [50] and this should be considered when interpreting loadings, especially when comparing models estimated with the two approaches. Moreover, while the general guidelines for the measurement properties discussed below are the same for both approaches, depending on the research context, the researcher may choose to evaluate PLS-SEM indicator loadings more conservatively given the tendency for somewhat higher loadings. At the same time, however, the higher loadings obtained with PLS-SEM enable the researcher to retain more items on the constructs, which generally results in higher content validity for the reflective measurement models.

**Construct Reliability** The next step is determining each construct's internal consistency reliability. While Cronbach's alpha is a widely used method of assessing reliability, it does not weight the individual indicators in the calculations. Jöreskog's [77] composite reliability overcomes this limitation since it weights the individual indicators based on their loadings and is therefore the preferred reliability approach.

Higher values indicate higher levels of reliability when interpreting internal consistency reliability results. For example, values between 0.60 and 0.70 are "acceptable in exploratory research," whereas results between 0.70 and 0.95 represent "satisfactory to good" reliability levels [49]. But reliability can also be so high (e.g., 0.95 or above) as to be unrealistic. When this situation arises, it most likely indicates the items are redundant (e.g., survey items are too similar or items are a slight variation of the same underlying data) or there is some systematic pattern in the responses (e.g., straight lining) [24]. When reliability is too high, particularly when it is examined in a pilot test, the scale should be redesigned to reduce redundancy before data is collected for the final study.

**Convergent Validity** Convergent validity is an overall metric of a reflective measurement model that measures the extent to which the indicators of a construct converge, thereby explaining the variance of the items. Many times referred to as communality, it is assessed by evaluating the average variance extracted (AVE) across all indicators associated with a particular construct. The AVE is the average (mean) of the squared loadings of all indicators associated with a particular construct. The rule of thumb for an acceptable AVE is 0.50 or higher. This level or higher indicates that on average the construct explains 50 percent or more of the variance of its indicators.

**Discriminant Validity** The final step in evaluating reflective measurement models is to assess their discriminant validity. This metric evaluates the extent to which a construct is distinct from other constructs. The underlying principle of discriminant validity is to assess how uniquely the indicators of a construct represent that construct (the shared variance within that construct) versus how much that construct is correlated with

all other constructs in the model (shared variance between constructs). Tests of discriminant validity are made for all pairs of reflective constructs within a model. Using the concept of AVE discussed above, discriminant validity is present when the shared variance within a construct (AVE) always exceeds the shared variance with all other constructs.

CB-SEM typically relies on the Fornell–Larcker [39] criterion, although an alternative method has been proposed [142]. The Fornell–Larcker method is a direct comparison of the AVEs of two constructs to the shared variance between the two constructs. In contrast, with PLS-SEM the recommended discriminant validity method is the Henseler et al. [69] heterotrait-monotrait ratio (HTMT) of correlations. The HTMT criterion is defined as the mean value of the indicator correlations across constructs (i.e., the heterotrait-heteromethod correlations) relative to the (geometric) mean of the average correlations of indicators measuring the same construct. The HTMT criterion is an estimate of the true correlation between two constructs if they were perfectly measured (i.e., if they were perfectly reliable). High HTMT values indicate a problem with discriminant validity. Based on simulation and previous research, Henseler et al. [69] recommend a value of 0.90 if the path model includes constructs that are conceptually similar (e.g., loyalty, cognitive satisfaction, and affective satisfaction). In other words, an HTMT value above 0.90 suggests a lack of discriminant validity. When the constructs are conceptually more distinct, a lower, more conservative threshold value of 0.85 is suggested [69]. Finally, in addition to examining the size of the HTMT value, researchers should use a bootstrapping procedure to determine whether the HTMT value is statistically significantly lower than one (1.0).

## ASSESSING FORMATIVE MEASUREMENT MODELS

As discussed in an earlier section, measurement models specified as formative are quite different in purpose and estimation than reflective models. As a result, they must also be evaluated differently from reflective measurement models. Evaluation of formative measurement models focuses on four aspects: (1) convergent validity, (2) indicator multicollinearity, (3) size and statistical significance of the indicator weights, and in some instances (4) relevance and statistical significance of the indicator loadings [49].

It should be noted that there is a difference in terminology in formative models relating to the magnitude of an indicator's relationship with the construct. In reflective measurement models we referred to loadings and they represented the correlation of each indicator with the construct. In formative models we obtain weights, similar to a regression weight for independent variables. This differs from a loading due to the different types of relationships between indicators and construct in reflective versus formative models. Reflective measurement models produce loadings because the relationship between the indicator and the construct is a bivariate correlation/regression, with

### Assessing Reflective Measurement Models

Indicator loadings should be a minimum of 0.708. Squared loadings (also referred to as item reliabilities) should be a minimum of 0.50.

Internal consistency reliability should be evaluated. Composite reliability is preferred but Cronbach's alpha is acceptable. The minimum recommended reliability is .70, except for exploratory studies, where .60 is considered the minimum. The maximum recommended reliability is .95, and preferably .90.

Convergent validity, as measured by average variance extracted (AVE), should be at least .50.

Discriminant validity should be evaluated using the HTMT method. For the HTMT method, the guideline is .90 for conceptually similar constructs, and .85 for conceptually distinct constructs. In addition, the HTMT value should be examined based on confidence intervals to determine if it is significantly different from one (1.0).

the indicator being the single dependent variable and the construct being the single independent variable. But formative measurement models have weights to represent the relationship between the construct and the indicators, because the construct is the dependent variable and the indicators are multiple independent variables in a regression equation.

**Convergent Validity** Convergent validity of reflectively measured constructs is based on communality, the degree to which the multiple indicators share variance with one another. For formatively measured constructs, however, this same method is not applicable since formative indicators are not assumed to exhibit multicollinearity (be correlated). Instead, redundancy analysis [15] assesses the degree to which the indicators relate to an additional reflectively measured construct (single or multi-item) that represents the same concept. As a general principle, a reflective indicator/construct representing the same concept is not likely to be included in a typical questionnaire. Researchers must, therefore, understand and plan for the assessment of convergent validity in the research design stage by including this reflectively measured construct or indicator in the final questionnaire. The construct, including its indicator(s), is separate from and in addition to the items in the formative construct. When possible researchers should avoid using single items for this additional construct, since compared to multi-item scales, single item measures have substantially lower levels of predictive validity [123], and they also have lower levels of variance that can create problems with a variance-based technique like PLS-SEM. Hair et al. [49] recommend that formatively measured constructs explain a minimum of 50 percent of the variance of the reflectively measured item(s), which corresponds to a path coefficient of approximately 0.708.

Note that when using secondary data with PLS-SEM, it is not possible to include a reflectively measured item as is possible in survey research. Therefore, to assess convergent validity with theoretical models that are based on secondary data, researchers should identify another indicator in the nomological net of the formative construct that is available and similar, and use it to complete the redundancy analysis.

**Indicator Multicollinearity** Formative constructs also need to be examined for high multicollinearity among the formative indicators. Assessing collinearity involves computing the variance inflation factor (VIF) for each indicator included in the formatively measured construct. As a guideline, the higher the VIF, the greater the level of collinearity, and VIF values above three are likely to indicate a problem, and above five are a definite indicator of high collinearity among the indicators, and thus a problem.

The VIF method is widely applied to assess collinearity, but is not a very rigorous test. In fact, collinearity can be a problem if VIF values are below three. An alternative approach to examining collinearity is to execute bivariate correlations that include all the indicators on each construct separately. If the results identify any bivariate correlation 0.50 or higher between any pair of construct indicators, this indicates collinearity is likely to be a problem. Whichever approach is chosen, researchers should interpret the results cautiously if collinearity appears to be present.

When collinearity is high among formative indicators, the weights will not be correct. The recommended solution in such situations is to retain all indicators, but to create two or more first order constructs that are connected with a higher order construct, which is the exogenous construct in the structural model [58].

**Statistical Significance of Indicator Weights** The third aspect in assessing formative measurement models is evaluating the size and statistical significance of the indicator weights. PLS-SEM is non-parametric and unlike CB-SEM and most other multivariate methods, it does not make any assumptions about the distribution of the error terms. To test statistical significance, therefore, researchers must apply bootstrapping. Bootstrapping is a resampling technique that creates subsamples (typically 1,000 or more) of the original sample, randomly with replacement, and re-estimates the model for each new subsample [49].

The result is a large number of subsamples that are used to construct a distribution of the parameters under consideration. The subsample distributions can be used to calculate standard errors, and ultimately to determine the statistical significance of the indicator weights, including  $t$  values and  $p$  values. Since bootstrapping is a random

process, the results will be different each time it is run. But when a large number of bootstrap samples are run, typically 1,000 or more, you can use the results to accurately define the distribution of the parameter estimates in calculating confidence intervals, and ultimately statistical significance. Note that the minimum number of bootstrap samples must be at least the size (number of observations) of the dataset, but preferably much larger since it is questionable how accurate the parameter estimates are when using a smaller number of bootstrap samples.

Bootstrapping can also be used to construct several different types of confidence intervals. We recommend bias-corrected and accelerated bootstrap confidence intervals [35] that adjust for skewness and other potential biases in the bootstrap distribution. If the confidence interval of an indicator weight includes zero, this means the weight is not statistically significant and the indicator should be considered for removal from the measurement model. But not until after the contribution of the indicator is assessed, as is described in the next section.

**Contribution of Indicator** The lack of significance for an indicator weight in a formative measurement model does not automatically lead to its removal. If an indicator weight is not significant the absolute contribution of the formative indicator to its construct should be evaluated. The absolute contribution of an indicator is the variance it shares with its construct, without considering any other indicators. The absolute contribution is based on the bivariate correlation of each indicator with the construct [14]. The absolute contribution to the construct is calculated through bivariate regressions of each indicator on its corresponding construct. The bivariate regression weight is also a standard output of most PLS software, and is generally referred to as an indicator loading.

To evaluate the absolute contribution in PLS-SEM, the following rules of thumb apply [49].

- Do not automatically remove an indicator if the weight (coefficient) is not statistically significant.
- When an indicator weight is not significant, the indicator is retained if the bivariate correlation (loading) is 0.50 or higher. But in general empirical support is not sufficient and keeping the indicator should also be based on theory and expert judgment.
- If the indicator weight is not significant and the bivariate correlation (loading) is below 0.50, the indicator has no empirical support to keep it and should be removed from the measurement model.

The absolute contribution of an indicator becomes even more useful as the number of indicators increases. Since a formative indicator weight is calculated in combination with all other indicators of the construct, the larger the number of indicators, the lower the expected average weight of the indicators [14]. This makes the use of only the significance and size of the indicator weights less relevant for establishing the contribution of an indicator. In addition, the absolute contribution measure should always be considered since it is unaffected by the number of formative indicators for a construct. Finally, an important consideration is if indicators are removed from a formative measurement model, the remaining indicators may not fully capture the entire domain of the construct. Recall that in contrast to reflective measurement models, formative indicators typically are not interchangeable. Thus, removing an indicator can change the meaning of the construct and reduce content validity [25]. As noted earlier, instead of removing weak formative indicators it may also be possible to create two or more first order constructs that are connected to a higher order construct, which is the exogenous construct in the structural model [58].

**Relevance of Indicators** The final step in evaluating formative measurement models is examining the relevance of each indicator in defining the construct. The indicator weights are standardized values that range from +1 to -1, with weights closer to +1 or -1 indicating either strong positive or strong negative relationships. At the same time, weights closer to zero (0) indicate weak relationships. Smaller weights are an indication of low relevance while larger weights are an indication of high relevance, at least for the context of the research being conducted. In the final analysis, researcher judgment is required to decide if removal of an item due to lack of relevance compromises the content validity of the formative construct. As a final note, it is possible to obtain values above +1 or below -1 if collinearity is high. When this situation arises, the researcher should revisit collinearity and reassess a previous decision to retain item(s) exhibiting high collinearity.

## Assessing Formative Measurement Models

Convergent validity is measured based on a measure of redundancy.

Multicollinearity between indicators should be minimal.

Indicator coefficients should be statistically significant and meaningful in size.

If indicator coefficients are not significant, then the size and significance of the indicator loadings is an alternative approach to evaluating whether formative indicators can be retained.

### SUMMARY

Reflective and formative measurement model approaches involve very different considerations when being evaluated. Measurement model assumptions for reflectively measured constructs assume the indicators are highly correlated and can therefore be evaluated based on internal consistency metrics. These metrics include internal consistency reliability, average variance extracted, and discriminant validity criteria all based on the correlations among the indicators. In contrast, measurement model assumptions for formatively measured constructs assume the indicators are seldom correlated so the same metrics cannot be used to evaluate this type of measurement model. The appropriate metrics for formative measurement models are redundancy to measure convergent validity, evaluation of indicator collinearity, and assessment of the significance and relevance of individual indicator weights, and if necessary the significance and relevance of individual indicator loadings.

## Stage 5: Assessing the Structural Model

Assuming the assessment of the measurement models is satisfactory, assessment of the structural model is next, which is Stage 5 of the PLS-SEM evaluation process. If the predictor constructs in the structural model exhibit collinearity, this can create problems in interpreting the results of PLS-SEM. Therefore, the first step in assessing the structural model is to examine the predictor constructs for collinearity. If collinearity is not a problem, the assessment of the structural model results is based primarily on its ability to predict the endogenous construct(s) and/or the indicators. The assessment of prediction is based on: the coefficient of determination ( $R^2$ ), the effect size ( $f^2$ ), cross-validated redundancy ( $Q^2$ ), and the sizes and significance of the path coefficients.

### COLLINEARITY AMONG PREDICTOR CONSTRUCTS

Just as in ordinary least squares regression, the path coefficients are affected by the presence of collinearity among the predictor constructs. This step is similar to the process followed with formative measurement model assessment, except in this analysis the latent variable scores of the exogenous constructs are used as input for the predictor constructs in a multiple regression to obtain VIF values. Recall that the higher the VIF, the greater the level of collinearity, and VIF values above 5 are a definite indication of collinearity among the predictor constructs. But just as with formative construct indicators, collinearity can be a problem sometimes even when VIF values are below 3.

In addition to examining the VIF values, the alternative approach of executing the bivariate correlations of the predictor construct scores is recommended. Again, if the results identify any bivariate correlations 0.50 or higher between any pair of constructs, this is an indication of the possibility of a problem with collinearity. After examining collinearity with both approaches, researchers should interpret the results cautiously when collinearity appears to be present. Note that if collinearity is a problem either in the measurement models or the structural model, in some situations this can be resolved by creating a higher order construct. For example, a predictor

construct exhibiting high collinearity among its indicators can be divided into two lower order constructs, or two constructs exhibiting high collinearity can be combined into a single higher order construct. This topic will be discussed in a later section.

## EXAMINING THE COEFFICIENT OF DETERMINATION

The primary statistical objective of PLS-SEM is prediction. If collinearity is not a problem, the initial criterion to examine is the coefficient of determination ( $R^2$ ). The coefficient of determination is a measure of in-sample predictive power [106, 126]. The  $R^2$  value ranges from 0 to 1, with 0 indicating no relationship and 1 indicating a perfect relationship. The higher the value of  $R^2$  the greater the explanatory power of the PLS structural model, and therefore the better the prediction of the endogenous constructs. As a guideline,  $R^2$  values of 0.75, 0.50, and 0.25 can be considered substantial, moderate, and weak, respectively [55, 71]. It should be noted, however, that in some research contexts  $R^2$  values of 0.10, and even lower, are considered satisfactory. Thus,  $R^2$  values should always be interpreted in the context of the study being conducted.

## EFFECT SIZE

The second criterion to examine is the effect size ( $f^2$ ). The effect size represents the change in the  $R^2$  value when a specified exogenous construct is omitted from the model. This metric is calculated to determine if removing a predictor construct from the structural model has a substantive impact on the endogenous constructs. To obtain this metric, the  $R^2$  value of the endogenous latent construct(s) is calculated when a selected predictor construct is included in the structural model, then when the predictor is not in the model, and the difference in the explanatory power is determined. Based on guidelines by Cohen [20],  $f^2$  values of 0.02, 0.15, and 0.35, respectively, represent small, medium, and large effects of an exogenous construct, and effect sizes of less than 0.02 indicate that there is no effect.

## BLINDFOLDING

The third criterion to examine is blindfolding ( $Q^2$ ). Blindfolding assesses the model's predictive power, also referred to as predictive relevance [48]. The  $Q^2$  value [40, 138] is obtained from the blindfolding procedure. To obtain the  $Q^2$  value, first the raw data values are omitted sequentially, the values are then imputed, and the model parameters are estimated. The parameter estimates are then used to predict the omitted raw data values. The process is repeated until every data point has been omitted and the model re-estimated. This is similar to other cross-validation measures such as the PRESS measure in multiple regression or the jack knife estimation approach in discriminant analysis.

When the difference between the original and predicted values is small, the result is a larger  $Q^2$ , which indicates the predictive accuracy is higher. As a guideline,  $Q^2$  values larger than zero for a particular endogenous construct indicate the path model's predictive accuracy is acceptable for that construct. At the same time, values less than zero indicate a lack of predictive relevance. There are two different approaches to calculating  $Q^2$  values. The cross-validated redundancy approach is considered a better criterion than the cross-validated communality approach.

## SIZE AND SIGNIFICANCE OF PATH COEFFICIENTS

The final criterion involves assessing the sizes and significance of the path coefficients. Recall that the bootstrapping procedure is executed to obtain significance. As with the assessment of formative indicator weights, the bootstrapping process uses standard errors to calculate  $t$  and  $p$  values for the path coefficients. The bias-corrected and accelerated confidence intervals [49] are also examined, and a path coefficient is significant at the 0.05 level if zero does not fall within the 95 percent (bias-corrected and accelerated) confidence interval. For example, if

the lower bound is  $-0.10$  and the upper bound is  $+0.27$ , the coefficient would not be significant, because the confidence interval includes zero. At the same time, if a path coefficient is  $0.18$  with  $0.12$  as the lower bound and  $0.21$  as the upper bound (95% confidence interval), it would be considered significant since zero does not fall into this confidence interval.

Examining relevance for the structural relationships involves evaluating the sizes of the path coefficients to see if they are large enough to be interpreted as meaningful. Recall that path coefficients are standardized and range from  $+1$  to  $-1$ , with  $+1$  indicating a perfect positive relationship,  $0$  indicating no relationship, and  $-1$  indicating a perfect negative relationship. We remind you again that values below  $-1$  and above  $+1$  may occur when collinearity is high, or some other error in the calculations has occurred. These “out of range” estimates should be resolved before accepting the final results.

Finally, when evaluating the size of a path coefficient, a value of  $0.50$  indicates when the independent construct score increases by one standard deviation unit, the dependent construct score will increase by  $0.50$  standard deviation units, assuming all other independent constructs remain constant. As with  $R^2$ , the context of the research should be considered in deciding if the coefficient is meaningful.

## SUMMARY

If measurement model metrics are satisfactory, assessment of the structural model follows. The predictor constructs in the path model are first examined for collinearity, since this can create problems in interpreting the results of PLS-SEM. If collinearity is not a problem, the structural model results are evaluated based primarily on the extent to which the exogenous constructs predict the endogenous construct(s). The assessment of prediction is based on: the coefficient of determination ( $R^2$ ), the effect size ( $f^2$ ), cross-validated redundancy ( $Q^2$ ), and the sizes and significance of the path coefficients.

## Assessing Structural Models

The first step in assessing the structural model is to examine multicollinearity among the predictor constructs. VIF is one alternative, but bivariate correlations between the latent variable (construct) scores should also be assessed. VIF values above three (3) suggest multicollinearity may be present, and bivariate correlations  $.50$  or higher may indicate problems with multicollinearity. Evaluation of multicollinearity is important because if present it can influence the size of the beta coefficients, and possibly change the sign.

The primary metric for assessing the structural model is the coefficient of determination ( $R^2$ ). The acceptable size for the  $R^2$  depends on the context of the research, and typically varies from one social science discipline to another.

The  $f^2$  effect size should be evaluated for in-sample predictive power.  $f^2$  values of  $0.02$ ,  $0.15$ , and  $0.35$ , respectively, represent small, medium, and large effects of an exogenous construct, and effect sizes of less than  $0.02$  indicate that there is no effect.

The predictive relevance of the structural model should be evaluated using the cross-validated redundancy method for blindfolding ( $Q^2$ ). As a general guideline,  $Q^2$  values larger than zero for a particular endogenous construct indicate the path model's predictive accuracy is acceptable.

To interpret the structural model paths, the coefficients should be meaningful in size and statistically significant. The acceptable size depends on the complexity of the path model and the context of the research. The statistical significance is obtained using the bootstrapping method.

## Stage 6: Advanced Analyses Using PLS-SEM

Obtaining a basic solution using PLS-SEM is often just the first step in the analysis. There are many advanced types of analysis that provide useful insights about the results. We will comment only on the major advanced analyses, and note that many methodological developments in PLS-SEM are rapidly emerging.

### MULTI-GROUP ANALYSIS OF OBSERVED HETEROGENEITY

Among the most useful types of analysis is **multi-group analysis** (PLS-MGA). Multi-group analysis is a process for examining separate groups of respondents to determine if there are differences in the model parameters between the groups [92]. This type of analysis is most often applied in situations where scholars include assessments of **observed heterogeneity** in their research design. A common example is the use of demographic characteristics to enable determining whether there are differences between groups based on age, gender, income, and so forth. Similarly, B-to-B researchers often collect data on firm size, market share, industry type, and so forth to be able to make group comparisons, or to include this type of data as a control variable when analyzing data. This type of information is observable and researchers plan ahead of time to collect it so they can make post hoc comparisons of the pre-defined groups.

The path coefficients of structural models are almost always different, but the important question is “Are the coefficients statistically significantly different?” PLS-MGA is an automated process available in the SmartPLS software that simultaneously tests all path relationships in the structural model to determine if they are significantly different.

### DETECTING UNOBSERVED HETEROGENEITY

PLS-MGA can also be applied to examine potential differences based on groups identified using methods to detect **unobserved heterogeneity**. Unobserved heterogeneity is a situation in which there are unobservable characteristics that cause differences in subgroups and thus the theoretical model should not be examined as a single homogeneous population. A similar approach has been applied in the past to identify subgroups in a single variable or multi-item construct using the traditional approach of cluster analysis, discussed in Chapter 4. The SmartPLS software includes several approaches to identify subgroups across the entire theoretical model simultaneously, if they exist. These techniques fall into the general class of models known as latent class techniques. We suggest you refer to the most recent discussions of this topic that summarize the benefits of these approaches [130, 131]. The software also easily examines construct measurement invariance by executing the MICOM procedure [69]. Multi-group analysis, invariance and latent class techniques can also be examined with CB-SEM models.

### CONFIRMATORY TETRAD ANALYSIS

Another type of analysis that may be useful for researchers applying PLS-SEM is confirmatory tetrad analysis (CTA-PLS) [45]. CTA-PLS is a method of empirically testing and evaluating the cause-effect relationships for latent variables as well as the specification of indicators in measurement models [49]. When applied, this test provides empirical evidence that can be used to minimize the likelihood of misspecification of formative and reflective indicators.

### MEDIATION EFFECTS

Scholars have assessed mediation for years, but recent developments in the concept are likely to result in a re-examination of the topic. Recent PLS-SEM software development facilitates application of procedures for examining mediation that overcome previous shortcomings noted by Zhao, Lynch and Chen [151] Moreover, PLS-SEM and CB-SEM are both superior approaches to examining mediation compared to the Preacher and Hayes [102] PROCESS approach that is limited to examining mediation with multiple regression applications.

Assessment of SEM relationships should not be limited to only direct effects. Instead, where applicable total effects, which are the sum of the direct effect and the indirect effects in the structural model, should be examined. Examination of total effects between constructs, including all their indirect effects, provides a more complete assessment of SEM relationships [94] and can easily be examined with most PLS-SEM software [49].

## MODERATION

Moderation is frequently examined with SEM. **Moderation** is a situation where the relationship between two variables/constructs is influenced by a third variable. Relationships that are potentially moderated by a third variable are often hypothesized a priori and subsequently tested. But moderation can also be tested on a post hoc basis when differences in relationships are identified from examining unobserved heterogeneity or applying multi-group analysis procedures. Moderated relationships are most often proposed for nominal/categorical variables, and can easily be examined using both PLS-SEM and CB-SEM. It is also possible, however, to have continuous moderating variables. Continuous moderators can easily be examined with the SmartPLS software, but are rather difficult to evaluate with CB-SEM, where the most often applied approach is to convert continuous variables to categorical variables before examining moderation.

## HIGHER-ORDER MEASUREMENT MODELS

Higher-order measurement models are relationships between constructs that simultaneously measure a concept at different levels of abstraction. Higher-order models are comprised of a single higher-order construct (HOCs) that represents the overall concept, and two or more lower-order constructs (LOCs) that measure more concrete facets of the HOC. Higher-order component models are increasingly appearing in scholarly journals. The application of higher order models most often involves second-order models, but HOCs are also possible with even higher levels of abstraction, such as third-order models [100]. The important caveat is that theory should support the use of higher order models in SEM. Application of higher-order measurement models when using PLS-SEM is very flexible and can be executed with two or more layers of LOCs. For more information on how to develop and apply higher-order models with PLS-SEM, see Hair et al. [58]. In contrast, CB-SEM HOCs require a minimum of three lower-order constructs to achieve identification [50].

## SUMMARY

There are other advanced applications that can be executed with PLS-SEM. Examples include important-performance analysis (IPMA) [73, 58] and cross-validation based on developing training and holdout samples to evaluate PLS-SEM predictions [136]. In the near future you can expect to see developments in PLS-SEM that evaluate endogeneity [JIM 2019], hierarchical linear modeling, and Bayesian approaches.

## PLS-SEM Illustration

As with CB-SEM, executing PLS-SEM is a two-stage process. The first stage is to examine and confirm the measurement model by executing a confirmatory composite analysis (CCA), and if confirmed, then the second stage is to assess the structural model. As with the CB-SEM illustration in previous chapters, HBAT would like to understand why some employees continue working for them longer than others. They know service quality and profitability can be improved when employees remain with the company longer. The six-stage SEM process begins with this goal in mind. For this illustration, we use the HBAT\_SEM 2018 dataset, available on the text's online resources. To refresh your memory, the five constructs are defined here:

- *Job satisfaction (JS)*. Reactions resulting from an appraisal of one's job situation.
- *Organizational commitment (OC)*. The extent to which an employee identifies and feels part of HBAT.

- *Staying intentions (SI)*. The extent to which an employee intends to continue working for HBAT and is not participating in activities that make quitting more likely.
- *Environmental perceptions (EP)*. Beliefs an employee has about day-to-day, physical working conditions.
- *Attitudes toward coworkers (AC)*. Attitudes an employee has toward the coworkers he/she interacts with on a regular basis.

HBAT is now ready to test the hypotheses using PLS-SEM. As part of the process, the results of PLS-SEM will be compared to the previous results using CB-SEM. The first three stages were examined in the CB-SEM chapters and do not need further discussion here. We first describe the path model and follow that with the illustration of Stage 4 of PLS-SEM.

### THEORETICAL PLS-SEM PATH MODEL

The application of PLS-SEM is conducted with the five latent variables (composites) used in the CB-SEM HBAT example. Recall the relevant theory leads HBAT to expect that the latent constructs EP, AC, JS, and OC are all related to SI, but in different ways. For example, a high EP score means that employees believe their work environment is comfortable and allows them to freely conduct their work. This environment is likely to create high job satisfaction, which in turn will facilitate a link between EP and SI.

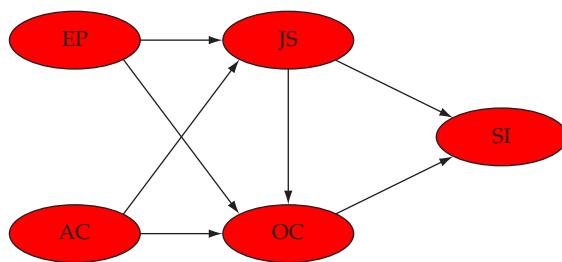
HBAT management will test the following hypotheses:

- $H_1$ : Environmental perceptions are positively related to job satisfaction.
- $H_2$ : Environmental perceptions are positively related to organizational commitment.
- $H_3$ : Attitudes toward coworkers are positively related to job satisfaction.
- $H_4$ : Attitudes toward coworkers are positively related to organizational commitment.
- $H_5$ : Job satisfaction is positively related to organizational commitment.
- $H_6$ : Job satisfaction is positively related to staying intentions.
- $H_7$ : Organizational commitment is positively related to staying intentions.

The theory can be expressed using a theoretical path model. The path model in Figure 13.2 depicts this theory. For simplicity, the measured indicator variables and their corresponding paths are not shown in the model. If a graphical interface is used with a PLS-SEM software program, then all measured variables would be shown on the path model. Note that different from CB-SEM, the analysis starts by drawing the theoretical path (structural) model that includes measurement models, not the measurement models connected by correlations as with CB-SEM. But when the analysis is run the first step is to examine the measurement models based on a process called confirmatory composite analysis (CCA). If the reliability and validity of the measurement models are confirmed, then the researcher moves on to the second step, which is analyzing the structural path model. Recall that reflective measurement models are assumed for all five HBAT constructs. See Hair et al. [54] for an in-depth illustration of formative measurement model assessment.

**Exogenous Constructs.** EP and AC are exogenous constructs in this model. These two constructs are considered to be determined by factors outside of this model. In practical terms, this means that no hypothesis predicts either of these constructs. Like independent variables in regression, they are used only to predict other variables/constructs. The two exogenous constructs—EP and AC—are drawn at the far left. No single-headed arrows are pointed to the exogenous constructs. Also, there is not a curved two-headed arrow between these two constructs, as was included in the CB-SEM model.

**Endogenous Constructs.** JS, OC, and SI are all endogenous constructs in the path model. Each is predicted by antecedent constructs in the model, and so each is also seen as an outcome based on the proposed hypotheses. Notice that both JS and OC are used as outcomes in some hypotheses and as predictors in others. This is acceptable



**Figure 13.2**  
Theoretical employee retention path model

#### Structural Model Paths and Hypotheses

Structural Relationships	Direction	Hypotheses
EP → JS	+	H1
EP → OC	+	H2
AC → JS	+	H3
AC → OC	+	H4
JS → OC	+	H5
JS → SI	+	H6
OC → SI	+	H7

in PLS-SEM, as well as CB-SEM, and a test for all hypotheses can be provided with one structural model test. This would not be possible with a single regression model because we would be limited to one dependent variable.

The structural path model begins to develop from the exogenous constructs. A path should connect any two constructs linked theoretically by a hypothesis. Therefore, after drawing the three endogenous constructs (JS, OC, and SI), single-headed arrows are placed connecting the predictor (exogenous) constructs with their respective outcomes based on the hypotheses. Each single-headed arrow represents a direct path and the legend in Figure 13.2 lists the seven hypotheses. For example,  $H_2$  hypothesizes a positive EP → OC relationship, and so on.

## Stage 4: Assessing Measurement Model Reliability and Validity

Once the SEM model is specified, a solution is estimated to provide an empirical measure of the relationships among measured variables and constructs represented by the theory. The results enable us to compare the theory against reality, so we see how well the theory can be confirmed by the data.

### PATH COEFFICIENTS

As noted earlier, the relationships between indicators and constructs (i.e., the *path estimates* linking constructs to indicator variables) are one of the most fundamental aspects in assessing measurement models. When testing a reflective measurement model, you expect to find relatively high loadings. After all, once the CCA is completed a good

conceptual understanding of the constructs and its items should exist. This knowledge, along with the preliminary empirical results, should enable the researcher to either accept or reject the validity of the measurement model and its indicators. We will focus on reflective measurement models in this discussion since that is the approach for all five constructs in the HBAT data.

**Size of Path Coefficients and Statistical Significance** Recall that rules of thumb suggest loadings should be at least .50 and ideally .708 or higher. For reflective measurement models, loadings of this size or larger confirm the indicators are strongly related to their associated constructs and are an indication of construct validity. These guidelines apply to the standardized loadings estimates that are a standard output in PLS-SEM software. If the measurement model is formative, then the researcher will examine the size and significance of the path coefficients. For more information on evaluating formative measurement model results, see Hair et al. [49].

In general, researchers should assess the statistical significance of all indicator loadings. Estimates that are not significant indicate an item should be removed. But a significant loading alone does not indicate an item is performing adequately. A loading can be significant (i.e.,  $p < .05$  but still considerably below  $.5$ ). Low loadings, even when significant, suggest that a variable is a candidate for deletion from reflective measurement models.

**Potential Problems** Loadings also should be examined for confusing estimates (coefficients) as indications of problems. An issue often overlooked is whether the loadings are logical. For example, items with the same valence (e.g., positive or negative wording) should produce the same sign. If an attitude scale consists of responses to four items—good, likeable, unfavorable, bad—then two items should carry positive loadings and two should carry negative loadings (unless they have previously been recoded). If the signs of the loadings are not opposite, the researcher should determine why and fix the problem or remove the offending items.

**Size of Loadings** The loadings for the HBAT PLS-SEM are shown in Table 13.1. All indicator loadings are above the recommended 0.708 level. Moreover, when compared to the loadings from CB-SEM shown in Chapter 11, the loadings are consistently higher. Some scholars have noted that the higher loadings of PLS-SEM are upwardly biased (i.e., consistency at large), but the basis of their comments is unfounded [124]. The loadings for CB-SEM and PLS-SEM are both proxies of the concepts being estimated. The statistical objective of the two approaches is different as is the algorithm, and that is the fundamental reason for the differences. Specifically, the CB-SEM loadings are based on the common factor model approach and the PLS-SEM loadings are based on the composite model (total variance) approach.

## CONSTRUCT RELIABILITY

As noted earlier, we know from both practical and theoretical perspectives that a concept cannot be measured perfectly and that **measurement error** is always present to some extent. Like CB-SEM, PLS-SEM is designed to assess and remove measurement error. **Reliability** is a measure of the degree to which a set of measured variables is internally consistent based on how highly interrelated the indicators are with each other. In other words, it represents the extent to which the indicators all measure the same thing. Reliability does not guarantee, however, that the measures indicate only one thing. In general, reliability is inversely related to measurement error. That is, as reliability increases the relationships between a construct and the indicators are larger, meaning that the construct explains more of the variance in each indicator. Thus, high reliability is associated with lower measurement error.

Reliability is assessed in PLS-SEM using two approaches. The traditional approach of Cronbach's alpha is used as well as composite reliability. Composite reliabilities for the five HBAT constructs range from a high of 0.93 for the AC construct to a low of 0.89 for both JS and OC, as shown in Table 13.2. All reliabilities exceed 0.70 indicating adequate reliability. Note that as anticipated, the composite reliability of the constructs is consistently higher than the Cronbach's alpha reliabilities. Moreover, the reliabilities are consistently higher than obtained with the CB-SEM approach.

**Table 13.1** HBAT PLS-SEM Loadings

	JS	OC	SI	EP	AC
JS <sub>1</sub>	0.78				
JS <sub>2</sub>	0.80				
JS <sub>3</sub>	0.76				
JS <sub>4</sub>	0.76				
JS <sub>5</sub>	0.82				
OC <sub>1</sub>		0.68			
OC <sub>2</sub>		0.90			
OC <sub>3</sub>		0.78			
OC <sub>4</sub>		0.88			
SI <sub>1</sub>			0.86		
SI <sub>2</sub>			0.90		
SI <sub>3</sub>			0.81		
SI <sub>4</sub>			0.89		
EP <sub>1</sub>				0.77	
EP <sub>2</sub>				0.87	
EP <sub>3</sub>				0.83	
EP <sub>4</sub>				0.87	
AC <sub>1</sub>					0.86
AC <sub>2</sub>					0.85
AC <sub>3</sub>					0.88
AC <sub>4</sub>					0.89

**Table 13.2** Reliability and Average Variance Extracted

	Cronbach's Alpha	Composite Reliability	Average Variance Extracted (AVE)
Job Satisfaction (JS)	0.84	0.89	0.61 (0.52)
Organizational Commitment (OC)	0.83	0.89	0.66 (0.56)
Staying Intentions (SI)	0.89	0.92	0.75 (0.67)
Environmental Perceptions (EP)	0.86	0.90	0.70 (0.60)
Attitude Toward Coworkers (AC)	0.89	0.93	0.76 (0.68)

Note: numbers in parentheses for AVEs are results from HBAT CB-SEM model.

## CONSTRUCT VALIDITY

Recall that in Chapter 3 *validity* was defined as the extent to which research is accurate, and the discussion centered on validating summated scales. Confirmatory Composite Analysis (CCA) eliminates the need to calculate summated scales because PLS-SEM programs compute latent construct scores for each respondent. This process allows relationships between constructs to be automatically corrected for the amount of error variance that exists in the construct measures.

One of the primary objectives of CCA/SEM is to assess the construct validity of a proposed measurement theory. **Construct validity** is the extent to which a set of measured items actually reflects the theoretical latent construct the items are designed to measure. Thus, it deals with the accuracy of measurement. Evidence of construct validity provides confidence that item measures taken from a sample are a good proxy of the population concept you are attempting to measure.

To assess construct validity, we examine convergent, discriminant, and nomological validity. Face validity was established in the CB-SEM chapters based on a qualitative review of the items.

**Convergent Validity** CCA evaluates convergent validity based on the average variance extracted (AVE), in a manner similar to CB-SEM. The AVE estimates range from 61 percent for JS to 76 percent for AC. All exceed the 50 percent rule of thumb, and all are higher than the AVEs obtained with the CB-SEM method. Overall, the AVE and reliability

evidence support convergent validity of the composite measurement models. Therefore, all the items are retained at this point and adequate evidence of convergent validity is provided.

**Discriminant Validity** **Discriminant validity** is the extent to which a construct is truly distinct from other constructs. Thus, high discriminant validity provides evidence that a construct is unique and captures some phenomena other measures do not. CCA provides two common ways of assessing discriminant validity when using PLS-SEM. Results for both methods are provided by the SmartPLS software.

In the past, the most commonly used criteria for discriminant validity with PLS-SEM was the Fornell–Larcker [39]. This test compares the average variance-extracted values for any two constructs with the square of the correlation estimate between these two constructs [14]. The variance-extracted estimates (AVEs) (shared variance within) should be greater than the squared interconstruct correlation estimate (shared variance between). This logic is based on the idea that a latent construct should explain more of the variance in its item measures than it shares with another construct. Passing this test provides some evidence of discriminant validity. The results of the Fornell–Larcker test are shown in Table 13.3 and indicate discriminant validity for the HBAT path model constructs.

The Fornell–Larcker test provides initial evidence of discriminant validity. The recommended method for assessing discriminant validity with PLS-SEM is the HTMT test [69]. The **HTMT (heterotrait-monotrait ratio** of the correlations) is the ratio of the between trait correlations to the within trait correlations. HTMT is the mean of all correlations of indicators between different constructs (i.e., the heterotrait-heteromethod correlations) relative to the (geometric) mean of the average correlations of indicators for the same construct (i.e., the monotrait-heteromethod correlations). The HTMT approach is thus an estimate of the true correlation between two constructs if they were perfectly measured (i.e., perfectly reliable).

The rule of thumb threshold for the HTMT is a value of 0.90 if the path model constructs are conceptually similar (e.g., affective satisfaction, cognitive satisfaction, and loyalty). That is, an HTMT value above 0.90 suggests a lack of discriminant validity. When the constructs in the path model are conceptually different, a lower threshold value of 0.85 is suggested [70]. The results for the HTMT test are shown in Table 13.4. All ratios are below the 0.85 level providing strong evidence of discriminant validity for the HBAT path model constructs.

Recall that PLS-SEM does not rely on any distributional assumptions and standard parametric significance tests cannot be applied to test whether the HTMT statistic is significantly different from 1. Instead, researchers use **bootstrapping** to derive a distribution of the HTMT statistic. Bootstrapping is applied with the HTMT statistic to derive standard errors for the estimates that are then used to develop **bootstrap confidence intervals**. The confidence

**Table 13.3** Fornell–Larcker Discriminant Validity Results

Constructs	AC	EP	JS	OC	SI
Attitudes toward Coworkers (AC)	<b>0.76</b>				
Environmental Perceptions (EP)	0.05	<b>0.70</b>			
Job Satisfaction (JS)	0.00	0.04	<b>0.61</b>		
Organizational Commitment (OC)	0.07	0.19	0.03	<b>0.66</b>	
Staying Intentions (SI)	0.08	0.25	0.04	0.21	<b>0.75</b>

Note: Bold numbers on the diagonal are AVEs and off diagonal numbers are squared interconstruct correlations

**Table 13.4** Results for HTMT Discriminant Validity

Constructs	AC	EP	JS	OC	SI
Attitudes Toward Coworkers (AC)					
Environmental Perceptions (EP)	0.257				
Job Satisfaction (JS)	0.066	0.244			
Organizational Commitment (OC)	0.275	0.495	0.209		
Staying Intentions (SI)	0.310	0.569	0.232	0.501	

interval is the range into which the true HTMT population value will fall, assuming a certain level of confidence (e.g., 95%). A confidence interval containing the value 1 indicates a lack of discriminant validity. But if the value 1 falls outside the range of the confidence interval, this indicates the two constructs are distinct.

The confidence intervals of the HTMT values derived from bootstrapping are shown in Table 13.5. None of the estimated confidence intervals contains a value of 1. Moreover, the bias associated with the bootstrapping estimates is very low. Thus, the confidence intervals for the HTMT values provide additional evidence of discriminant validity.

**Nomological Validity** Assessment of nomological validity is similar to the approach used with CB-SEM. Previous organizational behavior research suggests that more favorable evaluations of all constructs are generally expected to produce positive employee outcomes. For example, we expect these constructs to be positively related to whether an employee wishes to stay at HBAT. Moreover, satisfied employees are more likely to continue working for the same company.

Correlations between the latent variable scores produced by the PLS-SEM software for each construct are shown in Table 13.6. The results support the prediction that these constructs are positively related to one another. Specifically, satisfaction, organizational commitment, environmental perceptions, and attitudes toward coworkers all have significant positive correlations with staying intentions. In fact, only one correlation is inconsistent with this prediction. The correlation estimate between AC and JS is positive, but not significant. Because all other correlations are consistent, we are not concerned with this one exception.

Nomological validity can also be supported by demonstrating that the constructs are related to other constructs not included in the model in a manner that supports the theoretical framework. To do this, we can examine other constructs in the HBAT data that depict key relationships in the theoretical framework. In addition to the measured variables used as indicators for the constructs, several classification variables such as employee age, years of experience, and performance were also collected.

These other measures are helpful in further establishing nomological validity. Correlations between these three items and the factor scores for each measurement model construct provide additional validity information.

**Table 13.5** Bias Corrected Confidence Intervals Derived for HTMT Test

Path	Original Sample	Sample Mean	Bias	2.5%	97.5%
EP → AC	0.257	0.253	-0.004	0.138	0.366
JS → AC	0.066	0.095	0.029	0.042	0.080
JS → EP	0.244	0.242	-0.003	0.154	0.345
OC → AC	0.275	0.274	-0.001	0.178	0.373
OC → EP	0.495	0.497	0.002	0.398	0.609
OC → JS	0.209	0.213	0.004	0.111	0.316
SI → AC	0.310	0.308	-0.002	0.216	0.411
SI → EP	0.569	0.567	-0.002	0.459	0.670
SI → JS	0.232	0.230	-0.001	0.143	0.322
SI → OC	0.501	0.496	-0.005	0.400	0.603

**Table 13.6** HBAT Construct Latent Variable Scores Correlation Matrix

	JS	OC	SI	EP	AC
JS	1.00				
OC	0.184**	1.00			
SI	0.206**	0.462**	1.00		
EP	0.211**	0.437**	0.497**	1.00	
AC	0.053	0.259**	0.279**	0.228**	1.00

Significance level: \* = .05, \*\* = .01

Note: Values below the diagonal are correlations among latent variable scores; diagonal elements are construct variances.

Specifically, performance is positively correlated with four of the five constructs, age is positively correlated with three constructs, and experience is positively correlated with two of the constructs (all of these positive correlations are statistically significant). These results show the correlations are consistent with the theoretical expectations. Thus, the analysis of the correlations among the measurement model construct scores and the analysis of correlations between these constructs and other variables both support the nomological validity of the model.

### **HBAT CCA SUMMARY**

The HBAT CCA results provide strong support for the measurement models. Evidence of reliability is excellent, and construct validity is present based on convergent, discriminant, and nomological validity. Thus, HBAT can be fairly confident at this point that the five reflectively measured composite constructs are reliable and valid.

## **Stage 5: Assessing the Structural Model**

The structural model shown in the path diagram in Figure 13.2 can now be assessed. To do so, the emphasis will be on the predictive ability of the SEM model and then whether the structural relationships are consistent with theoretical expectations.

According to Hair et al. [49], the steps for assessing the structural model are: (1) examine collinearity, (2) evaluate the size and significance of the structural path relationships, (3) assess the  $R^2$ , (4) examine the  $f^2$  effect size, and (5) evaluate the predictive relevance based on  $Q^2$ .

The first step is to examine the exogenous constructs for collinearity. This is necessary since the path coefficients are based on OLS regressions and may be biased if multicollinearity is present [49]. All of the VIF values were well below the guideline of 3, with the highest VIF value being 1.1. Thus, collinearity is not a problem for the HBAT structural model.

Step two is evaluating the significance and size of the structural path coefficients. To obtain the significance levels, the bootstrapping option was run using 5,000 subsamples [49]. Table 13.7 shows the coefficients,  $t$  values, significance levels ( $p$  values), and 95 percent confidence intervals. All of the path coefficients except one (H3: AC  $\rightarrow$  JS) are statistically significant.

The path coefficients in the structural model may be significant even if their size is very small, typically due to large sample sizes. Therefore, it is important to determine if the size of the path coefficients is meaningful, which may vary depending on the research context. Examination of the sizes of the path coefficients for the significant HBAT relationships indicates they are all meaningful. When the sizes of the path coefficients are compared to the CB-SEM coefficients, two of the PLS-SEM path coefficients were considerably lower, two were slightly lower, one was the same, and two were slightly higher.

The third step is to examine the predictive ability of the structural model. This process begins by examining the  $R^2$  values. Using the PLS-SEM approach the explained variance is 4.0 percent for job satisfaction (JS), 22.6 percent for organizational commitment (OC), and 22.9 percent for staying intentions (SI). Using the CB-SEM approach the explained variance is 6.2 percent for job satisfaction, 31.9 percent for commitment, and 34.8 percent for staying

**Table 13.7** Structural Model Path Coefficients and Significance Testing

Hypotheses	Structural Relationships	Path Coefficients	T Statistics	P Values	95% Confidence Intervals
H1	EP $\rightarrow$ JS	0.21 (.25)	4.66	0.00	[0.125; 0.289]
H2	EP $\rightarrow$ OC	0.38 (.45)	7.27	0.00	[0.279; 0.473]
H3	AC $\rightarrow$ JS	0.01 (.01)	0.11	0.92	[-0.102; 0.097]
H4	AC $\rightarrow$ OC	0.17 (.20)	4.09	0.00	[0.085; 0.241]
H5	JS $\rightarrow$ OC	0.10 (.09)	2.12	0.04	[-0.007; 0.174]
H6	JS $\rightarrow$ SI	0.13 (.12)	3.176	0.00	[0.047; 0.199]
H7	OC $\rightarrow$ SI	0.44 (.55)	10.01	0.00	[0.355; 0.419]

Note: Numbers in parentheses for path coefficients are results from HBAT CB-SEM model.

intentions. At first glance, the implication is that the CB-SEM approach explained variance results are relatively higher. But in reality that is not true. Recall that the CB-SEM method is predicting the percentage of common variance in the indicators of the dependent constructs that are retained after GOF is achieved, while the PLS-SEM method is predicting the percentage of the total variance in the indicators of the dependent constructs. The only way to make a direct comparison is to know what percentage of the total variance is used in calculating the common variance used in the CB-SEM predictions of the endogenous constructs. Since this number is not provided as an output by the software a direct comparison of the predictive ability of the two methods is not possible. For a more detailed analysis and comparison of the predictive ability of PLS-SEM and CB-SEM, see [50].

If the focus of your research is on prediction of the variance in the dependent variable construct(s), in general PLS-SEM is the recommended method. In addition, simulation studies reported by Becker et al. [3] and Evermann and Tate [36] indicate PLS-SEM outperforms factor-based SEM in terms of prediction. Based on their results, PLS-SEM enables researchers to specify explanatory, theory-based models to aid in theory development, evaluation, prediction and confirmation.

Step four is examining the  $f^2$  effect sizes, which are shown in Table 13.8. The columns are results for the three dependent variables—JS, OC and SI (endogenous constructs). The predictor constructs are shown on the left, and the path coefficient for each predictor is shown first under its respective dependent variables, and then beside are the  $f^2$  effect sizes. Recall that the effect size represents the change in the  $R^2$  value as a result of predictive impact of a specific predictor variable. Guidelines by Cohen [20] indicate  $f^2$  values of 0.02, 0.15, and 0.35, respectively, represent small, medium, and large effects of an exogenous construct. In addition, effect sizes of less than 0.02 indicate that there is no effect. Four of the  $f^2$  effect sizes indicate a small effect (.044; .035; .042; .020), two are medium effects (.178; .242), and one is no effect (.000). Thus, the path model shows moderate in-sample predictive ability based on the sizes of the  $R^2$ , the path coefficients, and the  $f^2$  effect sizes.

The  $Q^2$  blindfolding, another estimate of path model prediction, was examined as the fifth step in evaluating the structural model. The construct cross-validated redundancy method was executed to obtain the  $Q^2$  blindfolding results.  $Q^2$  values larger than 0.0 suggest the path model has predictive relevance for the endogenous constructs; values below 0.0 indicate no predictive relevance. The  $Q^2$  values of 0.024, 0.136, and 0.158 for JS, OC, and SI, respectively, indicate meaningful predictive relevance for the HBAT path model using PLS-SEM [49].

**Table 13.8** Path Coefficients and  $f^2$  Effect Sizes

Predictor Construct	Endogenous Constructs					
	JS		OC		SI	
	Path Coefficient	$f^2$ Effect Size	Path Coefficient	$f^2$ Effect Size	Path Coefficient	$f^2$ Effect Size
EP	0.210	0.044	0.380	0.178		
AC	0.010	0.000	0.170	0.035		
JS			0.100	0.042	0.130	0.020
OC					0.440	0.242

## HBAT PLS-SEM Summary

SEM analysis involves assessment of both measurement theory and structural theory. In this chapter, we learned how to apply and interpret the findings of an alternative approach to examining structural equation models. PLS-SEM is not just another multivariate statistical procedure. It is a way of empirically exploring and testing theoretical relationships between multiple variables and constructs. Other statistical tools are available for examining relationships between variables. But when a researcher becomes knowledgeable enough about a subject matter to specify a set of theoretical relationships between constructs, in addition to the way these constructs are measured, PLS-SEM is an appropriate and very powerful tool. This chapter highlights numerous key points associated with PLS-SEM, including the following:

This chapter helps you to do the following:

**Understand the distinguishing characteristics of PLS-SEM.** PLS-SEM is suitable for analyzing complex structural models increasingly encountered by researchers in this era of Big Data, particularly from secondary data sources. PLS-SEM is an alternative approach to SEM whose primary statistical objective is different than covariance-based structural modeling, and can obtain solutions that are not possible with CB-SEM. While prediction is possible with CB-SEM, its primary statistical objective is confirming theory whereas the primary statistical objective of PLS-SEM is prediction that maximizes the variance explained in the dependent variable(s). Moreover, the assumptions of CB-SEM are quite rigorous, such as a normal distribution, while the assumptions of PLS-SEM are quite flexible since it is a non-parametric statistical method.

**Explain how the PLS-SEM algorithm is estimated.** PLS-SEM derives solutions using a three-stage approach. In the first stage the PLS-SEM algorithm iteratively determines the inner weights for the structural path relationships and the construct scores. The PLS-SEM algorithm uses two different estimation modes to compute the outer weights/loadings. Mode A uses the bivariate correlation between each indicator and the construct to determine the outer weights/loadings, typically associated with reflectively measured constructs. In contrast, Mode B uses the OLS regression results to obtain the outer weights, typically associated with formatively measured constructs. The second and third stages use the final latent variable scores as input for a series of ordinary least squares regressions. These OLS regressions compute the final outer loadings for reflective measurement models, the final outer weights for formative measurement models, and the final path coefficients as well as related elements.

**Describe the stages of the PLS-SEM decision process.** The PLS-SEM stages of the decision process are: (1) Defining research objectives and selecting constructs; (2) Designing a study to produce empirical results; (3) Specifying the measurement and structural models; (4) Assessing measurement model validity; and (5) Assessing the structural model. If advanced analyses are performed then an additional sixth stage of decisions, which varies by the type of analysis, is necessary.

**Distinguish between common factor modeling and composite modeling.** PLS-SEM is a composite modeling method that uses total variance in calculating solutions. The method easily accommodates both reflective and formative measurement model types without encountering identification problems. Whether the measurement models are specified as reflective or formative, PLS-SEM always uses linear combinations of indicators to represent the latent constructs when estimating model parameters. In contrast, CB-SEM, also referred to as common factor-based SEM (i.e., covariance-based SEM), uses only common variance and involves simultaneous maximum likelihood estimation of the relationships between constructs and measured indicator variables, as well as among latent constructs. CB-SEM is primarily designed for reflective measurement models.

**Explain how to assess reflective and formative measurement models.** Reflective measurement models assume the indicators are highly correlated, and are assessed based on internal consistency reliability, most often composite reliability, convergent validity by calculating the average variance extracted (AVE), and discriminant validity, with the HTMT method the recommended approach. In contrast, formative measurement models assume the construct indicators are not correlated and internal consistency metrics cannot be used to evaluate them. Formative measurement is assessed using a three-step approach: (1) redundancy to measure convergent validity; (2) collinearity to determine if there is a problem; and (3) bootstrapping to assess the significance of indicators as well as the size of the weights, and possibly loadings, to determine if they are meaningful.

**Interpret the results of the HBAT data used with PLS-SEM.** As with CB-SEM, executing PLS-SEM is a two-stage process. The first stage is to examine and confirm the measurement model by executing a confirmatory composite analysis (CCA), and if confirmed, then the second stage is to assess the structural model. All PLS-SEM indicator loadings are above the recommended 0.708 when executed with the HBAT data, and when compared to the CB-SEM results, the loadings are consistently higher. Composite reliabilities for the PLS-SEM solution all exceed 0.70 and AVEs are substantially above the recommended 0.50. Both PLS-SEM metrics are consistently higher than those obtained with

the CB-SEM approach. Discriminant validity is acceptable for the HBAT constructs for both PLS-SEM and CB-SEM. All of the PLS-SEM path coefficients except one (H3: AC → JS) are statistically significant. In addition, the sizes of the path coefficients for the significant relationships were all considered meaningful. When the sizes of the path coefficients are compared to the CB-SEM coefficients, two of the PLS-SEM path coefficients were considerably lower, two were slightly lower, one was the same, and two were slightly higher. Solutions for both methods are considered good.

**Describe when PLS-SEM and CB-SEM are the appropriate structural modeling method.** PLS-SEM and CB-SEM are two different approaches to structural equation modeling. Both SEM methods evaluate two models, the measurement model (representing how measured variables represent the constructs) and the structural model (showing how constructs are associated with each other), and theory provides the foundation upon which both the structural and measurement models are specified. While CB-SEM and PLS-SEM are comparable in their basic elements, there are some distinct differences. Prediction is possible with covariance-based structural modeling, but the primary statistical objective of CB-SEM is confirming theory by estimating a new covariance matrix that is not significantly different from the original observed covariance matrix. In contrast, the primary statistical objective of PLS-SEM is prediction that maximizes the explained variance in the dependent variable(s). The CB-SEM method is parametric and therefore requires a normal distribution in your data and other restrictive assumptions. But PLS-SEM is non-parametric and much more flexible in meeting the required assumptions. This flexibility means non normal data distributions and skewness are not a problem for PLS-SEM applications, nonmetric scales are more easily accommodated, and thus good solutions to a wide variety of research applications are possible much more often.

What are the fundamental differences between the PLS-SEM and CB-SEM methods of structural equation modeling?

Why does PLS-SEM not require a Goodness of Fit measure?

What are the similarities and differences in executing PLS-SEM and CB-SEM?

What is the difference between the common factor model and the composite model?

How does evaluation of construct validity differ between reflective and formative measurement models?

What advantages in specifying and executing higher order models does PLS-SEM have over CB-SEM?

Why can PLS-SEM explore both continuous and categorical moderator variables, while CB-SEM can only examine a categorical moderator variable?

A list of suggested readings and all of the other materials (i.e., datasets, etc.) relating to the techniques in this chapter are available in the online resources at the text's websites (Cengage Brain or [www.mvstats.com](http://www.mvstats.com))

- 1 Aaker, D. A. 1991. *Managing Brand Equity: Capitalizing on the Value of a Brand Name*. New York: Free Press.
- 2 Baumgartner, H., and C. Homburg. 1996. Applications of Structural Equation Modeling in Marketing and Consumer Research: A Review. *International Journal of Research in Marketing* 13: 139–61.
- 3 Becker, J.-M., and I. R. Ismail. 2016. Accounting for Sampling Weights in PLS Path Modeling: Simulations and Empirical Examples. *European Management Journal* 34: 606–17.
- 4 Becker, J.-M., A. Rai, and E. E. Rigdon. 2013a. Predictive Validity and Formative Measurement in Structural

- Equation Modeling: Embracing Practical Relevance. In *Proceedings of the International Conference on Information Systems*, Milan.
- 5 Becker, J.-M., A. Rai, C. M. Ringle, and F. Völckner. 2013b. Discovering Unobserved Heterogeneity in Structural Equation Models to Avert Validity Threats. *MIS Quarterly* 37: 665–94.
  - 6 Bentler, P. M., and W. Huang. 2014. On Components, Latent Variables, PLS and Simple Methods: Reactions to Rigdon's Rethinking of PLS. *Long Range Planning* 47: 138–45.
  - 7 Bollen, K. A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.
  - 8 Bollen, K. A. 2002. Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology* 53: 605–34.
  - 9 Bollen, K. A. 2011. Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models. *MIS Quarterly* 35: 359–72.
  - 10 Bollen, K. A., and S. Bauldry. 2011. Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates. *Psychological Methods* 16: 265–84.
  - 11 Bollen, K. A., and A. Diamantopoulos. 2017. In Defense of Causal–Formative Indicators: A Minority Report. *Psychological Methods*, 22: 581–96.
  - 12 Bollen, K. A., and R. Lennox. 1991. Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin* 110: 305–14.
  - 13 Borsboom, D., G. J. Mellenbergh, and J. van Heerden. 2003. The Theoretical Status of Latent Variables. *Psychological Review* 110: 203–19.
  - 14 Centefelli, R. T., and G. Bassellier. 2009. Interpretation of Formative Measurement in Information Systems Research. *MIS Quarterly* 33: 689–708.
  - 15 Chin, W. W. 1998. The Partial Least Squares Approach to Structural Equation Modeling. In G. A. Marcoulides (ed.), *Modern Methods for Business Research*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 295–358.
  - 16 Chin, W. W. 2010. How to Write Up and Report PLS Analyses. In Vinzi V. Esposito, W. W. Chin, J. Henseler, and H. Wang (eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Springer Handbooks of Computational Statistics Series, Vol. II. New York: Springer, pp. 655–90.
  - 17 Chin W. W., B. L. Marcolin, and P. R. Newsted. 2003. A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and an Electronic-Mail Emotion/Adoption Study. *Information Systems Research* 14: 189–217.
  - 18 Chou, C.-P., P. M. Bentler, and A. Satorra. 1991. Scaled Test Statistics and Robust Standard Errors for Non-Normal Data in Covariance Structure Analysis: A Monte Carlo Study. *British Journal of Mathematical and Statistical Psychology* 44: 347–57.
  - 19 Cochran, W. G. 1977. *Sampling Techniques*, 3rd edn. New York: Wiley.
  - 20 Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum Associates.
  - 21 Coltman, T., T. M. Devinney, D. F. Midgley, and S. Venaik. 2008. Formative Versus Reflective Measurement Models: Two Applications of Formative Measurement. *Journal of Business Research* 61: 1250–62.
  - 22 Diamantopoulos, A. 2006. The Error Term in Formative Measurement Models: Interpretation and Modeling Implications. *Journal of Modelling in Management* 1: 7–17.
  - 23 Diamantopoulos, A. 2011. Incorporating Formative Measures into Covariance-Based Structural Equation Models. *MIS Quarterly* 35: 335–58.
  - 24 Diamantopoulos, A., M. Sarstedt, C. Fuchs, P. Wilczynski, and S. Kaiser. 2012. Guidelines for Choosing Between Multi-Item and Single-Item Scales for Construct Measurement: A Predictive Validity Perspective. *Journal of the Academy of Marketing Science* 40: 434–49.
  - 25 Diamantopoulos, A., and H. M. Winklhofer. 2001. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research* 38: 269–77.
  - 26 Dijkstra, T. K. 2010. Latent Variables and Indices: Herman Wold's Basic Design and Partial Least Squares. In Vinzi V. Esposito, W. W. Chin, J. Henseler, and H. Wang (eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Springer Handbooks of Computational Statistics Series, Vol. II. New York: Springer, pp. 23–46.
  - 27 Dijkstra, T. K., and J. Henseler. 2011. Linear Indices in Nonlinear Structural Equation Models: Best Fitting Proper Indices and Other Composites. *Quality & Quantity* 45: 1505–18.
  - 28 Dijkstra, T. K., and J. Henseler. 2015a. Consistent and Asymptotically Normal PLS Estimators for Linear Structural Equations. *Computational Statistics & Data Analysis* 81: 10–23.
  - 29 Dijkstra, T. K., and J. Henseler. 2015b. Consistent Partial Least Squares Path Modeling. *MIS Quarterly* 39: 297–316.
  - 30 Dijkstra, T. K., and K. Schermelleh-Engel. 2014. Consistent Partial Least Squares for Nonlinear Structural Equation Models. *Psychometrika* 79: 585–604.
  - 31 do Valle, P. O., and G. Assaker. 2016. Using Partial Least Squares Structural Equation Modeling in Tourism Research: A Review of Past Research and Recommendations for Future Applications. *Journal of Travel Research* 55: 695–708.
  - 32 Eberl, M. 2010. An Application of PLS in Multi-Group Analysis: The Need for Differentiated Corporate-Level Marketing in the Mobile Communications Industry. In Vinzi V. Esposito, W. W. Chin, J. Henseler, and

- H. Wang (eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Springer Handbooks of Computational Statistics Series, Vol. II. New York: Springer, pp. 487–514.
- 33 Eberl, M., and M. Schwaiger. 2005. Corporate Reputation: Disentangling the Effects on Financial Performance. *European Journal of Marketing* 39: 838–54.
- 34 Edwards, J. R., and R. P. Bagozzi. 2000. On the Nature and Direction of Relationships Between Constructs and Measures. *Psychological Methods* 5: 155–74.
- 35 Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- 36 Evermann, J., and M. Tate. 2016. Assessing the Predictive Performance of Structural Equation Model Estimators. *Journal of Business Research* 69: 4565–82.
- 37 Falk, R. F., and N. B. Miller. 1992. *A Primer for Soft Modeling*. Akron, OH: University of Akron Press.
- 38 Fornell, C. G., and F. L. Bookstein. 1982. Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research* 19: 440–52.
- 39 Fornell, C., and D. Larcker. 1981. Structural Equation Models with Unobservable Variables and Measurement Error: Algebra and Statistics. *Journal of Marketing Research* 18: 382–8.
- 40 Geisser, S. 1974. A Predictive Approach to the Random Effects Model. *Biometrika* 61: 101–7.
- 41 Goodhue, D. L., W. Lewis, and R. Thompson. 2012. Does PLS Have Advantages for Small Sample Size or Non-Normal Data? *MIS Quarterly* 36: 981–1001.
- 42 Götz, O., K. Liehr-Gobbers, and M. Krafft. 2010. Evaluation of Structural Equation Models Using the Partial Least Squares (PLS) Approach. In Vinzi V. Esposito, W. W. Chin, J. Henseler, and H. Wang (eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Springer Handbooks of Computational Statistics Series, Vol. II. New York: Springer, pp. 691–711.
- 43 Grace, J. B., and K. A. Bollen. 2008. Representing General Theoretical Concepts in Structural Equation Models: The Role of Composite Variables. *Environmental and Ecological Statistics* 15: 191–213.
- 44 Gregor, S. 2006. The Nature of Theory in Information Systems. *MIS Quarterly* 30: 611–42.
- 45 Gudergan, S. P., C. M. Ringle, S. Wende, and A. Will. 2008. Confirmatory Tetrad Analysis in PLS Path Modeling. *Journal of Business Research* 61: 1238–49.
- 46 Haenlein, M., and A. M. Kaplan. 2004. A Beginner's Guide to Partial Least Squares Analysis. *Understanding Statistics* 3: 283–97.
- 47 Hahn, C., M. D. Johnson, A. Herrmann, and F. Huber. 2002. Capturing Customer Heterogeneity Using a Finite Mixture PLS Approach. *Schmalenbach Business Review* 54: 243–69.
- 48 Hair, J. F., C. L. Hollingsworth, A. B. Randolph, and A. Y. L. Chong. 2017. An Updated and Expanded Assessment of PLS-SEM in Information Systems Research. *Industrial Management & Data Systems* 117: 4442–58.
- 49 Hair, J. F., G. T. M. Hult, C. M. Ringle, and M. Sarstedt. 2017. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 2nd edn. Thousand Oaks, CA: Sage.
- 50 Hair, J. F., L. Matthews, R. Mathews, and M. Sarstedt. 2018. PLS-SEM or CB-SEM: Updated Guidelines on Which Method To Use. *International Journal of Multivariate Data Analysis* 1: 107–23.
- 51 Hair, J. F., C. M. Ringle, and M. Sarstedt. 2012. Partial Least Squares: The Better Approach to Structural Equations Modeling? *Journal of Long Range Planning* 46: 312–9.
- 52 Hair, J. F., M. Sarstedt, L. Hopkins, and V. Kuppelwieser. 2014. Partial Least Squares Structural Equation Modeling (PLS-SEM): An Emerging Tool in Business Research. *European Business Review* 26: 106–21.
- 53 Hair, J. F., M. Sarstedt, L. Matthews, and C. Ringle. 2016. Identifying and Treating Unobserved Heterogeneity with FIMIX-PLS: Part I—Method. *European Business Review* 28: 63–76.
- 54 Hair, J. F., G. T. M. Hult, C. M. Ringle, M. Sarstedt, and K. O. Thiele. 2017. Mirror, Mirror on the Wall: A Comparative Evaluation of Composite-based Structural Equation Modeling Methods *Journal of the Academy of Marketing Science*, <http://dx.doi.org/10.1007/s11747-017-0517-x>.
- 55 Hair, J. F., C. M. Ringle, and M. Sarstedt. 2011. PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice* 19: 139–51.
- 56 Hair, J. F., C. M. Ringle, and M. Sarstedt. 2013. Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results and Higher Acceptance. *Long Range Planning* 46: 1–12.
- 57 Hair, J. F., M. Sarstedt, T. M. Pieper, and C. M. Ringle. 2012. The Use of Partial Least Squares Structural Equation Modeling in Strategic Management Research: A Review of Past Practices and Recommendations for Future Applications. *Long Range Planning* 45: 320–40.
- 58 Hair, J. F., M. Sarstedt, C. M. Ringle, and S. P. Gudergan. 2018. *Advanced Issues in Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Thousand Oaks, CA: Sage.
- 59 Hair, J. F., M. Sarstedt, C. M. Ringle, and J. A. Mena. 2012. An Assessment of the Use of Partial Least Squares Structural Equation Modeling in Marketing Research. *Journal of the Academy of Marketing Science* 40: 414–33.
- 60 Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. 2015. The Extent and Consequences of P-Hacking in Science. *PLOS Biology* 13: e1002106, doi.org/10.1371/journal.pbio.1002106.
- 61 Helm, S., A. Eggert, and I. Garnefeld. 2010. Modeling the Impact of Corporate Reputation on Customer Satisfaction and Loyalty Using PLS.

- In Vinzi V. Esposito, W. W. Chin, J. Henseler, and H. Wang (eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Springer Handbooks of Computational Statistics Series, Vol. II. New York: Springer, pp. 515–34.
- 62 Henseler, J. 2010. On the Convergence of the Partial Least Squares Path Modeling Algorithm. *Computational Statistics* 25: 107–20.
- 63 Henseler, J. 2017. Using Variance-Based Structural Equation Modeling for Empirical Advertising Research at the Interface of Design and Behavioral Research. *Journal of Advertising* 46: 178–92.
- 64 Henseler, J., and W. W. Chin. 2010. A Comparison of Approaches for the Analysis of Interaction Effects Between Latent Variables Using Partial Least Squares Path Modeling. *Structural Equation Modeling* 17: 82–109.
- 65 Henseler, J., T. K. Dijkstra, M. Sarstedt, C. M. Ringle, A. Diamantopoulos, D. W. Straub, D. J. Ketchen, J. F. Hair, G. T. M. Hult, and R. J. Calantone. 2014. Common Beliefs and Reality about Partial Least Squares: Comments on Rönkkö & Evermann (2013). *Organizational Research Methods* 17: 182–209.
- 66 Henseler, J., G. Fassott, T. K. Dijkstra, and B. Wilson. 2012. Analyzing Quadratic Effects of Formative Constructs by Means of Variance-Based Structural Equation Modelling. *European Journal of Information Systems* 21: 99–112.
- 67 Henseler, J., G. S. Hubona, and P. A. Ray. 2016. Using PLS Path Modeling in New Technology Research: Updated Guidelines. *Industrial Management & Data Systems* 116: 1–19.
- 68 Henseler, J., C. M. Ringle, and M. Sarstedt. 2012. Using Partial Least Squares Path Modeling in International Advertising Research: Basic Concepts and Recent Issues. In S. Okazaki (ed.), *Handbook of Research in International Advertising*. Cheltenham: Edward Elgar, pp. 252–76.
- 69 Henseler, J., C. M. Ringle, and M. Sarstedt. 2015. A New Criterion for Assessing Discriminant Validity in Variance-Based Structural Equation Modeling. *Journal of the Academy of Marketing Science* 43: 115–35.
- 70 Henseler, J., C. M. Ringle, and M. Sarstedt. 2016. Testing Measurement Invariance of Composites Using Partial Least Squares. *International Marketing Review* 33: 405–31.
- 71 Henseler, J., C. M. Ringle, and R. R. Sinkovics. 2009. The Use of Partial Least Squares Path Modeling in International Marketing. In R. R. Sinkovics and P. N. Ghauri (eds.), *Advances in International Marketing*. Vol. 20. Bingley: Emerald, pp. 277–320.
- 72 Henseler, J., and M. Sarstedt. 2013. Goodness-of-Fit Indices for Partial Least Squares Path Modeling. *Computational Statistics* 28: 565–80.
- 73 Hock, C., C. M. Ringle, and M. Sarstedt. 2010. Management of Multi-Purpose Stadiums: Importance and Performance Measurement of Service Interfaces. *International Journal of Services Technology and Management* 14: 188–207.
- 74 Houston, M. B. 2004. Assessing the Validity of Secondary Data Proxies for Marketing Constructs. *Journal of Business Research* 57: 154–61.
- 75 Hu, L. T., and P. M. Bentler. 1998. Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification. *Psychological Methods* 3: 424–53.
- 76 Hui, B. S., and H. O. A. Wold. 1982. Consistency and Consistency at Large of Partial Least Squares Estimates. In H. O. A. Wold and K. G. Jöreskog (eds.), *Systems Under Indirect Observation, Part II*. Amsterdam: North-Holland, pp. 119–30.
- 77 Jöreskog, K. G. 1971. Simultaneous Factor Analysis in Several Populations. *Psychometrika* 36: 409–26.
- 78 Jöreskog, K. G. 1973. A General Method for Estimating a Linear Structural Equation System. In: A. S. Goldberger and O. D. Duncan (eds.), *Structural Equation Models in the Social Sciences*. New York: Seminar Press, pp. 255–84.
- 79 Jöreskog, K. G., and H. O. A. Wold. 1982. The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects. In H. O. A. Wold and K. G. Jöreskog (eds.), *Systems Under Indirect Observation, Part I*. Amsterdam: North-Holland, pp. 263–70.
- 80 Kaufmann, L., and J. Gaekler. 2015. A Structured Review of Partial Least Squares in Supply Chain Management Research. *Journal of Purchasing and Supply Management* 21: 259–72.
- 81 Kock, N., and P. Hadaya. 2017. Minimum Sample Size Estimation in PLS-SEM: The Inverse Square Root and Gamma-Exponential Methods. *Information Systems Journal*, doi. 10.1111/isj.12131.
- 82 Lee, L., S. Petter, D. Fayard, and S. Robinson. 2011. On the Use of Partial Least Squares Path Modeling in Accounting Research. *International Journal of Accounting Information Systems* 12: 305–28.
- 83 Lei, P.-W., and Q. Wu. 2012. Estimation in Structural Equation Modeling. In R. H. Hoyle (ed.), *Handbook of Structural Equation Modeling*. New York: Guilford Press, pp. 164–79.
- 84 Lohmöller, J.-B. 1989. *Latent Variable Path Modeling with Partial Least Squares*. Heidelberg: Physica.
- 85 MacCallum, R. C., and M. W. Brown. 1993. The Use of Causal Indicators in Covariance Structure Models: Some Practical Issues. *Psychological Bulletin* 114: 533–41.
- 86 Marcoulides, G. A., and W. W. Chin. 2013. You Write, but Others Read: Common Methodological Misunderstandings in PLS and Related Methods. In: H. Abdi, W. W. Chin, Vinzi V. Esposito, G. Russolillo, and L. Trinchera (eds.), *New Perspectives in Partial Least Squares and Related Methods*, Vol 56. Springer

- Proceedings in Mathematics & Statistics. New York: Springer, pp. 31–64.
- 87 Marcoulides, G. A., W. W. Chin, and C. Saunders. 2009. Foreword: A Critical Look at Partial Least Squares Modeling. *MIS Quarterly* 33: 171–5.
  - 88 Marcoulides, G. A., W. W. Chin, and C. Saunders (2012). When Imprecise Statistical Statements Become Problematic: A Response to Goodhue, Lewis, and Thompson. *MIS Quarterly* 36: 717–28.
  - 89 Marcoulides, G. A., and C. Saunders. 2006. PLS: A Silver Bullet? *MIS Quarterly* 30: III–IX.
  - 90 Mateos-Aparicio, G. 2011. Partial Least Squares (PLS) Methods: Origins, Evolution, and Application to Social Sciences. *Communications in Statistics—Theory and Methods* 40: 2305–17.
  - 91 Matthews, L., M. Sarstedt, J. F. Hair, and C. Ringle. 2016. Identifying and Treating Unobserved Heterogeneity with FIMIX-PLS: Part II—A Case Study. *European Business Review* 28: 208–24.
  - 92 Matthews, L. 2018. Applying Multi-Group Analysis in PLS-SEM: A Step-by-Step Process. In Hengky Latan and Richard Noonan (eds.), *Recent Developments on Partial Least Squares Structural Equation Modeling*. Berlin: Springer.
  - 93 McDonald, R. P. 1996. Path Analysis with Composite Variables. *Multivariate Behavioral Research* 31: 239–70.
  - 94 Nitzl, C. 2016. The Use of Partial Least Squares Structural Equation Modelling (PLS-SEM) in Management Accounting Research: Directions for Future Theory Development. *Journal of Accounting Literature* 37: 19–35.
  - 95 Nitzl, C., and W. W. Chin. 2017. The Case of Partial Least Squares (PLS) Path Modeling in Managerial Accounting. *Journal of Management Control* 28: 137–56.
  - 96 Nitzl, C., J. L. Roldán, and G. Cepeda Carrión. 2016. Mediation Analysis in Partial Least Squares Path Modeling: Helping Researchers Discuss More Sophisticated Models. *Industrial Management & Data Systems* 119: 1849–64.
  - 97 Noonan, R., and H. O. A. Wold. 1982. PLS Path Modeling with Indirectly Observed Variables: A Comparison of Alternative Estimates for the Latent Variable. In: K. G. Jöreskog and H. O. A. Wold (eds.), *Systems Under Indirect Observations: Part II*. Amsterdam: North-Holland, pp. 75–94.
  - 98 Nunnally, J. C., and I. Bernstein. 1994. *Psychometric Theory*, 3rd edn. New York: McGraw Hill.
  - 99 Olsson, U. H., T. Foss, S. V. Troye, and R. D. Howell. 2000. The Performance of ML, GLS, and WLS Estimation in Structural Equation Modeling Under Conditions of Misspecification and Nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal* 7: 557–95.
  - 100 Patel, V. K., S. Manley, O. C. Ferrell, J. F. Hair, and T. M. Pieper. 2017. Is Stakeholder Theory Relevant for European Firms? *European Management Journal* 36: 650–60.
  - 101 Peng, D. X., and F. Lai. 2012. Using Partial Least Squares in Operations Management Research: A Practical Guideline and Summary of Past Research. *Journal of Operations Management* 30: 467–80.
  - 102 Preacher, K. J., and A. F. Hayes. 2008. Asymptotic and Resampling Strategies for Assessing and Comparing Indirect Effects in Multiple Mediator Models. *Behavioral Research Methods* 40: 879–91.
  - 103 Raithel, S., M. Sarstedt, S. Scharf, and M. Schwaiger. 2012. On the Value Relevance of Customer Satisfaction. Multiple Drivers and Multiple Markets. *Journal of the Academy of Marketing Science* 40: 509–25.
  - 104 Raithel, S., and M. Schwaiger. 2015. The Effects of Corporate Reputation Perceptions of the General Public on Shareholder Value. *Strategic Management Journal* 36: 945–56.
  - 105 Reinartz, W. J., M. Haenlein, and J. Henseler. 2009. An Empirical Comparison of the Efficacy of Covariance-Based and Variance-Based SEM. *International Journal of Research in Marketing* 26: 332–44.
  - 106 Richter, N. F., R. R. Sinkovics, C. M. Ringle, and C. Schlägel. 2016. A Critical Look at the Use of SEM in International Business Research. *International Marketing Review* 33: 376–404.
  - 107 Rigdon, E. E. 2012. Rethinking Partial Least Squares Path Modeling: In Praise of Simple Methods. *Long Range Planning* 45: 341–58.
  - 108 Rigdon, E. E. 2013. Partial Least Squares Path Modeling. In: G. R. Hancock, and R. O. Mueller (eds.), *Structural Equation Modeling: A Second Course*, 2nd edn. Charlotte, NC: Information Age Publishing, pp. 81–116.
  - 109 Rigdon, E. E. 2013. Partial Least Squares Path Modeling. In: G. R. Hancock, and R. O. Mueller (eds.), *Structural Equation Modeling: A Second Course*, Vol. 1. Charlotte, NC: Information Age Publishing.
  - 110 Rigdon, E. E. 2014. Comment on “Improper Use of Endogenous Formative Variables”. *Journal of Business Research* 67: 2800–2.
  - 111 Rigdon, E. E. 2014. Rethinking Partial Least Squares Path Modeling: Breaking Chains and Forging Ahead. *Long Range Planning* 47: 161–7.
  - 112 Rigdon, E. E. 2016. Choosing PLS Path Modeling as Analytical Method in European Management Research: A Realist Perspective. *European Management Journal* 34: 598–605.
  - 113 Rigdon, E. E., J.-M. Becker, A. Rai, C. M. Ringle, A. Diamantopoulos, E. Karahanna, D. Straub, and T. K. Dijkstra. 2014. Conflating Antecedents and Formative Indicators: A Comment on Aguirre-Urreta and Marakas. *Information Systems Research* 25:780–4.
  - 114 Rigdon, E. E., M. Sarstedt, and C. M. Ringle. 2017. On Comparing Results from CB-SEM And PLS-SEM. Five Perspectives and Five Recommendations.

- Marketing ZFP—Journal of Research and Management*, forthcoming.
- 115 Ringle, C. M., M. Sarstedt, and R. Schlittgen. 2014. Genetic Algorithm Segmentation in Partial Least Squares Structural Equation Modeling. *OR Spectrum* 36: 251–76.
- 116 Ringle, C. M., M. Sarstedt, R. Schlittgen, and C. R. Taylor. 2013. PLS Path Modeling and Evolutionary Segmentation. *Journal of Business Research* 66: 1318–24.
- 117 Ringle, C. M., M. Sarstedt, and D. W. Straub. 2012. A Critical Look at the Use of PLS-SEM in MIS Quarterly. *MIS Quarterly* 36: iii–xiv.
- 118 Ringle, C. M., S. Wende, and A. Will. 2005. SmartPLS 2.0. Hamburg: [www.smartpls.de](http://www.smartpls.de).
- 119 Ringle, C. M., S. Becker, and J. Becker. 2015. SmartPLS 3. Boenningstedt, Germany: [www.smartpls.de](http://www.smartpls.de).
- 120 Roldán, J. L., and M. J. Sánchez-Franco. 2012. Variance-Based Structural Equation Modeling: Guidelines for Using Partial Least Squares in Information Systems Research. In M. Mora, O. Gelman, A. L. Steenkamp, and M. Raisinghani (eds.), *Research Methodologies, Innovations and Philosophies in Software Systems Engineering and Information Systems*. Hershey, PA: IGI Global, pp. 193–221.
- 121 Rönkkö, M., and J. Evermann. 2013. A Critical Examination of Common Beliefs About Partial Least Squares Path Modeling. *Organizational Research Methods* 16: 425–48.
- 122 Sarstedt, M., J.-M. Becker, C. M. Ringle, and M. Schwaiger. 2011. Uncovering and Treating Unobserved Heterogeneity with FIMIX-PLS: Which Model Selection Criterion Provides an Appropriate Number of Segments? *Schmalenbach Business Review* 63: 34–62.
- 123 Sarstedt, M., A. Diamantopoulos, T. Salzberger, and P. Baumgartner. 2016. Selecting Single Items to Measure Doubly-Concrete Constructs: A Cautionary Tale. *Journal of Business Research* 69: 3159–67.
- 124 Sarstedt, M., J. F. Hair, C. M. Ringle, K. O. Thiele, and S. P. Gudergan. 2016. Estimation Issues with PLS and CBSEM: Where the Bias Lies! *Journal of Business Research* 69: 3998–4010.
- 125 Sarstedt, M., J. Henseler, and C. M. Ringle. 2011. Multi-Group Analysis in Partial Least Squares (PLS) Path Modeling: Alternative Methods and Empirical Results. In: M. Sarstedt, M. Schwaiger, and C. R. Taylor (eds.), *Advances in International Marketing*, Vol. 22. Bingley: Emerald, pp. 195–218.
- 126 Sarstedt, M., and E. A. Mooi. 2014. *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*. Heidelberg: Springer.
- 127 Sarstedt, M., C. M. Ringle, D. Smith, R. Reams, and J. F. Hair. 2014. Partial Least Squares Structural Equation Modeling (PLS-SEM): A Useful Tool for Family Business Researchers. *Journal of Family Business Strategy* 5: 105–15.
- 128 Sarstedt, M., P. Wilczynski, and T. C. Melewar. 2013. Measuring Reputation in Global Markets—A Comparison of Reputation Measures' Convergent and Criterion Validities. *Journal of World Business* 48: 329–39.
- 129 Sarstedt, M., C. M. Ringle, and J. F. Hair. 2018. Partial Least Squares Structural Equation Modeling. *Handbook of Market Research*, C. Homburg, M. Klarman, and A. Vomberg (eds.). Berlin: Springer.
- 130 Sarstedt, M., C. M. Ringle, and J. F. Hair. 2018. Treating Unobserved Heterogeneity in PLS-SEM: A Multi-Method Approach. In Hengky Latan and Richard Noonan (eds.), *Recent Developments on Partial Least Squares Structural Equation Modeling*. Berlin: Springer.
- 131 Sarstedt, M., C. M. Ringle, and S. M. Gudergan. 2016. Guidelines for Treating Unobserved Heterogeneity in Tourism Research: A Comment on Marques and Reis. *Annals of Tourism Research* 57: 279–84.
- 132 Schlittgen, R., C. M. Ringle, M. Sarstedt, and J.-M. Becker. 2016. Segmentation of PLS Path Models by Iterative Reweighted Regressions. *Journal of Business Research* 69: 4583–92.
- 133 Schloderer, M. P., M. Sarstedt, and C. M. Ringle. 2014. The Relevance of Reputation in the Nonprofit Sector: The Moderating Effect of Socio-Demographic Characteristics. *International Journal of Nonprofit and Voluntary Sector Marketing* 19: 110–26.
- 134 Schwaiger, M. 2004. Components and Parameters of Corporate Reputation: An Empirical Study. *Schmalenbach Business Review* 56: 46–71.
- 135 Shah R., and S. M. Goldstein. 2006. Use of Structural Equation Modeling in Operations Management Research: Looking Back and Forward. *Journal of Operations Management* 24: 148–69.
- 136 Shmueli, G., and S. Ray. 2010. To Explain or to Predict? *Statistical Science* 25: 289–310.
- 137 Shmueli, G., S. Ray, J. M. Velasquez Estrada, and S. B. Chatla. 2016. The Elephant in the Room: Evaluating the Predictive Performance of PLS Models. *Journal of Business Research* 69: 4552–64.
- 138 Sosik, J. J., S. S. Kahai, and M. J. Piovoso. 2009. Silver Bullet or Voodoo Statistics? A Primer for Using the Partial Least Squares Data Analytic Technique in Group and Organization Research. *Group & Organization Management* 34: 5–36.
- 139 Stieglitz, S., L. Dang-Xuan, A. Bruns, and C. Neuberger. 2014. Social Media Analytics. *WIRTSCHAFTSINFORMATIK* 56: 101–9.
- 140 Stone, M. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society* 36: 111–47.
- 141 Tenenhaus, M., Vinzi V. Esposito, Y.-M. Chatelin, and C. Lauro. 2005. PLS Path Modeling. *Computational Statistics and Data Analysis* 48: 159–205.
- 142 Voorhees, C., M. Brady, R. Calantone, and E. Ramirez. 2016. Discriminant Validity Testing in Marketing: An Analysis, Causes for Concern, and Proposed Remedies.

- Journal of the Academy of Marketing Science* 44: 119–34.
- 143 Westland, J. C. 2015. Partial Least Squares Path Analysis. In: *Structural Equation Models: From Paths to Networks*. Cham: Springer, pp. 23–46.
- 144 Willaby, H. W., D. S. J. Costa, B. D. Burns, C. MacCann, and R. D. Roberts. 2015. Testing Complex Models With Small Sample Sizes: A Historical Overview and Empirical Demonstration of What Partial Least Squares (PLS) Can Offer Differential Psychology. *Personality and Individual Differences* 84: 73–8.
- 145 Wickens, M. R. 1972. A Note on the Use of Proxy Variables. *Econometrica* 40: 759–61.
- 146 Wold, H. O. A. 1975. Path Models with Latent Variables: The NIPALS Approach. In: H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi (eds.), *Quantitative Sociology: International Perspectives on Mathematical and Statistical Modeling*. New York: Academic Press, pp. 307–57.
- 147 Wold, H. O. A. 1980. Model Construction and Evaluation when Theoretical Knowledge is Scarce: Theory and Application of PLS. In: J. Kmenta and J. B. Ramsey (eds.), *Evaluation of Econometric Models*. New York: Academic Press, pp. 47–74.
- 148 Wold, H. O. A. 1982. Soft Modeling: The Basic Design and Some Extensions. In: K. G. Jöreskog and H. O. A. Wold (eds.), *Systems Under Indirect Observations: Part II*. Amsterdam: North-Holland, pp. 1–54.
- 149 Wold, H. O. A. 1985. Partial Least Squares. In: S. Kotz and N. L. Johnson (eds.), *Encyclopedia of Statistical Sciences Vol. 6*. New York: Wiley, pp. 581–91.
- 150 Wright, S. 1921. Correlation and Causation. *Journal of Agricultural Research* 20: 557–85.
- 151 Zhao, X., J. G. Lynch, Q. Chen. 2010. Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research* 37: 197–206.

# Index

Page numbers followed by f indicate figures; those followed by a t indicate tables

## A

a priori criterion, 122, 141, 168  
a priori test *see* planned comparison  
 $1-R^2$  ratio, 122  
absolute Euclidean distance *see* squared Euclidean distance  
absolute fit indices, 604, 636–638  
    goodness-of-fit index, 637  
    normed chi-square, 638  
    other absolute indices, 638  
root mean square error of approximation, 637  
root mean square residual, 637–638  
standardized root mean residual, 637–638  
 $X^2$  statistic, 636–637  
academic research, 5–6  
accuracy, 549, 566  
additional samples, 321  
adjusted coefficient of determination  
    (adjusted  $R^2$ ), 260, 300, 760  
adjusted goodness of fit index (AGFI), 639  
agglomerative methods, 190, 191, 196t, 212–215  
aggregated data, 556  
algorithmic models, 2, 7–8, 8f  
all-available approach, 46, 67, 604, 630  
all-possible-subsets regression, 260, 295, 352, 353t  
alpha ( $\alpha$ ), 372, 404  
    decreasing, 404–405  
    increasing, 404  
    *see also* Type I error  
analysis plan development, 33  
analysis sample, 472, 579  
analysis of variance (ANOVA), 333, 377  
    assumptions, 398–401  
    definition, 372  
    other designs, counterparts of, 397  
analyst, impact of Big Data, 6  
analytical sample, 549  
analytics, impact of Big Data, 6  
ANCOVA, 394–396  
anomalies *see* outliers  
ANOVA *see* analysis of variance  
anti-image correlation matrix, 122, 136  
association, new measures of, 55  
assumption of the strongly ignorable treatment assignment, 372, 423–424  
ATE, 372, 427  
atomatic fallacy, 260, 325  
ATT, 372, 428  
attitudes toward coworkers (AC), 682, 713, 784  
AUC (area under the curve), 549, 568–569  
auxiliary variables, 70  
average error variance indicators, 632  
average linkage, 190, 216  
average variance extracted (AVE), 659, 676, 760  
    *see also* communality

## B

backward elimination, 260, 297  
badness-of-fit, 604, 638  
balance, 372, 424  
balanced design, 372, 391  
Bartlett test of sphericity, 122, 136, 168  
basic residual plot, 288  
Bayesian SEM, 753–754, 754f

## Bayesian statistics, 727

beta ( $\beta$ ) *see* Type II error  
beta coefficient, 260, 310–311, 310f  
beta (standardized) weights, 318  
beta weights, 351–352  
between-construct error covariance, 659, 738  
between-group constraints, 727  
bias-corrected and accelerated (BCa) bootstrap confidence intervals, 760, 789t  
Big Data, 2  
    challenge of research efforts, 49–50  
    cluster analysis, 230  
    definition, 4–5  
    impacts on analytics and the analyst, 6  
    impacts on organization decisions and academic research, 5–6  
    moving forward, 7  
    multiple regression in era of, 265–266  
    problems in using, 6–7  
    rise, 4–7  
    summary, 51  
    variability and value, 5  
variety, 5  
velocity, 5  
veracity, 5  
volume, 4–5  
binary dependent variable, representation of, 553–554, 554f  
    assigning values, 553–554  
    unique nature of, 554  
binning, 46, 103–104  
bivariate correlations, 316–318, 350–351  
bivariate detection, 88–89, 91, 92t, 93  
bivariate partial correlation, 2, 26  
bivariate profiling  
    group differences, 53–54, 54f  
    relationship between variables, 52–53  
    scatterplot matrix of selected metric variables, 53f  
blindfolding, 760, 780  
blocking factor, 372, 390, 392, 425  
Bonferroni inequality, 372, 415–416  
bootstrap cases, 760  
bootstrap confidence interval, 760, 788–789  
bootstrap samples, 760  
bootstrapping, 2, 727, 746, 760, 788  
boundary point, 190, 225  
boxplot, 46, 53, 435f, 441f, 447f  
Box's M test, 372, 399–400, 472, 489  
C

## C/BAR, 550, 572

canonical correlation, 28  
cardinality, 46, 103  
case substitution, 68  
casewise diagnostics, 501–502, 519–520, 534  
    analyzing misclassified cases, 501–502  
    empirical measures, 502  
    graphical portrayal, 502  
    profiling on independent variables, 501  
    case-specific classification information, 534–535  
    in logistic regression, 571–572  
    influential measures, 572

## residuals, 571–572

in logistic regression illustrative example, 590–592t, 592  
misclassification of individual cases, 501  
categorical data, factor analysis of, 144–145  
categorical variable, 474, 551  
    *see also* nonmetric variable  
causal effects, 383, 421–422  
causal inference, 2, 9, 383, 421–422, 428–429, 604, 615  
causal model, 700  
causation, 604  
    establishing, 615–618  
    covariation, 616  
    nonspurious covariance, 616–618  
    sequence, 616  
    theoretical support, 618  
CB-SEM *see* covariance-based structural equation modeling  
cell means, 420f  
censored data, 46, 60  
    in ignorable missing data, 46  
centering, 47, 102, 190, 233t  
centroid, 472, 475  
centroid method, 190, 216  
chi-square difference, 550, 572, 727, 738  
chi-square test, 584  
chi-square ( $\chi^2$ ), 604  
chi-square ( $\chi^2$ ) difference statistic ( $\Delta\chi^2$ ), 604, 635, 645  
chi-square-based measure, 570  
city-block distance, 190  
classification  
    assessing accuracy, 515–518, 532–534  
    discriminant analysis assumptions, 489  
    evaluating accuracy, 518  
    two-group discriminant analysis, 516t  
classification accuracy  
    in overall model fit, 586–590  
    concordance, 590  
    cut-off value, determining appropriate, 586–587, 588t  
    Hosmer and Lemeshow test, 589–590  
    PPV and NPV, 587  
    predictive accuracy, overall, 587  
    ROC curve, 589  
    sensitivity and specificity, 587  
    summary, 590  
classification function, 472, 495  
classification function coefficients, 515  
classification matrix, 472, 498t, 550  
    constructing, 497–498, 516  
    individual observations, 495  
    predictive accuracy, 565–566  
    reasons for development, 494–495  
classification probability, 534–535  
cluster analysis, 122, 129  
    assumptions, 211–212, 235  
    multicollinearity, impact of, 211, 235  
    rules of thumb, 212  
    sample representativeness, 211, 235  
    structure exists, 211  
    between-and within-cluster variation, 214f  
chapter preview, 189

- conceptual development, 192–193  
 data reduction, 192–193  
 hypothesis generation, 193  
 conceptual support in, necessity of, 193
- D**
- DAG (directed acyclic graphs), 373, 426  
 data aggregation, 556  
 data collection, ignorable missing data, 60  
 data examination  
     challenge of Big Data research efforts, 49–51  
     chapter preview, 45–46  
     data transformations, 100–105  
     guidelines, 104–105  
     related to interpretation, 101–102  
     related to simplification, 103–104  
     related to specific relationship types, 102–103  
     related to statistical properties, 101  
     illustration of testing assumptions underlying  
         multivariate analysis, 105–112  
     homoscedasticity, 108  
     linearity, 108, 112  
     normality, 105, 108  
     summary, 112  
     incorporating nonmetric data with dummy  
         variables, 112–113  
     introduction, 49  
     key terms, 46–48  
     missing data, 56–85  
     four-step process for identifying and applying  
         remedies, 58–72  
     illustration of diagnosis with four-step  
         process, 72–85  
     impact, 56–57  
     recent development, 57  
     simple example, 57–58  
     outliers, 85–93  
     classifying, 87–88  
     detecting and handling, 88–93  
     impacts, 86  
     two different contexts for defining, 85–86  
     preliminary, 51  
     bivariate profiling, 52–54  
     multivariate profiles, 54–55  
     new measures of association, 55  
     summary, 55  
     univariate profiling, 51–52  
     summary, 114–115  
     testing assumptions of multivariate analysis,  
         93–100  
     assessing individual variables *vs.* the variate,  
         93–94  
     four important statistical assumptions, 94–100  
 data management, 47, 50  
 data mining models, 2, 7–8, 8f  
     distinguishing between models, 8–9  
     *vs.* statistical models, 7–9  
 data models, 7, 8f  
     *see also* statistical models  
 data in multivariate analysis and interpretation,  
     30–31  
 data quality, 47, 50–51  
 data reduction, 130–131  
     additional uses of exploratory factor analysis  
         results, 159–164  
     computing factor scores, 163  
     creating summated scales, 160–163  
     selecting among the three methods, 164  
     selecting surrogate variables for subsequent  
         analysis, 160  
     cluster analysis, 192–193  
     selecting between methods, 180–181  
     selecting in common factor analysis, 139  
 data summarization, 129–130  
     with interpretation, 130  
     without interpretation, 130  
 data transformations, 47, 100–105  
     data derived, 101  
     general guidelines, 104–105
- interpretation, 101  
 centering, 102  
 standardization, 101–102  
 in multiple regression, 281  
 related to simplification, 103  
 binning, 103–104  
 smoothing, 104  
 rules of thumb, 105  
 specific relationship types, 102–103  
 statistical properties, 101  
 achieving linearity, 101, 102f  
 achieving normality and homoscedasticity,  
     101  
 summary, 115  
 theoretical, 100–101
- data values, 194f  
 databases, 34  
     data warehouse classification variables, 36  
     other, 37  
     perceptions of, 36–37  
     primary, 35–37  
     purchase of outcomes, 37
- dCor, 47  
 decision flowchart, 34  
 degrees of freedom (df), 261, 280, 604–605, 673,  
     707, 761  
     as measure of generalizability, 280
- deleted residual, 261  
 dendrogram, 190, 191, 197, 238, 239f  
     hierarchical clustering, 214f  
     using, 238–239
- density, 190  
 density-based approach, 190, 225  
     noise point, 191
- density-reachable point *see* boundary point  
 dependence models  
     general linear model *vs.* generalized linear  
         model, 17–18  
     managing, 17–18, 17f, 21  
     single *vs.* multiple equation, 17
- dependence relationship, 605, 607, 764  
     examining hypothesized, 711
- dependence technique, 2, 21, 24, 613, 615f  
 dependent measures, selecting, 386–387  
 dependent variable, 2, 21, 261, 265  
     assumptions in MANOVA, 434  
     differences among combination of, 385  
     effects of multicollinearity on power, 406–407  
     linearity and multicollinearity among, 401  
     in logistic regression, 581  
     transforming, 558–559  
     sample size, 555–556  
     selection, 275–278  
     measurement error, 276  
     specification error, 276–278  
     strong theory, 276
- dependent variate, effects of MANOVA on,  
     412–414  
     example of interpreting interactions, 413  
     disordinal interaction, 413  
     no interaction, 413  
     ordinal interaction, 413  
     interaction terms, significance of, 412  
     main effects, 411–412  
     significant interactions, types of, 412–413  
     disordinal, 413  
     ordinal, 412–413
- desirability loadings, vector plot of, 505–506  
 deviance difference, 550, 572  
 deviance residuals, 572  
 dfbeta, 261, 305, 550, 572  
 DFFIT, 261, 306  
     *see also* SDFFIT
- diameter method *see* complete-linkage method  
 dichotomization, 47, 104  
 dimensional reduction, 2, 15, 317  
 dimensions, comparability of, 90  
 direct effect, 727, 745  
 directed acrylic graph (DAG), 2, 9
- directionality of relationship, 574–575  
     example of interpretation, 575  
     interpreting direction of exponentiated  
         coefficients, 575  
     interpreting direction of original coefficients,  
         575
- discriminant analysis, 380–382, 418, 437, 442–443  
 assumptions, 488–490  
 estimation and classification, impacts on, 489  
 interpretation, impacts on, 489–490  
     in three-group illustrative example, 524–525  
     in two-group illustrative example, 509  
     casewise diagnostics, 501–502  
     individual cases, misclassification, 501  
     misclassified cases, analyzing, 501–502  
     decision process, 494–507  
     stage 1: objectives, 484–485  
     stage 2: research design, 485–488  
     stage 3: assumptions, 488–490  
     stage 4: estimation and assessing overall fit,  
         490–502  
     stage 5: interpretation of results, 503–506  
     stage 6: validation of results, 506–507  
 estimation, 490–502, 491f  
 selecting method, 491–492  
 statistical significance, 492–493  
     in three-group illustrative example, 525–537  
     in two-group illustrative example, 508–520  
 explained, 474–476  
 hypothetical example, 476–483  
 geometric representation of two-group  
     discriminant function, 480–481
- three-group discriminant analysis (switching  
     intentions), 481–483
- two-group discriminant analysis (purchasers  
     *vs.* non-purchasers), 476–480
- interpretation of results, 503–506  
 choosing interpretive method, 506  
 discriminant loadings, 503  
 discriminant weights, 503  
 graphical display of discriminant scores and  
     loadings, 505–506  
 partial F values, 504  
 potency index, 504–505  
 rotation of discriminant functions, 504  
 two or more functions, 504–506  
 objectives  
     classification purposes, 485  
     descriptive profile analysis, 485  
     in three-group illustrative example, 524  
     in two-group illustrative example, 508  
     overall model fit, assessing  
         calculating discriminant Z scores, 493–494  
     group differences, evaluating, 494  
     group membership prediction accuracy,  
         assessing, 494–501
- rules of thumb, 502  
     in three-group illustrative example, 525–537  
     in two-group illustrative example, 509–520  
 research design  
     dependent and independent variables,  
         selecting, 485–487, 508
- division of sample, 488, 508–509  
 rules of thumb, 490  
 sample size, 487–488, 508  
     in three-group illustrative example, 523–544  
     in two-group illustrative example, 508–509  
 similarities to other multivariate  
     techniques, 476
- summary, 544–546  
 three-group illustrative example, 523–544  
 managerial overview, 543–544  
 stage 1: objectives, 524  
 stage 2: research design, 524  
 stage 3: assumptions, 524–525  
 stage 4: estimation and assessing of overall  
     fit, 525–537
- stage 5: interpretation of analysis results,  
     537–542

- discriminant analysis (*continued*)  
 stage 6: validation of results, 542–543  
 two-group illustrative example, 508–523  
 managerial overview, 523  
 stage 1: objectives, 508  
 stage 2: research design, 508–509  
 stage 3: assumptions, 509  
 stage 4: estimation and assessing overall fit, 509  
 stage 5: interpretation of results, 520–522  
 stage 6: validation of results, 522–523  
 validation of results, 506–507  
 managerial overview, 523, 543–544  
 procedures, 506–507  
 profiling group differences, 507  
 in three-group illustrative example, 542–543  
 in two-group illustrative example, 522–523  
 discriminant coefficient *see* discriminant weight  
 discriminant function, 373, 379, 472–475,  
 478f, 479t  
 estimation of, 526–532  
 rotation of, 504  
 discriminant loadings, 473, 503  
 graphical display, 541–542  
 in three-group illustrative example, 537,  
 541–542  
 in two-group illustrative example, 515,  
 521–522  
 discriminant score *see* discriminant Z score  
 discriminant validity, 122, 659, 676–677, 688–689,  
 761, 775–776, 788–789  
 discriminant weight, 473, 503, 515, 521  
 discriminant Z score, 473, 475f, 534  
 graphical display of, 506–507  
 discriminating variables, 522  
 disordinal interaction, 373  
 divisive method, 190, 191, 213  
 dominance analysis, 319, 352  
 dummy variables, 2, 47, 112, 122, 132, 261  
 coding, 113  
 effects, 113  
 indicator, 113  
 concept, 112–113  
 incorporating nonmetric data with, 281–282  
 representing nonmetric variables with,  
 112f, 113f  
 using, 113
- E**  
 ecological fallacy, 261, 325  
 effect size, 2, 19, 373, 780  
 effects coding, 47, 113, 261, 282  
*see also* indicator coding  
 eigenvalue, 122, 141  
*see also* latent root  
 elasticity, 47, 103  
 EM, 47  
 empirical measures of impact, 539, 541  
 endogenous constructs, 605, 610, 714–715,  
 784–785  
*see also* endogenous latent variables  
 endogenous latent variables, 761  
*see also* endogenous constructs  
 entropy group, 190, 198  
 environmental perceptions (EP), 682, 713, 784  
 equal variance dispersion, 98  
 EQUIMAX, 122, 150  
 equivalent models, 619  
 error probabilities, statistical inference  
 relationship, 19f  
 error term, 761  
 constant variance of, 290  
 diagnosis, 290  
 remedies, 290  
 independence of, 291  
 normality of distribution, 291  
 error term invariance, 727, 739  
 error variance, 122  
 esprit de corps, 704
- estimated covariance matrix, 605, 623–625, 624t  
 estimation, 308–309  
 in confirmatory factor analysis, 673  
 in structural equation modeling, 633–634  
 estimation sample, 2, 32  
 Euclidean distance, 190  
 evaluation criteria, 761  
 exogenous constructs, 605, 610, 714, 784  
*see also* exogenous latent variables  
 exogenous latent variables, 761  
*see also* exogenous constructs  
 experiment, 373  
 experimental design, 373, 376  
 experimental research  
 types of variables, 389–390  
 additional relationships, 390  
 basic relationship, 389–390  
 choosing, 390  
 external influences, 390  
 experimentwide error rate, 373, 385  
 explained variance *see* coefficient of determination  
 explanation, 309–311  
 interpreting with regression coefficients,  
 309–310  
 standardizing regression coefficients  
 (beta coefficients), 310–311  
 exploratory, 761  
 exploratory analysis, 605  
 exploratory approach, 122, 125  
 exploratory factor analysis, 25–26, 660–661  
 additional uses, 159–164  
 computing factor scores, 163  
 creating summated scales, 160–163  
 rules of thumb, 164  
 selecting among the three methods, 164  
 selecting surrogate variables for subsequent analysis, 160  
 assessing appropriateness, 166t  
 chapter preview, 121  
 decision diagram, 128f, 137f  
 design rules of thumb, 134  
 explanation, 124–126  
 hypothetical example, 126, 127f  
 key terms, 122–124  
 summary, 136, 184–186  
 testing assumptions, rules of thumb, 138  
*see also* common factor analysis; factor analysis  
 exponentiated logistic coefficient, 550, 574,  
 575, 576f  
 external validity, 373  
 extraordinary event, 87  
 extraordinary observation, 87  
 extreme groups approach, 47, 104, 473
- F**  
 F ratio, 333  
 F effect size, 761  
 face validity, 659, 677  
*see also* content validity  
 factor, 123, 373, 376, 389, 392  
 factor analysis, 25, 127–184  
 assessing appropriateness, 169t  
 assumptions, 135–136  
 conceptual issues, 135  
 statistical issues, 135–136  
 choosing models and number of factors, rules  
 of thumb, 144  
 data reduction and additional uses of results,  
 159–164  
 computing factor scores, 163  
 creating summated scales, 160–163  
 selecting among three methods, 164  
 selecting surrogate variables for subsequent analysis, 160  
 decision process, 127  
 stage 1: objectives, 127–132  
 stage 2: designing, 132–134
- stage 3: assumptions, 135–136  
 stage 4: deriving factors and assessing overall fit, 136–145  
 stage 5: interpreting, 146–158  
 stage 6: validation of, 158–164  
 deriving factors and assessing overall fit, 136–146  
 alternatives to principal components and common factor analysis, 144–146  
 factor extraction method, selecting, 138–140  
 stopping rules: criteria for number of factors to extract, 140–144  
 designing, 132–134  
 correlations among variables or respondents, 133–134, 134f  
 rules of thumb, 134  
 sample size, 132–133  
 variable selection and measurement issues, 132  
 illustrative example, 165–184  
 stage 1: objectives, 165  
 stage 2: designing, 165  
 stage 3: assumptions, 165, 168  
 stage 4: deriving factors and assessing overall fit, 168, 170–171  
 stage 5: interpreting the factors, 171–176  
 stage 6: validation of principal components analysis, 176–177  
 stage 7: additional uses, 178–181  
 interpreting factors, 146–158  
 factor extraction, 147  
 factor matrix, 153–158  
 judging significance of factor loading, 151–153  
 rotation of factors, 147–151  
 rules of thumb, 158  
 step 1: estimate factor matrix, 146  
 step 2: factor rotation, 147  
 step 3: factor interpretation and respecification, 147  
 objectives, 127–132  
 data summarization vs. data reduction, 129–131  
 specifying unit of analysis, 127, 129  
 variable selection, 131  
 with other multivariate techniques, 131–132  
 specifying unit of analysis, 127, 129  
 using with other multivariate techniques, 131–132  
 validation, 158–159  
 assessing factor structure stability, 159  
 detecting influential observations, 159  
 replication or confirmation perspective, 158–159  
*see also* common factor analysis; exploratory factor analysis  
 factor correlation matrix, 177t  
 factor covariance invariance, 727, 739  
 factor extraction, 147  
 factor extraction method, selecting, 138–140  
 common factor analysis vs principal component analysis, 139–140  
 partitioning variance of a variable, 138–139  
 common vs unique variance, 138  
 unique variance composed of specific and error variance, 138–139  
 factor indeterminacy, 123, 140  
 factor interpretation, 147  
 summary, 176  
 factor loadings, 123, 146  
 assessing, rules of thumb, 153  
 comparisons between rotated and unrotated, 149t  
 in confirmatory factor analysis, 675–676  
 interpretation of hypothetical matrix, 157t  
 matrix for unrotated factor matrix, 172–173  
 significance, judging of, 151–152, 152t  
 adjustments based on number of variables 152

- assessing statistical significance, 151–152  
ensuring practical significance, 151
- factor matrix, 123, 139f  
estimate, 146  
interpretation, example of, 156–158  
interpreting, 153–158  
step 1: examine loadings, 153–154, 157  
step 2: identify significant loading(s) for each variable, 154–155, 157  
step 3: assess communalities of the variables, 155, 157  
step 4: respecify factor model if needed, 156, 158  
step 5: label the factors, 156, 158
- factor matrix of loadings examining significant, identifying in unrotated factor matrix, 173  
for unrotated factor matrix, 172–173
- factor pattern matrix, 123, 153
- factor rotation, 123, 147–151  
oblique rotation methods, 150  
orthogonal rotation methods, 150  
orthogonal vs oblique rotation, 147–149  
selecting method, 150–151  
rules of thumb, 151
- factor score, 123, 179–180, 180t
- factor (score) indeterminacy, 761
- factor structure matrix, 123, 154
- factor structure stability, 159
- factorial design, 373, 391–394  
number of treatments, 393–394  
cells formed, 393  
interaction effects, creation of, 393–394  
selecting treatments, 392–393  
basic model, 392  
control as treatments, 392–393
- factors, 125  
naming, 175–176
- false negative, 550, 565
- false positive, 18, 550
- farthest-neighbour method, 190
- feedback loop, 700, 707
- field experiment, 373, 388
- fine tuning, 235
- first-order factor model, 727, 732, 734f
- Fisher's linear discriminant function, 473, 495
- fit indices, 635–642, 642t  
absolute, 636–638  
goodness-of-fit, 635–636  
guidelines for establishing acceptable and unacceptable fit, 641–642  
adjust index cut-off values based on model characteristics, 642  
compare models whenever possible, 642  
pursuit of better fit at expense of testing true model not a good trade-off, 642  
use multiple indices of differing types, 641–642  
incremental, 638–639  
parsimony, 639  
problems associated with using, 639–640  
unacceptable model specification to achieve fit, 641
- five-stage modeling strategy, 327–328
- fixed effect, 261, 326
- fixed parameter, 605, 633, 704
- forecasting, 309, 322
- formative measurement, 659, 729, 729f, 761  
operationalizing, 729–730  
rules of thumb, 733
- formative measurement model, 761  
assessing, 776–778  
contribution of indicator, 778  
convergent validity, 777  
indicator multicollinearity, 777  
relevance of indicators, 778  
statistical significance of indicator weights, 777–778  
rules of thumb, 779
- formative measurement theory, 669, 727, 730f, 731f
- formative measures *see* measurement models
- Fornell-Larcker criterion, 761, 788t
- forward addition, 261, 297
- free parameter, 633, 704
- full invariance, 727, 739
- full mediation, 727, 746
- functional relationship, 261, 274, 275f
- G**
- garbage in, garbage out phenomenon, 131
- general linear model (GLM), 2, 373, 403  
vs. generalized linear model (GLZ or GLIM), 17–18
- generalizability  
degrees of freedom, defining, 280  
degrees of freedom as measure, 280  
sample size and, 279–280
- generalized linear model (GLZ or GLIM), 2, 373  
vs general linear model (GLM), 17–18
- global null hypothesis, 585
- goodness-of-fit (GOF), 605, 715t  
basics, 635–636  
chi-square, 635  
degrees of freedom, 636  
statistical significance of  $\chi^2$ , 636  
detail on selected indices, 654–655  
in logistic regression model, 563–571  
model estimation fit, 563–564  
predictive accuracy, 565–571  
in structural equation modeling, 635–637
- goodness-of-fit index (GFI), 654
- graphical analyses, 95
- graphical diagnostics, 288
- graphical tests, equal variance dispersion, 98
- group centroids, 515
- group comparisons, six step process, 737–739  
stage 1: configural invariance, 738  
stage 2: metric invariance, 738  
stage 3: scalar invariance, 739  
stage 4: factor covariance invariance, 739  
stage 5: factor variance invariance, 739  
stage 6: error variance invariance, 739
- group differences, 442t, 532  
ANOVA vs. MANOVA, 377, 377f  
assessing, 509–510, 525–526  
by independent variable, 382–383  
evaluating, 494  
graphical display of group means of variate, 383f
- multivariate differences for assessing, 377–380
- K-group case: MANOVA, 379–380, 438–444
- two-group case: Hotelling's  $T^2$ , 378–379, 432–438
- overall, 382
- profiling, 507
- group membership, 534
- group membership prediction accuracy, assessing, 494–501, 532–533  
classification matrices, constructing, 497–498  
hit ratios, 498–500  
individual observations, classifying, 495  
matrices, 494–495, 497–498  
misclassification, costs of, 497  
prior probabilities, defining, 495–497  
statistically-based measures of classification accuracy relative to chance, 500–501
- group/cell sample size  
equal vs. unequal, 391  
minimum, 391  
number of dependent variables, 391  
recommended, 391
- H**
- hat matrix, 261, 305
- hat value *see* hat matrix
- Hausman test, 326
- HCM *see* hierarchical component model
- heat map, 47, 103
- heterogeneity, 190, 761  
change measures, 222–223, 223f  
percentage changes, 222  
statistical, 223
- variance, 222–223
- direct measures, 223, 224f
- comparative cluster heterogeneity, 223
- internal validation index, 223
- statistical significance of cluster variation, 223
- direct measures of, 242
- measuring, 198
- percentage changes in, 240–242
- statistical measures of change, 242
- heteroscedasticity, 97, 290, 345f  
remedies, 98–99  
scatterplot of relationship with homoscedasticity, 98f
- sources, 97
- skewed distribution of one or both variables, 97–98
- variable type, 97  
*see also* homoscedasticity
- heteroscedasticity-consistent standard errors, 290
- heterotrait-heteromethod correlation, 761
- heterotrait-monotrait ratio (HTMT), 761, 788, 789
- Heywood case, 659, 673
- hierarchical cluster analysis, 196t, 197f, 212–217, 214f, 235–245  
clustering algorithms, 215–217  
combination of methods, 220–221  
fine tuning, 235  
overview, 217  
partitioning, 235  
pros and cons, 219  
large samples, 219  
outliers, impact of, 219  
permanent combinations, 219  
similarity, measures of, 219  
simplicity, 219  
speed, 219  
step 1: selecting clustering algorithm, 236  
step 2: initial cluster results, 236–238  
step 3: respecified cluster results, 238–243  
step 4: profiling clustering variables, 244–245
- hierarchical clustering algorithm, 215–217  
average-linkage, 216  
centroid method, 216  
complete-linkage, 216  
selecting, 236  
single-linkage method, 191, 215  
Ward's method, 192, 216–217
- hierarchical component model (HCM), 761
- hierarchical procedures, 191
- higher-order component (HOC), 761
- higher-order factor models, 732–736  
decisions concerning use of, 736  
empirical concerns, 733–734  
rules of thumb, 737
- second-order measurement theories, using, 735–736  
summary, 755  
theoretical concerns, 734–735
- higher-order measurement models, 783
- higher-order model *see* hierarchical component model (HCM)
- histogram, 47, 51–52
- hit ratio, 495, 550, 566  
comparing to the standard, 500  
overall vs. group-specific, 500, 533–534  
standards of comparison, 498–500  
equal group sizes, 499  
holdout sample, 500  
sample size, 499–500  
unequal group sizes, 499
- HOC *see* higher-order component
- Hoeffding's D, 47
- holdout sample, 32, 473, 500, 506–507, 550, 579

holdout sample *see validation sample*  
 homogeneity of regression effect, 374, 396  
 homoscedasticity, 47, 97, 262, 343–344  
 assessment, 108  
 assumptions in MANOVA, 434, 440, 448  
 data transformations, 101  
 independent variable, 97  
 metric, 97  
 nonmetric, 97  
 moderation of distribution system, 453  
 scatterplot of relationship with  
 heteroscedasticity, 98f  
 tests, 98, 109t, 440t  
 graphical of equal variance dispersion, 98  
 statistical, 98  
 transformation of price elasticity, 110–111f  
*see also heteroscedasticity*  
 Hosmer and Lemeshow test, 550, 570, 570f,  
 589–590, 590t  
 hot deck, 67  
 hot deck imputation, 47  
 Hotelling's T<sup>2</sup>, 374, 378f, 379f, 403  
 HTMT *see heterotrait-monotrait ratio*  
 hypotheses testing, 475–476  
 hypothesis generation, 193

**I**

identification, 659, 666  
 issues in CFA, 671–673  
 avoiding identification problems, 671  
 incorrect indicator specification, 672  
 recognizing identification problems, 672  
 setting the scale of a construct, 672–673  
 sources and remedies, 672–673  
 three-indicator rule, 671–672  
 too few degrees of freedom, 673  
 ignorable cross-loading, 155  
 ignorable effects, 316  
 ignorable missing data, 47, 60  
 censored data, 46, 60  
 data collection, 60  
 sample of missing data, 60  
 illogical standardized parameters, 673  
 imputation, 47, 65–66, 605, 630  
 comparison of techniques for missing  
 data 71f  
 MAR missing data process, 69  
 maximum likelihood and EM, 69  
 maximum likelihood vs. multiple  
 imputation 70  
 multiple imputation, 69–70  
 MCAR by calculating replacement values, 68  
 mean substitution, 68  
 regression imputation, 68  
 MCAR using known replacement values, 67  
 case substitution, 68  
 hot or cold deck approach, 67  
 MCAR using only valid data, 66  
 complete case approach, 66  
 using all-available data, 66–67  
 overview of MCAR methods, 68–69  
 recap of missing value analysis  
 imputation is most logical course of  
 action, 84  
 imputed correlations differ across  
 techniques 84  
 missing data process is primarily MCAR,  
 83–84  
 regression parameter estimates for complete  
 case and six imputation methods, 84f  
 rules of thumb, 72  
 summary, 70, 72  
 in-sample predictive power *see coefficient of*  
*determination*  
 incremental fit indices, 605, 638–639  
 comparative fit index, 639  
 normed fit index, 638  
 relative non-centrality index, 639  
 Tucker Lewis index, 638–639

independence, 374, 399  
 independent variable, 2, 21, 262, 265, 360t,  
 538–539, 581t  
 for action, identifying, 289–290  
 calculating probabilities for specific value  
 of, 578  
 magnitude of relationship of metric  
 independent variable, 575–576  
 metric, 97  
 nonmetric, 97  
 in logistic regression, 556  
 profiling on, 501  
 relative importance, 317–320  
 selection, 275–278  
 measurement error, 276  
 specification error, 276–278  
 strong theory, 276  
 indicator coding, 47, 48, 113, 262, 281–282  
 reference category, 48  
*see also effects coding*  
 indicator loadings, 775  
 indicator multicollinearity, 777  
 indicator reliability, 761  
 indicator weights, statistical significance  
 of 777–778  
 indicators, 2, 14, 123, 605, 608, 610, 761  
 contribution of, 778  
 relevance of, 778  
*see also manifest variable*  
 indirect effect, 374, 396, 727, 745, 761  
 calculating, 408  
 mediated model, 747–748  
 relative strength of, 419–420  
 significance of, 408  
 individual constructs in confirmatory factor  
 analysis, defining, 682  
 individual groups, identifying differences  
 between, 415–417  
 multiple univariate tests adjusting for  
 experiment-wide error rate, 415–416  
 rules of thumb, 418  
 structured multigroup tests, 416–417  
 individual variables  
 assessing, 287–288  
 methods of diagnosis, 288  
 influential measures, 550, 572  
 influential observations, 159, 262, 302,  
 346f, 347t  
 deletion of, 347, 348t  
 identifying, 303–306  
 step 1: examining residuals and partial  
 regression plots, 304  
 step 2: identifying leverage points, 304–305  
 step 3: single-case diagnostics, 305–306  
 step 4: selecting, 306  
 impacts, 302–303  
 conflicting, 302–303  
 reinforcing, 302  
 shifting, 303  
 outliers, identifying, 345–348  
 patterns of, 303f  
 remedies, 307  
 types of, 302  
 leverage points, 302  
 outliers, 302  
 information value, 550, 578  
 inner model *see structural model*  
 instructional manipulation check, 374, 392  
 instrumental variable, 374, 390, 425  
 interaction effects, 374, 393–394, 394f, 402f, 449f  
 assessing, 448–449  
 interpreting, 315, 451  
 testing, 450, 451  
 interaction terms, impacts of MANOVA  
 412–414  
 interpreting interactions, example of, 413  
 significance of, 412  
 significant interactions, types of, 412–413  
 intercept ( $b_0$ ), 262, 267, 268

intercorrelation  
 overall measures, 135–136  
 variable-specific measures, 136  
 interdependence technique, 2, 21, 25, 613–614  
 overview, 119–120  
 internal validation index, 223  
 internal validity, 374  
 interobject similarity, 191  
 defining similarity, 233  
 measuring, 195  
 interpretational confounding, 700, 709  
 interval scales, 12  
 intervening effect *see indirect effect*  
 intraclass correlation (ICC), 262, 325–326  
 intrinsically multivariate questions, 385  
 inverse probability of treatment weighting  
 (IPTW), 374, 427  
 ipsatizing, 102  
 ipsatizing method, 48  
 item *see indicator*  
 items per construct, 665–668  
 items per construct and identification, 666–668  
 just-identified, 667  
 overidentified, 667  
 underidentified, 666

**J**

job satisfaction (JS), 682, 713, 783  
 just-identified, 659, 666f, 667

**K**

K group case, MANOVA, 379–380, 438–444  
 stage 1: objectives of MANOVA, 438–439  
 examining group profiles, 439  
 research questions, 439  
 stage 2: research design of MANOVA, 439  
 stage 3: assumptions in MANOVA, 439–440  
 homoscedasticity, 440  
 outliers, 440  
 stage 4: estimation of MANOVA model and  
 assessing overall fit, 440, 442–443  
 statistical significance testing, 440, 442–443  
 stage 5: interpretation of results, 443–444  
 main effect of  $X_i$ , assessing, 443  
 post hoc comparisons, making, 443–444  
 summary, 444  
 K-group analysis, 379–380, 415–416  
 k-means, 191  
 k-means algorithm, 218  
 Kaiser rule *see latent root criterion*  
 Kolmogorov-Smirnov test, 105  
 kurtosis, 48, 94

**L**

LASSO, 298  
 latent construct, 605, 608, 764  
 benefits of using, 608–609  
 statistical system, improving, 609  
 theoretical concepts, representing 608–609  
 distinguishing exogenous vs. endogenous  
 latent constructs, 610  
 latent factors, 607  
 latent growth models, 752–755  
 latent root, 141  
*see also eigenvalue*  
 latent root criterion, 123, 141, 170  
 latent variables, 761, 764  
*see also latent construct*  
 least squares, 262, 267  
 leptokurtic distribution, 94  
 level-1, 262, 323  
 level-2, 262, 323  
 leverage, 550, 572  
 leverage diagnostics, 307f  
 leverage plot, 346  
 leverage points, 262, 302  
 identifying, 304  
 likelihood value, 550  
 linear-log, 103

- linearity, 48, 99, 262, 343  
 among dependent variables, 401  
 assessment, 108, 112  
 data transformations, 101, 102f  
 identifying nonlinear relationships, 99  
 of the phenomenon, 288–290  
 remedies for nonlinearity, 99
- linearity of the logit, 556  
 link function, 373, 374  
 LISREL, 605, 607, 614  
 LISREL notation, 605, 628  
*listwise deletion* see complete case approach  
 Little's MCAR test, 65  
 loadings, 130, 786  
 LOC *see lower-order component*  
 log-linear, 103  
 log-log, 103  
 logistic coefficients, 559, 574f, 575  
 in logical regression illustrative example 593  
 direction of relationships, 593  
 magnitude of relationships, 593  
 predicting probabilities, 593
- logistic curve, 550, 554, 558f  
 logistic regression, 27, 473, 474, 548–596  
 assumptions, 556, 581  
 casewise diagnostics, 571–572  
 influential measures, 572  
 residuals, 571–572  
 chapter preview, 548–549  
 decision process, 552–579  
 stage 1: objectives, 552–553  
 stage 2: research design, 553–556  
 stage 3: assumptions, 556, 557  
 stage 4: estimation and assessing overall fit, 557–572  
 stage 5: interpretation of results, 572–579  
 stage 6: validation of results, 579  
 definition, 550  
 explained, 551–552  
 illustrative example, 580–596, 582t, 583–584t  
 managerial overview, 596  
 stage 1: objectives, 580  
 stage 2: research design, 580–581  
 stage 3: assumptions, 581  
 stage 4: estimation and assessing overall fit, 581–592  
 stage 5: interpretation of results, 592–596  
 stage 6: validation of results, 596  
 interpretation of results, 572–579  
 calculating probabilities for specific value of independent variable, 578–579  
 interpreting coefficients, 574–578  
 significance of coefficients, testing for, 573–574  
 key terms, 549–551  
 model estimation, 557–572  
 alternative models, 562  
 coefficients, estimating, 559  
 goodness-of-fit, assessing, 563–571  
 issues, 561  
 overview of assessing model fit, 571  
 transforming dependent variable, 558–559  
 transforming probability into odds and logit values, 559–561  
 objectives, 552–553  
 classification, 553  
 explanation, 552–553  
 research design, 553–556  
 aggregated data, use of, 556  
 representation of binary dependent variable, 553–554  
 sample size, 555–556  
 rules of thumb, 557  
 summary, 572, 596–597  
 validation of results, 579
- logit analysis *see logit regression*  
 logit transformation, 550
- logit values, 559–561  
 longitudinal data, 752  
 lower-order component (LOC), 762
- M**
- Mahalanobis distance ( $D^2$ ), 191, 262, 305, 534  
 main effect, 374, 389, 411–412, 414f  
 interpreting, 451–452  
 significance of mediated, 419  
 testing, 450, 451
- MANCOVA, 394–396  
 Manhattan distance *see city-block distance*  
 manifest variable, 605, 764  
*see also indicators*  
 manipulation check, 374, 392
- MANOVA, 26, 371–460  
 analysis, illustration of, 430–459  
 example 1: difference between two independent groups, 432–438  
 example 2: difference between K independent groups, 438–444  
 example 3: factorial design with two independent variables, 444–452  
 example 4: moderation and mediation, 452–459  
 research setting, 430–432  
 assumptions, 399–401  
 equality of variance-covariance matrices, 399–400  
 independence, 399  
 linearity and multicollinearity among dependent variables, 401  
 normality, 400–401  
 rules of thumb, 401  
 sensitivity to outliers, 401  
 chapter preview, 371–372  
 decision process, 383–384, 384f, 387, 402f  
 discriminant analysis difference, 380  
 estimation of model and assessing overall fit, 401–411  
 additional relationships: mediation and moderation, 407, 409  
 general linear model, 403  
 measures for significance testing, 403  
 rules of thumb, 410  
 statistical power of multivariate tests 403–407  
 experimental approaches *vs. other* multivariate methods, 376–377  
 extending univariate methods for assessing group differences, 377–380  
 multivariate procedures, 377–380  
 hypothetical illustration, 381–383  
 analysis design, 381  
 discriminant analysis differences, 381–382  
 forming variate and assessing differences, 382–383  
 interpretation of results, 410–421  
 covariates, evaluating, 410–411  
 dependent variate, assessing effects on, 411–414  
 individual groups, identifying differences between, 415–418  
 individual outcome variables, assessing significance for, 417–419  
 mediation and moderation, 419–421  
 key terms, 372–376  
 managerial overview of results, 459–460  
 objectives, 385–387  
 selecting dependent measures, 386–387  
 types of multivariate questions suitable for, 385–386  
 when to use, 385  
 re-emergence of experimentation, 376  
 research design, 387–398  
 covariates: ANCOVA and MANCOVA, 394–396
- factorial designs: two or more treatments, 391–394
- MANOVA counterparts of other ANOVA designs, 397  
 modeling other relationships between treatment and outcome, 396–397  
 rules of thumb, 398  
 sample size requirements: overall and by group, 381  
 selecting research approach, 389  
 special case of MANOVA: repeated measures, 397
- types of approaches, 387–389  
 variables in experimental research, 389–390  
 research setting, 430–432  
 research design, 431–432  
 research questions, 431  
 strategic objectives, 430–431  
 rules of thumb, 387  
 stage 1: objectives, 385–387  
 stage 2: research design issues, 387–398  
 stage 3: assumptions of ANOVA and MANOVA, 398–401
- stage 4: estimation of model and assessing overall fit, 401–409  
 stage 5: interpretation of results, 410–421  
 stage 6: validation of results, 421–428  
 summary, 430, 459–463  
 validation of results, 421–428  
 advanced issues: causal inference in nonrandomized situations, 421–422, 428–429  
 causality in social and behavioural sciences, 422  
 counterfactuals in non-experimental research designs, 423–424  
 extension to MANOVA, 428  
 overview, 428  
 potential outcome approach, 423  
 propensity score models, 424–428  
 matching, 374, 427  
 maximum chance criterion, 473, 518, 533, 550  
 maximum likelihood, 70  
 maximum likelihood estimation (MLE) 605, 632  
 maximum likelihood procedure, 550  
 maximum number of iterations, 762  
 mean substitution, 48, 68  
 measure of sampling adequacy (MSA), 123, 136, 168  
 measured variable, 605  
 measurement, 762  
*measurement equivalence* *see measurement invariance*  
 measurement error, 123, 262, 276, 605, 609, 762, 764, 786  
 composites, use of, 276  
 definition, 2, 13  
 error, 13  
 impact, 14  
 structural equation modeling, 276  
 validity, 13  
 measurement invariance, 727, 737, 743  
 measurement model, 605, 608, 646, 659, 704–705, 762, 764, 771–774  
*see also outer model*  
 measurement model comparisons, 737–739, 740t, 741t  
 alternative to partial invariance, 739  
 full *vs.* partial invariance, 739  
 HBAT invariance analysis, 740–741  
 measurement invariance conclusions, 741  
 six-stage invariance testing process, 740–741  
 level of invariance needed, 740  
 six-step process of group comparisons, 737–739  
 measurement model estimation, 766–767  
 measurement model misspecification, 762  
 measurement model validity in confirmatory factor analysis, assessing, 673–679, 685–695

measurement models in structural equation modeling, 627–629  
 measurement relationship, 605, 610  
 measurement scales, 762  
   in confirmatory factor analysis, 670  
   in multivariate analysis, 11–13  
 measurement theory, 659, 661, 727, 728, 762, 773–774  
 measurement type bias, 742–744  
   model interpretation, 744  
   model specification, 742–743  
 measures of association, 55  
 measures of intercorrelation, 135–136  
   visual inspection, 135–136  
   Bartlett test, 136  
   measure of sampling adequacy, 136  
 measures of variable importance, comparing, 319f  
 mediating effect, 727  
 mediation, 262, 286, 287, 396, 397, 407, 409, 414f, 458t, 744–748, 762  
   conceptual basis for, 745  
   distribution system by purchase level, 457–459  
   estimation, 457  
   interpretation, 457  
   summary, 459  
   estimating model, 407, 409  
   HBAT illustration of, 746–748, 748t  
   step 1: establish significant relationships between constructs, 747  
   step 2: estimate mediated model and assess level of mediation and indirect effects, 747–748  
   interpreting, 419–420  
   combined effects, 420  
   indirect effect, relative strength of, 419–420  
   indirect mediation effect, significance of, 419  
   in MANOVA, 409  
   mediated main effect, significance of, 419  
   other causes, 746  
   rules of thumb, 753  
   significance of indirect effect, 407  
   summary, 755–756  
   testing for, 745–746  
 mediation effect, 745, 782–783  
 mediation model *see* mediation  
 mediation variable *see* mediation  
 mediator, 374, 390, 425  
 metric data, 2, 12  
 metric data vs. nonmetric data, 629, 769  
 metric independent variable, 97  
   magnitude of relationship, 575–576  
   example of assessing magnitude of change, 575–576  
   exponentiated logistic coefficients, 575  
   original logistic coefficients, 575  
 metric invariance, 728, 738  
 metric measurement scales, 12  
   impact of choice, 13  
   interval scales, 12  
   ratio scales, 12  
 metric moderators, 749–750  
 metric variable, 473, 474  
 MIC (mutual information correlation), 48  
 minimum sample size, 762  
 misclassification cost, 497, 551, 567  
 misclassification of individual cases, 501  
 missing at random (MAR), 48, 606, 630  
   imputation of missing data process, 69–70  
   summary, 70, 72  
 missing completely at random (MCAR), 48, 64, 606, 630  
   imputation calculating replacement values, 68  
   imputation using known replacement values, 67–68  
   imputation using only valid data, 66–67  
   Little's test, 65  
   overview of methods, 68–69

missing data, 48, 56–85  
   assessing randomness through group comparisons of observations with missing vs valid data, 78t  
   comparing estimates of means and standard deviations across complete case and size imputation methods, 81t  
   comparing imputed values, 80f  
   comparison of correlations across imputation methods for selected variables, 82t  
   deleting individual cases and/or variables, 62–63  
   extent/patterns of, 61–62  
   identifying and applying remedies, four-step process, 58–72, 59f  
   step 1: determine type of missing data, 60–61  
   step 2: determine extent of missing data, 61–63  
   step 3: diagnose randomness of missing data processes, 63–65  
   step 4: select imputation method, 65–70, 72  
   impact, 56  
   need for concern, 56–57  
   practical, 56  
   substantive, 56  
   imputed variables for selected cases, 79t  
   nonrandom manner, 61  
   patterns, 74t, 75t  
   in PLS-SEM, 770  
   recap of missing value analysis  
   imputation most logical course of action, 84  
   imputed correlations differ across techniques, 84  
   multiple methods for replacing missing data available and appropriate, 85  
   primarily MCAR, 83–84  
 regression results for complete case and six imputation methods, 83t  
 rules of thumb, 62  
 in structural equation modeling, 630–632, 631t  
 extent and pattern of, 630  
 remedies, 630  
 selecting approach, 631–632  
 summary, 83, 114  
 summary statistics for original sample, 73t  
 summary statistics for reduced sample, 76t  
 missing data analysis  
   recent developments, 57  
   simple example, 57  
   overall impact, 58, 58f  
   practical impact, 57  
   substantive impact, 57–58  
 missing data process, 48, 56, 63f  
   known processes, 60–61  
   unknown processes, 61  
 missing not at random (MNAR), 64  
 missingness, 48  
   levels, 61  
   construct-level, 61  
   item-level, 61  
   person-level, 61  
 mixture model, 191, 226  
 mode A, 762  
 Mode B, 762  
 model complexity, 762, 770–771  
 model definition, 610–613  
   theory, importance of, 610  
   visual portrayal of model, 610–613  
   combining measurement and structural relationships, 612–613  
 model fit, 613  
 structural equations model, depicting constructs involved, 610–611  
 structural relationships, depicting 611–612  
 model development strategy, 619  
 model diagnostics  
   in confirmatory factor analysis, 677–679, 712–713  
   caveats in model respecification, 679  
   modification indices, 678  
   specification searches, 678–679  
   standardized residuals, 678  
 model estimation  
   in logistic regression, 557–562  
   alternative models, 562  
   coefficients, estimating, 561  
   complete separation, 562, 562f  
   estimating coefficients, 559  
   goodness-of-fit, assessing, 563–571  
   issues in, 561–562  
   quasi-complete separation, 562, 562f  
   rules of thumb, 563  
   transforming dependent variable, 558–559  
   transforming probability into odds and logit values, 559–561  
   using maximum likelihood for estimation, 561  
 in structural equation modeling, 633–635  
 computer programs, 634–635  
 estimation technique, 633–634  
 model structure, 633  
 rules of thumb, 634  
 model estimation fit, 563–564  
   between model comparisons, 563–564  
 -2LL difference, 563  
   null model, 563  
   proposed model, 563  
   comparison to multiple regression, 564  
   global null hypothesis test, 564  
   overview of assessing, 571  
   pseudo R<sup>2</sup> measures, 564  
   rules of thumb, 573  
 model fit  
   in confirmatory regression model, 355–356  
   in discriminant analysis, 493–502, 509–520, 525–537  
   in logistic regression, 571, 584–587, 589–590  
   in multiple regression, 292–307, 564  
   in multivariate model, 34  
   in regression analysis, 332–348  
   in stepwise estimation, 332–333, 337  
   in structural equation modeling, 614, 623–625  
   in structural model, 710–711  
 model overlap, 374, 426  
 model respecification, 606, 679  
 model-based approach, 191, 226, 630  
 modeling strategy in structural equation modeling theory, 618–619  
 competing models strategy, 619  
 confirmatory modeling strategy, 618  
 model development strategy, 619  
 models, 606  
   assessing overall fit, 34  
   complexity, 632  
   simplify by separation, 31–32  
   specifying in CFA, 670–671  
   striving for parsimony, 31  
 models-based approach, 606  
 moderating effect, 728, 748  
 moderation, 396, 397, 409, 420, 420f, 452, 748–752, 751f, 752t, 762, 783  
   distribution system by firm size, 453–456  
   group profiles, 453  
   homoscedasticity, 453  
   moderation effect, interpreting, 456  
   statistical significance testing, 454–456  
   summary, 456  
   graphical displays of effects, 454f  
   HBAT illustration, 751–752  
   interpreting effect, 456  
   in MANOVA, 409  
   metric moderators, 749–750  
   multigroup SEM to test moderation, using, 750–751  
   nonmetric moderators, 749  
   rules of thumb, 753  
   summary, 755–756  
 moderator, 374, 390, 425

- moderator effect, 287  
 adding, 285  
 defined, 262, 284–285  
 examples, 284–285  
 interpreting, 285
- modification index, 659, 678, 692
- multicollinearity, 2, 14–15, 123, 135, 606, 762  
 among dependent variables, 401  
 decomposition of, 313  
 diagnosing, 349  
 effects of dependent variable on power, 406–407  
 increase in, 357  
 in logistic regression, 578  
 measuring degree and impact of, 349–350  
 in multiple regression analysis, 270  
 assessing, 311–317  
 how much is too much, 316–317  
 identifying, 312  
 measure of correlation incorporating multicollinearity, 311–312  
 predictive power, impact measures of, 270  
 remedies for, 317  
 variables with low multicollinearity, favors, 270
- in multiple regression analysis, effects of, 313–316  
 coefficients, 314–315  
 on estimation, 313–314  
 on explanation, 315–316  
 ignorable effects, 316  
 shared variances, interpretation of, 316  
 singularity, 313–314  
 standard errors, increases in, 314–315  
 suppression, 314  
 in sequential search methods, 297  
 stepwise estimation, impact in, 337, 339–341  
 summated scales as remedies for, 357, 359, 361  
*see also* collinearity
- multigroup analysis, 762, 782
- multigroup structural equations modeling, 750–751  
 summary, 755–756
- multilevel model (MLM), 262, 323–328  
 basic concepts and issues, 325–327  
 fixed vs. random effects, 326–327  
 intraclass correlation, 325–326  
 matching measurement properties to level, 325  
 sample size by level, 327  
 summary, 327  
 benefits, 324  
 brief history of, 324–325  
 five-stage modeling strategy, 327–328  
 review, 328  
 step 1: sufficient variation at level 2, 327–328  
 step 2: level-1 model with level-2 effects, 328  
 step 3: introduce level-2 independent variables, 328  
 step 4: test for random coefficients, 328  
 step 5: add cross-level interactions to explain variations in coefficients, 328  
 introduction, 323–325  
 resources for, 325  
 simple example, 323–324  
 separate equations by region, 323–324  
 summary, 328
- multiple analysis of covariance (MANCOVA), 26
- multiple discriminant analysis (MDA), 26–27, 471–544  
 chapter preview, 471–472  
 explained, 474–476  
 key terms, 472–474
- multiple equation vs single equation, 17
- multiple groups analysis, 728, 736–742  
 measurement model comparisons, 737–741  
 full vs. partial invariance, 739  
 HBAT invariance analysis, 740–741
- level of invariance, 740  
 partial invariance, alternative to, 739  
 six-step process of group comparisons, 737–739  
 rules of thumb, 743
- multiple imputation, 48, 69  
 combining results from multiple imputed datasets, 70  
 estimate the model, 70  
 generate set of imputed datasets, 69–70
- multiple R, 332
- multiple regression, 26, 262, 271f, 362t  
 decision diagram, 293–294f  
 in era of Big Data, 265–266  
 example of, 266–271  
 explained, 265  
 explanation with, 273–274  
 dependent variables, nature of relationship with, 274  
 independent variables, nature of relationship with, 274  
 independent variables, relative importance of, 274  
 extending, 322–330  
 multilevel models, 323–328  
 panel models, 328–330  
 model fit, comparison to logistic regression, 563–564, 564f  
 prediction using several independent variables, 269–271  
 adding third independent variable, 271  
 impact of multicollinearity, 270  
 multiple regression equation, 270–271  
 research problems, 273–274  
 explanation, 273–274  
 prediction, 273  
 rules of thumb, 278  
 selection of dependent and independent variables, 275–278  
 measurement error, 276  
 specification error, 276–278  
 strong theory, 276  
 specifying statistical relationship, 274–275  
 summary, 363–366
- multiple regression analysis  
 assumptions, 287–292  
 assessing individual variables vs variate, 287–288  
 error terms, 290, 291  
 linearity of the phenomenon, 288–290  
 methods of diagnosis, 288  
 rules of thumb, 292  
 summary, 292  
 chapter preview, 259–260  
 decision process, 272  
 stage 1: objectives of multiple regression, 273–278  
 stage 2: research design of multiple regression analysis, 278–286  
 stage 3: assumptions in multiple regression analysis, 287–292  
 stage 4: estimating regression model and assessing overall model fit, 292–307  
 stage 5: interpreting regression variate, 308–320  
 stage 6: validation of results, 321–330  
 definition, 265  
 key terms, 260–265  
 research design, 278–286  
 overview, 286  
 rules of thumb, 287  
 sample size, 278–280  
 variables, creating additional, 281–287  
 validation of results, 321–330  
 additional samples, 321  
 forecasting with model, 322  
 PRESS statistic, calculating, 321  
 regression models, comparing, 322  
 split samples, 321
- multiple univariate questions, 385
- multiple univariate tests, 415–416
- multisample confirmatory factor analysis (MCFA), 728, 737
- multivariate analysis, 3, 9–10  
 assessing individual variables vs. the variate, 93–94  
 chapter preview, 1  
 errors, looking at, 31  
 key terms, 2–3  
 measurement error and multivariate measurement, 13–14  
 measurement scales, 11–13  
 model parsimony, striving for, 31  
 model simplification by separation, 31–32  
 remedies for assumption violations, applying, 344–345
- statistical assumptions, 94  
 absence of correlated errors, 99–100  
 homoscedasticity, 97–99  
 linearity, 99  
 normality, 94–97  
 summary, 39–41, 55  
 testing assumptions, 93–100, 105–112  
 homoscedasticity, 108, 109f, 110–111f  
 linearity, 108, 112  
 normality, 105, 106f, 107–108, 107f, 110–111f  
 summary, 112  
 validity of results, 32
- variante, 10–11
- multivariate analysis and interpretation  
 data, knowing, 30–31  
 guidelines, 29–32  
 practical significance, establishing, 30  
 sample size affects all results, recognizing, 30  
 statistical significance, establishing, 30
- multivariate analysis of variance *see* MANOVA
- multivariate dependence methods, relationship between, 24f
- multivariate detection, 89, 92t, 93
- multivariate graphical display, 48, 54–55
- multivariate measurement, 3, 448t  
 employment, 13–14
- multivariate model  
 estimating and assessing overall model fit, 34  
 managing the variate, 14–16  
 validating, 34
- multivariate model building, structured approach, 32–33
- decision flowchart, 34
- define research problem, objectives, and multivariate technique to be used (stage 1), 33
- develop analysis plan (stage 2), 33
- estimate multivariate model and assess overall model fit (stage 4), 34
- evaluate assumptions underlying multivariate technique (stage 3), 33
- interpret variates (stage 5), 34
- miltivariate normal distribution, 374
- multivariate normality, 769
- multivariate procedures, 371–372
- group difference assessment, 377–380
- K-group case: MANOVA, 379–380
- two-group case: Hotelling's T<sup>2</sup>, 378–379
- multivariate profiles, 54–55
- multivariate questions, 385–386
- intrinsically multivariate questions, 386
- multiple univariate questions, 386
- structured multivariate questions, 386
- multivariate significance tests, 455
- multivariate statistical testing, 435, 436t, 440–442, 440t, 442t, 450t
- multivariate techniques, 476  
 canonical correlation, 28  
 classification, 21  
 cluster analysis, 26  
 cluster analysis as, 192  
 common factor analysis, 25–26

multivariate techniques (*continued*)

- confirmatory factor analysis, 27–28
- conjoint analysis, 28–29
- correspondence analysis, 29
- defining, 33
- evaluating assumptions underlying, 33
- with factor analysis, 131–132
- logistic regression, 27
- multiple discriminant analysis, 26–27
- multiple regression, 26
- multivariate analysis of covariance, 26
- multivariate analysis of variance, 26
- partial least squares structural equation modeling, 28
- perceptual mapping, 29
- principal components analysis, 25–26
- selecting, 22–23f
- structural equation modeling, 27–28
- structural equation modeling comparisons, 613–614
- types, 25–29
- using power with, 20
- multivariate tests, statistical power of, 403–407
- effects of dependent variable multicollinearity on Power, 406–407
- impacts, 404–406
- calculating power levels, 406
- sample size, 405
- statistical significance level alpha ( $\alpha$ ), 404–405
- unique issues with MANOVA, 405–406
- review of power in MANOVA, 407
- using power in planning and analysis, 406
- multivariate variance, 3–4

**N**

- natural experiment, 374, 388
- nearest-neighbour method *see* single-linked method
- negative predictive value (NPV), 551, 566, 587
- neighbourhood, 191, 225
- nested model, 606, 645
- noise point, 191, 225
- nominal scales, 11
- nomological validity, 123, 659, 677, 689–690, 789
- non-recursive model, 700, 707–708, 707f
- nonhierarchical cluster analysis, 212, 217–218
  - cluster seeds, selecting, 217–218
  - clustering algorithms, 218
  - combination of methods, 220–221
  - emergence of, 219–220
  - fine tuning, 235
  - optimizing procedure, 191
  - partitioning, 235
  - stage 4: deriving clusters and assessing overall fit, 245–247
  - stage 5: profiling clustering variables, 247–248
  - stage 6: validation and profiling the clusters, 248–251
- nonhierarchical clustering algorithms, 218
  - optimizing method, 218
  - parallel threshold method, 218
  - sequential threshold method, 218
- nonhierarchical procedures, 191
- nonlinear effects, assessing, 595, 595t
- nonlinearity, 99, 284f
- nonmetric data, 3
  - incorporating dummy variables with, 112–113, 281
  - vs. metric data, 629, 769
- nonmetric independent variable, 97, 112f, 361, 577f
  - interpreting magnitude for, 576–577
- nonmetric measurement scales, 11
  - nominal scales, 11
  - ordinal scales, 11–12
- nonmetric moderators, 749
- nonmetric variable, 473, 474, 551
  - see also* categorical variable

## nonrandom, 61

- nonrandomized situations, causal inference in, 421–428
- causality in social and behavioural sciences, 422
- counterfactuals in non-experimental research designs, 423–424
- overview, 428
- potential outcomes approach, 423
- propensity score models, 424–428
- rules of thumb, 428–429
- nonspurious covariance, 616–618
- collinearity, impact of, 617
- spurious relationships, testing for, 617–618
- normal distribution, 48, 94
- normal probability plots (NPP), 48, 95, 96f, 262, 291, 344f
  - of non-normal metric variables, 107f
- normality, 344
  - among dependent variables in MANOVA and ANOVA, 400–401
  - assessing impact of violating the assumption, 94
  - sample size, 95
  - shape of the distribution, 94–95
  - assessment of metric variables, 105, 107–108
  - assumptions in MANOVA, 434
  - checking for, 400
  - data transformations, 101
  - definition, 48, 94
  - remedies for non-normality, 97
  - tests, 95, 106f
  - graphical analyses, 95
  - statistical, 95–97
  - transformation of price elasticity, 110–111f
  - univariate vs. multivariate, 94
- normed chi-square, 638
- normed fit index (NFI), 638
- nuisance factor, 374, 392, 728, 742
- null hypothesis, 374, 377, 378f, 585
- null model, 563, 606
- null plot, 262, 288

**O**

- object, 191, 192
- objectives, defining, 33
- oblique factor rotation, 123, 148f, 150
  - vs. orthogonal rotation, 147–149
- oblique rotation, 176, 177t
- observation indicators used in HBAT and CFA of employee behaviour, 683t
- observational study, 374, 388
- observed heterogeneity, 782
- observed sample covariance matrix, 606, 621–622
- odds, 551, 559
- operationalization, 659
- operationalizing a construct, 606, 633
- optimal cutting score, 473
- optimal scaling, 123, 145
- optimizing procedure, 191, 218
- order condition, 659
- ordinal interaction, 374
- ordinal scales, 11–12
- organizational commitment (OC), 682, 713, 783
- organizational decision, 5–6
- original logistic coefficient, 551, 574, 575
- orthogonal, 123, 141, 374, 380
- orthogonal factor rotation, 123, 148f
  - method, 150
  - EQUIMAX, 150
  - QUARTIMAX, 150
  - VARIMAX, 150
  - vs. oblique rotation, 147–149
- out-of-sample predictive power *see* Q<sup>2</sup> value
- outcome, 374, 389
  - assessing significance for individual variables, 417–419
  - correlations of variables, 432
- outer loadings, 762
- outer model, 764
  - see also* measurement model
- outer weights, 762
- outliers, 85–93, 263, 307f
  - analysis, 91
  - outlier detection, 91, 93
  - retention or deletion, 93
  - assumptions in MANOVA, 434, 440, 448
  - classifying, 87–88
  - cluster analysis, 232–233, 236–238
  - contexts, 85
  - post-analysis: meeting analysis expectations, 86
  - pre-analysis: a member of a population, 85
  - definition, 48, 85
  - description and profiling, 90–91
  - designation, 87, 90
  - extraordinary event, 87
  - extraordinary observation, 87
  - procedural error, 87
  - summary, 88
  - unique combination, 88
  - detecting, 88
  - bivariate, 88–89, 91, 92f, 92t, 93
  - multivariate, 89, 92t, 93
  - rules of thumb, 90
  - univariate, 88, 91, 92t
  - identifying as influential observations, 345–348
  - impacts, 86
  - of dimensionality, 89–90
  - good or bad, 86
  - practical, 86
  - substantive, 86
  - influential observations, 302
  - largest dissimilarity values for identifying potential outliers, 234t
  - retention or deletion, 91, 93
  - sensitivity to, 401
  - summary, 86, 114–115
  - types of impacts on analysis, 87
  - error outliers, 87
  - influential outliers, 87
  - interesting outliers, 87
- overall fit, 532
  - in confirmatory factor analysis, 685–687
  - in discriminant model, 490, 493–501
  - calculating discriminant Z scores, 493–494
  - group differences, evaluating, 494
  - group membership prediction accuracy, assessing, 494–501
  - two-group illustrative example, 509–520
  - estimation of MANOVA model, 401–410
  - additional relationships: mediation and moderation, 407, 409
  - general linear model, 403
  - measures for significance testing, 403
  - statistical power of multivariate tests, 403–407
  - in logistic regression, 557–572
  - in logistic regression illustrative example, 581–592, 586t
  - assessing, 584–590
  - casewise diagnostics, 590, 592
  - stepwise model estimation, 582–584
  - nonhierarchical cluster analysis, 245–247
  - in principal component factor analysis, 168, 170–171
  - in regression analysis, 292–307
  - examining statistical significance of model, 299–302
  - influential observations, understanding, 302–307
  - managing the variate, 292
  - testing regression variate for meeting regression assumptions, 298–299
  - variable selection, 295–298
  - variable specification, 294

overall measurement model in confirmatory factor analysis, 663–669, 682–684  
congeneric measurement model, 665  
items per construct, 665–668  
reflective vs. formative measurement, 668–669  
rules of thumb, 669  
unidimensionality, 664–665  
overfitting, 3, 32  
overidentified model, 659, 667, 667f

**P**

p value, 762  
pairwise deletion, 630  
panel analysis *see* panel models  
panel models, 263, 328–330  
background, 329  
basic issues, 329–330  
adding time, 330  
fixed vs. random effects, selecting between, 330  
model types, 330  
variables, types of, 330  
benefits, 329  
similarity to multilevel models, 329  
summary, 330  
parallel analysis, 123, 143, 170, 171t  
as stopping rule for common factor analysis, 182t  
parallel coordinates graph *see* profile diagram  
parallel threshold method, 218  
parameter, 263, 280, 659  
parsimonious models, 762  
parsimony fit indices, 606, 639  
adjusted goodness of fit index, 639  
parsimony normed fit index, 639  
parsimony normed fit index (PNFI), 639  
parsimony ratio (PR), 655  
part correlation, 263  
partial correlation coefficient, 263, 296  
partial correlations, 167t, 169t, 311, 336  
partial F (or f) values, 263, 296, 504  
partial invariance, 728, 739  
partial least squares structural equation modeling (PLS-SEM), 28, 759–791, 772f, 785f, 787t  
chapter preview, 759  
decision process, 768–791  
stage 1: research objectives and constructs, defining and selecting, 768–769  
stage 2: designing study to produce empirical results, 769–771  
stage 3: measurement and structural models, specifying, 771–774  
stage 4: measurement model validity, assessing, 774–779  
stage 5: structural model, assessing, 779–781  
stage 6: advanced analyses using PLS-SEM, 782–783  
stage 7: measurement model reliability and validity, assessing, 785–790  
stage 8: structural model, assessing, 790–791  
definition, 762  
designing study to produce empirical results, 769–771  
metric vs. nonmetric data and multivariate normality, 769  
missing data, 770  
model complexity and sample size, 770–771  
statistical power, 770  
emergence of structural equations modeling, 765  
estimation of path models with, 766–767  
measurement model estimation, 766–767  
structural model estimation, 767  
using PLS-SEM algorithm, 767  
explained, 764–766  
general guidelines using, 771  
HBAT summary, 791  
illustration, 783–785

key terms, 760–763  
measurement model, 764  
measurement model validity, assessing, 774–779  
formative measurement models, 776–778  
reflective measurement models, 775–776  
summary, 779  
research objectives and constructs, defining and selecting, 768–769  
role of PLS-SEM vs. CB-SEM, 766  
rules of thumb, 771  
structural model, 764  
summary, 792–793  
theory and path models in, 765  
partial mediation, 374, 728, 745  
*see also* complete mediation  
partial regression plot, 263, 289–290, 304, 346  
partitioning procedures, 212, 235  
path analysis, 606, 622  
estimating relationships using, 650–652  
structural equation model, setting up for, 620–621, 621f  
path coefficients, 762, 785–786, 791t  
loadings, size of, 786  
potential problems, 786  
size and significance, 780–781, 786  
path diagrams, 606, 610, 615f, 644f, 661–663, 662f, 734f  
*see also* visual diagram  
path estimates  
in confirmatory factor analysis, 674–675, 690–691  
problems—identifying, 675  
size of, and statistical significance, 674–675  
*see also* structural parameter estimate  
path identification, 650  
path model, 765, 774  
estimating using PLS-SEM algorithm, 767  
stage 1: initial measurement model estimates, 767  
stages 2 and 3: initial structural model and final model estimates, 767  
theoretical PLS-SEM path model, 784–785  
Pearson residual, 551  
percentage correctly classified *see* accuracy; hit ratio  
percentage of variance criterion, 123, 142, 170  
perceptual mapping, 29  
Pillai's criterion, 375, 403  
planned comparison, 375, 416  
platykurtic distribution, 94  
PLS path modeling *see* partial least squares structural equation modeling  
PLS regression, 762  
PLS-SEM algorithm, 763, 767  
PLS-SEM bias, 763  
polar extremes approach *see* extreme groups approach  
polynomial, 263, 284f  
positive predictive value (PPV), 566, 587  
post hoc analysis, 700  
post hoc comparisons, 443–444, 444t  
post hoc test, 375, 416  
potency index, 473, 504–505, 538–539, 540t  
potential covariates, 452  
potential cross loading, 155  
potential outcomes, 375, 423  
power, 3, 19, 263, 375, 403  
analysis, 437  
calculating levels, 406  
effects of dependent variable multicollinearity on, 406–407  
in planning and analysis, 406  
review of power in MANOVA, 407  
in various regression models, 278–279  
practical significance  
defined, 3  
establishing, 30  
practical standpoint, 57  
prediction, 763  
misclassified, 535t  
multiple regression, 273  
maximize accuracy, 273  
model comparison, 273  
in regression analysis, 308–309  
using several independent variables, 269–271  
using single independent variable, 267–269  
assessing accuracy, 269  
prediction error, 263, 267, 763  
predictive accuracy, in logistic regression model  
estimation, 565–571, 566f, 568f, 596t  
Chi-square-based measure, 570–571  
classification matrix, 565–566  
concordance, 569–570  
cut-off value, selecting, 565  
measures, 566–568  
of actual outcomes, 567  
overall predictive accuracy, 566–567  
of predicted outcomes, 567  
summary, 567–568  
ROC curve, 568–569  
predictive relevance ( $Q^2$ ), 763  
predictor constructs, collinearity among, 779–780  
predictor variable *see* independent variable  
PRESS statistic, 263, 321, 354–355  
Press's Q statistic, 473, 500–501, 518, 534  
principal component analysis, 123, 139, 168–181  
alternatives, 144–146  
managerial overview of results, 183–184  
stage 4: deriving factors and assessing overall fit, 168, 169t, 170–171  
stopping rules, 168, 170–171  
stage 5: interpreting the factors, 171–176  
step 1: examine factor matrix of loadings for unrotated factor matrix, 172–173  
step 2: identify significant loadings in unrotated factor matrix, 173  
step 2 and, 3: assess significant factor loading(s) and communalities of rotated factor matrix, 173–175  
step 3: assess communalities of the variables in unrotated factor matrix, 173  
step 4: respecify factor model if needed, 175  
step 5: naming the factors, 175–176  
stage 6: validation, 176–177  
stage 7: additional uses of exploratory factor analysis results, 178–181  
creating summated scales, 179  
selecting between data reduction methods, 180–181  
selecting surrogate variables for subsequent analysis, 179  
use of factor scores, 179–180  
principal components analysis, 763  
probability, 560f  
calculating for specific value of independent variable, 578  
in logistic regression illustrative example, 593, 594t  
transforming into odds and logit values, 559–561  
calculating logit value, 560–561  
calculating probabilities, 559–560  
restating probability as odds, 560  
probit model, 551, 562  
problematic cross-loading, 155  
procedural error, 87  
PROCESS macro, 375, 409  
profile diagram, 191  
propensity score, 375, 426  
propensity scoring model, 375, 424–428  
developing and applying, 424–428  
step 1: specify covariates, 424–426  
step 2: estimate, 426  
step 3: balance of covariate set and model overlap, estimate, 426–427

- propensity scoring model (*continued*)  
 step 4: application of propensity scores, 427  
 step 5: causal effect, estimate, 427–428  
 popularity of, 424
- proportional chance criterion, 473, 518, 533, 551, 567
- pseudo R<sup>2</sup>, 551, 564, 586
- psychologic fallacy, 263, 323
- psychometrics, 659, 661
- purchase outcome measures, 439t, 441f, 442t, 446t, 447f, 450t, 453t, 455t, 456t
- Q**
- Q factor analysis, 123, 129
- Q<sup>2</sup> value, 763
- QUARTIMAX, 123, 150
- quasi-complete separation, 551, 562
- quasi-experiment, 375, 388
- R**
- R factor analysis, 123, 129
- R square (R<sup>2</sup>), 333  
*see also* coefficient of determination (R2)
- R<sup>2</sup> values, 763
- random effect, 263, 326
- randomness of missing data process, 63  
 diagnostic tests, 64–65  
 is it MAR or MCAR, 65  
 Little's MCAR test, 65  
 tests of missingness, 65  
 levels, 63  
 defining type of missing data process, 64  
 missing completely at random (MCAR), 64  
 missing data at random (MAR), 63–64  
 net missing at random (NMAR), 64
- rank condition, 659, 671
- rank order, 11–12
- ratio scales, 12
- ratio of variances, 300
- recursive models, 700, 707–708
- reduced form equation, 606, 607
- redundancy analysis, 763
- reference category, 48, 113, 263, 281
- reflective measurement, 659, 763
- reflective measurement models  
 assessing, 775–776  
 construct reliability, 775  
 convergent validity, 775  
 discriminant validity, 775  
 indicator loadings, 775  
 rules of thumb, 776
- reflective measurement theory, 724–729
- reflective vs. formative measurement, 668–669, 728–732, 732t  
 choosing, 732  
 differences, 730–731
- reflective vs. formative theory, 728–729
- regression analysis, 68  
 alternative models, evaluating, 355–357  
 estimates with multicollinear data, 315f  
 illustration of, 231  
 stage 1: objectives of multiple regression  
 331  
 stage 2: research design of multiple regression analysis, 331  
 stage 3: assumptions in multiple regression analysis, 332  
 stage 4: estimating regression model and assessing overall model fit, 332–348  
 stage 5: interpreting regression variate, 348–353  
 stage 6: validating results, 353–355  
 managerial overview of results, 361–363  
 nonmetric independent variable, including, 361  
 regression parameter estimates for complete case and six imputation methods, 84f  
 summated scales as remedies for multicollinearity, use of, 357–361
- variante, estimating for assumptions of, 342–345
- regression coefficient (b<sub>n</sub>), 263, 268, 310f, 313, 334, 359, 361  
 interpretation of, 348–349  
 interpretation with, 309–310  
 significance tests for, 301–302  
 standardizing, 310–311  
 using, 308–311  
 explanation, 309–311  
 prediction, 308–309  
 variance-decomposition matrix, 264
- regression equation, 265
- regression imputation, 48
- regression model, 265  
 evaluating alternative models, 355–363  
 evaluating variate for assumptions of regression analysis, 342–345  
 outliers as influential observations, identifying, 345–348  
 overall fit, estimating and assessing 292–307  
 examining statistical significance of model, 299–302, 308  
 influential observations, understanding, 302–308  
 managing the variate, 292  
 rules of thumb, 308  
 testing regression variate for meeting regression assumptions, 298–299  
 variable selection, 295–298
- variable specification, 294
- power levels in, 278–279
- stepwise estimation, 1: selecting first variable, 332–337
- stepwise estimation, 2: adding second variable (X<sub>g</sub>), 337–338
- stepwise estimation, 3: third variable (X<sub>12</sub>) added, 338–340
- stepwise estimation, 4: fourth and fifth variables (X<sub>7</sub> and X<sub>11</sub>) added, 340–341
- stepwise estimation overview, 341–342
- regression variate, 264, 265, 308–320  
 interpreting, 320  
 rules of thumb, 320  
 testing for meeting regression assumptions, 298–299
- regression weights, 318, 351–352
- relative importance, 317–320, 551  
 direct measures of variable importance, 317–318  
 beta (standardized) weights, 318  
 bivariate correlations, 317–318  
 regression weights, 318  
 squared semi-partial correlation, 318  
 measures of relative importance, 318–320  
 all possible subsets regression, 318  
 choice of measure, 319–320  
 commonality analysis, 318–319  
 dominance analysis, 319  
 relative weights, 319  
 structured coefficients, 318
- relative non-centrality index (RNI), 639
- relative weights, 319, 352
- reliability, 3, 13, 123, 606, 609, 676, 763
- repeated measures, 375, 397
- replacement values, 68  
 mean substitution, 68  
 regression analysis, 68
- replication, 158–159, 375
- research approaches  
 experimental: non-randomization, 388  
 natural experiment, 388  
 quasi-experiment, 388  
 experimental: randomization of groups, 387–388  
 controlled, 388  
 field, 388  
 non-experimental, 388
- observational study, 388  
 selecting, 389
- research design, 423–424  
 issues in SEM, 629–633  
 covariance vs. correlation, 629–630  
 metric vs. nonmetric data, 629  
 missing data, 630–632  
 sample size, 632–633
- research problem, defining, 33
- researcher, 4  
 control as preferred, 16  
 loss of control in sequential search, 297  
 multivariate use, 10
- residual, 606
- residual (e or ε), 48, 99, 264, 267, 289f  
 vs. leverage plot, 346
- residuals, 659  
 analysis of, 304, 345–346  
 in confirmatory factor analysis, 678  
 independence of, 344  
 in logistic regression, 571–572
- respecification, 147
- respondents  
 correlation among, in factor analysis, 133  
 heterogeneity of, 143
- response surface, 48, 104
- response-style effect, 191
- reverse scoring, 124
- ridge regression, 298
- robustness, 48, 93
- ROC curve, 551, 568–569, 569f, 570f, 589, 589t
- root mean square error of approximation (RMSEA), 637, 654
- root mean square residual (RMR), 637–638
- root mean square standard deviation (RMSSTD), 191, 222–223
- rotated factor matrix, assess significant factor loadings and communalities in, 173–175
- rotation, 537–538  
*see also* oblique factor rotation; orthogonal factor rotation
- row-centering standardization *see* within-case standardization
- Roy's greatest characteristic root (gcr), 375, 403
- S**
- sample representativeness, cluster analysis, 235
- sample size, 447t, 555–556  
 by label, 327  
 cluster analysis, 233–234  
 considerations, 445, 447  
 effect on results, 30  
 exploratory factor analysis, 132–133  
 impact on statistical power, 19–20  
 in logistic regression, 555–557, 581
- nometric independent variables, impact of, 556
- overall, 555
- per category of dependent variable, 555–556
- in MANOVA
- group/cell, 391
- overall, 391
- in MANOVA illustration, 432
- in multiple regression analysis, 278–280
- considerations, 280
- generalizability and, 279–280
- requirements for desired power, 279
- statistical power and, 278–279
- per category of dependent variable, 555–556
- low frequency of occurrence, 555
- low rate of occurrence, 555–556
- in PLS-SEM, 770–771
- power and, 406
- standards of comparison, 499–500
- in structural equation modeling, 632–633, 670
- average error variance of indicators, 632
- estimation technique, 632
- model complexity, 632
- summary on sample size, 633

- sampling adequacy, measure of, 136, 167t, 169t  
 sampling error, 18, 264  
 saturated structural model, 700, 711  
 scalar invariance, 728, 739  
 scale development, 124, 130, 627  
 scatterplot, 48, 52  
     matrix of selected metric variables, 53f  
     of seven observations based on two clustering variables, 194f  
 scoring procedure, 124  
 scree test, 124, 170  
 scree test criterion, 142–143  
*SDFBETA* *see DFBETA*  
*SDFFIT*, 306  
     *see also DFIT*  
 second-order factor model, 728, 733, 734f  
 second-order measurement theories, using, 735–736  
 selection strategy, 425–426  
*SEM* *see structural equation modeling*  
 semi-partial or part correlation, 311  
 sensitivity, 551, 567, 587  
 sequence, 616  
 sequential search methods, 295–298  
     caveats, 297–298  
     impact of multicollinearity, 297  
     increased alpha level, 298  
     loss of researcher control, 297  
     forward addition and backward elimination, 297  
     stepwise estimation, 295–297  
 sequential threshold method, 218  
 significance level (alpha), 19, 264, 301  
 significance tests, 403  
     Hotelling's  $T^2$ , 403  
     Pillai's criterion, 403  
     Roy's greatest characteristic root, 403  
     Wilks' lambda, 403  
     *see also critical value*  
 significance tests of regression coefficients, 301–302  
     confidence interval, applying, 301–302  
     confidence interval, establishing, 301  
     sampling error, 301  
     significance level, 301  
     standard error, 301  
 similarity *see interobject similarity*  
 simple regression, 264, 268f  
     example of, 266–271  
     prediction using single independent variable, 267–269  
     assessing accuracy, 269  
     interpreting model, 268–269  
 simultaneous estimation, 473  
 single equation, *vs* multiple equation, 17  
 single-case diagnostics, 305–306  
     individual coefficients, influences on, 305  
     overall influence measures, 305–306  
 single-item construct, 668, 763  
 single-item measures, 708–709  
 single-linkage method, 191, 215, 215f, 216f  
 singularity, 264, 313–314  
 skewness, 48, 94, 97–98  
 smoothing, 104  
 Sobel test, 375  
 specific variance, 124, 763  
 specification error, 3, 264, 276–278, 277f  
     irrelevant or redundant variable, inclusion of, 277  
     problem avoidance, 277–278  
     relevant variable, exclusion of, 277  
 specification search, 659, 678–679  
 specificity, 551, 567, 587  
 sphericity assumption, 375, 397  
 split-sample, 321, 354t  
     estimation, 353–354  
     validation, 32  
 split-sample validation *see cross-validation*
- spurious relationship, 606, 617–618  
     testing for, 617, 617f  
 squared Euclidean distance, 191  
 squared multiple correlation, 659, 675  
 squared semi-partial correlation, 318  
 stable unit treatment value assumption (SUTVA), 375  
 standard error, 264, 301  
     heteroscedasticity-consistent, 290  
     increases in, 314  
 standard error of the coefficient, 335  
 standard error of the estimate ( $SE_e$ ), 264, 269, 333  
 standardization, 49, 101–102, 264, 310  
     cluster analysis, 234–235  
 standardized coefficients, 336  
 standardized data, 763  
 standardized partial regression plots, 343f  
 standardized residuals, 264, 304, 638, 660, 678, 691  
 standardized root mean residual (SRMR), 637–638  
 statistical assumptions, testing, 100  
 statistical error, types of, 18–19  
 statistical inference, error probabilities  
     relationship, 19f  
 statistical issues  
     in exploratory factor analysis, 135–136  
     overall measures of intercorrelation, 135–136  
     variable-specific measures of intercorrelation, 136  
 statistical measure, 404, 584–586  
 statistical models, 3, 7, 8f, 279f  
     distinguishing between models, 8–9  
     *vs* data mining models, 7–9  
     *see also* data models  
 statistical power, 763, 770  
     adequacy of, 447  
     analysis, 21  
     comparison of two means, 20f  
     effects on, 19–20  
     managing three elements of, 20  
     of multivariate tests, 403–407  
     effects of dependent variable multicollinearity on, 406–407  
 impacts, 404–406  
 review of power in MANOVA, 407  
 using power in planning and analysis, 406  
 sample size and, 278–279  
 types, 18–19  
 using with multivariate techniques, 20  
     *vs* statistical significance, 18–20  
 statistical relationship, 264, 275, 275f  
 statistical significance  
     of cluster variation, 223, 242  
     of the coefficients, 592  
     of indicator weights, 777–778  
     path estimates in CFA, size of, 674–675  
     rules of thumb, 308  
     testing, 435, 437, 440–443, 454  
     *vs.* statistical power, 18–20  
 statistical testing, 379  
     assumptions of multivariate analysis, 94–100  
     homoscedasticity, 98  
 staying intentions (SI), 682, 713, 784  
 stepdown analysis, 375, 418–419  
 stepdown tests, 437, 442–443  
 stepwise estimation, 264, 295–296, 342t, 351t, 473  
     adding first variable, 332–337, 334t, 510–511, 511t, 526–527, 527t  
     adding second variable, 337–338, 338t, 512–513, 512t, 527, 528t  
     adding third variable, 338–340, 339t, 513–514, 513t, 527–528  
     adding fourth and fifth variables, 340–341, 340t, 527–528, 529t, 530t  
     estimated coefficients, 337  
     flowchart, 296f  
     identifying variables to add, 338  
     impact of multicollinearity, 337  
     in logistic regression illustrative example, 582–584t, 585t
- adding first variable, 583–584  
 adding second variable, 584  
 base model estimation, 582–583  
 looking ahead, 337  
 overall model fit, 332–333, 337  
 overview of process, 341–342  
 summary, 514–515, 514t, 530  
 variables in equation (step 1), 333–335  
 variables not in equation (step 1), 335–337  
 stop criterion *see convergence*  
 stopping rules, 124, 221, 224, 242t, 243f  
     alternative, 239  
 criteria for the number of factors to extract, 140–144  
 heterogeneity of the respondents, 143  
 latent root criterion, 141  
 parallel analysis, 143  
 percentage of variance criterion, 142  
 a priori criterion, 141  
 scree test criterion, 142–143  
 summary, 143–144  
 definition, 191  
 deriving factors and assessing overall fit, 168, 170–171  
 parallel analysis, 182t  
 stratification, 375, 427  
 stretched vector, 474, 505  
 strong theory, 276  
 structural equation model testing, 699–721  
     chapter preview, 699  
     illustration, 713–721  
     key terms, 700  
     overview, 702–703  
     stages in, 703  
     one-step *vs.* two-step approaches, 703  
     summary, 722  
 structural equation modeling notation, 628, 628f  
 structural equation modeling (SEM), 18, 27–28, 276, 603–655, 601f, 612f, 622f, 644f, 763  
 abbreviations, 653  
 advanced approaches, 752–755  
 latent growth models, 752–755  
 longitudinal data, 752  
 assessing measurement model validity, 635–642  
 absolute fit indices, 636–638  
 establishing acceptable and unacceptable fit, 641–642  
 guidelines for, 641–642  
 goodness-of-fit, basics, 635–636  
 incremental fit indices, 638–639  
 parsimony fit indices, 639  
 unacceptable model specification to achieve fit, 641  
 using fit indices, problems associated with, 639–640  
 assessing structural model validity, 644–647  
 competitive fit, 645–647  
 testing structural relationships, 647  
 chapter preview, 603–604  
 decision-process stages, 625–647, 626f  
 rules of thumb, 647  
 stage 1: individual constructs, defining, 627  
 stage 2: measurement model, developing and specifying, 627–629  
 stage 3: empirical results, designing study to produce, 629–635  
 stage 4: model validity, assessing measurement, 635–642  
 stage 5: specifying, 643–644  
 stage 6: validity, assessing, 644–647  
 defining individual constructs, 627  
 new scale development, 627  
 operationalizing, 627  
 pretesting, 627  
 scales from prior research, 627  
 defining a model, 610–613  
 model fit, 613  
 theory, importance of, 610  
 visual portrayal, 610–613

structural equation modeling (SEM) (*continued*)  
 definition, 606  
 designing study to produce empirical results, 629–635  
 model estimation, issues in, 633–635  
 research design, issues in, 629–633  
 rules of thumb, 634  
 developing and specifying measurement model, 627–629  
 creating, 629  
 SEM notation, 628–629  
 emergence of, 614, 765  
 explained, 607  
 key terms, 604–607  
 latent variables not measured directly, 608–610  
 exogenous vs. endogenous latent constructs, distinguishing, 610  
 latent constructs, benefits of using, 608–609  
 multiple interrelated dependence relationships, 607–608  
 multivariate techniques, other, 613  
 dependence techniques, similarity to, 613  
 interdependence techniques, similarity to, 613–614  
 simple example, 619–625, 661  
 estimation and assessment, basics of, 621–625  
 path analysis, setting up for, 620–621  
 rules of thumb, 625  
 theory, 619–620  
 specifying structural model, 643–644  
 summary, 648–649  
 theory in, 614–619  
 causation, establishing, 615–618  
 modeling strategy, developing, 618–619  
 specifying relationships, 614–615  
 using multigroup SEM to test moderation, 750–751  
 structural model, 606, 700, 714f, 716f, 717t, 718t, 719t, 720t, 721t, 763, 764, 771–772  
 assessing, 779–781, 790–791  
 blindfolding, 780  
 coefficient of determination, examining, 780  
 collinearity among predictor constructs, 779–780  
 effect size, 780  
 path coefficients, size and significance of, 780–781  
 rules of thumb, 781  
 summary, 781  
 comparisons, 741–742  
 estimation, 767  
 explained, 700–701  
 illustration  
 specifying structural model, 713–715  
 validity, assessing, 715–721  
 simple example, 701–702  
 specifying, 703–710  
 model specification using path diagram, 704–708  
 rules of thumb, 710  
 study, designing, 708–710  
 unit of analysis, 704  
 summary, 722  
 transforming to, 705–707  
 degrees of freedom, 707  
 notational changes, 706  
 theoretical changes, 705–706  
 validity, assessing, 710–713  
 model diagnostics, examining, 712–713  
 structural model fit vs. confirmatory factor analysis fit, understanding, 710–711  
 structural model fit vs. confirmatory factor analysis fit, 710–711  
 hypothesized dependence relationships, examining, 711  
 overall structure model fit, assessing, 711  
 saturated theoretical models, 711

structural parameter estimate, 700  
*see also* path estimates  
 structural relationship, 606, 611–612, 700  
 combining measurement with, 612–613  
 structural theory, 700, 709–710, 763, 774  
 structure coefficients, 352  
 structure correlations, 474, 503  
 structure matrix, 177t  
 structured multigroup tests, 416–417  
 structured multivariate questions, 385  
 studentized residual, 264, 304, 345f  
 substantive perspective, 56–58  
 sum of squared errors ( $SS_E$ ), 264, 267  
 sum of squares regression ( $SS_R$ ), 264  
 summated scales, 3, 13–14, 124  
 creating, 179  
 multicollinearity, as remedies for, 357, 359, 361  
*see also* composite measure  
 suppression effect, 265, 314  
 surrogate variable, 124  
 selecting for subsequent analysis, 179

**T**

t statistic, 375  
 t test, 375, 377  
 t value, 335–337  
 tau equivalence, 660  
 taxonomy, 191, 201  
 Ten (10) times rule, 763  
 territorial map, 474, 502, 505, 535–537, 536f  
 TF *see* totally free multiple group model  
 theoretical PLS-SEM path model, 784–785  
 exogenous constructs, 780–785  
 theory, 607, 610, 763, 765  
 importance of, 610  
 in structural equation modeling, 614–620, 713–714  
 causation, establishing, 615–618  
 modelling strategy, developing, 618–619  
 relationships, specifying, 614–615  
 simple example, 619–620  
 visual diagram, 714–715  
 three-group discriminant analysis: switching intentions, 481–483, 482t, 483f, 531t, 533t, 535t, 541t  
 calculating two discriminant functions, 481–483  
 identifying discriminant variables, 481  
 illustrative example, 523–543  
 three-indicator rule, 660, 671–672  
 tolerance, 265, 312, 316–317, 474, 489  
 total effect, 763  
 total indirect effect, 763  
 total sum of squares ( $SS_T$ ), 265  
 totally free multiple group model (TF), 728, 738  
 trace, 124  
 transformation, 265, 281, 287  
 treatments, 3, 26, 375, 376, 389  
 counterfactuals in non-experimental research designs, 423  
 main effects, 411  
 modeling other relationships between treatment and outcome, 396–397  
 number of, 393–394  
 cells formed, 393  
 creation of interaction effects, 393–394  
 selecting  
 basic model, 392  
 controls as, 392–393  
 selecting covariates, 395  
 correlated, 395  
 uncorrelated, 395  
 true negative, 551  
 true positives, 565  
 Tucker Lewis index (TLI), 638–639, 655  
 two independent groups, difference between, 432–438

stage 1: objectives of analysis, 432–433  
 examining group profiles, 433  
 research questions, 433  
 stage 2: research design of MANOVA, 433  
 stage 3: assumptions in MANOVA, 433–434  
 correlation and normality of dependent variables, 434  
 homoscedasticity, 434  
 independence of observations, 433  
 outliers, 434  
 stage 4: estimation of MANOVA model and assessing overall fit, 434–437  
 power analysis, 437  
 statistical significance testing, 435, 437  
 stage 5: interpretation of results, 437–438  
 summary, 438  
 two independent variables, factorial design for MANOVA, 444–452  
 stage 1: objectives of MANOVA, 445  
 group profiles, examining, 445  
 research questions, 445  
 stage 2: research design of MANOVA, 445, 447  
 sample size considerations, 445, 447  
 statistical power, adequacy of, 447  
 stage 3: assumptions in MANOVA, 447–448  
 homoscedasticity, 448  
 outliers, 448  
 stage 4: estimation of MANOVA model and assessing overall fit, 448–451  
 interaction effect, assessing, 448–449  
 interaction and main effects, testing, 450–451  
 stage 5: interpretation of results, 451–452  
 interaction and main effects, 451–452  
 potential covariates, 452  
 summary, 452  
 two-factor analysis, 431t  
 two-group analyses, 415  
 two-group discriminant analysis: purchasers vs. non-purchasers, 476–481, 480f, 510t, 511t, 512t, 513t, 514t, 517t, 519t, 521t  
 calculating discriminant function, 479  
 geometric representation, 480–481  
 identifying discriminant variables, 477–478  
 illustrative example, 508–523  
 two-step SEM process, 700, 703  
 Type I error, 3, 18–19, 375, 378–379  
 Type II error, 3, 19, 375, 403  
 typology, 191, 201

**U**

U statistic *see* Wilks' lambda  
 unbalanced design, 375, 391  
 underidentified model, 660, 666, 666f  
 unidentified model, 660, 666  
 unidimensional, 124  
 unidimensional measures, 660, 664–665  
 unique combination, 88  
 unique variance, 124  
 unit of analysis, 127, 129, 700, 704  
 univariate analysis of variance (ANOVA), 3, 26  
 univariate detection, 88, 91  
 univariate distribution, graphical representation, 52f  
 univariate F, 520–521  
 univariate measures, 448t  
 univariate procedure, 371, 377f  
 univariate profiling, examining shape of distribution, 51–52  
 univariate significance tests, 417, 455–456  
 univariate statistical tests, 435, 437, 440t, 442, 442t, 450t  
 unobserved heterogeneity, 782  
 unrotated common factor loadings matrix, 183t  
 unrotated component analysis factor matrix, 172t  
 unrotated factor matrix  
 assess communalities of variables in, 173  
 identify significant loadings in, 173

**V**

validation sample, 3, 32, 506, 551, 579  
 validity, 3, 13, 124, 763, 787  
   confirmatory factor analysis, 673–679, 685–695  
   criterion, 228–229, 249–250  
   cross-validation, 2, 32, 228  
   discriminant analysis, 506–507  
   in PLS-SEM, 774–779  
   principal components analysis, 176–177, 178t  
   split-sample, 32  
   structural model, 710–713, 715–721

value, 5

variability, 5

variable clustering, 124, 145–146

variable importance  
   assessing, 349  
   measures, 350–353, 357, 577–578  
   all possible subsets regression, 352  
   bivariate correlations, 350–351  
   commonality analysis, 352  
   dominance analysis, 352  
   regression and beta weights, 351–352  
   relative importance, 577–578  
   relative weights, 352  
   squared semi-partial correlations, 351  
   structure coefficients, 352  
   summary, 352–353  
   weight of evidence/information value, 578

## variable selection, 131

  factors are always produced, 131  
   factors require multiple variables, 131  
   measurement issues, 132  
   specification, 131

## variable transformation, 290

## variables

  assess communalities of, 155  
   assessing individual *vs.* the variate, 93–94  
   auxiliary, 70  
   confounding, 423  
   correlation among, in factor analysis, 133  
   database description, 35f  
   identifying, 338

## impact of irrelevant, 89–90

irrelevant or redundant, inclusion of, 277  
 metric independent, 97  
 nonmetric independent, 97, 112f  
 profiling discriminating variables, 522  
 relevant, exclusion of, 277  
 selection, 16, 295–299  
 combinatorial approach, 295  
 confirmatory or simultaneous, 295  
 constrained methods, 298  
 as necessary, 16  
 review of, 298  
 sequential search methods, 295–298  
 skewed distribution, 97–98  
 specification, 294, 299  
 specifying the variate, 15–16  
 summary, 114  
 type, 97

## variance

  common, 122  
   common *vs* unique, 138  
   compare as a ratio of variances, 154–155  
   compare variance not loadings, 154  
   error, 122  
   graphical tests of equal variance dispersion, 98  
   partitioning of a variable, 138–139  
   specific, 124  
   types included in factor matrix, 139f  
   unique, 124

variance extracted, 660, 676  
*see also* communality

variance inflation factor (VIF), 265, 312–313, 316–317, 763

variance-based SEM *see* partial least squares  
   structural equation modeling

variance-covariance matrices, equality of, 399–400  
   testing for, 399–400  
   when assumption not met, 400  
   check for normality, 400  
   equal cell sizes, 400  
   unequal cell sizes, 400  
   variable transformation, 400

variante, 3, 49, 93–94, 124, 129, 376, 377, 474–475, 551, 607, 763  
 evaluating for assumptions of regression analysis, 342–345  
 forming, 382–383  
 interpreting, 34, 268, 356–357  
 managing, 14–16, 15f, 21  
 in multiple regression equation, 270–271  
 reconciling two decisions, 16  
 selection, 16  
 specification is critical, 16  
 specifying variables, 15–16  
 variable reduction methods in managing, 125f

*vs* assessing individual variables, 287

## variety, 5

VARIMAX, 124, 150, 173–175, 174t, 184t  
 vector, 376, 474, 505

## veracity, 5

VIF *see* variance inflation factor

visual diagram, 714–715

*see also* path diagram

## volume, 4–5

**W**

Wald statistic, 551

Ward's method, 192, 216–217

weight of evidence, 551, 578

weighted least squares, 290

weighting, 427

whiskers, 53

whithin-construct error covariance, 660

Wilk's Lambda, 376, 403, 520–521

within-case standardization, 192

**Y**

Youden index, 551, 566

**Z**

Z score, 474

  calculating discriminant Z scores, 493–494

zero-order correlation, 311



