

Advanced and Multivariate Statistical Methods

Practical Application
and Interpretation

Sixth Edition

Craig A. Mertler
Arizona State University

Rachel Vannatta Reinhart
Bowling Green State University

Sixth edition published 2017
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2017 Taylor & Francis

First edition published by Pyrczak 2001
Fifth edition published by Pyrczak 2013

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN: 978-1-138-28971-0 (hbk)
ISBN: 978-1-138-28973-4 (pbk)
ISBN: 978-1-315-26697-8 (ebk)

Typeset in Times New Roman

Visit the eResources: www.routledge.com/9781138289734

CONTENTS

Detailed Chapter Contents	<i>v</i>
Preface	<i>xi</i>
Acknowledgments	<i>xiv</i>
Dedications	<i>xv</i>
Chapter 1 Introduction to Multivariate Statistics	1
Chapter 2 A Guide to Multivariate Techniques	13
Chapter 3 Pre-Analysis Data Screening	27
Chapter 4 Factorial Analysis of Variance	71
Chapter 5 Analysis of Covariance	99
Chapter 6 Multivariate Analysis of Variance and Covariance	125
Chapter 7 Multiple Regression	169
Chapter 8 Path Analysis	203
Chapter 9 Factor Analysis	247
Chapter 10 Discriminant Analysis	279
Chapter 11 Logistic Regression	307
Appendix A SPSS Data Sets	327
Appendix B The Chi-Square Distribution	357
Glossary	359
References	369
Subject Index	371

DETAILED CHAPTER CONTENTS

Chapter 1 Introduction to Multivariate Statistics	1
Section 1.1 Multivariate Statistics: Some Background	2
Research Designs	2
The Nature of Variables	3
Data Appropriate for Multivariate Analyses	4
Standard and Sequential Analyses	5
Difficulties in Interpreting Results	7
Section 1.2 Review of Descriptive and Inferential Statistics	7
Descriptive Statistics	7
Inferential Statistics.....	9
Section 1.3 Organization of the Book	12
Chapter 2 A Guide to Multivariate Techniques	13
Section 2.1 Degree of Relationship Among Variables.....	14
Bivariate Correlation and Regression.....	14
Multiple Regression	14
Path Analysis.....	14
Section 2.2 Significance of Group Differences.....	15
t Test.....	15
One-Way Analysis of Variance.....	15
One-Way Analysis of Covariance.....	15
Factorial Analysis of Variance	16
Factorial Analysis of Covariance	16
One-Way Multivariate Analysis of Variance	16
One-Way Multivariate Analysis of Covariance	17
Factorial Multivariate Analysis of Variance	17
Factorial Multivariate Analysis of Covariance.....	17
Section 2.3 Prediction of Group Membership.....	17
Discriminant Analysis	17
Logistic Regression	18
Section 2.4 Structure	18
Factor Analysis and Principal Components Analysis	18
Section 2.5 The Table of Statistical Tests	19
Section 2.6 The Decision-Making Tree for Statistical Tests.....	20
Summary	21
Exercises for Chapter 2	25
Chapter 3 Pre-Analysis Data Screening	27
Section 3.1 Why Screen Data?.....	27
Section 3.2 Missing Data	28
Section 3.3 Outliers.....	29
Section 3.4 Normality	32
Section 3.5 Linearity	34
Section 3.6 Homoscedasticity	35
Section 3.7 Using SPSS to Examine Data for Univariate Analysis	36

Univariate Example With Grouped Data	36
Univariate Example With Ungrouped Data	52
Section 3.8 Using SPSS to Examine Grouped Data for Multivariate Analysis	58
Summary	62
<i>Steps for Screening Grouped Data</i>	64
<i>Steps for Screening Ungrouped Data</i>	66
Exercises for Chapter 3	68

Chapter 4 Factorial Analysis of Variance 71

I. Univariate Analysis of Variance	71
II. Factorial Analysis of Variance	74
Section 4.1 Practical View	75
Purpose	75
Sample Research Questions	78
Section 4.2 Assumptions and Limitations	79
Methods of Testing Assumptions	79
Section 4.3 Process and Logic	80
The Logic Behind Factorial ANOVA	80
Interpretation of Results	81
Writing Up Results	84
Section 4.4 Sample Study and Analysis	84
Problem	84
Methods and SPSS “How To”	85
Output and Interpretation of Results	90
Presentation of Results	92
Section 4.5 Variations of the Two-Factor Design	93
Summary	94
<i>Checklist for Conducting a Factorial ANOVA</i>	95
Exercises for Chapter 4	96

Chapter 5 Analysis of Covariance 99

I. Analysis of Variance Versus Analysis of Covariance	99
II. Analysis of Covariance	100
Section 5.1 Practical View	100
Purpose	100
Sample Research Questions	103
Section 5.2 Assumptions and Limitations	103
Methods of Testing Assumptions	104
Section 5.3 Process and Logic	105
The Logic Behind ANCOVA	105
Interpretation of Results	107
Writing Up Results	111
Section 5.4 Sample Study and Analysis	112
Problem	112
Methods and SPSS “How To”	113
Output and Interpretation of Results	119
Presentation of Results	121
Summary	122

<i>Checklist for Conducting ANCOVA</i>	123
Exercises for Chapter 5	124

Chapter 6 Multivariate Analysis of Variance and Covariance 125

I. MANOVA	125
Section 6.1 Practical View	126
Purpose	126
Sample Research Questions	129
Section 6.2 Assumptions and Limitations	129
Methods of Testing Assumptions	130
Section 6.3 Process and Logic	131
The Logic Behind MANOVA	131
Interpretation of Results	132
Writing Up Results	135
Section 6.4 MANOVA Sample Study and Analysis	136
Problem	136
Methods and SPSS “How To”	137
Output and Interpretation of Results	140
Presentation of Results	144
II. MANCOVA	145
Section 6.5 Practical View	145
Purpose	145
Sample Research Questions	146
Section 6.6 Assumptions and Limitations	147
Methods of Testing Assumptions	148
Section 6.7 Process and Logic	148
The Logic Behind MANCOVA	148
Interpretation of Results	149
Writing Up Results	153
Section 6.8 MANCOVA Sample Study and Analysis	154
Problem	154
Methods and SPSS “How To”	154
Output and Interpretation of Results	159
Presentation of Results	161
Summary	162
<i>Checklist for Conducting MANOVA</i>	163
<i>Checklist for Conducting MANCOVA</i>	165
Exercises for Chapter 6	166

Chapter 7 Multiple Regression 169

Section 7.1 Practical View	169
Purpose	169
Sample Research Questions	177
Section 7.2 Assumptions and Limitations	177
Methods of Testing Assumptions	178
Section 7.3 Process and Logic	181
The Logic Behind Multiple Regression	181
Interpretation of Results	182
Writing Up Results	188

Section 7.4 Sample Study and Analysis.....	189
Problem	189
Methods and SPSS “How To”	189
Output and Interpretation of Results	196
Presentation of Results.....	197
Summary	197
<i>Checklist for Conducting Multiple Regression</i>	199
Exercises for Chapter 7	200

Chapter 8 Path Analysis 203

Section 8.1 Practical View	203
Purpose.....	203
Sample Research Questions	208
Section 8.2 Assumptions and Limitations.....	208
Methods of Testing Assumptions.....	210
Section 8.3 Process and Logic	210
The Logic Behind Path Analysis.....	210
Interpretation of Results	219
Writing Up Results.....	224
Section 8.4 Sample Study and Analysis.....	226
Problem	226
Methods, SPSS “How To,” Output, and Interpretation.....	227
Presentation of Results.....	241
Summary	241
<i>Checklist for Conducting Path Analysis</i>	243
Exercises for Chapter 8	244

Chapter 9 Factor Analysis..... 247

Section 9.1 Practical View	247
Purpose.....	247
Sample Research Questions	255
Section 9.2 Assumptions and Limitations.....	255
Methods of Testing Assumptions.....	257
Section 9.3 Process and Logic	257
The Logic Behind Factor Analysis.....	257
Interpretation of Results	258
Writing Up Results.....	263
Section 9.4 Sample Study and Analysis.....	264
Problem	264
Methods, SPSS “How To,” Output, and Interpretation.....	264
Presentation of Results.....	272
Summary	273
<i>Checklist for Conducting Factor Analysis</i>	274
Exercises for Chapter 9	275

Chapter 10 Discriminant Analysis 279

Section 10.1 Practical View	279
Purpose	279
Sample Research Questions	284
Section 10.2 Assumptions and Limitations.....	284

Methods of Testing Assumptions.....	285
Section 10.3 Process and Logic	286
The Logic Behind Discriminant Analysis.....	286
Interpretation of Results	287
Writing Up Results.....	293
Section 10.4 Sample Study and Analysis.....	294
Problem	294
Methods and SPSS “How To”	294
Output and Interpretation of Results	299
Presentation of Results	303
Summary	303
<i>Checklist for Conducting Discriminant Analysis</i>	305
Exercises for Chapter 10	306

Chapter 11 Logistic Regression 307

Section 11.1 Practical View	307
Purpose.....	307
Sample Research Questions	311
Section 11.2 Assumptions and Limitations.....	311
Section 11.3 Process and Logic	312
The Logic Behind Logistic Regression	312
Interpretation of Results	313
Writing Up Results.....	315
Section 11.4 Sample Study and Analysis.....	316
Problem	316
Methods and SPSS “How To”	316
Output and Interpretation of Results	320
Presentation of Results	322
Summary	322
<i>Checklist for Conducting Binary Logistic Regression</i>	323
Exercises for Chapter 11	324

PREFACE

Nearly all graduate programs in the social and behavioral sciences (i.e., education, psychology, sociology, nursing, criminal justice, and so on) offer an advanced or multivariate statistics course that covers concepts beyond those addressed in an introductory-level statistics course. This text addresses these advanced statistical concepts.

Purpose of The Text and Its Intended Audience

This text provides conceptual and practical information regarding multivariate statistical techniques to students who do not *necessarily* need technical and/or mathematical expertise in these methods. The students for whom we have written this text are required to understand the basic concepts and practical applications of these advanced methods in order to interpret the results of research studies that have utilized such methods, as well as to apply these analytical techniques as part of their thesis or dissertation research.

This text was written for use by students taking a multivariate statistics course as part of a graduate degree program in which the course is viewed as a research tool. Examples of degree programs for which this text would be appropriate include—but are not limited to—psychology, education, sociology, criminal justice, social work, mass communication, and nursing. Although the text is not primarily intended for students who are majoring in statistical or research methodologies, they could certainly use it as a reference.

This text has three main purposes. The first purpose is to facilitate conceptual understanding of multivariate statistical methods by limiting the technical nature of the discussion of those concepts and focusing on their practical applications. The multivariate statistical methods covered in this text are:

- factorial analysis of variance (ANOVA),
- analysis of covariance (ANCOVA),
- multivariate analysis of variance (MANOVA),
- multivariate analysis of covariance (MANCOVA),
- multiple regression,
- path analysis,
- factor analysis,
- discriminant analysis, and
- logistic regression.

The second purpose is to provide students with the skills necessary to interpret research articles that have employed multivariate statistical techniques. A critical component of graduate research projects is a review of research literature. It is crucial for students to be able to understand not only what multivariate statistical techniques were used in a particular research study, but also to appropriately interpret the results of that study for the purposes of synthesizing the existing research as background for their own study. The acquisition of these skills will benefit students not only during the time that they are conducting their graduate research studies, but also long after that time as they review current research as part of their professional career activities.

The third purpose of this text is to prepare graduate students to apply multivariate statistical methods to the analysis of their own quantitative data or that of their institutions, such that they are able to complete the following for each particular technique:

- understand the limitations of the technique,

- fulfill the basic assumptions of the technique,
- conduct the appropriate steps (including the selection of various options available through the use of computer analysis software),
- interpret the results, and
- write the results in the appropriate research reporting format.

Special Features

There are currently no texts available that take a more practical, truly *applied* approach to multivariate statistics. Most texts tend to be highly technical and mathematical in their orientation. While this approach is appropriate for students who are majoring in statistical and research methodologies, it is not nearly as appropriate for students representing other majors in the social and behavioral sciences.

We see a mismatch between what professors are trying to teach in applied advanced statistics courses and the content presented in current texts. Often, professors are required to water down the content coverage of the statistical techniques when using one of the existing texts. An alternative is for professors to utilize an introductory statistics text and supplement the content with lectures and examples. This approach ultimately harms students because they often do not have access to these supplementary resources in order to process the information at their own pace outside of class.

Our main reason for writing this book was that we were unable to find an appropriate text to use in teaching our courses in Advanced Statistical Methods. Our students had taken only an introductory statistics course and found many of the texts previously available to be far too technical. To our knowledge, no text was previously available that combined advanced/multivariate statistical methods with a practical approach to conducting and interpreting such tests.

Approach to Content

To facilitate student understanding of the practical applications and interpretation of advanced/multivariate statistical methods, each of Chapters 4 through 11 of this text focuses on a specific statistical test and includes sections on the following:

- Practical View
 - Purpose
 - Sample Research Questions
- Assumptions and Limitations
 - Methods of Testing Assumptions
- Process and Logic
 - The Theory Behind the Test
 - Interpretation of Results
 - Writing Up Results
- Sample Study and Analysis
 - Problem
 - Methods and SPSS “How To”
 - Output and Interpretation of Results
 - Presentation of Results
- Summary

The first section of each chapter, **Practical View**, describes the purpose of the test in very conceptual terms and provides sample research questions and scenarios.

The section titled **Assumptions and Limitations** presents the assumptions that must be fulfilled in order to conduct the specific test, suggested methods for testing those assumptions, and limitations of the test.

Process and Logic explains the theory behind the test in very practical terms, the statistics that are generated from the test, how to determine significance, and how the results are written.

The **Sample Study and Analysis** section provides a demonstration of the process of the test, how it was conducted using SPSS, the SPSS output generated, how to interpret the output, and how to summarize the results.

Please note that in addition to the Base program, SPSS also offers several modules. Most institutions have at least the SPSS Base program—with the modules Regression and Advanced Models installed. With these components, you can conduct all of the analyses in this text. If you are using the student version of the SPSS Base program only, some procedures will not be available. However, the SPSS Graduate Pack will provide all the necessary modules.

A unique feature of our text is the highlighting of key test statistics and their implications for the test within the SPSS output. This has been accomplished by highlighting relevant test statistics and describing them with dialogue boxes within example outputs.

The sample studies are based on analysis of SPSS data sets, which can be downloaded from the publisher's web page:

www.routledge.com/9781138289734

Data sets are also described in **Appendix A** near the end of this book, which includes specification of variables and their measurement scales. In addition, this section demonstrates the steps used in conducting the test by means of SPSS.

Within each chapter, pertinent SPSS menu screens are pictured to display the sequence of steps involved in the analysis, as well as the options available in conducting the test.

Finally, the **Summary** provides a step-by-step checklist for conducting the specific statistical procedure covered in that chapter.

Pedagogical Features

All chapters, with the exception of Chapters 1 and 2, include **SPSS example output** as well as **menu screens** to demonstrate various program options. **Tables** that display sample study results have been included. **Figures**, wherever appropriate, have been used to graphically represent the process and logic of the statistical test.

Because the data sets are accessible via the web address shown above, students can conduct the sample analyses themselves as well as practice the procedures using other data sets. Assignments that utilize these data sets have also been included.

New to the Sixth Edition

Several important changes have been made to the sixth edition of *Advanced and Multivariate Statistical Methods*. These include the following:

- All references to SPSS (e.g., directions and screenshots) have been **updated to Version 23** of the software. This incremental revision is reflected throughout the book.¹ For users of the previous edition of the text, it is important to note that you may not see substantive changes to various SPSS screenshots; however, these remain compatible with the most current version of SPSS.²
- Approximately 8 to 10 Student Learning Objectives have been added to the beginning of each chapter as a means for students to target their learning and for instructors to focus their instruction.
- Lists of keywords that appear in each chapter have also been added to the end of each chapter.
- All the SPSS data sets utilized in this edition are **available for download** from the following web address: www.routledge.com/9781138289734. From this web address, you may download the data set files individually as you proceed through the book, or you may download them all at once in a single compressed file that contains all data sets. For users of previous editions of this text, it is important to note that these data sets are identical to those used in previous editions.

ACKNOWLEDGMENTS

We would like to sincerely thank the late Fred Pyrczak for believing in us and in this project from its inception. Without his assistance and timely responsiveness, this text would not exist. We would also like to express our gratitude to Monica Lopez and Edward Brancieri of Pyrczak Publishing, who had a hand in the production of this book. Thanks also to the numerous professors and students at institutions across the country who have sent unsolicited feedback to us over the years. It has been sincerely appreciated. Doctoral student Keenan Colquitt was instrumental in creating updated SPSS screenshots, and Dr. Bernadette R. Hadden of The City University of New York Hunter College created test banks for the sixth edition. Finally, we would certainly be remiss if we did not acknowledge the editorial feedback—both grammatical and substantive in nature—provided by our doctoral students in the College of Education and Human Development at Bowling Green State University.

*Craig A. Mertler
Rachel Vannatta Reinhart*

¹ Due to the considerable flexibility of the SPSS program and the many options it offers that affect its appearance, you may notice slight differences on your screen from the screenshots in the text (e.g., charts may display variable labels rather than variable names). Please consult SPSS online help for guidance if you wish to modify your display's appearance to better conform to the included screenshots. Remember to check the SPSS website for the latest updates and patches to the SPSS program.

² At the time this text was published, the current version of SPSS was Version 23.

CHAPTER 1

INTRODUCTION TO MULTIVARIATE STATISTICS

STUDENT LEARNING OBJECTIVES

After studying Chapter 1, students will be able to:

1. Distinguish between multivariate and univariate analyses.
2. Explain distinctions between experimental and nonexperimental research designs.
3. Apply various categorization schemes for classifying variables.
4. Explain differences between various data matrices.
5. Provide appropriate examples of orthogonal and non-orthogonal relationships between variables.
6. Distinguish between standard and sequential analyses.
7. Explain the process of hypothesis testing and determining statistical significance, using appropriate terminology.
8. Evaluate the difference between Type I and Type II errors.
9. Discuss the relationship between α (alpha) and β (beta).
10. Describe what is reported by effect size.

For many years, multivariate statistical techniques have simplified the analysis of complex sets of data. As a collective group, these techniques enable researchers, evaluators, policy analysts, and others to analyze data sets that include numerous independent variables (IVs) and dependent variables (DVs). In other words, they allow researchers to analyze data sets where the participants have been described by several demographic variables and also have been measured on a variety of outcome variables. For instance, a researcher may want to compare the effectiveness of four alternative approaches to reading instruction on measures of reading comprehension, word recognition, and vocabulary, while controlling for initial reading ability. The most appropriate method of analyzing these data is to examine the relationships and potential interactions between all variables simultaneously. Relying on univariate statistical procedures would prevent proper examination of these data. Due to the increasingly complex nature of research questions in the social sciences, and to the advent—and continued refinement—of computer analysis programs (e.g., SPSS[®], SAS[®], BMDP[®], and SYSTAT[®]), the results of multivariate analyses are appearing more and more frequently in academic journals.

The purpose of this book is to provide the reader with an overview of multivariate statistical techniques by examining each technique in terms of its purpose, the logic behind the test, practical applications of the technique, and the interpretations of results. The authors' major goal is to prepare students to apply and interpret the results of various multivariate statistical analysis techniques. It is not our intent to inundate

the student with mathematical formulae, but rather to provide an extremely practical approach to the use and interpretation of multivariate statistics.

SECTION 1.1 MULTIVARIATE STATISTICS: SOME BACKGROUND

Multivariate statistical techniques are used in a variety of fields, including research in the social sciences (e.g., education, psychology, and sociology), natural sciences, and medical fields. Their use has become more commonplace due largely to the increasingly complex nature of research designs and related research questions. It is often unrealistic to examine the effects of an isolated treatment condition on a single outcome measure—especially in the social sciences, where the participants in research studies are nearly always human beings.

As we all know, human beings are complex entities, complete with knowledge, beliefs, feelings, opinions, attitudes, and so on. Studying human participants by examining a single independent variable and a single dependent variable is truly impractical because these variables do not co-exist in isolation as part of the human mind or set of behaviors. These two variables may affect or be affected by several other variables. In order to draw conclusions and offer accurate explanations of the phenomenon of interest, the researcher should be willing to examine many variables simultaneously.

Stevens (2001) offers three reasons for using multiple outcome measures (i.e., DVs) in research studies, specifically those involving examinations of the effects of varying treatments (e.g., teaching methods, counseling techniques, etc.).

1. Any treatment will usually affect participants in more than one way. Examining only one criterion measure is too limiting. To fully understand the effects of a treatment condition, the researcher must look at various ways that participants may respond to the conditions.
2. By incorporating multiple outcome measures, the researcher is able to obtain a more complete and detailed description of the phenomenon under investigation.
3. Treatments can be expensive to implement, but the cost of obtaining measures on several dependent variables (within the same study) is often quite small and allows the researcher to maximize the information gain.

It should be noted that a study appropriate for ***multivariate*** statistical analysis is typically defined as one with several dependent variables (as opposed to ***univariate*** studies, which have only one dependent variable). However, the authors have included several techniques in this book that are typically classified as ***advanced univariate*** techniques (e.g., multiple regression, factorial analysis of variance, analysis of covariance, etc.). The reason for their inclusion here is that they are ordinarily not included in an introductory course in statistical analysis but are nonetheless important techniques for students to understand.

Research Designs

The basic distinction between experimental and nonexperimental research designs is whether the levels of the independent variable(s) have been manipulated by the researcher. In a true experiment, the researcher has control over the levels of the IVs. That is, the researcher decides to which conditions participants will be exposed. For instance, if a researcher was conducting an experiment to investigate the effectiveness of three different counseling techniques, she would randomly assign each subject to one of the three conditions. In essence, she has ***controlled*** which participants receive which treatment condition.

In nonexperimental research (e.g., descriptive, correlational, survey, or causal-comparative designs), the researcher has no control over the levels of the IVs. The researcher can define the IV, but cannot assign participants to its various levels. The participants enter the study already “belonging” to one of the levels. For instance, suppose a researcher wanted to determine the extent to which groups differed on some outcome measure. A simple scenario might involve an examination of the extent to which boys and girls differed with respect to their scores on a statewide proficiency test. The independent variable, ***gender*** in

this case, cannot be manipulated by the researcher. All participants enter the study already categorized into one of the two levels of the IV. However, notice that in both experimental and nonexperimental research designs, the levels of the independent variable have defined the groups that will ultimately be compared on the outcome DV.

Another important distinction between these two types of research designs lies in the ability of the researcher to draw conclusions with respect to causality. In an experimental research study, if the researcher finds a statistically significant difference between two or more of the groups representing different treatment conditions, he can have some confidence in attributing causality to the IV. Manipulating the levels of the IV by randomly assigning participants to those levels permits the researcher to draw causal inferences from the results of his study. However, because there is no manipulation or random assignment in a nonexperimental research study, the researcher is able to conclude that the IV and DV are related to each other, but causal inference is limited.

The choice of statistical analysis technique is extraneous to the choice of an experimental or non-experimental design. The various multivariate statistical techniques described in this book are appropriate for situations involving experimental as well as nonexperimental designs. The computer analysis programs will run and the statistics will work in either case. However, the decision of the researcher to attribute causality from the IV(s) to the DV(s) is ultimately dependent upon the initial decision of whether the study will be experimental or nonexperimental.

The Nature of Variables

The authors have been using the terms *independent* and *dependent* variables throughout the beginning of this chapter, so a review of these terms—and others related to the nature of variables—is in order. Variables can be classified in many ways. The most elementary classification scheme dichotomizes variables into either independent or dependent variables. Independent variables consist of the varying treatment conditions (e.g., a new medication vs. a standard medication) to which participants are exposed or differing characteristics that the participants bring into the study with them (e.g., school location, defined as urban, suburban, or rural). In an experimental situation, the IVs may also be referred to as *predictor* or *causal* variables because they have the potential of causing differing scores on the DV, which is sometimes referred to as the *criterion* or *outcome* variable. The reader should also be aware that a specific variable is not inherently an IV or a DV. An IV in one study might be a DV in another study, and vice versa. *Univariate* statistics refers to analyses where there are one or more IVs and only one DV. *Factorial* analyses are appropriate in situations where there are two or more IVs and one DV. *Bivariate* statistics refers to analyses that involve two variables where neither is identified as an IV or a DV. Finally, *multivariate* statistics refers to situations where there is more than one DV and there may be one or more IVs.

Another way to classify variables refers to the level of measurement represented by the variable. Variables may be quantitative, categorical, or dichotomous. *Quantitative* variables are measured on a scale that has a smooth transition across all possible values. The numerical value represents the amount of the variable possessed by the subject. Examples of quantitative variables include age, income, and temperature. Quantitative variables are also referred to as *continuous* or *interval* variables.

Categorical variables consist of separate, indivisible categories. There are no values between neighboring categories of a categorical variable. Categorical variables are often used to classify participants. Examples of categorical variables include gender (male or female), type of school (urban, suburban, or rural), and categories of religious affiliation. Categorical variables may also be referred to as *nominal*, *ordinal*, *discrete*, or *qualitative*. A specific type of categorical variable is one that is *dichotomous*. A dichotomous variable is one that has only two possible levels or categories. For instance, gender is a categorical variable that is also dichotomous. Often, for purposes of addressing specific research questions, quantitative or categorical variables may be dichotomized. For instance, age is a quantitative variable, but one could recode the values so that it would be transformed into a dichotomous variable. Age could be dichotomized

into two categories, such as “less than 35 years of age” and “35 years of age and older.” Often, a transformation of data such as these allows the researcher to be more flexible in terms of the analysis techniques she can use.

When conducting a multivariate analysis, researchers sometimes have a tendency to include too many variables. Prior consideration of the analysis is crucial in determining the variables on which to collect and include data. The best recommendation is to obtain the solution with the fewest number of variables (Tabachnick & Fidell, 2007). This is known as a *parsimonious* solution. Arguments for the inclusion of variables should be based on the feasibility (i.e., cost and availability) of collecting data on them and the nature of the theoretical relationships among the variables being considered.

Data Appropriate for Multivariate Analyses

Obviously, the data for multivariate analyses must be numerical. Quantitative variables consist of the scores themselves on specific variables. The values for categorical variables consist of the codes assigned by the researcher. For instance, for the variable of school location, urban schools might be assigned a 1, suburban schools assigned a 2, and rural schools assigned a 3.

There are many forms in which data can be submitted for analysis using multivariate techniques. The majority of the time, a data matrix will be analyzed. A **data matrix** is an organization of raw scores or data, where the rows represent participants, or cases, and the columns represent variables. Another possible format in which data may appear for analysis is a correlation matrix. Readers who have completed an introductory course in statistics are probably somewhat familiar with this type of matrix. A **correlation matrix** is a square, symmetrical matrix where each row and each column represents a different variable and the intersecting cells contain the correlation coefficient between two variables. A third option is a **variance-covariance matrix**, which is also a square, symmetrical matrix where the elements on the main diagonal (i.e., the intersection of a variable with itself) represent the variance of each variable and the elements on the off-diagonals represent the covariances between variables. Finally, a **sum-of-squares and cross-products matrix** is the precursor to the variance-covariance matrix. Specifically, it is a matrix consisting of deviation values that have not yet been averaged.

The mathematical calculations involved in multivariate statistical analyses may be performed on any of the previously mentioned matrices. However, the calculations are rather complex and involve a set of skills known as matrix algebra. **Matrix algebra** is somewhat different from scalar algebra—that is, addition, subtraction, multiplication, and division of a single number—with which the reader is more familiar. Matrix algebra is an extension of scalar algebra where mathematical operations are performed on an ordered array of numerical values. Because, as stated earlier in this chapter, it is not the intent of the authors to deluge the reader with intricate, and often convoluted, mathematical calculations, matrix algebra will not be discussed further in this text. If the reader is interested in learning more about matrix algebra and its applications in multivariate statistical analyses, several excellent resources include Johnson and Wichern (2008), Tabachnick and Fidell (2007), Stevens (2001), and Tatsuoka (1988).

In multivariate statistics, as in univariate statistics, the quality of the data is crucial. Fortunately, advanced computer analysis programs make the computations easy. However, there is a downside to this wonderful feature: The programs will provide output to the requested analysis, including beautifully formatted graphs and tables, regardless of the quality of the data on which the analyses were performed. For instance, assume a researcher has data that have not been reliably collected, contain data-entry errors, and include strange values that will surely influence the results. In this case, the adage “garbage in, garbage out” holds true. However, by simply examining the output, the researcher usually is unable to discern that the results are of poor quality. Prior to analysis, the researcher must take measures to ensure that the data are of the highest possible quality (techniques will be discussed in Chapter 3). Only by doing so can one be assured of the quality of the results and confident of the subsequent conclusions drawn.

Standard and Sequential Analyses

The benefits of—and the disadvantages associated with—multivariate statistics are often direct results of the relationships among the variables in a given data set. A lack of relationship among variables typically enables the researcher to interpret the results of an analysis with more clarity. For this reason, orthogonality is an important concept in the application of multivariate statistical analyses. **Orthogonality** is perfect nonassociation between variables. If we know the value for an individual on a given variable, and if that variable has an orthogonal relationship with a second variable, knowing the value of the first variable provides no information in determining the value of the second variable. In other words, the correlation between the two variables is equal to zero.

Orthogonality is often a desirable quality for multivariate statistical analyses. For instance, assume we are interested in examining the nature of the relationships among a set of IVs and a single DV. If all pairs of IVs in the set are orthogonal, then each IV adds a distinctively unique component to the prediction of the DV. As a simple example, assume that we are investigating the effects that two IVs (years of education and motivation) have on a single DV (income). If years of education and motivation are orthogonal, then each contributes separately, and in additive fashion, to the prediction of income. For instance, if 25% of the variability in income can be predicted by years of education and 40% can be predicted by motivation, then 65% of the variability in income can be predicted from years of education and motivation taken together. This relationship can easily be shown through the use of a Venn diagram (see Figure 1.1).

Figure 1.1. Venn Diagram for Orthogonal Relationship Among Income, Years of Education, and Motivation.

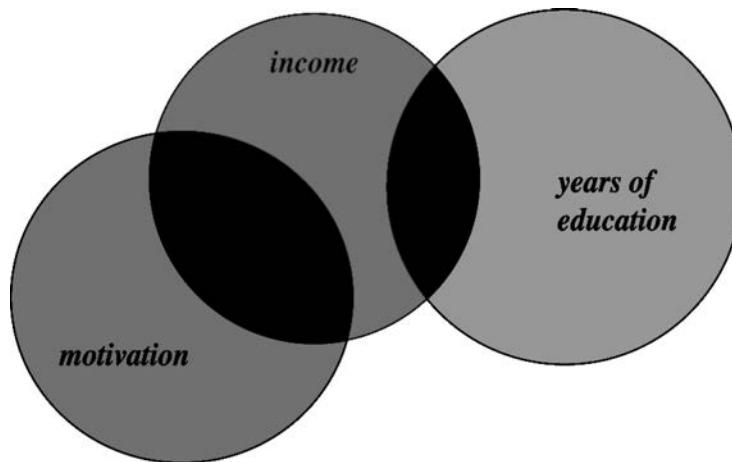
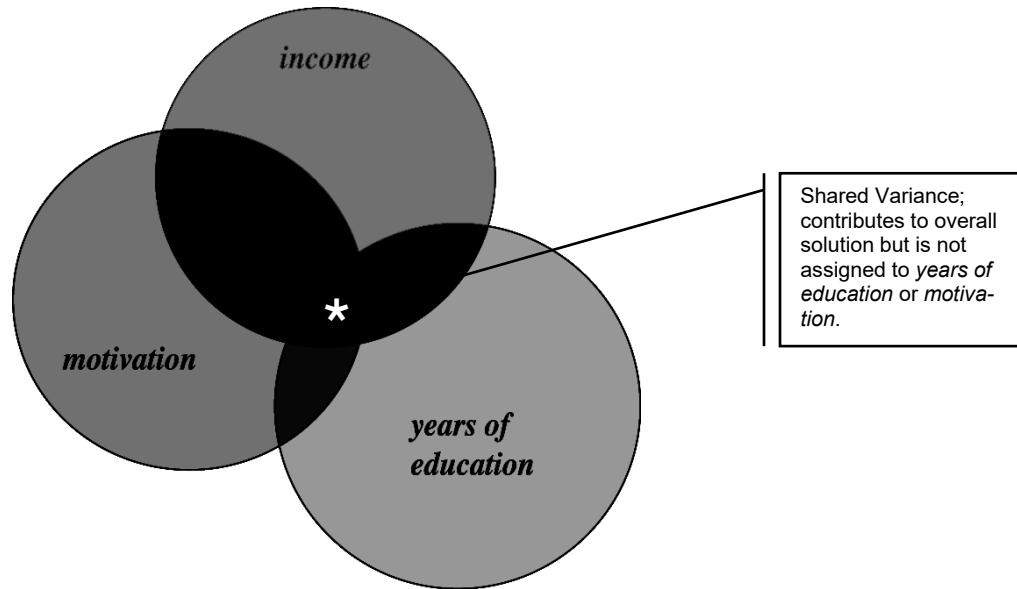
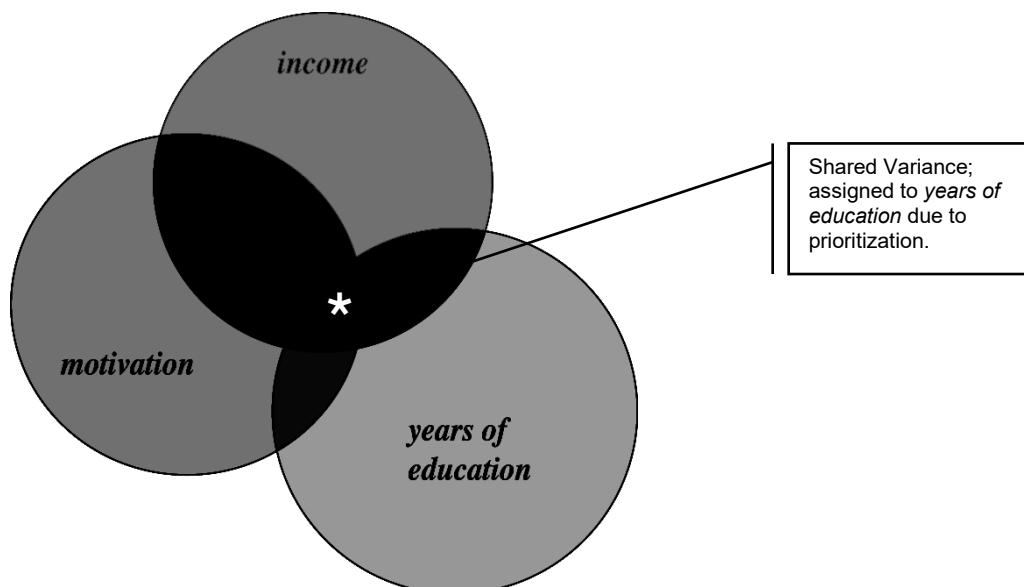


Figure 1.2. Venn Diagram Resulting From a Standard Analysis of the Relationship Among Income, Years of Education, and Motivation.



Having a data set with orthogonal variables is the ideal situation. However, most variables with which social science researchers work are correlated to some degree. That is, they are nonorthogonal. When variables are correlated, they have overlapping, or shared, variance. Returning to our previous example, if years of education and motivation were correlated, the contribution of years of education to income would still be 25%, and the contribution of motivation would remain at 40%. However, their combined contribution—which could no longer be determined by means of an additive procedure—would be less than 65%. This is due to the fact that the two IVs share some amount of variance. There are two basic strategies for handling this situation in order to determine the contribution of individual IVs to a DV.

Figure 1.3. Venn Diagram Resulting From a Sequential Analysis of the Relationship Among Income, Years of Education, and Motivation.



Using a **standard analysis** approach, the overlapping portion of variance is included in the overall summary statistics of the relationship of the set of IVs to the DV, but that portion is not assigned to either of the IVs as part of their individual contribution. The overlapping variance is completely disregarded when evaluating the contribution of each IV, taken separately, to the overall solution. Figure 1.2 is a Venn diagram of a situation where years of education and motivation are nonorthogonal and share some variance. For this shared portion, we are investigating the effects that two IVs (years of education and motivation) have on a single DV (income). Using a standard analysis approach, the shared variance is included in the total variability explained by the set of IVs but is not assigned to either years of education or motivation when examining their *individual* contributions.

An alternative approach, **sequential analysis**, requires the researcher to prioritize the entry of IVs into the equation or solution. The first variable entered into the equation will be assigned both its unique variance and any additional variance that it shares with any lower-priority variables; the second variable will be assigned its unique variance and any overlapping variance with lower-priority variables, and so on. Figure 1.3 is a Venn diagram showing this type of approach, where years of education has been assigned the highest priority and therefore is credited both with its unique variance and that which it shares with motivation. Notice that, in this situation, the total amount of variance remains the same. However, years of education now has a stronger relationship to income than it did in the standard analysis, while the contribution of motivation remains the same.

Difficulties in Interpreting Results

The need to understand the nature of relationships among numerous variables measured simultaneously makes multivariate analysis an inherently difficult subject (Johnson & Wichern, 2008). One of the major difficulties in using multivariate statistical analyses is that it is sometimes nearly impossible to get a firm statistical answer to your research questions (Tabachnick & Fidell, 2007). This is due largely to the increased complexity of the techniques. Often, results are ambiguous. Two or more statistical indices resulting from one computer run may contradict each other. The researcher must then determine the most appropriate way to interpret the results of the analysis. This introduces some subjectivity into the process. But rest assured, we believe that the benefits of being able to examine complex relationships among a large set or sets of variables make multivariate procedures well worth the time and effort required to master them.

SECTION 1.2 REVIEW OF DESCRIPTIVE AND INFERRENTIAL STATISTICS

The purpose of the remainder of this chapter is to provide the reader with a brief review of descriptive and inferential statistics. While it certainly is not our intention to provide thorough coverage of these topics, the discussions should serve as a good refresher of material already mastered by the reader prior to beginning a course in multivariate statistics.

Descriptive Statistics

The first step in nearly any data analysis situation is to describe or summarize the data collected on a set of participants that constitute the sample of interest. In some studies, such as simple survey research, the entire analysis may involve only descriptive statistics. However, most studies begin with a summarization of the data using descriptive techniques and then move on to more advanced techniques in order to address more complex research questions. There are four main types of descriptive statistics: measures of central tendency, variability, relative position, and relationship.

Measures of Central Tendency. Measures of central tendency permit the researcher to describe a set of data with a single, numerical value. This value represents the average, or typical, value. The three most commonly used measures of central tendency are the mode, the median, and the mean. The **mode** is the most frequently occurring score in a distribution. There is no calculation involved in obtaining the mode. One simply examines the distribution of scores and determines which score was obtained by the participants

most often. The mode does have limited use and is most appropriately used for variables measured at a nominal level.

The **median** is the score in the distribution that divides the upper 50% of scores from the lower 50%. Like the mode, the median is also of limited use because it does not take into consideration all values in the distribution. The values of extreme scores, both positive and negative, are completely ignored. The median is most appropriate for ordinal measures.

The most frequently used measure is the **mean**, which is simply the arithmetic average of a set of scores. It is the preferred measure of central tendency because it takes into account the actual values of all scores in a distribution. If there are extreme scores in the distribution, the mean can be unduly influenced (e.g., an extremely high score will increase the value of the mean, thus making it less representative of the distribution). In this case, the median may be the more appropriate measure. However, when data are measured on an interval or ratio scale, the mean remains the favored method of describing central tendency.

Measures of Variability. Often, a measure of central tendency is not enough to adequately describe a distribution of scores. A researcher may also want to know the degree to which the scores are spread around the mean, or another measure of central tendency. The amount of spread is indicated by one of three measures of variability. The most basic measure of variability is the range. The **range** is simply the difference between the highest score and the lowest score in the distribution. The range is not a good indicator of the amount of spread in scores because it is based solely on the largest and smallest values. It is typically used only as a rough estimate of the variability in a set of scores.

When a distribution of scores contains some extreme values, an alternative to the range is the quartile deviation. The **quartile deviation** is defined as one-half of the difference between the 3rd quartile (i.e., the 75th percentile) and the 1st quartile (i.e., the 25th percentile). The resulting value is the amount of spread in the scores that are located within a range defined by the median $\pm 12.5\%$ of the cases. A limitation of the quartile deviation is that it does not take into consideration all values in the distribution.

The **standard deviation** is an appropriate measure of variability when variables are measured on an interval or ratio scale. The standard deviation is defined as a special type of average distance of scores away from the mean. It is the most stable measure of variability because it takes into account every score in the distribution. It is obtained by first subtracting the mean from each score, squaring the resulting differences, summing the squared differences, and finally finding the average of that summed value. This value is called the **variance**, and one must simply find the square root of the variance in order to obtain the standard deviation. A large standard deviation indicates that the scores in the distribution are spread out away from the mean, and a small standard deviation indicates that the scores are clustered closer together around the mean. The mean and standard deviation taken together do a fairly good job of describing a set of scores.

Measures of Relative Position. Measures of relative position indicate where a specific score is located in relation to the rest of the scores in the distribution. Interpretation of these measures allows a researcher to describe how a given individual performed when compared to all others measured on the same variable(s). The two most common measures of relative position are percentile ranks and standard scores.

Many of us have seen our performances on standardized tests reported as percentile ranks. A **percentile rank** indicates the percentage of scores that fall at or below a given score. If a raw score of 75 points corresponds to a percentile rank of 88, then 88% of the scores in the distribution were equal to or less than 75 points. Percentile ranks are most appropriate for ordinal measures, although they are often used for interval measures as well.

There are several types of standard scores that can be used to report or describe relative position. A **standard score** is derived from the manipulation of a raw score that expresses how far away from the mean a given score is located, usually reported in standard deviation units. Because the calculation of a standard score involves some algebraic manipulation, the use of standard scores is appropriate when data are measured at an interval or ratio level. Two of the most common types of standard scores are *z*-scores and *T*-scores. A ***z*-score** indicates the distance away from the mean a score is in terms of standard deviation units

and is calculated by subtracting the mean from the raw score and then dividing the value by the standard deviation. If a raw score was equal to the mean, it would have a *z*-score equal to 0. If a raw score was two standard deviations greater than the mean, it would have a *z*-score equal to +2.00. If a raw score was one standard deviation below the mean, it would have a *z*-score equal to -1.00. Note that the sign is an important component of a reported *z*-score because it serves as a quick indicator of whether the score is located above or below the mean.

A ***T*-score** is simply a *z*-score expressed on a different scale. In order to convert a *z*-score to a *T*-score, simply multiply the *z*-score by 10 and add 50. For instance, if we had a distribution with a mean of 65 and a standard deviation of 5, an individual who obtained a raw score of 75 would have a *z*-score equal to +2.00 and a *T*-score equal to 70. The reader should be aware that all three measures used in this example (i.e., the raw score, the *z*-score, and the *T*-score) indicate a score that is equivalent to two standard deviations above the mean.

Measures of Relationship. Measures of relationship indicate the degree to which two quantifiable variables are related to each other. These measures do not describe—or even imply—a causal relationship. They only verify that a relationship exists. Degree of relationship between two variables is expressed as a correlation coefficient ranging from -1.00 to +1.00. If the two variables in question are not related, a coefficient at or near zero will be obtained. If they are highly related, a coefficient near +1.00 or -1.00 will be obtained. Although there are many different types of correlation coefficients, depending on the scale of measurement being used, two commonly used measures of relationship are the *Spearman rho* and the *Pearson r*.

If data for one or both of the variables are expressed as ranks (i.e., ordinal data) instead of scores, the *Spearman rho* is the appropriate measure of correlation. The interpretation is the same as previously discussed, with values ranging from -1.00 to +1.00. If a group of participants produced identical ranks on the two variables of interest, the correlation coefficient would be equal to +1.00, indicating a perfect relationship.

If data for both variables represent interval or ratio measures, the *Pearson r* is the appropriate measure of correlation. Like the mean and standard deviation, the *Pearson r* takes into account the value of every score in both distributions. The *Pearson r* assumes that the relationship under investigation is a linear one; if in reality it is not, then the *Pearson r* will not yield a valid measure of the relationship.

Inferential Statistics

Inferential statistics deal with collecting and analyzing information from samples in order to draw conclusions, or inferences, about the larger population. The adequacy, or representativeness, of the sample is a crucial factor in the validity of the inferences drawn as the result of the analyses. The more representative the sample, the more generalizable the results will be to the population from which the sample was selected. Assume we are interested in determining whether or not two groups differ from each other on some outcome variable. If we take appropriate measures to ensure that we have a representative sample (i.e., use a random sampling technique), and we find a difference between the group means at the end of our study, the ultimate question in which we are interested is whether a similar difference exists in the population from which the samples were selected. It is possible that no real difference exists in the population and that the one that we found between our samples was due simply to chance. Perhaps if we had used two different samples, we would not have discovered a difference. However, if we do find a difference between our samples and conclude that the difference is large enough to infer that a real difference exists in the population (i.e., the difference was *statistically significant*), then what we really want to know is, “How likely is it that our inference is incorrect?” This idea of “how likely is it” is the central concept in inferential statistics. In other words, if we inferred that a true difference exists in the population, how many times out of 100 would we be wrong? Another way of looking at this concept is to think of selecting 100 random samples, testing each of them, and then determining for how many our inference would be wrong.

There are several key underlying concepts to the application of inferential statistics. One of those is the concept of standard error. Any given sample will, in all likelihood, not perfectly represent the population. In fact, if we selected several random samples from the same population, each sample would probably have a different sample mean and probably none of them would be equal to the population mean. This expected, chance variation among sample means is known as **sampling error**. Sampling error is inevitable and cannot be eliminated. Even though sampling errors are random, they behave in a very orderly fashion. If enough samples are selected and means are calculated for each sample, all samples will not have the same mean, but those means will be normally distributed around the population mean. This is called the **distribution of sample means**. A mean of this distribution of sample means can be calculated and will provide a good estimate of the population mean. Furthermore, as with any distribution of scores, a measure of variability can also be obtained. The standard deviation of the sample means is usually referred to as the standard error. The **standard error of the mean** tells us by how much we would expect our sample means to differ if we used other samples from the same population. This value, then, indicates how well our sample represents the population from which it was selected. Obviously, the smaller the standard error, the better. With a smaller standard error, we can have more confidence in the inferences that we draw about the population based on sample data. In reality, we certainly would not have the time or resources to select countless random samples, nor do we need to. Only the sample size and the sample standard deviation are required in order to calculate a good estimate of the standard error.

The main goal of inferential statistics is to draw inferences about populations based on sample data, and the concept of standard error is central to this goal. In order to draw these inferences with confidence, a researcher must ensure that a sample is representative of the population. In **hypothesis testing**, we are testing predictions we have made regarding our sample. For instance, suppose the difference between two means was being examined. The **null hypothesis** (H_0) explains the chance occurrence that we have just discussed and predicts that the only differences that exist are chance differences that represent only random sampling error. In other words, the null hypothesis states that there is no true difference to be found in the population. In contrast, the **research** or **alternative hypothesis** (H_1) states that one method is expected to be better than the other or, to put it another way, that the two group means are not equal and therefore represent a true difference in the population. In inferential statistics, it is the null hypothesis that we are testing because it is easier to disprove the null than to prove the alternative.

Null hypotheses are tested through the application of specific statistical criteria known as **significance tests**. Significance tests are the procedures used by the researcher to determine if the difference between sample means is substantial enough to rule out sampling error as an explanation for the difference. A test of significance is made at a predetermined **probability level** (i.e., the probability that the null hypothesis is correct), which obviously allows the researcher to pass judgment on the null hypothesis. For instance, if the difference between two sample means is not large enough to convince us that a real difference exists in the population, the statistical decision would be to “fail to reject the null hypothesis.” In other words, we are not rejecting the null hypothesis, which stated that there was no real difference, other than a difference due to chance, between the two population means. On the other hand, if the difference between sample means was substantially large (i.e., large enough to surpass the statistical criteria), we would “reject the null hypothesis” and conclude that a real difference, beyond chance, exists in the population.¹ There are a number of tests of significance that can be used to test hypotheses, including, but not limited to, the *t* test, analysis of variance, the chi-square test, and tests of correlation.

Based on the results of these tests of significance, the researcher must decide whether to reject or fail to reject the null hypothesis. The researcher can never know with 100% certainty whether the statistical decision was correct, only that he or she was *probably* correct. There are four possibilities with respect to statistical decisions—two reflect correct decisions and two reflect incorrect decisions:

¹ For testing the difference(s) between mean(s), the size of the difference(s) is a major consideration. Sample size and the amount of variability within the samples also play a major role in significance testing.

1. The null hypothesis is actually true (i.e., there is no difference), and the researcher concludes that it is true (*fail to reject H_0*) — correct decision.
2. The null hypothesis is actually false (i.e., a real difference exists), and the researcher concludes that it is false (*reject H_0*) — correct decision.
3. The null hypothesis is actually true, and the researcher concludes that it is false (*reject H_0*) — incorrect decision.
4. The null hypothesis is actually false, and the researcher concludes that it is true (*fail to reject H_0*) — incorrect decision.

If it is concluded that a null hypothesis is false when it is actually true (Number 3), a **Type I error** has been committed by the researcher. If a null hypothesis is actually false when it is concluded to be true (Number 4), a **Type II error** has been made.

When a researcher makes a decision regarding the status of a null hypothesis, she or he does so with a pre-established (*a priori*) probability of being incorrect. This probability level is referred to as the **level of significance** or **alpha (α) level**. This value determines how large the difference between means must be in order to be declared significantly different, thus resulting in a decision to reject the null hypothesis. The most common probability levels used in behavioral science settings are $\alpha = .05$ or $\alpha = .01$. The selected significance level (α) determines the probability of committing a Type I error, which is the risk of being wrong that is assumed by a researcher. The probability of committing a Type II error is symbolized by β (beta) but is not arbitrarily set, as is alpha. To determine the value for β , a complex series of calculations is required. Many beginning researchers assume that it is best to set the alpha level as small as possible (thereby reducing the risk of a Type I error to almost zero). However, the probability levels of committing Type I and Type II errors have a complimentary relationship. If one reduces the probability of committing a Type I error, the probability of committing a Type II error increases (Harris, 1998). These factors must be weighed, and levels established, prior to the implementation of a research study.

The **power** of a statistical test is the probability of rejecting H_0 when H_0 is, in fact, false. In other words, making a correct decision (Number 2). Power is appropriately named because this is exactly what the researcher hopes to accomplish during hypothesis testing. Therefore, it is desirable for a test to have high power (Agresti & Finlay, 2009). Power is determined in the following manner:

$$\text{Power} = 1 - \beta$$

Power, as with α , is established arbitrarily and should be set at a high level because the researcher is hoping to reject a null hypothesis that is not true and wants to have a high probability of doing so (Brewer, 1978).

Another factor related to hypothesis testing is **effect size**. **Effect size** (often denoted as *ES* or partial η^2) is defined as the size of the treatment effect the researcher wishes to detect with respect to a given level of power. In an experimental study, *ES* is equal to the difference between the population means of the experimental and control groups divided by the population standard deviation for the control group. In other words, it is a measure of the amount of difference between the two groups reported in standard deviation units (it is a standardized or transformed score and is, therefore, metric-free). Effect sizes can also be calculated for correlation coefficients or for mean differences resulting from nonexperimental studies (Harris, 1998). Effect size, like α and power, is set *a priori* by the researcher, but it also involves strong consideration of what the researcher hopes to find in the study as well as what constitutes important and trivial differences.

A more powerful statistical test will be able to detect a smaller effect size. Cohen (1988) established a rule of thumb for evaluating effect sizes: An *ES* of .2 is considered small, one of .5 is considered medium, and one of .8 is considered large. A researcher would want to design a study and statistical analysis procedures that would be powerful enough to detect the smallest effect size that would be of interest and non-trivial (Harris, 1998).

Sample size (*n*) is a final factor whose value must be considered when conducting a research study and which must be considered prior to data collection. The required sample size for a study is a function of

α , power, and effect size. Because sample size has several relationships with these three factors, values for the factors must be set prior to the selection of a sample. For instance, for a fixed α -level, the probability of a Type II error decreases when the sample size increases. In addition, for a fixed α -level, power increases as sample size increases. If sample size is held constant and α is lowered (e.g., in an attempt to reduce the probability of committing a Type I error), power will also decrease. This fact provides partial justification for not setting α near zero. The power of a test would be too low and the researcher may be much less likely to reject a null hypothesis that is really false (Agresti & Finlay, 2009). However, the solution to this dilemma is not to obtain the largest sample possible. A huge sample size might produce such a powerful test that even the slightest, most trivial difference could be found to be statistically significant (Harris, 1998). In summation, as n increases, ES, α , and β will decrease, causing power to increase (Brewer, 1978). However, obtaining the largest possible sample size need not be the goal because the most appropriate sample involves a balanced combination of α , ES, and power. Tables have been developed to provide optimum sample sizes for a diverse range of values for n , α , ES, and power (Cohen, 1969).

SECTION 1.3 ORGANIZATION OF THE BOOK

The remainder of this textbook is organized in the following manner. Chapter 2 presents a guide to various multivariate techniques, in addition to a review of several univariate techniques. Included for each is an overview of the technique and descriptions of research situations appropriate for its use. Chapter 3 addresses the assumptions associated with multivariate statistical techniques and also discusses methods for determining if any of those assumptions have been violated by the data and, if so, how to deal with those violations. The concepts and procedures discussed in Chapter 3 are requisite to conducting any of the statistical analyses in subsequent chapters.

Chapters 4 through 11 present specific multivariate statistical procedures. Included in each chapter (i.e., for each technique) are a practical description of the technique, examples of research questions, assumptions and limitations, the logic behind the technique, and how to interpret and present the results. A sample research study, from problem statement through analyses and presentation of results, is also included in each chapter. Finally, a step-by-step guide to conducting the analysis procedure using SPSS is presented.

KEYWORDS

- correlation matrix
- data matrix
- distribution of sample means
- effect size
- hypothesis testing
- level of significance [alpha (α) level]
- matrix algebra
- mean
- median
- mode
- multivariate statistical analyses
- null hypothesis (H_0)
- orthogonality
- Pearson r
- percentile rank
- power
- probability level
- quartile deviation
- range
- research or alternative hypothesis (H_1)
- sampling error
- sequential analysis
- significance tests
- Spearman ρ
- standard analysis
- standard deviation
- standard error of the mean
- standard score
- sum-of-squares and cross-products matrix
- T -score
- Type I error
- Type II error
- univariate
- variance
- variance-covariance matrix
- z -score

CHAPTER 2

A GUIDE TO MULTIVARIATE TECHNIQUES

STUDENT LEARNING OBJECTIVES

After studying Chapter 2, students will be able to:

1. Describe various techniques for measuring the degree of relationships between variables, including bivariate correlation and regression, multivariate regression, and path analysis.
2. Describe various techniques for testing the significance of group differences, including *t* tests, one-way ANOVAs and ANCOVAs, factorial ANOVAs and ANCOVAs, one-way MANOVAs and MANCOVAs, and factorial MANOVAs and MANCOVAs.
3. Describe techniques used to predict group membership, including discriminant analysis and logistic regression.
4. Describe techniques for determining the underlying structure of a set of variables, including factor analysis and principal components analysis.
5. Develop research questions, applicable to their respective fields of study, that are appropriate for each statistical technique described.
6. Accurately identify independent and dependent variables, as well as the appropriate number of independent variable categories, for a given research question.
7. Apply the Table of Statistical Tests to determine an appropriate technique for a given research question.
8. Apply the Decision-Making Tree for Statistical Tests to determine an appropriate technique for a given research question.

One of the most difficult tasks for students conducting quantitative research is identifying the appropriate statistical technique to utilize for a particular research question. Fortunately, if an accurate and appropriate research question has been generated, the process of determining the statistical technique is really quite simple. The primary factor that determines the statistical test students should use is the variable—more specifically, the type or scale of variables (categorical or quantitative) and the number of independent and dependent variables, both of which influence the nature of the research question being posed. To facilitate this identification process, we provide two decision-making tools so that the reader may select whichever is more comfortable. The Table of Statistical Tests begins with the identification of the numbers and scales of independent and dependent variables. In contrast, the Decision-Making Tree is organized around the four different types of research questions: degree of relationship among variables, significance of group differences, prediction of group membership, and structure. This chapter presents these decision-making tools and provides an overview of the statistical techniques addressed in this text as well as basic

univariate tests, all of which will be organized by the four types of research questions. Please note that there are additional multivariate techniques that are not addressed in this book.

SECTION 2.1 DEGREE OF RELATIONSHIP AMONG VARIABLES

When investigating the relationship between two or more quantitative variables, correlation and/or regression is the appropriate test. Three statistical tests are presented that address this type of research question.

Bivariate Correlation and Regression

Bivariate correlation and regression evaluate the degree of relationship between two quantitative variables. The Pearson correlation coefficient (r), the most commonly used **bivariate correlation** technique, measures the association between two quantitative variables without distinction between the independent and dependent variables (e.g., What is the relationship between SAT scores and college freshmen GPAs?). In contrast, **bivariate regression** utilizes the relationship between the independent and dependent variables to predict the score of the dependent variable from the independent variable (e.g., To what degree do SAT scores [IV] predict college freshmen GPAs [DV]?). An overview of bivariate correlation and regression is provided in Chapter 7.

When to use bivariate correlation or regression

Number of Variables by Type	Type of Variable	Nature of Evaluation
One	IV (quantitative) →	relationship/prediction
One	DV (quantitative)	

Multiple Regression

Multiple regression identifies the best combination of predictors (IVs) of the dependent variable. Consequently, it is used when there are several independent quantitative variables and one dependent quantitative variable (e.g., Which combination of risk-taking behaviors [amount of alcohol use, drug use, sexual activity, and violence—IVs] best predicts the amount of suicide behavior [DV] among adolescents?). To produce the best combination of predictors of the dependent variable, a sequential multiple regression selects independent variables, one at a time, by their ability to account for the most variance in the dependent variable. As a variable is selected and entered into the group of predictors, the relationship between the group of predictors and the dependent variable is reassessed. When no more variables are left that explain a significant amount of variance in the dependent variable, then the regression model is complete. Multiple regression is discussed in Chapter 7.

When to use multiple regression

Number of Variables by Type	Type of Variable	Nature of Evaluation
Two or more	IVs (quantitative) →	relationship/prediction
One	DV (quantitative)	

Path Analysis

Path analysis utilizes multiple applications of multiple regression to estimate causal relations, both direct and indirect, among several variables and to test the acceptability of the causal model hypothesized by the researcher (e.g., What are the direct and indirect effects of reading ability, family income, and parents' education [IVs] on students' GPA [DV]?). Before any data analysis is conducted, the researcher must first hypothesize the causal model, which is usually based upon theory and previous research. This model is then graphically represented in a path diagram. Path coefficients are calculated to estimate the strength

of the relationships in the hypothesized causal model. A further discussion of path analysis is presented in Chapter 8.

When to use path analysis

Number of Variables by Type	Type of Variable	Nature of Evaluation
Two or more	IVs (quantitative) →	relationship/causal
One or more	DV (quantitative)	

SECTION 2.2 SIGNIFICANCE OF GROUP DIFFERENCES

A primary purpose of testing for group differences is to determine a causal relationship between the independent and dependent variables. Comparison groups are created by the categories identified in the IV(s). The number of categories in the IV, the number of IVs, and the number of DVs determine the appropriate test.

***t* Test**

The most basic statistical test that measures group differences is the ***t* test**, which analyzes significant differences between two group means. Consequently, a *t* test is appropriate when the IV is defined as having two categories and the DV is quantitative (e.g., Do males and females [IV] have significantly different SAT scores [DV]?). Further explanation of *t* tests is provided in most introductory-level statistical texts and therefore is not included in this text.

When to use a t test

Number of Variables by Type	Type of Variable	Nature of Evaluation
One	IV (2 categories)	group differences
One	DV (quantitative)	

One-Way Analysis of Variance

One-way analysis of variance (ANOVA) tests the significance of group differences between two or more means as it analyzes variation between and within each group. ANOVA is appropriate when the IV is defined as having two or more categories and the DV is quantitative (e.g., Do adolescents from low, middle, and high socioeconomic status families [IV] have different scores on an AIDS knowledge test [DV]?). Because ANOVA only determines the significance of group differences and does not identify which groups are significantly different, post hoc tests are usually conducted in conjunction with ANOVA. An overview of ANOVA is provided in Chapter 4.

When to use a one-way ANOVA

Number of Variables by Type	Type of Variable	Nature of Evaluation
One	IV (2+ categories)	group differences
One	DV (quantitative)	

One-Way Analysis of Covariance

One-way analysis of covariance (ANCOVA) is similar to ANOVA in that two or more groups are being compared on the mean of some DV, but ANCOVA additionally controls for a variable (covariate) that may influence the DV (e.g., Do preschoolers of low, middle, and high socioeconomic status [IV] have different literacy test scores [DV] after adjusting for family type [covariate]?). Many times the covariate may be pretreatment differences in which groups are equated in terms of the covariate(s). In general, ANCOVA is appropriate when the IV is defined as having two or more categories, the DV is quantitative,

and the effects of one or more covariates need to be removed. Further discussion of one-way ANCOVA is provided in Chapter 5.

When to use a one-way ANCOVA

Number of Variables by Type	Type of Variable	Nature of Evaluation
One	IV (2+ categories)	group differences
One	DV (quantitative)	
One or more	covariate	

Factorial Analysis of Variance

Factorial analysis of variance (factorial ANOVA) extends ANOVA to research scenarios with two or more IVs that are categorical (e.g., Do third grade students have different math achievement scores [DV] based upon instructional treatment [IV-treatment vs. control] and gender [IV]?). Post hoc tests are used to determine specific group differences.

When to use a factorial ANOVA

Number of Variables by Type	Type of Variable	Nature of Evaluation
Two or more	IVs (categorical) →	group differences
One	DV (quantitative)	

Factorial Analysis of Covariance

An extension of factorial ANOVA, **factorial analysis of covariance (factorial ANCOVA)** examines group differences in a single quantitative dependent variable based upon two or more categorical independent variables, while controlling for a covariate that may influence the DV (e.g., Do third grade students have different math achievement scores [DV] based upon instructional treatment [IV-treatment vs. control] and gender [IV], while controlling for pretreatment math achievement [covariate]?). Post hoc tests are again used to determine specific group differences.

When to use a factorial ANCOVA

Number of Variables by Type	Type of Variable	Nature of Evaluation
Two or more	IVs (categorical) →	group differences
One	DV (quantitative)	
One or more	covariate	

One-Way Multivariate Analysis of Variance

Similar to ANOVA in that both techniques test for differences among two or more groups as defined by a single IV, **one-way multivariate analysis of variance (MANOVA)** is utilized to simultaneously study two or more related DVs while controlling for the correlations among the DVs (Vogt, 2005). If DVs are not correlated, then it is appropriate to conduct separate ANOVAs. Because groups are being compared on several DVs, a new DV is created from the set of DVs that maximizes group differences. After this linear combination of the original DVs is created, an ANOVA is then conducted to compare groups based on the new DV. A MANOVA example follows: Does ethnicity [IV] significantly affect reading achievement, math achievement, and overall achievement [DVs] among sixth grade students? Chapter 6 discusses one-way and factorial models of MANOVA and MANCOVA.

When to use a one-way MANOVA

Number of Variables by Type	Type of Variable	Nature of Evaluation
One	IV (2+ categories)	group differences
Two or more	DVs (quantitative)	

One-Way Multivariate Analysis of Covariance

An extension of ANCOVA, *one-way multivariate analysis of covariance (MANCOVA)* investigates group differences among several DVs while also controlling for covariates that may influence the DVs (e.g., Does ethnicity [IV] significantly affect reading achievement, math achievement, and overall achievement [DVs] among sixth grade students after adjusting for family income [covariate]?).

When to use a one-way MANCOVA

Number of Variables by Type	Type of Variable	Nature of Evaluation
One	IV (2+ categories)	group differences
Two or more	DVs (quantitative)	
One or more	covariate	

Factorial Multivariate Analysis of Variance

Factorial multivariate analysis of variance (factorial MANOVA) extends MANOVA to research scenarios with two or more IVs that are categorical (e.g., Do ethnicity and learning preference [IVs] significantly affect reading achievement, math achievement, and overall achievement [DVs] among sixth grade students?). Because several independent variables are used, different combinations of DVs are created for each main effect and interaction of the IVs.

When to use a factorial MANOVA

Number of Variables by Type	Type of Variable	Nature of Evaluation
Two or more	IVs (categorical)	group differences
Two or more	DVs (quantitative)	

Factorial Multivariate Analysis of Covariance

Factorial multivariate analysis of covariance (factorial MANCOVA) extends factorial MANOVA to research scenarios that require the adjustment of one or more covariates on the DVs (e.g., Do ethnicity and learning preference [IVs] significantly affect reading achievement, math achievement, and overall achievement [DVs] among sixth grade students after adjusting for family income [covariate]?).

When to use a factorial MANCOVA

Number of Variables by Type	Type of Variable	Nature of Evaluation
Two or more	IVs (categorical) →	group differences
Two or more	DVs (quantitative)	
One or more	covariate	

SECTION 2.3 PREDICTION OF GROUP MEMBERSHIP

The primary purpose of predicting group membership is to identify specific IVs that best predict group membership as defined by the DV. Consequently, the following statistical techniques are appropriate when the DV is categorical.

Discriminant Analysis

Discriminant analysis is often seen as the reverse of MANOVA in that it seeks to identify which combination of quantitative IVs best predicts group membership as defined by a single DV that has two or more categories (e.g., Which risk-taking behaviors [amount of alcohol use, drug use, sexual activity, violence—IVs] distinguish suicide attempters from nonattempters [DV]?). In contrast, MANOVA identifies group differences on a combination of quantitative DVs. *Discriminant analysis* seeks to interpret the pattern of differences among the predictors (IVs). Consequently, the analysis will often produce several sets

or combinations of IVs that predict group membership. Each IV set, referred to as a *function*, represents a mathematical attempt to maximize a linear combination of the IVs to discriminate among groups. Discriminant analysis is best used when groups are formed naturally, based on some characteristic, and not randomly. Chapter 10 discusses discriminant analysis in further detail.

When to use discriminant analysis

Number of Variables by Type	Type of Variable	Nature of Evaluation
Two or more	IVs (quantitative)	
One	DV (2+ categories) →	group prediction

Logistic Regression

Logistic regression is similar to discriminant analysis in that both identify a set of IVs that best predicts group membership. Although SPSS provides both binary and multinomial logistic regression, our discussion will address only the binary logistic regression, in which the DV is a ***dichotomous*** (having only two categories) ***variable***. The IVs may be categorical and/or quantitative. Because the DV consists of only two categories, ***logistic regression*** estimates the probability of the DV occurring as the values of the IVs change. For instance, a research question that would utilize logistic regression is as follows: To what extent do certain risk-taking behaviors (amount of alcohol use, drug use, sexual activity, and the presence of violent behavior—IVs) increase the odds of a suicide attempt (DV) occurring? Logistic regression is discussed in Chapter 11.

When to use logistic regression

Number of Variables by Type	Type of Variable	Nature of Evaluation
Two or more	IVs (categorical/quantitative)	
One	DV (2 categories) →	group prediction

SECTION 2.4 STRUCTURE

When the researcher questions the underlying structure of an instrument or is interested in reducing the number of IVs, factor analysis and/or principal components analysis are appropriate methods. Although factor analysis and principal components analysis are different techniques, they are very similar and will be presented together under the heading of factor analysis. Both of these techniques will be discussed in Chapter 9.

Factor Analysis and Principal Components Analysis

Factor analysis allows the researcher to explore the underlying structures of an instrument or data set and is often used to develop and test a theory. These underlying structures are typically referred to as *latent factors*. Latent factors are essentially unobservable variables, traits, or characteristics. One of the important benefits of factor analysis is that researchers can try to measure variables that are unobservable (e.g., IQ). ***Principal components analysis*** is generally used to reduce the number of IVs, which is advantageous when conducting multivariate techniques in which the IVs are highly correlated. For instance, principal components analysis can reduce a 100-item instrument to 10 factors that will then be utilized as IVs in subsequent data analysis. This IV reduction can also aid the researcher in exploring, developing, and testing theories based upon how the items are grouped. Consequently, factor analysis/principal components analysis combines several related IVs into fewer, more basic underlying factors. Independent variables that share common variance are grouped together. Once factors are created, they are often adjusted (rotated) so that these factors are not highly related to one another and more accurately represent the combined IVs. Because research questions that utilize factor analysis/principal components analysis typically only address IVs, this statistical technique is not included in the Table of Statistical Tests, which relies upon the identification of both IVs and DVs. An example research question that would utilize factor analysis/principal

components analysis is as follows: What underlying structure exists among the variables of male life expectancy, female life expectancy, birth rate, infant mortality rate, fertility rate among women, number of doctors, number of radios, number of telephones, number of hospital beds, and gross domestic product?

SECTION 2.5 THE TABLE OF STATISTICAL TESTS

The Table of Statistical Tests is presented in Figure 2.1 (p. 23). This tool organizes statistical methods by the number and type (categorical vs. quantitative) of IVs and DVs. Steps for using this table are as follows:

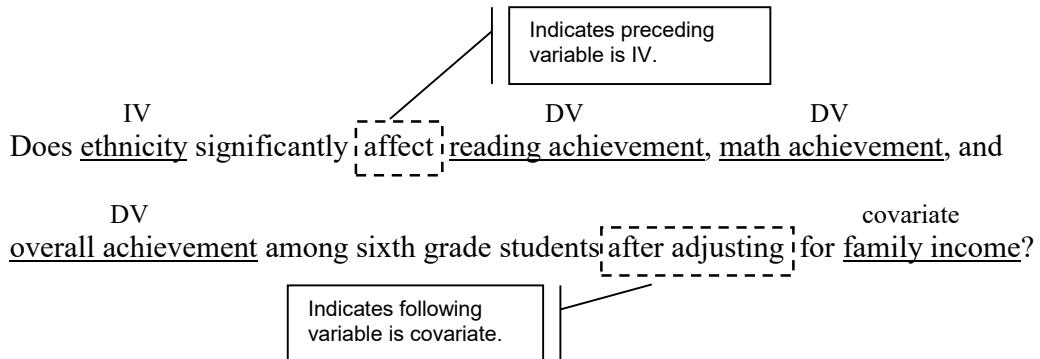
1. Identify the variables in the research question.
2. Indicate which variables are the independent and dependent variables and covariates.
3. Determine the type (categorical or quantitative) of all variables. If a variable is categorical, determine the number of categories.
4. Use the table to identify
 - the appropriate row for the IVs,
 - the appropriate column for the DVs, and
 - the row and column intersection that indicates the statistical test to be used.

These steps are applied to the following research question: Does ethnicity significantly affect reading achievement, math achievement, and overall achievement among sixth grade students after adjusting for family income?

Step 1: Identify the variables in the research question.

Does ethnicity significantly affect reading achievement, math achievement, and overall achievement among sixth grade students after adjusting for family income?

Step 2: Indicate which variables are the independent and dependent variables and covariates. It is helpful to examine the sentence order of variables because the first variables are usually the IVs. The verb of the research question can also help in the identification process.



Step 3: Determine the type (categorical or quantitative) of all variables. This is dependent upon how you decide to operationalize your variables.

(IV) 2+ categories

(DV) quantitative

(DV) quantitative

Does ethnicity significantly affect reading achievement, math achievement, and

(DV) quantitative

(covariate) quantitative

overall achievement among sixth grade students after adjusting for family income?

Consequently, this research question includes the following: one IV (2+ categories), 3 DVs (all quantitative), and one covariate (quantitative).

Step 4: Use the table to

- identify the appropriate row for the IVs
(Example: IV → categorical → one IV → 2+ categories with one covariate),
- identify the appropriate column for the DVs
(Example: DV → quantitative → several DVs), and
- identify the row and column intersection that indicates the statistical test to be used
(Example: The intersection of the preceding row and column indicates that a one-way MANCOVA should be conducted).

SECTION 2.6 THE DECISION-MAKING TREE FOR STATISTICAL TESTS

The Decision-Making Tree for Statistical Tests is presented in Figure 2.2 (p. 24). This tool organizes statistical methods by the purpose of the research question. Once the purpose has been identified, the process is then guided by the number and type of variables. Although the Decision-Making Tree begins with the purpose of the research question, we recommend first identifying the number and types of variables, as this will guide the process of determining the purpose. The steps for using the Decision-Making Tree are as follows:

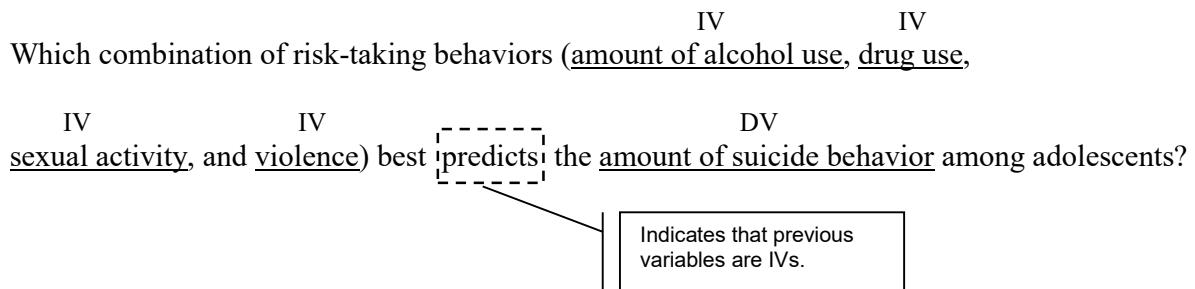
1. Identify the variables in the research question.
2. Indicate which variables are the independent and dependent variables and covariates.
3. Determine the type (categorical or quantitative) of all variables. If a variable is categorical, determine the number of categories.
4. Determine the purpose of the research question: degree of relationship, group differences, prediction of group membership, or structure. Here are a few helpful hints in using the variable information to determine the research question purpose.
 - When the IVs and DVs are all quantitative, the purpose is degree of relationship.
 - When the IVs are categorical and the DVs are quantitative, the purpose is group differences.
 - When the DVs are categorical, the purpose is predicting group membership.
5. Apply the information from the preceding steps to the Decision-Making Tree, following the process of decisions—research question, number and type of DV, number and type of IV, and covariates—to the appropriate test.

These steps are applied to the following research question: Which combination of risk-taking behaviors (amount of alcohol use, drug use, sexual activity, and violence) best predicts the amount of suicide behavior among adolescents?

Step 1: Identify the variables in the research question.

Which combination of risk-taking behaviors (amount of alcohol use, drug use, sexual activity, and violence) best predicts the amount of suicide behavior among adolescents?

Step 2: Indicate which variables are the independent and dependent variables and covariates.



Step 3: Determine the type (categorical or quantitative) of all variables.

(IV) quantitative (IV) quantitative
Which combination of risk-taking behaviors (amount of alcohol use, drug use,

(IV) quantitative (IV) quantitative (DV) quantitative
sexual activity, and violence) best predicts the amount of suicide behavior among adolescents?

Consequently, this research question includes the following: four IVs (all quantitative) and one DV (quantitative).

Step 4: Determine the purpose of the research question: degree of relationship, group differences, prediction of group membership, or structure. Because all our variables are quantitative, the purpose of the research question is degree of relationship.

Step 5: Apply the information from the preceding steps to the Decision-Making Tree: research question, number and type of DV, number and type of IV, and covariates. Continuing with the example, the decisions would be as follows:

degree of relationship → 1 DV (quant.) → 2+ IVs (quant.) → multiple regression

SUMMARY

Determining the appropriate statistical technique relies upon the identification of the type of variables (categorical or quantitative) and the number of IVs and DVs, all of which influence the nature of the research questions being posed. This chapter introduced the statistical tests to be presented in the upcoming chapters. The statistical methods are organized under four purposes of research questions: degree of relationship, significance of group differences, prediction of group membership, and structure. Statistical tests that analyze the degree of relationship include bivariate correlation and regression, multiple regression, and path analysis. Research questions addressing degree of relationship all have quantitative variables. Methods that examine the significance of group differences are *t* test, one-way and factorial ANOVA, one-way and factorial ANCOVA, one-way and factorial MANOVA, and one-way and factorial MANCOVA. Research questions that address group differences have categorical IVs. Statistical tests that predict group membership are discriminant analysis and logistic regression. Research questions that address prediction of group membership have a categorical DV. Statistical tests that address the purpose of structure are factor analysis and principal components. Questions that address structure usually do not distinguish between independent and dependent variables.

Two decision-making tools are provided to assist in identifying which statistical method to utilize—the Table of Statistical Tests and the Decision-Making Tree for Statistical Tests. The Table of Statistical Tests is organized by the type and number of IVs and DVs, while the Decision-Making Tree for Statistical Tests is organized by the purpose of the research question.

KEYWORDS

- bivariate correlation
- bivariate regression
- dichotomous variable
- discriminant analysis
- factor analysis
- factorial analysis of covariance (factorial ANCOVA)
- factorial analysis of variance (factorial ANOVA)
- factorial multivariate analysis of covariance (factorial MANCOVA)
- factorial multivariate analysis of variance (factorial MANOVA)
- logistic regression
- multiple regression
- one-way analysis of covariance (ANCOVA)
- one-way analysis of variance (ANOVA)
- one-way multivariate analysis of covariance (MANCOVA)
- one-way multivariate analysis of variance (MANOVA)
- path analysis
- principal components analysis
- *t* test

Figure 2.1. Table of Statistical Tests.

INDEPENDENT VARIABLE(S)		DEPENDENT VARIABLE(S)	
Categorical		Quantitative	
Quantitative	Categorical	One DV	Several DVs
2 categories	2 categories	<i>t</i> Test	One-way MANOVA
2+ categories	2+ categories	One-way ANOVA	One-way MANOVA
One IV	One IV	One-way ANCOVA	One-way MANCOVA
Several IVs	No covariate	Factorial ANOVA	Factorial MANOVA
With covariate	With covariate	Factorial ANCOVA	Factorial MANCOVA
One IV	One IV	Bivariate Correlation	
Several IVs	Several IVs	Bivariate Regression	
	Discriminant Analysis	Multiple Regression	
	Logistic Regression	Path Analysis	Path Analysis

Figure 2.2. Decision-Making Tree for Statistical Tests.

Research Question	Number & Type of DVs	Number & Type of IVs	Covariates	Test	Goal of Analysis
Degree of Relationship	1 quantitative	1 quantitative	1 quantitative	Bivariate Correlation and/or Regression	Determine relationship and prediction
	1 quantitative	2+ quantitative		Multiple Regression	Create linear combination that best predicts DV
	1+ quantitative	2+ quantitative		Path Analysis	Estimate causal relations among variables in a hypothesized model
Group Differences	1 quantitative	1 categorical (2 categories)		<i>t</i> Test	
	1 quantitative	1 categorical (2+ categories)	None	One-way ANOVA	Determine significance of mean group differences
			Some	One-way ANCOVA	
	2+ categorical	None	Factorial ANOVA		
		Some	Factorial ANCOVA		
	2+ quantitative	1 categorical	None	One-way MANOVA	
		Some	One-way MANCOVA		
		2+ categorical	None	Factorial MANOVA	Create linear combo of DVs to maximize mean group differences
			Some	Factorial MANCOVA	
Prediction of Group Membership	1 categorical (2 categories)	2+ mixed		Logistic Regression	Create linear combo of IVs of the log of odds of being in one group
	1 categorical (2+ categories)	2+ quantitative		Discriminant Analysis	Create best linear combo to predict group membership
Structure	3+ quantitative			Factor Analysis (theoretical)	
				Principal Components (empirical)	Create linear combinations of observed variables to represent latent variable

Exercises for Chapter 2

Directions: The research questions that follow are used as examples throughout this chapter. Identify the appropriate statistical test for each question by using both decision-making tools. Determine the tool with which you are most comfortable.

1. To what degree do SAT scores predict college freshmen GPAs?
2. Does ethnicity significantly affect reading achievement, math achievement, and overall achievement among sixth grade students?
3. What are the causal effects (direct and indirect) among number of school absences due to illness, reading ability, semester GPA, and total score on the Iowa Test of Basic Skills among eighth grade students?
4. Do males and females have significantly different SAT scores?
5. What is the relationship between SAT scores and college freshmen GPAs?
6. Which risk-taking behaviors (amount of alcohol use, drug use, sexual activity, violence) distinguish suicide attempters from nonattempters?
7. Do adolescents from low, middle, and high socioeconomic status families have different literacy test scores after adjusting for family type?
8. Does ethnicity significantly affect reading achievement, math achievement, and overall achievement among sixth grade students after adjusting for family income?
9. Which combination of risk-taking behaviors (amount of alcohol use, drug use, sexual activity, and violence) best predicts the amount of suicide behavior among adolescents?
10. Do preschoolers of low, middle, and high socioeconomic status have different literacy test scores?
11. Do ethnicity and learning preference significantly affect reading achievement, math achievement, and overall achievement among sixth grade students?

12. To what extent do certain risk-taking behaviors (amount of alcohol use, drug use, and sexual activity, and the presence of violent behavior) increase the odds of a suicide attempt occurring?
13. Do ethnicity and learning preference significantly affect reading achievement, math achievement, and overall achievement among sixth grade students after adjusting for family income?
14. What underlying structure exists among the following variables: amount of alcohol use, drug use, sexual activity, school misconduct, cumulative GPA, reading ability, and family income?

CHAPTER 3

PRE-ANALYSIS DATA SCREENING

STUDENT LEARNING OBJECTIVES

After studying Chapter 3, students will be able to:

1. Discuss the importance of screening data prior to any substantive data analysis.
2. Describe four main purposes for screening data.
3. Present various alternatives for handling missing data.
4. Distinguish between univariate and multivariate outliers.
5. Interpret the proper application of Mahalanobis distance when evaluating potential outliers.
6. Compare and contrast various forms of departure from normality in a variable's (or set of variables') distribution.
7. Discuss the purpose and use of data transformations.
8. Describe the use of residuals in evaluating violations to the assumption of linearity.
9. Explain how the assumption of homoscedasticity is assessed and violations are determined.
10. Test data set(s) for various pre-analysis conditions by following the appropriate SPSS guidelines provided.

In this chapter, we discuss several issues related to the quality of data that a researcher wishes to subject to a multivariate analysis. These issues must be carefully considered and addressed *prior* to the actual statistical analysis—they are essentially an analysis *within* the analysis! Only after these quality assurance issues have been examined can the researcher be confident that the main analysis will be an honest one, which will ultimately result in valid conclusions being drawn from the data.

SECTION 3.1 WHY SCREEN DATA?

There are four main purposes for screening data prior to conducting a multivariate analysis. The first of these deals with the accuracy of the data collected. Obviously, the results of any statistical analysis are only as good as the data analyzed. If inaccurate data are used, the computer program will run the analysis (in all likelihood), and the researcher will obtain her output. However, the researcher will not be able to discern the extent to which the results are valid simply by examining the output—the results will appear to be legitimate (i.e., values for test statistics will appear, accompanied by significance values, etc.). The researcher will then proceed to interpret the results and draw conclusions. However, unknown to her, they will be erroneous conclusions because they will have been based on the analysis of inaccurate data.

With a small data file, simply printing the entire data set and proofreading it against the actual data is an easy and efficient method of determining the accuracy of data. This can be accomplished using the

SPSS List procedure. However, if the data set is rather large, this process would be overwhelming. In this case, examination of the data using frequency distributions and descriptive statistics is a more realistic method. Both frequency distributions and descriptive statistics can be obtained by using the **SPSS Frequencies** procedure. For quantitative variables, a researcher might examine the range of values to be sure that no cases have values outside the range of possible values. Assessment of the means and standard deviations (i.e., are they plausible?) is also beneficial. For categorical variables, the researcher should also make sure that all cases have values that correspond to the coded values for the possible categories.

The second purpose deals with missing data and attempts to assess the effect of and ways to deal with incomplete data. Missing data occur when measurement equipment fails, participants do not complete all trials or respond to all items, or errors occur during data entry. The amount of missing data is less crucial than the pattern of missing data (Tabachnick & Fidell, 2007). Missing values that are randomly scattered throughout a data set sometimes are not serious because their pattern is random. Nonrandom missing data, on the other hand, create problems with respect to the generalizability of the results. Because these missing values are nonrandom, there is likely some underlying reason for their occurrence. Unfortunately, there are no firm guidelines for determining what quantity of missing data is too much for a given sample size. Those decisions still rest largely on the shoulders of the researcher. Methods for dealing with missing data are discussed in Section 3.2.

The third purpose deals with assessing the effects of extreme values (i.e., *outliers*) on the analysis. Outliers are cases with such extreme values on one variable or on a combination of variables that they distort the resultant statistics. Outliers often create critical problems in multivariate data analyses. There are several causes for a case to be defined as an extreme value, some of which are far more serious than others. These various causes and methods for addressing each will be discussed in Section 3.3.

Finally, all multivariate statistical procedures are based to some degree on assumptions. The fourth purpose of screening data is to assess the adequacy of fit between the data and the assumptions of a specific procedure. Some multivariate procedures have unique assumptions (which will be discussed in other chapters) upon which they are based, but nearly all techniques include three basic assumptions: **normality**, **linearity**, and **homoscedasticity**. These assumptions will be defined and methods for assessing the adequacy of the data with respect to each will be discussed in Sections 3.4, 3.5, and 3.6, respectively. Techniques for implementing these methods using SPSS will be described in Sections 3.7 and 3.8.

SECTION 3.2 MISSING DATA

Many researchers tend to assume that any missing data that occur within their data sets are random in nature. This may or may not be the case. If it is not the case, serious problems can arise when trying to generalize to the larger population from which the sample was obtained. The best thing to do when a data set includes missing data is to examine it. Using data that are available, a researcher should conduct tests to see if patterns exist in the missing data. To do so, one could create a dichotomous dummy variable, coded so that one group includes cases with values on a given variable and the other group contains cases with missing values on that variable. For instance, if respondents on an attitudinal survey are asked to provide their income and many do not supply that information (for reasons unknown to us at this time), those who provided an income level would be coded 0 and those who did not would be coded 1. Then the researcher could run a simple independent samples *t* test to determine if there are significant mean differences in attitude between the two groups. If significant differences do exist, there is an indication that those who did not provide income information possess attitudes different from those who did report their income. In other words, a pattern exists in the missing responses.

If a researcher decides that the missing data are important and need to be addressed, there are several alternative methods for handling these data. (For a discussion on additional techniques when there are missing data, the reader is advised to refer to Tabachnick and Fidell, 2007.) The first of these alternatives involves deleting the cases or variables that have created the problems. Any case that has a missing value is

simply dropped from the data file. If only a few cases have missing values, this is a good alternative. Another option involves a situation where the missing values may be concentrated to only a few variables. In this case, an entire variable may be dropped from the data set, provided it is not central to the main research questions and subsequent analysis. However, if missing values are scattered throughout the data and are abundant, deletion of cases and/or variables may result in a substantial loss of data, in the form of either participants or measures. Sample size may begin to decrease rather rapidly and, if the main analysis involves group comparisons, some groups may approach dangerously low sample sizes that are inappropriate for some multivariate analyses.

A second alternative for handling missing data is to estimate the missing values and then use these values during the main analysis. There are three main methods of estimating missing values. The first of these is for the researcher to use *prior knowledge*, or a well-educated guess, for a replacement value. This method should be used only when a researcher has been working in the specific research area for quite some time and is very familiar with the variables and the population being studied.

Another method of estimating missing values involves the calculation of the means, using available data, for variables with missing values. Those mean values are then used to replace the missing values prior to the main analysis. When no other information is available to the researcher, the mean is the best estimate for the value on a given variable. This is a somewhat conservative procedure because the overall mean does not change by inserting the mean value for a case and no guessing on the part of the researcher is required. However, the variance is reduced somewhat because the “real” value probably would not have been precisely equal to the mean. This is usually not a serious problem unless there are numerous missing values. In this situation, a possible concession is to insert a group mean, as opposed to the overall mean, for a missing value. This procedure is more appropriate for situations involving group comparison analyses.

Finally, a third alternative for handling missing data also estimates the missing value, but it does so using a *regression* approach. Regression is discussed extensively in Chapter 7. In regression, several IVs are used to develop an equation that can be used to predict the value of a DV. For missing data, the variable with missing values becomes the DV. Cases with complete data are used to develop this prediction equation. The equation is then used to predict missing values of the DV for incomplete cases. An advantage to this procedure is that it is more objective than a researcher’s guess and factors in more information than simply inserting the overall mean. One disadvantage of regression is that the predicted scores are better than they actually would be. Because the predicted values are based on other variables in the data set, they are more consistent with those scores than a real score would be. Another disadvantage of regression is that the IVs must be good predictors of the DV in order for the estimated values to be accurate. Otherwise, this amounts to simply inserting the overall mean in place of the missing value (Tabachnick & Fidell, 2007).

If any of the previous methods are used to estimate missing values, a researcher should consider repeating the analysis using only complete cases (i.e., conduct the main analysis with the missing values and repeat the analysis with no missing values). If the results are similar, one can be confident in the results. However, if they are different, an examination of the reasons for the differences should be conducted. The researcher should then determine which of the two represents the real world more accurately, or consider reporting both sets of results.

SECTION 3.3 OUTLIERS

Cases with unusual or extreme values at one or both ends of a sample distribution are known as *outliers*. There are three fundamental causes for outliers: (1) data-entry errors were made by the researcher, (2) the participant is not a member of the population for which the sample is intended, or (3) the participant is simply different from the remainder of the sample (Tabachnick & Fidell, 2007).

The problem with outliers is that they can distort the results of a statistical test. This is due largely to the fact that many statistical procedures rely on squared deviations from the mean (Aron, Aron, & Coups, 2006). If an observation is located far from the rest of the distribution (and, therefore, far from the mean), the value of its deviation is large. Imagine by how much a deviation increases when squared! Generally

speaking, statistical tests are quite sensitive to outliers. An outlier can exert a great deal of influence on the results of a statistical test. A single outlier, if extreme enough, can cause the results of a statistical test to be significant when, in fact, they would not have been if they had been based on all values other than the outlier. The complementary situation can also occur: An outlier can cause a result to be insignificant when, without the outlier, it would have been significant. Similarly, outliers can seriously affect the values of correlation coefficients. It is vital that the results of researchers' statistical analyses represent the majority of the data and are not largely influenced by one, or a few, extreme observations. It is for this reason that it is crucial for researchers to be able to identify outliers and decide how to handle them (Stevens, 2001).

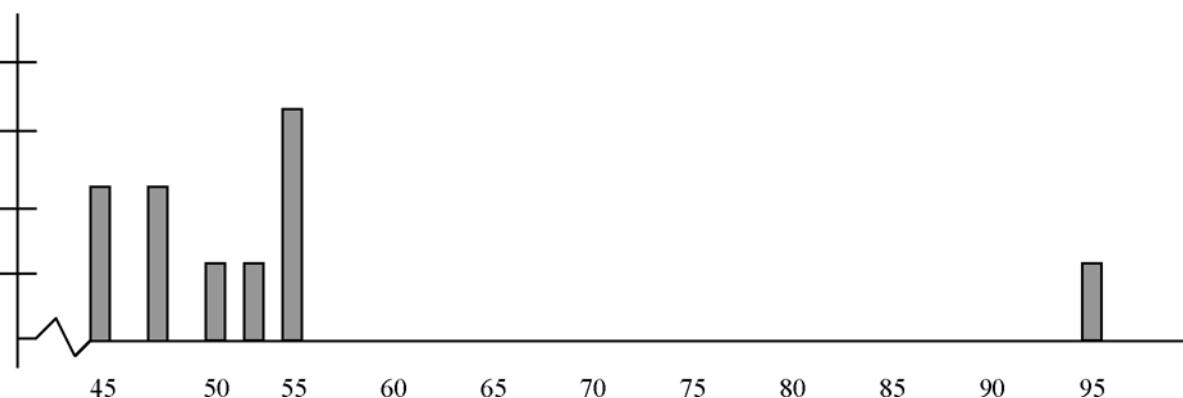
Outliers can exist in both univariate and multivariate situations, among dichotomous and **continuous variables**, and among IVs as well as DVs (Tabachnick & Fidell, 2007). **Univariate outliers** are cases with extreme values on one variable. **Multivariate outliers** are cases with unusual combinations of scores on two or more variables. With data sets consisting of a small number of variables, detection of univariate outliers is relatively simple. It can be accomplished by visually inspecting the data, either by examining a frequency distribution or by obtaining a histogram and looking for unusual values. One would simply look for values that appear far from the others in the data set. In Figure 3.1, Case 3 would clearly be identified as an outlier because it is located far from the rest of the observations.

Figure 3.1. (a) Sample Data Set, and (b) Corresponding Histogram Indicating One Outlier.

(a)

Case No.	X_1
1	55
2	48
3	95
4	48
5	51
6	55
7	45
8	53
9	55
10	45

(b)



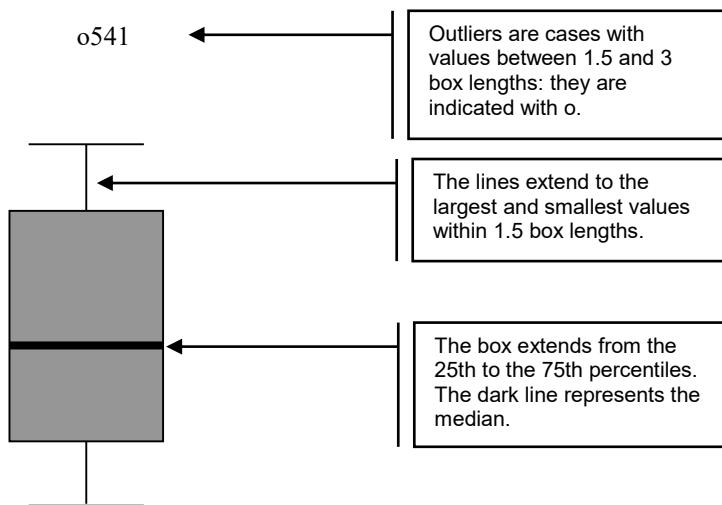
Univariate outliers can also be detected through statistical methods by standardizing all raw scores in the distribution. This is most easily accomplished by transforming the data into z -scores. If a normal distribution is assumed, approximately 99% of the scores will lie within 3 standard deviations of the mean. Therefore, any z value greater than +3.00 or less than -3.00 indicates an unlikely value, and the case should

be considered an outlier. However, with large sample sizes (i.e., $n > 100$), it is likely that a few participants could have z -scores in excess of ± 3.00 . In this situation, the researcher might want to consider extending the rule to $z > +4.00$ and $z < -4.00$ (Stevens, 2001). For small sample sizes (i.e., $n \leq 10$), any data point with a z value greater than 2.50 should be considered a possible outlier.

Univariate outliers can also be detected using graphical methods (Tabachnick & Fidell, 2007). Box plots literally “box in” cases that are located near the median value. Extreme values are located far away from the box. Figure 3.2 presents a sample box plot.

As shown in Figure 3.2, the box portion of the plot extends from the 25th to the 75th percentile, with the dark line representing the median value for the distribution. The lines above and below the box include all values within 1.5 box lengths. Cases with values between 1.5 and 3 box lengths from the upper or lower edges of the box are outliers and are designated by a small circle (o). The specific case number is also listed next to the symbol. Although not depicted in the figure, cases with values greater than 3 box lengths from the edges are also identified and designated with asterisks (*) and the specific case number. It should be noted that these designations are conventions within SPSS.

Figure 3.2. Sample Box Plot Indicating One Outlier.



Multivariate outliers, as previously stated, consist of unusual combinations of scores on two or more variables. Individual z -scores may not indicate that the case is a univariate outlier (i.e., for each variable, the value is within the expected range), but the combination of variables clearly separates the particular case from the rest of the distribution. Multivariate outliers are more subtle and, therefore, more difficult to identify, especially by using any of the previously mentioned techniques. Fortunately, a statistical procedure (known as **Mahalanobis distance**) exists that can be used to identify outliers of any type (Stevens, 2001). Mahalanobis distance is defined as the distance of a case from the centroid of the remaining cases, where the centroid is the point created by the means of all the variables (Tabachnick & Fidell, 2007).

For multivariate outliers, Mahalanobis distance is evaluated as a chi-square (χ^2) statistic with degrees of freedom equal to the number of variables in the analysis (Tabachnick & Fidell, 2007). The accepted criterion for outliers is a value for Mahalanobis distance that is significant beyond $p < .001$, determined by comparing the obtained value for Mahalanobis distance to the chi-square critical value.

Once outliers have been identified, it is necessary to investigate them further. First, the researcher must determine whether the outlier was due to an error in data entry. In this situation, of course, the value would be corrected and the data reanalyzed. However, if the researcher determines that the extreme value was correctly entered and that it may be due to an instrumentation error or due to the fact that the participant is simply different from the rest of the sample, then it is appropriate to drop the case from the analysis. If it cannot be determined that either of these situations resulted in the extreme value, one should not drop the

case from the analysis, but rather should consider reporting two analyses—one with the outlying case included and the other after the case has been deleted (Stevens, 2001). Remember that outliers should not be viewed as being “bad” because they often represent interesting cases. Care must be taken so that the outlying case is not automatically dropped from the analysis. The case and its value(s) on the variable(s) may be perfectly legitimate.

If the researcher decides that a case with unusual values is legitimate and should remain in the sample, steps may be taken to reduce the relative influence of those cases. Variables may be transformed (i.e., the scales may be changed so that the distribution appears more normal), thus reducing the impact of extreme values. Data transformations are discussed in greater detail in the next section. For a more thorough discussion of variable transformations, refer to Johnson and Wichern (2008), Stevens (2001), and Tabachnick and Fidell (2007).

SECTION 3.4 NORMALITY

As previously mentioned, there are three general assumptions involved in multivariate statistical testing: normality, linearity, and homoscedasticity. There are consequences for applying statistical analyses—particularly inferential testing—to data that do not conform to these assumptions. If one or more assumptions are violated, the results of the analysis may be biased (Kennedy & Bush, 1985). It is critical, then, to assess the extent to which the sample data meet the assumptions. The issue at hand is one of test robustness. **Robustness** refers to the relative insensitivity of a statistical test to violations of the underlying inferential assumptions. In other words, it is the degree to which a statistical test is still appropriate to apply when some of its assumptions are not met:

If in the presence of marked departures from model assumptions, little or no discrepancy between nominal and actual levels of significance occurs, then the statistical test is said to be robust with respect to that particular violation (Kennedy & Bush, 1985, p. 144).

The first of these assumptions is that of a normal sample distribution. Prior to examining multivariate normality, one should first assess univariate normality. Univariate normality refers to the extent to which all observations in the sample for a given variable are distributed normally. There are several ways, both graphical and statistical, to assess univariate normality. A simple graphical method involves the examination of the histogram for each variable. Although somewhat oversimplified, this does give an indication as to whether or not normality might be violated. One of the most popular graphical methods is the *normal probability plot*. In a normal probability plot, also known as a *normal Q-Q plot*, the observations are arranged in increasing order of magnitude and plotted against the expected normal distribution values (Stevens, 2001). The plot shows the variable’s observed values along the *x*-axis and the corresponding predicted values from a standard normal distribution along the *y*-axis (Norusis, 1998). If normality is defensible, the plot should resemble a straight line.

Among the statistical options for assessing univariate normality are the use of skewness and kurtosis coefficients. **Skewness** is a quantitative measure of the degree of symmetry of a distribution about the mean. **Kurtosis** is a quantitative measure of the degree of peakedness of a distribution. A variable can have significant skewness, kurtosis, or both. When a distribution is normal, the values for skewness and kurtosis are both equal to zero.¹ If a distribution has a positive skew (i.e., a skewness value > 0), there is a clustering of cases to the left, and the right tail is extended with only a small number of cases. In contrast, if a distribution has a negative skew (i.e., a skewness value < 0), there is a clustering of cases to the right, and the left tail is extended with only a small number of cases. Values for kurtosis that are positive indicate that the distribution is too peaked with long, thin tails (a condition known as *leptokurtosis*). Kurtosis values that are negative indicate that the distribution is too flat, with many cases in the tails (a condition known as *platykurtosis*). Significance tests for both skewness and kurtosis values should be evaluated at an alpha level of .01 or

¹ The mathematical equation for kurtosis gives a value of 3 when the distribution is normal, but statistical packages subtract 3 before printing so that the expected value is equal to zero.

.001 for small to moderate sample sizes, using a table of critical values for skewness and kurtosis, respectively. Larger samples may show significant skewness and/or kurtosis values, but they often do not deviate enough from normal to make a meaningful difference in the analysis (Tabachnick & Fidell, 2007).

Another specific statistical test used to assess univariate normality is the **Kolmogorov-Smirnov statistic**, with Lilliefors significance level. The Kolmogorov-Smirnov statistic tests the null hypothesis that the population is normally distributed. A rejection of this null hypothesis based on the value of the Kolmogorov-Smirnov statistic and associated observed significance level serves as an indication that the variable is not normally distributed.

Multivariate normality refers to the extent to which all observations in the sample for all combinations of variables are distributed normally. Similar to the univariate examination, there are several ways, both graphical and statistical, to assess multivariate normality. It is difficult to completely describe multivariate normality but suffice it to say, “normality on each of the variables separately is a necessary but not sufficient condition for multivariate normality to hold” (Stevens, 2001). Because univariate normality is a necessary condition for multivariate normality, it is recommended that all variables be assessed based on values for skewness and kurtosis, as previously described.

Characteristics of multivariate normality include the following:

1. Each of the individual variables must be normally distributed.
2. Any linear combination of the variables must be normally distributed.
3. All subsets of the set of variables (i.e., every pairwise combination) must have a multivariate normal distribution (this is known as *bivariate normality*).

Bivariate normality implies that the scatterplots for each pair of variables will be elliptical. An initial check for multivariate normality would consist of an examination of all bivariate scatterplots to check that they are approximately elliptical (Stevens, 2001). A specific graphical test for multivariate normality exists, but it requires that a special computer program be written because it is not available in standard statistical software packages (Stevens, 2001).

If the researcher determines that the data have substantially deviated from normal, he or she can consider transforming the data. **Data transformations** involve the application of mathematical procedures to the data in order to make them appear more normal. Once data have been transformed, provided all other assumptions have been met, the results of the statistical analyses will be more accurate. It should be noted that there is nothing unethical about transforming data. Transformations are nothing more than a reexpression of the data in different units (Johnson & Wichern, 2008). The transformations are performed on every participant in the data set, so the order and relative position of observations are not affected (Aron, Aron, & Coups, 2006).

A variety of data transformations exist, depending on the shape (i.e., extent of deviation from normal) of the original raw data. For instance, if a distribution differs only moderately from normal, a square root transformation should be tried initially. If the deviation is more substantial, a log transformation is obtained. Finally, if a distribution differs severely, an inverse transformation is tried. The direction of the deviation must also be considered. The above transformations are appropriate for distributions with positive skewness. If the distribution has a negative skew, the appropriate strategy is to reflect the variable and then apply the transformation procedure listed above. Reflection involves finding the largest score in the distribution and adding 1 to it to form a constant that is larger than any score in the distribution. A new variable is then created by subtracting each score from the constant. In effect, this process converts a distribution with negative skewness to one with positive skewness. It should be noted that interpretation of the results of analyses of this variable must also be reversed (Tabachnick & Fidell, 2007). Transformations can be easily obtained in various statistical packages, including SPSS. The transformations discussed here, and the SPSS language for the computation of new variables, are summarized in Figure 3.3.

Once variables have been transformed, it is important to reevaluate the normality assumption. Following the confirmation of a normal or near-normal distribution, the analysis may proceed typically, resulting in vastly improved results (Tabachnick & Fidell, 2007). In addition, the researcher should be cognizant

of the fact that any transformations performed on the data must be discussed in the Methods section of any research report.

It should be understood that the topic of data transformation is much too broad to be adequately addressed here. Should one require further details and examples of these various transformations, refer to Tabachnick and Fidell (2007).

Figure 3.3. Summary of Common Data Transformations to Produce Normal Distributions.

Original Shape	Transformation	SPSS Compute Language
Moderate positive skew	Square root	NEWX = SQRT(X)
Substantial positive skew	Logarithm	NEWX = LG10(X)
– With value ≤ 0	Logarithm	NEWX = LG10(X + C) ^a
Severe positive skew	Inverse	NEWX = 1/X
– With value ≤ 0	Inverse	NEWX = 1/(X + C) ^a
Moderate negative skew	Reflect & square root	NEWX = SQRT(K - X) ^b
Substantial negative skew	Reflect & logarithm	NEWX = LG10(K - X) ^b
Severe negative skew	Reflect & inverse	NEWX = 1/(K - X) ^b

^a C = a constant added to each score in order to bring the smallest value to at least 1.

^b K = a constant from which each score is subtracted so that the smallest score equals 1.

SECTION 3.5 LINEARITY

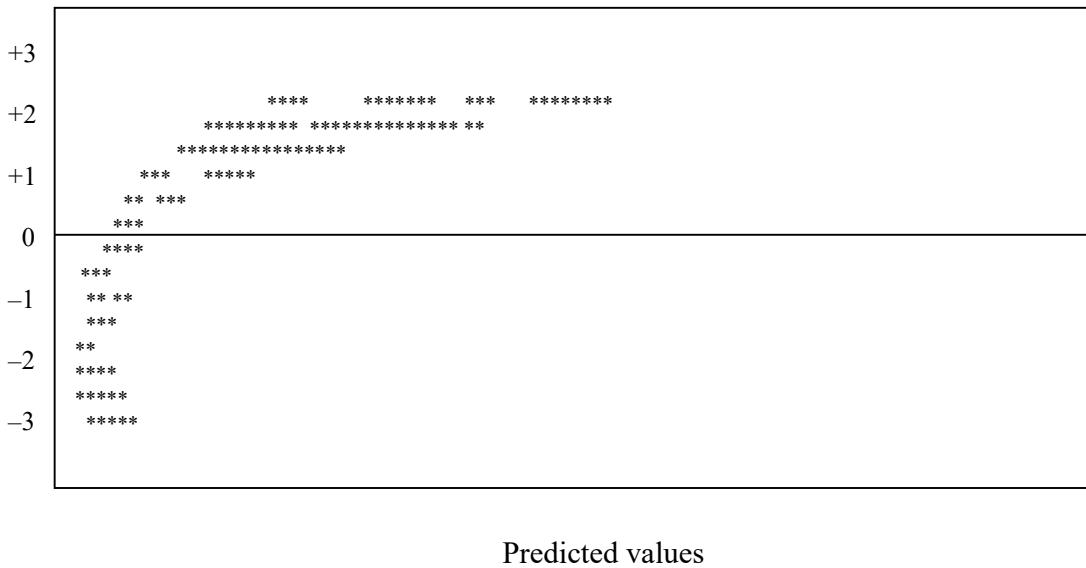
The second assumption, that of **linearity**, presupposes that there is a straight-line relationship between two variables. These two variables can be individual raw data variables (e.g., *drug dosage* and *length of illness*) or combinations of several raw data variables (i.e., *composite* or *subscale scores*, such as eight items additively combined to arrive at a score for “self-esteem”). The assumption of linearity is important in multivariate because many of the analysis techniques are based on linear combinations of variables. Furthermore, statistical measures of relationship such as Pearson’s r capture only linear relationships between variables and ignore any substantial nonlinear relationships that may exist (Tabachnick & Fidell, 2007).

There are essentially two methods of assessing the extent to which the assumption of linearity is supported by data. In analyses that involve predicted variables (e.g., multiple regression as presented in Chapter 7), nonlinearity is determined through the examination of residuals plots. **Residuals** are defined as the portions of scores not accounted for by the multivariate analysis. They are also referred to as *prediction errors* because they serve as measures of the differences between obtained and predicted values on a given variable. If standardized residual values are plotted against the predicted values, nonlinearity will be indicated by a curved pattern to the points (Norusis, 1998). In other words, residuals will fall above the zero line for some predicted values and below the line for other predicted values (Tabachnick & Fidell, 2007). Therefore, a relationship that does not violate the linearity assumption will be indicated by the points clustering around the zero line. Both a nonlinear and a linear relationship are depicted in Figure 3.4(a).

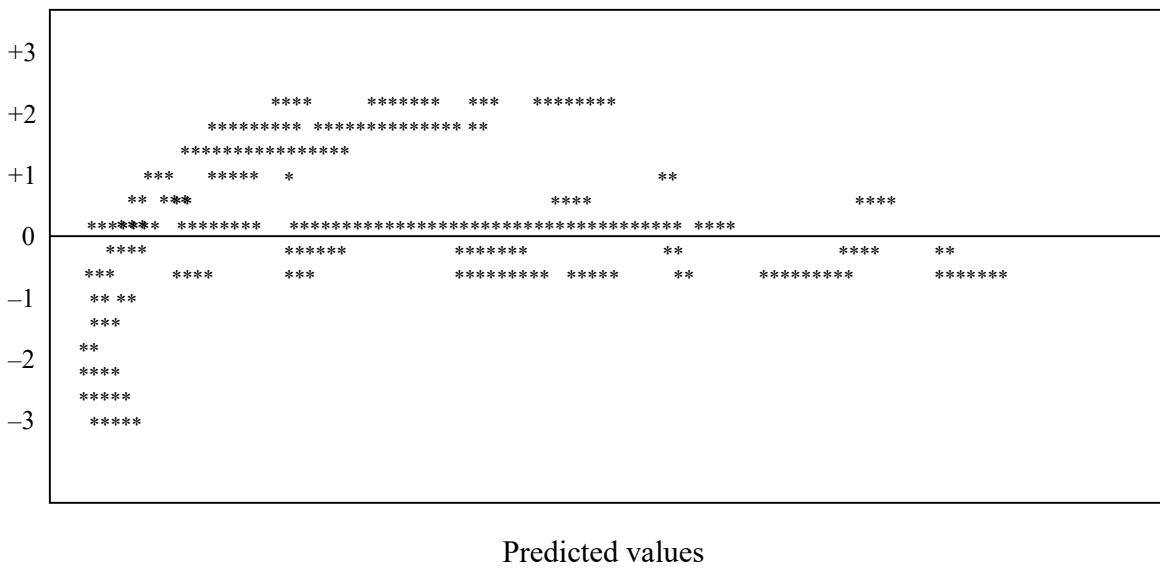
A second, and more crude, method of assessing linearity is accomplished by inspection of bivariate scatterplots. If both variables are normally distributed and linearly related, the shape of the scatterplot will be elliptical. If one of the variables is not normally distributed, the relationship will not be linear, and the scatterplot between the two variables will not be oval-shaped. Assessing linearity by means of bivariate scatterplots is an extremely subjective procedure, at best. The process can become even more cumbersome when data sets with numerous variables are being examined. In situations where nonlinearity between variables is apparent, the data can once again be transformed in order to enhance the linear relationship.

Figure 3.4. Sample Standardized Residuals Plots Showing a Strong Nonlinear Relationship (a) and a Linear Relationship (b).

(a) nonlinear relationship



(b) linear relationship



SECTION 3.6 HOMOSCEDASTICITY

The third and final assumption is that of homoscedasticity. **Homoscedasticity** is the assumption that the variability in scores for one continuous variable is roughly the same at all values of another continuous variable. This concept is analogous to the univariate assumption of homogeneity of variance (i.e., the variability in a continuous dependent variable is expected to be roughly consistent at all levels of the independent, or discrete grouping, variable). In the univariate case, homogeneity of variances is assessed statistically with **Levene's test**. This statistic provides a test of the hypothesis that the samples come from populations

with the same variances. If the observed significance level for Levene's test is small (i.e., $p < .05$), one should reject the null hypothesis that the variances are equal. It should be noted that a violation of this assumption, based on a "reject" decision of Levene's test, is not fatal to the analysis. Furthermore, Levene's test provides a sound means for assessing univariate homogeneity because it is not affected by violations of the normality assumption (Kennedy & Bush, 1985).

Homoscedasticity is related to the assumption of normality because if the assumption of multivariate normality is met, the two variables must be homoscedastic (Tabachnick & Fidell, 2007). The failure of the relationship between two variables to be homoscedastic is caused either by the nonnormality of one of the variables or by the fact that one of the variables may have some sort of relationship to the transformation of the other variable (Tabachnick & Fidell, 2007). Errors in measurement, which are greater at some levels of the independent variable than at others, may also cause a lack of homoscedasticity.

Heteroscedasticity, or the violation of the assumption of homoscedasticity, can be assessed through the examination of bivariate scatterplots. Within the scatterplot, the collection of points between variables should be approximately the same width across all values with some bulging toward the middle. Although subjective in nature, homoscedasticity is best assessed through the examination of bivariate scatterplots. In multivariate situations, homoscedasticity can be assessed statistically by using **Box's M test for equality of variance-covariance matrices**. This test allows the researcher to evaluate the hypothesis that the covariance matrices are equal. If the observed significance level for Box's M test is small (i.e., $p < .05$), one should reject the null hypothesis that the covariance matrices are equal. It should be noted, however, that Box's M test is very sensitive to nonnormality. Thus, one may reject the assumption that covariance matrices are equal due to a lack of multivariate normality, not because the covariance matrices are different (Stevens, 2001). Therefore, it is recommended that the tenability of the multivariate normality assumption be assessed prior to examining the results of the Box's M test as a means of assessing possible violations of the assumption of homoscedasticity. Violations of this assumption can be corrected by transformation of variables. However, it should be noted that a violation of the assumption of homoscedasticity, similar to a violation of homogeneity, will not prove fatal to an analysis (Kennedy & Bush, 1985; Tabachnick & Fidell, 2007). The linear relationship will still be accounted for, although the results will be greatly improved if the heteroscedasticity is identified and corrected (Tabachnick & Fidell, 2007).

Because screening data prior to multivariate analysis requires univariate screening, we have provided two univariate examples and one multivariate example.

SECTION 3.7 USING SPSS TO EXAMINE DATA FOR UNIVARIATE ANALYSIS

The following univariate examples explain the steps for using SPSS to examine missing values, outliers, normality, linearity, and homoscedasticity for both grouped data and ungrouped data. Both examples utilize the data set *career-a.sav* from the website that accompanies this book (see p. *xiii*).

Univariate Example With Grouped Data

Suppose one is interested in investigating income (*rincom91*) differences between individuals who are either satisfied or not satisfied with their job (*satjob2*). Because this research question compares groups, screening procedures must also examine data for each group.

Before the screening of data begins, we must first address a coding problem within the variable *rincom91*. This variable represents income levels ranging from 1 to 21. However, 22 represents "refusal to report," and 98 and 99 represent "not applicable." Because these values could be misinterpreted as income levels, they should be recoded as missing values. To do so, open the following menus:

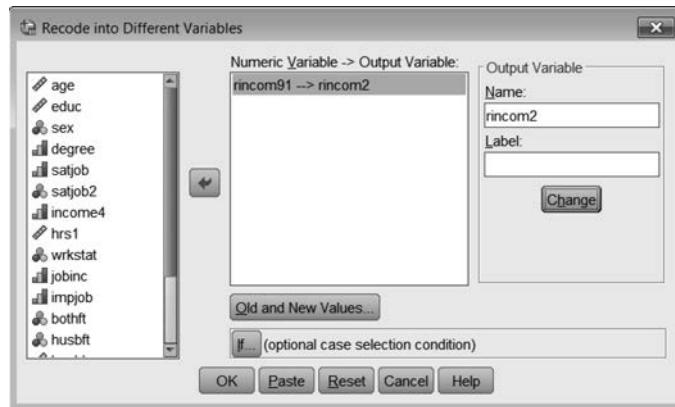
Transform

Recode Into Different Variable

Recode into Different Variables dialog box (see Figure 3.5)²

We recommend recoding *rincom91* into a different variable because this provides a record of both the original and altered variables. (Because variables may be transformed numerous times, we will name our new variable *rincom2*.) Once in this dialog box, indicate the new name for the variable, then click **Change**. Then click **Old and New Values** to specify the transformations.

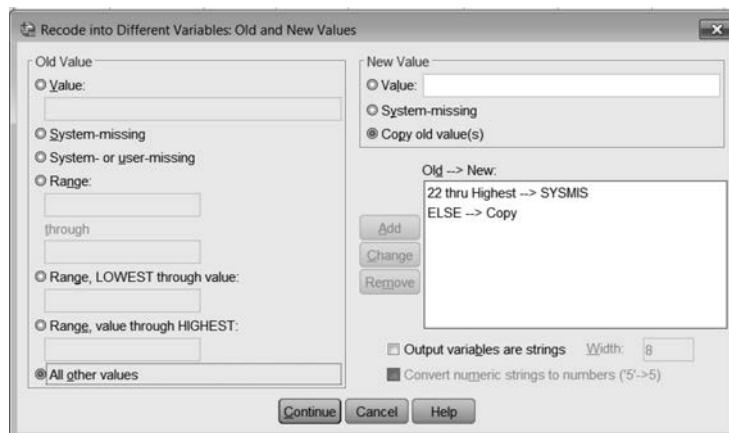
Figure 3.5. Recode Into Different Variables Dialog Box.



Recode into Different Variables: Old and New Values dialog box (see Figure 3.6)

The only cases to be changed are those with values of 22, 98, and 99; therefore, values equal to or greater than 22 will be transformed to missing values while all other values will remain the same. To indicate these transformations, click **Range**, **value through HIGHEST** under **Old Value**. In the blank box below your check, type the value 22. Under **New Value**, click **System-missing**, then click **Add**. Once this transformation has been made, be sure to indicate that all other values should be copied. (Specifically, click **All other values**, click **Copy old value(s)**, click **Add**, then click **Continue** and **OK**.) Now examination for missing data may begin.

Figure 3.6. Recode Into Different Variables: Old and New Values Dialog Box.



² By default, SPSS displays variable labels. To display variable names as shown here, go to the **Edit**, **Options**, **General** tab and, under **Variable Lists**, check **Display names**. Click on the **Output Labels** tab, then select **Names**, both from **Variables in item labels shown as** in the **Outline Labeling** box and from **Variables in labels shown as** in the **Pivot Table Labeling** box.

Missing Data

SPSS has several procedures within the analysis process for deleting cases or participants that have missing values. Examination of missing data in categorical variables can be done by creating a frequency table using **Frequencies**. To determine the extent of missing values within the variable of *satjob2*, the following menus should be selected:

Analyze
 Descriptive Statistics
 Frequencies

Quantitative variables with missing data can be examined by creating a table of **Descriptive Statistics**. To evaluate missing values in *rincom2*, open the following menus:

Analyze
 Descriptive Statistics
 Explore

For our example, the Frequencies and Explore tables reveal zero missing values for *satjob2* and 37 missing values for *rincom2*. Typically, if a categorical variable has 5% or fewer cases missing, the **Listwise** default would be utilized to delete the cases during the analysis. If a categorical variable has more than 5% but fewer than 15% of cases with missing data, an additional level or category would be created within the variable so that missing data would be recoded with this new level. Because SPSS no longer detects the missing values and does not recognize the new category as providing meaningful information for the variable being analyzed, these cases would not be included in the analysis.

SPSS also provides a variety of options for handling missing values in quantitative data. For most analyses, the **Explore: Option** dialog box typically displays the default of **Listwise** in which participants with missing values are removed for any of the variables identified in the analysis. **Pairwise** is another method of deleting participants with missing values. This method removes participants with missing values from each pair of variables being analyzed. Most researchers utilize the **Listwise** method because it is the default and provides a consistent sample size (*n*) across the analyses. However, when you need to maximize the number of participants within each analysis, **Pairwise** should be selected. In our example, data are missing for 37 cases in the variable *rincom2*. Because 5% of the cases have missing values, the **Listwise** default will be used to delete the missing cases. If more than 5% of the values were missing, an alternate method of replacement would need to be utilized. The most common method is to replace the missing values with the mean score of available cases for that variable. The replacement procedure also allows for other types of replacement values (e.g., median of nearby points, mean of nearby points). Typically, replacing 15% or fewer of the participants will have little effect on the outcome of the analysis. However, if a certain subject or variable has more than 15% missing data, you may want to consider dropping the participant or variable from the analysis. In such a case, to replace missing values with an estimated value, you would select the following menus:³

Transform
 Replace Missing Values

Replace Missing Values dialog box (see Figure 3.7)

Once in this dialog box, identify the targeted variable and move it to the **New Variable(s)** box. Notice that a name for the new variable has been generated. This may be changed accordingly. Next,

³ The steps illustrated in Figure 3.7 are for illustrative purposes only and are not performed in our example. Continue working on our example from outliers on the next page.

select the method of replacement. Five options are available in which missing values are replaced by the following:

Series Mean—The mean of all available cases for the specific variable. This is the default.

Mean of Nearby Points—The mean of surrounding values. You can designate the number of surrounding values to use under **Span of Nearby Points**. The default span is two values.

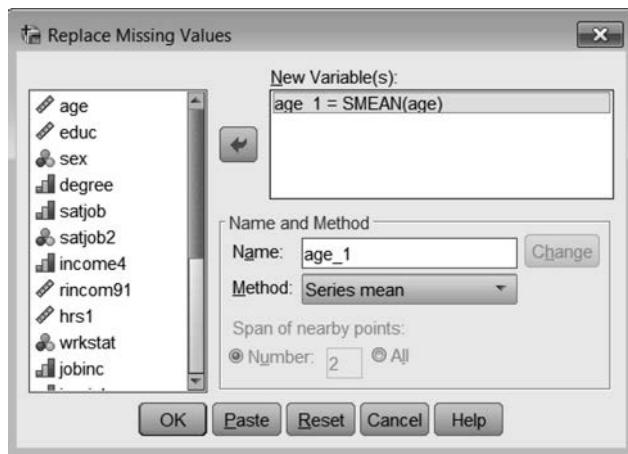
Median of Nearby Points—The median of surrounding values.

Linear Interpolation—The value midway between the surrounding two values.

Linear Trend at Point—A value consistent with a trend that has been established (e.g., values increasing from the first to the last case).

Once the method of replacement has been determined for the variable, you may also identify additional variables for replacement by using the **Change** button. This will allow you to identify another variable as well as another replacement method. Missing values in quantitative variables can also be estimated by creating a regression equation in which the variable with missing data serves as the dependent variable. Because this method is fairly sophisticated, we will discuss how to use predicted values in Chapter 7 (on multiple regression). Finally, if publishing the results of analyses that have utilized replacement of missing values, one should present the procedure(s) for handling such data.

Figure 3.7. Replace Missing Values Dialog Box (for illustration only).



Outliers

Because univariate outliers are participants or cases with extreme values for one variable, identification of such cases is fairly easy. The **Explore** menu under **Descriptive Statistics** offers several options for such examination. To identify outliers in the categorical variable of *satjob2*, **Frequencies** could be used to detect very uneven splits in categories, splits that typically produce outliers. Categorical variables with 90–10 splits between categories are usually deleted from the particular analysis because scores in the category with 10% of the cases influence the analysis more than those in the category with 90% of the cases. Because our example research question investigates group differences in income, both the IV (*satjob2*) and DV (*rincom2*) can be examined for outliers using **Explore**. This procedure will allow us to identify outliers for income within each group. To do so, select the following menus:

Analyze

Descriptive Statistics

Explore

Explore dialog box (see Figure 3.8)

Within this dialog box, move the DVs into the **Dependent List**. Move IVs into the **Factor List**. One button that you will see within many analyses is that of **Bootstrap**. Bootstrapping allows you to use your sample data to create “phantom” samples through replacement. Your analysis is then conducted with these copy samples to estimate the sampling distribution of the statistic. While this technique has become more popular in recent years and certainly has advantages, the use and interpretation of bootstrapping will not be addressed in this text. After you have defined the variables, click the **Statistics** button.

Figure 3.8. Explore Dialog Box.



Explore: Statistics dialog box (see Figure 3.9)

This box provides the following options for examining outliers:

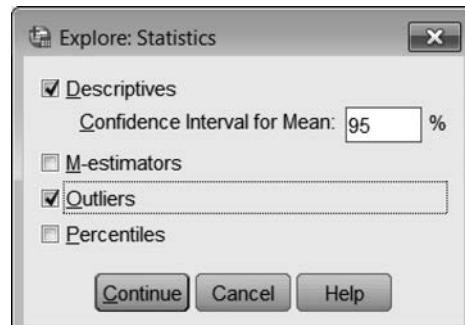
Descriptives—Calculates descriptive statistics for all participants and identified categories in the data. This is selected by default.

M-estimators—Assigns weights to cases depending upon their distance from the center.

Outliers—Identifies the five highest and five lowest cases for the DV by group.

Percentiles—Displays 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles for DV by group.

Figure 3.9. Explore: Statistics Dialog Box.



For our example, we selected **Descriptives** and **Outliers**. Click **Continue**, then click **Plots**.

Explore: Plots dialog box (see Figure 3.10)

This box provides several options for creating graphic representations of the data. For our example, we will select **Boxplots:Factor levels together** and **Descriptive:Stem-and-leaf**. Because it is best to examine normality after outliers have been addressed, other selections, such as **Normality plots with tests** and **Histogram**, will be conducted later. Click **Continue**, then click **OK**.

Figure 3.10. Explore: Plots Dialog Box.



The output reveals some outlier problems within the example. The case summary shows category splits in that 44% (313/710) of the sample is very satisfied while 56% (397/710) is not satisfied. This split is not severe enough to delete this variable. The table generated on extreme values (see Figure 3.11) identifies the five highest and lowest scores for each group. Keep in mind that these values are not necessarily outliers. The boxplot (see Figure 3.12) generated reveals that both groups have some outliers. The stem-and-leaf plots (see Figure 3.13) support this finding but provide more information regarding the number of outliers. The first plot indicates that 11 participants who are very satisfied reported extreme income values of 0. In contrast, the second plot displays that 22 participants who are not very satisfied reported extreme income values of 3 or less. Because the number of outlying cases for both groups is fairly small, these outliers could either be deleted using the case numbers identified in the boxplot or be altered to a value that is within the extreme tail in the accepted distribution. In this example, outliers will be altered by replacing them with a maximum/minimum value (depending on the direction of outliers) that falls within the accepted distribution. To alter the outliers in *rincom2*, the stem-and-leaf plots (Figure 3.13) help one identify the specific outlying values to be altered and the accepted minimum value to be used as the replacement value. Cases that have an income level of 3 or less will be replaced with the accepted value of 4. To alter outliers, complete the following steps:

Transform

Recode Into Different Variable

Figure 3.11. Extreme Values Table for Income (*rincom2*) by Job Satisfaction (*satjob2*).

Extreme Values			
<i>satjob2</i>		Case Number	Value
rincom2	Very satisfied	Highest 1	62 21.00
		2	69 21.00
		3	78 21.00
		4	86 21.00
		5	108 21.00 ^a
	Lowest	Highest 1	738 .00
		2	716 .00
		3	691 .00
		4	663 .00
		5	649 .00 ^b
Not very satisfied	Highest	Highest 1	12 21.00
		2	41 21.00
		3	51 21.00
		4	54 21.00
		5	63 21.00 ^a
	Lowest	Highest 1	734 .00
		2	732 .00
		3	670 .00
		4	668 .00
		5	551 .00 ^b

a. Only a partial list of cases with the value 21.00 are shown in the table of upper extremes.

b. Only a partial list of cases with the value .00 are shown in the table of lower extremes.

Figure 3.12. Boxplot for Income (*rincom2*) by Job Satisfaction (*satjob2*).

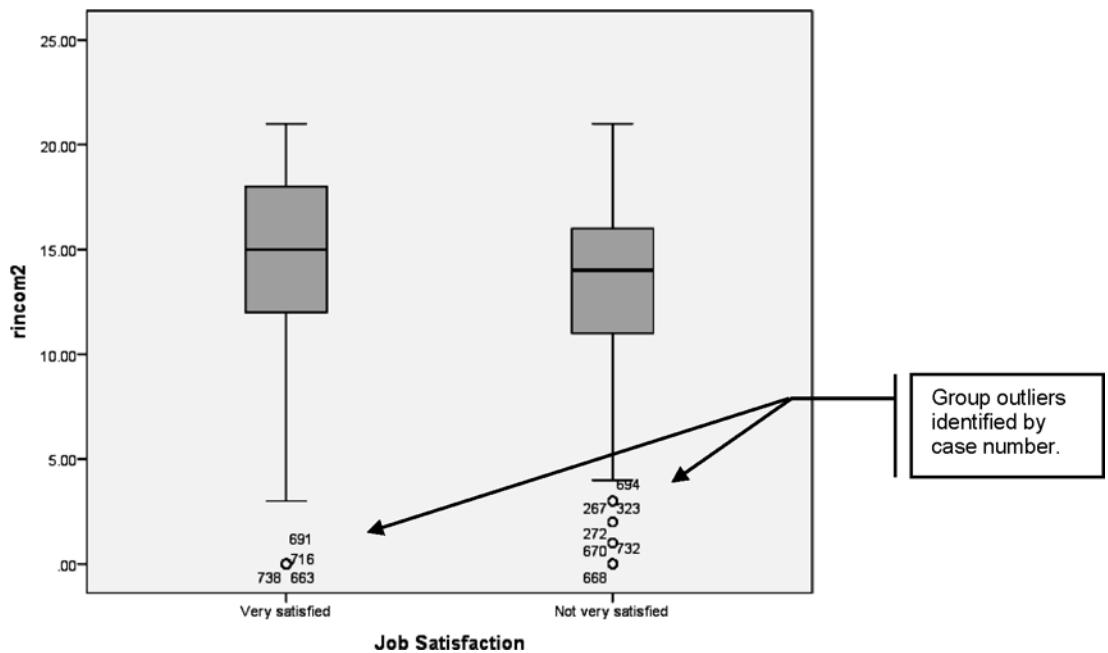


Figure 3.13. Stem-and-Leaf Plots for Income (*rincom2*) by Job Satisfaction (*satjob2*).

rincom2 Stem-and-Leaf Plot for
satjob2= Very satisfied

Frequency Stem & Leaf

11 participants
are outliers with
values of 0.

Stem width: 1.00
Each leaf: 1 case(s)

rincom2 Stem-and-Leaf Plot for
satjob2= Not very satisfied

Frequency Stem & Leaf

22.00	Extremes	(=<3.0)	22 ou of
8.00	4 .	00000000	
4.00	5 .	0000	
4.00	6 .	0000	
4.00	7 .	0000	
9.00	8 .	000000000	
24.00	9 .	0000000000000000000000000000	
24.00	10 .	0000000000000000000000000000	
33.00	11 .	00000000000000000000000000000000	
29.00	12 .	00000000000000000000000000000000	
27.00	13 .	00000000000000000000000000000000	
35.00	14 .	00000000000000000000000000000000	
41.00	15 .	00000000000000000000000000000000	
36.00	16 .	00000000000000000000000000000000	
24.00	17 .	00000000000000000000000000000000	
27.00	18 .	00000000000000000000000000000000	
15.00	19 .	0000000000000000	
14.00	20 .	0000000000000000	
17.00	21 .	0000000000000000	

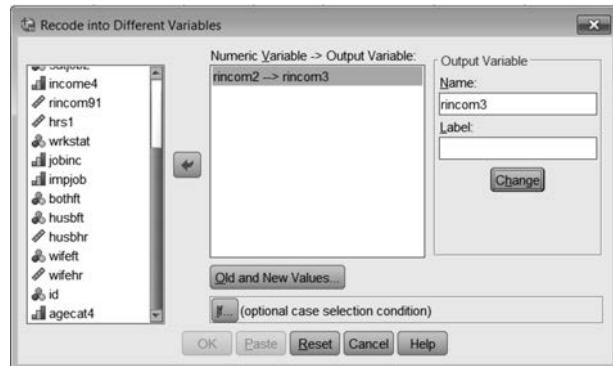
22 participants are outliers with values of 3 or less.

Stem width: 1.00
Each leaf: 1 case(s)

Recode Into Different Variables dialog box (see Figure 3.14)

Recoding *rincom2* into a different variable will allow us to conduct our analysis with the original variable (*rincom91*), the first altered variable (*rincom2*), and the second altered variable (*rincom3*). This joint analysis helps to determine if altering the outliers had an impact on the results. Once in this dialog box, identify the variable (*rincom2*) to be altered and move it to the Input (or Numeric Variable → Output Variable) box. Indicate a new name for the variable (*rincom3*), then click **Change**. Click **Old and New Values** to specify the transformations.

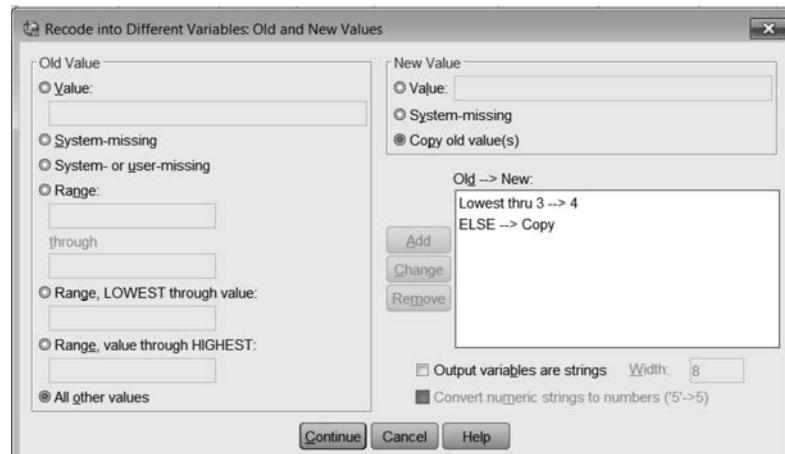
Figure 3.14. Recode Into Different Variables Dialog Box.



Recode Into Different Variables:Old and New Values dialog box (see Figure 3.15)

The only cases to be changed are those with values of 3 or less. All other values will remain the same. To indicate these transformations, under **Old Value**, click **Range: LOWEST through value**. In the blank box beneath, indicate the cutoff value of 3. Under **New Value**, check **Value**, type in 4 and click **Add**. These commands have transformed the outliers (those 3 or less) to a value of 4. The next step is to indicate that all other values will stay the same. To do so, under **Old Value** click **All other values**. Under **New Value**, click **Copy old value(s)**, then click **Add**. Click **Continue**, and then click **OK** in the next window. Once cases have been altered, you can proceed with further data examination and analysis. But remember, when conducting the analyses, do so with both original and altered variables.

Figure 3.15. Recode Into Different Variables: Old and New Values Dialog Box.



Normality, Linearity, and Homoscedasticity

The **Explore** procedure also provides several options for examining normality and is usually conducted after addressing outliers. To conduct this procedure using the DV (*rincom3*) and IV (*satjob2*), return to the previous directions for **Explore**. Within the **Explore: Statistics** dialog box, be sure to check **Descriptives**. In the **Explore: Plots** dialog box, check **Boxplots: None**. In **Descriptive**, you may uncheck **Stem-and-Leaf**, but be sure to check **Histogram**. Also check **Normality plots with tests**. These are most helpful in examining normality. Click **Continue**, and then click **OK** in the next window. Descriptive statistics (see Figure 3.16) present skewness and kurtosis values, which also imply negative distributions. Typically, skewness and kurtosis values should lie between +1 and -1.

Figure 3.16. Descriptive Statistics for Income (*rincom3*) by Job Satisfaction (*satjob2*).

Descriptives				
		Statistic	Std. Error	
satjob2	rincom3	Mean	14.6166	
		95% Confidence Interval for Mean		
		Lower Bound	14.1297	
		Upper Bound	15.1035	
		5% Trimmed Mean	14.8518	
		Median	15.0000	
		Variance	19.167	
		Std. Deviation	4.37797	
		Minimum	4.00	
		Maximum	21.00	
		Range	17.00	
		Interquartile Range	6.00	
		Skewness	-.739	
	Not very satisfied	Kurtosis	.078	
satjob2		Mean	.22378	
		95% Confidence Interval for Mean		
		Lower Bound	12.8573	
		Upper Bound	13.7372	
		5% Trimmed Mean	13.3938	
		Median	14.0000	
		Variance	19.881	
		Std. Deviation	4.45883	
		Minimum	4.00	
		Maximum	21.00	
		Range	17.00	
		Interquartile Range	5.50	
		Skewness	-.395	
		Kurtosis	.404	
		.138	.275	
		.122	.244	

For a normal distribution, kurtosis and skewness values will be close to zero but can range between -1 and +1.

Histograms (see Figure 3.17) display moderate, negatively skewed distributions for both groups. The normal Q-Q plots support this finding as the observed values deviate somewhat from the straight line (see Figure 3.18). Tests of normality were also calculated. Specifically, the Kolmogorov-Smirnov test (see Figure 3.19) significantly rejects the hypothesis of normality of income for the populations of both groups. Thus, the variable of *rincom3* must be transformed again. To decrease the moderate negative skewness, the transformation procedure will reflect and take the square root of the variable. Steps for such transformation follow:

Transform

Compute Variable

Figure 3.17. Histograms for Income (*rincom3*) by Job Satisfaction (*satjob2*).

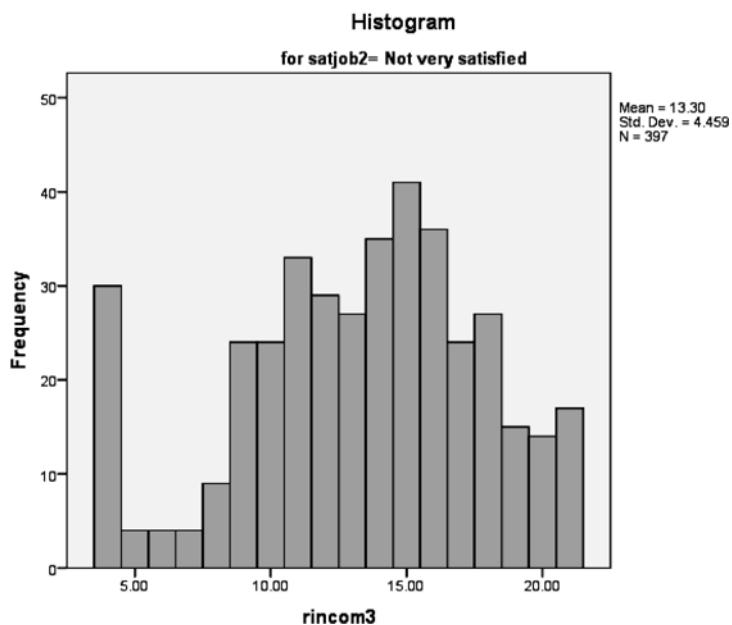
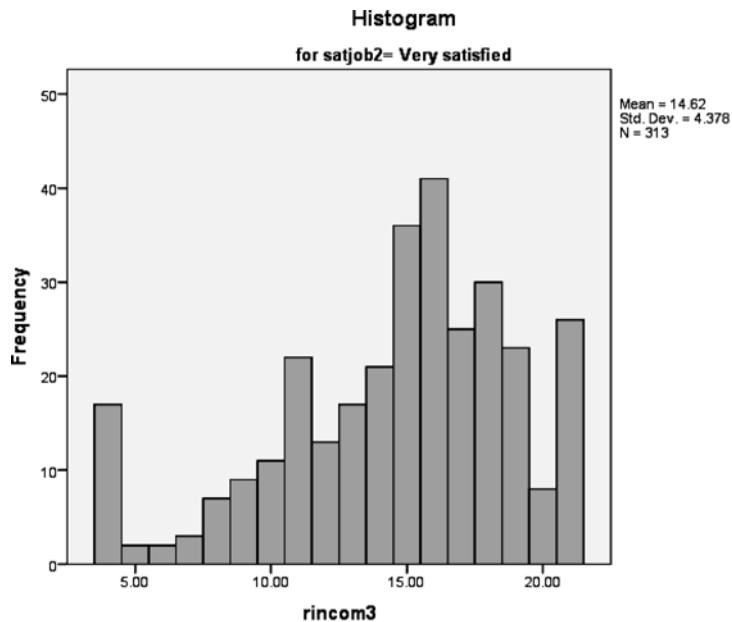


Figure 3.18. Normal Q-Q Plots for Income (*rincom3*) by Job Satisfaction (*satjob2*).

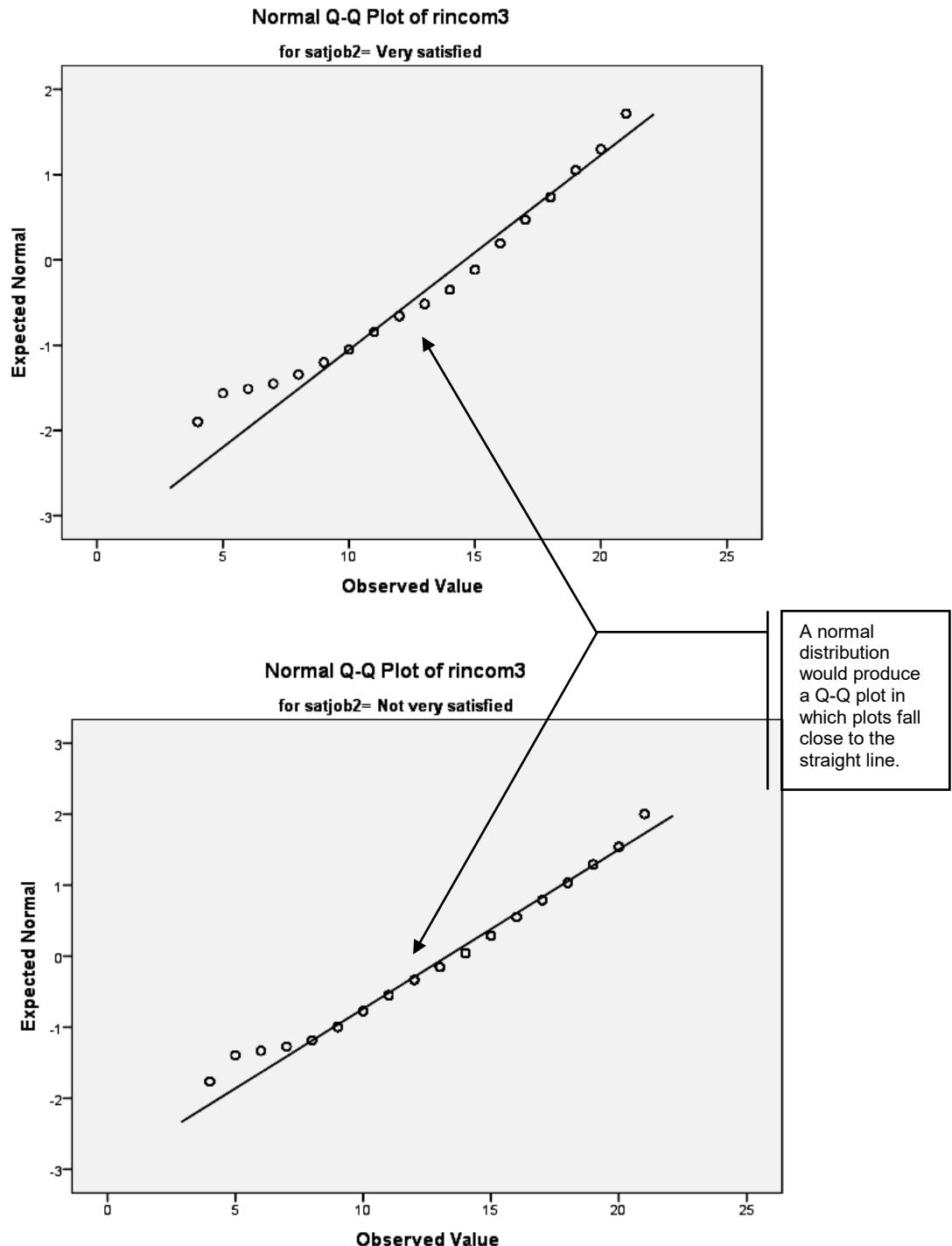


Figure 3.19. Tests of Normality for Income (*rincom3*) by Job Satisfaction (*satjob2*).

		Tests of Normality					
		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
rincom3	Very satisfied	.139	313	.000	.937	313	.000
	Not very satisfied	.089	397	.000	.962	397	.000

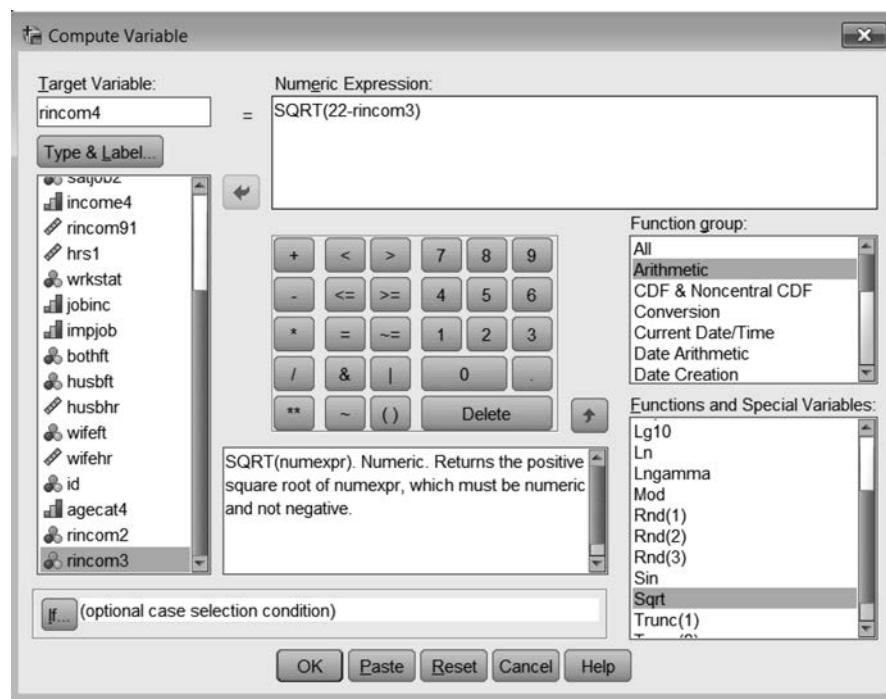
a. Lilliefors Significance Correction

Significance indicates a nonnormal distribution.

Compute Variable dialog box (see Figure 3.20)

Within the **Transform:Compute Variable** dialog box, create a name for the new variable you are creating by typing it in the Target Variable box. For our example, we have used *rincom4*. Then identify the appropriate function to be used in the transformation and move it to the **Numeric Expression** box. Because the example calls for reflect with square root, the equation to apply is as follows: NewVar = $\text{SQRT}(K-\text{OldVar})$ in which *K* is the largest score of the OldVar plus one. For the *rincom3*, the largest value is 21; thus, *K* = 22. Once the function has been inserted into the **Numeric Expression** box, the remaining parts of the transformation equation must be inserted, creating the expression *rincom4* = $\text{SQRT}(22-\text{rincom3})$.

Figure 3.20. Compute Variable Dialog Box.



Once data have been transformed, examination of normality should be conducted again using the **Explore** procedure. Although the Kolmogorov-Smirnov test is still significant, the skewness and kurtosis values (see Figure 3.21) are much closer to zero. In addition, histograms (see Figure 3.22) and normal Q-Q plots (see Figure 3.23) reveal distributions for both groups that are much more normal. Consequently, we will assume the transformation was successful.⁴

⁴ At this point, a new data set named *career-b.sav* was created, reflecting all transformations and recoding of variables performed thus far in this chapter upon the original *career-a.sav* data set.

Figure 3.21. Descriptive Statistics for Income (*rincom4*) by Job Satisfaction (*satjob2*).

Descriptives

satjob2		Statistic	Std. Error
rincom4	Very satisfied	Mean	2.5877
		95% Confidence Interval for Mean	Lower Bound 2.4953 Upper Bound 2.6800
		5% Trimmed Mean	2.5839
		Median	2.6458
		Variance	.690
		Std. Deviation	.83042
		Minimum	1.00
		Maximum	4.24
		Range	3.24
		Interquartile Range	1.16
		Skewness	-.013
		Kurtosis	-.362
	Not very satisfied	Mean	2.8384
		95% Confidence Interval for Mean	Lower Bound 2.7590 Upper Bound 2.9178
		5% Trimmed Mean	2.8593
		Median	2.8284
		Variance	.648
		Std. Deviation	.80476
		Minimum	1.00
		Maximum	4.24
		Range	3.24
		Interquartile Range	.94
		Skewness	-.291
		Kurtosis	-.265

Skewness and kurtosis values are closer to zero, indicating a more normal distribution.

Figure 3.22. Histograms for Income (*rincom4*) by Job Satisfaction (*satjob2*).

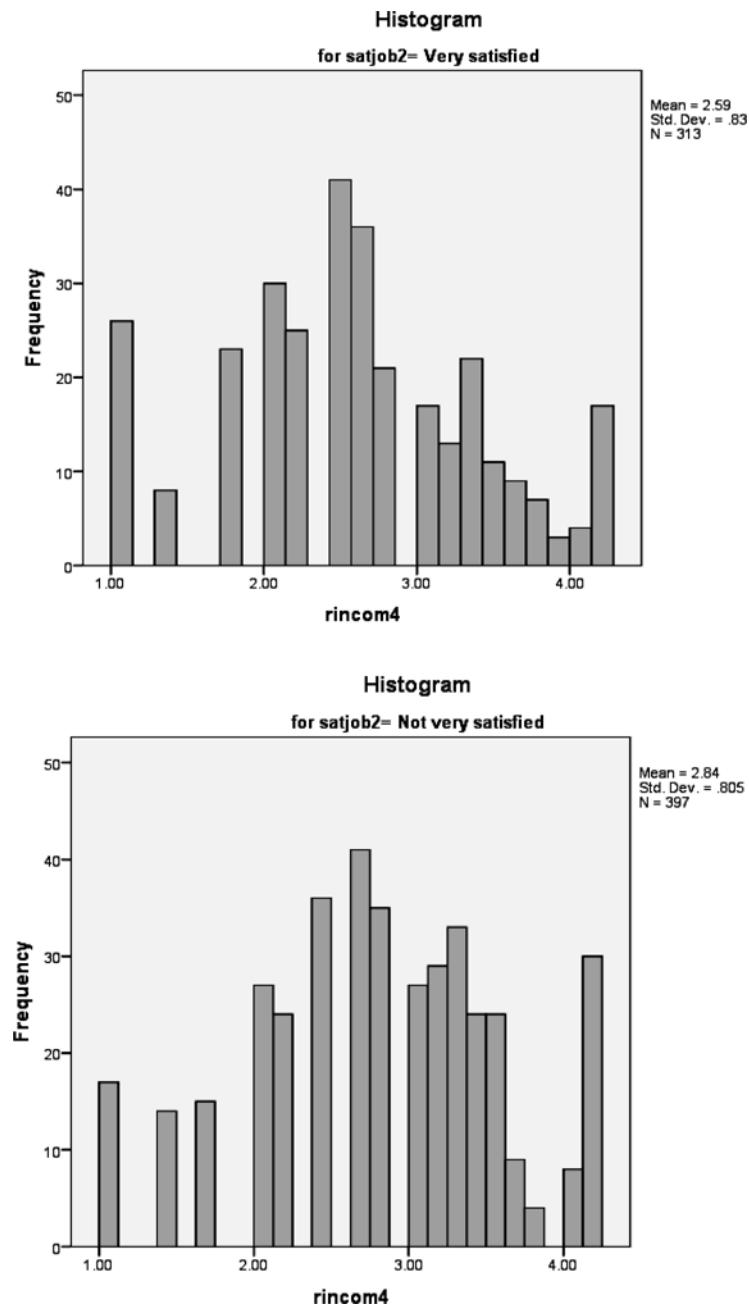
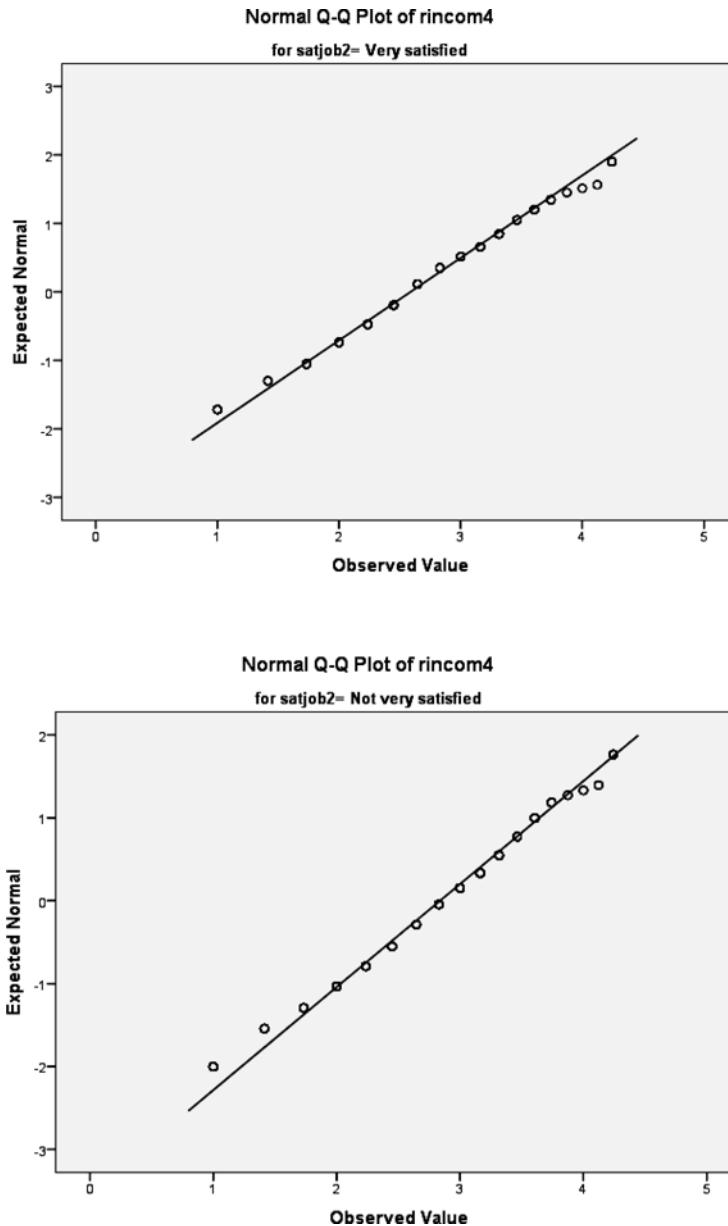


Figure 3.23. Normal Q-Q Plots for Income (*rincom4*) by Job Satisfaction (*satjob2*).



Because our research question involves comparing groups on a single quantitative variable, linearity cannot be examined. However, homoscedasticity, also known as *homogeneity of variance* when comparing groups, can be assessed by determining if the variability for the DV (*rincom4*) is about the same within each category of the IV (*satjob2*). This can be completed when conducting the group comparison analyses (e.g., *t* test, ANOVA). Within these statistical procedures, Levene's test for equal variances is automatically calculated. Levene's test may also be conducted within **Explore**. Within the **Explore: Plots** dialog box, be sure to select **Untransformed** under **Spread vs Level with Levene Test** (see Figure 3.24). The **Untransformed** option calculates the Levene's statistic with the raw data. Figure 3.25 presents output for this Levene's test. The Levene's statistic is 0.139 with a *p* value of 0.709. Thus, the hypothesis for equal variances is not rejected, which indicates that variances are fairly equivalent between the groups.

Figure 3.24. Explore: Plots Dialog Box.

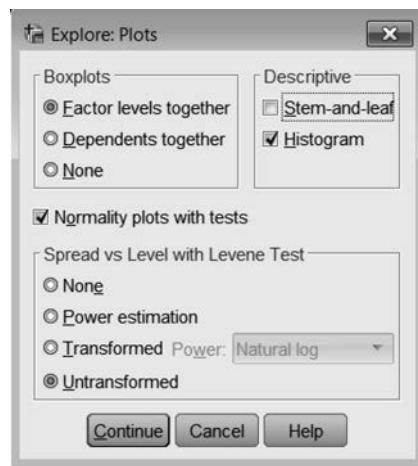


Figure 3.25. Levene's Test for Equality of Variances.

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
rincom4	Based on Mean	.139	1	708	.709
	Based on Median	.123	1	708	.726
	Based on Median and with adjusted df	.123	1	705.840	.726
	Based on trimmed mean	.120	1	708	.729

Nonsignificant value indicates homogeneity of variance.

Univariate Example With Ungrouped Data

In this example, we utilize the *career-b.sav* data set and seek to investigate the degree to which the variables of years of education (*educ*), age (*age*), and hours worked weekly (*hrs1*) predict income levels (*rincom4*).

Missing Data and Outliers

Missing data are analyzed for each of the four variables using the methods described in the previous example. However, the identification of outliers requires a different method because several variables are in question. Although this is not a multivariate example, analysis for multivariate outliers will occur in order to examine the variables together with respect to outliers. The most common method is calculating Mahalanobis distance within the **Regression** procedure. To calculate Mahalanobis distance, complete the following steps to conduct the regression:

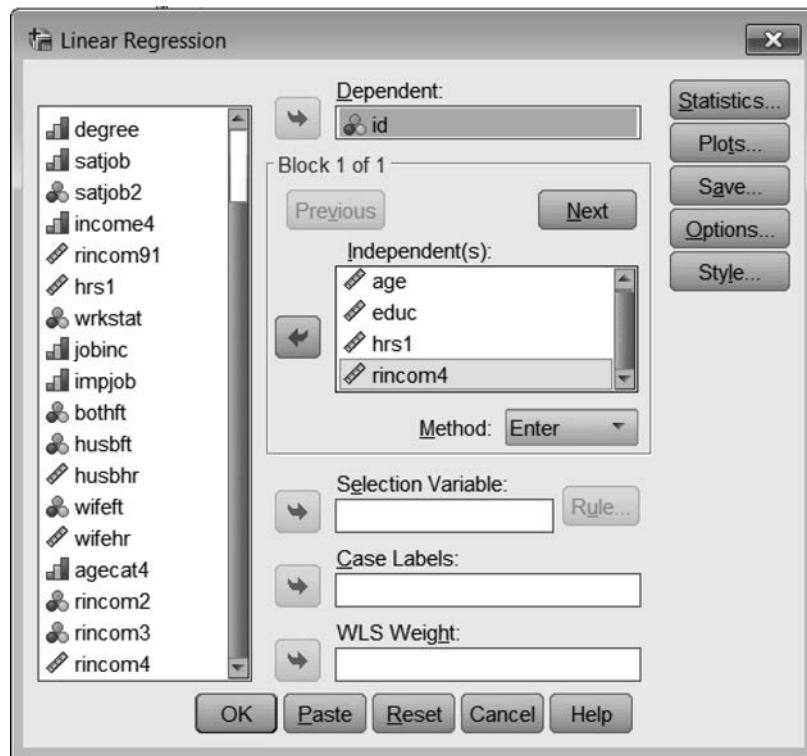
Analyze

Regression
Linear

Linear Regression dialog box (see Figure 3.26)

Identify all four quantitative variables to be analyzed and move them to the **Independent(s)** box. Utilize a case number or the *id* variable for the Dependent because this procedure is not influenced by the DV. Next, click the **Save** button from the choices at the right side of the window. All other procedures regarding regression will be discussed further in Chapter 7.

Figure 3.26. Linear Regression Dialog Box.



Linear Regression: Save dialog box (see Figure 3.27)

Once in this box, check **Mahalanobis** under **Distances**, click **Continue**, and then click **OK**. Although this procedure does not produce output that is especially helpful in identifying multivariate outliers, a new variable (*MAH_I*) is created for Mahalanobis distances, which is tested using chi-square (χ^2) criteria. Outliers are indicated by chi-square values that are significant at $p < .001$ with the respective degrees of freedom. The number of variables being examined for outliers is used as the degrees of freedom. To determine the critical value for χ^2 , one must utilize a table of critical values for chi-square, available on page 357 of this textbook. For our example, the critical value of χ^2 at $p < .001$ and $df = 4$ is 18.467. Consequently, cases with a Mahalanobis distance greater than 18.467 are considered multivariate outliers for the variables of *age*, *educ*, *hrs1*, and *rincom4*. Identification of the outlying cases can now be easily achieved using the **Explore** procedure, specifying *MAH_I* as the DV. In this case, the Factor list is left blank. Within **Explore**, all that is necessary is to check **Outliers** within the **Statistics** dialog box, as previously demonstrated. The Extreme Values for Mahalanobis Distance (see Figure 3.28) that was generated lists the five highest and lowest values for *MAH_I*.

Figure 3.27. Linear Regression: Save Dialog Box.

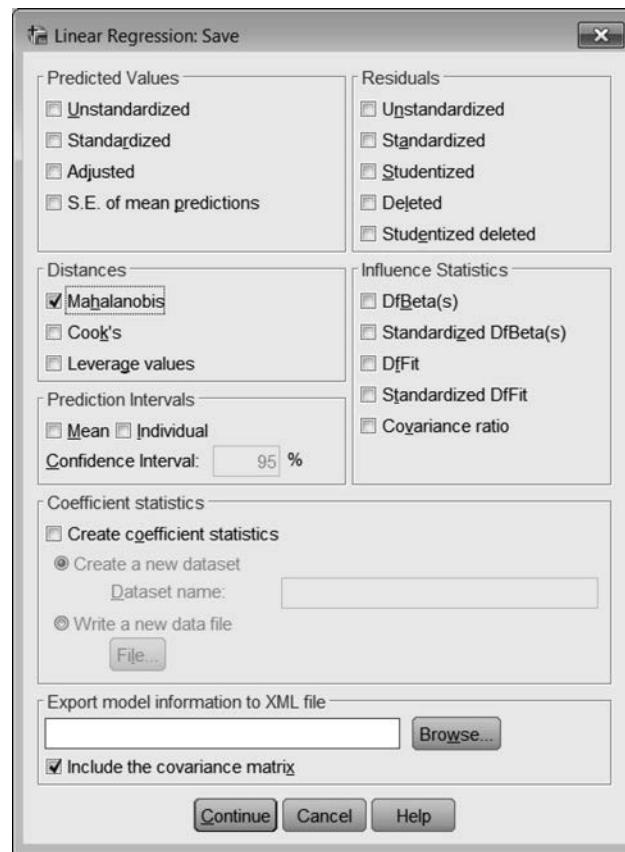


Figure 3.28. Extreme Values for Mahalanobis Distance.

Extreme Values		
	Case Number	Value
MAH_1 Highest	222	29.93848
	24	22.03483
	616	18.71248
	208	18.52947
	729	18.15252
Lowest	292	.23228
	146	.24275
	550	.30986
	126	.32204
	443	.33112

Only cases (222, 24, 616, 208) with values that exceed the critical value of chi-square are considered outliers.

Four cases (222, 24, 616, and 208) are identified as outliers because they exceed 18.467. These cases are eliminated from future analysis by the following:

Data

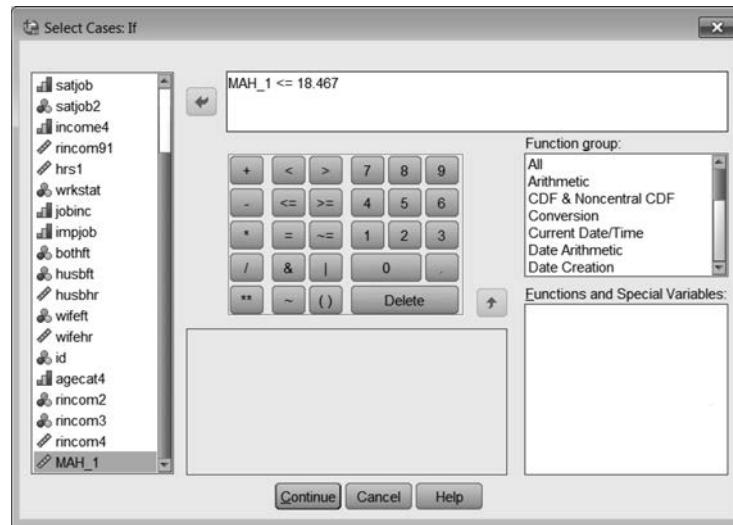
Select Cases

Within the **Select Cases** dialog box, select **If condition is satisfied** and click **If**.

Select Cases: If dialog box (see Figure 3.29)

Select *MAH_1* and move it to the function window. Indicate that we are selecting cases in which $MAH_1 \leq 18.467$. Clicking **Continue** and then **OK** eliminates cases from all future analysis where *MAH_1* is greater than 18.467. Keep in mind that even missing cases for this variable (*MAH_1*) are eliminated as well, which may significantly reduce your sample size.

Figure 3.29. Select Cases: If Dialog Box.



Normality, Linearity, and Homoscedasticity

Because our continuing example includes several quantitative variables, univariate normality should be examined for each individual variable. However, multivariate normality will need to be assessed as well. To assess univariate normality, the **Explore** procedure is conducted for each of these variables. Histograms, normal Q-Q plots, and descriptive statistics reveal the following: *age* has moderate, positive skewness; *hrs1* and *educ* are fairly normal but very peaked. *Age* is therefore transformed into *age2* by taking the square root of *age*. *Hrs1* and *educ* will not be transformed at this point.

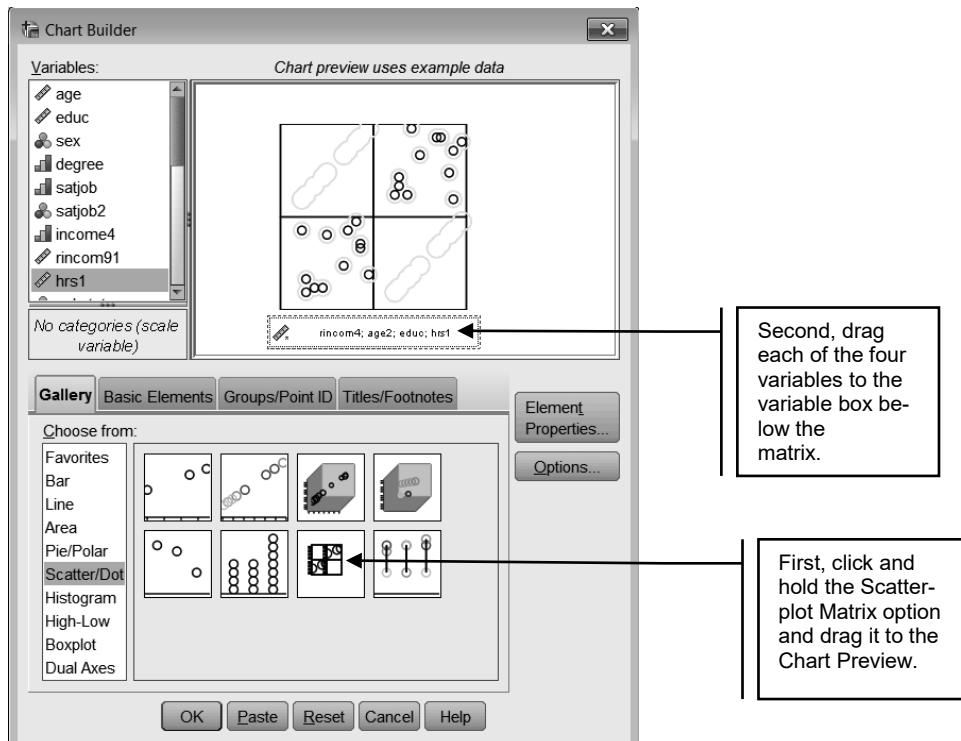
The next step is to analyze for multivariate normality and linearity. The most common method of evaluating multivariate normality is creating scatterplots of all variables in relation to one another. If variable combinations are normal, scatterplots will display elliptical shapes. To create scatterplots of the four variables, open the following menus:

Graphs
Chart Builder

Chart Builder dialog box (see Figure 3.30)

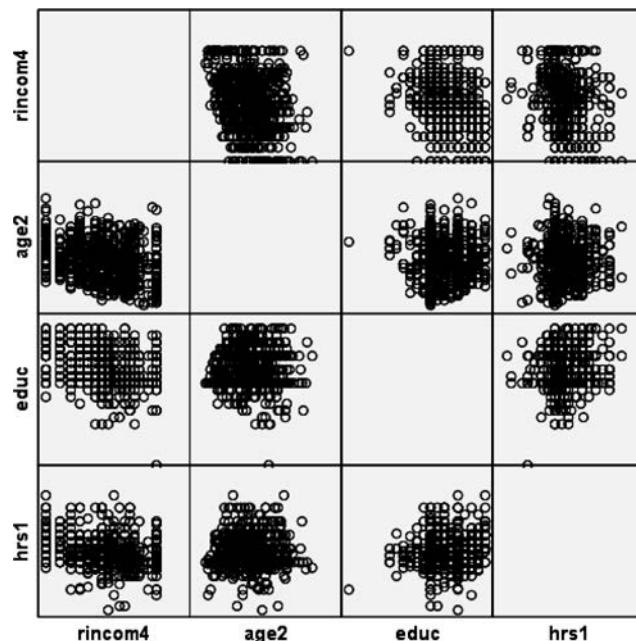
Under **Choose from**, select **Scatter/Dot**. From the scatterplot options that appear, click and hold the **Scatterplot Matrix** icon and drag it to the **Chart Preview** area. Then select each variable in turn and drag it to the **Scattermatrix** area one at a time. This will define the variables to be presented in the matrix. Notice that we are utilizing the *age2* variable as transformed in the paragraphs above. Click **OK**.

Figure 3.30. Chart Builder Dialog Box.



The output (see Figure 3.31) for our example displays non-elliptical shapes for all combinations, which implies failure of normality and linearity. For such a situation, two options are available: (1) recheck univariate normality for each variable or (2) only utilize variables for univariate analyses.

Figure 3.31. Scatterplot Matrix of *rincom4*, *age2*, *educ*, and *hrs1*.



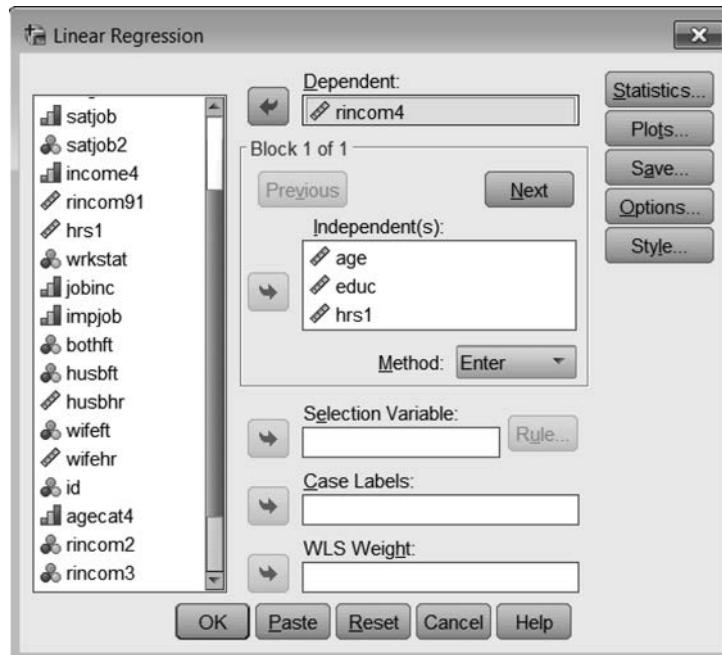
Because the use of bivariate scatterplots is fairly subjective in examining linearity, we recommend a more sophisticated method that compares standardized residuals to the predicted values of the DV. This method also provides some information regarding homoscedasticity. To create the residual plot for these variables, open the following menus:

Analyze
Regression
Linear

Linear Regression dialog box (see Figure 3.32)

Move *rincom4* to the **Dependent** box. Select the three IVs and move them to the **Independent(s)** box. Then click **Plots**.

Figure 3.32. Linear Regression Dialog Box.



Linear Regression: Plots dialog box (see Figure 3.33)

Within this menu, select the standardized residuals (ZRESID) for the *y*-axis. Select the standardized predicted values (ZPRED) for the *x*-axis, click **Continue**, and then click **OK**. When the assumptions of linearity, normality, and homoscedasticity are met, residuals will create an approximate rectangular distribution with a concentration of scores along the center. Figure 3.34 displays fairly consistent scores throughout the plot with concentration in the center. When assumptions are not met, residuals may be clustered on the top or bottom of the plot (nonnormality), may be curved (nonlinearity), or may be clustered on the right or left side (heteroscedasticity). Because such extreme clustering is not displayed, we will conclude that the assumptions of normality, linearity, and homoscedasticity are met for these variables.

Figure 3.33. Linear Regression: Plots Dialog Box.

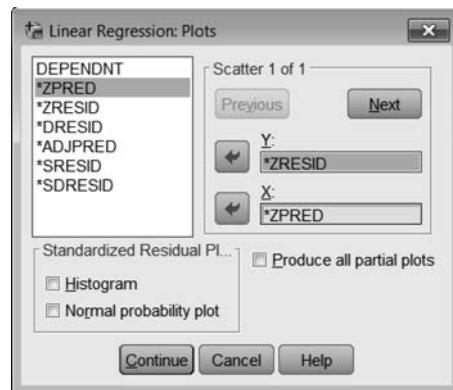
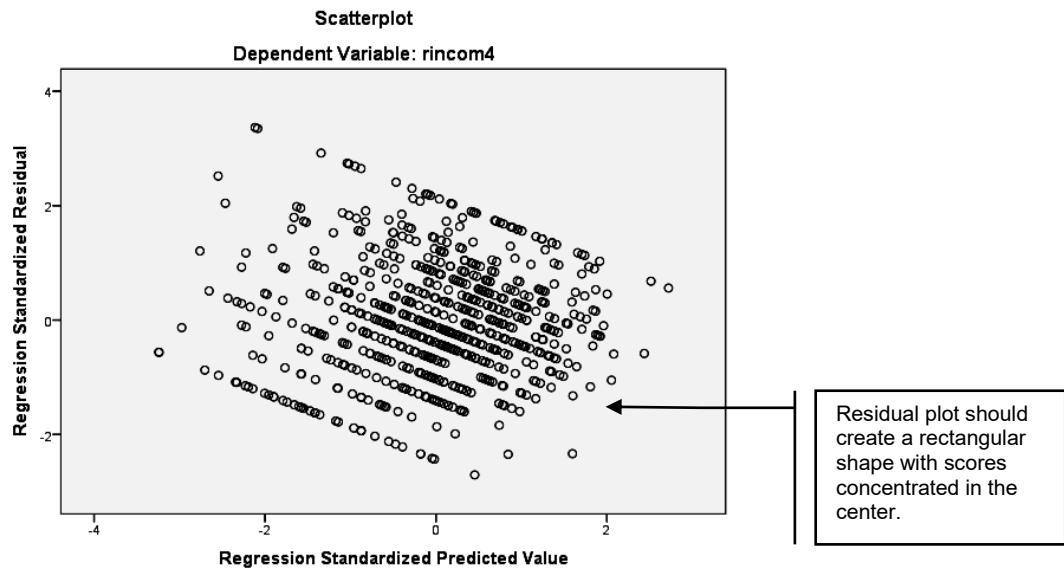


Figure 3.34. Scatterplot of Standardized Predicted Values by Standardized Residuals.



SECTION 3.8 USING SPSS TO EXAMINE GROUPED DATA FOR MULTIVARIATE ANALYSIS

The following example describes the process of examining missing values, outliers, normality, linearity, and homoscedasticity for grouped multivariate data. A nongrouped multivariate scenario would typically follow the univariate nongrouped example that was previously presented. For our example, we will continue to utilize the data set *career-b.sav* (see footnote, page 48), as we are interested in investigating group differences (*satjob2*) in *rincom4*, *age2*, *educ*, and *hrs1*. Notice that we are utilizing previously transformed variables of *rincom4* and *age2*. However, because our research question has changed a bit, all cases have been selected. In other words, the previously eliminated cases (222, 24, 616, and 208) are back in the analysis. To return these cases to the analysis, undo what we did on pages 54 and 55, using the following steps:

Data
 Select Cases
 Select
 All cases

Missing Data and Outliers

Missing data would be assessed for each variable. Multivariate outliers would be examined using Mahalanobis distances within **Regression** with *rincom4*, *age2*, *educ*, and *hrs1* as the independent variables, **Explore** would be conducted with *MAH_1* as the DV and *satjob2* as the factor. Please refer to the previous example on methods for conducting this procedure, beginning on page 52. Table of Extreme Values (see Figure 3.35) presents chi-square values for each possible outlier within each group. The critical value at $p < .001$ for chi-square is again 18.467 with $df = 4$. Thus, the very satisfied group has four outliers (222, 24, 616, and 208), while the unsatisfied group has none.⁵ Notice that the particular cases identified as outliers are the same from the previous example where data was ungrouped. Identified outliers will once again be filtered from further analysis by selecting cases where $MAH_1 \leq 18.467$, as shown on pages 53 and 54.

Figure 3.35. Table of Extreme Values.

Extreme Values			
satjob2		Case Number	Value
MAH_1	Very satisfied	Highest	1
			222 29.93848
			24 22.03483
			616 18.71248
			208 18.52947
			344 17.19371
		Lowest	1
			146 .24275
			550 .30986
			126 .32204
			443 .33112
			405 .37152
	Not very satisfied	Highest	1
			729 18.15252
			545 16.99042
			551 16.85647
			575 16.49552
			427 15.57695
		Lowest	1
			292 .23228
			261 .34476
			328 .41852
			637 .42747
			677 .43391

Cases 222, 24, 616, and 208 are outliers because their values exceed chi-square critical.

Normality, Linearity, and Homoscedasticity

Because groups are being compared, assumptions of normality, linearity, and homoscedasticity must be examined for all the quantitative variables together by each group. Prior to multivariate examination, univariate examination should take place for each variable within each group. Although these variables have been assessed for assumptions in the previous example, the examination was with ungrouped data. Consequently, assessment of normality and homoscedasticity will need to be conducted for each variable within each group. Using the **Explore** procedure provides the histograms, tests of normality, descriptive statistics, and normal Q-Q plots, all of which indicate that the four quantitative variables are fairly normal. Homoscedasticity (homogeneity of variance) will be assessed using the Levene's test within the *t* test of independent samples. These results indicate equality of variance for each variable between groups.

Multivariate normality, linearity, and homoscedasticity can now be assessed. Multivariate normality and linearity are examined with a matrix of scatterplots for each group. Because we are creating a matrix of scatterplots for each group, complete the following steps:

Graphs

Legacy Dialogs

Scatter/Dot

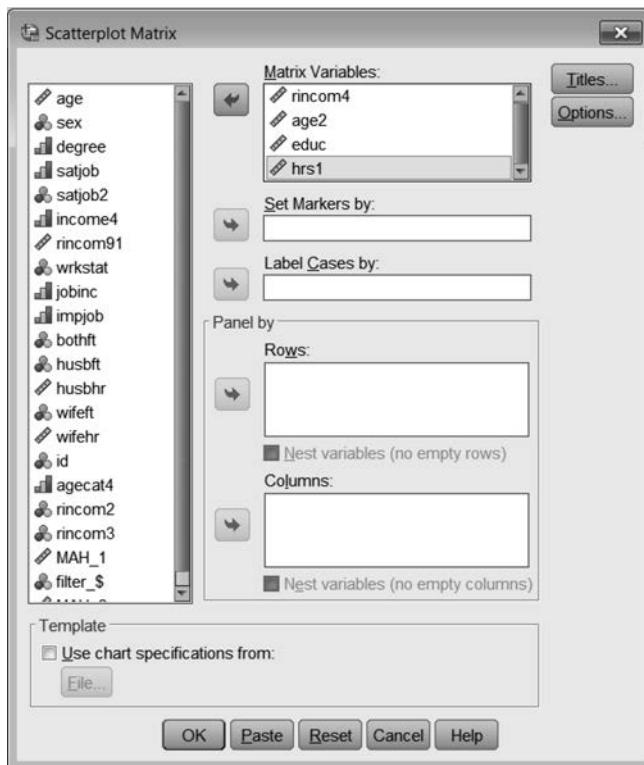
Within the **Scatter/Dot** dialog box, select **Matrix Scatter**, then **Define**.

⁵ Data set *career-c.sav* on website created, at this point, for reference.

Scatterplot Matrix dialog box (see Figure 3.36)

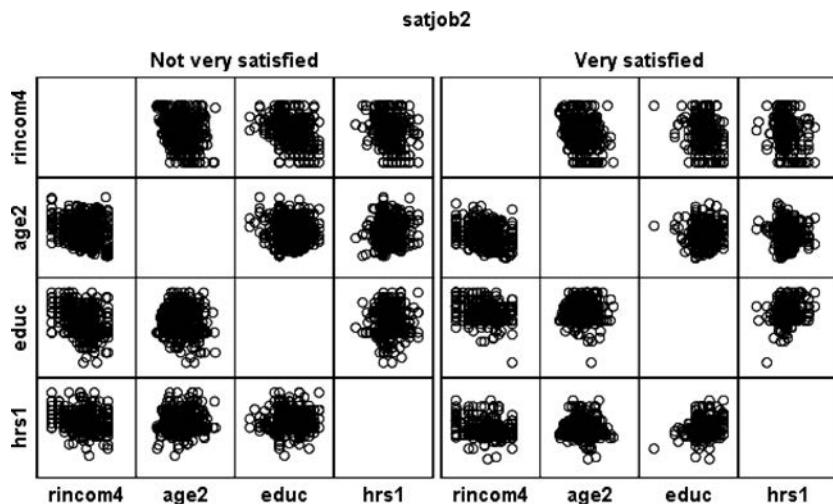
Move all continuous variables (*rincom4*, *age2*, *educ*, and *hrs1*) to **Matrix Variables**. Move the grouping variable (*satjob2*) to **Columns**, under **Panel by**.

Figure 3.36. Scatterplot Matrix Dialog Box.



The results (see Figure 3.37) are quite similar to the previously produced scatterplot matrix of the same variables but with ungrouped data (see Figure 3.31). Although some plots display enlarged oval shapes, multivariate normality and linearity are questionable.

Figure 3.37. Scatterplot Matrices for *rincom4*, *age2*, *educ*, and *hrs1* by *satjob2*.



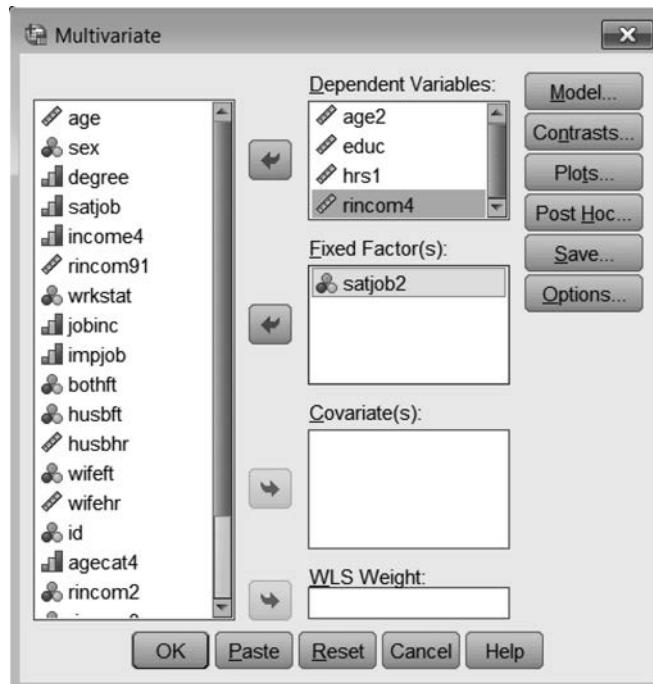
Homogeneity of variance-covariance matrices is evaluated within MANOVA by calculating Box's test of equality of covariance. To do so, open the following menus:

Analyze
General Linear Model
Multivariate

Multivariate dialog box (see Figure 3.38)

Move the DVs into the **Dependent Variables** box. Identify the IV and move it to the **Fixed Factor(s)** box. Once variables have been identified, click **Options**.

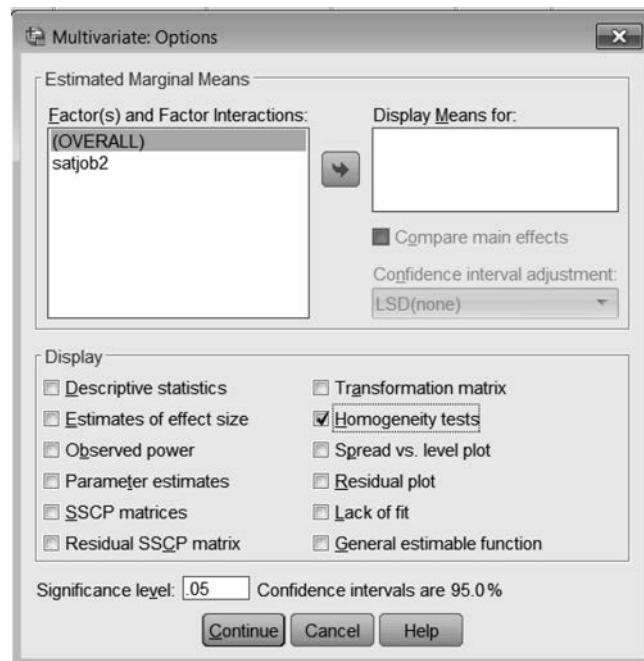
Figure 3.38. Multivariate Dialog Box.



Multivariate: Options dialog box (see Figure 3.39)

Under **Display**, check **Homogeneity tests**. Because tests of homogeneity of variance-covariance matrices are quite strict, a more stringent critical value of .025 or .01 is often used, rather than .05. Thus, when interpreting the results from the Box's test (see Figure 3.40), the probability value was calculated at .207, which would lead us to conclude that the covariance matrices for the dependent variable are fairly equivalent at the .025 level of significance.

Figure 3.39. Multivariate: Options Dialog Box.



Click **Continue**, then **OK**.

Figure 3.40. Box's Test of Equality of Covariance.

Box's Test of Equality of Covariance Matrices ^a	
Box's M	13.385
F	1.330
df1	10
df2	2029679.692
Sig.	.207

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + satjob2

Significance is NOT found at .025 or .01. Equality of covariance is concluded.

SUMMARY

Screening data for missing data, outliers, and the assumptions of normality, linearity, and homoscedasticity is an important task prior to conducting statistical analyses. If data are not screened, conclusions drawn from statistical results may be erroneous. Figure 3.41 presents the steps for univariate and multivariate examination of grouped data, while Figure 3.42 presents the steps for univariate and multivariate examination of ungrouped data.

KEYWORDS

- Box's M test for equality of variance-covariance matrices
- continuous variables
- data transformations
- heteroscedasticity
- homoscedasticity
- Kolmogorov-Smirnov statistic
- kurtosis
- leptokurtosis
- Levene's test
- linearity
- Mahalanobis distance
- multivariate outliers
- normality
- outliers
- platykurtosis
- residuals
- robustness
- skewness
- univariate outliers

Figure 3.41. Steps for Screening Grouped Data.

Examination & Process	SPSS Procedure	Technique for “Fixing”
<p>Missing Data</p> <ul style="list-style-type: none"> Examine missing data for each variable. 	<ul style="list-style-type: none"> Run Frequency for categorical variables. <ol style="list-style-type: none"> <input checked="" type="checkbox"/> Analyze...Descriptive Statistics...Frequencies. Move IVs to Variables box. <input checked="" type="checkbox"/> OK. Run Explore for quantitative variables. <ol style="list-style-type: none"> <input checked="" type="checkbox"/> Analyze...Descriptive Statistics...Explore. Move quantitative variables to Variables box. <input checked="" type="checkbox"/> Continue, then <input checked="" type="checkbox"/> OK. 	<ul style="list-style-type: none"> Less than 5% missing cases → use Listwise default. 5%–15% missing cases → replace missing values with estimated value by conducting Transform. <ol style="list-style-type: none"> <input checked="" type="checkbox"/> Transform...Replace Missing Values. Identify variable to be transformed and move to New Variable box. Identify new variable name (this occurs automatically). Select method of replacement (e.g., mean, median). <input checked="" type="checkbox"/> OK. More than 15% missing cases → delete variable from analysis.
<p>Univariate Outliers</p> <ul style="list-style-type: none"> Examine outliers for quantitative variable within each group. 	<ul style="list-style-type: none"> Run Explore. <ol style="list-style-type: none"> <input checked="" type="checkbox"/> Analyze...Descriptive Statistics...Explore. Move DVs to Dependent List box. Move IVs to Factor List box. <input checked="" type="checkbox"/> Statistics. Check Descriptives and Outliers. <input checked="" type="checkbox"/> Continue, then Plots. Check Boxplots:Factor levels together and Stem-and-Leaf. <input checked="" type="checkbox"/> Continue, then <input checked="" type="checkbox"/> OK. 	<ul style="list-style-type: none"> More than 90–10 split between categories → delete variable from analysis. Small number of outliers → delete severe outliers. Small to moderate number of outliers → replace with accepted minimum or maximum value by conducting Recode. <ol style="list-style-type: none"> <input checked="" type="checkbox"/> Transform...Recode Into Different Variables. Select variable to be transformed and move to Input Variable → Output Variable box. Type in new variable name under Output Variable Name box. <input checked="" type="checkbox"/> Change. <input checked="" type="checkbox"/> Old and New Values. Identify value to be changed under Old Value. Under New Value, identify appropriate new value; <input checked="" type="checkbox"/> Add. After all necessary values have been recoded, check All Other Values under Old Value. Check Copy Old Value(s) under New Value; <input checked="" type="checkbox"/> Add. <input checked="" type="checkbox"/> Continue, then <input checked="" type="checkbox"/> OK.
<p>Univariate Normality</p> <ul style="list-style-type: none"> Examine normality for quantitative variable within each group. 	<ul style="list-style-type: none"> Run Explore. <ol style="list-style-type: none"> <input checked="" type="checkbox"/> Analyze...Descriptive Statistics...Explore. Move DVs to Dependent List box. Move IVs to Factor List box. <input checked="" type="checkbox"/> Statistics. Check Descriptives. <input checked="" type="checkbox"/> Continue, then <input checked="" type="checkbox"/> OK. <input checked="" type="checkbox"/> Plots. Check Histograms and Normality Plots with tests. <input checked="" type="checkbox"/> Continue, then <input checked="" type="checkbox"/> OK. 	<ul style="list-style-type: none"> Transform variable (see Figure 3.3) using Compute. <ol style="list-style-type: none"> <input checked="" type="checkbox"/> Transform...Compute Variable. Under Target Variable, identify new variable name. Identify appropriate function and move to Numeric Expression(s) box. Identify variable to be transformed and move within the function equation <input checked="" type="checkbox"/> OK.

Figure 3.41. Steps for Screening Grouped Data (continued).

<p>Univariate Homoscedasticity</p> <ul style="list-style-type: none"> Examine homogeneity of variances between/among groups. 	<ul style="list-style-type: none"> Run Explore to conduct Levene's Test. <ol style="list-style-type: none"> Analyze...Descriptive Statistics...Explore. Move DVs to Dependent List box. Move IVs to Factor List box. Plots. Check Untransformed under Spread vs. Level with Levene Test. Continue, then OK. 	<ul style="list-style-type: none"> <i>p</i> value is significant at .05 → reevaluate univariate normality and consider transformations.
<p>Multivariate Outliers</p> <ul style="list-style-type: none"> Examine quantitative variables together by group for outliers. 	<ul style="list-style-type: none"> Conduct Regression to test Mahalanobis distance. <ol style="list-style-type: none"> Analyze...Regression...Linear. Identify a variable that serves as a case number and move to Dependent Variable box. Identify all appropriate quantitative variables and move to Independent(s) box. Save. Check Mahalanobis under Distances. Continue, then OK. Determine chi-square (χ^2) critical value at $p < .001$. Conduct Explore to test outliers for Mahalanobis chi-square (χ^2). <ol style="list-style-type: none"> Analyze...Descriptive Statistics...Explore. Move MAH_1 to Dependent List box. Leave Factor List box empty. Statistics. Check Outliers. Continue, then OK. 	<ul style="list-style-type: none"> Delete outliers for participants when χ^2 exceeds critical χ^2 at $p < .001$.
<p>Multivariate Normality, Linearity</p> <ul style="list-style-type: none"> Examine normality and linearity of variable combinations by group. 	<ul style="list-style-type: none"> Create Scatterplot Matrix. <ol style="list-style-type: none"> Graphs...Legacy Dialogs... Scatter/Dot. Scatter Matrix. Define. Identify appropriate quantitative variables and move to Matrix Variables. Move grouping variable to Columns, under Panel by. OK. 	<ul style="list-style-type: none"> Scatterplot shapes are not close to elliptical shapes → reevaluate univariate normality and consider transformations.
<p>Multivariate Homogeneity of Variance—Covariance</p> <ul style="list-style-type: none"> Examine homogeneity of variance-covariance between/among groups. 	<ul style="list-style-type: none"> Conduct MANOVA using Multivariate to run homogeneity tests. <ol style="list-style-type: none"> Analyze...General Linear Model...Multivariate. Move DVs to Dependent Variables box. Move IVs to Fixed Factor(s) box. Options. Check Homogeneity Tests. Continue, then OK. 	<ul style="list-style-type: none"> <i>p</i> value is significant at .025 or .01 → reevaluate univariate normality and consider transformations.

Figure 3.41 continues on the next page.

Figure 3.42. Steps for Screening Ungrouped Data (continued).

Examination & Process	SPSS Procedure	Technique for “Fixing”
<p>Missing Data</p> <ul style="list-style-type: none"> Examine missing data for each variable. 	<ul style="list-style-type: none"> Run Explore for quantitative variables. 1. Analyze...Descriptive Statistics...Explore. 2. Move quantitative variables to Variables box. 3. Continue, then OK. 	<ul style="list-style-type: none"> Less than 5% missing cases → use Listwise default. 5%–15% missing cases → replace missing values with estimated value by conducting Transform. <ol style="list-style-type: none"> Transform...Replace Missing Values. Identify variable to be transformed and move to New Variable box. Identify new variable name (this occurs automatically). Select method of replacement (e.g., mean, median). OK. More than 15% missing cases → delete variable from analysis.
<p>Univariate Outliers</p> <ul style="list-style-type: none"> Examine outliers for quantitative variable within each group. 	<ul style="list-style-type: none"> Conduct Regression to test Mahalanobis distance. <ol style="list-style-type: none"> Analyze...Regression...Linear. Identify a variable that serves as a case number and move to Dependent Variable box. Identify all appropriate quantitative variables and move to Independent(s) box. Save. Check Mahalanobis under Distances. Continue, then OK. Determine chi-square (χ^2) critical value at $p < .001$. Conduct Explore to test outliers for Mahalanobis chi-square (χ^2). <ol style="list-style-type: none"> Analyze...Descriptive Statistics...Explore. Move MAH_1 to Dependent List box. Leave Factor List box empty. Statistics. Check Outliers. Continue, then OK. 	<ul style="list-style-type: none"> Small number of outliers → delete severe outliers. Small to moderate number of outliers → replace with accepted minimum or maximum value by conducting Recode. <ol style="list-style-type: none"> Transform...Recode Into Different Variables. Select variable to be transformed and move to Input Variable → Output Variable box. Type in new variable name under Output Variable Name box. Change. Old and New Values. Identify value to be changed under Old Value. Under New Value, identify appropriate new value. Add. After all necessary values have been recoded, check All Other Values under Old Value. Check Copy Old Value(s) under New Value. Add. Continue, then OK.
<p>Univariate Normality</p> <ul style="list-style-type: none"> Examine normality for quantitative variable within each group. 	<ul style="list-style-type: none"> Run Explore. <ol style="list-style-type: none"> Analyze...Descriptive Statistics...Explore. Move DVs to Dependent List box. Statistics. Check Descriptives. Continue. Plots. Check Histograms and Normality Plots with tests. Continue, then OK. 	<ul style="list-style-type: none"> Transform variable (see Figure 3.3) using Compute. <ol style="list-style-type: none"> Transform...Compute Variable. Under Target Variable, identify new variable name. Identify appropriate function and move to Numeric Expression(s) box. Identify variable to be transformed and move within the function equation OK.

Figure 3.42. Steps for Screening Ungrouped Data (continued).

<p>Multivariate Outliers</p> <ul style="list-style-type: none"> Examine quantitative variables together for outliers. 	<ul style="list-style-type: none"> Conduct Regression to test Mahalanobis distance. <ol style="list-style-type: none"> Analyze...Regression...Linear. Identify a variable that serves as a case number and move to Dependent Variable box. Identify all appropriate quantitative variables and move to Independent(s) box. Save. Check Mahalanobis under Distances. Continue, then OK. Determine chi-square (χ^2) critical value at $p < .001$. Conduct Regression to test Mahalanobis distance. <ol style="list-style-type: none"> Analyze...Descriptive Statistics...Explore. Move <i>MAH_1</i> to Dependent List box. Leave Factor List box empty. Statistics. Check Outliers. Continue, then OK. 	<ul style="list-style-type: none"> Delete outliers for participants when χ^2 exceeds critical χ^2 at $p < .001$.
<p>Multivariate Normality, Linearity</p> <ul style="list-style-type: none"> Examine normality and linearity of variable combinations. 	<ul style="list-style-type: none"> Create Scatterplot Matrix. <ol style="list-style-type: none"> Graphs...Chart Builder. Select Scatter/Dot. Select Scatterplot Matrix icon and drag to Chart Preview area. Identify appropriate quantitative variables (one at a time) and drag to Scattermatrix Area. OK. 	<ul style="list-style-type: none"> Scatterplot shapes are not close to elliptical shapes → reevaluate univariate normality and consider transformations.
<p>Multivariate Homogeneity of Variance—Covariance</p> <ul style="list-style-type: none"> Examine standardized residuals to predicted values. 	<ul style="list-style-type: none"> Create residual plot using Regression. <ol style="list-style-type: none"> Analyze...Regression...Linear. Move DV to Dependent Variable box. Move IVs to Independent(s) Variable box. Plots. Select ZRESID for Y-axis. Select ZPRED for X-axis. Continue, then OK. 	<ul style="list-style-type: none"> Residuals are clustered at the top, bottom, left, or right area in plot → reevaluate univariate normality and consider transformations.

Exercises for Chapter 3

This exercise utilizes the data set *schools-a.sav*, which can be downloaded from this website:

www.routledge.com/9781138289734

1. You are interested in investigating if being above or below the median income (*medloinc*) impacts ACT means (*act94*) for schools. Complete the necessary steps to examine univariate grouped data in order to respond to the questions below. Although deletions and/or transformations may be implied from your examination, all steps will examine original variables.
 - a. How many participants have missing values for *medloinc* and *act94*?
 - b. Is there a severe split in frequencies between groups?
 - c. What are the cutoff values for outliers in each group?
 - d. Which outlying cases should be deleted for each group?
 - e. Analyzing histograms, normal Q-Q plots, and tests of normality, what is your conclusion regarding normality? If a transformation is necessary, which one would you use?
 - f. Do the results from Levene's test for equal variances indicate homogeneity of variance? Explain.
2. Examination of the variable of *scienc93* indicates a substantial to severe positively skewed distribution. Transform this variable using the two most appropriate methods. After examining the distributions for these transformed variables, which produced the better alteration?
3. You are interested in studying predictors (*math94me*, *loinc93*, and *read94me*) of the percentage graduating in 1994 (*grad94*).
 - a. Examine univariate normality for each variable. What are your conclusions about the distributions? What transformations should be conducted?
 - b. After making the necessary transformations, examine multivariate outliers using Mahalanobis distance. Which cases should be deleted?

- c. After deleting the multivariate outliers, examine multivariate normality and linearity by creating a Scatterplot Matrix.
- d. Examine the variables for homoscedasticity by creating a residuals plot (standardized vs. predicted values). What are your conclusions about homoscedasticity?

CHAPTER 4

FACTORIAL ANALYSIS OF VARIANCE

STUDENT LEARNING OBJECTIVES

After studying Chapter 4, students will be able to:

1. Compare and contrast a one-way ANOVA and a factorial ANOVA.
2. Explain the generic form of the null hypothesis in a one-way ANOVA.
3. Interpret the basic formula for the calculation of an F ratio.
4. Discuss the purpose of *post hoc* testing in ANOVAs.
5. Describe all of the possible hypotheses in a two-way analysis of variance.
6. Differentiate between ordinal and disordinal interactions.
7. Develop research questions appropriate for both one-way and factorial ANOVAs.
8. Describe the nature of the partitioning of sums of squares variability in a two-way ANOVA.
9. Test data sets for group differences by following the appropriate SPSS guidelines provided.

Many students enrolled in an introductory statistics course find univariate analysis of variance to be one of the most challenging topics that is covered. This is probably due to the complex nature of the formulae and the related hand calculations. After all, compared to a simple t test, the nine separate equations and subsequent calculations that make up a univariate analysis of variance seem absolutely overwhelming. The advice we provide to our students is that they should not get lost in the equations and/or the numbers as they proceed through the calculations. Rather, they should try to remain focused on the overall purpose of the test itself. (By the way, we also provide the same advice as we progress through our discussions of multivariate statistical techniques.) With that in mind, a brief review of univariate analysis of variance (ANOVA) is undoubtedly warranted.

I. UNIVARIATE ANALYSIS OF VARIANCE

The **univariate** case of ANOVA is a hypothesis-testing procedure that simultaneously evaluates the significance of mean differences on a dependent variable (DV) between two or more treatment conditions or groups (Agresti & Finlay, 2009). The treatment conditions or groups are defined by the various levels of the independent variable (IV), or *factor* in ANOVA terminology. We will limit our discussion here to a *one-way ANOVA*, which studies the effect that one factor has on one dependent variable. For instance, we might be interested in examining mean differences in achievement test scores (DV) for various school settings (IV), namely urban, suburban, and rural. Our research question might be stated, “Are there

significant differences in achievement test scores for urban, suburban, and rural schools?" A second example could involve an investigation of mean differences in student performance in a college course based on four different styles of instruction. Here, the research question could be, "Do college students enrolled in the same college course perform differently based on the type of instruction they receive?"

The null hypothesis in a one-way ANOVA states that there is no difference among the treatment conditions or groups. Using the first example, the null hypothesis would state that there is no difference in achievement test scores between urban, suburban, and rural schools. Restating this hypothesis using proper statistical notation would give us

$$H_0: \mu_{\text{Urban}} = \mu_{\text{Suburban}} = \mu_{\text{Rural}}$$

where μ represents the various group means.

Similarly, the alternative, or research, hypothesis for this scenario says that at least one of the group or treatment means is significantly different from the others. It is not necessary to state the alternative hypothesis using proper notation because that would require accounting for every pairwise comparison (e.g., $\mu_{\text{Urban}} = \mu_{\text{Suburban}}$, $\mu_{\text{Urban}} = \mu_{\text{Rural}}$, $\mu_{\text{Suburban}} = \mu_{\text{Rural}}$), which can become extremely cumbersome when there are more than three levels of the IV. In addition, in order to disprove the null hypothesis, we need only to find one pair that is significantly different, and at this point, we are not concerned with which pair that might be. Thus, restating the alternative hypothesis would give us

$$H_1: \text{At least one group mean is different from the others.}$$

Therefore, this leaves the researcher with two possible interpretations of the results of a one-way ANOVA:

- There really are no differences between the treatment conditions or groups. Any observed differences are due only to chance or sampling error (*fail to reject H_0*).
- Any observed differences between the conditions or groups represent real differences in the population (*reject H_0*).

In order to decide between these two possible outcomes, the researcher must rely on the specific and appropriate test statistic. The test statistic for ANOVA is the ***F* ratio**, and it has the following structure:

$$F = \frac{\text{variance between participants}}{\text{variance expected due to chance (error)}} \quad (\text{Equation 4.1})$$

Note that the *F* ratio is based on *variances* as opposed to mean *differences* (Gravetter & Wallnau, 2008) because we are now dealing with more than two groups. In a research situation involving two groups, for which a *t* test would be appropriate, it is relatively simple to determine the size of the differences between two sample means. Simply subtract the mean of the first group from the mean of the second group. However, if a third group is added, it becomes much more difficult to describe the difference among the sample means. The solution is to compute the variance for the set of sample means. If the three sample means cluster closely together (i.e., have small differences), the variance will be small. If the means are spread out (i.e., have large differences), the variance will be larger. Analysis of variance actually takes the total variability and analyzes, or *partitions*, it into two separate components. The variance calculated in the numerator in the equation for the *F* ratio, as indicated above, provides a singular value that describes the differences between the three sample means. Thus, the numerator is referred to as the ***between-groups variability***. The variance in the denominator of the *F* ratio, often referred to as the ***error variance*** (or ***variability***) or ***within-groups variability***.

groups variability, provides a measure of the variance that the researcher could expect due simply to chance factors, such as sampling error.

There are two possible causes or explanations for the differences that occur between groups or treatments (Gravetter & Wallnau, 2008).

1. The differences are due to the *treatment effect*. In this case, the various treatments, or group characteristics, actually cause the differences. For instance, students in the three different school settings (i.e., urban, suburban, and rural) performed differently on their achievement tests due to the setting in which they received instruction. Therefore, changing the school locations results in an increase or decrease in achievement-test performance.
2. The differences occur *simply due to chance*. Because all possible random samples from a population consist of different individuals with different scores, even if there is no treatment effect at all, you would still expect to see some differences—most likely *random* differences—in the scores.

Therefore, when the between-groups variability is measured, we are actually measuring differences due to either the effect of the treatment or to chance. In contrast, the within-groups variability is caused only by chance differences. Within each treatment or group, all participants in that sample have been exposed to the same treatment or share the same characteristic. The researcher has not done anything that would result in different scores. Obviously, individuals within the same group will likely have different scores, but these differences are due to random effects. The within-groups variability provides a measure of the differences that exist when there is no treatment effect that could have caused those differences (Gravetter & Wallnau, 2008).

If we substitute these explanations for variability into Equation 4.1, expressing each component in terms of its respective sources, we obtain the following equation:

$$F = \frac{\text{treatment effect} + \text{differences due to chance}}{\text{differences due to chance}} \quad (\text{Equation 4.2})$$

Thus, the *F* ratio becomes a comparison of the two partitioned components of the total variability. The value of the ratio will be used to determine whether differences are large enough to be attributed to a treatment effect, or if they are due simply to chance effects. When the treatment has no effect, any differences between treatment groups are due only to chance. In this situation, the numerator and denominator of the *F* ratio are measuring the same thing and should be roughly equivalent (Gravetter & Wallnau, 2008). That is,

$$F = \frac{0 + \text{differences due to chance}}{\text{differences due to chance}}$$

In this case, we would expect the value of the *F* ratio to be approximately equal to 1.00. A value near 1.00 indicates that the differences between the groups are roughly the same as would be expected due to chance. There is no evidence of a treatment effect; therefore, our statistical decision would be to “fail to reject H_0 ” (i.e., the null hypothesis cannot be rejected and remains a viable explanation for the differences). However, if a treatment effect exists, the numerator of the *F* ratio will be larger than the denominator (due to the nonzero value of the treatment effect), which will result in an *F* value greater than 1.00. If that value is substantially larger than 1.00 (i.e., larger than the critical value for *F*), a statistically significant treatment effect is indicated. In this situation, we would conclude that at least one of the sample means is significantly different from the others, permitting us to “reject H_0 ” (i.e., the null hypothesis is not true).

Earlier, in our discussions of the alternative hypothesis, we stated that we were not concerned with which groups were different from which other groups. Because we have now rejected the null hypothesis, it is important for us to be able to discuss specifically where the differences lie. Are students in urban schools different from those in rural schools? Or from students in suburban *and* rural schools? Or do the differences occur between suburban and rural schools? In this situation, a post hoc comparison is the appropriate procedure to address these questions.

Post hoc tests, also known as ***multiple comparisons***, enable the researcher to compare individual treatments two at a time, a process known as ***pairwise comparisons***. If the goal of the analysis was to determine whether any groups differed from each other, one might ask why we did not just conduct a series of independent-samples *t* tests. The reason is fairly simple: Conducting a series of hypothesis tests within a single research or analysis situation results in the accumulation of the risk of a Type I error. For instance, in our school setting analysis problem, there would be three pairwise comparisons: urban versus suburban, urban versus rural, and suburban versus rural. If each test was conducted at an α level equal to .05, the overall risk of a Type I error for this analysis would be equal to three times the pre-established α level. In other words, the ***experimentwise alpha level*** would be equal to .15. Now we have increased our probability of making an error in judgment from 5% of the time to 15% of the time. Most of us would not be willing to assume a risk of that magnitude. Several techniques, which basically involve the simultaneous testing of all pairwise comparisons, exist for addressing this problem (Agresti & Finlay, 2009). These include the ***Scheffé test***, the ***Bonferroni test***, and ***Tukey's Honest Significant Difference (HSD)***. An excellent discussion of the use, advantages, and disadvantages of each of these multiple-comparison tests is provided by Harris (1998, pp. 371–398).

As with any inferential statistical test, assumptions must be met for there to be proper use of the test and interpretation of the subsequent results. You will recall that the assumptions for a one-way analysis of variance are identical to those for an independent-samples *t* test. Specifically, these assumptions are as follows:

1. The observations within each sample must be independent of one another.
2. The populations from which the samples were selected must be normal.
3. The populations from which the samples were selected must have equal variances (also known as *homogeneity of variance*).

Generally speaking, the one-way analysis of variance is robust to violations of the normality and homogeneity of variance assumptions (Harris, 1998).

II. FACTORIAL ANALYSIS OF VARIANCE

The goal of most research studies is to determine the extent to which variables are related and, perhaps more specifically, to determine if the various levels of one variable differ with respect to values on a given dependent variable. We have just finished examining one of the simplest cases of the latter situation—a one-way analysis of variance. However, examining variables in relative isolation (i.e., two variables at a time) is seldom very informative. When conducting social science research, human beings are nearly always the participants in the research study. When studying human behavior, it would be naïve to think that only one variable could influence another. Behaviors are usually influenced by “a variety of different variables acting and interacting simultaneously” (Gravetter & Wallnau, 2008). Research designs that include more than one factor are called ***factorial designs***. In the remainder of this chapter, we will investigate the simplest of factorial designs—a ***two-way analysis of variance***. A two-way analysis of variance, as the name suggests, consists of two IVs and one DV. We will limit our discussion here to designs where there

is a separate sample for each treatment condition or characteristic. Specifically, this is referred to as a two-factor, independent-measures design.

SECTION 4.1 PRACTICAL VIEW

Purpose

The purpose of factorial analysis of variance is to test for mean differences with respect to some DV, similar to a simple one-way analysis of variance. However, in the case of a two-way analysis of variance design, there are now two IVs. Factorial analysis of variance allows the researcher not only to test the significance of group differences (based on levels of the two IVs), but also to test for any interaction effects between levels of IVs. The two-way ANOVA actually tests three separate hypotheses simultaneously in one analysis. Two hypotheses test the significance of the levels of the two IVs separately, and the third tests the significance of the interaction of the levels of the two IVs.

For instance, assume we have a research design composed of two IVs (Factor *A* and Factor *B*, each with two levels) and one DV. Furthermore, we are interested in evaluating the mean differences on the DV produced by either Factor *A* or Factor *B*, or by Factor *A* and Factor *B* in combination. The structure of this two-factor design is shown in Figure 4.1.

Figure 4.1. Basic Structure for a Two-Factor Design, With Each IV Having Two Levels.

		Factor <i>B</i>	
		Level 1	Level 2
Factor <i>A</i>	Level 1	Scores for n participants in A_1B_1	Scores for n participants in A_1B_2
	Level 2	Scores for n participants in A_2B_1	Scores for n participants in A_2B_2

Any differences produced by either Factor *A* or Factor *B*, independent of the other, are called **main effects**. In Figure 4.1, there are two main effects: (1) If there are differences between Level 1 and Level 2 of Factor *A*, there is a main effect due to Factor *A*; (2) If there are differences between Level 1 and Level 2 of Factor *B*, there is a main effect due to Factor *B*. Remember that we still do not know if these main effects are statistically significant—we must still conduct hypothesis tests to determine significance. These main effects provide us with two of the three hypotheses in a two-factor design. The main effect due to Factor *A* involves the comparison of its two levels. The null hypothesis for this main effect states that there is no difference in the scores due to the level of *A*. Symbolically, the null hypothesis would be

$$H_0: \mu_{A_1} = \mu_{A_2}$$

The alternative hypothesis states that there is a difference in scores due to the level of *A*. The alternative hypothesis would appear as

$$H_1: \mu_{A_1} \neq \mu_{A_2}$$

In similar fashion, the hypotheses corresponding to the main effects for Factor *B* would be represented by

$$H_0: \mu_{B_1} = \mu_{B_2}$$
$$H_1: \mu_{B_1} \neq \mu_{B_2}$$

The test statistic for the hypothesis tests of the main effects is again the *F* ratio. The formula used for the calculation of the *F* ratio in this case is identical to Equation 4.1.

In addition to testing for the effects of each factor individually, a two-factor ANOVA allows the researcher to evaluate mean differences that are the result of unique combinations of levels of the two factors. This is accomplished by comparing all of the cell means. A *cell* is defined as each combination of a particular level of one factor and a particular level of the second factor. In the example in Figure 4.1, there are four cells. The *cell means* are simply the averages of all the scores that fall into each cell. When mean differences exist in a given design that are not explained by the main effects, that design contains an interaction between factors. **Interaction between factors** occurs when the effect of one factor depends on different levels of the other factor (Gravetter & Wallnau, 2008).

In order to evaluate whether a significant interaction effect exists, the ANOVA procedure first computes any mean differences that cannot be explained by the main effects. After any additional mean differences are identified, they are also evaluated using an *F* ratio. However, the nature of the *F* ratio changes somewhat:

$$F = \frac{\text{variance not explained by main effects}}{\text{variance expected due to chance (error)}} \quad (\text{Equation 4.3})$$

The null hypothesis for the test of an interaction effect states that there is no interaction between Factor *A* and Factor *B*. In other words, there are no differences between any of the cell means that cannot be explained by simply summing the individual effects of Factor *A* and Factor *B* (Harris, 1998). The alternative hypothesis states that there is a significant interaction effect. That is, there are mean differences that cannot be explained by Factor *A* or Factor *B* alone.

Another way of identifying interaction effects between factors is to examine line plots of the cell means. Essentially, whenever there is an interaction, the lines on the graph will not be parallel (Aron, Aron, & Coups, 2006). Line plots are a graphical method of depicting a pattern of differences among different levels of the factors. Realize that the lines do not have to cross in order to be significantly nonparallel. For instance, Figure 4.2 shows two possible situations that could occur when examining interaction effects.

If there is an interaction between the levels of Factor *A* and Factor *B*, the distance between B_1 and B_2 on A_1 will be significantly different from the distance between B_1 and B_2 on A_2 (Newman, Newman, Brown, & McNeely, 2005). Clearly in Figure 4.2(a), these two distances are not significantly different; therefore, there is no interaction in this case. However, in Figure 4.2(b), a substantial discrepancy between the two distances exists and represents an interaction between Factor *A* and Factor *B*.

There are two types of interaction that can be displayed graphically—ordinal and disordinal interactions. An interaction is said to be an **ordinal interaction** when the lines plotted on the graph do not cross within the values of the graph (Kennedy & Bush, 1985; Newman, Newman, Brown, & McNeely, 2005). In contrast, an interaction is a **disordinal interaction** when the lines plotted on the graph actually cross within the values of the graph. Figure 4.2(b) shows an interaction that is ordinal. An example of a disordinal interaction plot is presented in Figure 4.3.

When interpreting the results of a two-way analysis of variance, one should always examine the significance of the interaction first. If the interaction is significant, it does not make much sense to interpret any main effects. Knowing that two IVs in combination result in a significant effect on the DV is more informative than determining that one and/or the other IV has individual effects. If the interaction is significant, there may be situations when it is appropriate to evaluate the main effects. Specifically, it may make sense to evaluate the main effects when the interaction is ordinal, simply because the interaction is not as pronounced as in a disordinal interaction (Newman, Newman, Brown, & McNeely, 2005). If the interaction is not significant, then the researcher should proceed to the evaluation of the main effects. This is done separately for each factor. Whenever significant group differences are identified, it is appropriate to conduct follow-up post hoc tests in order to determine where specific differences lie.

Another indicator of the strength of relation is effect size, represented by *eta squared*. Effect size is calculated for each factor and for the interaction, and it indicates the amount of total variance that is explained by the IVs. An effect size of .50 or greater indicates a substantial result.

Figure 4.2. Cell Mean Plots Show (a) No Interaction and (b) Ordinal Interaction Between Factors.

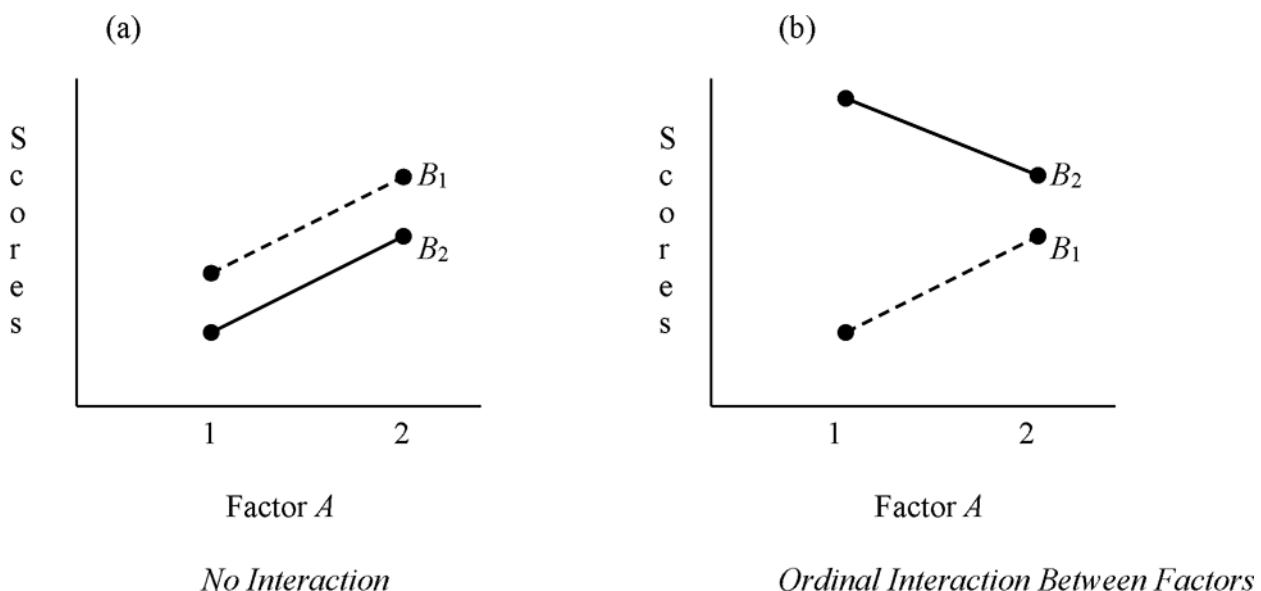
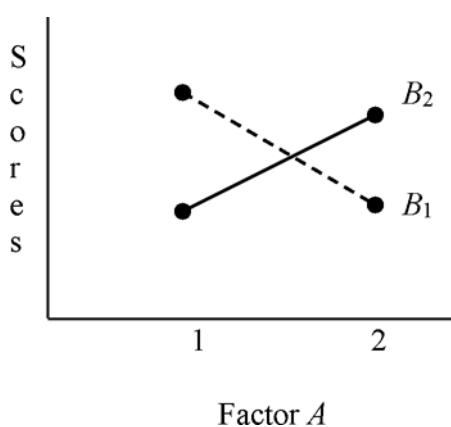


Figure 4.3. Cell Mean Plot Showing a Disordinal Interaction Between Factors.



Sample Research Questions

Suppose we were interested in analyzing the effect of gender and age category on income. Eight different groups of employees would be created, one for each combination of gender and age category (see Figure 4.4). The independent variables for this study would consist of gender and age category, and the dependent variable would be income.

Research questions should parallel the hypotheses that will be tested by this two-way ANOVA. Thus, this study should address the following research questions:

1. Are there significant mean differences for income between male and female employees?
2. Are there significant mean differences for income by age category among employees?
3. Is there a significant interaction on income between gender and age category?

Figure 4.4. Structure of a 4×2 , Two-Factor Design.

		Gender	
		<i>Female</i>	<i>Male</i>
Age Category	18–29	Income for n participants <i>female</i> —(18–29)	Income for n participants <i>male</i> —(18–29)
	30–39	Income for n participants <i>female</i> —(30–39)	Income for n participants <i>male</i> —(30–39)
	40–49	Income for n participants <i>female</i> —(40–49)	Income for n participants <i>male</i> —(40–49)
	50+	Income for n participants <i>female</i> —(50+)	Income for n participants <i>male</i> —(50+)

SECTION 4.2 ASSUMPTIONS AND LIMITATIONS

The validity of the results of a factorial ANOVA is dependent upon three assumptions that should be familiar after having studied previous independent-measures designs (e.g., independent-samples t test, one-way ANOVA, etc.). These assumptions are as follows:

1. The observations within each sample must be randomly sampled and must be independent of one another.
2. The distributions of scores on the dependent variable must be normal in the populations from which the data were sampled.
3. The distributions of scores on the dependent variable must have equal variances.

The assumption of independence is primarily a design issue, as opposed to a statistical one. Provided the researcher has randomly sampled and, more important, randomly assigned participants to treatments, it is usually safe to believe that this particular assumption has not been violated.

Methods of Testing Assumptions

Generally speaking, analysis of variance is robust to violations of the normality assumption, especially when the sample size is relatively large, and should not be cause for substantial concern (Gravetter & Wallnau, 2008). Slight departures from normality are to be expected, but even larger deviations will seldom have much effect on the interpretation of results (Kennedy & Bush, 1985). However, if the researcher wishes to have a greater degree of confidence concerning this assumption, there are methods of testing for violations of it. Initially, one can simply “eyeball” the distributions of the data in each cell by obtaining histograms and boxplots of the dependent variable. If the distribution appears to be relatively normal (i.e., there are no marked departures from normality in the form of extreme values), it is safe to assume that the assumption of normality has not been violated. If, however, the shape of the distribution is clearly nonnormal or there appear to be cases with extreme values, the researcher will probably want to submit the data to a more stringent test. For instance, the researcher may want to obtain numerical values for skewness and kurtosis and test their significance (as discussed in Chapter 3). Whether or not a distribution is normally distributed can also be tested by means of a chi-square goodness-of-fit test (Kennedy & Bush, 1985). A goodness-of-fit test purports to test the null hypothesis that a population has a specific shape or proportions. If one specifies in the null hypothesis that the distribution is normal, then the results of the chi-square test will indicate statistically whether or not the distribution is normal. A decision to reject the null hypothesis would lead the researcher to the conclusion that the assumption of normality had been violated and, thus, could possibly call into question the results of the analysis of variance. On the other hand, a decision to fail to reject the null hypothesis would permit the researcher to conclude that there was not enough evidence to support the claim that the distribution is not normal.

Violation of the assumption of homogeneity of variance is more crucial than a violation of the other assumptions. Initially, one should examine the standard deviations in the boxplots for each cell. If one suspects that the data do not meet the homogeneity requirement, a specific test should be conducted prior to beginning the analysis of variance. **Hartley's F_{\max}** test allows the researcher to use sample variances to determine if there are any differences among population variances. Hartley's test is simply a ratio of the largest group variance to the smallest group variance. The F_{\max} statistic is then compared to a critical value obtained from a table of the distribution of F_{\max} statistics. Another possible statistical test of homogeneity is **Cochran's test**, which is also a test of the ratio between variances. A third option is **Levene's test**, which

essentially consists of performing a one-way analysis of variance on data that has been transformed (Kennedy & Bush, 1985). All three of these methods are statistical tests of inference and involve the evaluation of a significance level (if calculated by means of a computer program) or of a test statistic by comparing the obtained value to a critical value (if calculated by hand). One disadvantage of these three tests is that they are appropriate only for situations that involve groups with equal sample sizes. **Bartlett's test** is an alternative appropriate for unequal ns , but it can provide misleading results because it is also extremely sensitive to nonnormality (Kennedy & Bush, 1985).

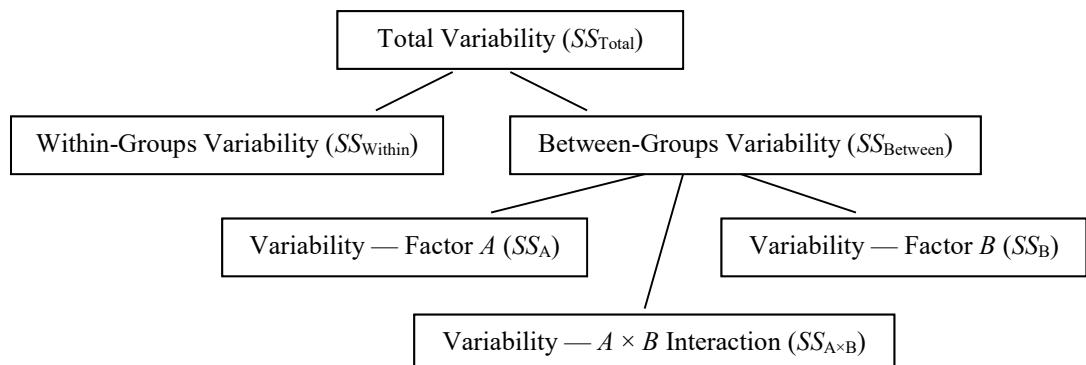
SECTION 4.3 PROCESS AND LOGIC

The Logic Behind Factorial ANOVA

The logic of a two-way analysis of variance is quite similar to the logic behind its one-way counterpart, with the addition of a few more computational complexities (Harris, 1998). Recall that in a one-way ANOVA, two estimates of variance are computed. One of these consists of only random error (i.e., the variance within groups, or MS_w). The other consists of random error as well as group differences (i.e., the variance between groups, or MS_B). If there are no differences between the population means of the various groups in the design, the estimates of the two variances will be approximately equal, resulting in the ratio of one (MS_B) to the other (MS_w) being approximately equal to 1.00. This ratio, of course, is the F ratio.

In a two-way ANOVA, the total variance is again partitioned into separate components. The within-groups variability remains the same (SS_w and, subsequently, MS_w). However, the between-groups variability is further partitioned into three subcomponents: variability due to Factor A (SS_A), variability due to Factor B (SS_B), and variability due to the interaction of Factor A and Factor B ($SS_{A \times B}$). This subsequent partitioning of sums of squares is shown in Figure 4.5.

Figure 4.5. Partitioning of Sums of Squares Variability in a Two-Way ANOVA.



In a two-way ANOVA, all between-groups variability components (i.e., MS_A , MS_B , and $MS_{A \times B}$) are compared to the within-groups variability (i.e., MS_w) individually. If the group means in the population are different for the various levels of Factor A (after removing or controlling for the effects of Factor B), then MS_A will be greater than MS_w . That is, F_A will be significantly greater than 1.00. Similarly, if the group means in the population are different for the various levels of Factor B (after removing the effects of Factor A), then MS_B will be greater than MS_w . That is, F_B will be significantly greater than 1.00. Finally, if the group means in the population are different for the various combinations of levels for Factor A and Factor B (after removing the individual effects of Factor A and Factor B), then $MS_{A \times B}$ will be greater than MS_w . That is, $F_{A \times B}$ will be significantly greater than 1.00.

The F ratio is clearly a measure of the statistical significance of the differences between group means or differences between combinations of levels of the IVs. There is, however, another measure that one can obtain that more directly examines the magnitude of the relationship between the independent and dependent variables. This measure is called **eta squared** (η^2) and is commonly viewed as the proportion of variance in the dependent variable explained by the independent variable(s) in the sample. Eta squared is viewed as a descriptive statistic and is interpreted as a measure of effect size (Harris, 1998).

Interpretation of Results

Because groups in factorial ANOVA are created by two or more factors or independent variables, it is important to determine if factors are interacting (working together) to affect the dependent variable. Typically, a line plot is created to graphically display any factor interaction. If lines overlap and criss-cross, factor interaction is present. Although a line plot may reveal some factor interaction, the ANOVA results may show that the interaction is not statistically significant. Consequently, it is important to determine interaction significance using the F ratio and p level for interaction generated from the factorial ANOVA test. If factors significantly interact such that they are working together to affect the dependent variable, we cannot determine the influence that each separate factor has on the dependent variable by looking at the main effects for each factor. Although factor main effects may be significant even while factor interaction is significant, caution should be used when drawing inferences about factor main effects. Effect size should also be analyzed to determine the strength of such effects.

In summary, the first step in interpreting the factorial ANOVA results is to determine if an interaction is present among factors by looking at the F ratio and its level of significance for the interaction. If no interaction is present, then look at the F ratio and its level of significance for each factor's main effect. This will indicate the degree to which each factor affects the dependent variable. If there is an interaction among factors, then you cannot discuss how each factor affects the dependent variable. Instead, you can only draw conclusions about how the factors work together to affect the dependent variable.

Continuing with the example of gender (*sex*), age category (*agecat4*), and differences in income (*rincom91*) using data set *career-d.sav*, data were first screened for missing data and outliers and then examined for test assumptions. Data screening led to the transformation of *rincom91* into *rincom2*, which eliminated all cases with income greater than or equal to 22. Although group distributions indicate moderate negative skewness, no further transformations were conducted¹ because ANOVA is not highly sensitive to nonnormality if group sample sizes are fairly large. Using techniques that will be presented in detail later in this chapter, homogeneity of variance was tested within the ANOVA. A line plot was then created for income means for each group. The nonoverlapping lines in Figure 4.6 demonstrate the lack of factor interaction between gender and age category. Figures 4.7 through 4.9 present the output from conducting the **Univariate ANOVA**. Levene's test for equal variances is presented in Figure 4.7 and indicates homogeneity of variance among groups, $F(7, 701) = 1.06, p = .387$. The ANOVA summary table (see Figure 4.8) indicates no significant factor interaction, $F(3, 701) = .97, p = .408$, partial $\eta^2 = .004$. Significant group differences were found in gender [$F(1, 701) = 40.48, p < .001$, partial $\eta^2 = .055$] and age category [$F(3, 701) = 21.64, p < .001$, partial $\eta^2 = .085$]. Although significant group differences were found, one should note the small effect size for each factor. Effect size indicates that a very small proportion of variance in income is accounted for by the IVs.

¹ Note that previously transformed variables that are no longer needed have been removed from data set *career-d.sav*.

Figure 4.6. Visual Representation of Interaction Between Factors.

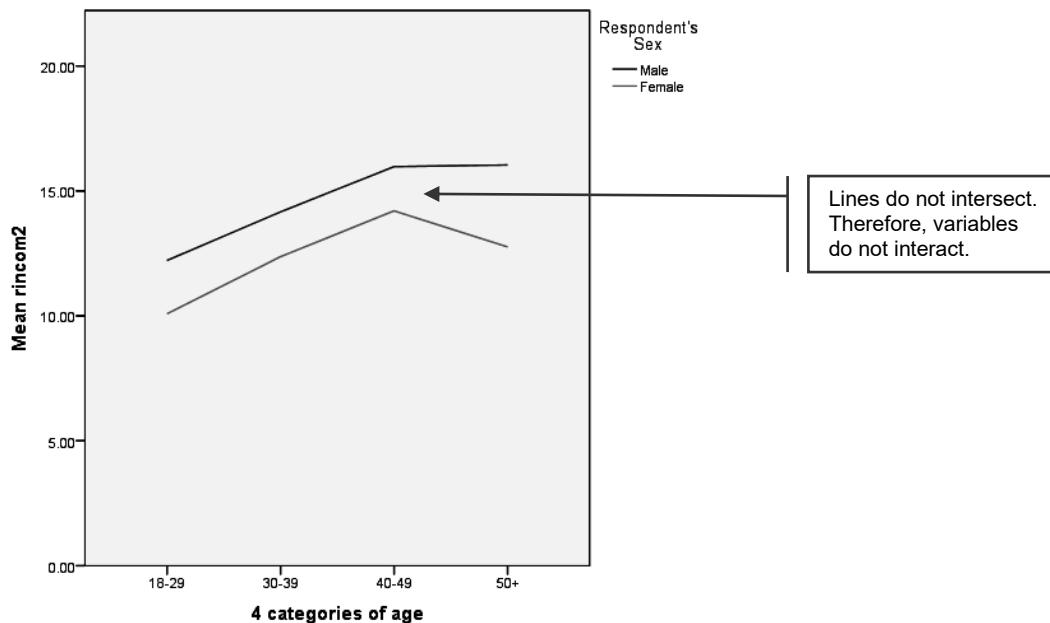


Figure 4.7. Levene's Test of Equality of Error Variances.

**Levene's Test of Equality of Error
Variances^a**

Dependent Variable: *rincom2*

F	df1	df2	Sig.
1.061	7	701	.387

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + sex + agecat4 +
sex * agecat4

Nonsignificance indicates homogeneity of variance.

Figure 4.8. ANOVA Summary Table for Interaction and Main Effects.

Tests of Between-Subjects Effects

Dependent Variable: rincom2

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2236.798 ^a	7	319.543	15.360	.000	.133
Intercept	121426.954	1	121426.954	5836.953	.000	.893
sex	842.073	1	842.073	40.478	.000	.055
agecat4	1350.516	3	450.172	21.640	.000	.085
sex * agecat4	60.316	3	20.105	.966	.408	.004
Error	14583.002	701	20.803			
Total	149966.000	709				
Corrected Total	16819.800	708				

a. R Squared = .133 (Adjusted R Squared = .124)

Main effects for each factor.

Interaction between factors.

Effect sizes are very small.

F ratios and *p* values show no significant interaction between factors. Age category and gender are significant.

Figure 4.9. Scheffé Multiple Comparisons for Student Level.

Multiple Comparisons

Dependent Variable: rincom2

Scheffé

(I) 4 categories of age	(J) 4 categories of age	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
18-29	30-39	-2.1172*	.49334	.000	-3.4997	-.7348
	40-49	-3.9165*	.51013	.000	-5.3461	-2.4870
	50+	-3.2930*	.54550	.000	-4.8216	-1.7643
30-39	18-29	2.1172*	.49334	.000	.7348	3.4997
	40-49	-1.7993*	.44207	.001	-3.0381	-.5605
	50+	-1.1757	.48245	.116	-2.5277	.1762
40-49	18-29	3.9165*	.51013	.000	2.4870	5.3461
	30-39	1.7993*	.44207	.001	.5605	3.0381
	50+	.6235	.49961	.669	-.7765	2.0236
50+	18-29	3.2930*	.54550	.000	1.7643	4.8216
	30-39	1.1757	.48245	.116	-.1762	2.5277
	40-49	-.6235	.49961	.669	-2.0236	.7765

Based on observed means.

The error term is Mean Square(Error) = 20.803.

*. The mean difference is significant at the .05 level.

Asterisk indicates which groups are significantly different.

Because the ANOVA results can only indicate group differences and cannot identify which groups are different, the **Scheffé post hoc test** was conducted to compare all group combinations and identify any significantly different pairs. Figure 4.9 presents the output from the Scheffé post hoc test for age category. A post hoc test was not conducted for gender because it has only two categories. Results indicate that the age category of 18–29 significantly differs in income from all other age categories. In addition, those 30–39 are significantly different in income from those 40–49 years of age.

Writing Up Results

Because several assumptions must be fulfilled when conducting a factorial ANOVA, the summary of results should first include the steps taken to screen the data. Group means and standard deviations of the dependent variable should first be reported in a table. A line plot of group means, as seen in Figure 4.6, should be created to provide a visual representation of any interaction among factors. ANOVA results (F ratios, degrees of freedom for the particular factor and error, levels of significance, and effect sizes) are then presented in a narrative format in the following order: main effects of each factor, interaction of factors, effect size of factors and interaction, and post hoc results. The following results statement applies the results from Figures 4.6 through 4.9.

The two-way analysis of variance was conducted to investigate income differences in gender and age category among employees. ANOVA results, presented in Table 1, show a significant main effect for gender [$F(1, 701) = 40.48, p < .001$, partial $\eta^2 = .055$] and age category [$F(3, 701) = 21.64, p < .001$, partial $\eta^2 = .085$]. Interaction between factors was not significant [$F(3, 701) = .97, p = .408$, partial $\eta^2 = .004$]. However, the calculated effect size for each factor indicates a small proportion of income variance is accounted for by each factor. The Scheffé post hoc test was conducted to determine which age categories were significantly different. Results revealed that the age category of 18–29 significantly differed in income from all other age categories. In addition, those 30–39 were significantly different in income from those 40–49 years of age.

Table 1
Two-way ANOVA Summary Table

Source	SS	df	MS	F	p	ES
Between treatments	2236.80	7	9.54			
Age category	1350.52	3	0.17	21.64	< .001	.085
Gender	842.07	1	2.07	40.48	< .001	.055
Age category × Gender	60.32	3	20.10	.97	.408	.004
Within treatments	14583.00	701	20.803			
Total	149966.00	709				

Note that each of the obtained F ratios is reported with its degrees of freedom and the degrees of freedom for error in parentheses. If space allows, an ANOVA summary table as shown in Table 1 should be created. In addition, a narrative results statement should be presented.

SECTION 4.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example that applies the entire process of conducting a factorial ANOVA: development of research questions and hypotheses, data-screening methods, test methods, interpretation of output, and a presentation of results. This example utilizes the data set *career-d.sav* from the website that accompanies this book (see p. *xiii*).

Problem

Employers and employees may be interested in determining if income is different for those with differing degrees and whether the differences are the same for individuals who are very satisfied or not very satisfied with their jobs. The following research questions and respective null hypotheses address the main effects for each factor and the possible interaction between factors.

Research Questions

Null Hypotheses

RQ1: Is there a difference in income between employees with different degrees?



H_01 : Income will not differ between employees with different degrees.

RQ2: Is there a difference in income between satisfied and unsatisfied employees?



H_02 : Income will not differ between satisfied and unsatisfied employees.

RQ3: Is there a difference in income based on degree held and level of satisfaction of employees?



H_03 : Income will not differ based on degree held and level of satisfaction of employees.

The independent variables are nominal and include highest degree attained (*degree*) and job satisfaction (*satjob2*). As seen in Figure 4.10, these two factors create 10 different groups for comparison on the dependent variable of respondents' income in 1991 (*rincom91*), which is an interval/ratio variable. Note that the variable of *rincom91* was previously transformed to *rincom2* to eliminate participants who reported income greater than or equal to a value of 22, and it was further transformed to eliminate outliers of values of 3 or less. This transformed variable (*rincom3*) will be utilized in the following example.

Figure 4.10. 2 × 5 Factor Design of Respondents' Income by Degree and Job Satisfaction.

Highest Degree Attained

	< H.S. Diploma	H.S. Diploma	Junior College	Bachelor's	Graduate
Very Satisfied With Job	Income for n employees who are very satisfied with jobs and have less than a H.S. diploma.	Income for n employees who are very satisfied with jobs and have a H.S. diploma.	Income for n employees who are very satisfied with jobs and have a junior college degree.	Income for n employees who are very satisfied with jobs and have a bachelor's degree.	Income for n employees who are very satisfied with jobs and have a graduate degree.
Not Very Satisfied With Job	Income for n employees who are not very satisfied with jobs and have less than a H.S. diploma.	Income for n employees who are not very satisfied with jobs and have a H.S. diploma.	Income for n employees who are not very satisfied with jobs and have a junior college degree.	Income for n employees who are not very satisfied with jobs and have a bachelor's degree.	Income for n employees who are not very satisfied with jobs and have a graduate degree.

Methods and SPSS "How To"

This section demonstrates the steps for conducting a factorial ANOVA using the **Univariate** procedure with the preceding *career-d.sav* example. This procedure analyzes the independent effects of the factors and covariates as well as the interaction among factors. In addition, different models may be used to estimate the sum of squares.

The first step in conducting a factorial ANOVA was to examine the data for missing participants and outliers and ensure that test assumptions were fulfilled. The **Explore** procedure was conducted to evaluate outliers and normality. Stem-and-Leaf plots revealed that several groups within *degree* and *satjob2* have outlying values of 3 or less. Therefore, *rincom2* was transformed again to *rincom3* in

order to eliminate participants who reported income of 3 or less. **Explore** was conducted again with *rincom3*. Figure 4.11 displays the Stem-and-Leaf plots for *satjob2*. Tests of normality indicated nonnormal distributions for the majority groups. Group histograms revealed slight to substantial negative skewness. However, *rincom3* will not be transformed again because ANOVA is not heavily dependent upon fulfilling the normality assumption as long as group sample sizes are adequate. Homogeneity of variance will be examined within the ANOVA.

Figure 4.11. Stem-and-Leaf Plots for Job Satisfaction.

rincom3 Stem-and-Leaf Plot for
satjob2= Not very satisfied

Frequency Stem & Leaf

8.00	4 . 00000000
4.00	5 . 0000
4.00	6 . 0000
4.00	7 . 0000
9.00	8 . 000000000
24.00	9 . 00000000000000000000000000
24.00	10 . 00000000000000000000000000
33.00	11 . 00000000000000000000000000000000
29.00	12 . 00000000000000000000000000000000
27.00	13 . 00000000000000000000000000000000
35.00	14 . 00000000000000000000000000000000
41.00	15 . 00000000000000000000000000000000
36.00	16 . 00000000000000000000000000000000
24.00	17 . 00000000000000000000000000000000
27.00	18 . 00000000000000000000000000000000
15.00	19 . 0000000000000000
14.00	20 . 0000000000000000
17.00	21 . 0000000000000000

Stem width: 1.00
 Each leaf: 1 case(s)

rincom3 Stem-and-Leaf Plot for
satjob2= Very satisfied

Frequency Stem & Leaf

3.00	Extremes (=<5.0)
2.00	6 . 00
3.00	7 . 000
7.00	8 . 0000000
9.00	9 . 000000000
11.00	10 . 00000000000
22.00	11 . 000000000000000000000000
13.00	12 . 000000000000
17.00	13 . 000000000000000
21.00	14 . 000000000000000
36.00	15 . 0000000000000000000000000000
41.00	16 . 0000000000000000000000000000
25.00	17 . 0000000000000000000000000000
30.00	18 . 0000000000000000000000000000
23.00	19 . 0000000000000000000000000000
8.00	20 . 00000000
26.00	21 . 0000000000000000000000000000

Stem width: 1.00
 Each leaf: 1 case(s)

The next step was to create a line plot using **Chart Builder** that displays any interaction between the factors. To produce this line plot, the following steps were conducted:

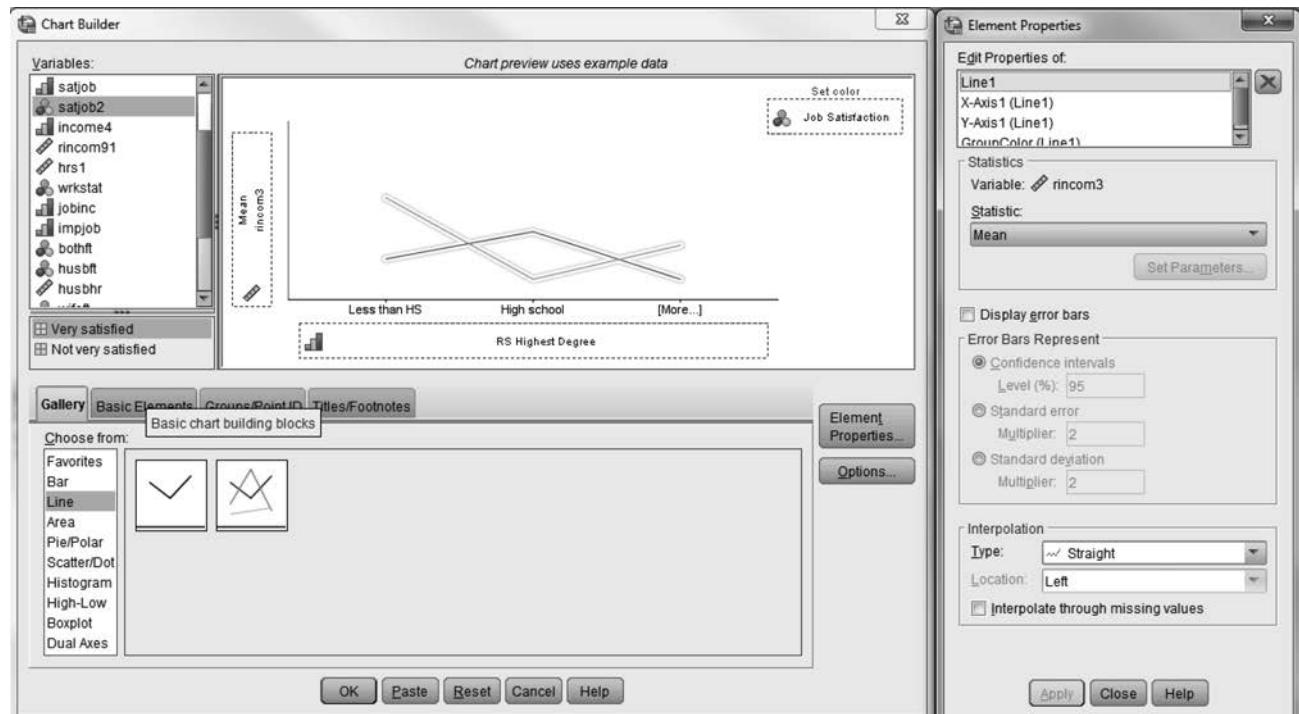
Graphs

Chart Builder

Chart Builder dialog box (see Figure 4.12)

Under **Gallery**, select **Line**. Click the graph with multiple lines and drag it to the **Chart Preview Area**. Select *degree* and move it to the *x*-axis area. Select *rincom3* and move it to the *y*-axis area. Select *satjob2* and move it to the **Set Color** area. Click **OK**. (The Chart Builder output and the outputs for all the steps that follow are presented and discussed beginning on p. 90.)

Figure 4.12. Chart Builder Dialog Box.



We will now conduct the factorial ANOVA using **Univariate**. Select the following menus:

Analyze

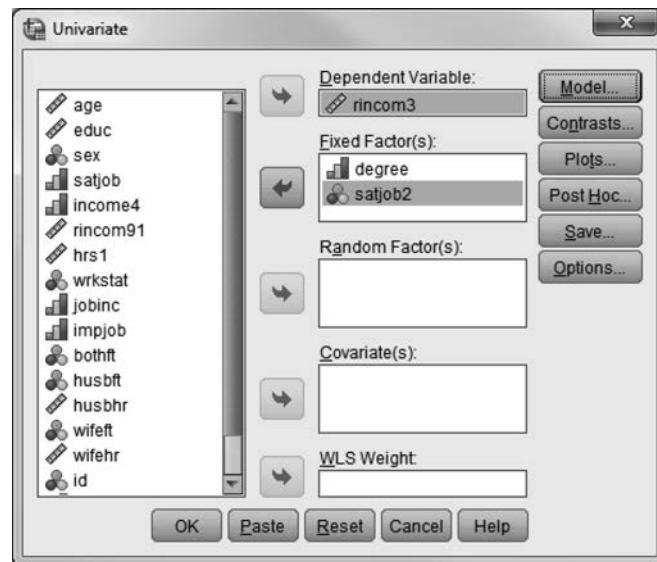
General Linear Model

Univariate

Univariate dialog box (see Figure 4.13)

Once in this dialog box, identify one DV (*rincom3*) and move it into the **Dependent Variable** box. Then select the independent variables (*degree* and *satjob2*) that create the groups and move them into the **Fixed Factor(s)** box. After you have defined the variables, click the **Model** button.

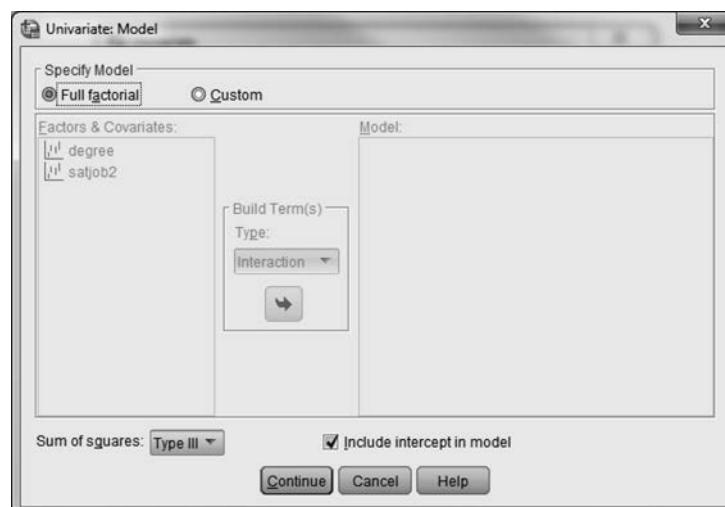
Figure 4.13. Univariate Dialog Box.



Univariate:Model dialog box (see Figure 4.14)

This box allows you to choose between a full factorial model and a custom model. The full factorial model, which we will use here, is the default and is usually the most appropriate because it will test all main effects and interactions. The custom model will test only the main effects and interactions you select. If building a custom model, be sure to select interaction variables together and click the **Build Term(s)** button. The **Model** dialog box also allows you to identify the way in which Sum of Squares will be calculated. Type III is the default and the most commonly used because it is useful when subgroups vary in sample size but have no missing cells/subgroups. Type IV is used for models with empty cells/groups. Click **Continue**.

Figure 4.14. Univariate: Model Dialog Box.



Univariate: Options dialog box (see Figure 4.15)

This dialog box provides several options for display that can help you in examining your data. Three commonly used options are as follows:

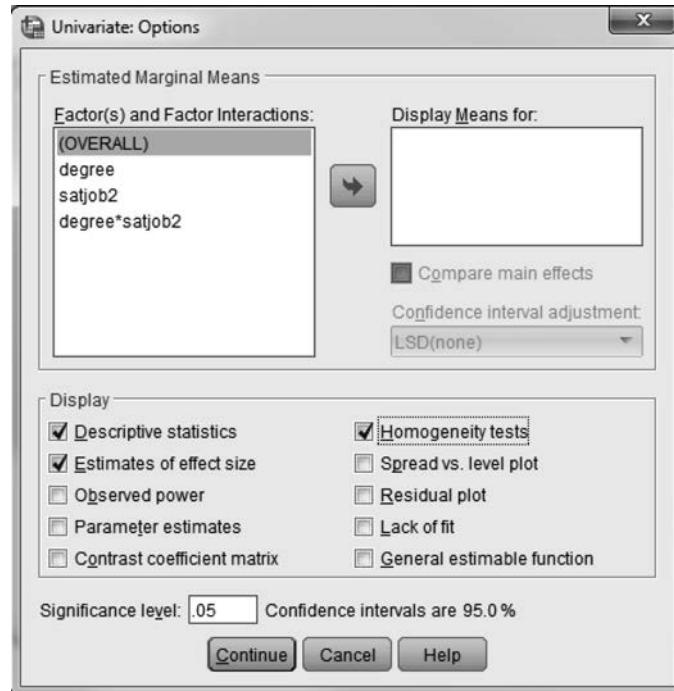
Descriptive statistics—produces means, standard deviations, and counts for subgroups.

Estimates of effect size—calculates eta squared; represents the amount of total variance that is explained by the independent variables.

Homogeneity tests—calculates Levene's statistic, the equality of variance test for subgroups.

Select these by placing a check in the box for each. Then click **Continue**.

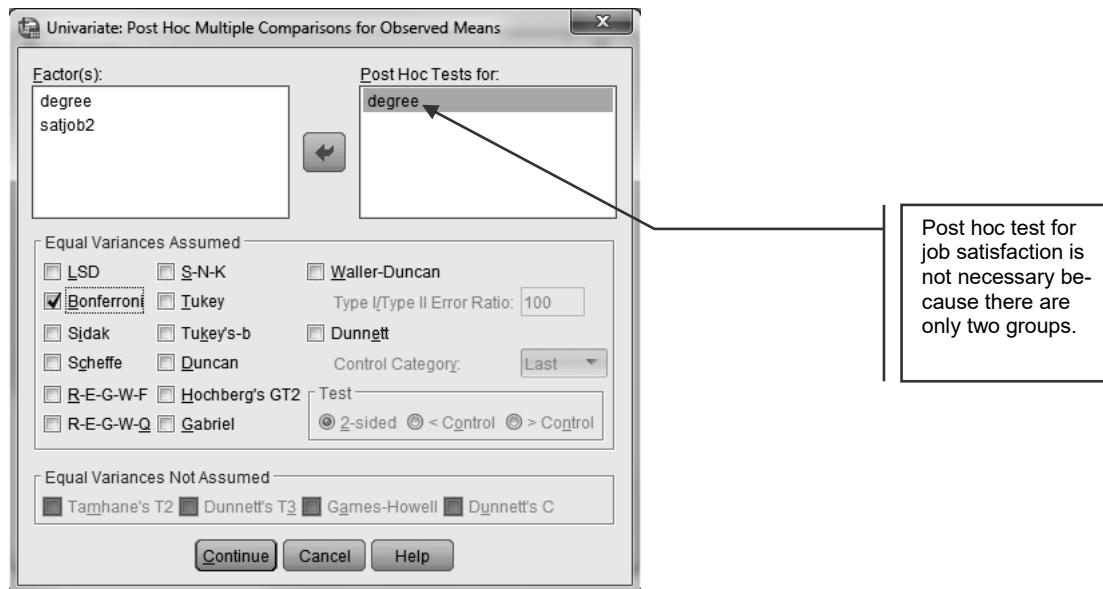
Figure 4.15. Univariate: Options Dialog Box.



Univariate: Post Hoc Multiple Comparisons for Observed Means dialog box (see Figure 4.16)

Post hoc tests allow one to identify which groups are different from one another within a single factor. However, you can test only those factors that have three or more groups. SPSS provides a variety of post hoc tests, grouped according to whether or not equal variances are assumed. Scheffé and Bonferroni tests are the most conservative. For our example, we have chosen to display means for *degree* and will utilize Bonferroni. Click **Continue** to exit this window, and then click **OK**.

Figure 4.16. Univariate: Post Hoc Multiple Comparisons for Observed Means Dialog Box.



Output and Interpretation of Results

Levene's test of equality of variances was conducted within ANOVA and indicates homogeneity of variance within groups (see Figure 4.17). The line plot of degree and job satisfaction (Figure 4.18) shows interaction between factors. Figure 4.19 reveals that the factor interaction is not statistically significant. Consequently, we can then determine if the main effects of each factor are significant. *F* ratios and levels of significance reveal that income is significantly different with respect to the levels of job satisfaction and degree. However, effect size (eta squared) is quite small for both factors. Figure 4.20 displays the results of Bonferroni's post hoc test for highest degree attained. This figure presents mean differences for every possible paired combination of highest degree attained. Significant mean differences at the .05 level are indicated by an asterisk. Results reveal that income of individuals with a graduate degree is significantly different from all other degree groups. Those with a high school diploma are not significantly different from those with no high school diploma or those with a junior college degree. In addition, income is not significantly different between junior college and bachelor degree holders. A post hoc test on job satisfaction is not necessary because this variable has only two categories.

Figure 4.17. Levene's Test of Equality of Variances.

Levene's Test of Equality of Error Variances^a

Dependent Variable: rincom3

F	df1	df2	Sig.
.700	9	662	.709

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

Levene's test is not significant, which indicates equal variances.

a. Design: Intercept + degree + satjob2
+ degree * satjob2

Figure 4.18. Line Plot of Interaction Between Degree and Job Satisfaction.

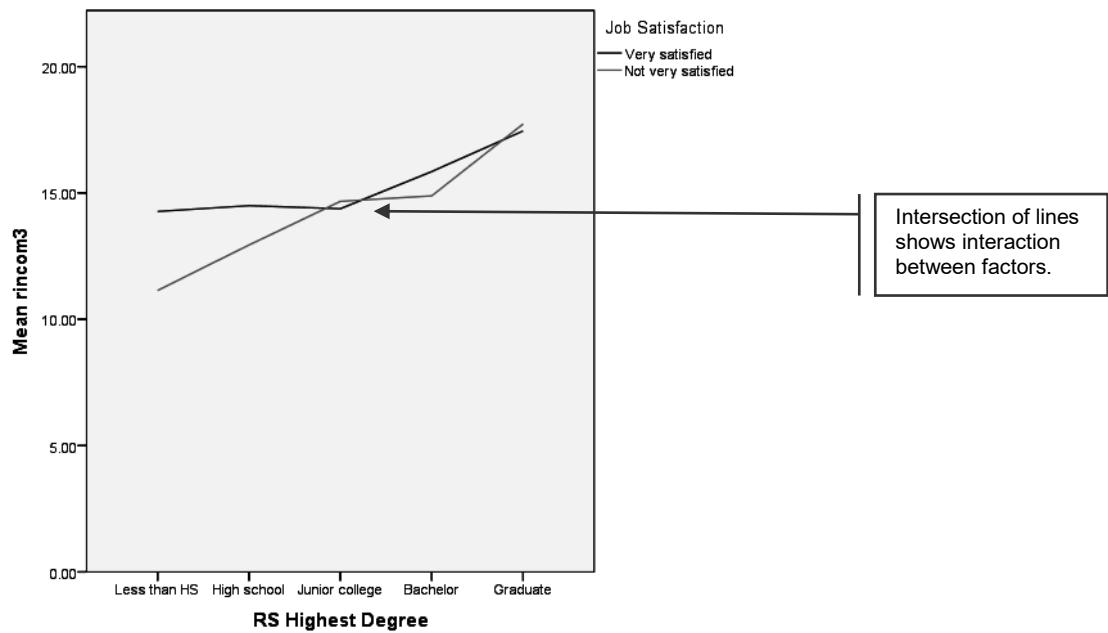


Figure 4.19. Univariate ANOVA Table.

Tests of Between-Subjects Effects

Dependent Variable: rincom3

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1687.354 ^a	9	187.484	14.513	.000	.165
Intercept	81072.058	1	81072.058	6275.675	.000	.905
degree	1197.372	4	299.343	23.172	.000	.123
satjob2	95.785	1	95.785	7.415	.007	.011
degree * satjob2	120.681	4	30.170	2.335	.054	.014
Error	8552.021	662	12.918			
Total	150312.000	672				
Corrected Total	10239.375	671				

a. R Squared = .165 (Adjusted R Squared = .153)

Income is significantly different among degrees and job satisfaction.

Effect size does not exceed .50.

The interaction between degree and satisfaction level is NOT significant.

Figure 4.20. Bonferroni Multiple Comparisons for Degree.

Multiple Comparisons						
		Dependent Variable: rincom3				
		Bonferroni				
(I) RS Highest Degree		Mean Difference (I-J)			95% Confidence Interval	
(J) RS Highest Degree		Std. Error	Sig.		Lower Bound	Upper Bound
Less than HS	High school	-.4122	.57518	.143	-.30321	.2078
	Junior college	-.23249*	.73953	.017	-.44077	-.2420
	Bachelor	-.30935*	.61575	.000	-.48277	-.13592
	Graduate	-.53777*	.67611	.000	-.72820	-.34735
High school	Less than HS	1.4122	.57518	.143	-.2078	3.0321
	Junior college	-.9127	.53901	.909	-.24308	.6054
	Bachelor	-.16813*	.35040	.000	-.26682	-.6944
	Graduate	-.39656*	.44806	.000	-.52275	-.27036
Junior college	Less than HS	2.3249*	.73953	.017	.2420	4.4077
	High school	.9127	.53901	.909	-.6054	2.4308
	Bachelor	-.7686	.58211	1.000	-.24081	.8709
	Graduate	-.30529*	.64562	.000	-.48712	-.12345
Bachelor	Less than HS	3.0935*	.61575	.000	1.3592	4.8277
	High school	1.6813*	.35040	.000	.6944	2.6682
	Junior college	.7686	.58211	1.000	-.8709	2.4081
	Graduate	-.22843*	.49908	.000	-.36899	-.8786
Graduate	Less than HS	5.3777*	.67611	.000	3.4735	7.2820
	High school	3.9656*	.44806	.000	2.7036	5.2275
	Junior college	3.0529*	.64562	.000	1.2345	4.8712
	Bachelor	2.2843*	.49908	.000	.8786	3.6899

Based on observed means.

The error term is Mean Square(Error) = 12.918.

*. The mean difference is significant at the .05 level.

Presentation of Results

The following report briefly summarizes the two-way ANOVA results from the previous example. Notice that Table 2 was created using data from Figure 4.19 and that Figure 4.17 is used to display factor interaction.

Data were screened to ensure that the assumptions of factorial ANOVA were fulfilled. To eliminate outliers, participants with income values of less than or equal to 3 or greater than or equal to 22 were removed. A univariate ANOVA was conducted. A summary of results is presented in Table 2. Main effect results revealed that income was significantly different among employees with differing degrees, $F(4, 662) = 23.17, p < .001$, partial $\eta^2 = .123$. Income was also significantly different for employees who were very satisfied or not very satisfied with their jobs, $F(1, 662) = 7.42, p = .007$, partial $\eta^2 = .011$. Although the interaction between degree and job satisfaction was not statistically significant, $F(4, 662) = 2.34, p = .054$, Figure 4.17 reveals some interaction. Estimates of effect size revealed low strength in associations. Bonferroni's post hoc test was conducted to determine which degree groups were significantly different in income. Results reveal that income of individuals with a graduate degree is significantly different from all other degree groups. Those with a high school diploma are not significantly different from those with no high school diploma or those with a junior college degree. In addition, income is not significantly different between junior college and bachelor's degree holders.

Table 2*Two-way ANOVA Summary Table*

Source	SS	df	MS	F	p	ES
Between treatments	1687.35	9	187.48			
Job satisfaction	95.79	1	95.79	7.42	.007	.011
Degree	1197.37	4	299.34	23.17	< .001	.123
Job satisfaction × Degree	120.68	4	30.17	2.34	.054	.014
Within treatments	8552.02	662	12.92			
Total	150312.0	672				

SECTION 4.5 VARIATIONS OF THE TWO-FACTOR DESIGN

Many variations of this design come under the heading of factorial designs. For instance, measures taken on participants may be dependent rather than independent. In other words, participants may each have a score at each level of the factor(s), as in a design that involves a pretest, a posttest, and an extended follow-up test. For this reason, these designs are often referred to as within-groups, repeated-measures, or randomized-block ANOVAs.

Another variation of the factorial ANOVA design involves fixed and random effects. A **fixed effect** is a factor in which all levels of interest of the IV have been included. For instance, if our study of interest involved an examination of the effect of gender, we would most likely include both females and males. In contrast, a **random effect** is a factor for which the levels represent only a sample of all possible levels to which we hope to be able to generalize the results. For instance, suppose we are interested in teachers' effectiveness in using a new approach to classroom instruction. We might include five secondary science teachers as the five levels of the variable called *teacher* in our study. Because we are interested not only in these five teachers (obviously, we hope to be able to generalize our results to the entire population of secondary science teachers), the variable *teacher* and its subsequent levels would be considered a random effect.

Another important distinction in factorial designs is that of crossed and nested designs. A **crossed** design is one in which all combinations of the levels of the IVs are included, or represented, by at least one observation. For instance, if we are concerned with how employees' scores on some motivation inventory are affected by gender (i.e., female or male) and education level (i.e., high school degree, college degree, or graduate degree), we would have at least one person in each of the six possible combined-level categories. In contrast, a **nested** design is one in which some possible combinations of levels are missing. For instance, we might study the effectiveness of three teachers using curriculum *A* and three different teachers using curriculum *B*. Notice that each teacher is paired with only one of the two curricula.

Finally, there are designs in which the *cells*, or combination of levels of IVs, do not contain equal numbers of participants. This is often the case in nonexperimental research designs. Unequal *ns* are usually a concern only when some of the assumptions of factorial analysis of variance have been violated. In this case, the nature of the statistical analysis becomes much more complex (Harris, 1998).

If interested in a more detailed account of any of the previously mentioned variations on factorial designs, many excellent resources include, but are not limited to, Agresti and Finlay (2009); Aron, Aron, and Coups (2006); Gravetter and Wallnau (2008); Harris (1998); and Kennedy and Bush (1985).

SUMMARY

The purpose of a factorial ANOVA is to determine group differences when two or more factors create these groups. Factorial ANOVA will test the main effect of each factor on the dependent variable and the interaction among factors. Usually, post hoc tests are conducted in conjunction with the ANOVA to determine which groups are significantly different. Prior to conducting the factorial ANOVA, data should be screened to ensure the fulfillment of test assumptions—*independence of observations, normal distributions of subgroups, and equal variances among subgroups*. The SPSS *Univariate ANOVA* table provides F ratios and p values that indicate the significance of factor main effects and interaction. If factor interaction is significant, then conclusions about the main effects of each factor are limited because factors are working together to affect the dependent variable. Post hoc results will identify which group combinations are significantly different. Figure 4.21 provides a checklist for conducting a factorial ANOVA.

KEYWORDS

- Bartlett's test
- between-groups variability
- Bonferroni test
- cell
- Cochran's test
- disordinal interaction
- error variance (or variability)
- eta squared (η^2)
- experimentwise alpha level
- F ratio
- factor
- factorial designs
- fixed effect
- Hartley's F_{\max} test
- interaction between factors
- Levene's test
- main effects
- multiple comparisons
- one-way ANOVA
- ordinal interaction
- pairwise comparisons
- random effect
- Scheffé post hoc test
- Scheffé test
- Tukey's Honest Significant Difference (HSD)
- two-way analysis of variance
- within-groups variability

Figure 4.21. Checklist for Conducting a Factorial ANOVA.

I. Screen Data

- a. Missing Data?
- b. Outliers?
 1. Run Outliers and review **Stem-and-Leaf plots** and **Boxplots** within **Explore**.
 2. Eliminate or transform outliers if necessary.
- c. Normality?
 1. Run Normality Plots with Tests within **Explore**.
 2. Review boxplots and histograms.
 3. Transform data if necessary.
- d. Homogeneity of Variance?
 1. Run Levene's Test of Equality of Variances within **Univariate**.
 2. Transform data if necessary.
- e. Factor Interaction?
 1. Create line plot of DV by IVs using **Chart Builder**.

II. Conduct ANOVA

- a. Run Factorial ANOVA with post hoc test.
 1. **Analyze... General Linear Model... Univariate.**
 - Move DV to **Dependent Variable** box.
 - Move IVs to **Fixed Factor** box.
 2. **Model.**
 - Full Factorial**; **Continue**.
 3. **Options.**
 - Check **Descriptive statistics**, **Estimates of effect size**, and **Homogeneity tests**; **Continue**.
 4. **Post hoc.**
 - Select Bonferroni; **Continue**; then **OK**.
- b. Interpret factor interaction.
- c. If no factor interaction, interpret main effects for each factor.

III. Summarize Results

- a. Describe any data elimination or transformation.
- b. Present line plot of factor interaction.
- c. Narrate main effects for each factor and interaction (F ratio, p value, and effect size).
- d. Draw conclusions.

Exercises for Chapter 4

The data in Questions 1 and 2 come from *profile-a.sav*, which can be downloaded from this website:

www.routledge.com/9781138289734

1. The table below presents means for the number of hours worked last week (*hrs1*) for individuals by general happiness (*happy*) and job satisfaction (*satjob2*). Using the data below, draw a line plot. Use the line plot to complete the following steps to estimate the factor main effects on the dependent variable and the interaction between factors.

	Very Satisfied With Job	Not Very Satisfied With Job
Very happy	43	40
Pretty happy	42	42
Not too happy	38	40

- a. Develop the appropriate research questions for main effects and interaction.
 - b. Do the factors interact? If so, do you think the interaction will be statistically significant? Explain.
 - c. Do you think that there will be a significant main effect for the factor of general happiness? If so, which groups do you think will be significantly different?
 - d. Do you think that there will be a significant main effect for the factor of job satisfaction?
2. Using your research questions from Question 1, complete a factorial ANOVA analysis and Bonferroni's post hoc test. The variables used were hours worked last week (*hrs1*), general happiness (*happy*), and job satisfaction (*satjob2*). Use the questions below to guide your interpretation of the output.
- a. Is factor interaction significant? Explain.
 - b. Are main effects significant? Explain.
 - c. How do these results compare with your estimation in Question 1?

3. Use the *salary-a.sav* data file and the *salary-b.sav* data file to determine if current salaries (*salnow*) are related to gender (*sex*) and minority status (*minority*).
- a. Develop the appropriate research questions for main effects and interaction.
 - b. Using *salary-a.sav*, evaluate your data to ensure that they meet the necessary assumptions.
 - c. Use *salary-b.sav*, which includes the transformed variable of *salnow3*, to run the appropriate analyses and interpret your results.
 - d. Write a results statement.

CHAPTER 5

ANALYSIS OF COVARIANCE

STUDENT LEARNING OBJECTIVES

After studying Chapter 5, students will be able to:

1. Describe what is meant by the term *concomitant variable*.
2. Discuss the process of partialing out the effects of a covariate.
3. Examine the differences between an analysis of variance and an analysis of covariance.
4. Describe the three main purposes for the use of analysis of covariance.
5. Discuss possible reasons for limiting the number of covariates used in a single analysis of covariance.
6. Develop research questions appropriate for both one-way and factorial ANOVAs.
7. Describe the logic behind the use of analysis of covariance.
8. Test data sets for group differences, and incorporating a covariate, by following the appropriate SPSS guidelines provided.

As mentioned in the previous chapter, there are numerous variations, extensions, and elaborations of the one-way analysis of variance technique. In this chapter, we discuss one such technique, analysis of covariance, which can be used to improve research design efficiency by adjusting the effect of variables that are related to the dependent variable. We begin by examining a basic comparison between analysis of variance and analysis of covariance, followed by the specific details of analysis of covariance.

I. ANALYSIS OF VARIANCE VERSUS ANALYSIS OF COVARIANCE

One-way analysis of variance compares the means on some dependent variable for several groups, as defined by the various levels of the independent variable. In many situations, especially in the social sciences, it is difficult to imagine that differences on a dependent variable could be attributed only to the effect of *one* independent variable. It would likely take very little persuasion to convince the reader that *many* variables may, in fact, affect that particular dependent variable. Often, a researcher may be able to identify one or more variables that also demonstrate an effect on the DV. If one is still interested only in the effect of the original IV on the DV, the effects of these “accompanying” variables, also known as *concomitant variables*, can be controlled for, or *partialed out* of, the results. In analysis of variance, the effects of any concomitant variables are simply ignored. The covariance analysis itself mirrors the ordinary analysis of variance, but only after the effect of the unwanted variable has been partialled out. The variable whose effects have been partialled out of the results is called the *covariate*. The results of the analysis are then interpreted just like any other analysis of variance.

In the previous chapter, we presented an example that investigated income level differences. Here, we will apply an analysis of covariance design to these same variables. Let us assume, for instance, that we want to determine whether or not job satisfaction, gender, or the interaction between them has an effect on

income level. However, our data set also includes a measure for level of education, and we would probably be justified in assuming that level of education has an effect on income level. In other words, income level differences may exist due to differences in the level of education of individuals in the sample. In general, individuals who have higher levels of education tend to have higher levels of income. If this is the case, we will never be able to accurately determine income differences due to job satisfaction or gender. Furthermore, the problem is that we are not concerned with the education level variable—it may not have been central to our research interests and/or research questions. Therefore, we would like to control for the effect of level of education on income level. Stated another way, level of education has been identified as the covariate for our analysis.

II. ANALYSIS OF COVARIANCE

While analysis of variance is similar to factorial analysis of variance, the use of analysis of covariance provides researchers with a technique that allows them to more appropriately analyze data collected in social science settings. The examination of the relationships among variables considered in relative isolation can be somewhat troubling, especially when dealing with human beings as the participants in a research study. Often, there are extraneous variables present that may influence the dependent measures. Analysis of covariance is an extension of analysis of variance where the main effects and interactions are assessed *after* the effects of some other concomitant variable have been removed. The effects of the covariate are removed by adjusting the scores on the DV in order to reflect initial differences on the covariate. In this chapter, we will discuss the purposes, proper applications, and interpretations of analysis of covariance.

SECTION 5.1 PRACTICAL VIEW

Purpose

The use of analysis of covariance essentially has three major purposes (Tabachnick & Fidell, 2007). The first of these purposes is to increase the sensitivity of the *F* tests of main effects and interactions by reducing the error variance, primarily in experimental studies. This is accomplished by removing from the error term (i.e., the within-groups variability) any unwanted, *predictable* variance associated with the covariate(s). Recall, for a moment, the generic equations for an *F* test as discussed in Chapter 4. The error term corresponds to the denominator of the formula for the calculation of an *F* statistic:

$$F = \frac{\text{variance between participants}}{\text{variance due to chance (error)}} = \frac{\text{variance between participants}}{\text{variance within participants}}$$

This predictable variance (or *systematic bias*) results from the use of intact groups that differ systematically on several variables and is best addressed through random assignment of participants to groups (Stevens, 2001). When random assignment is not possible due to constraints within the research setting, the inclusion of a covariate in the analysis can be helpful in reducing the error variance. The covariate is used to assess any undesirable variance in the DV. This variance is actually estimated by scores on the covariate. If the covariate has a substantial effect on the DV, a portion of the within-groups variability will be statistically removed, resulting in a smaller error term (i.e., a smaller denominator). This ultimately produces a larger value for the *F* statistic, therefore producing a more sensitive test (Stevens, 2001). In other words, we are now more likely to reject the null hypotheses concerning the main and interaction effects. An example of a source of this undesirable variance would be individual differences possessed by the participants prior to entry into the research study.

The second purpose of analysis of covariance involves a statistical adjustment procedure (Tabachnick & Fidell, 2007). This is most appropriately used in nonexperimental situations where participants cannot be randomly assigned to treatments. In this situation, analysis of covariance is used as a statistical matching procedure where the means of the DV for each group are adjusted to what they would be if all

groups had scored equally on the covariate. In this manner, instead of the ideal situation consisting of participants randomly assigned to groups/treatments, the researcher now has two (or more) groups containing participants who have been matched based on the covariate scores. Although this condition is certainly not as advantageous as having random assignment, it does improve the research design when random assignment is not feasible. Analysis of covariance is also used primarily for descriptive purposes in nonexperimental studies: The covariate improves the predictability of the DV, but there is no implication of causality. Tabachnick and Fidell (2007) warn that “...if the research question to be answered involves causality, [the use of] ANCOVA is no substitute for running an experiment” (p. 322).

The third purpose of analysis of covariance is to interpret differences in levels of the IV when several DVs are included in the analysis. This procedure is known as ***multivariate analysis of covariance (MANCOVA)*** and will be discussed in greater detail in the next chapter. Briefly, it is often the goal of a research study to assess the contribution of each DV to the significant differences in the IVs. One method of accomplishing this is to remove the effects of all other DVs by treating them as covariates in the analysis.

The most common use of analysis of covariance is its use for the first purpose described (i.e., increase the sensitivity of the F tests of main effects and interactions by reducing the error variance, primarily in experimental studies). A generic example of the classical application of ANCOVA involves the random assignment of participants to various levels of one or more IVs. The participants are then measured on one or more covariates. A commonly used covariate consists of scores on some pretest, measured in identical fashion as the DV (i.e., a posttest) but prior to the manipulation of the levels of the IV(s). This pretest measure is followed by manipulation of the IV (i.e., implementation of the treatment), which is then followed by the measurement of the DV. It is important to keep in mind that the covariate need not consist of a pretest. Demographic characteristics (e.g., level of education, socioeconomic level, gender, IQ, etc.) that differ from, but are related to, the DV can serve as covariates (Tabachnick & Fidell, 2007). Covariates can be incorporated in all variations of the basic ANOVA design, including factorial between-subjects, within-subjects, crossed, and nested designs, although analyses of these designs are not readily available in most computer programs (Tabachnick & Fidell, 2007).

Let us now apply this generic example to our current, concrete example. The example that we began discussing at the beginning of this chapter was actually a factorial analysis of covariance design. Recall that in factorial analysis of variance, the researcher not only tests the significance of group differences (based on levels of the two IVs), but he or she also tests for any interaction effects between levels of IVs. The additional component of the design of a factorial analysis of covariance is that the significance of group differences as well as the interaction effects is tested only *after* the effects of some covariate have been removed. Similar to the two-way ANOVA, the two-way ANCOVA actually tests three separate hypotheses simultaneously in one analysis. Two of the hypotheses test the significance of the levels of the two IVs separately, after removing the effects of the covariate, and the third tests the significance of the interaction of the levels of the two IVs, also after removing the effects of the covariate. Specifically, these three null hypotheses can be stated as follows:

$$H_0: \mu_{A_1} = \mu_{A_2} \text{ (main effect of variable A)}$$

$$H_0: \mu_{B_1} = \mu_{B_2} \text{ (main effect of variable B)}$$

$$H_0: \mu_{A_1 B_1} = \mu_{A_1 B_2} = \mu_{A_2 B_1} = \mu_{A_2 B_2} \text{ (interaction effect of variables A and B)}$$

where μ refers to the income level group means and the subscripts 1 and 2 indicate the two levels of the independent variables, or factors, A and B.

In our example, recall that we wanted to determine whether or not job satisfaction, gender, or the interaction between them has an effect on income level, after controlling for education level. The DV is the actual income as reported by each participant in our data set. Our two IVs—gender and job satisfaction—have two levels each. Therefore, we have a four-cell design, as depicted in Figure 5.1. Because we have identified education level as a covariate in our analysis, the variability in income levels that is attributable to level of education has been partialled out. In other words, the effect of education level on income has

been controlled. Remember that controlling for the effects of a covariate is a statistical procedure accomplished only by the computer during the analysis procedure.

A critical issue that needs to be addressed by the researcher in any analysis of covariance design is the choice of covariate(s). Generally speaking, any variables that should theoretically correlate with the DV or that have been shown to correlate with the DV on similar types of participants should be considered possible covariates (Stevens, 2001). Ideally, if quantitative variables are being considered, one should choose as covariates those variables that are significantly correlated with the DV and that have low correlations among themselves (in cases where more than one covariate is being used). If there exists a weak correlation between two covariates, they will each remove relatively unique portions of the error variance from the DV, which is advantageous because we want to obtain the greatest amount of total error reduction through the inclusion of covariates. On the other hand, if there exists a strong correlation between two covariates (e.g., $r > .80$), then those two covariates will essentially remove the same error variance from the DV. In other words, the inclusion of the second covariate will have contributed very little to improving the design and resultant analysis (Stevens, 2001). If categorical variables are being considered as covariates, the degree of relationship with the DV or with other covariates is not an issue.

Figure 5.1. Structure for a Two-Factor Analysis of Covariance Design.

Job Satisfaction	Gender	
	Female	Male
Very Satisfied With Job	Income* for $n = 140$ participants <i>female</i> — <i>very satisfied</i>	Income* for $n = 173$ participants <i>male</i> — <i>very satisfied</i>
Not Very Satisfied With Job	Income* for $n = 180$ participants <i>female</i> — <i>not very satisfied</i>	Income* for $n = 217$ participants <i>male</i> — <i>not very satisfied</i>

*Indicates that degree (education level) has been partialled out of income level means.

The *number* of covariates to be included in an analysis is also a decision that should not be taken lightly by the researcher. Huitema (1980, p. 161) provided a formula that can be used as guidance in determining the number of covariates to be included in a study, based on the number of groups and participants. The formula recommends limiting the number of covariates such that

$$\frac{C + (J - 1)}{N} < .10$$

where C is the number of covariates, J is the number of groups, and N is the total number of participants in the study. For instance, if we wanted to conduct a research study consisting of three groups and a total of 45 participants, then $(C + 2)/45 < .10$, or $C < 2.5$. Thus, we should probably include fewer than 2.5 (i.e., 1 or 2) covariates in this study. If the ratio is greater than .10, the adjusted means resulting from the analysis of covariance are likely to be unstable—that is, the adjusted means for our sample would be substantially different from other samples drawn from the same population (Stevens, 2001).

Sample Research Questions

In our study, as represented in Figure 5.1, we have two IVs with two levels each, creating four groups. After controlling for the effect of education level, we are concerned with investigating three components of the design:

- the main effect of gender on income level,
- the main effect of job satisfaction on income level, and
- the interaction effect between gender and job satisfaction on income level.

Remember that our research questions should parallel our null hypotheses as we have previously stated them. Therefore, this study should address the following research questions:

1. Are there significant mean differences for income level between males and females, after controlling for education level?
2. Are there significant mean differences for income level between individuals who are very satisfied with their jobs and those who are not very satisfied with their jobs, after controlling for education level?
3. Is there a significant interaction on income level between gender and job satisfaction, after controlling for education level?

SECTION 5.2 ASSUMPTIONS AND LIMITATIONS

The results of an analysis of covariance are valid only to the extent to which the assumptions are not violated. Analysis of covariance is subject to the same assumptions as analysis of variance. As a reminder, these are the assumptions:

1. The observations within each sample must be randomly sampled and must be independent of one another.
2. The distributions of scores on the dependent variable must be normal in the population from which the data were sampled.
3. The distributions of scores on the dependent variable must have equal variances.

Analysis of covariance also rests on three additional assumptions.

4. A linear relationship exists between the dependent variable and the covariate(s).
5. The regression slopes for a covariate are homogeneous (i.e., the slope for the regression line is the same for each group).
6. The covariate is reliable and is measured without error.

Methods of Testing Assumptions

Potential violations of the first three ANCOVA assumptions are examined in the same manner as they are in an analysis of variance design. Recall that the assumption of random sampling and independent observations really constitute issues of design and should be addressed prior to any collection of data. The assumption of normally distributed scores on the dependent variable can be assessed initially by inspection of histograms, boxplots, and normal Q-Q plots, but it is probably best tested statistically by examining the values (and the associated significance tests) for skewness and kurtosis and through the use of the Kolmogorov-Smirnov test. Finally, the assumption of homogeneity of variances is best tested using Box's test or one of three different statistical tests, namely Hartley's F_{\max} test, Cochran's test, or Levene's test.

The three additional assumptions, unique to analysis of covariance, address two issues that are familiar and one issue that is new to discussions of screening data. The first of these additional assumptions, that of a linear relationship between the covariate(s) and the dependent variable, is analogous to the general multivariate assumption of linearity as discussed in Chapter 3. The researcher needs to be concerned with this assumption only when the covariate is quantitative. If a nonlinear relationship exists between the quantitative covariate and DV, the error terms are not reduced as fully as they could be, and therefore, the group means are incompletely adjusted (Tabachnick & Fidell, 2007). A violation of this assumption could ultimately result in errors in statistical decision making. This assumption of linearity is roughly assessed by inspecting the bivariate scatterplots between the covariate and the DV. More precisely, however, one should test this assumption by obtaining and examining residuals plots comparing the standardized residuals to the predicted values for the dependent variable. If it is apparent that serious curvilinearity is present, one should then examine the within-cells scatterplots of the DV with each covariate (Tabachnick & Fidell, 2007). If curvilinear relationships are indicated, they may be corrected by transforming some or all of the variables. If transforming the variables would create difficulty in interpretations, one might consider eliminating the covariate that appears to produce nonlinearity and replacing it with another appropriate covariate (Tabachnick & Fidell, 2007).

Similar to the ANOVA/ANCOVA assumption of random selection and independent observations, a second ANCOVA assumption, that of the inclusion of a reliable covariate, is an issue of research design and is most appropriately addressed prior to data collection. Steps should be taken at the outset of a research study to ensure that covariates—and all variables—are measured as error-free as possible. Failure to reliably measure covariates can again result in a less sensitive F test, which could cause errors in statistical decision making.

The final ANCOVA assumption presents a new issue to the reader. This assumption states that the regression slopes for a covariate are homogeneous (i.e., that the slope for the regression line is the same for each group). This assumption is also referred to as **homogeneity of regression**. Regression, which is discussed in detail in Chapter 7, is a statistical technique in which the relationship between two variables can be used so that the DV can be predicted from the IV. Briefly, the method used in the prediction process is often referred to as the *regression line*, or the best-fitting line through a series of points, as defined by a mathematical equation. If the values for n participants on two variables are plotted as scatterplots, those series of points will form a unique shape. There will be an equally unique line that will mathematically satisfy the requirements to be the best-fitting line through those points. Because the selection of a covariate in ANCOVA is based upon the extent to which there exists a predictable relationship (i.e., a strong relationship) between the covariate and the DV, regression is incorporated into ANCOVA in order to predict scores on the DV based on knowledge of scores on the covariate (Kennedy & Bush, 1985). The homogeneity of regression slopes assumption states that the regression slopes (i.e., those best-fitting lines between the covariate and the DV) are equal for each group in the analysis. In a design consisting of three groups, each containing 15 participants, the assumption of homogeneous regression slopes is met if the slopes appear like those pictured in Figure 5.2(a). Figure 5.2(b) depicts a violation of the assumption of homogeneous regression slopes.

A violation of the assumption of homogeneous regression slopes is crucial with respect to the validity of the results of an analysis of covariance. If the slopes are unequal, it is implied that there is a different

DV-covariate slope in some cells of the design or that there is an interaction between the IV and the covariate (Tabachnick & Fidell, 2007). If an IV-covariate interaction exists, the relationship between the covariate and the DV is different at different levels of the IV(s). If one again examines Figure 5.2(b), it should be clear that, at low scores on the covariate, Group 2 outscores Group 1 on the DV; however, at higher covariate scores, Group 1 outscores Group 2 (and does so at an increasing rate). Therefore, the covariate adjustment needed for the various cells is different.

Various computer programs will provide a statistical test (specifically, an F test) of the assumption of homogeneity of regression slopes. The null hypothesis being tested in this case is that all regression slopes are equal. If the researcher is to continue in the use of analysis of covariance, he or she would hope to fail to reject that particular null hypothesis, thus indicating that the assumption is tenable and that analysis of covariance is an appropriate technique to apply. In SPSS, this is determined by examining the results of the F test for the interaction of the IV(s) by the covariate(s). If the F test is significant, then ANCOVA should *not* be conducted.

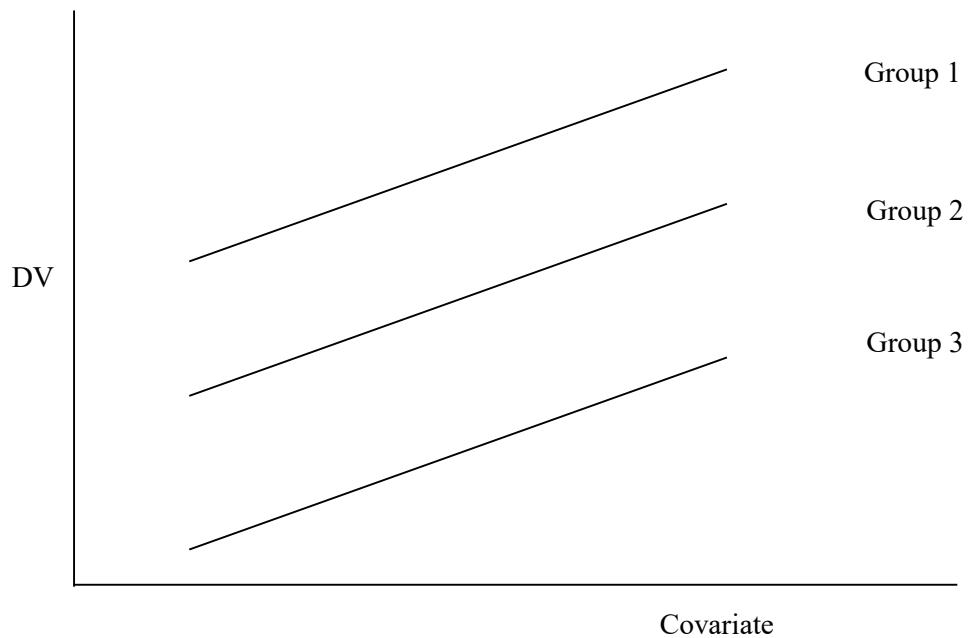
SECTION 5.3 PROCESS AND LOGIC

The Logic Behind ANCOVA

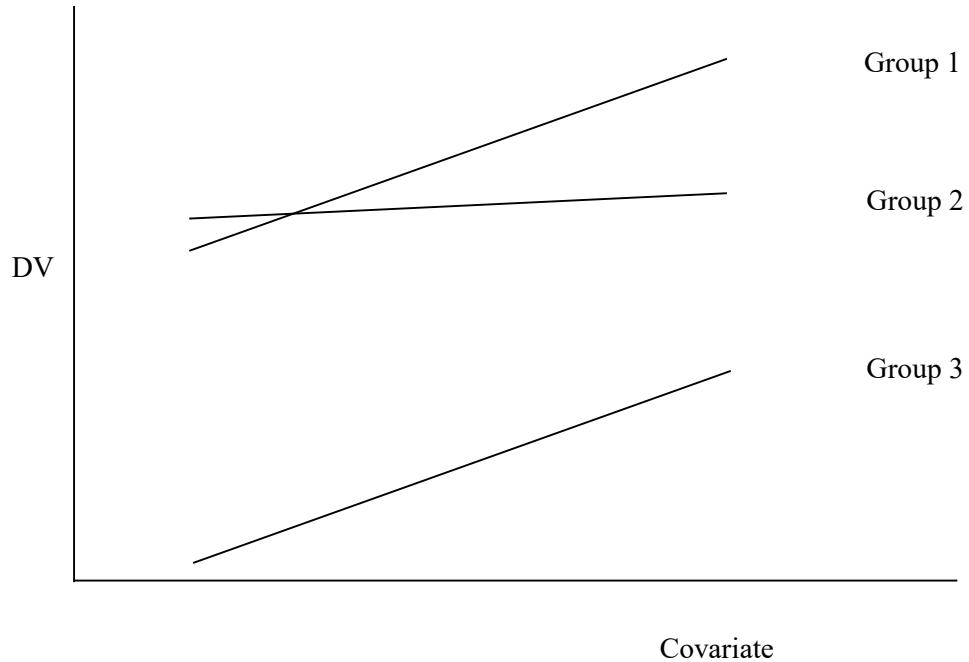
The logic of analysis of covariance is nearly identical to the logic behind analysis of variance. Recall that in an ANOVA, two estimates of variance are computed. One of these consists of only random error (i.e., the variance within groups, or MS_w). The other consists of random error as well as group differences (i.e., the variance between groups, or MS_B). If there are no differences between the population means of the various groups in the design, the estimates of the two variances will be approximately equal, resulting in the ratio of one (MS_B) to the other (MS_w) approximately equal to 1.00. This ratio, of course, is the F ratio.

Figure 5.2. Regression Lines Between a Dependent Variable and Covariate for Three Groups Depicting (a) Homogeneous Regression Slopes and (b) Heterogeneous Regression Slopes.

(a)



(b)



Analysis of covariance parallels the previous procedure, with one additional component: the adjustment of DV scores. Initially, the covariate is measured prior to the manipulation of the independent variable (i.e., implementation of the treatments). Following the implementation of the treatments, the dependent measures are collected. The initial phase of the analysis involves the statistical adjustment of the DV group means in order to control for the effects of the covariate on the DV. From this point on, the logic behind analysis of covariance is the same as that behind analysis of variance. A ratio of random errors plus treatment effect (MS_B) to random error (MS_w —reduced due to the adjustment of DV group means) is obtained. The magnitude of the resultant F ratio is then evaluated in order to determine the significance of the effect of group differences, after controlling for the covariate, on the DV.

As an extension of this one-way ANCOVA design, a factorial ANCOVA again mirrors the factorial ANOVA design, following the adjustment of DV means. The total variance is partitioned into separate components. The within-groups variability remains the same (SS_w and, subsequently, MS_w) but has been reduced due to the inclusion of the covariate. The between-groups variability is partitioned into the variability due to Factor A (SS_A), variability due to Factor B (SS_B), and variability due to the interaction of Factor A and Factor B ($SS_{A \times B}$). Refer to Figure 4.5 in Chapter 4 (p. 80) for a review.

Then, similar to the two-way ANOVA, all between-groups variability components (i.e., MS_A , MS_B , and $MS_{A \times B}$) are compared to the reduced within-groups variability (i.e., MS_w) individually. If the group means in the population are different for the various levels of Factor A (after removing or controlling for the effects of Factor B), then MS_A will be greater than MS_w . That is, F_A will be significantly greater than 1.00. The same logic applies to testing the effects of Factor B (after removing the effects of Factor A) and for testing the various combinations of levels for Factor A and Factor B (after removing the individual effects of Factor A and Factor B).

In addition to the F ratio, we can again obtain a measure of effect size. Recall that eta squared (η^2) is a measure of the magnitude of the relationship between the independent and dependent variables and is interpreted as the proportion of variance in the dependent variable explained by the independent variable(s) in the sample after partialing out the effects of the covariate(s).

Interpretation of Results

Interpretation of ANCOVA results is similar to that of ANOVA. However, with the inclusion of covariates, interpretation of a preliminary or custom ANCOVA is necessary in order to test the assumption of homogeneity of regression slopes. Basically, this preliminary analysis tests for the interaction between the factors (IVs) and covariates. In addition, homogeneity of variance will also be tested in this preliminary analysis. If the F test of factor-covariate interaction is significant, then the full ANCOVA should not be conducted. If factor-covariate interaction is not significant, then one can proceed with interpreting the Levene's test as well as proceeding with the full ANCOVA. If the Levene's test is not significant, then homogeneity of variance is assumed. Once homogeneity of regression slopes and homogeneity of variance have been established, then the full factorial can be conducted.

Interpretation of the full ANCOVA results must also take into account the interaction among the factors (IVs). Consequently, line graphs are typically created to graphically represent any interaction between/among factors based upon the DV. Overlapping lines indicate factor interaction, which may limit the inferences drawn from the analysis. Fortunately, the ANCOVA procedure also calculates the significance of such interaction. If interaction is statistically significant, the main effect for each factor on the DV is not a valid indicator of effect because factors are working together to affect the DV. Effect size (η^2) for each factor and interaction is also a good indicator of the strength of these effects.

Because ANCOVA is adjusting group means as if participants scored equally on the covariate(s), the utility of each covariate is also analyzed within ANCOVA. F ratios and p values for each covariate indicate the degree to which the covariate significantly influences the DV. Care must be taken here because if the covariate was measured after the introduction of the treatment—and if the covariate is affected by the treatment variable—the change on the covariate may be related to a change in the DV. Therefore, when the

covariate adjustment is made, a portion of the treatment effect will be removed. This will lead to lower F values and higher p values for the main effects and interactions. Although a significant covariate decreases the likelihood of significant factor main effects, the results should present a more accurate picture of group differences when adjusted for some covariate(s). Effect size for the covariate(s) should also be examined to determine the amount of variance accounted for by each covariate.

In summary, the first step in interpreting the results of ANCOVA is to determine if factor interaction is present by examining the F ratio and p value for the interaction. If no interaction is present, then each factor's main effect can be reliably interpreted. F ratios, p values, and effect sizes are examined for each factor's main effect and indicate the degree to which each adjusted factor affects the DV. A comparison of adjusted group means can also indicate which groups differ from one another. This is often helpful because post hoc analyses are not available in *SPSS ANCOVA*. The influence of covariate(s) should also be assessed by examining F ratios, p values, and η^2 .

We continue with our example from Chapters 3 and 4 using the *career-e.sav* data set, which seeks to determine the effect of gender (*sex*) and job satisfaction (*satjob2*) on respondents' income (*rincom91*) while controlling for highest degree attained (*degree*). Data were previously screened for missing data and outliers and were then examined for testing assumptions. Data screening led to the transformation of *rincom91* to *rincom2* in order to eliminate all cases with income equal to or exceeding 22. Although group distributions of *rincom2* are slightly negatively skewed, no further transformation was done. Linearity of the covariate and DV was not assessed because *degree* is categorical. Factor interaction was then investigated by creating a line plot of income for gender and job satisfaction (see Figure 5.3). The line plot reveals no factor interaction. Next, homogeneity of regression slopes and homogeneity of variance were evaluated by conducting a preliminary **Univariate** ANCOVA. The ANCOVA summary table indicates no significant interaction between the factors and covariate, $F(3, 703) = 2.18, p = .089$ (see Figure 5.4). In addition, the Levene's test reveals equal variances among groups, $F(3, 706) = .602, p = .614$ (see Figure 5.5). With test assumptions fulfilled, **Univariate** ANOVA was conducted, and it produced the output in Figures 5.6 through 5.8. Figures 5.6 and 5.7 present the unadjusted and adjusted means for each group. The ANCOVA summary table is presented in Figure 5.8 and indicates no factor interaction. Main effects for gender [$F(1, 705) = 37.68, p < .001$, partial $\eta^2 = .051$] and job satisfaction [$F(1, 705) = 10.87, p = .001$, partial $\eta^2 = .015$] are significant. However, one should be cautious in drawing inferences regarding the effect of each IV because effect sizes are very small.

Figure 5.3. Line Plot of Gender and Job Satisfaction for Income.



Figure 5.4. ANCOVA Summary Table for Homogeneity of Regression Slopes.

Tests of Between-Subjects Effects

Dependent Variable: rincom2

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2733.846 ^a	6	455.641	22.655	.000	.162
Intercept	28091.978	1	28091.978	1396.740	.000	.665
sex	108.004	1	108.004	5.370	.021	.008
satjob2	261.418	1	261.418	12.998	.000	.018
degree	1295.667	1	1295.667	64.421	.000	.084
sex * satjob2 * degree	131.716	3	43.905	2.183	.089	.009
Error	14139.113	703	20.113			
Total	150407.000	710				
Corrected Total	16872.959	709				

a. R Squared = .162 (Adjusted R Squared = .155)

No significant interaction. Assumption of homogeneity of regression slopes is fulfilled.

Figure 5.5. Levene's Test of Homogeneity of Variance.

Levene's Test of Equality of Error Variances^a

Dependent Variable: rincom2

F	df1	df2	Sig.
.602	3	706	.614

Not significant. Assumption of equal variances is fulfilled.

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + sex + satjob2 + degree + sex * satjob2 * degree

Figure 5.6. Unadjusted Descriptive Statistics for Income by Gender and Job Satisfaction.

Descriptive Statistics

Dependent Variable: rincom2

sex	satjob2	Mean	Std. Deviation	N
Male	Very satisfied	15.3584	4.52498	173
	Not very satisfied	14.1198	4.98186	217
	Total	14.6692	4.81811	390
Female	Very satisfied	13.3500	4.91646	140
	Not very satisfied	11.9278	4.44335	180
	Total	12.5500	4.70216	320
Total	Very satisfied	14.4601	4.80176	313
	Not very satisfied	13.1259	4.86372	397
	Total	13.7141	4.87834	710

Figure 5.7. Adjusted Descriptive Statistics for Income by Gender and Job Satisfaction.

1. Respondent's Sex

Dependent Variable: rincom2

Respondent's Sex	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Male	14.737 ^a	.228	14.289	15.185
Female	12.666 ^a	.252	12.171	13.161

a. Covariates appearing in the model are evaluated at the following values: degree = 1.79.

2. Job Satisfaction

Dependent Variable: rincom2

Job Satisfaction	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Very satisfied	14.286 ^a	.255	13.784	14.787
Not very satisfied	13.118 ^a	.227	12.673	13.563

a. Covariates appearing in the model are evaluated at the following values: degree = 1.79.

Figure 5.8. Univariate ANCOVA Summary Table.

Tests of Between-Subjects Effects						
Dependent Variable: rincom2						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2602.188 ^a	4	650.547	32.138	.000	.154
Intercept	28459.544	1	28459.544	1405.949	.000	.666
sex	762.681	1	762.681	37.678	.000	.051
satjob2	220.068	1	220.068	10.872	.001	.015
degree	1505.805	1	1505.805	74.389	.000	.095
sex * satjob2	.057	1	.057	.003	.958	.000
Error	14270.771	705	20.242			
Total	150407.000	710				
Corrected Total	16872.959	709				

a. R Squared = .154 (Adjusted R Squared = .149)

Significance of the covariate, each factor, and factor interaction. The covariate and factors of gender and job satisfaction are significant.

Effect size for the covariate, each factor, and interaction. All are very small.

Writing Up Results

ANCOVA requires the fulfillment of several assumptions. If data screening leads to any participant elimination and/or variable transformation, these should be reported in the summary of results. The narrative of results should also include ANCOVA results (F ratios, degrees of freedom, p values, and effect sizes) for the main effect of each factor and covariate as well as the interaction of factors. Figures and tables may also be generated to support results (e.g., line graph of factor interactions, ANCOVA summary table, and a table comparing unadjusted and adjusted group means). The following results statement applies the results from Figures 5.3 through 5.8.

A 2×2 analysis of covariance was conducted to determine the effect of gender and job satisfaction on income when controlling for highest degree attained. Initial data screening led to the transformation of income by eliminating all values greater than or equal to 22. Interaction of factors was first analyzed by creating a line plot, which demonstrates no interaction (see Figure 5.3). ANCOVA results (see Table 1) indicate a significant main effect for gender [$F(1, 705) = 37.68, p < .001$, partial $\eta^2 = .051$] and a significant main effect for job satisfaction [$F(1, 705) = 10.87, p = .001$, partial $\eta^2 = .015$]. Interaction between gender and job satisfaction was not significant [$F(1, 705) = 0.00, p = .958$]. The covariate of degree significantly influenced the dependent variable of income [$F(1, 705) = 74.39, p < .001$, partial $\eta^2 = .095$]. Table 2 presents the adjusted and unadjusted group means for gender and job satisfaction, which indicate that males ($M = 14.74$) have significantly higher income than females ($M = 12.67$) and that satisfied ($M = 14.29$) employees earn more than unsatisfied ($M = 13.12$) employees.

Table 1
ANCOVA Summary Table

Source	SS	df	MS	F	p	η^2
Between treatments	2602.19	4	650.55	32.14	< .001	.154
Degree	1505.81	1	1505.81	74.39	< .001	.095
Gender	762.68	1	762.68	37.68	< .001	.051
Job satisfaction	220.07	1	220.07	10.87	.001	.015
Gender × Satisfaction	0.06	1	0.06	0.00	.958	.000
Error	14270.77	705	20.24			
Total	150407.00	710				

Table 2*Adjusted and Unadjusted Group Means for Income*

	Adjusted <i>M</i>	Unadjusted <i>M</i>
Males	14.74	14.67
Females	12.67	12.55
Very satisfied	14.29	14.46
Not very satisfied	13.12	13.13

SECTION 5.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example that applies the entire process of conducting ANCOVA: development of research questions and hypotheses, data-screening methods, test methods, interpretation of output, and presentation of results. The SPSS data set of *career-e.sav* is utilized.

Problem

Suppose you are interested in determining if hours worked per week is different by gender and for those who are satisfied or not satisfied with their jobs, while equalizing these groups on income. The following research questions and respective null hypotheses are generated for the main effects of each factor and the possible interaction between factors.

Research Questions

RQ1: Is there a difference in hours worked per week by gender among employees when controlling for income differences?



Null Hypotheses

H_01 : Hours worked per week will not differ by gender among employees when controlling for income differences.

RQ2: Is there a difference in hours worked per week between satisfied and unsatisfied employees when controlling for income differences?



H_02 : Hours worked per week will not differ between satisfied and unsatisfied employees when controlling for income differences.

RQ3: Is there a difference in hours worked per week based on gender and level of satisfaction of employees when controlling for income differences?



H_03 : Hours worked per week will not differ by gender or satisfaction level when controlling for income differences.

The IVs are categorical and include gender (*sex*) and job satisfaction (*satjob2*). The DV is hours worked per week (*hrs1*). The covariate is *rincom2*. Both the DV and covariate are quantitative variables. The IVs create a 2×2 factor design for hours worked per week, as seen in Figure 5.9.

Figure 5.9. Structure of 2×2 , Two-Factor Design for Hours Worked per Week.

Job Satisfaction	Gender	
	Female	Male
Very Satisfied With Job	Hours worked per week* for $n = 140$ females—very satisfied	Hours worked per week* for $n = 173$ males—very satisfied
Not Very Satisfied With Job	Hours worked per week* for $n = 180$ females—not very satisfied	Hours worked per week* for $n = 217$ males—not very satisfied

* Indicates that income has been partialled out of group means for hours worked per week.

Methods and SPSS “How To”

Before ANCOVA can be conducted, an examination of data for missing cases, outliers, and fulfillment of test assumptions must occur. The SPSS steps for data screening and fulfilling test assumptions are presented in Chapters 3 and 4.

The **Explore** procedure is conducted to identify missing values and outliers and to evaluate normality. Stem-and-Leaf plots indicate extreme values in *hrs1* for each group in job satisfaction (see Figure 5.10) and gender (see Figure 5.11). Because the cutoff for extreme values differs for each group, the more conservative criteria (≤ 16 and ≥ 80) will be used in eliminating outliers. Thus, *hrs1* has been transformed into *hrs2* so that cases less than or equal to 16 will be transformed to 17 and cases greater than or equal to 80 have been transformed to 79. This procedure will only *reduce* the number of outliers and does not eliminate all outliers.

Because the DV has been transformed, the **Explore** procedure will be conducted again to examine normality of *hrs2*. Although the tests of normality indicate nonnormal distributions for all groups (see Figure 5.12), ANCOVA is not highly sensitive to nonnormality as long as group sizes are large and fairly equivalent. Another assumption to test is the linear relationship between the covariate and the DV. This is appropriate only if the covariate is quantitative. Because income is quantitative, a simple scatterplot will be created to determine the linearity of the relationship between *hrs2* and *rincom2*. The scatterplot (see Figure 5.13) indicates a linear trend.

Figure 5.10. Stem-and-Leaf Plots for Hours Worked by Job Satisfaction.

Figure 5.11. Stem-and-Leaf Plots for Hours Worked by Gender.

Figure 5.12. Tests of Normality for Job Satisfaction and Gender.

Tests of Normality						
satjob2	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
hrs2	.174	327	.000	.886	327	.000
	.209	420	.000	.912	420	.000

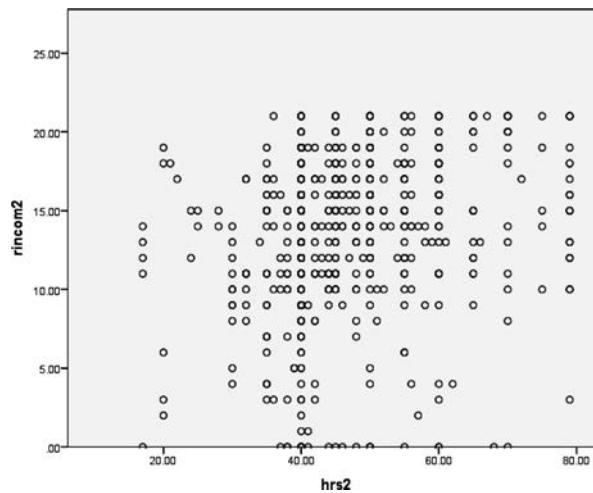
• Lilliefors Significance Correction

Tests of Normality						
sex	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
hrs2	Male	.149	408	.000	.935	408
	Female	.266	239	.000	.926	239

a. Lilliefors Significance Correction

Indicates nonnormal distributions for all groups.

Figure 5.13. Scatterplot of Income and Hours Worked per Week.



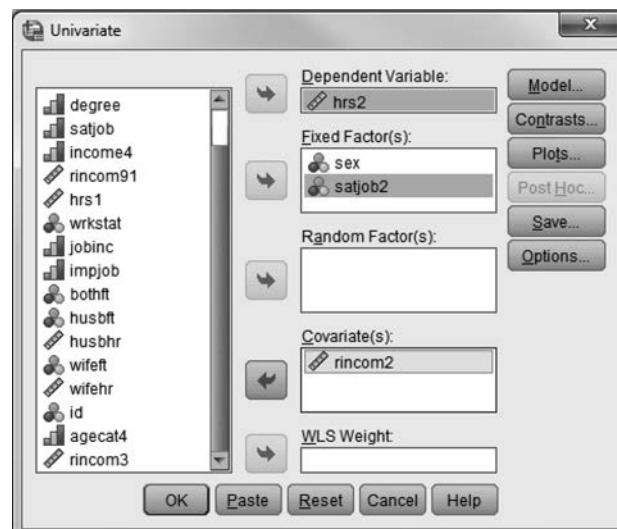
The next assumption to be tested is homogeneity of regression slopes to determine if significant interaction between the covariate and the factors is present. If significant interaction is found, ANCOVA results are not meaningful because the interaction implies that differences on the dependent variable among groups vary as a function of the covariate. In such a situation, ANCOVA should not be conducted. To evaluate homogeneity of regression slopes, a preliminary ANCOVA is conducted using **Univariate** analysis. This preliminary ANCOVA can also be used to evaluate homogeneity of variance. To open the **Univariate** dialog box as shown in Figure 5.14, select the following:

Analyze
General Linear Model
Univariate

Univariate dialog box (see Figure 5.14)

Once in this box, click the DV (*hrs2*) and move it to the **Dependent Variable** box. Click the IVs (*sex* and *satjob2*) and move each to the **Fixed Factor(s)** box. Click the covariate (*rincom2*) and move it to the **Covariate(s)** box. Then click **Model**.

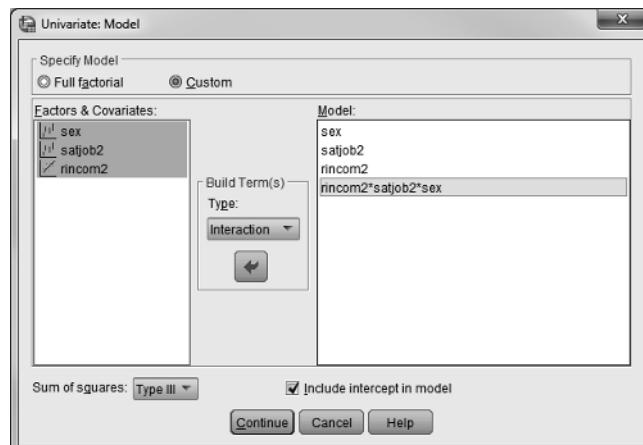
Figure 5.14. Univariate ANOVA Dialog Box.



Univariate:Model dialog box (see Figure 5.15)

Under **Specify Model**, click **Custom**. Move each IV and covariate(s) to the **Model** box. Then hold down the **Ctrl** key and highlight all IVs and covariate(s). Once highlighted, continue to hold down the **Ctrl** key and move this selection to the **Model** box. This should create the interaction between all IVs and covariate(s) (i.e., *rincom2*satjob2*sex*). Also check to make sure that **Interaction** is specified in the **Build Term(s)** box. Click **Continue** and then **Options**.

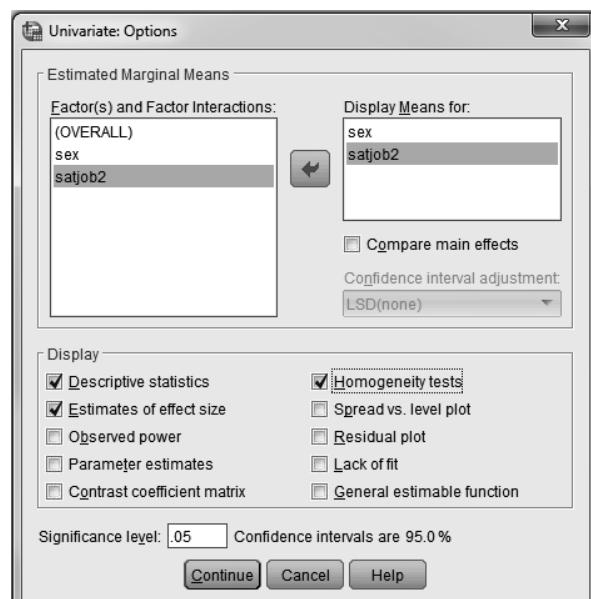
Figure 5.15. Univariate: Model Dialog Box for Homogeneity-of-Slopes Test.



Univariate:Options dialog box (see Figure 5.16)

Click each IV (*sex* and *satjob2*) in the **Factor(s) and Factor Interactions** box and move each to the **Display Means for** box. Select **Descriptive statistics**, **Estimates of effect size**, and **Homogeneity tests** in the **Display** box. Note that these options are described in the previous chapter on ANOVA. Click **Continue**. Back in the **Univariate** dialog box, click **OK**.

Figure 5.16. Univariate: Options Dialog Box.



This process will create the output to evaluate homogeneity of regression slopes. If no interaction between IVs and covariate(s) has been found, then the following steps for ANCOVA can be conducted. For our example, results (see Figure 5.17) indicate that the interaction of gender, job satisfaction, and income is not significant [$F(3, 703) = 1.72, p = .161$, partial $\eta^2 = .007$]. In addition, the Levene's test for equal variances (see Figure 5.18) indicates that variances between groups are fairly equivalent [$F(3, 706) = 2.27, p = .079$]. Because interaction between the factors and covariate was not found, factor interaction can now be analyzed by creating a line plot (see Figure 5.19).

Figure 5.17. ANCOVA Summary Table for Homogeneity of Regression Slopes.

Tests of Between-Subjects Effects						
Dependent Variable: hrs2						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	7025.503 ^a	6	1170.917	9.731	.000	.077
Intercept	121798.226	1	121798.226	1012.219	.000	.590
sex	65.529	1	65.529	.545	.461	.001
satjob2	12.804	1	12.804	.106	.744	.000
rincom2	2419.342	1	2419.342	20.106	.000	.028
sex * satjob2 * rincom2	621.925	3	207.308	1.723	.161	.007
Error	84590.527	703	120.328			
Total	1634449.000	710				
Corrected Total	91616.030	709				

a. R Squared = .077 (Adjusted R Squared = .069)

Indicates interaction between factors and covariate is NOT significant.

Figure 5.18. Levene's Test of Homogeneity of Variance.

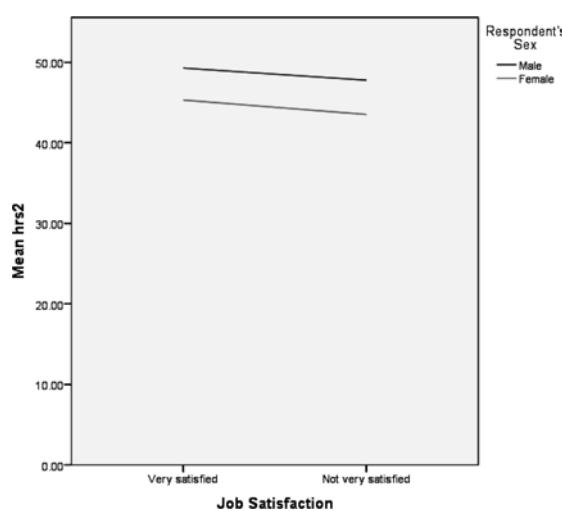
Levene's Test of Equality of Error Variances ^a			
Dependent Variable: hrs2			
F	df1	df2	Sig.
2.268	3	706	.079

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + sex + satjob2 + rincom2 + sex * satjob2 * rincom2

Indicates equal variances among groups.

Figure 5.19. Line Plot of Hours Worked by Gender and Job Satisfaction.



Factor interaction is not an issue, so we will continue conducting ANCOVA using **Univariate**. Note that the steps within the Univariate dialog box and Options dialog box are very similar to those used when conducting the homogeneity of regression slopes test earlier in this chapter.

Analyze

General Linear Model

Univariate

Univariate dialog box (see Figure 5.14)

Once in this box, click the DV (*hrs2*) and move to the **Dependent Variable** box. Click the IVs (*sex* and *satjob2*) and move each to the **Fixed Factor(s)** box. Click the covariate (*rincom2*) and move to the **Covariate(s)** box. (Note: No changes have been made in this step from previous analysis.) Then click **Model**.

Univariate:Model dialog box (see Figure 5.15)

This time, under **Specify Model**, click **Full Factorial** (*not Custom*, as shown in Figure 5.15). Within this box, you can also identify the method of calculating the Sum of Squares. Type III is the default and the most commonly used because it is best used when subgroups vary in sample size but have no missing cells. Back in the **Univariate** dialog box, click **Options**.

Univariate:Options dialog box (see Figure 5.16)

Our steps in this dialog box replicate the steps we took on page 117. Click each IV (*sex* and *satjob2*) in the **Factor(s) and Factor Interactions** box and move it to the **Display Means for** box. Select **Descriptive statistics**, **Estimates of effect size**, and **Homogeneity tests** in the **Display** box. Click **Continue**. Back in the **Univariate:Options** dialog box, click **OK**.

Output and Interpretation of Results

Figure 5.20 presents the means and standard deviations for each group prior to adjustment. Figure 5.21 displays the adjusted means. The summary table of ANCOVA results (see Figure 5.22) indicates that factor interaction is not significant. Therefore, the main effects of each factor and covariate can be more accurately interpreted. *F* ratios and *p* values indicate that after adjustment for income, hours worked per week is significantly different for males and females. However, the effect size ($\eta^2 = .021$) for gender is quite small in that gender accounts for only 2.1% of the variance in hours worked per week. No significant differences in hours worked per week were found in job satisfaction after adjustment for income. Results also indicate that the covariate (*rincom2*) significantly adjusted the DV.

Figure 5.20. Unadjusted Descriptive Statistics for Hours Worked by Gender and Job Satisfaction.

Descriptive Statistics				
Dependent Variable: hrs2				
sex	satjob2	Mean	Std. Deviation	N
Male	Very satisfied	49.4624	11.70197	173
	Not very satisfied	47.7465	11.37473	217
	Total	48.5077	11.53774	390
Female	Very satisfied	45.2071	11.53568	140
	Not very satisfied	43.6111	10.03813	180
	Total	44.3094	10.73102	320
Total	Very satisfied	47.5591	11.80112	313
	Not very satisfied	45.8715	10.97132	397
	Total	46.6155	11.36744	710

Figure 5.21. Adjusted Descriptive Statistics for Hours Worked by Gender and Job Satisfaction.

1. Respondent's Sex				
Dependent Variable: hrs2				
Respondent's Sex	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Male	48.171 ^a	.567	47.057	49.285
Female	44.864 ^a	.627	43.634	46.094

a. Covariates appearing in the model are evaluated at the following values: rincom2 = 13.7141.

2. Job Satisfaction				
Dependent Variable: hrs2				
Job Satisfaction	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Very satisfied	47.064 ^a	.627	45.832	48.296
Not very satisfied	45.971 ^a	.557	44.876	47.065

a. Covariates appearing in the model are evaluated at the following values: rincom2 = 13.7141.

Figure 5.22. ANCOVA Summary Table.

Tests of Between-Subjects Effects						
Dependent Variable: hrs2						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	6405.268 ^a	4	1601.317	13.249	.000	.070
Intercept	124454.483	1	124454.483	1029.687	.000	.594
rincom2	2823.075	1	2823.075	23.357	.000	.032
sex	1807.177	1	1807.177	14.952	.000	.021
satjob2	203.103	1	203.103	1.680	.195	.002
sex * satjob2	1.690	1	1.690	.014	.906	.000
Error	85210.761	705	120.866			
Total	1634449.000	710				
Corrected Total	91616.030	709				

The factor of gender is significant, but effect size is quite small.

Interaction is not significant.

a. R Squared = .070 (Adjusted R Squared = .065)

Presentation of Results

The following report summarizes the two-way ANCOVA results from the previous example.

A 2×2 analysis of covariance was conducted on hours worked per week. Independent variables consisted of gender and job satisfaction (very satisfied versus not very satisfied). The covariate was income. Data screening for outliers led to the transformation of hours worked per week. Participants with hours less than or equal to 16 were recoded 17, while participants with hours greater than or equal to 80 were recoded 79. After significant adjustment by the covariate of income, hours worked per week varied significantly with gender [$F(1, 705) = 14.95, p < .001$, partial $\eta^2 = .021$]. Table 3 presents a summary of the ANCOVA results. Comparison of adjusted group means, as displayed in Table 4, reveals that males work significantly more hours per week than females. No statistically significant difference was found for job satisfaction.

Table 3
ANCOVA Summary Table

Source	SS	df	MS	F	p	η^2
Between treatments	6405.27	4	1601.32	13.25	< .001	.070
Income	2823.08	1	2823.08	23.36	< .001	.032
Gender	1807.18	1	1807.18	14.95	< .001	.021
Job satisfaction	203.10	1	203.10	1.68	.195	.002
Gender \times Satisfaction	1.69	1	1.69	0.01	.906	.000
Error	85210.76	705	120.87			
Total	1634449.00	710				

Table 4
Adjusted and Unadjusted Group Means for Hours Worked per Week

	Adjusted M	Unadjusted M
Males	48.17	48.51
Females	44.86	44.31
Very satisfied	47.06	47.56
Not very satisfied	45.97	45.87

SUMMARY

Analysis of covariance allows the researcher to determine group differences while controlling for the effect of one or more variables. Essentially, the influence that the covariate(s) has on the DV is partitioned out before groups are compared. Thus, ANCOVA compares group means that have been adjusted to control for any influence the covariate(s) may have on the DV. Prior to conducting ANCOVA, data should be screened for missing data, outliers, normality of subgroups, homogeneity of variance, and homogeneity of regression slopes. The SPSS Univariate procedure generates an ANCOVA summary table that provides F ratios, p values, and effect sizes, which indicate the significance of factor and covariate main effects as well as factor interaction. If factor interaction is significant, then conclusions are limited about the main effects for each factor. Figure 5.23 provides a checklist for conducting analysis of covariance. SPSS steps for screening missing data, outliers, normality, and homogeneity of variance were presented in Chapter 3. Steps for creating a line plot were presented in Chapter 4.

KEYWORDS

- concomitant variable
- covariate
- homogeneity of regression
- multivariate analysis of covariance (MANCOVA)
- partialled out
- systematic bias

Figure 5.23. Checklist for Conducting ANCOVA.

I. Screen Data

- a. Missing Data?
 - Run Outliers and review Stem-and-Leaf plots and boxplots within **Explore**.
 - Eliminate or transform outliers if necessary.
- b. Outliers?
 - Run Outliers and review Stem-and-Leaf plots and boxplots within **Explore**.
 - Eliminate or transform outliers if necessary.
- c. Normality?
 - Run Normality Plots with Tests within **Explore**.
 - Review boxplots and histograms.
 - Transform data if necessary.
- d. Homogeneity of Variance?
 - Run Levene's Test of Equality of Variances within **Univariate**.
 - Transform data if necessary.
- e. Homogeneity of regression slopes?
 - Run Univariate ANOVA for Homogeneity of regression slopes.
 1. **Analyze**... **General Linear Model**... **Univariate**.
 - Move DV to **Dependent Variable** box.
 - Move IVs to **Fixed Factor** box.
 - Move covariate(s) to **Covariate** box.
 2. **Model**, **Custom**.
 - Move each IV and covariate to the **Model** box.
 - Hold down **Ctrl** key and highlight all IVs and covariate(s), ► while still holding down the **Ctrl** key in order to move interaction to **Model** box.
 3. **Continue**.
 4. **Continue**, then **OK**.
 - If factors and covariates interact, do not conduct ANCOVA.
- f. Factor Interaction.
 - Create lineplot of DV by IVs.

II. Conduct ANCOVA

- a. Run ANCOVA using Univariate ANOVA.
 1. **Analyze**... **General Linear Model**... **Univariate**.
 - Move DV to **Dependent Variable** box.
 - Move IVs to **Fixed Factor** box.
 - Move covariate(s) to **Covariate(s)** box.
 2. **Model**, **Full Factorial**, **Continue**.
 3. **Options**.
 - Move IVs to **Display Means for** box.
 - Check **Descriptive statistics**, **Estimates of effect size**, and **Homogeneity tests**.
 4. **Continue**, then **OK**.
- b. Interpret factor interaction.
- c. If no factor interaction, interpret main effects for each factor.

III. Summarize Results

- a. Describe any data elimination or transformation.
- b. Present line plot of factor interaction.
- c. Present table of adjusted and unadjusted group means.
- d. Narrate main effects for each factor and interaction (*F* ratio, *p* value, and effect size).
- e. Draw conclusions.

Exercises for Chapter 5

The exercises below utilize the data sets *career-a.sav* and *career-e.sav*, which can be downloaded from the following website:

www.routledge.com/9781138289734

You are interested in evaluating the effect of gender (*sex*) and age (*agecat4*) on respondents' income (*rincom91*) while controlling for hours worked per week (*hrs1*).

1. Develop the appropriate research questions and/or hypotheses for main effects and interaction.
2. Use *career-a.sav* to screen data for missing data and outliers. What steps, if any, are necessary for reducing missing data and outliers?

For all subsequent analyses, use *career-e.sav*, which eliminates outliers in *rincom2*.

3. Test the assumptions of normality, homogeneity of regression slopes, and homogeneity of variance.
 - a. What steps, if any, are necessary for increasing normality?
 - b. Do the covariate and factors interact? Can you conclude homogeneity of regression slopes?
 - c. Can you conclude homogeneity of variance?
4. Create a line plot of the factors. Do factors interact?
5. Conduct ANCOVA.
 - a. Is factor interaction significant? Explain.
 - b. Are main effects significant? Explain.
 - c. Does the covariate significantly influence the DV? Explain.
 - d. What can you conclude from the effect size for each main effect?
6. Write a results statement.

CHAPTER 6

MULTIVARIATE ANALYSIS OF VARIANCE AND COVARIANCE

STUDENT LEARNING OBJECTIVES

After studying Chapter 6, students will be able to:

1. Compare and contrast ANOVA/ANCOVA with MANOVA/MANCOVA.
2. Explain the advantages of using multivariate ANOVAs over using univariate ANOVAs.
3. Specify the order in which the results of a multivariate analysis are interpreted.
4. Explain the essential assumptions and limitations in using both MANOVAs and MANCOVAs.
5. Describe the logic behind the use of both MANOVAs and MANCOVAs.
6. Create an entire set of research questions appropriate for a multivariate analysis of variance.
7. Test data sets for multivariate analysis of variance group differences by following the appropriate SPSS guidelines provided.
8. Develop an entire set of research questions appropriate for a multivariate analysis of covariance.
9. Test data sets for multivariate analysis of covariance group differences by following the appropriate SPSS guidelines provided.

All of the statistical analysis techniques discussed to this point have involved only one dependent variable. In this chapter, for the first time, we consider ***multivariate statistics***—statistical procedures that involve more than one dependent variable. The focus of this chapter is on two of the most widely used multivariate procedures: the multivariate variations of analysis of variance and analysis of covariance. These versions of analysis of variance and covariance are designed to handle two or more dependent variables within the standard ANOVA/ANCOVA designs. We begin by discussing multivariate analysis of variance in detail, followed by a discussion of the application of covariance analysis in the multivariate setting.

I. MANOVA

Like ANOVA, multivariate analysis of variance (MANOVA) is designed to test the significance of group differences. The only substantial difference between the two procedures is that MANOVA can include several dependent variables, whereas ANOVA can handle only one. Often, these multiple dependent variables consist of different measures of essentially the same thing (Aron, Aron, & Coups, 2006), but this

need not always be the case. At a minimum, the DVs should have some degree of linearity and share a common conceptual meaning (Stevens, 2001). They should make sense as a group of variables. As you will soon see, the basic logic behind a MANOVA is essentially the same as in a univariate analysis of variance.

SECTION 6.1 PRACTICAL VIEW

Purpose

The clear advantage of a multivariate analysis of variance over a univariate analysis of variance is the inclusion of multiple dependent variables. Stevens (2001) provides two reasons why a researcher should be interested in using more than one DV when comparing treatments or groups based on differing characteristics:

1. Any worthwhile treatment or substantial characteristic will likely affect participants in more than one way, hence the need for additional criterion (dependent) measures.
2. The use of several criterion measures permits the researcher to obtain a more holistic picture, and therefore a more detailed description, of the phenomenon under investigation. This stems from the idea that it is extremely difficult to obtain a “good” measure of a trait (e.g., math achievement, self-esteem, etc.) from one variable. Multiple measures on variables representing a common characteristic are bound to be more representative of that characteristic.

ANOVA tests whether mean differences among k groups on a single DV are significant or are likely to have occurred by chance. However, when we move to the multivariate situation, the multiple DVs are treated in combination. In other words, MANOVA tests whether mean differences among k groups on a *combination of DVs* are likely to have occurred by chance. As part of the actual analysis, a new DV is created. This new DV is, in fact, a linear combination of the original measured DVs, combined in such a way as to maximize the group differences (i.e., separate the k groups as much as possible). The new DV is created by developing a linear equation where each measured DV has an associated weight and, when combined and summed, creates maximum separation of group means with respect to the new DV:

$$Y_{\text{new}} = a_1 Y_1 + a_2 Y_2 + a_3 Y_3 + \dots + a_n Y_n \quad (\text{Equation 6.1})$$

where Y_n is an original DV, a_n is its associated weight, and n is the total number of original measured DVs. An ANOVA is then conducted on this newly created variable.

Let us consider the following example: Assume we wanted to investigate the differences in worker productivity, as measured by income level (DV_1) and hours worked (DV_2), for individuals of different age categories (IV). Our analysis would involve the creation of a new DV, which would be a linear combination (DV_{new}) of our participants’ income levels and numbers of hours worked that maximizes the separation of our age-category groups. Our new DV would then be subjected to a univariate ANOVA by comparing variances on DV_{new} for the various groups as defined by age category.

One could also have a *factorial MANOVA*—a design that would involve multiple IVs as well as multiple DVs. In this situation, a different linear combination of DVs is formed for each main effect and each interaction (Tabachnick & Fidell, 2007). For instance, we might consider investigating the effects of gender (IV_1) and job satisfaction (IV_2) on employee income (DV_1) and years of education (DV_2). Our analysis would actually provide three new DVs—the first linear combination would maximize the separation between males and females (IV_1), the second linear combination would maximize the separation among job-satisfaction categories (IV_2), and the third would maximize the separation among the various cells of the interaction between gender and job satisfaction.

At this point, one might be inclined to question why a researcher would want to engage in a multivariate analysis of variance, as opposed to simply doing a couple of comparatively simple analyses of variance. MANOVA has several advantages over its simpler univariate counterpart (Tabachnick & Fidell, 2007). First, as previously mentioned, by measuring several DVs instead of only one, the chances of discovering what actually changes as a result of the differing treatments or characteristics (and any interactions) improve immensely. If we want to know what measures of work productivity are affected by gender and age, we improve our chances of uncovering these effects by including hours worked as well as income level. There are also several statistical reasons for preferring a multivariate analysis to a univariate one (Stevens, 2001).

A second advantage is that, under certain conditions, MANOVA may reveal differences not shown in separate ANOVAs (Stevens, 2001; Tabachnick & Fidell, 2007). Assume we have a one-way design with two levels on the IV and two DVs. If separate ANOVAs are conducted on two DVs, the distributions for each of the two groups (and for each DV) might overlap sufficiently, such that a mean difference probably would not be found. However, when the two DVs are considered in combination with each other, the two groups may differ substantially and could result in a statistically significant difference between groups. Therefore, a MANOVA may sometimes be more powerful than separate ANOVAs.

Third, the use of several univariate analyses leads to a greatly inflated overall Type I error rate. Consider a simple design with one IV (with two levels) and five DVs. If we assume that we wanted to test for group differences on each of the DVs (at $\alpha = .05$ level of significance), we would have to conduct five univariate tests. Recall that at an α level of .05, we are assuming a 95% chance of no Type I errors. Because of the assumption of independence, we can multiply the probabilities. The effect of these error rates is compounded over all of the tests such that the overall probability of *not* making a Type I error becomes

$$(.95)(.95)(.95)(.95)(.95) = .77$$

In other words, the probability of at least one false rejection (i.e., Type I error) becomes

$$1 - .77 = .23$$

which is an unacceptably high rate of possible statistical decision error (Stevens, 2001). Therefore, using this approach of fragmented univariate tests results in an overall error rate that is entirely too risky. The use of MANOVA includes a condition that maintains the overall error rate at the .05 level, or whatever α level is preselected (Harris, 1998).

Finally, the use of several univariate tests ignores some very important information. Recall that if several DVs are included in an analysis, they should be correlated to some degree. A multivariate analysis incorporates the intercorrelations among DVs into the analysis (this is essentially the basis for the linear combination of DVs).

Keep in mind, however, that there are disadvantages to the use of MANOVA. The main disadvantage is the fact that MANOVA is substantially more complicated than ANOVA (Tabachnick & Fidell, 2007). In the use of MANOVA, there are several important assumptions that need to be met. Furthermore, the results are sometimes ambiguous with respect to the effects of IVs on individual DVs. Finally, situations in which MANOVA is more powerful than ANOVA, as discussed earlier in this chapter, are quite limited. Often, the multivariate procedure is much *less* powerful than ANOVA (Tabachnick & Fidell, 2007). It has been recommended that one carefully consider the need for additional DVs in an analysis in light of the added complexity (Tabachnick & Fidell, 2007).

In the univariate case, the null hypothesis stated that the population means were equal:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

The calculations for MANOVA, however, are based on matrix algebra (as opposed to scalar algebra). The null hypothesis in MANOVA states that the population *mean vectors* are equal (note that **bold** font indicates that the variables are vector, not scalar):

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \dots = \boldsymbol{\mu}_k$$

For the univariate analysis of variance, recall that the F statistic is used to test the tenability of the null hypothesis. This test statistic is calculated by dividing the variance between the groups by the variance within the groups. There are several available test statistics for multivariate analysis of variance, but the most commonly used criterion is **Wilks' Lambda (Λ)**. (Other test statistics for MANOVA include **Pillai's Trace**, **Hotelling's Trace**, and **Roy's Largest Root**.) Without going into great detail, Wilks' Lambda is obtained by calculating $|\mathbf{W}|$ (a measure of the within-groups sum-of-squares and cross-products matrix—a multivariate generalization of the univariate sum-of-squares within [SS_W]) and dividing it by $|\mathbf{T}|$ (a measure of the total sum-of-squares and cross-products matrix—also a multivariate generalization, this time of the total sum-of-squares [SS_T]). The obtained value of Wilks' Λ ranges from zero to one. It is important to note that Wilks' Λ is an **inverse criterion**: that is, the smaller the value of Λ , the more evidence for treatment effects or group differences (Stevens, 2001). This is the opposite relationship that F has to the amount of treatment effect.

In conducting a MANOVA, one first tests the overall multivariate hypothesis (i.e., that all groups are equal on the combination of DVs). This is accomplished by evaluating the significance of the test associated with Λ . If the null hypothesis is retained, it is common practice to stop the interpretation of the analysis at this point and conclude that the treatments or conditions have no effect on the DVs. However, if the overall multivariate test is significant, the researcher would then likely wish to discover which of the DVs is being affected by the IV(s). To accomplish this, one conducts a series of univariate analyses of variance on the individual DVs. This will undoubtedly result in multiple tests of significance, which will result in an inflated Type I error rate.

To counteract the potential of an inflated error rate due to multiple ANOVAs, an adjustment must be made to the alpha level used for the tests. This *Bonferroni-type* adjustment involves setting a more stringent alpha level for the test of each DV so that the alpha for the *set* of DVs does not exceed some critical value (Tabachnick & Fidell, 2007). That critical value for testing each DV is usually the overall alpha level for the analysis (e.g., $\alpha = .05$) divided by the number of DVs. For instance, if one had three DVs and wanted an overall α equal to $.05$, each univariate test could be conducted at $\alpha = .016$ because $.05/3 = .0167$. Note that rounding down is necessary to create an overall alpha less than $.05$. The following equation may be used to check adjustment decisions:

$$\alpha = 1 - [(1 - \alpha_1)(1 - \alpha_2)\dots(1 - \alpha_p)]$$

where the overall error rate (α) is based on the error rate for testing the first DV (α_1), the second DV (α_2), and all others to the p^{th} DV (α_p). All alphas can be set at the same level, or more important DVs can be given more liberal alphas (Tabachnick & Fidell, 2007).

Finally, for any univariate test of a DV that results in significance, one then conducts univariate post hoc tests (as discussed in Chapter 4) in order to identify where specific differences lie (i.e., which levels of the IV are different from which other levels). To summarize the analysis procedure for MANOVA, follow these steps:

1. Examine the overall multivariate test of significance—if the results are significant, proceed to the next step. If not, stop.

2. Examine the univariate tests of individual DVs—if any are significant, proceed to the next step. If not, stop.
3. Examine the post hoc tests for individual DVs.

Sample Research Questions

In our first sample study in this chapter, we are concerned with investigating differences in worker productivity, as measured by income level (DV_1) and hours worked (DV_2), for individuals of different age categories (IV)—a one-way MANOVA design. Therefore, this study should address the following research questions:

1. Are there significant mean differences in worker productivity (as measured by the combination of income and hours worked) for individuals of different ages?
2. Are there significant mean differences in income levels for individuals of different ages?
 - (2a) If so, which age categories differ?
3. Are there significant mean differences in hours worked for individuals of different ages?
 - (3a) If so, which age categories differ?

Our second sample study will demonstrate a two-way MANOVA where we investigate the gender (IV₁) and job satisfaction (IV₂) differences in income level (DV₁) and years of education (DV₂). Note that the following questions address the MANOVA analysis, univariate ANOVA analyses, and post hoc analyses:

1. a. Are there significant mean differences in the combined DV of income and years of education for males and females?
 - b. Are there significant mean differences in the combined DV of income and years of education for different levels of job satisfaction? If so, which job-satisfaction categories differ?
 - c. Is there a significant interaction between gender and job satisfaction on the combined DV of income and years of education?
2. a. Are there significant mean differences on income between males and females?
 - b. Are there significant mean differences on income between different levels of job satisfaction? If so, which job-satisfaction categories differ?
 - c. Is there a significant interaction between gender and job satisfaction on income?
3. a. Are there significant mean differences in years of education between males and females?
 - b. Are there significant mean differences in years of education among different levels of job satisfaction? If so, which job-satisfaction categories differ?
 - c. Is there a significant interaction between gender and job satisfaction on years of education?

SECTION 6.2 ASSUMPTIONS AND LIMITATIONS

Because we are introducing our first truly multivariate technique in this chapter, we have a new set of statistical assumptions to discuss. They are new in that they apply to the multivariate situation; however, they are quite analogous to the assumptions for univariate analysis of variance, which we have already examined (see Chapter 4). For multivariate analysis of variance, these assumptions are as follows:

1. The observations within each sample must be randomly sampled and must be independent of each other.

2. The observations on all dependent variables must follow a multivariate normal distribution in each group.
3. The population covariance matrices for the dependent variables in each group must be equal (this assumption is often referred to as the *homogeneity of covariance matrices assumption* or the *assumption of homoscedasticity*).
4. The relationships among all pairs of DVs for each cell in the data matrix must be linear.

Remember that the assumption of independence is primarily a design issue, not a statistical one. Provided the researcher has randomly sampled and assigned participants to treatments, it is usually safe to believe that this assumption has not been violated. We will focus our attention on the assumptions of multivariate normality, homogeneity of covariance matrices, and linearity.

Methods of Testing Assumptions

As discussed in Chapter 3, multivariate normality implies that the sampling distribution of the means of each DV in each cell and all linear combinations of DVs are normally distributed (Tabachnick & Fidell, 2007). Multivariate normality is a difficult entity to describe and even more difficult to assess. Initial screening for multivariate normality consists of assessments for univariate normality (see Chapter 3) for all variables, as well as examinations of all bivariate scatterplots (see Chapter 3) to check that they are approximately elliptical (Stevens, 2001). Specific graphical tests for multivariate normality do exist, but they are not available in standard statistical software packages (Stevens, 2001) and will not be discussed here.

It is probably most important to remember that both ANOVA and MANOVA are robust to moderate violations of normality, provided the violation is created by skewness and not by outliers (Tabachnick & Fidell, 2007). With equal or unequal sample sizes and only a few DVs, a sample size of about 20 in the smallest cell should be sufficient to ensure robustness to violations of univariate and multivariate normality. If it is determined that the data have substantially deviated from normal, transformations of the original data should be considered.

Recall that the assumption of equal covariance matrices (i.e., homoscedasticity) is a necessary condition for multivariate normality (Tabachnick & Fidell, 2007). The failure of the relationship between two variables to be homoscedastic is caused either by the nonnormality of one of the variables or by the fact that one of the variables may have some sort of relationship to the transformation of the other variable. Therefore, checking for univariate and multivariate normality is a good starting point for assessing possible violations of homoscedasticity. Specifically, possible violations of this assumption may be assessed by interpreting the results of Box's test. Note that a violation of the assumption of homoscedasticity, similar to a violation of homogeneity, will not prove fatal to an analysis (Tabachnick & Fidell, 2007; Kennedy & Bush, 1985). However, the results will be greatly improved if the heteroscedasticity is identified and corrected (Tabachnick & Fidell, 2007) by means of data transformations. On the other hand, if homogeneity of variance-covariance is violated, a more robust multivariate test statistic, Pillai's Trace, can be selected when interpreting the multivariate results.

Linearity is best assessed through inspection of bivariate scatterplots. If both variables in the pair are normally distributed and linearly related, the shape of the scatterplot will be elliptical. If one of the variables is not normally distributed, the relationship will not be linear and the scatterplot between the two variables will not appear oval-shaped. As mentioned in Chapter 3, assessing linearity by means of bivariate scatterplots is an extremely subjective procedure. In situations where nonlinearity between variables is apparent, the data can once again be transformed in order to enhance the linear relationship.

SECTION 6.3 PROCESS AND LOGIC

The Logic Behind MANOVA

As previously mentioned, the calculations for MANOVA somewhat parallel those for a univariate ANOVA, although they exist in multivariate form (i.e., they rely on matrix algebra). Because several variables are involved in this analysis, calculations are based on a *matrix* of values, as opposed to the mathematical manipulations of a single value. Specifically, the matrix used in the calculations is the sum-of-squares and cross-products (SSCP) matrix, which, you will recall, is the precursor to the variance-covariance matrix (see Chapter 1).

In univariate ANOVA, recollect that the calculations are based on a partitioning of the total sum-of-squares into the sum-of-squares between the groups and the sum-of-squares within the groups:

$$SS_{\text{Total}} = SS_{\text{Between}} + SS_{\text{Within}}$$

In MANOVA, the calculations are based on the corresponding matrix analogue (Stevens, 2001), in which the total sum-of-squares and cross-products matrix (**T**) is partitioned into a between sum-of-squares and cross-products matrix (**B**) and a within sum-of-squares and cross-products matrix (**W**):

$$SSCP_{\text{Total}} = SSCP_{\text{Between}} + SSCP_{\text{Within}}$$

or

$$\mathbf{T} = \mathbf{B} + \mathbf{W} \quad (\text{Equation 6.2})$$

Wilks' Lambda (Λ) is then calculated by using the **determinants**—a sort of generalized variance for an entire set of variables—of the SSCP matrices (Stevens, 2001). The resulting formula for Λ becomes

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} \quad (\text{Equation 6.3})$$

If there is no treatment effect or group difference, then **B** = 0 and Λ = 1, indicating no differences between groups on the linear combination of DVs. Whereas if **B** were very large (i.e., substantially greater than 0), then Λ would approach 0, indicating significant group differences on the combination of DVs.

As in all of our previously discussed ANOVA designs, we can again obtain a measure of strength of association, or effect size. Recall that eta squared (η^2) is a measure of the magnitude of the relationship between the independent and dependent variables and is interpreted as the proportion of variance in the dependent variable explained by the independent variable(s) in the sample. For MANOVA, eta squared is obtained in the following manner:

$$\eta^2 = 1 - \Lambda$$

In the multivariate situation, η^2 is interpreted as the variance accounted for in the best linear combination of DVs by the IV(s) and/or interactions of IVs.

Interpretation of Results

The MANOVA procedure generates several test statistics to evaluate group differences on the combined DV: Pillai's Trace, Wilks' Lambda, Hotelling's Trace, and Roy's Largest Root. When the IV has only two categories, the F test for Pillai's Trace, Wilks' Lambda, and Hotelling's Trace will be identical. When the IV has three or more categories, the F test for these three statistics will differ slightly but will maintain consistent significance or nonsignificance. Although these test statistics may vary only slightly, Wilks' Lambda is the most commonly reported MANOVA statistic. Pillai's Trace is used when homogeneity of variance-covariance is in question. If two or more IVs are included in the analysis, factor interaction must be evaluated before main effects.

In addition to the multivariate tests, the output for MANOVA typically includes the test for homogeneity of variance-covariance (Box's test), univariate ANOVAs, and univariate post hoc tests. Because homogeneity of variance-covariance is a test assumption for MANOVA and has implications for how to interpret the multivariate tests, the results of Box's test should be evaluated first. Highly sensitive to the violation of normality, Box's test should be interpreted with caution. Typically, if Box's test is significant at $p < .001$ and group sample sizes are extremely unequal, then robustness cannot be assumed, due to unequal variances among groups (Tabachnick & Fidell, 2007). In such a situation, a more robust MANOVA test statistic, Pillai's Trace, is utilized when interpreting the MANOVA results. If equal variances are assumed, Wilks' Lambda is commonly used as the MANOVA test statistic. Once the test statistic has been determined, factor interaction (F ratio and p value) should be assessed if two or more IVs are included in the analysis. Like two-way ANOVA, if interaction is significant, then inferences drawn from the main effects are limited. If factor interaction is *not* significant, then one should proceed to examine the F ratios and p values for each main effect. When multivariate significance is found, the univariate ANOVA results can indicate the degree to which groups differ for each DV. A more conservative alpha level should be applied using the Bonferroni adjustment. Post hoc results can then indicate which groups are significantly different for the DV if univariate significance is found for that particular DV.

In summary, the first step in interpreting the MANOVA results is to evaluate the Box's test. If homogeneity of variance-covariance is assumed, utilize the Wilks' Lambda statistic when interpreting the multivariate tests. If the assumption of equal variances is violated, use Pillai's Trace. Once the multivariate test statistic has been identified, examine the significance (F ratios and p values) of factor interaction. This is necessary only if two or more IVs are included. Next, evaluate the F ratios and p values for each factor's main effect. If multivariate significance is found, interpret the univariate ANOVA results to determine significant group differences for each DV. If univariate significance is revealed, examine the post hoc results to identify which groups are significantly different for each DV.

Recalling the example introduced previously¹ that investigates age category (*agecat4*) differences in respondents' income (*rincom91*) and hours worked per week (*hrs1*), data were screened for missing data and outliers and then examined for fulfillment of test assumptions. Data screening led to the transformation of *rincom91* to *rincom2* in order to eliminate all cases with income equal to zero and cases equal to or exceeding 22. *Hrs1* was also transformed to *hrs2* as a means of reducing the number of outliers. Those less than or equal to 16 were recoded 17, and those greater than or equal to 80 were recoded 79. Although normality of these transformed variables is still questionable, group sample sizes are quite large and fairly equivalent. Therefore, normality will be assumed. Linearity of the two DVs was then tested by creating a scatterplot and calculating the Pearson correlation coefficient. Results indicate a linear relationship. Although the correlation coefficient is statistically significant, it is still quite low ($r = .253$, $p < .001$).

¹ The *careerf* data set contains all raw and transformed variables used to create Figures 6.1–6.4.

Unadjusted means for income (*rincom2*) and hours worked (*hrs2*) by age category (*agecat4*) were tested utilizing the **Multivariate** procedure (see Figure 6.1). The last assumption, homogeneity of variance-covariance, will be tested within MANOVA. Thus, MANOVA was conducted utilizing the **Multivariate** procedure. The Box's test (see Figure 6.2) reveals that equal variances can be assumed, $F(9, 2886561) = .766, p = .648$. Therefore, Wilks' Lambda will be used as the test statistic. Figure 6.3 presents the MANOVA results. The Wilks' Lambda criteria indicates significant group differences in age category with respect to income and hours worked per week, Wilks' $\Lambda = .909, F(6, 1360) = 11.04, p < .001$, multivariate $\eta^2 = .046$. Univariate ANOVA results (see Figure 6.4) were interpreted using a more conservative alpha level ($\alpha = .025$). Results reveal that age category significantly differs only for income [$F(3, 681) = 21.00, p < .001$, partial $\eta^2 = .085$] and not for hours worked per week [$F(3, 681) = .167, p = .919$, partial $\eta^2 = .001$]. Examination of post hoc results reveals that income of those 18–29 years of age significantly differs from all other age categories (see Figure 6.5). In addition, income for individuals 30–39 years differs from those 40–49 years.

Figure 6.1. Unadjusted Means for Income and Hours Worked by Age Category.

Descriptive Statistics				
	agecat4	Mean	Std. Deviation	N
rincom2	18-29	11.8672	4.14381	128
	30-39	14.0315	3.88102	222
	40-49	15.3247	3.86598	194
	50+	14.9574	4.41729	141
	Total	14.1839	4.21547	685
hrs2	18-29	46.3203	10.32002	128
	30-39	47.0315	11.41817	222
	40-49	46.4897	11.75450	194
	50+	46.3262	11.51489	141
	Total	46.6000	11.31890	685

Figure 6.2. Box's Test for Homogeneity of Variance-Covariance.

**Box's Test of Equality
of Covariance Matrices^a**

Box's M	6.936
F	.766
df1	9
df2	2886560.794
Sig.	.648

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design:
Intercept +
agecat4

Box's test is not significant. Use Wilks' Lambda criteria.

Figure 6.3. Multivariate Tests for Income and Hours Worked by Age Category.

Multivariate Tests^a

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.957	7507.272 ^b	2.000	680.000	.000	.957
	Wilks' Lambda	.043	7507.272 ^b	2.000	680.000	.000	.957
	Hotelling's Trace	22.080	7507.272 ^b	2.000	680.000	.000	.957
	Roy's Largest Root	22.080	7507.272 ^b	2.000	680.000	.000	.957
agecat4	Pillai's Trace	.091	10.791	6.000	1362.000	.000	.045
	Wilks' Lambda	.909	11.035 ^b	6.000	1360.000	.000	.046
	Hotelling's Trace	.100	11.279	6.000	1358.000	.000	.047
	Roy's Largest Root	.099	22.457 ^c	3.000	681.000	.000	.090

Indicates that age category significantly differs for the combined DV.

a. Design: Intercept + agecat4

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Figure 6.4. Univariate ANOVA Summary Table.

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	rincom2	1029.016 ^a	3	343.005	20.995	.000	.085
	hrs2	64.281 ^b	3	21.427	.167	.919	.001
Intercept	rincom2	128493.515	1	128493.515	7864.965	.000	.920
	hrs2	1410953.564	1	1410953.564	10972.708	.000	.942
agecat4	rincom2	1029.016	3	343.005	20.995	.000	.085
	hrs2	64.281	3	21.427	.167	.919	.001
Error	rincom2	11125.807	681	16.337			
	hrs2	87568.119	681	128.588			
Total	rincom2	149966.000	685				
	hrs2	1575151.000	685				
Corrected Total	rincom2	12154.823	684				
	hrs2	87632.400	684				

Indicates that age category significantly affects income but NOT hours worked.

a. R Squared = .085 (Adjusted R Squared = .081)

b. R Squared = .001 (Adjusted R Squared = -.004)

Figure 6.5. Post Hoc Results for Income and Hours Worked by Age Category.

Multiple Comparisons							
Dependent Variable	() 4 categories of age	(J) 4 categories of age	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
rincom2	18-29	30-39	-2.1643*	.44859	.000	-3.3513	-.9774
		40-49	-3.4576*	.46027	.000	-4.6754	-2.2397
		50+	-3.0903*	.49346	.000	-4.3960	-1.7846
	30-39	18-29	2.1643*	.44859	.000	.9774	3.3513
		40-49	-1.2932*	.39725	.007	-2.3443	-.2421
		50+	.9259	.43527	.203	-2.0776	.2258
	40-49	18-29	3.4576*	.46027	.000	2.2397	4.6754
		30-39	1.2932*	.39725	.007	.2421	2.3443
		50+	.3673	.44731	1.000	-.8163	1.5509
	50+	18-29	3.0903*	.49346	.000	1.7846	4.3960
		30-39	.9259	.43527	.203	-.2258	2.0776
		40-49	-.3673	.44731	1.000	-1.5509	.8163
hrs2	18-29	30-39	-.7112	1.25850	1.000	-4.0412	2.6187
		40-49	-.1694	1.29128	1.000	-3.5861	3.2473
		50+	-.0059	1.38440	1.000	-3.6690	3.6572
	30-39	18-29	.7112	1.25850	1.000	-2.6187	4.0412
		40-49	.5418	1.11447	1.000	-2.4070	3.4907
		50+	.7053	1.22114	1.000	-2.5258	3.9364
	40-49	18-29	.1694	1.29128	1.000	-3.2473	3.5861
		30-39	-.5418	1.11447	1.000	-3.4907	2.4070
		50+	.1634	1.25491	1.000	-3.1570	3.4839
	50+	18-29	.0059	1.38440	1.000	-3.6572	3.6690
		30-39	-.7053	1.22114	1.000	-3.9364	2.5258
		40-49	-.1634	1.25491	1.000	-3.4839	3.1570

Based on observed means.

The error term is Mean Square(Error) = 128.588.

*. The mean difference is significant at the .05 level.

Writing Up Results

Once again, any data transformations utilized to increase the likelihood of fulfilling test assumptions should be reported in the summary of results. The summary should then report the results from the multivariate tests by first indicating the test statistic utilized and its respective value and then reporting the F ratio, degrees of freedom, p value, and effect size for each IV main effect. If follow-up analysis was conducted using univariate ANOVA, these results should be summarized next. Report the F ratio, degrees of freedom, p value, and effect size for the main effect on each DV. Utilize the post hoc results to indicate which groups were significantly different within each DV. Finally, you may want to create a table of means and standard deviations for each DV by the IV categories. In summary, the MANOVA results narrative should address the following:

1. Participant elimination and/or variable transformation
2. MANOVA results (test statistic, F ratio, degrees of freedom, p value, and effect size)
 - a. Main effects for each IV on the combined DV
 - b. Main effect for the interaction between IVs

3. Univariate ANOVA results (F ratio, degrees of freedom, p value, and effect size)
 - a. Main effect for each IV and DV
 - b. Comparison of means to indicate which groups differ on each DV
4. Post hoc results (mean differences and levels of significance)

Utilizing our previous example, the following statement applies the results from Figures 6.1 through 6.5.

A one-way multivariate analysis of variance (MANOVA) was conducted to determine age category differences in income and hours worked per week. Prior to the test, variables were transformed to eliminate outliers. Cases with income equal to zero or equal to or exceeding 22 were eliminated. Hours worked per week was also transformed; those less than or equal to 16 were recoded 17 and those greater than or equal to 80 were recoded 79. MANOVA results revealed significant differences among the age categories on the dependent variables [Wilks' $\Lambda = .909$, $F(6, 1360) = 11.04$, $p < .001$, multivariate $\eta^2 = .046$]. Analysis of variance (ANOVA) was conducted on each dependent variable as a follow-up test to MANOVA. Age category differences were significant for income [$F(3, 681) = 21.00$, $p < .001$, partial $\eta^2 = .085$]. Differences in hours worked per week were not significant [$F(3, 681) = .167$, $p = .919$, partial $\eta^2 = .001$]. The Bonferroni post hoc analysis revealed that income of those 18–29 years of age significantly differs from all other age categories. In addition, income for individuals 30–39 years of age differs from those 40–49 years of age. Table 1 presents means and standard deviations for income and hours worked per week by age category.

Table 1

Means and Standard Deviations for Income and Hours Worked per Week by Age Category

Age	Income		Hours Worked per Week	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
18–29 years	11.87	4.14	46.32	10.32
30–39 years	14.03	3.88	47.03	11.42
40–49 years	15.32	3.87	46.49	11.75
50+ years	14.96	4.42	46.33	11.51

SECTION 6.4 MANOVA SAMPLE STUDY AND ANALYSIS

This section provides a complete example that applies the entire process of conducting MANOVA: development of research questions and hypotheses, data-screening methods, test methods, interpretation of output, and presentation of results. The SPSS data set *career-f.sav* is utilized. Our previous example demonstrates a one-way MANOVA, while this example will present a two-way MANOVA.

Problem

This time, we are interested in determining the degree to which gender and job satisfaction affect income and years of education among employees. Because two IVs are tested in this analysis, questions must also take into account the possible interaction between factors. The following research questions and respective null hypotheses address the multivariate main effects for each IV (gender and job satisfaction on income and hours worked) and the possible interaction between factors. Both IVs are categorical and include gender (*sex*) and job satisfaction (*satjob*). One should note that *satjob* represents four levels: very satisfied, moderately satisfied, a little dissatisfied, and very dissatisfied. The DVs are respondents' income

(*rincom2*) and years of education (*educ*). Both are quantitative. The variable, *rincom2*, is a transformation of *rincom91* from the previous example.

Research Questions

RQ1: Do income and years of education differ by gender among employees?

RQ2: Do income and years of education differ by job satisfaction among employees?

RQ3: Do gender and job satisfaction interact in the effect on income and years of education?

Null Hypotheses

→ H_01 : Income and years of education will not differ by gender among employees.

→ H_02 : Income and years of education will not differ by job satisfaction among employees.

→ H_03 : Gender and job satisfaction will not interact in the effect on income and years of education.

Methods and SPSS “How To”

Data should first be examined for missing data, outliers, and fulfillment of test assumptions. The **Explore** procedure was conducted to identify outliers and evaluate normality. Boxplots (see Figure 6.6) indicate extreme values in *educ*. Consequently, *educ* was transformed to *educ2* in order to eliminate participants with 6 years of education or less. **Explore** was conducted again to evaluate normality. Tests indicate significant nonnormality for both *rincom2* and *educ2* in many categories (see Figure 6.7). Because MANOVA is fairly robust to nonnormality, no further transformations will be performed. However, the significant nonnormality coupled with the unequal group sample sizes, as in this example, may lead to violation of homogeneity of variance-covariance. The next step in examining test assumptions was to determine linearity between the DVs. A scatterplot was created. Pearson correlation coefficients were calculated (see Figure 6.8) by selecting the following:

Analyze

Correlate

Bivariate...

Within the **Bivariate...** dialogue box (not pictured), move pertinent variables (*educ2* and *rincom2*) to the Variables box. Then click **OK**. Scatterplot and correlation output indicate a linear relationship. Although the correlation coefficient is significant, it is still fairly weak ($r = .337, p < .001$).

Figure 6.6. Boxplots for Years of Education by (a) Gender and (b) Job Satisfaction.

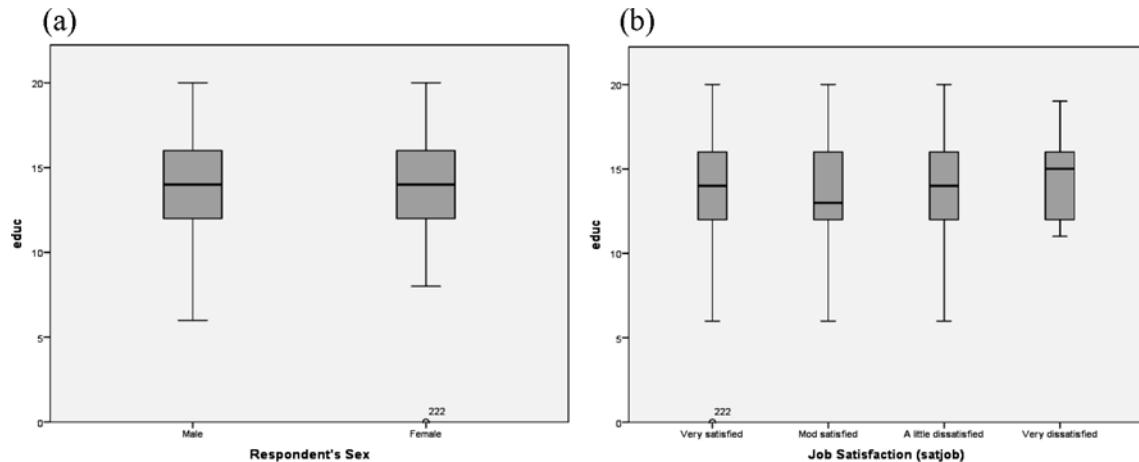


Figure 6.7. Tests of Normality of Income and Years of Education by Gender and Job Satisfaction.

Tests of Normality

sex	Kolmogorov-Smirnov ^a			Shapiro-Wilk			
	Statistic	df	Sig.	Statistic	df	Sig.	
rincom2	Male	.100	374	.000	.950	374	.000
	Female	.110	307	.000	.970	307	.000
educ2	Male	.162	374	.000	.954	374	.000
	Female	.177	307	.000	.922	307	.000

a. Lilliefors Significance Correction

Tests of Normality

satjob	Kolmogorov-Smirnov ^a			Shapiro-Wilk			
	Statistic	df	Sig.	Statistic	df	Sig.	
rincom2	Very satisfied	.123	299	.000	.956	299	.000
	Mod satisfied	.080	291	.000	.972	291	.000
	A little dissatisfied	.086	67	.200*	.964	67	.047
	Very dissatisfied	.203	24	.012	.954	24	.333
educ2	Very satisfied	.148	299	.000	.945	299	.000
	Mod satisfied	.197	291	.000	.935	291	.000
	A little dissatisfied	.167	67	.000	.936	67	.002
	Very dissatisfied	.193	24	.021	.925	24	.074

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure 6.8. Correlation Coefficients for Income and Years of Education.

Correlations

	educ2	rincom2	
educ2	Pearson Correlation	1	.337**
	Sig. (2-tailed)		.000
	N	742	681
rincom2	Pearson Correlation	.337**	1
	Sig. (2-tailed)	.000	
	N	681	686

**. Correlation is significant at the 0.01 level (2-tailed).

Correlation coefficient indicates fairly weak relationship.

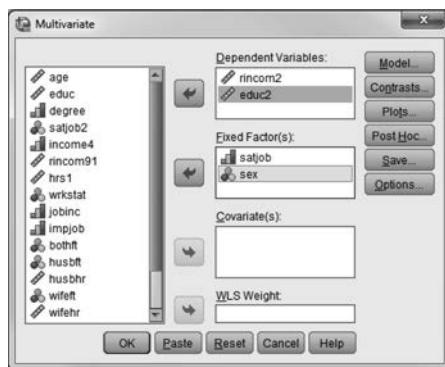
The final test assumption of homogeneity of variance-covariance will be tested within MANOVA using the **Multivariate** procedure. To open the Multivariate dialog box as shown in Figure 6.9, select the following:

Analyze
General Linear Model
Multivariate

Multivariate dialog box (see Figure 6.9)

Once in this box, click the DVs (*rincom2* and *educ2*) and move each to the **Dependent Variables** box. Click the IVs (*satjob* and *sex*) and move each to the **Fixed Factor(s)** box. Then click **Options**.

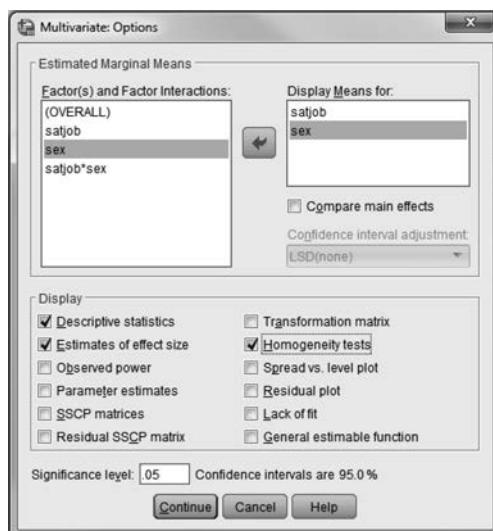
Figure 6.9. Multivariate Dialog Box.



Multivariate: Options dialog box (see Figure 6.10)

Move both IVs to the **Display Means for** box. Select **Descriptive statistics**, **Estimates of effect size**, and **Homogeneity tests** under **Display**. These options are described in Chapter 4. Click **Continue**. Back in the **Multivariate** dialog box, click **Post Hoc**.

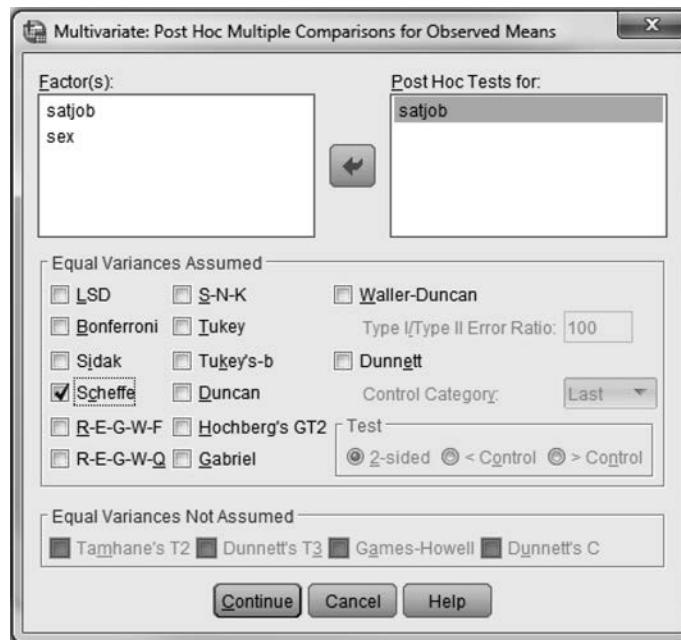
Figure 6.10. Multivariate: Options Dialog Box.



Multivariate: Post Hoc Multiple Comparisons for Observed Means dialog box (see Figure 6.11)

Under **Factor(s)**, select the IVs (*satjob*) and move to **Post Hoc Tests for** box. For our example, only *satjob* was selected because gender has only two categories. Under **Equal Variances Assumed**, select the desired post hoc test. We selected Scheffé. Click **Continue** then **OK**.

Figure 6.11. Multivariate: Post Hoc Multiple Comparisons for Observed Means Dialog Box.



Output and Interpretation of Results

Figures 6.12 through 6.17 present some of the MANOVA output. The Box's test (see Figure 6.12) is not significant and indicates that homogeneity of variance-covariance is fulfilled [$F(21, 20370) = 1.245$, $p = .201$], so the Wilks' Lambda test statistic will be used in interpreting the MANOVA results. The multivariate tests are presented in Figure 6.13. Factor interaction was then examined, and it revealed nonsignificance [$F(6, 1344) = .749$, $p = .610$, $\eta^2 = .003$]. The main effects of job satisfaction [$F(6, 1344) = 3.98$, $p = .001$, $\eta^2 = .017$] and gender [$F(2, 672) = 8.14$, $p < .001$, $\eta^2 = .024$] are both significant. However, multivariate effect sizes are very small. Prior to examining the univariate ANOVA results, the alpha level was adjusted to $\alpha = .025$ because two DVs were analyzed. Univariate ANOVA results (see Figure 6.14) indicate that income significantly differs for job satisfaction [$F(3, 673) = 7.17$, $p < .001$, $\eta^2 = .031$] and gender [$F(1, 673) = 16.14$, $p < .001$, $\eta^2 = .023$]. Years of education do not significantly differ for job satisfaction [$F(3, 673) = 2.18$, $p = .089$, $\eta^2 = .010$] or gender [$F(1, 673) = 1.03$, $p = .310$, $\eta^2 = .002$]. Scheffé post hoc results (see Figure 6.15) for income and job satisfaction indicate that very satisfied individuals significantly differ from those with only moderate satisfaction. Figures 6.16 and 6.17 present the unadjusted and adjusted group means for income and years of education.

Figure 6.12. Box's Test for Homogeneity of Variance-Covariance.

Box's Test of Equality of Covariance Matrices	
Box's M	26.935
F	1.245
df1	21
df2	20370.100
Sig.	.201

a. Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design:
Intercept +
satjob + sex
+ satjob * sex

Box's test is NOT significant. Use Wilks' Lambda criteria.

Figure 6.13. MANOVA Summary Table.

Multivariate Tests ^a						
Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.923	4042.110 ^b	2.000	.672.000	.000 .923
	Wilks' Lambda	.077	4042.110 ^b	2.000	.672.000	.000 .923
	Hotelling's Trace	12.030	4042.110 ^b	2.000	.672.000	.000 .923
	Roy's Largest Root	12.030	4042.110 ^b	2.000	.672.000	.000 .923
satjob	Pillai's Trace	.035	3.967	6.000	1346.000	.001 .017
	Wilks' Lambda	.965	3.984 ^b	6.000	1344.000	.001 .017
	Hotelling's Trace	.036	4.002	6.000	1342.000	.001 .018
	Roy's Largest Root	.033	7.291 ^c	3.000	.673.000	.000 .031
sex	Pillai's Trace	.024	8.135 ^b	2.000	.672.000	.000 .024
	Wilks' Lambda	.976	8.135 ^b	2.000	.672.000	.000 .024
	Hotelling's Trace	.024	8.135 ^b	2.000	.672.000	.000 .024
	Roy's Largest Root	.024	8.135 ^b	2.000	.672.000	.000 .024
satjob * sex	Pillai's Trace	.007	.750	6.000	1346.000	.609 .003
	Wilks' Lambda	.993	.749 ^b	6.000	1344.000	.610 .003
	Hotelling's Trace	.007	.748	6.000	1342.000	.611 .003
	Roy's Largest Root	.005	1.051 ^c	3.000	.673.000	.370 .005

a. Design: Intercept + satjob + sex + satjob * sex

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Indicates that job satisfaction significantly affects the combined DV.

Indicates that gender significantly affects the combined DV.

Indicates that factor interaction is NOT significantly affecting the combined DV.

Figure 6.14. Univariate ANOVA Summary Table.

Tests of Between-Subjects Effects							
Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	rincom2	1101.382 ^a	7	157.340	9.676	.000	.091
	educ2	64.156 ^b	7	9.165	1.358	.220	.014
Intercept	rincom2	47955.386	1	47955.386	2949.195	.000	.814
	educ2	49956.968	1	49956.968	7404.472	.000	.917
satjob	rincom2	349.728	3	116.576	7.169	.000	.031
	educ2	44.078	3	14.693	2.178	.089	.010
sex	rincom2	262.463	1	262.463	16.141	.000	.023
	educ2	6.952	1	6.952	1.030	.310	.002
satjob * sex	rincom2	26.942	3	8.981	.552	.647	.002
	educ2	15.304	3	5.101	.756	.519	.003
Error	rincom2	10943.317	673	16.261			
	educ2	4540.639	673	6.747			
Total	rincom2	149640.000	681				
	educ2	139907.000	681				
Corrected Total	rincom2	12044.699	680				
	educ2	4604.796	680				

a. R Squared = .091 (Adjusted R Squared = .082)

b. R Squared = .014 (Adjusted R Squared = .004)

Indicates that job satisfaction significantly affects income but NOT years of education.

Indicates that gender significantly affects income but NOT years of education.

Figure 6.15. Post Hoc Tests for Income and Years of Education by Job Satisfaction.

Multiple Comparisons								
Dependent Variable	(I) Job Satisfaction (satjob)		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval		
						Lower Bound	Upper Bound	
	rincom2	Very satisfied	Mod satisfied	.15045	.33206	.000	.5739	2.4351
			A little dissatisfied	1.2174	.54505	.174	-.3101	2.7450
			Very dissatisfied	1.3151	.85551	.501	-1.0826	3.7127
		Mod satisfied	Very satisfied	-.15045	.33206	.000	-2.4351	-.5739
			A little dissatisfied	-.2871	.54642	.964	-1.8184	1.2443
			Very dissatisfied	-.1894	.85639	.997	-2.5895	2.2107
		A little dissatisfied	Very satisfied	-.2174	.54505	.174	-2.7450	.3101
			Mod satisfied	.2871	.54642	.964	-1.2443	1.8184
			Very dissatisfied	.0976	.95928	1.000	-2.5908	2.7861
		Very dissatisfied	Very satisfied	-.3151	.85551	.501	-3.7127	1.0826
			Mod satisfied	.1894	.85639	.997	-2.2107	2.5895
			A little dissatisfied	-.0976	.95928	1.000	-2.7861	2.5908
	educ2	Very satisfied	Mod satisfied	.4929	.21389	.152	-.1066	1.0923
			A little dissatisfied	.3211	.35109	.841	-.6629	1.3050
			Very dissatisfied	-.4706	.55108	.866	-2.0150	1.0738
		Mod satisfied	Very satisfied	-.4929	.21389	.152	-1.0923	.1066
			A little dissatisfied	-.1718	.35197	.971	-1.1582	.8146
			Very dissatisfied	-.9635	.55164	.385	-2.5095	.5825
		A little dissatisfied	Very satisfied	-.3211	.35109	.841	-1.3050	.6629
			Mod satisfied	.1718	.35197	.971	-.8146	1.1582
			Very dissatisfied	-.7917	.61791	.650	-2.5234	.9401
		Very dissatisfied	Very satisfied	.4706	.55108	.866	-1.0738	2.0150
			Mod satisfied	.9635	.55164	.385	-.5825	2.5095
			A little dissatisfied	.7917	.61791	.650	-.9401	2.5234

Based on observed means.

The error term is Mean Square(Error) = 6.747.

*. The mean difference is significant at the .05 level.

Figure 6.16. Unadjusted Group Means for Income and Years of Education by Job Satisfaction and Gender.

Descriptive Statistics				
	satjob	sex	Mean	Std. Deviation
rincom2	Very satisfied	Male	15.8193	3.73889
		Female	14.0301	4.00934
		Total	15.0234	3.95649
	Mod satisfied	Male	14.5157	4.32370
		Female	12.3182	4.03670
		Total	13.5189	4.32979
	A little dissatisfied	Male	15.2571	4.13978
		Female	12.2187	3.75658
		Total	13.8060	4.21843
	Very dissatisfied	Male	14.2143	5.05628
		Female	13.0000	2.86744
		Total	13.7083	4.24755
educ2	Total	Male	15.1524	4.11833
		Female	13.0717	4.03758
		Total	14.2144	4.20866
	Very satisfied	Male	14.2590	2.88556
		Female	14.3985	2.20859
		Total	14.3211	2.60303
	Mod satisfied	Male	13.7484	2.67659
		Female	13.9242	2.47621
		Total	13.8282	2.58471
educ2	A little dissatisfied	Male	14.1429	2.78803
		Female	13.8438	2.55405
		Total	14.0000	2.66288
	Very dissatisfied	Male	15.3571	2.46848
		Female	14.0000	2.16025
		Total	14.7917	2.39527
	Total	Male	14.0722	2.78595
		Female	14.1238	2.36346
		Total	14.0954	2.60226

Figure 6.17. Adjusted Group Means for Income and Years of Education by Job Satisfaction and Gender.**1. Job Satisfaction (satjob)**

Dependent Variable	Job Satisfaction (satjob)	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
rincom2	Very satisfied	14.925	.235	14.464	15.385
	Mod satisfied	13.417	.237	12.951	13.883
	A little dissatisfied	13.738	.493	12.770	14.706
	Very dissatisfied	13.607	.835	11.968	15.246
educ2	Very satisfied	14.329	.151	14.032	14.626
	Mod satisfied	13.836	.153	13.536	14.137
	A little dissatisfied	13.993	.318	13.370	14.617
	Very dissatisfied	14.679	.538	13.623	15.734

2. Respondent's Sex

Dependent Variable	Respondent's Sex	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
rincom2	Male	14.952	.338	14.288	15.615
	Female	12.892	.386	12.135	13.649
educ2	Male	14.377	.218	13.950	14.804
	Female	14.042	.248	13.554	14.529

Presentation of Results

The following narrative summarizes the results for the two-way MANOVA example.

A two-way MANOVA was conducted to determine the effect of job satisfaction and gender on the two dependent variables of respondents' income and years of education. Data were first transformed to eliminate outliers. Respondents' income was transformed to remove cases with income of zero or equal to or exceeding 22. Years of education was also transformed to eliminate cases with 6 or fewer years. MANOVA results indicate that job satisfaction [Wilks' $\Lambda = .965$, $F(6, 1344) = 3.98$, $p = .001$, $\eta^2 = .017$] and gender [Wilks' $\Lambda = .976$, $F(2, 672) = 8.14$, $p < .001$, $\eta^2 = .024$] significantly affect the combined DV of income and years of education. However, multivariate effect sizes are very small. Univariate ANOVA and Scheffé post hoc tests were conducted as follow-up tests. ANOVA results indicate that income significantly differs for job satisfaction [$F(3, 673) = 7.17$, $p < .001$, $\eta^2 = .031$] and gender [$F(1, 673) = 16.14$, $p < .001$, $\eta^2 = .023$]. Years of education does not significantly differ for job satisfaction [$F(3, 673) = 2.18$, $p = .089$, $\eta^2 = .010$] or gender [$F(1, 673) = 1.03$, $p = .310$, $\eta^2 = .002$]. Scheffé post hoc results for income and job satisfaction indicate that very satisfied individuals significantly differ from those with only moderate satisfaction. Table 2 presents the adjusted and unadjusted group means for income and years of education by job satisfaction and gender.

Table 2

Adjusted and Unadjusted Group Means for Income and Years of Education by Gender and Job Satisfaction

	Income		Years of Education	
	Adjusted <i>M</i>	Unadjusted <i>M</i>	Adjusted <i>M</i>	Unadjusted <i>M</i>
Gender				
Male	14.95	15.15	14.38	14.07
Female	12.89	13.07	14.04	14.12
Job Satisfaction				
Very satisfied	14.93	15.02	14.33	14.32
Mod. satisfied	13.42	13.52	13.84	13.83
Little dissatisfied	13.74	13.81	13.99	14.00
Very dissatisfied	13.61	13.71	14.68	14.79

II. MANCOVA

As with univariate ANCOVA, researchers often wish to control for the effects of concomitant variables in a multivariate design. The appropriate analysis technique for this situation is a multivariate analysis of covariance, or MANCOVA. Multivariate analysis of covariance is essentially a combination of MANOVA and ANCOVA. MANCOVA asks if there are statistically significant mean differences among groups after adjusting the newly created DV (a linear combination of all original DVs) for differences on one or more covariates.

SECTION 6.5 PRACTICAL VIEW

Purpose

The main advantage of MANCOVA over MANOVA is the fact that the researcher can incorporate one or more covariates into the analysis. The effects of these covariates are then removed from the analysis, leaving the researcher with a clearer picture of the true effects of the IV(s) on the multiple DVs. There are two main reasons for including several (i.e., more than one) covariates in the analysis (Stevens, 2001). First, the inclusion of several covariates will result in a greater reduction in error variance than would result from incorporation of one covariate. Recall that in ANCOVA, the main reason for including a covariate is to remove from the error term unwanted sources of variability (variance within the groups), which could be attributed to the covariate. This ultimately results in a more sensitive *F* test, which increases the likelihood of rejecting the null hypothesis. By including more covariates in a MANCOVA analysis, we can reduce this unwanted error by an even greater amount, thus improving the chances of rejecting a null hypothesis that is really false.

A second reason for including more than one covariate is that it becomes possible to make better adjustments for initial differences in situations where the research design includes the use of intact groups (Stevens, 2001). The researcher has even more information upon which to base the statistical matching procedure. In this case, the means of the linear combination of DVs for each group are adjusted to what they would be if all groups had scored equally on the combination of covariates.

Again, the researcher needs to be cognizant of the choice of covariates in a multivariate analysis. There should exist a significant relationship between the set of DVs and the covariate or set of covariates (Stevens, 2001). Similar to ANCOVA, if more than one covariate is being used, there should be relatively low intercorrelations among all covariates (roughly $< .40$). In ANCOVA, the amount of error reduction was

a result of the magnitude of the correlation between the DV and the covariate. In MANCOVA, if several covariates are being used, the amount of error reduction is determined by the magnitude of the multiple correlation (R^2) between the newly created DV and the set of covariates (Stevens, 2001). A higher value for R^2 is directly associated with low intercorrelations among covariates, which means a greater degree of error reduction.

The null hypothesis being tested in MANCOVA is that the *adjusted* population mean vectors are equal:

$$H_0: \boldsymbol{\mu}_1^{\text{adj}} = \boldsymbol{\mu}_2^{\text{adj}} = \boldsymbol{\mu}_3^{\text{adj}} = \dots = \boldsymbol{\mu}_k^{\text{adj}}$$

Wilks' Lambda (Λ) is again the most common test statistic used in MANCOVA. However, in this case, the sum-of-squares and cross-products (SSCP) matrices are first adjusted for the effects of the covariate(s).

The procedure to be used in conducting MANCOVA mirrors that used in conducting MANOVA. Following the statistical adjustment of newly created DV scores, the overall multivariate null hypothesis is evaluated using Wilks' Λ . If the null hypothesis is retained, interpretation of the analysis ceases at this point. However, if the overall null hypothesis is rejected, the researcher then examines the results of univariate ANCOVAs in order to discover which DVs are being affected by the IV(s). A Bonferroni-type adjustment to protect from the potential of an inflated Type I error rate is again appropriate at this point.

In Chapter 5, we mentioned a specific application of MANCOVA that is used to assess the contribution of each individual DV to any significant differences in the IVs. This procedure is accomplished by removing the effects of all other DVs by treating them as covariates in the analysis.

Sample Research Questions

In the sample study presented earlier in this chapter, we investigated the differences in worker productivity (measured by income level, DV₁, and hours worked, DV₂) for individuals in different age categories (IV). Assume that the variable of years of education has been shown to relate to both DVs and we want to remove its effect from the analysis. Consequently, we decide to include years of education as a covariate in our analysis. The design we now have is a one-way MANCOVA. Accordingly, this study should address the following research questions:

1. Are there significant mean differences in worker productivity (as measured by the combination of income and hours worked) for individuals of different ages, after removing the effect of years of education?
2. Are there significant mean differences in income levels for individuals of different ages, after removing the effect of years of education?
 - 2a. If so, which age categories differ?
3. Are there significant mean differences in hours worked for individuals of different ages, after removing the effect of years of education?
 - 3a. If so, which age categories differ?

For our second MANCOVA example, we will add a covariate to the two-factor design presented earlier. This two-way MANCOVA will investigate differences in the combined DV of income level (DV₁) and years of education (DV₂) for individuals of different gender (IV₁) and of different levels of job satisfaction (IV₂), while controlling for age. Again, one should note that the following research questions address both the multivariate and the univariate analyses within MANCOVA:

1.
 - a. Are there significant mean differences in the combined DV of income and years of education between males and females, after removing the effect of age?
 - b. Are there significant mean differences in the combined DV of income and years of education for different levels of job satisfaction, after removing the effect of age? If so, which job-satisfaction categories differ?
 - c. Is there a significant interaction between gender and job satisfaction on the combined DV of income and years of education, after removing the effect of age?
2.
 - a. Are there significant mean differences on income between males and females, after removing the effect of age?
 - b. Are there significant mean differences on income among different levels of job satisfaction, after removing the effect of age? If so, which job satisfaction categories differ?
 - c. Is there a significant interaction between gender and job satisfaction on income, after removing the effect of age?
3.
 - a. Are there significant mean differences in years of education between males and females, after removing the effect of age?
 - b. Are there significant mean differences in years of education among different levels of job satisfaction, after removing the effect of age? If so, which job satisfaction categories differ?
 - c. Is there a significant interaction between gender and job satisfaction on years of education, after removing the effect of age?

SECTION 6.6 ASSUMPTIONS AND LIMITATIONS

Multivariate analysis of covariance rests on the same basic assumptions as univariate ANCOVA. However, the assumptions for MANCOVA must accommodate multiple DVs. The following list presents the assumptions for MANCOVA, with an asterisk indicating modification from the ANCOVA assumptions.

1. The observations within each sample must be randomly sampled and must be independent of each other.
- 2.* The distributions of scores on the dependent variables must be normal in the populations from which the data were sampled.
- 3.* The distributions of scores on the dependent variables must have equal variances.
- 4.* Linear relationships must exist between all pairs of DVs, all pairs of covariates, and all DV-covariate pairs in each cell.
- 5.* If two covariates are used, the regression planes for each group must be homogeneous or parallel. If more than two covariates are used, the regression hyperplanes must be homogeneous.
6. The covariates are reliable and are measured without error.

The first and sixth assumptions essentially remain unchanged. Assumptions 2 and 3 are simply modified in order to include multiple DVs. Assumption 4 has a substantial modification in that we must now assume linear relationships not only between the DV and the covariate, but also among several other pairs of variables (Tabachnick & Fidell, 2007). There also exists an important modification to Assumption 5. Recall that if only one covariate is included in the analysis, there exists the assumption that covariate regression slopes for each group are homogeneous. However, if the MANCOVA analysis involves more than one covariate, the analogous assumption involves homogeneity of regression planes (for two covariates) and hyperplanes (for three or more covariates).

Our discussion of assessing MANCOVA assumptions will center on the two substantially modified assumptions (i.e., 4 and 5). Similar procedures, as have been discussed earlier, are used for testing the remaining assumptions.

Methods of Testing Assumptions

The assumption of normally distributed DVs is assessed in the usual manner. Initial assessment of normality is done through inspection of histograms, boxplots, and normal Q-Q plots. Statistical assessment of normality is accomplished by examining the values (and the associated significance tests) for skewness and kurtosis and through the use of the Kolmogorov-Smirnov test. The assumption of homoscedasticity is assessed primarily with Box's test or using one of three different statistical tests discussed in Chapters 3 and 5, namely Hartley's F_{\max} test, Cochran's test, or Levene's test.

The assumption of linearity among all pairs of DVs and covariates is crudely assessed by inspecting the within-cells bivariate scatterplots between all pairs of DVs, all pairs of covariates, and all DV-covariate pairs. This process is feasible if the analysis includes only a small number of variables. However, the process becomes much more cumbersome (and potentially unmanageable) with analyses involving the examination of numerous DVs and/or covariates—just imagine all the possible bivariate pairings! If a researcher is involved in such an analysis, one recommendation is to engage in “spot checks” of random bivariate relationships or bivariate relationships in which nonlinearity may be likely (Tabachnick & Fidell, 2007).

Once again, if curvilinear relationships are indicated, they may be corrected by transforming some or all of the variables. Bear in mind that transforming the variables may create difficulty in interpretations. One possible solution might be to eliminate the covariate that appears to produce nonlinearity, replacing it with another appropriate covariate (Tabachnick & Fidell, 2007).

Remember that a violation of the assumption of homogeneity of regression slopes (as well as regression planes and hyperplanes) is an indication that there is a covariate by treatment (IV) interaction, meaning that the relationship between the covariate and the newly created DV is different at different levels of the IV(s). A preliminary or custom MANCOVA can be conducted to test the assumption of homogeneity of regression planes (in the case of two covariates) or regression hyperplanes (in the case of three or more covariates). If the analysis contains more than one covariate, there is an interaction effect for each covariate. The effects are lumped together and tested as to whether the combined interactions are significant (Stevens, 2001).

The null hypothesis being tested in these cases is that all regression planes/hyperplanes are equal and parallel. Rejecting this hypothesis means that there is a significant interaction between covariates and IVs and that the planes/hyperplanes are not equal. If a researcher is to continue in the use of multivariate analysis of covariance, he or she would hope to fail to reject this particular null hypothesis. In SPSS, this is determined by examining the results of the F test for the interaction of the IV(s) by the covariate(s).

SECTION 6.7 PROCESS AND LOGIC

The Logic Behind MANCOVA

The calculations for MANCOVA are nearly identical to those for MANOVA. The only substantial difference is that the sum-of-squares and cross-products (SSCP) matrices must first be adjusted for the effects of the covariate(s). The adjusted matrices are symbolized by T^* (adjusted total sum-of-squares and cross-products matrix), W^* (adjusted within sum-of-squares and cross-products matrix), and B^* (adjusted between sum-of-squares and cross-products matrix).

Wilks' Λ is again calculated by using the SSCP matrices (Stevens, 2001). We can compare the MANOVA and MANCOVA formulas for Λ :

MANOVA

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}$$

MANCOVA

$$\Lambda^* = \frac{|\mathbf{W}^*|}{|\mathbf{T}^*|} = \frac{|\mathbf{W}^*|}{|\mathbf{B}^* + \mathbf{W}^*|} \quad (\text{Equation 6.4})$$

The interpretation of Λ remains as it was in MANOVA. If there is no treatment effect and there are no group differences, then $\mathbf{B}^* = 0$ and $\Lambda^* = 1$, indicating no differences between groups on the linear combination of DVs after removing the effects of the covariate(s). Whereas if \mathbf{B}^* were very large, then Λ^* would approach 0, indicating significant group differences on the combination of DVs, after controlling for the covariate(s).

As in MANOVA, eta squared for MANCOVA is obtained in the following manner:

$$\eta^2 = 1 - \Lambda$$

In the multivariate analysis of covariance situation, η^2 is interpreted as the variance accounted for in the best linear combination of DVs by the IV(s) and/or interactions of IV(s), after removing the effects of any covariate(s).

Interpretation of Results

Interpretation of MANCOVA results is quite similar to that of MANOVA. However, with the inclusion of covariates, interpretation of a preliminary MANCOVA is necessary in order to test the assumption of homogeneity of regression slopes. Essentially, this analysis tests for the interaction between the factors (IVs) and covariates. This preliminary or custom MANCOVA will also test homogeneity of variance-covariance (Box's test), which is actually interpreted first because it helps in identifying the appropriate test statistic to be utilized in examining the homogeneity of regression and the final MANCOVA results. If the Box's test is significant at $p < .001$ and group sample sizes are extremely unequal, then Pillai's Trace is utilized when interpreting the homogeneity of regression test and the MANOVA results. If equal variances are assumed, Wilks' Lambda should be used as the multivariate test statistic. Once the test statistic has been determined, then the homogeneity of regression slopes or planes results are interpreted by examining the F ratio and p value for the interaction. If factor-covariate interaction is significant, then MANCOVA is not an appropriate analysis technique. If interaction is not significant, then one can proceed with conducting the full MANCOVA analysis. Using the F ratio and p value for a test statistic that was identified in the preliminary analysis through the Box's test, factor interaction should be examined if two or more IVs are utilized in the analysis. If factor interaction is significant, then main effects for each factor on the combined DV is not a valid indicator of effect. If factor interaction is not significant, the main effects for each IV can be accurately interpreted by examining the F ratio, p value, and effect size for the appropriate test statistic. When main effects are significant, univariate ANOVA results indicate group differences

for each DV. Because MANCOVA does not provide post hoc analyses, examining group means (before and after covariate adjustment) for each DV can assist in determining how groups differed for each DV.

In summary, the first step in interpreting the MANCOVA results is to evaluate the preliminary MANCOVA results that include the Box's test and the test for homogeneity of regression slopes. If Box's test is not significant, utilize the Wilks' Lambda statistic when interpreting the homogeneity of regression slopes and the subsequent multivariate tests. If Box's test is significant, use Pillai's Trace. Once the multivariate test statistic has been identified, examine the significance (F ratios and p values) of factor-covariate interaction (homogeneity of regression slopes). If factor-covariate interaction is not significant, then proceed with the full MANCOVA. To interpret the full MANCOVA results, examine the significance (F ratios and p values) of factor interaction. This is necessary only if two or more IVs are included. Next, evaluate the F ratio, p value, and effect size for each factor's main effect. If multivariate significance is found, interpret the univariate ANOVA results to determine significant group differences for each DV.

For our example² that investigates age category (*agecat4*), differences in respondents' income (*rincom91*), and hours worked per week (*hrs1*) when controlling for education level (*educ*), the previously transformed variables of *rincom2*, *hrs2*, and *educ2* were utilized. These transformations are described in Section 6.3. Linearity of the two DVs (*rincom2*, *hrs2*) and the covariate (*educ2*) was then tested by creating a matrix scatterplot and calculating Pearson correlation coefficients. Results indicate linear relationships. Although the correlation coefficients are statistically significant, all are quite low. The last assumption, homogeneity of variance-covariance, was tested within a preliminary MANCOVA analysis utilizing **Multivariate** (see steps for conducting a preliminary/custom MANCOVA in Figure 6.35). The Box's test (see Figure 6.18) reveals that equal variances can be assumed [$F(9, 2827520) = .634, p = .769$]. Therefore, Wilks' Lambda will be used as the multivariate statistic. Figure 6.19 presents the MANCOVA results for the homogeneity of regression test. The interaction between *agecat4* and *educ2* is not significant [Wilks' $\Lambda = .993, F(6, 1342) = .815, p = .558$]. A full MANCOVA was then conducted using **Multivariate** (see Figure 6.20). Wilks' Lambda criteria indicate significant group differences in age category with respect to income and hours worked per week [Wilks' $\Lambda = .898, F(6, 1348) = 12.36, p < .001$, multivariate $\eta^2 = .052$]. Univariate ANOVA results (see Figure 6.21) reveal that age category significantly differs for only income [$F(3, 675) = 24.18, p < .001$, partial $\eta^2 = .097$] and not hours worked per week [$F(3, 675) = .052, p = .984$, partial $\eta^2 = .000$]. A comparison of unadjusted and adjusted means shows that individuals 18–29 years of age have income that is more than 3 points lower than those 40–49 and older than 50 (see Figure 6.22).

² Once again, the data set *career-f.sav* is used.

Figure 6.18. Box's Test for Homogeneity of Variance-Covariance.

Box's Test of Equality of Covariance Matrices ^a	
Box's M	5.740
F	.634
df1	9
df2	2827520.026
Sig.	.769

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + agecat4 + educ2 + agecat4 * educ2

Box's test is not significant. Use Wilks' Lambda criteria.

Figure 6.19. MANCOVA Summary Table: Test for Homogeneity of Regression Slopes.

Multivariate Tests ^c						
Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.284	133.096 ^a	2.000	671.000	.000
	Wilks' Lambda	.716	133.096 ^a	2.000	671.000	.000
	Hotelling's Trace	.397	133.096 ^a	2.000	671.000	.000
	Roy's Largest Root	.397	133.096 ^a	2.000	671.000	.000
agecat4	Pillai's Trace	.004	.504	6.000	1344.000	.805
	Wilks' Lambda	.996	.504 ^a	6.000	1342.000	.806
	Hotelling's Trace	.005	.503	6.000	1340.000	.806
	Roy's Largest Root	.004	.833 ^b	3.000	672.000	.476
educ2	Pillai's Trace	.106	39.974 ^a	2.000	671.000	.000
	Wilks' Lambda	.894	39.974 ^a	2.000	671.000	.000
	Hotelling's Trace	.119	39.974 ^a	2.000	671.000	.000
	Roy's Largest Root	.119	39.974 ^a	2.000	671.000	.000
agecat4 * educ2	Pillai's Trace	.007	.816	6.000	1344.000	.558
	Wilks' Lambda	.993	.815 ^a	6.000	1342.000	.558
	Hotelling's Trace	.007	.814	6.000	1340.000	.559
	Roy's Largest Root	.005	1.169 ^b	3.000	672.000	.321

a. Exact statistic.

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept + agecat4 + educ2 + agecat4 * educ2

Indicates that factor-covariate interaction is not significant.

Figure 6.20. MANCOVA Summary Table.

Multivariate Tests ^c						
Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.298	142.742 ^a	2.000	674.000	.000
	Wilks' Lambda	.702	142.742 ^a	2.000	674.000	.000
	Hotelling's Trace	.424	142.742 ^a	2.000	674.000	.000
	Roy's Largest Root	.424	142.742 ^a	2.000	674.000	.000
educ2	Pillai's Trace	.126	48.428 ^a	2.000	674.000	.000
	Wilks' Lambda	.874	48.428 ^a	2.000	674.000	.000
	Hotelling's Trace	.144	48.428 ^a	2.000	674.000	.000
	Roy's Largest Root	.144	48.428 ^a	2.000	674.000	.000
agecat4	Pillai's Trace	.102	12.037	6.000	1350.000	.000
	Wilks' Lambda	.898	12.356 ^a	6.000	1348.000	.000
	Hotelling's Trace	.113	12.673	6.000	1346.000	.000
	Roy's Largest Root	.113	25.371 ^b	3.000	675.000	.000

Indicates that job satisfaction significantly affects the combined DV.

Indicates that job satisfaction significantly affects the combined DV.

a. Exact statistic.

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept + educ2 + agecat4

Figure 6.21. Univariate ANOVA Summary Table.

Tests of Between-Subjects Effects							
Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	rincom2	2411.930 ^a	4	602.983	42.456	.000	.201
	hrs2	1490.452 ^b	4	372.613	2.948	.020	.017
Intercept	rincom2	922.346	1	922.346	64.943	.000	.088
	hrs2	33502.664	1	33502.664	265.018	.000	.282
educ2	rincom2	1355.655	1	1355.655	95.452	.000	.124
	hrs2	1439.392	1	1439.392	11.386	.001	.017
agecat4	rincom2	1030.099	3	343.386	24.177	.000	.097
	hrs2	19.820	3	6.607	.052	.984	.000
Error	rincom2	9586.657	675	14.202			
	hrs2	85331.025	675	126.416			
Total	rincom2	149199.000	680				
	hrs2	1567026.000	680				
Corrected Total	rincom2	11998.587	679				
	hrs2	86821.476	679				

a. R Squared = .201 (Adjusted R Squared = .196)

b. R Squared = .017 (Adjusted R Squared = .011)

Indicates that age category significantly affects income but NOT hours worked.

Figure 6.22. Unadjusted and Adjusted Means for Income and Hours Worked per Week by Age Category.

Descriptive Statistics				
	agecat4 4 categories...	Mean	Std. Deviation	N
rincom2	18-29	11.8672	4.14381	128
	30-39	14.0315	3.88102	222
	40-49	15.3333	3.88491	192
	50+	15.0797	4.31440	138
	Total	14.2044	4.20368	680
hrs2	18-29	46.3203	10.32002	128
	30-39	47.0315	11.41817	222
	40-49	46.5052	11.80381	192
	50+	46.5725	11.40488	138
	Total	46.6559	11.30782	680

4 categories of age

Dependent Variable	4 categories of age	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
rincom2	18-29	11.993 ^a	.333	11.339	12.648
	30-39	13.887 ^a	.253	13.389	14.384
	40-49	15.356 ^a	.272	14.822	15.890
	50+	15.165 ^a	.321	14.535	15.795
hrs2	18-29	46.450 ^a	.995	44.497	48.403
	30-39	46.882 ^a	.756	45.398	48.366
	40-49	46.528 ^a	.811	44.935	48.122
	50+	46.660 ^a	.957	44.780	48.540

a. Covariates appearing in the model are evaluated at the following values: educ2 = 14.0985.

Writing Up Results

The process of summarizing MANCOVA results is almost identical to MANOVA. However, MANCOVA results will obviously include a statement of how the covariate influenced the DVs. One should note that although the preliminary MANCOVA results are quite important in the analysis process, these results are not reported because it is understood that if a full MANCOVA has been conducted, such assumptions have been fulfilled. Consequently, the MANCOVA results narrative should address the following:

1. Participant elimination and/or variable transformation
2. Full MANCOVA results (test statistic, F ratio, degrees of freedom, p value, and effect size)
 - a. Main effects for each IV and covariate on the combined DV
 - b. Main effect for the interaction between IVs
3. Univariate ANOVA results (F ratio, degrees of freedom, p value, and effect size)
 - a. Main effect for each IV and DV
 - b. Comparison of means to indicate which groups differ on each DV

Often, a table is created that compares the unadjusted and adjusted group means for each DV. For our example, the results statement includes all of these components with the exception of factor interaction because only one IV is utilized. The following results narrative applies the results from Figures 6.18 through 6.22.

Multivariate analysis of covariance (MANCOVA) was conducted to determine the effect of the age category on employee productivity as measured by income and hours worked per week while controlling for years of education. Prior to the test, variables were transformed to eliminate outliers. Cases with income equal to zero and equal to or exceeding 22 were eliminated. Hours worked per week was transformed. Those less than or equal to 16 were recoded 17, and those greater than or equal to 80 were recoded 79. Years of education was also transformed to eliminate cases with 6 or fewer years. MANCOVA results revealed significant differences among the age categories on the combined dependent variable [Wilks' $\Lambda = .898$, $F(6, 1348) = 12.36$, $p < .001$, multivariate $\eta^2 = .052$]. The covariate (years of education) significantly influenced the combined dependent variable [Wilks' $\Lambda = .874$, $F(2, 674) = 48.43$, $p < .001$, multivariate $\eta^2 = .126$]. Analysis of covariance (ANCOVA) was conducted on each dependent variable as a follow-up test to MANCOVA. Age category differences were significant for income [$F(3, 675) = 24.18$, $p < .001$, partial $\eta^2 = .097$], but not hours worked per week [$F(3, 675) = .052$, $p = .984$, partial $\eta^2 = .000$]. A comparison of adjusted means revealed that income of those 18–29 years of age differs by more than 3 points from those 40–49 years of age and those 50 years and older. Table 3 presents adjusted and unadjusted means for income and hours worked per week by age category.

Table 3

Adjusted and Unadjusted Means for Income and Hours Worked per Week by Age Category

Age	Income		Hours Worked per Week	
	Adjusted M	Unadjusted M	Adjusted M	Unadjusted M
18–29 years	11.99	11.87	46.45	46.32
30–39 years	13.89	14.03	46.88	47.03
40–49 years	15.36	15.33	46.53	46.51
50+ years	15.17	15.08	46.66	46.57

SECTION 6.8 MANCOVA SAMPLE STUDY AND ANALYSIS

This section provides a complete example that applies the entire process of conducting MANCOVA: development of research questions and hypotheses, data-screening methods, test methods, interpretation of output, and presentation of results. The SPSS data set *career-f.sav* is utilized. Our previous example demonstrates a one-way MANCOVA, while this example will present a two-way MANCOVA.

Problem

Utilizing the two-way MANOVA example previously presented, in which we examined the degree to which gender and job satisfaction affect income and years of education among employees, we are now interested in adding the covariate of age. Because two IVs are tested in this analysis, questions must also take into account the possible interaction between factors. The following research questions and respective null hypotheses address the multivariate main effects for each IV and the possible interaction between factors.

Research Questions

RQ1: Do income and years of education differ by gender among employees when controlling for age?

RQ2: Do income and years of education differ by job satisfaction among employees when controlling for age?

RQ3: Do gender and job satisfaction interact in the effect on income and years of education when controlling for age?

Null Hypotheses

- H_01 : Income and years of education will not differ by gender among employees when controlling for age.
- H_02 : Income and years of education will not differ by job satisfaction among employees when controlling for age.
- H_03 : Gender and job satisfaction will not interact in the effect on income and years of education when controlling for age.

Both IVs are categorical and include gender (*sex*) and job satisfaction (*satjob*). The DVs are respondents' income (*rincom2*) and years of education (*educ2*); both are quantitative. The covariate is years of age (*age*) and is quantitative. Note that the variables *rincom2* and *educ2* are transformed variables of *rincom91* and *educ*, respectively. Transformations of these variables are described in Section 6.3 of this chapter.

Methods and SPSS “How To”

Because variables were previously transformed to eliminate outliers, data screening is complete. MANCOVA test assumptions should now be examined. Linearity between the DVs and covariate is first assessed by creating a scatterplot matrix and calculating Pearson correlation coefficients. Scatterplots and correlation coefficients indicate linear relationships. Although three of the four correlation coefficients are significant ($p < .001$), coefficients are still fairly weak. The final test assumptions of homogeneity of variance-covariance and homogeneity of regression slopes will be tested in a preliminary MANCOVA using **Multivariate**. To open the Multivariate dialog box, select the following:

Analyze

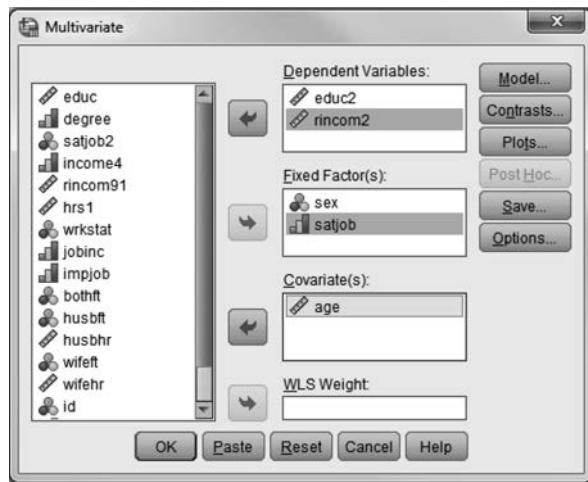
General Linear Model

Multivariate

Multivariate dialog box (see Figure 6.23)

Once in this dialog box, click each DV (*educ2* and *rincom2*) and move them to the **Dependent Variables** box. Click each IV (*sex* and *satjob*) and move them to the **Fixed Factor(s)** box. Then click each covariate (*age*) and move it to the **Covariate(s)** box. Then click **Model**.

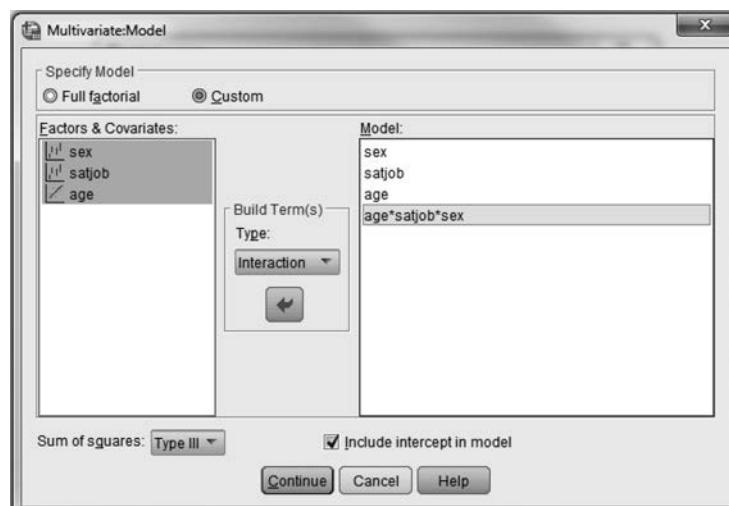
Figure 6.23.



Multivariate:Model dialog box (see Figure 6.24)

Under **Specify Model**, click **Custom**. Move each IV and covariate to the **Model** box. Then hold down the **Ctrl** key and highlight all IVs and covariate(s). Once highlighted, continue to hold down the **Ctrl** key and move them to the **Model** box. This should create the interaction between all IVs and covariate(s) (i.e., *age*satjob*sex*). Also, check to make sure that **Interaction** is specified in the **Build Term(s)** box. Click **Continue**. Back in the **Multivariate: Model** dialog box, click **Options**.

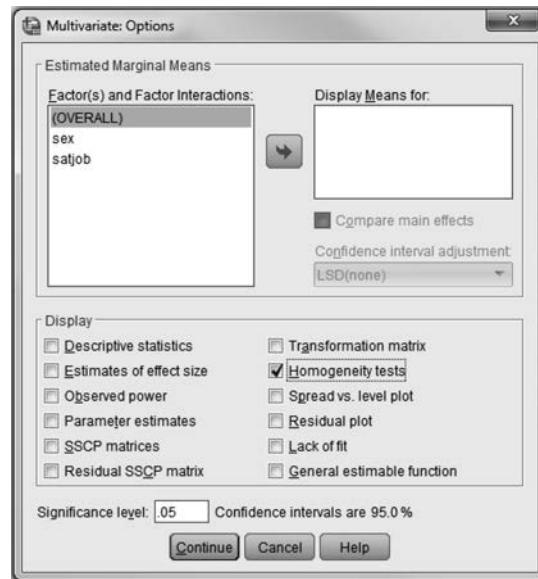
Figure 6.24. Multivariate: Model Dialog Box.



Multivariate: Options dialog box (see Figure 6.25)

Under **Display Means for**, click **Homogeneity tests**. Click **Continue**. Back in the **Multivariate** dialog box, click **OK**.

Figure 6.25. Multivariate: Options Dialog Box.



These steps will create the output to evaluate homogeneity of variance-covariance and homogeneity of regression slopes (see Figure 6.27). If interaction between the factors and covariates is not significant, then proceed with the following steps for conducting the full MANCOVA. For our example, Box's test (see Figure 6.26) indicates homogeneity of variance-covariance [$F(21, 20374) = 1.24, p = .204$]. Therefore, Wilks' Lambda will be utilized as the test statistic for all the multivariate tests. Figure 6.27 reveals that factor and covariate interaction are not significant [Wilks' $\Lambda = .976, F(14, 1332) = 1.143, p = .315$]. Full MANCOVA will then be conducted using **Multivariate**.

Figure 6.26. Box's Test for Homogeneity of Variance-Covariance.

Box's Test of Equality of Covariance Matrices ^a	
Box's M	26.868
F	1.242
df1	21
df2	20374.494
Sig.	.204

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + sex + satjob + age + sex * satjob * age

Box's test is not significant. Use Wilks' Lambda criteria.

Figure 6.27. MANCOVA Summary Table: Test for Homogeneity of Regression Slopes.

Multivariate Tests ^a						
Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.399	220.855 ^b	2.000	666.000	.000
	Wilks' Lambda	.601	220.855 ^b	2.000	666.000	.000
	Hotelling's Trace	.663	220.855 ^b	2.000	666.000	.000
	Roy's Largest Root	.663	220.855 ^b	2.000	666.000	.000
sex	Pillai's Trace	.001	.173 ^b	2.000	666.000	.841
	Wilks' Lambda	.999	.173 ^b	2.000	666.000	.841
	Hotelling's Trace	.001	.173 ^b	2.000	666.000	.841
	Roy's Largest Root	.001	.173 ^b	2.000	666.000	.841
satjob	Pillai's Trace	.010	1.100	6.000	1334.000	.360
	Wilks' Lambda	.990	1.100 ^b	6.000	1332.000	.360
	Hotelling's Trace	.010	1.100	6.000	1330.000	.360
	Roy's Largest Root	.009	1.947 ^c	3.000	667.000	.121
age	Pillai's Trace	.029	9.912 ^b	2.000	666.000	.000
	Wilks' Lambda	.971	9.912 ^b	2.000	666.000	.000
	Hotelling's Trace	.030	9.912 ^b	2.000	666.000	.000
	Roy's Largest Root	.030	9.912 ^b	2.000	666.000	.000
sex * satjob * age	Pillai's Trace	.024	1.143	14.000	1334.000	.315
	Wilks' Lambda	.976	1.143 ^b	14.000	1332.000	.315
	Hotelling's Trace	.024	1.142	14.000	1330.000	.315
	Roy's Largest Root	.018	1.702 ^c	7.000	667.000	.105

a. Design: Intercept + sex + satjob + age + sex * satjob * age

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Factor-covariate interaction is NOT significant.

To conduct the full Multivariate analysis, the same dialog boxes are opened, but different commands will be used. Open the **Multivariate** dialog box by selecting the following:

Analyze
General Linear Model
Multivariate

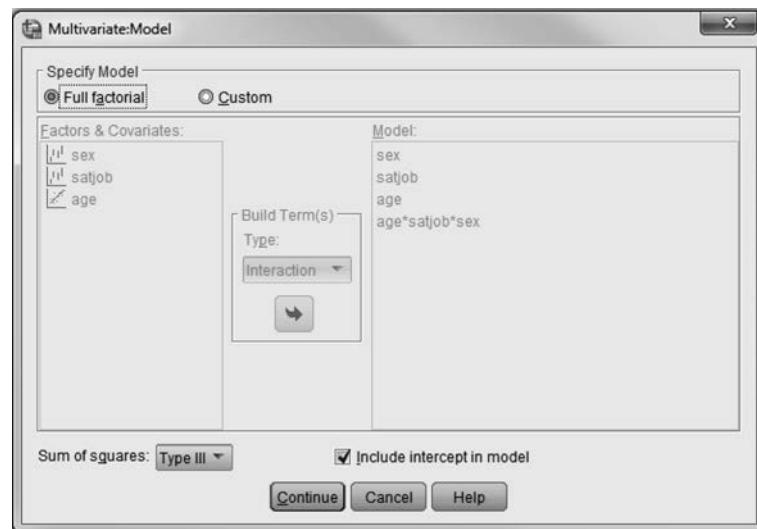
Multivariate dialog box (see Figure 6.23)

If you have conducted the preliminary MANCOVA, variables should already be identified. If not, proceed with the following. Click each DV (*educ2* and *rincom2*) and move each to the **Dependent Variables** box. Click each IV (*sex* and *satjob*) and move each to the **Fixed Factor(s)** box. Then click the covariate (*age*) and move it to the **Covariate(s)** box. Then click **Model**.

Multivariate: Model dialog box (see Figure 6.28)

Under **Specify Model**, we will now click **Full factorial**. Click **Continue**. Back in the **Multivariate** dialog box, click **Options**.

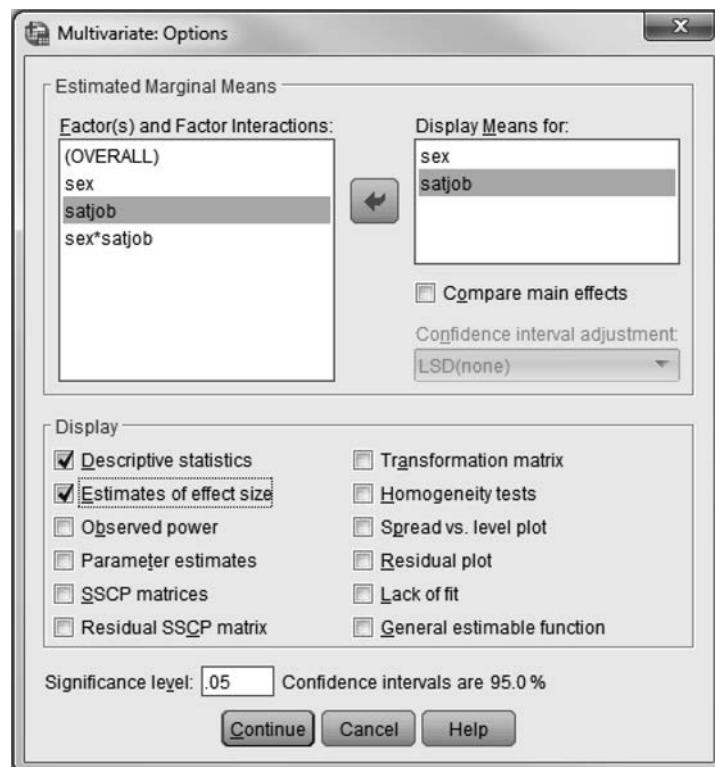
Figure 6.28. Multivariate: Model Dialog Box.



Multivariate: Options dialog box (see Figure 6.29)

Under **Factor(s) and Factor Interactions**, click both IVs and move them to the **Display Means for** box. Under **Display**, click **Descriptive statistics** and **Estimates of effect size**. Click **Continue**. Back in the **Multivariate** dialog box, click **OK**.

Figure 6.29. Multivariate: Options Dialog Box.



Output and Interpretation of Results

Figure 6.30 presents the unadjusted group means for each DV, while Figure 6.31 displays the adjusted group means. MANCOVA results are presented in Figure 6.32 and indicate no significant interaction between the two factors of gender and job satisfaction [Wilks' $\Lambda = .993$, $F(6, 1340) = .839$, $p = .539$]. The main effects of gender [Wilks' $\Lambda = .974$, $F(2, 670) = 9.027$, $p < .001$, multivariate $\eta^2 = .026$] and job satisfaction [Wilks' $\Lambda = .972$, $F(6, 1340) = 3.242$, $p = .004$, multivariate $\eta^2 = .014$] indicate significant effect on the combined DV. However, note the extremely small effect sizes for each IV. The covariate significantly influenced the combined DV [Wilks' $\Lambda = .908$, $F(2, 670) = 33.912$, $p < .001$, multivariate $\eta^2 = .092$]. Univariate ANOVA results (see Figure 6.33) indicate that only the DV of income was significantly affected by the IVs and covariate.

Figure 6.30. Unadjusted Group Means for Years of Education and Income.

Descriptive Statistics

	sex	satjob	Mean	Std. Deviation	N
educ2	Male	Very satisfied	14.2590	2.88556	166
		Mod satisfied	13.7484	2.67659	159
		A little dissatisfied	14.1429	2.78803	35
		Very dissatisfied	15.3571	2.46848	14
		Total	14.0722	2.78595	374
	Female	Very satisfied	14.4167	2.20700	132
		Mod satisfied	13.9242	2.47621	132
		A little dissatisfied	13.8438	2.55405	32
		Very dissatisfied	14.0000	2.16025	10
		Total	14.1307	2.36419	306
rincom2	Male	Very satisfied	14.3289	2.60392	298
		Mod satisfied	13.8282	2.58471	291
		A little dissatisfied	14.0000	2.66288	67
		Very dissatisfied	14.7917	2.39527	24
		Total	14.0985	2.60293	680
	Female	Very satisfied	15.8193	3.73889	166
		Mod satisfied	14.5157	4.32370	159
		A little dissatisfied	15.2571	4.13978	35
		Very dissatisfied	14.2143	5.05628	14
		Total	15.1524	4.11833	374
	Total	Very satisfied	13.9773	3.97793	132
		Mod satisfied	12.3182	4.03670	132
		A little dissatisfied	12.2187	3.75658	32
		Very dissatisfied	13.0000	2.86744	10
		Total	13.0458	4.01855	306
	Total	Very satisfied	15.0034	3.94789	298
		Mod satisfied	13.5189	4.32979	291
		A little dissatisfied	13.8060	4.21843	67
		Very dissatisfied	13.7083	4.24755	24
		Total	14.2044	4.20368	680

Figure 6.31. Adjusted Group Means for Years of Education and Income by Gender and Job Satisfaction.

1. Respondent's Sex

Dependent Variable	Respondent's Sex	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
educ2	Male	14.374 ^a	.218	13.946	14.802
	Female	14.043 ^a	.249	13.555	14.532
rincom2	Male	14.996 ^a	.325	14.358	15.633
	Female	12.921 ^a	.371	12.193	13.649

a. Covariates appearing in the model are evaluated at the following values: age = 40.34.

2. Job Satisfaction (satjob)

Dependent Variable	Job Satisfaction (satjob)	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
educ2	Very satisfied	14.345 ^a	.152	14.046	14.643
	Mod satisfied	13.830 ^a	.153	13.530	14.131
	A little dissatisfied	13.994 ^a	.318	13.370	14.618
	Very dissatisfied	14.666 ^a	.538	13.609	15.723
rincom2	Very satisfied	14.798 ^a	.226	14.353	15.242
	Mod satisfied	13.507 ^a	.229	13.058	13.955
	A little dissatisfied	13.727 ^a	.474	12.797	14.658
	Very dissatisfied	13.801 ^a	.803	12.225	15.378

a. Covariates appearing in the model are evaluated at the following values: age = 40.34.

Figure 6.32. MANCOVA Summary Table.

Multivariate Tests^a

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.670	679.424 ^b	2.000	670.000	.000	.670
	Wilks' Lambda	.330	679.424 ^b	2.000	670.000	.000	.670
	Hotelling's Trace	2.028	679.424 ^b	2.000	670.000	.000	.670
	Roy's Largest Root	2.028	679.424 ^b	2.000	670.000	.000	.670
age	Pillai's Trace	.092	33.912 ^b	2.000	670.000	.000	.092
	Wilks' Lambda	.908	33.912 ^b	2.000	670.000	.000	.092
	Hotelling's Trace	.101	33.912 ^b	2.000	670.000	.000	.092
	Roy's Largest Root	.101	33.912 ^b	2.000	670.000	.000	.092
sex	Pillai's Trace	.026	9.027 ^b	2.000	670.000	.000	.026
	Wilks' Lambda	.974	9.027 ^b	2.000	670.000	.000	.026
	Hotelling's Trace	.027	9.027 ^b	2.000	670.000	.000	.026
	Roy's Largest Root	.027	9.027 ^b	2.000	670.000	.000	.026
satjob	Pillai's Trace	.028	3.231	6.000	1342.000	.004	.014
	Wilks' Lambda	.972	3.242 ^b	6.000	1340.000	.004	.014
	Hotelling's Trace	.029	3.252	6.000	1338.000	.004	.014
	Roy's Largest Root	.026	5.894 ^c	3.000	671.000	.001	.026
sex * satjob	Pillai's Trace	.007	.840	6.000	1342.000	.539	.004
	Wilks' Lambda	.993	.839 ^b	6.000	1340.000	.539	.004
	Hotelling's Trace	.008	.838	6.000	1338.000	.540	.004
	Roy's Largest Root	.005	1.189 ^c	3.000	671.000	.313	.005

a. Design: Intercept + age + sex + satjob + sex * satjob

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

The covariate of age significantly influences the combined DV.

Gender significantly influences the combined DV.

Job satisfaction significantly influences the combined DV.

Factor interaction is NOT significant.

Figure 6.33. Univariate ANOVA Summary Table.

Tests of Between-Subjects Effects							
Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	educ2	69.142 ^a	8	8.643	1.280	.251	.015
	rincom2	1917.923 ^b	8	239.740	15.958	.000	.160
Intercept	educ2	9098.497	1	9098.497	1347.329	.000	.668
	rincom2	4331.310	1	4331.310	288.305	.000	.301
age	educ2	3.586	1	3.586	.531	.466	.001
	rincom2	813.705	1	813.705	54.163	.000	.075
sex	educ2	6.758	1	6.758	1.001	.317	.001
	rincom2	266.291	1	266.291	17.725	.000	.026
satjob	educ2	46.615	3	15.538	2.301	.076	.010
	rincom2	254.331	3	84.777	5.643	.001	.025
sex * satjob	educ2	15.551	3	5.184	.768	.512	.003
	rincom2	29.773	3	9.924	.661	.577	.003
Error	educ2	4531.257	671	6.753			
	rincom2	10080.664	671	15.023			
Total	educ2	139763.000	680				
	rincom2	149199.000	680				
Corrected Total	educ2	4600.399	679				
	rincom2	11998.587	679				

a. R Squared = .015 (Adjusted R Squared = .003)

b. R Squared = .160 (Adjusted R Squared = .150)

Gender significantly affects income but NOT years of education.

Job satisfaction significantly affects income but NOT years of education.

Presentation of Results

The following narrative summarizes the results from this two-way MANCOVA example.

A two-way MANCOVA was conducted to determine the effect of gender and job satisfaction on income and years of education while controlling for years of age. Data were first transformed to eliminate outliers. Respondents' income was transformed to eliminate cases with income of zero and equal to or exceeding 22. Years of education was also transformed to eliminate cases with 6 or fewer years. The main effects of gender [Wilks' $\Lambda = .974$, $F(2, 670) = 9.027$, $p < .001$, multivariate $\eta^2 = .026$] and job satisfaction [Wilks' $\Lambda = .972$, $F(6, 1340) = 3.242$, $p = .004$, multivariate $\eta^2 = .014$] indicate significant effect on the combined DV (Figure 6.32). The covariate significantly influenced the combined DV [Wilks' $\Lambda = .908$, $F(2, 670) = 33.912$, $p < .001$, multivariate $\eta^2 = .092$]. Univariate ANOVA results (Figure 6.33) indicate that only the DV of income was significantly affected by gender [$F(1, 671) = 17.73$, $p < .001$, partial $\eta^2 = .026$]; job satisfaction [$F(3, 671) = 5.64$, $p = .001$, partial $\eta^2 = .025$]; and the covariate of age [$F(1, 671) = 54.16$, $p < .001$, partial $\eta^2 = .075$]. Table 4 presents the adjusted and unadjusted group means for income and years of education. Comparison of adjusted income means indicates that those who are very satisfied have higher incomes than those less satisfied.

Table 4*Adjusted and Unadjusted Group Means for Income and Years of Education*

	Income		Years of Education	
	Adjusted <i>M</i>	Unadjusted <i>M</i>	Adjusted <i>M</i>	Unadjusted <i>M</i>
Gender				
Male	15.00	15.15	14.37	14.07
Female	12.92	13.05	14.04	14.13
Job Satisfaction				
Very satisfied	14.80	15.00	14.35	14.33
Mod. satisfied	13.51	13.52	13.83	13.83
Little dissatisfied	13.73	13.81	13.99	14.00
Very dissatisfied	13.80	13.71	14.67	14.79

SUMMARY

Multivariate analysis of variance (MANOVA) allows the researcher to examine group differences within a set of dependent variables. Factorial MANOVA will test the main effect for each factor on the combined DV as well as the interaction among factors on the combined DV. Usually, follow-up such as univariate ANOVA and post hoc tests are conducted within MANOVA to determine the specificity of group differences. Prior to conducting MANOVA, data should be screened for missing data and outliers. Data should also be examined for fulfillment of test assumptions: normality, homogeneity of variance-covariance, and linearity of DVs. Box's test for homogeneity of variance-covariance will help to determine which test statistic (e.g., Wilks' Lambda, Pillai's Trace) to utilize when interpreting the multivariate tests. The SPSS MANOVA table provides four different test statistics (Wilks' Lambda, Pillai's Trace, Hotelling's Trace, and Roy's Largest Root) with the *F* ratio, *p* value, and effect size that indicate the significance of factor main effects and interaction. Wilks' Lambda is the most commonly used criterion. If factor interaction is significant, then conclusions about main effects are limited. Univariate ANOVA and post hoc results determine group differences for each DV. Figure 6.34 provides a checklist for conducting MANOVA.

KEYWORDS

- Hotelling's Trace
- Pillai's Trace
- Wilks' Lambda (*A*)
- inverse criterion
- Roy's Largest Root
- multivariate statistics

Figure 6.34. Checklist for Conducting MANOVA.

I. Screen Data

- a. Missing Data?
 - Run Outliers and review Stem-and-Leaf plots and boxplots within **Explore**.
 - Eliminate or transform outliers if necessary.
- b. Outliers?
 - Run Outliers and review Stem-and-Leaf plots and boxplots within **Explore**.
 - Eliminate or transform outliers if necessary.
- c. Normality?
 - Run Normality Plots with Tests within **Explore**.
 - Review boxplots and histograms.
 - Transform data if necessary.
- d. Linearity of DVs?
 - Create Scatterplots.
 - Calculate Pearson correlation coefficients.
 - Transform data if necessary.
- e. Homogeneity of Variance-Covariance?
 - Run Box's test within **Multivariate**.

II. Conduct MANOVA

- a. Run MANOVA with post hoc test.
 1. **Analyze... General Linear Model... Multivariate.**
 - Move DVs to **Dependent Variables** box.
 - Move IVs to **Fixed Factor(s)** box.
 2. **Options.**
 - Move each IV to the **Display Means for** box.
 - Check **Descriptive statistics**, **Estimates of effect size**, and **Homogeneity tests**. **Continue**.
 3. **Post hoc.**
 - Move each IV to the **Post Hoc Test(s) for** box.
 - Select post hoc method. **Continue**.
 4. **OK.**
- b. Homogeneity of Variance-Covariance?
 - Examine F ratio and p value for Box's test.
 - If significant at $p < .001$ with extremely unequal group sample sizes, use Pillai's Trace for the test statistic.
 - If NOT significant at $p < .001$ with fairly equal group sample sizes, use Wilks' Lambda for the test statistic.
- c. Interpret factor interaction.
 - If factor interaction is significant, main effects are erroneous.
 - If factor interaction is NOT significant, interpret main effects.
- d. Interpret main effects for each IV on the combined DV.
- e. Interpret univariate ANOVA results.
- f. Interpret post hoc results.

III. Summarize Results

- a. Describe any data elimination or transformation.
- b. Narrate Full MANOVA results.
 - Main effects for each IV on the combined DV (test statistic, F ratio, p value, effect size).
 - Main effect for factor interaction (test statistic, F ratio, p value, effect size).
- c. Narrate univariate ANOVA results.
 - Main effects for each IV and DV (F ratio, p value, effect size).
- d. Narrate post hoc results.
- e. Draw conclusions.

Multivariate analysis of covariance (MANCOVA) allows the researcher to examine group differences within a set of dependent variables while controlling for covariate(s). Essentially, the influence that the covariate(s) has on the combined DV is partitioned out before groups are compared, such that group means of the combined DV are adjusted to eliminate the effect of the covariate(s). One-way MANCOVA will test the main effects for the factor on the combined DV while controlling for the covariate(s). Factorial MANCOVA will do the same but will also test the interaction among factors on the combined DV while controlling for the covariate(s). Usually, univariate ANCOVA is conducted within MANCOVA to determine the specificity of group differences. Prior to conducting MANCOVA, data should be screened for missing data and outliers. Data should also be examined for fulfillment of test assumptions: normality, homogeneity of variance-covariance, homogeneity of regression slopes, and linearity of DVs and covariates. A preliminary or custom MANCOVA must be conducted to test the assumptions of homogeneity of variance-covariance and homogeneity of regression slopes. Box's test for homogeneity of variance-covariance will help to determine which test statistic (e.g., Wilks' Lambda, Pillai's Trace) to utilize when interpreting the test for homogeneity of regression slopes and the full MANCOVA analyses. The test for homogeneity of regression slopes will indicate the degree to which the factors and covariate(s) interact to affect the combined DV. If interaction is significant, as indicated by the F ratio and p value for the appropriate test statistic, then the full MANCOVA should *NOT* be conducted. If interaction is not significant, then the full MANCOVA can be conducted. Once the full MANCOVA has been completed, factor interaction should be examined when two or more IVs are utilized. If factor interaction is significant, then conclusions about main effects are limited. Interpretation of the multivariate main effects and interaction is similar to MANOVA. Univariate ANOVA results determine the significance of group differences for each DV. Figure 6.35 provides a checklist for conducting MANCOVA.

Figure 6.35. Checklist for Conducting MANCOVA.

I. Screen Data

- a. Missing Data?
 - Run Outliers and review Stem-and-Leaf plots and boxplots within **Explore**.
 - Eliminate or transform outliers if necessary.
- b. Outliers?
 - Run Normality Plots with Tests within **Explore**.
 - Review boxplots and histograms.
 - Transform data if necessary.
- c. Normality?
 - Run Normality Plots with Tests within **Explore**.
 - Review boxplots and histograms.
 - Transform data if necessary.
- d. Linearity of DVs and covariate(s)?
 - Create Scatterplots.
 - Calculate Pearson correlation coefficients.
 - Transform data if necessary.
- e. Test remaining assumptions by conducting preliminary MANCOVA.

II. Conduct Preliminary (Custom) MANCOVA

- a. Run Custom MANCOVA.
 1. **Analyze... General Linear Model... Multivariate.**
 - Move DVs to **Dependent Variables** box.
 - Move IVs to **Fixed Factor(s)** box.
 - Move covariate(s) to **Covariate(s)** box.
 2. **Model.**
 3. **Custom.**
 - Move each IV and covariate to the **Model** box.
 - Hold down **Ctrl** key and highlight all IVs and covariate(s), **►** while still holding down the **Ctrl** key in order to move interaction to **Model** box. **Continue.**
 4. **Options.**
 - Check **Homogeneity tests**. **Continue.**
 5. **OK.**
- b. Homogeneity of Variance-Covariance?
 - Examine *F* ratio and *p* value for Box's test.
 - If significant at *p* < .001 with extremely unequal group sample sizes, use Pillai's Trace for the test statistic.
 - If *NOT* significant at *p* < .001 with fairly equal group sample sizes, use Wilks' Lambda for the test statistic.
- c. Homogeneity of Regression Slopes?
 - Using the appropriate test statistic, examine *F* ratio and *p* value for the interaction among IVs and covariates.
 - If interaction is significant, do not proceed with Full MANCOVA.
 - If interaction is *NOT* significant, proceed with Full MANCOVA.

III. Conduct MANCOVA

- a. Run Full MANCOVA.
 1. **Analyze... General Linear Model... Multivariate.**
 - Move DVs to **Dependent Variables** box.
 - Move IVs to **Fixed Factor(s)** box.
 - Move covariate(s) to **Covariate(s)** box.
 2. **Model.**
 3. **Full factorial.** **Continue.**
 4. **Options.**
 - Move each IV to the **Display Means for** box.
 - Check **Descriptive statistics** and **Estimates of effect size**. **Continue.**
 5. **OK.**
- b. Interpret factor interaction.
 - If factor interaction is significant, main effects are erroneous.
 - If factor interaction is *NOT* significant, interpret main effects.
- c. Interpret main effects for each IV on the combined DVs.
- d. Interpret univariate ANOVA results.

Figure 6.35 continues on the next page.

Figure 6.35. Checklist for Conducting MANCOVA (*continued*).

IV. Summarize Results

- a. Describe any data elimination or transformation.
- b. Narrate Full MANCOVA results.
 - Main effects for each IV and covariate on the combined DV (test statistic, F ratio, p value, effect size).
 - Main effect for factor interaction (test statistic, F ratio, p value, effect size).
- c. Narrate univariate ANOVA results.
 - Main effects for each IV and DV (F ratio, p value, effect size).
- d. Compare group means to indicate which groups differ on each DV.
- e. Draw conclusions.

Exercises for Chapter 6

The two exercises that follow utilize the data sets *career-a.sav* and *career-f.sav*, which can be downloaded from this website:

www.routledge.com/9781138289734

1. You are interested in evaluating the effect of job satisfaction (*satjob2*) and age category (*agecat4*) on the combined DV of hours worked per week (*hrs1*) and years of education (*educ*). Use *career-a.sav* for steps a and b.
 - a. Develop the appropriate research questions and/or hypotheses for main effects and interaction.
 - b. Screen data for missing data and outliers. What steps, if any, are necessary for reducing missing data and outliers?
- For all subsequent analyses in Question 1, use *career-f.sav* and the transformed variables of *hrs2* and *educ2*.
- c. Test the assumptions of normality and linearity of DVs.
 - i. What steps, if any, are necessary for increasing normality?
 - ii. Are DVs linearly related?

- d. Conduct MANOVA with post hoc (be sure to test for homogeneity of variance-covariance).
- i. Can you conclude homogeneity of variance-covariance? Which test statistic is most appropriate for interpretation of multivariate results?
 - ii. Is factor interaction significant? Explain.
 - iii. Are main effects significant? Explain.
 - iv. What can you conclude from univariate ANOVA and post hoc results?
- e. Write a results statement.
2. Building on the previous problem, in which you investigated the effects of job satisfaction (*satjob2*) and age category (*agecat4*) on the combined dependent variable of hours worked per week (*hrs1*) and years of education (*educ*), you are now interested in controlling for respondents' income such that *rincom91* will be used as a covariate. Complete the following using *career-a.sav*.
- a. Develop the appropriate research questions and/or hypotheses for main effects and interaction.
 - b. Screen data for missing data and outliers. What steps, if any, are necessary for reducing missing data and outliers?

For all subsequent analyses in Question 2, use *career-f.sav* and the transformed variables of *hrs2*, *educ2*, and *rincom2*.

- c. Test the assumptions of normality and linearity of DVs and covariate.

- i. What steps, if any, are necessary for increasing normality?

- ii. Are DVs and covariate linearly related?

- d. Conduct a preliminary MANCOVA to test the assumptions of homogeneity of variance-covariance and homogeneity of regression slopes/planes.
 - i. Can you conclude homogeneity of variance-covariance? Which test statistic is most appropriate for interpretation of multivariate results?
 - ii. Do factors and covariate significantly interact? Explain.
 - e. Conduct MANCOVA.
 - i. Is factor interaction significant? Explain.
 - ii. Are main effects significant? Explain.
 - iii. What can you conclude from univariate ANOVA results?
 - f. Write a results statement.
3. Compare the results from Question 1 and Question 2. Explain the differences in main effects.

CHAPTER 7

MULTIPLE REGRESSION

STUDENT LEARNING OBJECTIVES

After studying Chapter 7, students will be able to:

1. Explain differences between simple regression and multiple regression.
2. Describe what is meant by the terms *regression line* and *centroid*.
3. Summarize the concept of a least-squares solution in a multiple regression analysis.
4. Explain the relationship between multicollinearity and orthogonality.
5. Compare and contrast standard, sequential, and stepwise approaches to multiple regression.
6. Compare and contrast forward selection, stepwise selection, and backward deletion as approaches to stepwise multiple regression.
7. Explain why the assumptions associated with prediction errors are essential in obtaining the best linear estimates.
8. Develop research questions appropriate for multiple regression analysis.
9. Test data sets using multiple regression by following the appropriate SPSS guidelines provided.

Up to this point, we have focused our attention on statistical analysis techniques that investigate the existence of differences between groups. In this chapter, we begin to redirect the focus of our attention to a second grouping of advanced/multivariate techniques—those that describe and test the existence of predictable relationships among a set of variables. Our discussion will include a brief review of simple linear regression, followed by an in-depth examination of multiple regression.

SECTION 7.1 PRACTICAL VIEW

Purpose

Regression analysis procedures have as their primary purpose the development of an equation that can be used for *predicting* values on some DV for all members of a population. (A secondary purpose is to use regression analysis as a means of *explaining* causal relationships among variables, the focus of Chapter 8.) The most basic application of regression analysis is the bivariate situation, to which the reader was undoubtedly exposed in an introductory statistics course. This is often referred to as *simple linear regression*, or just *simple regression*. Simple regression involves a single IV and a single DV. The goal of simple regression is to obtain a linear equation so that we can predict the value of the DV if we have the value of the IV. Simple regression capitalizes on the correlation between the DV and the IV in order to make specific

predictions about the DV (Sprinthall, 2007). The correlation tells us how much information about the DV is contained in the IV. If the correlation is perfect (i.e., $r = \pm 1.00$), the IV contains everything we need to know about the DV, and we will be able to perfectly predict one from the other. However, this is seldom, if ever, the case.

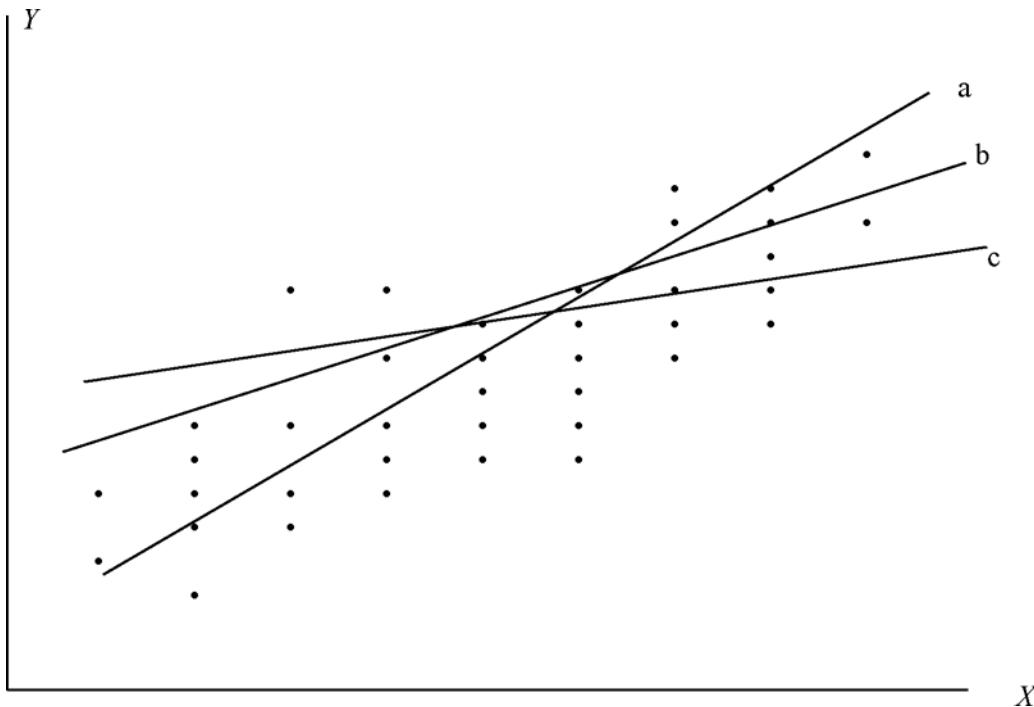
The idea behind simple regression is that we want to obtain the equation for the best-fitting line through a series of points. If we were to view a bivariate scatterplot for our fictitious IV (X) and DV (Y), we could then envision a line drawn through those points. Theoretically, an infinite number of lines could be drawn through those points (see Figure 7.1). However, only one of these lines would be the best-fitting line. Regression analysis is the means by which we determine the best-fitting line, called the **regression line**.

The regression line is the single straight line that lies closest to all points in a given scatterplot—this line is sometimes said to pass through the **centroid** of the scatterplot (Sprinthall, 2007). In order to make predictions, three important facts about the regression line must be known:

1. The extent to which points are scattered around the line
2. The slope of the regression line
3. The point at which the line crosses the Y -axis (Sprinthall, 2007)

These three facts are so important to regression that they serve as the basis for the calculation of the regression equation itself. The extent to which the points are scattered around the line is typically indicated by the degree of relationship between the IV (X) and the DV (Y). This relationship is measured by a correlation coefficient (e.g., the Pearson correlation, symbolized by r)—the stronger the relationship, the higher the degree of predictability between X and Y . (You will see in Section 7.3 just how important r is to the regression equation calculation.) The slope of the regression line can greatly affect prediction (Sprinthall, 2007). The degree of slope is determined by the amount of change in Y that accompanies a unit change (e.g., one point, one inch, one degree, etc.) in X . It is the slope that largely determines the predicted values of Y from known values for X . Finally, it is important to determine exactly where the regression line crosses the Y -axis (this value is known as the Y -intercept). Said another way, it is crucial to know what value is expected for Y when $X = 0$.

Figure 7.1. Bivariate Scatterplot Showing Several Possible Regression Lines.



The three facts we have just discussed actually define the regression line. The regression line is essentially an equation that expresses Y as a function of X (Tate, 1992). The basic equation for simple regression is

$$\hat{Y} = bX + a \quad (\text{Equation 7.1})$$

where \hat{Y} is the predicted value for the DV, X is the known raw score value on the IV, b is the slope of the regression line, and a is the Y -intercept. The significant role that both the slope and the Y -intercept play in the regression equation should now be apparent. Often, you will see the above equation presented in the following analogous, although more precise, form:

$$Y = B_0 + B_1 X_1 + e \quad (\text{Equation 7.2})$$

where Y is the value for the DV, X_1 is the raw score value on the IV, B_1 is the slope of the regression line, and B_0 is the Y -intercept. We have added one important term, e , in Equation 7.2, which is the symbol for the errors of prediction, also referred to as the residuals. As previously mentioned, unless we have a perfect correlation between the IV and the DV, the predicted values obtained by our regression equation will also be less than perfect—that is, there will be errors. The residuals constitute an important measure of those errors and are essentially calculated as the difference between the actual value and the predicted value for the DV (i.e., $e_i = y_i - \hat{y}_i$).

Let us return momentarily to the concept of the best-fitting line (see Figure 7.2). The reason that we obtain the best-fitting line as our regression equation is that we mathematically calculate the line with the smallest amount of total squared error. This is commonly referred to as the **least-squares solution** (Stevens, 2001; Tate, 1992) and actually provides us with values for the constants in the regression equation, β_1 and

β_0 (also known as the **regression coefficients**, **beta coefficients**, or **beta weights** [β]), that minimize the sum of squared residuals—that is, $\sum(y_i - \hat{y}_i)^2$ is minimized. In other words, the *total* amount of prediction error, both positive and negative, is as small as possible, giving us the best mathematically achievable line through the set of points in a scatterplot.

Multiple regression is merely an extension of simple linear regression involving more than one IV, or predictor variable. This technique is used to predict the value of a single DV from a weighted, linear combination of IVs (Harris, 1998). A multiple regression equation appears similar to its simple regression counterpart except that there are more coefficients, one for the *Y*-intercept and one for each of the IVs:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + e \quad (\text{Equation 7.3})$$

where there is a corresponding β coefficient for each IV (X_k) in the equation and the best linear combination of weights and raw score X values will again minimize the total squared error in our regression equation.

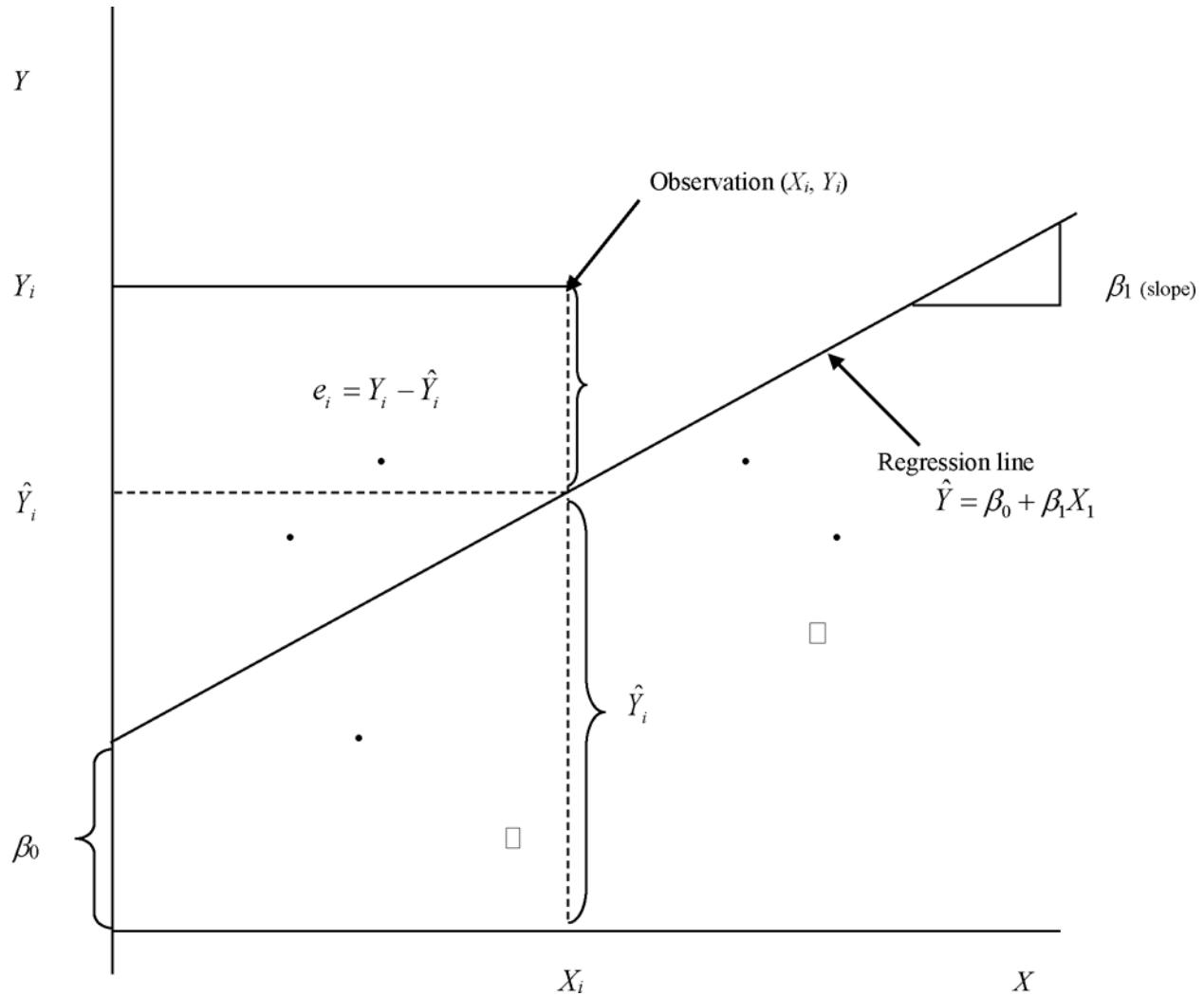
Let us consider a concrete example: Suppose we wanted to determine the extent to which we could predict female life expectancy from a set of predictor variables for a selected group of countries throughout the world. The predictor variables we have selected include percent urban population; gross domestic product per capita; birthrate per 1,000; number of hospital beds per 10,000; number of doctors per 10,000; number of radios per 100; and number of telephones per 100. In our analysis, we would be looking to obtain the regression coefficients for each IV that would provide us with the best linear combination of IVs—and their associated weights—in order to predict, as accurately as possible, female life expectancy. The regression equation predicting female life expectancy is as follows:

$$\text{Female life exp.} = \beta_0 + \beta_{\text{urban}} X_{\text{urban}} + \beta_{\text{GDP}} X_{\text{GDP}} + \beta_{\text{birthrate}} X_{\text{birthrate}} + \beta_{\text{beds}} X_{\text{beds}} + \beta_{\text{docs}} X_{\text{docs}} + \beta_{\text{radios}} X_{\text{radios}} + \beta_{\text{phones}} X_{\text{phones}} + e_i$$

We will return to this example in greater detail later in the chapter, but first, there are several important issues related to multiple regression that warrant our attention.

A first issue of interest is a set of measures unique to multiple regression. Another way of looking at the previously mentioned concept of the minimization of total error is to consider multiple regression as a means of seeking the linear combination of IVs that *maximally* correlate with the DV (Stevens, 2001). This maximized correlation is called the **multiple correlation** and is symbolized by R . The multiple correlation is essentially equivalent to the Pearson correlation between the actual, or observed, values and the predicted values on the DV (i.e., $R = r_{y_i \hat{y}_i}$). Analogous to our earlier interpretation of the Pearson correlation, the multiple correlation tells us how much information about a DV (e.g., female life expectancy) is contained in the combination of IVs (e.g., percent urban population, gross domestic product, birthrate, number of hospital beds, number of doctors, number of radios, and number of telephones). In multiple regression, there is a test of significance (F test) to determine whether the relationship between the set of IVs and the DV is large enough to be meaningful.

Figure 7.2. Graphical Representation of a Linear Regression Model and the Least-Squares Criterion.



You may recall from an earlier course in statistics the term **coefficient of determination**, or r^2 . For the Pearson r , this value was interpreted as the proportion of one variable in the pair that can be explained (or accounted for) by the other variable. In multiple regression, R^2 is also called the coefficient of determination and has a similar interpretation. The coefficient of determination is the proportion of DV variance that can be explained by the combination of the IVs (Levin & Fox, 2006; Sprinthall, 2007). In our example, an obtained value for R^2 would be interpreted as the proportion of variability in female life expectancy that could be accounted for by the combination of the seven predictor variables. If one multiplies this value by 100, R^2 becomes the percentage of explained variance (Sprinthall, 2007).

A second issue is one that has an associated word of caution, which we will address for a moment. The issue at hand is that of **multicollinearity**. Multicollinearity is a problem that arises when moderate to high intercorrelations exist among predictor variables (IVs) to be used in a regression analysis. (Recall from Chapter 1 that the opposite of multicollinearity is *orthogonality*, or complete independence among variables.) The underlying problem of multicollinearity is that if two variables are highly correlated, they essentially contain the same—or at least much of the same—information and are therefore measuring the same

thing (Sprinthall, 2007). Not only does one gain little by adding to regression analysis variables that are measuring the same thing, but multicollinearity can cause real problems for the analysis itself. Stevens (2001) points out three reasons why multicollinearity can be problematic for researchers:

1. Multicollinearity severely limits the size of R because the IVs are “going after” much of the same variability on the DV.
2. When trying to determine the importance of individual IVs, multicollinearity causes difficulty because individual effects are confounded due to the overlapping information.
3. Multicollinearity tends to increase the variances of the regression coefficients, which ultimately results in a more unstable prediction equation.

Multicollinearity should be addressed by the researcher prior to the execution of the regression analysis. The simplest method for diagnosing multicollinearity is to examine the correlation matrix for the predictor variables, looking for moderate to high intercorrelations. However, it is preferable to use one of two statistical methods to assess multicollinearity. First, tolerance statistics can be obtained for each IV.

Tolerance is a measure of collinearity among IVs, where possible values range from 0 to 1. A value for tolerance close to zero is an indication of multicollinearity. Typically, a value of 0.1 serves as the cutoff point—if the tolerance value for a given IV is less than 0.1, multicollinearity is a distinct problem (Norusis, 1998). A second method is to examine values for the **variance inflation factor (VIF)** for each predictor. The variance inflation factor for a given predictor “indicates whether there exists a strong linear association between it and all remaining predictors” (Stevens, 2001). The VIF is defined by the quantity $1/(1-R_i^2)$ and is obtainable on most computer programs. Although there is no steadfast rule of thumb, values of VIF that are greater than 10 are generally cause for concern (Stevens, 2001).

There are several methods for combating multicollinearity in a regression analysis. Two of the most straightforward methods are presented here. The simplest method is to delete the problematic variable from the analysis (Sprinthall, 2007). If the information in one variable is being captured by another, no real information is being lost by deleting one of them. A second approach is to combine the variables involved to create a single measure that addresses a single construct, thus deleting the repetition (Stevens, 2001). One might consider this approach for variables with intercorrelations of .80 or higher. Several other approaches to dealing with multicollinear relationships exist, but they are beyond the scope of this text. If interested, pursue the discussion in Stevens (2001).

A third issue that is of great importance in multiple regression is the method of specifying the regression model—in other words, determining or selecting a good set of predictor variables. Keeping in mind that the goal of any analysis should be to achieve a *parsimonious* solution, we want to select IVs that will give us an efficient regression equation without including everything under the sun. Initially, one of the most efficient methods of selecting a group of predictors is to rely on the researcher’s substantive knowledge (Stevens, 2001). Being familiar with and knowledgeable about your population, sample, and data will provide you with meaningful information about the relationships among variables and the likely predictive power of a set of predictors. Furthermore, for reasons we will discuss later, a recommended ratio of participants to IVs (i.e., n/k) of at least 15 to 1 will provide a reliable regression equation (Stevens, 2001). Keeping the number of predictor variables low tends to improve this ratio because most researchers do not have the luxury of increasing their sample size at will, which would be necessary if one were to continue to add predictors to the equation.

Once a set of predictors has been selected, there are several methods by which they may be incorporated into the regression analysis and subsequent equation. Tabachnick and Fidell (2007) identify three

such strategies: standard multiple regression, sequential multiple regression, and stepwise multiple regression. (Recall—and possibly revisit—the discussion of standard and sequential analyses as presented in Chapter 1.) It should be noted that decisions about model specification can and do affect the nature of the research questions being investigated. In **standard multiple regression**, all IVs are entered into the analysis simultaneously. The effect of each IV on the DV is assessed as if it had been entered into the equation after all other IVs had been entered. Each IV is then evaluated in terms of what it adds to the prediction of the DV, as specified by the regression equation (Tabachnick & Fidell, 2007).

In **sequential multiple regression**, sometimes referred to as **hierarchical multiple regression**, a researcher may want to examine the influence of several predictor IVs in a specific order. Using this approach, the researcher specifies the order in which variables are entered into the analysis. Substantive knowledge, as previously mentioned, may lead the researcher to believe that one variable may be more influential than others in the set of predictors, and that variable is entered into the analysis first. Subsequent variables are then added in order to determine the specific amount of variance they can account for, above and beyond what has been explained by any variables entered prior (Aron, Aron, & Coups, 2008). Individual effects are assessed at the point of entry of a given variable (Tabachnick & Fidell, 2007).

Finally, **stepwise multiple regression**, also sometimes referred to as **statistical multiple regression**, is often used in studies that are exploratory in nature (Aron, Aron, & Coups, 2008). The researcher may have a large set of predictors and may want to determine which specific IVs make meaningful contributions to the overall prediction. There are essentially three variations of stepwise multiple regression:

1. **Forward selection** — The bivariate correlations among all IVs and the DV are calculated. The IV that has the highest correlation with the DV is entered into the analysis first. It is assessed in terms of its contribution (in terms of R^2) to the DV. The next variable to be entered into the analysis is the IV that contributes most to the prediction of the DV, after partialing out the effects of the first variable. This effect is measured by the increase in R^2 (ΔR^2) due to the second variable. This process continues until, at some point, predictor variables stop making significant contributions to the prediction of the DV. It is important to remember that once a variable has been entered into the analysis, it remains there (Pedhazur, 1982; Stevens, 2001).
2. **Stepwise selection** — Stepwise selection is a variation of forward selection. It is an improvement over the previous method in that, at each step, tests are performed to determine the significance of each IV already in the equation as if it were to enter last. In other words, if a variable entered into the analysis is measuring much of the same construct as another, this reassessment may determine that the first variable to enter may no longer contribute anything to the overall analysis. In this procedure, that variable would then be dropped out of the analysis. Even though it was at one time a good predictor, in conjunction with others, it may no longer serve as a substantial contributor (Pedhazur, 1982).
3. **Backward deletion** — The initial step here is to compute an equation with all predictors included. Then, a significance test (a partial F test) is conducted for every predictor, as if each were entered last, in order to determine the level of contribution to overall prediction. The smallest partial F is compared to a preselected “ F to remove” value. If the value is less than the “ F to remove” value (not significant), that predictor is removed from the analysis and a new equation with the remaining variables is computed, followed by another test of the resulting smallest partial F . This process continues until only significant predictors remain in the equation (Stevens, 2001).

It is important to note that both sequential and stepwise approaches to regression contain a distinct advantage over standard multiple regression: One variable is added at a time, and each is continually

checked for significant improvement to prediction. However, the important difference between these two is that sequential regression orders and adds variables based on some *theory or plan by the researcher*, whereas in stepwise regression, those decisions are being made by a computer based *solely on statistical analyses* (Aron, Aron, & Coups, 2008). Sequential regression should be used in research based on theory or some previous knowledge. Stepwise regression should be used where exploration is the purpose of the analysis.

A fourth issue of consequence in multiple regression is that of model validation, sometimes called model ***cross-validation***. A regression equation is developed in order to predict DV values for individuals in a population, but remember that the equation itself was developed based only on a sample from that population. The multiple correlation, R , will be at its maximum value for the sample from which the equation was derived. If the predictive power drops off drastically when applied to an independent sample from the same population, the regression equation is of little use because it has little or no generalizability (Stevens, 2001). If the equation is not predicting well for other samples, it is not fulfilling its designed and intended purpose.

In order to obtain a reliable equation, substantial consideration must be given to the sample size (n) and the number of predictors (k). As mentioned earlier, a recommended ratio of these two factors is about 15 participants for every predictor (Stevens, 2001). This results in an equation that will cross-validate with relatively little loss in its ability to predict. Another recommendation for this ratio is identified by Tabachnick and Fidell (2007). The simplest rule of thumb they offer is that $n \geq 50 + 8k$, for testing multiple correlations, and $n \geq 104 + k$, for testing individual predictors. They suggest calculating n both ways and using the larger value.

Cross-validation can be accomplished in several ways. Ideally, one should wait a period of time, select an independent sample from the same population, and test the previously obtained regression equation (Tatsuoka, 1988). This is not always feasible, so an alternative would be to split the original sample into two subsamples. Then, one subsample can be used to develop the equation, while the other is used to cross-validate it (Stevens, 2001). This would be feasible only if one had a large enough sample, based on the criteria set forth above.

A final issue of importance in regression is the effect that outliers can have on a regression solution. Recall that regression is essentially a maximization procedure (i.e., we are trying to maximize the correlation between observed and predicted DV scores). Because of this fact, multiple regression can be very sensitive to extreme cases. One or two outliers have been shown to adversely affect the interpretation of regression analysis results (Stevens, 2001). It is therefore recommended that outliers be identified and dealt with appropriately prior to running the regression analysis. This is typically accomplished by initial screenings of boxplots, but is more precisely accomplished with the statistical procedure known as *Mahalanobis distance* (as described in Chapter 3).

Note that there does exist a multivariate version of multiple regression (i.e., multivariate multiple regression) does exist, but it is so similar in its approach and conduct that it will not be discussed in detail in this text. Basically, multivariate multiple regression involves the prediction of several DVs from a set of predictor IVs. This procedure is a variation of multiple regression in that the regression equations realized are those that would be obtained if each DV were regressed *separately* on the set of IVs. The actual correlations among DVs in the analysis are ignored (Stevens, 2001).

Sample Research Questions

Building on the example we began discussing in the previous section, we can now specify the research questions to be addressed by our multiple regression analysis. The methods by which the regression model is developed often dictate the type of research question(s) to be addressed. For instance, if we were entering all seven IVs from our data set into the model, the appropriate research questions would be as follows:

1. Which of the seven predictor variables (i.e., percent urban population, GDP, birthrate, number of hospital beds, number of doctors, number of radios, and number of telephones) are most influential in predicting female life expectancy? Are there any predictor variables that do not contribute significantly to the prediction model?
2. Does the obtained regression equation resulting from a set of seven predictor variables allow us to reliably predict female life expectancy?

However, if we were using a stepwise method of specifying the model, the revised questions would be as follows:

1. Which of the possible seven predictor variables (i.e., percent urban population, GDP, birthrate, number of hospital beds, number of doctors, number of radios, and number of telephones) are included in an equation for predicting female life expectancy?
2. Does the obtained regression equation resulting from a subset of the seven predictor variables allow us to reliably predict female life expectancy?

SECTION 7.2 ASSUMPTIONS AND LIMITATIONS

In multiple regression, there are actually two sets of assumptions—assumptions about the raw scale variables and assumptions about the residuals (Pedhazur, 1982). With respect to the raw scale variables, the following conditions are assumed:

1. The independent variables are fixed (i.e., the same values of the IVs would have to be used if the study were to be replicated).
2. The independent variables are measured without error.
3. The relationship between the independent variables and the dependent variable is linear (in other words, the regression of the DV on the combination of IVs is linear).

The remaining assumptions concern the residuals. Recall from Chapter 3 that *residuals*, or *prediction errors*, are the portions of scores not accounted for by the multivariate analyses. Meeting these assumptions is necessary in order to achieve the best linear estimations (Pedhazur, 1982).

4. The mean of the residuals for each observation on the dependent variable over many replications is zero.
5. Errors associated with any single observation on the dependent variable are independent of (i.e., not correlated with) errors associated with any other observation on the dependent variable.
6. The errors are not correlated with the independent variables.
7. The variance of the residuals across all values of the independent variables is constant (i.e., homoscedasticity of the variance of the residuals).
8. The errors are normally distributed.

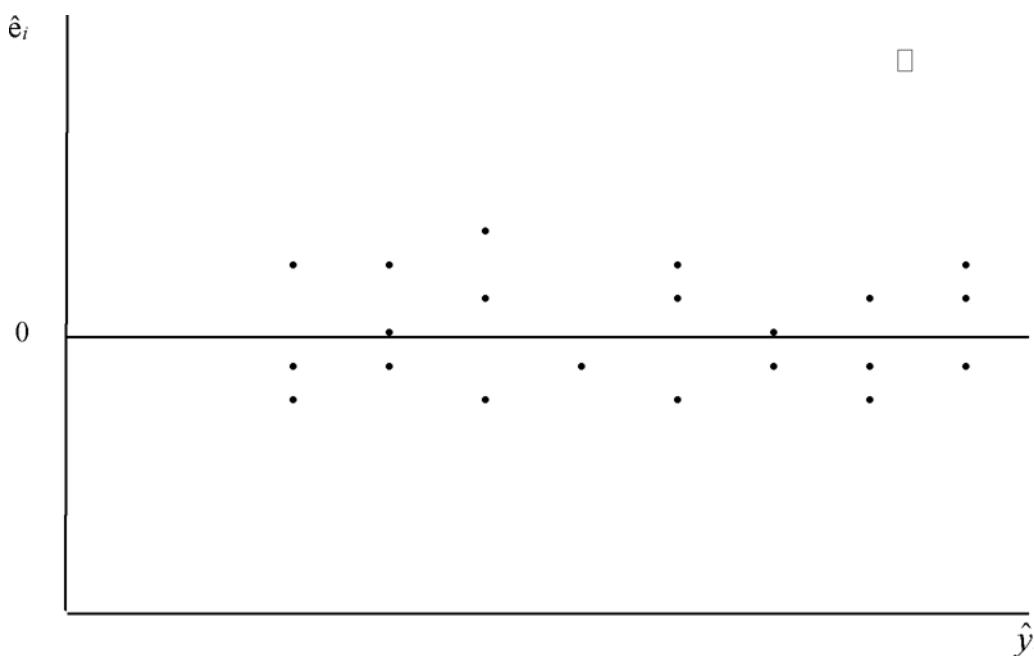
Assumptions 1, 2, and 4 are largely research design issues. We will focus our attention on Assumptions 3, 5, and 6—which address the issue of linearity—and Assumptions 7 and 8—which address homoscedasticity and normality, respectively.

Methods of Testing Assumptions

There are essentially two approaches to testing the assumptions in multiple regression (Tabachnick & Fidell, 2007). The first approach involves the routine pre-analysis data-screening procedures that have been discussed in several of the preceding chapters. As a reminder, linearity can be assessed through examination of the various bivariate scatterplots. Normality is evaluated in similar fashion, as well as through the assessment of the values for skewness, kurtosis, and Kolmogorov-Smirnov statistics. Finally, homoscedasticity is assessed by interpreting the results of Box's test.

The alternative approach to the routine procedure is to examine the residuals scatterplots. These scatterplots resemble bivariate scatterplots in that they are plots of values on the combination of two variables—in this case, these are the predicted values of the DV (\hat{y}) and the standardized residuals or prediction errors (\hat{e}_i). Examination of these residual scatterplots provides a test of *all three* of these crucial assumptions (Tabachnick & Fidell, 2007). If the assumptions of linearity, normality, and homoscedasticity are tenable, we would expect to see the points cluster along the horizontal line defined by $\hat{e}_i = 0$, in a somewhat rectangular pattern (see Figure 7.3).

Figure 7.3. Residuals Plot of Standardized Residuals (\hat{e}_i) Versus Predicted Values (\hat{y}_i) When Assumptions Are Met.



Any systematic, differential patterns or clusters of points are an indication of possible model violations (Stevens, 2001; Tabachnick & Fidell, 2007). Examples of residuals plots depicting violations of the three assumptions are shown in Figure 7.4. (It is important to note that the plots shown in this figure are idealized and have been constructed to show clear violations of assumptions. A word of caution—with real data, the patterns are seldom this obvious.) If the assumption of linearity is tenable, we would expect to see

a relatively straight-line relationship among the points in the plot. This typically appears as a rectangle (Tabachnick & Fidell, 2007), as depicted in Figure 7.3. However, as shown in Figure 7.4(a), the points obviously appear in a nonlinear pattern. In fact, this example is so extreme as to depict a clearly curvilinear pattern. This is an unmistakable violation of the assumption of linearity.

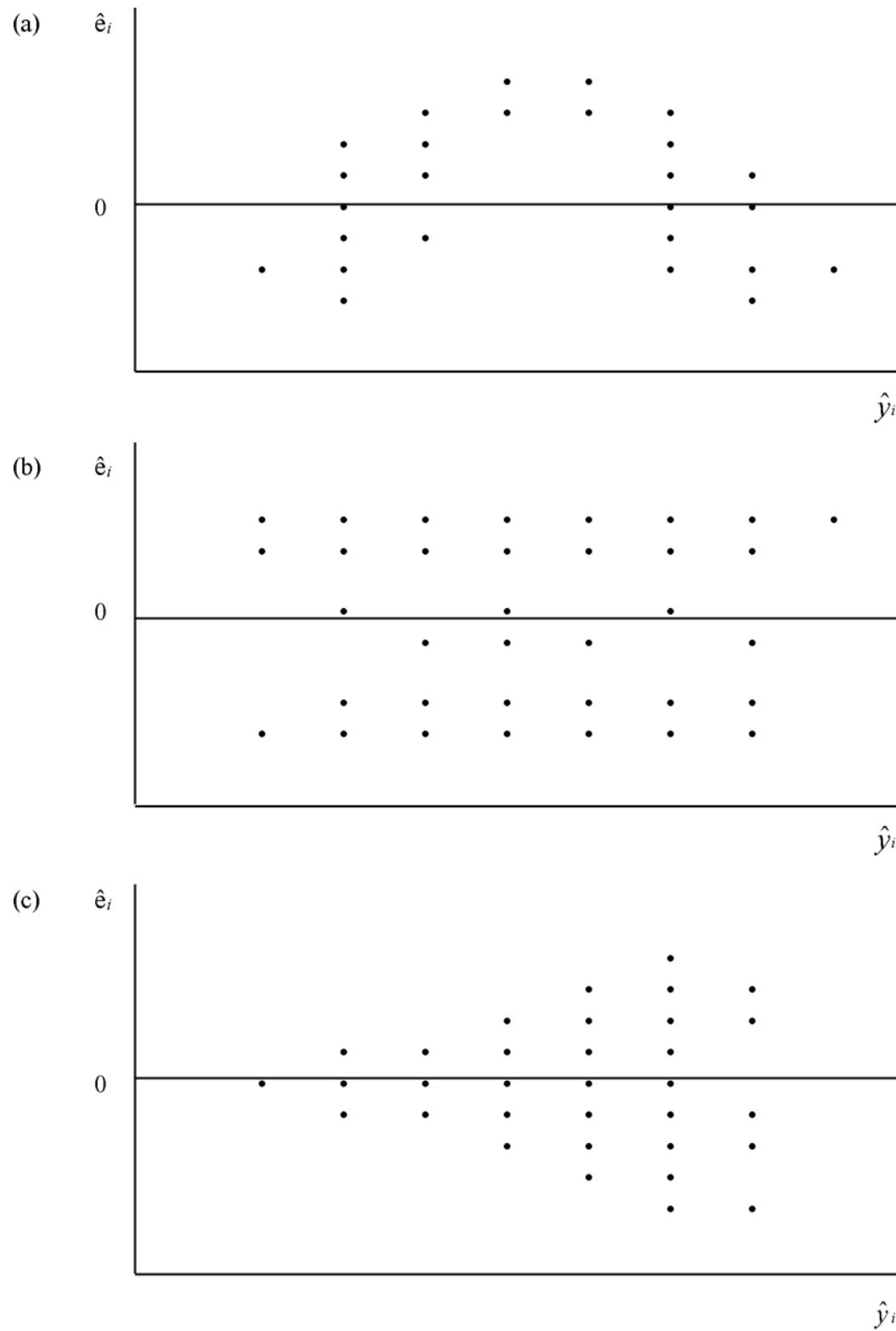
If the assumption of normality is defensible, we would expect to see an even distribution of points both above and below the line defined by $\hat{e}_i = 0$. In Figure 7.4(b), there appears to be a clustering of points the farther we move above and below that reference line, indicating a nonnormal (in this case, bimodal) distribution of residuals (Tate, 1992).

Finally, Figure 7.4(c) shows a violation of the assumption of homoscedasticity. If this assumption is tenable, we would expect to see the points dispersed evenly about the reference line—again, defined by $\hat{e}_i = 0$ —across all predicted values for the DV. In Figure 7.4(c), notice that the width is very narrow at small predicted values for the DV. However, the width increases rapidly as the predicted DV value increases. This is a clear indication of heteroscedasticity, or a lack of constant variance.

Residuals scatterplots may be examined *in place of* the routine pre-analysis data screening or *following* those procedures (Tabachnick & Fidell, 2007). If examination of the residuals scatterplots is conducted instead of the routine procedures—and if no violations are evident, no outliers exist, there are a sufficient number of cases, and there is no evidence of multicollinearity—then one is safe in interpreting that single regressions run on the computer. However, if the initial residuals scatterplots do not look “clean,” then further data screening using the routine procedures is warranted (Tabachnick & Fidell, 2007). In many cases, this may involve the transformation of one or more variables in order to meet the assumptions. If a curvilinear pattern appears, one possible remedy is to use a polynomial (i.e., nonlinear) model (Stevens, 2001), which is beyond the scope of this book.

In cases that involve moderate violations of linearity and homoscedasticity, one should be aware that these violations merely weaken the regression analysis, but they do not invalidate it (Tabachnick & Fidell, 2007). Furthermore, moderate violations of the normality assumption may often be ignored—especially with larger sample sizes—because there are no adverse effects on the analysis (Tate, 1992). It may still be possible to proceed with the analysis, depending on the subjective judgments of the researcher. Unfortunately, however, there are no rules to explicitly define that which constitutes a moderate violation. In reality, we would probably be justified in expecting some slight departures from the ideal situation, as depicted in Figure 7.3, due to sampling fluctuations (Tate, 1992).

Figure 7.4. Residuals Plots Showing Violations of (a) Linearity, (b) Normality, and (c) Homoscedasticity.



SECTION 7.3 PROCESS AND LOGIC

The Logic Behind Multiple Regression

You will recall from your previous exposure to simple regression that the statistical calculations basically involve the determination of the constants a and b . The slope of the line (i.e., b) is first calculated by multiplying the correlation coefficient between X and Y —recall that we discussed earlier in this chapter the important role played by the correlation between X and Y —by the standard deviation of Y and dividing that term by the standard deviation of X :

$$b = \frac{(r)(SD_Y)}{(SD_X)} \quad (\text{Equation 7.4})$$

The constant a (the Y -intercept) is then calculated in the following manner:

$$a = \bar{Y} - b\bar{X} \quad (\text{Equation 7.5})$$

There are analogous equations for the multivariate regression situation. They appear slightly more ominous and, therefore, will not be shown here. Recall from Equation 7.3 that in multiple regression there are *at least* two regression coefficients (specifically, the slope coefficients B_1 and B_2) that must be calculated. The calculations mirror Equation 7.4. The only substantial difference is that they incorporate a concept known as partial correlation. **Partial correlation** is a measure of the relationship between an IV and a DV, holding all other IVs constant. For instances, the calculated value for B_1 tells us how much of a change in Y can be expected for a given change in X_1 when the effects of X_2 are held constant (Sprinthall, 2007).

The other main calculation in multiple regression is the determination of the value for R^2 and its associated significance test. Recall that R^2 is a measure of variance accounted for in the DV by the predictors. One can think of this as being similar to analysis of variance in that we must partition the sum of squares variability. In regression analysis, we separate the total variability into variability due to regression (see Equation 7.6) and variability about the regression, also known as the sum of squares residual (see Equation 7.7).

$$SS_{reg} = \sum (\hat{y}_i - \bar{y})^2 \quad (\text{Equation 7.6})$$

$$SS_{res} = \sum (y_i - \hat{y})^2 \quad (\text{Equation 7.7})$$

The total sum of squares is simply the sum of these two terms and is symbolized by $\sum (y_i - \bar{y})^2$. The squared multiple correlation is then calculated by dividing the sum of squares due to regression (SS_{reg}) by the sum of squares total (SS_{tot}): □

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \quad (\text{Equation 7.8})$$



The standard F test from analysis of variance can be written making some simple algebraic substitutions:

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)} \quad (\text{Equation 7.9})$$

where k and $n - k - 1$ are the appropriate degrees of freedom for the numerator and denominator, respectively (Stevens, 2001). From this point, the significance of the obtained value for R^2 can be tested using the standard F test criteria, or by simply examining the associated p value from the computer printout. This then tells the researcher whether the set of IV predictor variables is accounting for, or explaining, a statistically significant amount of variance in the DV.

Interpretation of Results

Interpretation of multiple regression focuses on determining the adequacy of the regression model(s) that has been developed. Conducting multiple regression typically generates output that can be divided into three parts: model summary, ANOVA, and coefficients. Our discussion on how to interpret regression results will address these three parts. The first part of the regression output, *model summary*, displays three multiple correlation indices—multiple correlation (R), squared multiple correlation (R^2), and adjusted squared multiple correlation (R^2_{adj})—all of which indicate how well an IV or combination of IVs predicts the criterion variable (DV). The multiple correlation (R) is a Pearson correlation coefficient between the predicted and actual scores of the DV. The squared multiple correlation (R^2) represents the degree of variance accounted for by the IV or combination of IVs. Unfortunately, R and R^2 typically overestimate their corresponding population values—especially with small samples; thus R^2_{adj} is calculated to account for such bias. Change in R^2 (ΔR^2) is also calculated for each step and represents the change in variance that is accounted for by the set of predictors once a new variable has been added to the model. Change in R^2 is important because it is used to determine which variables significantly contribute to the model, or in the case of a stepping method, which variables are added or removed from the model. If a stepping method is used, the model summary will present these statistics for each model or step that is generated.

The ANOVA table presents the F test and corresponding level of significance for each step or model generated. This test examines the degree to which the relationship between the DV and IVs is linear. If the F test is significant, then the relationship is linear and, therefore, the model significantly predicts the DV.

The final part of the output is the coefficients table that reports the following: the unstandardized regression coefficient (B), the standardized regression coefficient (beta or β), t and p values, and three correlation indices. The ***unstandardized regression coefficient*** (B), also known as the ***partial regression coefficient***, represents the slope weight for each variable in the model and is used to create the regression equation. B weights also indicate how much the value of the DV changes when the IV increases by 1 and the other IVs remain the same. A positive B specifies a positive change in the DV when the IV increases, whereas a negative B indicates a negative change in the DV when the IV increases. Because it is difficult to interpret the relative importance of the predictors when the slope weights are not standardized, ***beta weights*** (β) or ***standardized regression coefficients*** are often utilized to create a prediction equation for the standardized variables. Beta weights are based upon z -scores with a mean of 0 and standard deviation of 1. The coefficients table also presents t and p values, which indicate the significance of the B weights, beta weights, and the subsequent part and partial correlation coefficients. Actually, three correlation coefficients are displayed in the coefficients table. The zero-order correlation represents the bivariate correlation between the IV and the DV. The partial correlation coefficient indicates the relationship between the IV and

DV after partialing out all other IVs. The part correlation, rarely used when interpreting the output, represents the correlation between the DV and the IVs after partialing only one of the IVs.

The final important statistic in the coefficient table is tolerance, which is a measure of multicollinearity among the IVs. Because the inclusion of IVs that are highly dependent upon each other can create an erroneous regression model, determining which variables account for a high degree of common variance in the DV is critical. Tolerance is reported for all the IVs included and excluded in the generated model. This statistic represents the proportion of variance in a particular IV that is not explained by its linear relationship with the other IVs. Tolerance ranges from 0 to 1, with 0 indicating multicollinearity. Typically, if tolerance of an IV is less than .1, the regression procedure should be repeated without the violating IV.

As one can see, there is a lot to interpret when conducting multiple regression. Because tolerance is an indicator of the appropriateness of IVs utilized in the regression, this statistic should be interpreted first. If some IVs violate the tolerance criteria, regression should be conducted again without the violating variables. If the value for tolerance is acceptable, one should proceed with interpreting the model summary, ANOVA table, and table of coefficients.

Let us now apply this process to our example using the data set *country-a.sav*. Because we will utilize the Forward stepping method, our research question is more exploratory in nature: Which IVs (% urban population [*urban*]; gross domestic product per capita [*gdp*]; birthrate per 1,000 [*birthrat*]; hospital beds per 10,000 [*hospbed*]; doctors per 10,000 [*docs*]; radios per 100 [*radio*]; and telephones per 100 [*phone*]) are predictors of female life expectancy (*lifeexpf*)? Data were first screened for missing data and outliers and then examined for test assumptions. Outliers were identified by calculating Mahalanobis distance in a preliminary **Regression** procedure (see Chapter 3, p. 52 for SPSS “How To”). **Explore** was then conducted on the newly generated Mahalanobis variable (*MAH_1*) to determine which cases exceeded the chi-square (χ^2) criteria (see Figure 7.5). Using a chi-square table, we found the critical value of chi-square at $p < .001$ with $df = 8$ to be 26.125. Two cases (67 and 72) exceeded this critical value and therefore were deleted from our present analysis by using **Select Cases, If** $MAH_1 \leq 26.125$. Linearity was then analyzed by creating a scatterplot matrix (see Figure 7.6). Scatterplots display nonlinearity for the following variables: *gdp*, *hospbed*, *docs*, *radio*, and *phone*. These variables were transformed by taking the natural log of each. Note that the data set already includes these transformations as *lngdp*, *lnbeds*, *lndocs*, *lnradio*, and *lnphone*. A scatterplot matrix (see Figure 7.7) with the transformed variables displays elliptical shapes that indicate linearity and normality. Univariate normality was also assessed by conducting **Explore**. Histograms and normality tests (see Figure 7.8) indicate some nonnormal distributions. However, the distributions are not extreme. Multivariate normality and homoscedasticity were examined through the generation of a residuals plot within another preliminary **Regression**. The residuals plot is somewhat scattered but again is not extreme (see Figure 7.9). Thus, multivariate normality and homoscedasticity will be assumed.

Regression was then conducted using the Forward method. The three major parts of the output—model summary, ANOVA table, and coefficients table—are presented in Figures 7.10 through 7.12, respectively. Tolerance among the IVs is adequate because coefficients for all IVs included and excluded are above .1 (see Figure 7.12). Because the Forward method was utilized, only some of the IVs were entered into the model. The model summary (see Figure 7.10) indicates that three of the seven IVs were entered into the model with **R squared change, Part and partial correlations**, and **Collinearity diagnostics** selected in the **Statistics** dialog box. For the first step, *lnphone* was entered as it accounted for the most unique variance in the DV ($R^2 = .803$). The variables of *birthrat* and *lndocs* were entered in the next two steps, respectively, creating a model that accounted for 86.9% of the variance in female life expectancy. The ANOVA table (see Figure 7.11) presents the *F* test for each

step/model. The final model significantly predicts the DV [$F(3, 101) = 223.85, p < .001$]. The table of coefficients (see Figures 7.12 and 7.13) is then utilized to create a prediction equation for the DV. The following equation is generated using the B weights.

$$\text{Female life expectancy} = 2.346X_{\ln\text{phone}} - .231X_{\text{birthrat}} + 2.190X_{\ln\text{docs}} + 67.691$$

If we utilize the beta weights, we develop the following equation for predicting the standardized DV.

$$Z_{\text{Female life expectancy}} = .409 Z_{\ln\text{phone}} - .272 Z_{\text{birthrat}} + .303 Z_{\ln\text{docs}}$$

Bivariate and partial correlation coefficients should also be noted in the coefficients table.

Figure 7.5. Outliers for Mahalanobis Distance.

Extreme Values			
		Case Number	Value
MAH_1	Highest	1	67
		2	72
		3	69
		4	108
		5	83
	Lowest	1	.68285
		2	40
		3	87
		4	18
		5	53

Figure 7.6. Scatterplot Matrix for Original IVs and DV.

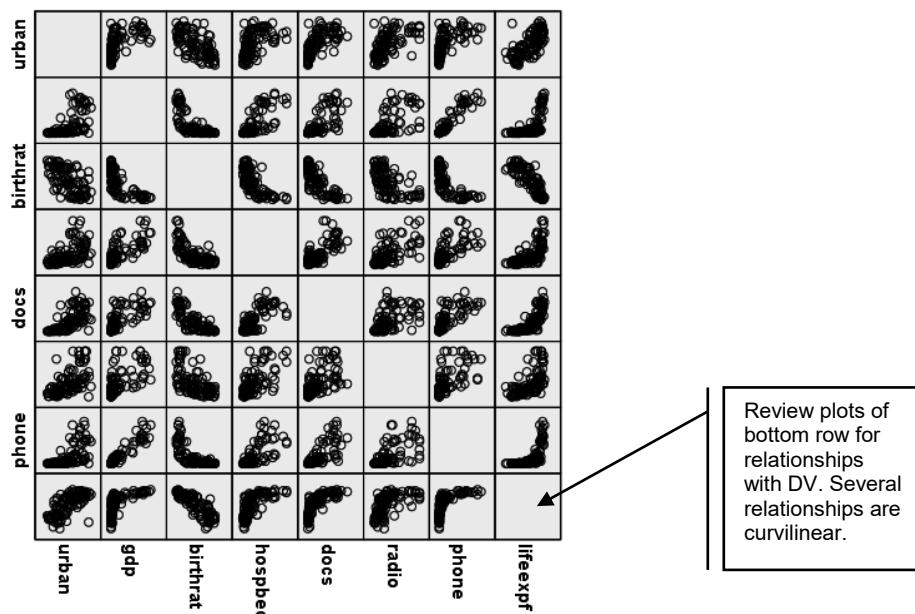


Figure 7.7. Scatterplot Matrix of Transformed IVs With DV.

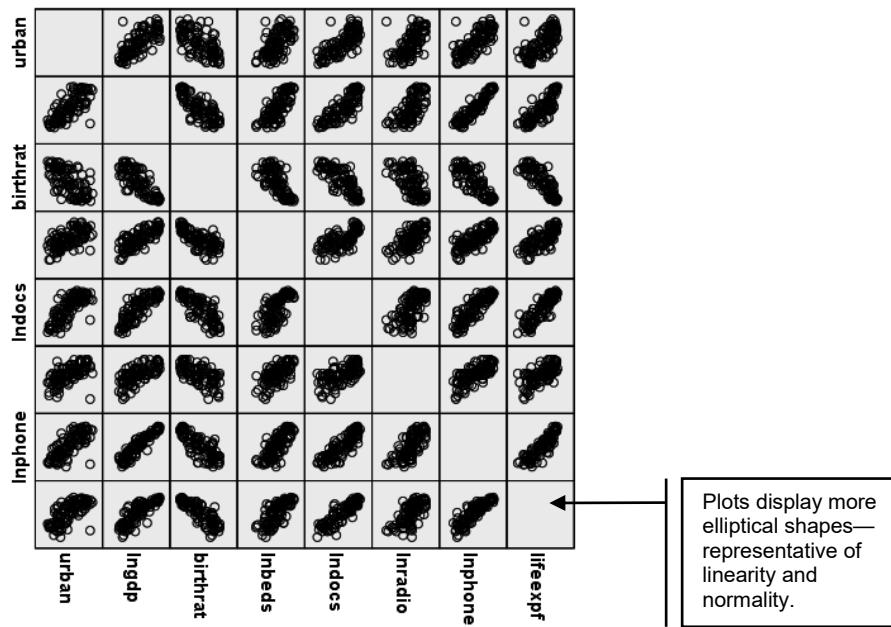


Figure 7.8. Test of Normality.

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
urban	.091	105	.033	.966	105	.009
Ingdp	.097	105	.016	.940	105	.000
birthrat	.135	105	.000	.924	105	.000
Inbeds	.062	105	.200*	.984	105	.234
Inradio	.100	105	.011	.967	105	.011
Inphone	.088	105	.042	.951	105	.001
lifeexpf	.126	105	.000	.933	105	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Indicates that most distributions are nonnormal.

Figure 7.9. Residuals Plot.

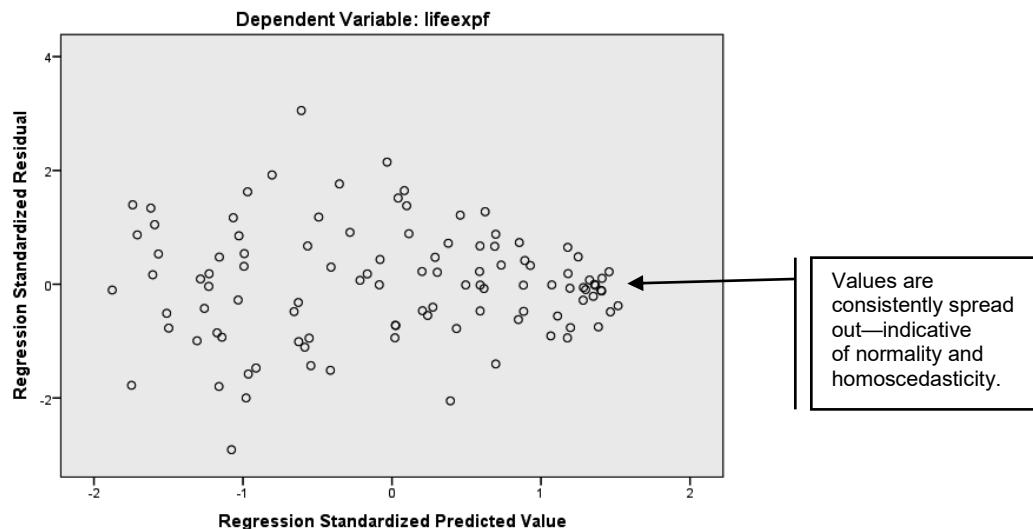


Figure 7.10. Model Summary Table for Female Life Expectancy.

Model	Model Summary ^d									
	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.896 ^a	.803	.801	5.001	.803	420.591	1	103	.000	
2	.922 ^b	.850	.847	4.382	.047	32.138	1	102	.000	
3	.932 ^c	.869	.865	4.117	.019	14.565	1	101	.000	

- a. Predictors: (Constant), Inphone
 b. Predictors: (Constant), Inphone, birthrat
 c. Predictors: (Constant), Inphone, birthrat, Indocs
 d. Dependent Variable: lifeexpf

Represents each step in the model building.

Figure 7.11. ANOVA Summary Table.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10519.284	1	10519.284	420.591	.000 ^b
	Residual	2576.106	103	25.011		
	Total	13095.390	104			
2	Regression	11136.489	2	5568.244	289.938	.000 ^c
	Residual	1958.902	102	19.205		
	Total	13095.390	104			
3	Regression	11383.379	3	3794.460	223.854	.000 ^d
	Residual	1712.012	101	16.951		
	Total	13095.390	104			

a. Dependent Variable: lifeexpf

b. Predictors: (Constant), Inphone

c. Predictors: (Constant), Inphone, birthrat

d. Predictors: (Constant), Inphone, birthrat, Indocs

Indicates that the final model is significant in predicting the DV.

Figure 7.12. Coefficients Table for Variables Included in the Model.

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error				Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	60.444	.581	103.964	.000	.896	.896	.896	1.000	1.000
	Inphone	5.137	.250							
2	(Constant)	72.372	2.165	33.431	.000	.896	.659	.339	.332	3.012
	Inphone	3.372	.381							
	birthrat	-.319	.056							
	Indocs	2.190	.574							
3	(Constant)	67.691	2.375	28.502	.000	.896	.463	.189	.212	4.711
	Inphone	2.346	.448							
	birthrat	-.231	.058							
	Indocs	2.190	.574							

a. Dependent Variable: lifeexpf

Coefficients used to develop regression equation.

Coefficients used to develop regression equation.

Tolerance statistics should be greater than .1.

Figure 7.13. Coefficients Table for Variables Excluded in the Model.

Model		Excluded Variables ^a					Collinearity Statistics			Tolerance statistics should be greater than .1.
		Beta ln	t	Sig.	Partial Correlation	Tolerance	VIF	Minimum Tolerance		
1	birthrat	-.377 ^b	-5.669	.000	-.489	.332	3.012	.332		
	Indocs	.431 ^b	5.517	.000	.479	.244	4.103	.244		
	urban	.057 ^b	.839	.403	.083	.418	2.390	.418		
	Ingdp	-.017 ^b	-.129	.897	-.013	.111	8.979	.111		
	Inbeds	.097 ^b	1.376	.172	.135	.383	2.610	.383		
	Inradio	.080 ^b	1.254	.213	.123	.463	2.162	.463		
2	Indocs	.303 ^c	3.816	.000	.355	.205	4.883	.205		
	urban	.045 ^c	.763	.447	.076	.418	2.393	.233		
	Ingdp	-.154 ^c	-1.318	.191	-.130	.107	9.357	.103		
	Inbeds	.007 ^c	.115	.909	.011	.358	2.793	.257		
	Inradio	.066 ^c	1.181	.240	.117	.462	2.166	.247		
3	urban	-.042 ^d	-.700	.486	-.070	.357	2.805	.175		
	Ingdp	-.159 ^d	-1.455	.149	-.144	.107	9.358	.088		
	Inbeds	.025 ^d	.419	.676	.042	.356	2.810	.173		
	Inradio	.068 ^d	1.298	.197	.129	.462	2.166	.173		

a. Dependent Variable: lifeexpf

b. Predictors in the Model: (Constant), Inphone

c. Predictors in the Model: (Constant), Inphone, birthrat

d. Predictors in the Model: (Constant), Inphone, birthrat, Indocs

Writing Up Results

The summary of multiple regression results should always include a description of how variables have been transformed or cases deleted. Typically, descriptive statistics (e.g., correlation matrix, means, and standard deviations for each variable) are presented in tables unless only a few variables are analyzed. Note that our example of a results summary will not include these descriptive statistics due to space limitations. The overall regression results are summarized in the narrative by identifying the variables in the model: R^2 , R^2_{adj} , F , and p values with degrees of freedom. If a step approach has been utilized, you may want to report each step (R^2 , R^2_{adj} , R^2 change, and level of significance for change) within a table. Finally, you may want to report the B weight, beta weight, bivariate correlation coefficients, and partial correlation coefficients of the predictors with the DV in a table. If you do not present these coefficients in a table, you may want to report the prediction equation, either standardized or unstandardized. The following results statement applies to the results presented in Figures 7.10 through 7.13.

Forward multiple regression was conducted to determine which independent variables (% urban population [*urban*]; gross domestic product per capita [*gdp*]; birthrate per 1,000 [*birthrat*]; hospital beds per 10,000 [*hospbed*]; doctors per 10,000 [*docs*]; radios per 100 [*radio*]; and telephones per 100 [*phone*]) were the predictors of female life expectancy. Data screening led to the elimination of one case. Evaluation of linearity led to the natural log transformation of *gdp*, *beds*, *docs*, *radios*, and *phones*. Regression results indicate an overall model of three predictors (phone, birthrate, and *docs*) that significantly predict female life expectancy [$R^2 = .869$, $R^2_{adj} = .865$, $F(3, 101) = 223.85$, $p < .001$]. This model accounted for 86.9% of variance in female life expectancy. A summary of the regression model is presented in Table 1. In addition, bivariate and partial correlation coefficients between each predictor and the dependent variable are presented in Table 2.

Table 1*Model Summary*

Step	R	R ²	R ² adj	ΔR ²	F _{chg}	p	df ₁	df ₂
1. Phones	.896	.803	.801	.803	420.59	< .001	1	103
2. Birthrate	.922	.850	.847	.047	32.14	< .001	1	102
3. Doctors	.932	.869	.865	.019	14.57	< .001	1	101

Table 2*Coefficients for Final Model*

	B	β	t	Bivariate r	Partial r
Phones per 100	2.346	.409	5.242*	.896	.463
Birthrate per 1,000	-.231	-.273	-4.006*	-.858	-.370
Doctors per 10,000	2.190	.303	3.816*	.884	.355

Note. * Indicates significance at $p < .001$.

SECTION 7.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example of the process of conducting multiple regression. This process includes the development of research questions and hypotheses, data-screening methods, test methods, interpretation of output, and presentation of results. The example utilizes the data set *country-a.sav* from the website that accompanies this book (see p. *xiii*).

Problem

In the previous example, we identified predictors of female life expectancy. For this example, we will utilize the same IVs but change the DV to male life expectancy. In addition, the Enter method will be used, such that all IVs will be entered into the model. The following research question was generated to address this scenario:

How accurately do the IVs (% urban population [*urban*]; gross domestic product per capita [*gdp*]; birthrate per 1,000 [*birthrat*]; hospital beds per 10,000 [*hospbed*]; doctors per 10,000 [*docs*]; radios per 100 [*radio*]; and telephones per 100 [*phone*]) predict male life expectancy (*lifeexpm*)?

Methods and SPSS “How To”

Data were screened to identify missing data and outliers and to evaluate the fulfillment of test assumptions. Outliers were identified by calculating Mahalanobis distance in a preliminary **Regression** procedure. **Explore** was then conducted on the newly generated Mahalanobis variable (*MAH_2*) to determine which cases exceeded the chi-square (χ^2) criteria (see Figure 7.14). Using a chi-square table, we found the critical value of chi-square at $p < .001$ with $df = 8$ to be 26.125. Cases 67, 72, and 69 exceeded this critical value and so were deleted from the analysis by using **SELECT CASES, IF MAH_2 ≤ 26.125**. Linearity was then analyzed by creating a scatterplot matrix (see Figure 7.15). Scatterplots display nonlinearity for the following variables: *gdp*, *hospbed*, *docs*, *radio*, and *phone*. These variables were transformed by taking the natural log of each. The data set already includes these transformations as *ln gdp*, *ln beds*,

lndocs, *lnradio*, and *lnphone*. A scatterplot matrix of the transformed variables indicates linearity and normality (see Figure 7.16). Univariate normality was also assessed by conducting **Explore**. Histograms and normality tests (see Figure 7.17) indicate some nonnormal distributions. However, the distributions are not too extreme. Multivariate normality and homoscedasticity were examined through the generation of a residuals plot within another preliminary **Regression**. The residuals plot is somewhat scattered but again is not extreme (see Figure 7.18). Thus, multivariate normality and homoscedasticity will be assumed.

Figure 7.14. Outliers for Mahalanobis Distance.

Extreme Values			
		Case Number	Value
MAH_2	Highest	1	67 50.81753
		2	72 27.23509
		3	69 26.69299
		4	108 25.67930
		5	83 21.00443
	Lowest	1	21 1.19032
		2	40 1.29138
		3	87 1.59614
		4	53 1.73529
		5	25 1.81816

Outliers exceed χ^2 critical value.

Figure 7.15. Scatterplot Matrix of Original IVs with DV.

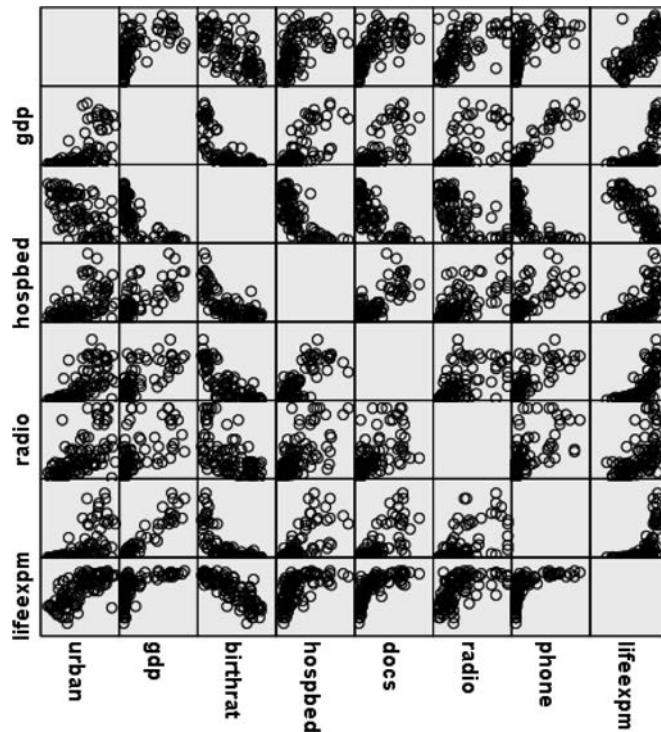


Figure 7.16. Scatterplot Matrix of Transformed IVs with DV.

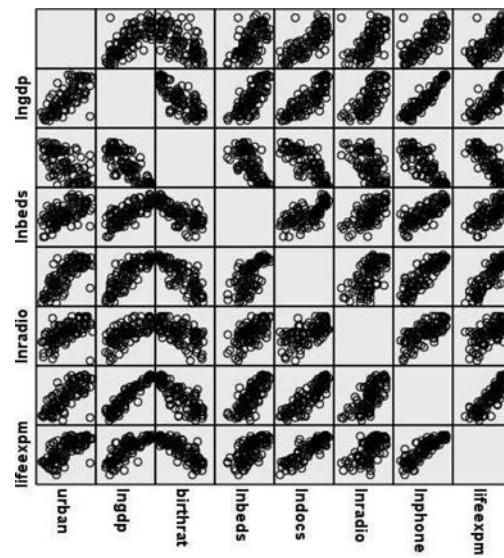


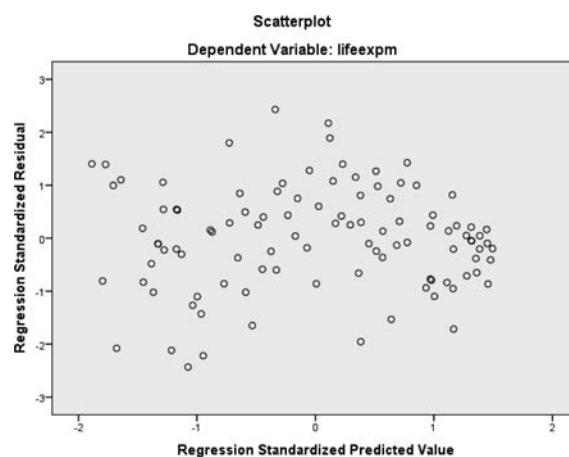
Figure 7.17. Tests of Normality.

	Tests of Normality					
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
urban	.091	104	.034	.966	104	.010
lnradio	.097	104	.017	.941	104	.000
lnbeds	.136	104	.000	.924	104	.000
lnspcs	.062	104	.200*	.984	104	.240
lntrab	.134	104	.000	.931	104	.000
lntrab	.102	104	.009	.968	104	.013
lntrab	.088	104	.044	.953	104	.001
lifeexpm	.136	104	.000	.934	104	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure 7.18. Residuals Plot.



Multiple **Regression** was then conducted using the Enter method. To conduct regression, select the following menus:

Analyze
Regression
Linear

Linear Regression dialog box (see Figure 7.19)

Once in this dialog box, identify the DV (*lifeexpm*) and move it to the **Dependent** box. Identify each of the IVs and move each to the **Independent(s)** box. Next, select the appropriate regression method. SPSS provides five different methods:

Enter—Enters all IVs, one at a time, into the model regardless of significant contribution.

Stepwise—Combines Forward and Backward methods. Two criteria are utilized—one for entering and one for removing. Basically, at each step, an IV is entered that meets the criteria for entering. Then the model is analyzed to determine if any IVs should be removed. If an entered variable meets the removal criteria, it is removed. The process continues until no more IVs meet the enter or removal criteria. Although Stepwise is quite common, it has recently come under great criticism for not creating a model of the best combination of predictors (Thompson, 1998).

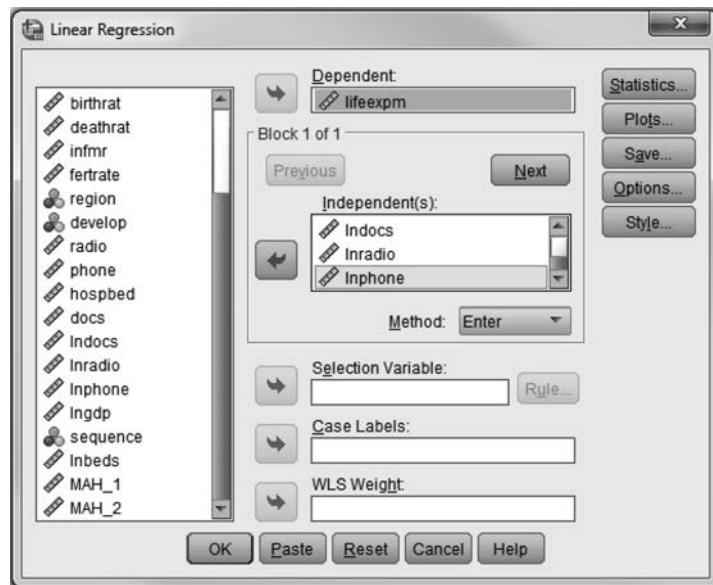
Remove—This method first utilizes the **Enter** method, after which a specified variable(s) is removed from the model and **Enter** is conducted again.

Backward—Enters all IVs one at a time and then removes them one at a time based upon a level of significance for removal (default is $p \geq .10$). The process ends when no more variables meet the removal requirement.

Forward—Only enters IVs that significantly contribute to the model (i.e., account for a significant amount of unique variance in the DV). Variables are entered one variable at a time. When no more variables account for a significant amount of variance, the process ends.

For this example, we selected **Enter**. Once the method is selected, click **Statistics**.

Figure 7.19. Linear Regression Dialog Box.



Linear Regression: Statistics dialog box (see Figure 7.20)

This box provides several statistical options. Under **Regression Coefficients**, three options are provided:

Estimates—The default. This option produces B and beta weights with associated standard error, t , and p values.

Confidence intervals—Calculates confidence intervals for B weights at 95%.

Covariance matrix—Creates a covariance-variance-correlation matrix that can help to assess collinearity. The matrix is organized with covariances below the diagonal, variances on the diagonal, and correlations above the diagonal.

For our example, we used the default of **Estimates**. Just to the right of **Regression Coefficients** is a list of more statistical options. These are described below.

Model fit—Produces Multiple R , R^2 , an ANOVA table, and corresponding F and p values.

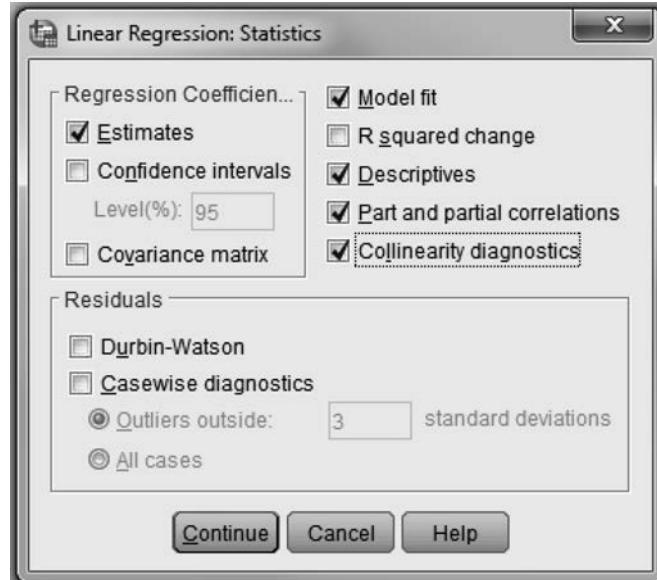
R squared change—Reports the change in R^2 as a new variable is entered into the model. This is appropriate when using a stepping method.

Descriptives—Calculates variable means, standard deviations, and correlation matrix.

Part and partial correlations—Calculates part and partial correlation coefficients.

Collinearity diagnostics—Calculates tolerance of each IV.

Figure 7.20. Linear Regression: Statistics Dialog Box.

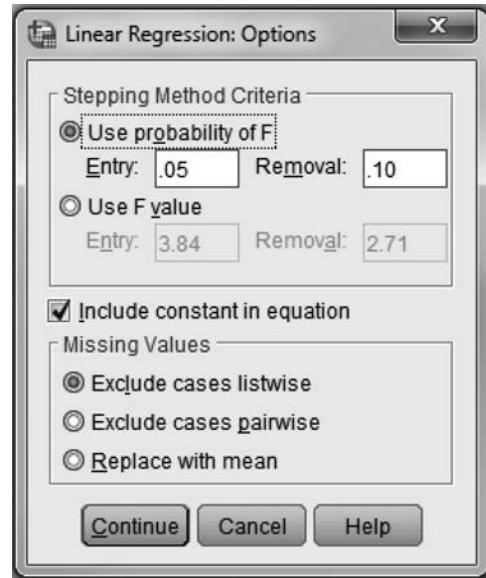


For this example, the following were checked: **Model fit**, **Descriptives**, **Part and partial correlations**, and **Collinearity diagnostics**. **R squared change** was not selected because a stepping method was not utilized. The last set of statistical options is under **Residuals** and is not utilized a great deal. Click **Continue**. Back in the **Regression** box, click **Options**.

Linear Regression: Options dialog box (see Figure 7.21)

This box provides criteria options for entering or removing variables from the model. Two criteria options are provided: **Use probability of F** or **Use F value**. The default method is **Use probability of F**, with .05 criteria for entry and .10 for removal. Because this example utilizes the **Enter** method, selection of stepping criteria is unnecessary. Another option in this box is removing the constant from the regression equation. Including the constant is the default. Click **Continue**. Without making any changes, click **OK** to exit this box. Although this completes the steps for our example problem, two dialog boxes provide additional options that will be described for future use. (Actually, we have already used both in evaluating test assumptions.) We will look first at **Plots**.

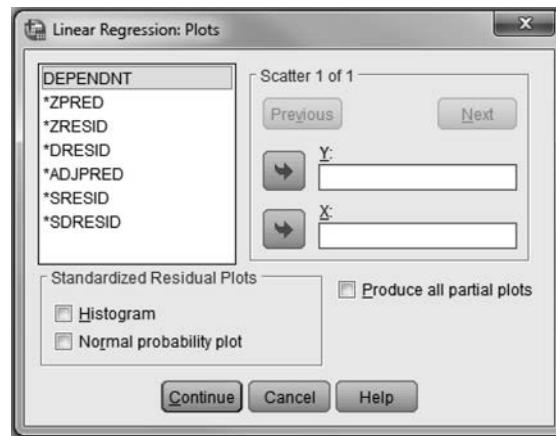
Figure 7.21. Linear Regression: Options Dialog Box.



Linear Regression: Plots dialog box (see Figure 7.22)

The **Plots** function provides several options for graphically analyzing the residuals or, in other words, for evaluating how well the generated model predicts the DV. Scatterplots can be created for any combination of the following: the DV, standardized predicted values (ZPRED), standardized residuals (ZRESID), deleted residuals (DRESID), adjusted predicted values (ADJPRED), studentized residuals (SRESID), or studentized deleted residuals (SDRESID). Earlier, we plotted the standardized predicted values (ZPRED) against the standardized residuals (ZRESID) to test linearity and homoscedasticity. Histograms of standardized residuals and normal probability plots comparing the distribution of the standardized residuals to a normal distribution can also be generated. The final option is **Produce all partial plots**. Checking this option will create scatterplots of residuals for each IV with the residuals of the DV when both variables are separately regressed on the remaining IVs. A minimum of two IVs must be in the equation to generate a partial plot. For our example, no plots were selected. We will now move to **Save**.

Figure 7.22. Linear Regression: Plots Dialog Box.



Linear Regression: Save dialog box (see Figure 7.23)

Several types of residuals can be saved in the data file with new variable names. Saving these residuals is necessary if you seek to create plots with these residuals. Types of residuals include **Unstandardized** and **Standardized**. Three other types are provided and described as follows:

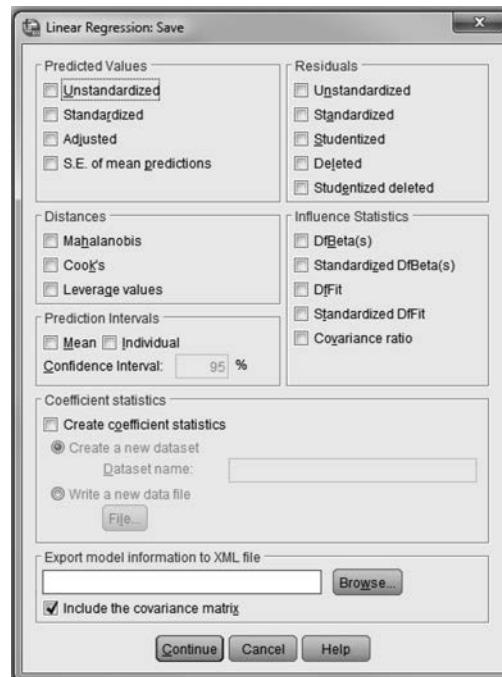
Studentized—Residuals standardized on a case-by-case basis, depending on how far the case is from the mean.

Deleted—Residual for a case if the case was deleted from the analysis.

Studentized deleted—Combines the concepts of studentized and deleted residuals.

You can also save predicted values and distances, such as Mahalanobis distance that was utilized to screen for outliers. For our current example, residuals were not saved. Click **Continue**, then **OK**.

Figure 7.23. Linear Regression: Save Dialog Box.



Output and Interpretation of Results

Figures 7.24 through 7.26 present the three primary parts of regression output: model summary, ANOVA summary table, and coefficients table. Review of the tolerance statistics presented in the coefficients table (see Figure 7.26) indicates that all but one of the IVs were tolerated in the model. The model summary (see Figure 7.24) and the ANOVA summary (see Figure 7.25) indicate that the overall model of the seven IVs significantly predicts male life expectancy [$R^2 = .845$, $R^2_{\text{adj}} = .834$, $F(7, 96) = 74.69$, $p < .001$]. However, a review of the beta weights in Figure 7.26 specifies that only three variables, *birthrat* $\beta = -.241$, $t(96) = -3.02$, $p = .003$; *Indocs* $\beta = .412$, $t(96) = 4.26$, $p < .001$; and *lnphone* $\beta = .548$, $t(96) = 3.88$, $p < .001$, significantly contributed to the model. Note that although the same three variables created the model for predicting female life expectancy, despite a different method being utilized, the significance of the model predicting male life expectancy is much lower because all seven variables were entered into the model.

Figure 7.24. Model Summary Predicting Male Life Expectancy.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.919 ^a	.845	.834	3.987

a. Predictors: (Constant), *lnphone*, *Inradio*, *urban*, *Inbeds*, *birthrat*, *Indocs*, *Ingdp*

Indicates the amount of variance in the DV that is accounted for by the model.

Figure 7.25. ANOVA Summary Table.

ANOVA ^a					
Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8311.675	7	1187.382	74.690
	Residual	1526.162	96	15.898	
	Total	9837.837	103		

a. Dependent Variable: *lifeexpr*

b. Predictors: (Constant), *lnphone*, *Inradio*, *urban*, *Inbeds*, *birthrat*, *Indocs*, *Ingdp*

F ratio and level of significance indicate the degree to which the model predicts the DV.

Figure 7.26. Coefficients Table.

Model	Coefficients ^a										
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics		
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF	
1	(Constant)	71.271	6.205	11.486	.000	.701	-.025	-.010	.342	2.927	
	urban	-.007	.027	-.017	.808	.147	.816	-.148	.101	9.898	
	Ingdp	-1.182	.808	-.185	-1.462	.003	-.829	-.295	.253	3.849	
	birthrat	-.179	.059	-.241	-3.021	.681	.086	-.121	.341	2.929	
	Inbeds	-.627	.738	-.058	-.850	.398	.399	.171	.173	5.791	
	Indocs	2.586	.607	.412	4.258	.000	.004	.448	.448	2.232	
	Inradio	.065	.628	.006	.104	.918	.639	.011	.004		
	Inphone	2.745	.707	.548	3.881	.000	.874	.368	.156	.081	

a. Dependent Variable: *lifeexpr*

Tolerance statistics exceed .1 for all but one of the variables.

Presentation of Results

Standard multiple regression was conducted to determine the accuracy of the independent variables (% urban population [*urban*]; gross domestic product per capita [*gdp*]; birthrate per 1,000 [*birthrat*]; hospital beds per 10,000 [*hospbed*]; doctors per 10,000 [*docs*]; radios per 100 [*radio*]; and telephones per 100 [*phone*]) predicting male life expectancy. Data screening led to the elimination of three cases. Evaluation of linearity led to the natural log transformation of *gdp*, *hospbed*, *docs*, *radio*, and *phone*. Regression results indicate that the overall model significantly predicts male life expectancy [$R^2 = .845$, $R^2_{\text{adj}} = .834$, $F(7, 96) = 74.69$, $p < .001$]. This model accounts for 84.5% of variance in male life expectancy. A summary of regression coefficients is presented in Table 3 and indicates that only three (birthrate, doctors, and phones) of the seven variables significantly contributed to the model.

Table 3
Coefficients for Model Variables

	<i>B</i>	β	<i>t</i>	<i>p</i>	Bivariate <i>r</i>	Partial <i>r</i>
Urban	-.007	-.017	-.243	.808	.701	-.025
GDP	-1.182	-.185	-1.462	.147	.816	-.148
Birthrate	-.179	-.241	-3.021	.003	-.829	-.295
Beds	-.627	-.058	-.850	.398	.681	-.086
Doctors	2.586	.412	4.258	<.001	.883	.399
Radios	.065	.006	.104	.918	.639	.011
Phones	2.745	.548	3.881	<.001	.874	.368

SUMMARY

The purpose of multiple regression is to model or group variables that best predict a criterion variable (DV). The procedure examines the significance of each IV to predict the DV, as well as the significance of the entire model to predict the DV. A variety of methods (enter, forward, backward, remove, and step-wise) can be used to develop and test different models. Prior to conducting the regression, data should be screened for missing data and outliers, as well as evaluated for test assumptions—linearity, normality, and homoscedasticity. Regression output typically includes three parts: model summary, ANOVA summary table, and coefficients table. The model summary table displays several multiple correlation indices—multiple correlation (*R*), squared multiple correlation (R^2), adjusted squared multiple correlation (R^2_{adj}), and change in R^2 (ΔR^2)—all of which indicate how well an IV or combination of IVs predicts the criterion variable (DV). The ANOVA summary table presents the *F* test and corresponding level of significance for each step or model generated. This test examines the degree to which the relationship between the IVs and DV is linear. The coefficients table reports the following: unstandardized regression coefficient (*B*), the standardized regression coefficient (beta or β), *t* and *p* values, three correlation indices (bivariate *r*, partial *r*, and part *r*), and tolerance coefficient. When interpreting the output, tolerance should be examined first because this measures the degree to which IVs account for unique variance in the DV. If tolerance for an IV is less than .1, the regression analysis should be conducted again without the violating IV. If tolerance is acceptable, proceed with interpreting the model summary, ANOVA summary table, and table of coefficients. Figure 7.27 provides a checklist for conducting multiple regression.

KEYWORDS

- backward deletion
- beta coefficients
- beta weights (β)
- centroid
- coefficient of determination (r^2)
- cross-validation
- forward selection
- hierarchical multiple regression
- least-squares solution
- multicollinearity
- multiple correlation (R)
- partial correlation
- partial regression coefficient
- regression coefficients
- regression line
- sequential multiple regression
- simple linear regression
- standard multiple regression
- standardized regression coefficient
- statistical multiple regression
- stepwise multiple regression
- stepwise selection
- tolerance
- unstandardized regression coefficient
- variance inflation factor (VIF)

Figure 7.27. Checklist for Conducting Multiple Regression.

I. Screen Data

- Missing Data?
 - Run preliminary Regression to calculate Mahalanobis distance.
 - Analyze... Regression... Linear.**
 - Identify a variable that serves as a case number and move to **Dependent Variable** box.
 - Identify all appropriate quantitative variables and move to **Independent(s)** box.
 - Save.**
 - Check **Mahalanobis** under **Distances**.
 - Continue, OK.**
 - Determine chi-square (χ^2) critical value at $p < .001$.
 - Conduct **Explore** to test outliers for Mahalanobis chi-square (χ^2).
 - Analyze... Descriptive statistics... Explore.**
 - Move **mah_1** to **Dependent Variable** box.
 - Leave **Factor** box empty.
 - Statistics.**
 - Check **Outliers**.
 - Continue, OK.**
 - Delete outliers for participants when χ^2 exceeds critical χ^2 at $p < .001$.
- Linearity, Normality, Homoscedasticity?
 - Create Scatterplot Matrix of all IVs and DV.
 - Scatterplot shapes are not close to elliptical shapes → reevaluate univariate normality and consider transformations.
 - Run Normality Plots with Tests within **Explore**.
 - Run preliminary Regression to create residual plot.
 - Analyze... Regression... Linear.**
 - Move DV to **Dependent Variable** box.
 - Move IVs to **Independent(s) Variable** box.
 - Plot.**
 - Select **ZRESID** for *y*-axis.
 - Select **ZPRED** for *x*-axis.
 - Continue, OK.**
 - If residuals are clustered at the top, bottom, left, or right area in plot → reevaluate univariate normality and consider transformations.

II. Conduct Multiple Regression

- Run Regression using **Linear Regression**.
 - Analyze... Regression... Linear.**
 - Move DV to **Dependent Variable** box.
 - Move IVs to **Independent(s)** box.
 - Select appropriate method.
 - Statistics.**
 - Check **Estimates, Model fit, R squared change** (only if a stepping method is utilized), **Descriptives, Part and partial correlations**, and **Collinearity diagnostics**.
 - Continue.**
 - Options.**
 - Select appropriate criteria.
 - Continue, OK.**
- Interpret tolerance.
- If tolerance for each IV is greater than .1, interpret model summary, ANOVA summary table, and coefficients table.

III. Summarize Results

- Describe any data elimination or transformation.
- Present descriptive statistics in tables (correlation matrix, means, and standard deviations).
- Narrate the significance of the overall regression (R^2 , R^2_{adj} , F and p values with degrees of freedom).
- If stepping method was used, summarize steps in a table (R^2 , R^2_{adj} , R^2 change, and level of significance for change).
- Create a table that reports the B weights, β weights, bivariate r , and partial r for each IV in the model.
- Draw conclusions.

Exercises for Chapter 7

1. The following output was generated from conducting a forward multiple regression to identify which IVs (*urban*, *birthrat*, *lnphone*, and *lnradio*) predict *lngdp*. The data analyzed were from the SPSS *country-a.sav* data file.

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	lnphone		Forward (Criterion: Probability-of- F-to-enter <= . 050)
2	birthrat		Forward (Criterion: Probability-of- F-to-enter <= . 050)

a. Dependent Variable: *lngdp*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.941 ^a	.886	.885	.51798	.886	862.968	1	111	.000
2	.943 ^b	.890	.888	.51091	.004	4.095	1	110	.045

a. Predictors: (Constant), *lnphone*

b. Predictors: (Constant), *lnphone*, *birthrat*

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error				Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	6.389	.058	110.662	.000					
	<i>lnphone</i>	.736	.025	.941	29.376	.000	.941	.941	1.000	1.000
2	(Constant)	6.878	.248	27.744	.000					
	<i>lnphone</i>	.663	.044	.849	15.238	.000	.941	.824	.482	.322
	<i>birthrat</i>	-.013	.006	-.113	-2.024	.045	-.811	-.189	-.064	.322

a. Dependent Variable: *lngdp*

Excluded Variables^a

Model	Beta ln	t	Sig.	Partial Correlation	Collinearity Statistics		
					Tolerance	VIF	Minimum Tolerance
1	urban	.095 ^b	1.901	.060	.178	.404	.404
	birthrat	-.113 ^b	-2.024	.045	-.189	.322	.322
	Inradio	.026 ^b	.557	.579	.053	.461	.461
2	urban	.091 ^c	1.848	.067	.174	.403	.225
	Inradio	.021 ^c	.455	.650	.044	.459	.243

a. Dependent Variable: *lngdp*b. Predictors in the Model: (Constant), *Inphone*c. Predictors in the Model: (Constant), *Inphone*, *birthrat***ANOVA^a**

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	231.539	1	231.539	862.968
	Residual	29.782	111	.268	^b
	Total	261.321	112		
2	Regression	232.608	2	116.304	445.561
	Residual	28.713	110	.261	^c
	Total	261.321	112		

a. Dependent Variable: *lngdp*b. Predictors: (Constant), *Inphone*c. Predictors: (Constant), *Inphone*, *birthrat*

- Evaluate the tolerance statistics. Is multicollinearity a problem?
- What variables create the model to predict *lngdp*? What statistics support your response?
- Is the model significant in predicting *lngdp*? Explain.
- What percentage of variance in *lngdp* is explained by the model?
- Write the regression equation for *lngdp*.

2. This question utilizes the data sets *profile-a.sav* and *profile-b.sav*, which can be downloaded from this website:

www.routledge.com/9781138289734

You are interested in examining whether the variables shown here in brackets [years of age (*age*), hours worked per week (*hrs1*), years of education (*educ*), years of education for mother (*maeduc*), and years of education for father (*paeduc*)] are predictors of individual income (*rincmdol*). Complete the following steps to conduct this analysis:

- a. Using *profile-a.sav*, conduct a preliminary regression to calculate Mahalanobis distance. Identify the critical value for chi-square. Conduct **Explore** to identify outliers. Which cases should be removed from further analysis?

For all subsequent analyses, use *profile-b.sav*. Make sure that only cases where $MAH_I \leq 22.458$ are selected.

- b. Create a scatterplot matrix. Can you assume linearity and normality?
- c. Conduct a preliminary regression to create a residual plot. Can you assume normality and homoscedasticity?
- d. Conduct multiple regression using the Enter method. Evaluate the tolerance statistics. Is multicollinearity a problem?
- e. Does the model significantly predict *rincmdol*? Explain.
- f. Which variables significantly predict *rincmdol*? Which variable is the best predictor of the DV?
- g. What percentage of variance in *rincmdol* is explained by the model?
- h. Write the regression equation for the standardized variables.
- i. Explain why the variables of mother's and father's education are not significant predictors of *rincmdol*.

CHAPTER 8

PATH ANALYSIS

STUDENT LEARNING OBJECTIVES

After studying Chapter 8, students will be able to:

1. Explain how path analysis is related to multiple regression.
2. Explain what is depicted in a path diagram.
3. Differentiate between direct, indirect, and spurious effects in a path model.
4. Describe the difference between exogenous and endogenous variables.
5. Summarize the various components contained in a structural equation.
6. Explain the process of path tracing or path decomposition.
7. Describe how model fit is assessed in path analysis.
8. Develop research questions appropriate for path analysis.
9. Use a data set to develop a path model and test its fit.

In the previous chapter, we discussed in detail one of the main purposes of multiple regression—*prediction*. In this chapter, we present a discussion of another use of multiple regression—providing *explanations* of possible causal relationships among a set of variables. Path analysis is actually one of two techniques classified under the broad heading of causal modeling. Following a brief introduction to causal modeling, and the distinctions between the two major types of causal modeling, we present a detailed discussion of path analysis, focusing on appropriate uses and proper interpretations of the technique.

SECTION 8.1 PRACTICAL VIEW

Purpose

Regression can be used to establish the possibility of cause-and-effect relationships among a set of variables (Sprinthall, 2007). Using regression analysis in this manner constitutes a specific set of statistical analysis techniques known as causal modeling. ***Causal modeling*** techniques examine whether a pattern of intercorrelations among variables fits the researcher's underlying theory of which variables are causing other variables (Aron, Aron, & Coups, 2006). It is important to remember, however, that in causal modeling we are attempting to draw causal inferences from correlational data—the degree of confidence in the validity of causal inference from correlational data is typically much weaker than inference drawn from data resulting from a well-designed experimental study where the important concept of random assignment to treatments has been incorporated (Tate, 1992). Conclusions drawn from causal modeling with correlational data must be confined to the following limitation: The results of causal modeling are valid and unbiased *only if* the assumed model adequately represents the *real* causal processes (Tate, 1992).

In causal modeling, the causal interrelationships are examined among a set of variables that have been logically ordered on the basis of time (Sprinthall, 2007). Logically, a causal variable must precede any variable that it supposedly affects—this establishes the causal ordering of the variables (Sprinthall, 2007). There are two types of causal modeling techniques: path analysis and structural equation modeling (the latter will be described at the end of this section). **Path analysis** begins with the researcher developing a diagram with arrows connecting variables and depicting the *causal flow*, or the direction of cause-and-effect. The precursor to path analysis is a simpler version of causal modeling in which the only effects represented are direct causal effects. Path analysis has a substantial advantage over the simpler model in that both *direct* and *indirect* causal effects can be estimated. We will first examine the simplest form of causal modeling, followed by a presentation of the more involved form—path analysis.

As we have mentioned, the simplest version of the causal modeling technique is one in which only direct causal effects are represented (Tate, 1992). This version is quite similar to multiple regression, as discussed in the previous chapter. The direct causal effect of an IV (X) on a DV (Y) is defined as the amount of change in Y resulting from a unit change in X , holding constant all other causal determinants of Y . The causal model is represented by a single regression equation in which the IVs are the causal determinants of the DV. For instance, using the *country-a.sav* data set from the preceding chapter, we might want to determine the direct causal paths of three IVs on a single DV. Assume we wanted to investigate the effects of a country's location in the world (*region*), its status as a developing nation (*develop*), and the number of doctors per 10,000 individuals (*docs*) on male life expectancy (*lifeexpm*). Because we are assuming only direct causal paths, our single equation would simply attempt to explain the direct causes of each of the three IVs (*region*, *develop*, *docs*) on the DV (*lifeexpm*).

The development of a causal model is probably the most difficult aspect of conducting any causal modeling study (Tate, 1992). The specification of the model is a formal declaration of the researcher's beliefs regarding the causal links among the variables. What was the basis of our decision to order the four variables as described above? These beliefs are typically influenced by several sources of information, including the research literature, formal and informal theories, personal observations and experiences with the phenomenon of interest, expert opinions, and last, but certainly not least, common sense and logic (Tate, 1992). Specification of a hypothesized model is often complicated by three sources of difficulty:

- the vagueness of many theories in social science research,
- the potentially infinite number of possible causal determinants that are often posited in the related research literature, and
- the complexity of nearly all phenomena of interest in social science research (Tate, 1992)—which has been discussed on several prior occasions in this text.

The specified causal model can be represented in two ways: as an equation or in diagrammatic form. The assumed causal model, when stated as an equation, is often referred to as a **structural equation** and is typically stated in its standardized form. If we were to define our variables using z -score coefficients—that is, the standardized form where $z_1 = \text{region}$, $z_2 = \text{develop}$, $z_3 = \text{docs}$, and $z_4 = \text{lifeexpm}$ —the structural equation for our working example would be

$$z_4 = p_{41}z_1 + p_{42}z_2 + p_{43}z_3 + e_4 \quad (\text{Equation 8.1})$$

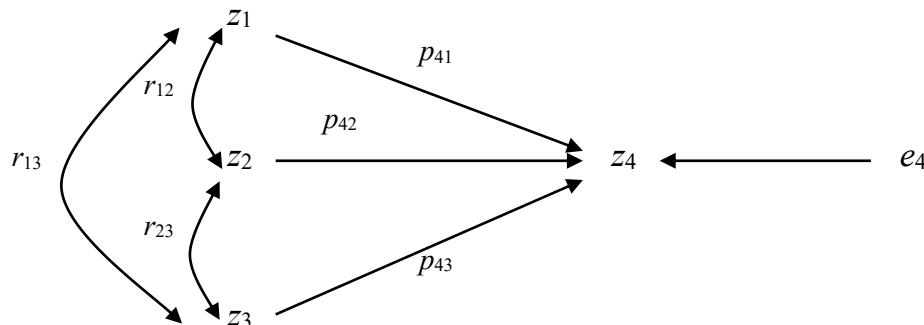
In this structural equation, the direct causal effects are represented by the p coefficients, often called **path coefficients** or **structural coefficients**. These coefficients are analogous to standardized regression coefficients, β , resulting from a multiple regression analysis (Agresti & Finlay, 2009), and their interpretation is similar (Aron, Aron, & Coups, 2006, 2008; Asher, 1983; Tate, 1992). In other words, they are interpreted

as the estimated change in the DV, expressed in standard deviation units, associated with one standard deviation change in each IV, holding the other IVs constant. The subscripts that accompany the path coefficients indicate the direction of causation, with the first subscript indicating the variable being determined and the second indicating the direct cause (Tate, 1992). The z_s indicate the standardized raw score value on each variable. The final component in the structural equation is the residual, or e_t . This residual term, called the ***disturbance term*** in causal modeling parlance, represents the composite effect of any other direct determinants of z_4 , which have not been included in the causal model, plus any measurement error in z_4 (Tate, 1992; Tatsuoka, 1988).

Although we are really dealing with three IVs and one DV in our working example, it is not accurate to refer to them as such when conducting a causal modeling study. In the specific language of causal modeling, the variable that is being explained by the model (i.e., the DV, the effect, or, in our example, z_4) is referred to as the ***endogenous variable***, while all variables not explained by the model (i.e., the IVs, the causes, or z_1 , z_2 , and z_3) are referred to as ***exogenous variables*** (Tate, 1992; Tatsuoka, 1988). Endogenous variables are assumed to have their variance explained by the exogenous variables included in the model. The variability of exogenous variables is assumed to be explained by other variables outside the causal model under consideration (Pedhazur, 1982).

The second way that a specified causal model can be portrayed is with a path diagram. A ***path diagram*** is a pictorial representation of the theoretical explanations of cause-and-effect relationships among a set of variables (Agresti & Finlay, 2009). The path diagram for our simple working example is shown in Figure 8.1. It is important to note that a path diagram is not necessary for causal modeling analysis, but it is helpful in presenting the results of the analysis (Pedhazur, 1982).

Figure 8.1. Sample Path Diagram for a Single Equation Causal Model.



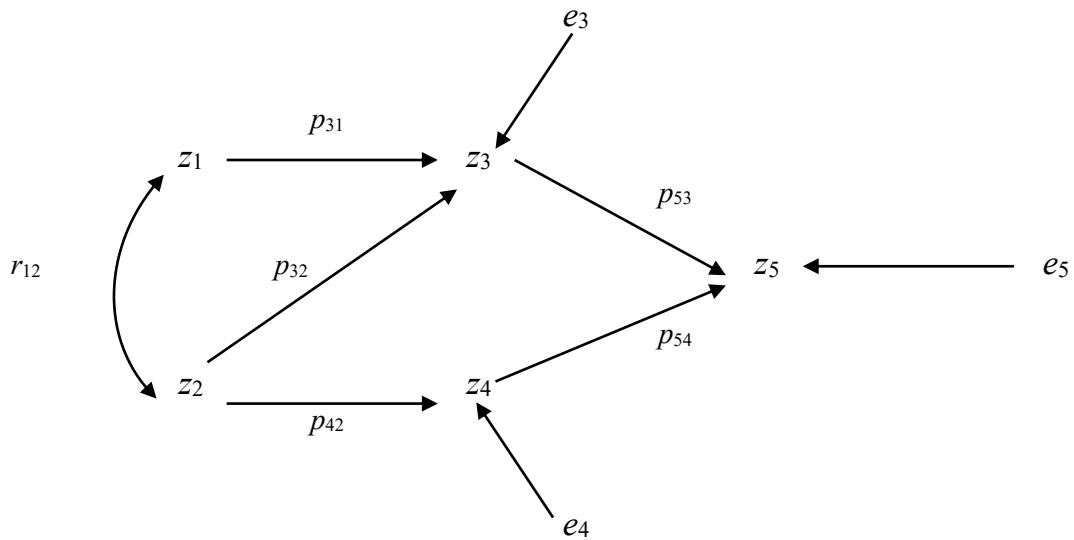
The direct causal effects of the exogenous variables z_1 , z_2 , and z_3 on the endogenous variable z_4 are shown with straight arrows, with the arrowheads indicating the assumed direction of causation (Tate, 1992). These arrows are often referred to as ***causal paths*** and are labeled with the associated path coefficients. Notice that the effect of the disturbance term is also included. Finally, the curved, double-headed arrows simply represent the bivariate correlations between exogenous variables in the model.

As we have previously mentioned, path analysis builds upon this simpler version by modeling both direct *and* indirect causal effects among the variables. An ***indirect effect*** occurs when a variable affects an endogenous variable through its effect on some other variable, known as an ***intervening variable*** (Agresti & Finlay, 2009). As in any causal modeling analysis, the first step is to specify the model of direct causal links among variables. This model will then imply indirect and total causal effects, which is a critical element that is missing in the simpler, single equation model previously discussed (Tate, 1992). Another distinct advantage of path analysis over the single equation model is the fact that it is now possible to test the

overall fit of the model to the data in order to ascertain if the model (theory) is consistent with the observed correlations (actual data). The method by which we assess model fit will be described and elaborated upon in Section 8.3. If serious inconsistencies between the model and the data exist, it is recommended that the model be revised prior to describing any of the causal effects (Tate, 1992). It is important to note that consistency between the model and the observed correlations does not prove the validity of the model, but it does represent support of the model (Tate, 1992).

We can now expand our single equation model into a path model. To do so, we decide to add another variable to our model—the number of deaths per 1,000 individuals (*deathrat*). Our initial path model is presented in Figure 8.2. The arrows indicate that *deathrat* and *docs* (although we will actually use the natural log of *docs*, or *Indocs*) are the only important direct causal determinants of *lifeexprm*. Furthermore, we hypothesize that *region* and *develop* have a direct causal effect on *deathrat* and that *develop* has a direct causal effect on *Indocs*. Notice that in our path model, *Indocs* has changed from an exogenous (unexplained) variable to an endogenous (explained) variable.

Figure 8.2. Path Diagram for the Initial Model (Male Life Expectancy).



$z_1 = \text{region}$

$z_2 = \text{develop}$

$z_3 = \text{deathrat}$

$z_4 = \text{Indocs}$

$z_5 = \text{lifeexprm}$

Model specification in path analysis becomes a much more convoluted process than in the single equation model (Tate, 1992). Correct specification in the single equation model necessitates that we have identified all causal effects of the lone endogenous variable. In our path analysis model, we have likely investigated the possible causes of our ultimate variable of interest—male life expectancy. After all, it is this variable in which we are most interested, in terms of explaining or describing its causal determinants. However, correct specification of the overall model in path analysis requires that *each* endogenous variable in the model be correctly specified. In other words, we have assumed that the model for *lifeexprm* is correctly specified, but what about the models for *deathrat* and *Indocs*? We must ensure that the models for these

additional endogenous variables are also correctly specified in order for our overall model to be valid and accurate (Tate, 1992).

Notice that we have also made informed decisions about excluding certain paths from the model. Specifically, our model has four missing paths—those from z_1 to z_4 , z_1 to z_5 , z_2 to z_5 , and z_3 to z_4 . It is important to note that these missing paths should also be consistent with theory and the associated literature (Tate, 1992). For instance, excluding the path from *region* (z_1) to *lifeexpm* (z_5) indicates the belief that *region* only affects *lifeexpm* through its effect on *deathrat* (i.e., an indirect effect). Tate (1992) notes that a final theoretical path model should be “represented as much by the excluded paths as by the included paths.”

Finally, when we are reasonably comfortable with our theoretical model, we can formally represent that model with a system of structural equations. This system of structural equations must include one for *each* endogenous variable in the model. For our current example, as depicted in Figure 8.2, these equations would be

$$z_3 = p_{31}z_1 + p_{32}z_2 + e_3 \quad (\text{Equation 8.2})$$

$$z_4 = p_{42}z_2 + e_4 \quad (\text{Equation 8.3})$$

$$z_5 = p_{53}z_3 + p_{54}z_4 + e_5 \quad (\text{Equation 8.4})$$

One should notice that in a path diagram, indirect effects are identified by a chain of two or more straight arrows all going in the same direction (Tate, 1992). The value of an indirect path coefficient is determined by finding the product of all path coefficients in the chain. In Figure 8.2, for instance, the paths from *region* to *deathrat* (p_{31}) and from *deathrat* to *lifeexpm* (p_{53}) combine to produce an indirect effect of *region* on *lifeexpm* (equal to $p_{31}p_{53}$).

Multiple regression analysis provides the values for the unbiased estimates of the path coefficients. In order to obtain the coefficients, a separate regression run must be completed for each structural equation, each including only the direct causal effects for its associated endogenous variable. Using Equation 8.2 as an example, in order to obtain p_{31} and p_{32} , one must regress z_3 (*deathrat*) on z_1 (*region*) and z_2 (*develop*). In other words, a multiple regression analysis is conducted with *deathrat* as the DV and *region* and *develop* as the IVs. Similar procedures are then conducted for the two remaining structural equations.

Probably the most crucial part of the analysis in a causal modeling study is the assessment of model fit. Before the obtained estimates of path coefficients can be used to describe the causal effects among the variables, one should determine whether or not the model is consistent with the observed, empirical correlations among the variables. This is typically accomplished by obtaining the **reproduced correlations**—those logically implied by the hypothetical or theoretical model—and comparing them to the empirical correlations (Agresti & Finlay, 2009; Tate, 1992). The reproduced correlations, therefore, are the bivariate correlations that *would* be produced *if* the causal model were correctly specified. If the observed and the reproduced correlations are reasonably close (say, within roughly .05 of each other), it can be assumed that the model is consistent with the empirical data (Tate, 1992). Larger discrepancies indicate that the model is not consistent with the data and that model revisions should be considered. Unfortunately, the reproduced correlations, and subsequent comparisons to observed correlations, cannot be obtained via SPSS computer analysis and must be computed by hand. The procedures for doing so are described in detail in Section 8.3.

Earlier in this chapter, we alluded to a second type of causal modeling strategy, and we briefly introduce it here. This second type of causal modeling offers several advantages over path analysis. **Structural equation modeling**, sometimes referred to as *latent variable modeling*, also involves diagrams with arrows showing causal flows among variables. However, one major advantage is that the computer analysis

procedure provides an overall indication of the fit between the model and the theory. We have briefly mentioned, and will see later in some detail, how this assessment of model fit must be done by hand in path analysis. A second major advantage of structural equation modeling over path analysis is that it can incorporate latent variables. A **latent variable** is a variable that cannot actually be measured but can only be *approximated* with actual measures (Aron, Aron, & Coups, 2008). For instance, intelligence is a latent variable. We would be hard-pressed to find a single measure for intelligence, but we can approximate measures for intelligence by obtaining values on several observable variables such as IQ, performance on academic achievement tests, and so on. In structural equation modeling, a diagram is set up such that latent variables are combinations of observable, measurable variables. Path diagrams in structural equation modeling are much more involved, incorporating several additional components over and above those included in a path analysis. A disadvantage of structural equation modeling is that standard statistical analysis software packages (such as the SPSS base program) are not able to conduct the required procedures. Special statistics programs or the optional SPSS add-on module AMOS (*Analysis of Moment Structures*) are required in order to conduct this type of analysis. One such program, LISREL, takes its name from the purpose of the technique—that is, to uncover linear structural relations. Discussions of structural equation modeling and the LISREL program are beyond the scope and purpose of this text. Those interested in reading further about structural equation modeling can find brief descriptions and examples in Aron, Aron, and Coups (2006, 2008) and Johnson and Wichern (2008). If detailed information is required, the reader is directed to Tabachnick and Fidell (2007), Long (1983), and Pedhazur (1982).

Sample Research Questions

Returning to our working example for path analysis, we can specify our research questions for the study as follows:

1. Is our model—which describes the causal effects among the variables *region of the world*, *status as a developing nation*, *number of deaths*, *number of doctors*, and *male life expectancy*—consistent with our observed correlations among these variables?
2. If our model is consistent, what are the estimated direct, indirect, and total causal effects among the variables?

SECTION 8.2 ASSUMPTIONS AND LIMITATIONS

Because path analysis is essentially an extension and specific application of multiple regression, the assumptions discussed in the previous chapter are also appropriate here. They are listed here as a reminder:

1. The independent variables are fixed (i.e., the same values of the IVs would have to be used if the study were to be replicated).
2. The independent variables are measured without error.
3. The relationship between the independent variables and the dependent variable is linear (in other words, the regression of the DV on the combination of IVs is linear).
4. The mean of the residuals for each observation on the dependent variable over many replications is zero.
5. Errors associated with any single observation on the dependent variable are independent of (i.e., not correlated with) errors associated with any other observation on the dependent variable.
6. The errors are not correlated with the independent variables.

7. The variance of the residuals across all values of the independent variables is constant (i.e., homoscedasticity of the variance of the residuals).
8. The errors are normally distributed.

If additional specific information regarding the assumptions associated with multiple regression is required, revisit Chapter 7.

As we have previously mentioned, valid causal inference requires the correct specification of the structural equation(s) in a path analysis. If, *and only if*, the model is correctly specified, the estimates of the various causal effects will be accurate and unbiased (Tate, 1992). In contrast, any specification errors that exist will cause the estimates of causal effects to be biased to some unknown degree. In order to use multiple regression in a manner to estimate the path coefficients, the following assumptions regarding correct model specifications must be met:

1. The model must accurately reflect the actual causal sequence.
2. The structural equation for each endogenous variable includes all variables that are direct causes of that particular endogenous variable (i.e., variables that are not included in the model, and whose effects are therefore assumed to be captured by the residuals, are also assumed not to be correlated with any of the determinant variables).
3. There is a one-way causal flow in the model (i.e., there can be no reciprocal causation between variables).
4. The relationships among variables are assumed to be linear, additive, and causal in nature. Any curvilinear relations, and so on, are to be excluded.
5. All exogenous variables are measured without error (Pedhazur, 1982; Tate, 1992).

Note that Assumptions 1 through 4 for path analysis deal directly with the specifications of the model, which, as we have previously mentioned, can be based on a combination of factors (theory, experience, research literature, opinion, etc.). As we have seen with previous techniques, Assumption 5 is largely an issue of research design and data collection.

We would be remiss if we did not discuss several limitations of path analysis. Earlier in the chapter, we referred to the fact that with path analysis we are attempting to estimate and describe causal relationships through the use of correlational data. Because of this fact, the degree of confidence we can have in the causal inferences drawn from the results of the analysis is bound to be much less than the confidence in inferences drawn from an experimental study.

Furthermore, if it is concluded that a model is not consistent with the empirical data, the model has been *misspecified*, which is a matter of degree. This degree of misspecification is subjective, to say the least, and must be evaluated by the researcher. Tate (1992) describes this limitation in the following manner:

A model, which omits several relatively unimportant causes, ignores a real but weak causal feedback, and is based on measures with some modest measurement error may still produce estimates which are technically biased but still reasonable (and valuable) approximations to the true causal effects. On the other hand, completely misleading conclusions may result from a model which is perfect in every way except for the omission of a single important variable. (p. 319)

There is no statistical test that will definitively indicate whether or not the misspecification is within reasonable limits—those decisions are left to the researcher.

Due to the above limitations, it has been suggested (Tate, 1992) that the use of conditional statements in reporting the results of a path analysis study is warranted. For instance, one might state obtained

results in the following manner: “If this model accurately reflects reality, the estimated causal effects are....” This serves as an appropriate reminder to ourselves—and to the readers of our research reports—of the limitations associated with drawing causal inferences from correlational data.

Methods of Testing Assumptions

With respect to the initial eight assumptions associated with the use of multiple regression analysis, a thorough discussion of the methods of assessing the tenability of those assumptions was presented in Chapter 7. As a reminder, these assumptions may be assessed through the use of routine data-screening procedures (see Chapter 3), but they are most appropriately tested through inspection of bivariate scatter-plots and more accurately through inspection of the residuals plots (see Chapter 7). Recall that residuals plots may be used to assess assumptions of linearity, normality, and constant variance (homoscedasticity).

The method of assessing the validity of the assumptions specific to path analysis differs greatly from the assessment of assumptions for statistical inference (Tate, 1992). No statistical procedures exist for evaluating these assumptions because they deal specifically with the degree to which the causal model has been correctly specified. There is no empirical test that can tell us the extent to which we have selected and described the correct model. In order to evaluate these five assumptions, Tate (1992) suggests that we focus our attention on the credibility, reasonableness, and utility of a proposed model. In other words,

- a model should be plausible to those who are expert in the particular field of inquiry,
- the results should be reasonable within the context of the current research literature, and
- a model should be useful in predicting future events.

The responsibility for assessing the assumptions in this manner ultimately rests with the researcher and his or her subjective judgments.

SECTION 8.3 PROCESS AND LOGIC

The Logic Behind Path Analysis

You will recall that in Chapter 7 we provided a brief overview of the calculations involved in conducting a multiple regression analysis. The same logic and associated calculations hold true here because we are again applying a regression analysis, albeit within the analysis of a causal model as opposed to a straightforward multiple regression. Therefore, we will again be calculating the β coefficients (i.e., the standardized versions) in order to represent the path coefficients for each causal determinant, the squared multiple correlation (R^2) for each structural equation, and the associated significance tests. This will be done in the same manner as described in the previous chapter.

Typically, we reserve this section of each chapter to explain the logic behind the calculations of each technique, without overwhelming the reader with mathematical equations and hand calculations, because the calculations are obtained via computer analysis. However, you will recall that we mentioned earlier that the assessment of model fit in a path analysis can be accomplished only through the use of hand calculations. The assessment of model fit is conducted by obtaining the reproduced correlations and comparing them to the empirical correlations, then evaluating them against the difference criterion of .05. Again, if all reproduced and observed correlations are relatively close to each other, the model is consistent with the empirical data. In other words, the model fits the data.

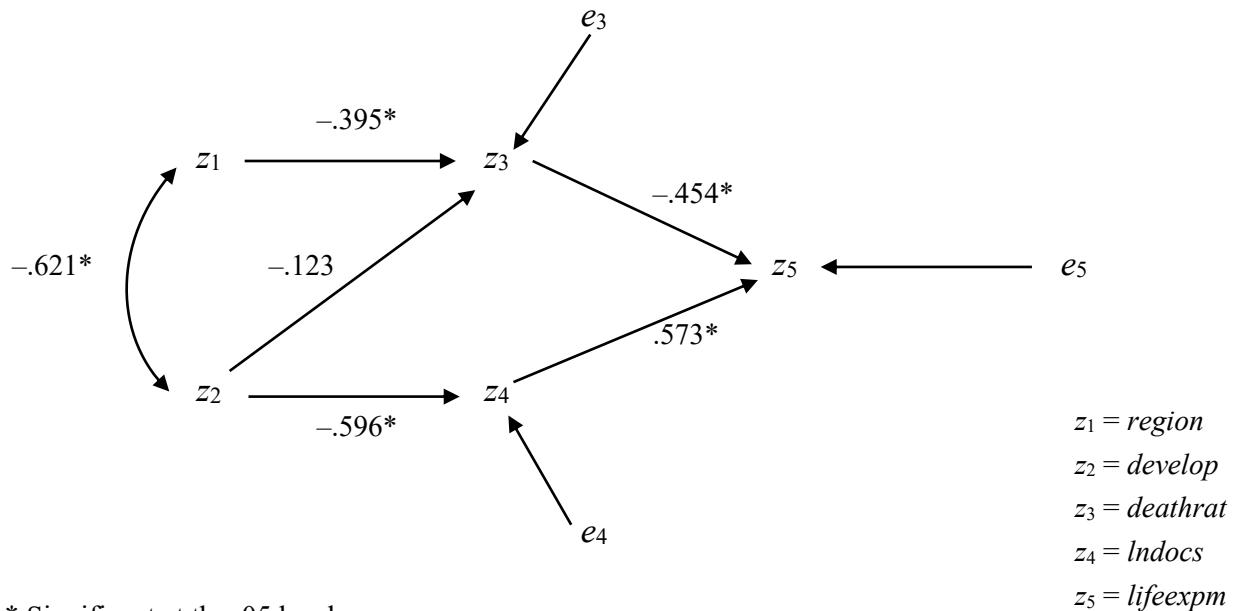
One commonly used approach to determining the reproduced correlations between two variables (and, therefore, among all variables in the set) involves the identification of all legitimate paths between the variables in the model in a process referred to as *path tracing* (Tate, 1992) or *path decomposition*

(Pedhazur, 1982). Path tracing is a process that results in a correlation coefficient for each path, which is equal to the product of all coefficients in the path. A key is that one may only use legitimate paths, which are those paths that do not violate any of the following three rules:

1. No path may pass through the same variable more than once,
2. no path may go backward on an arrow after going forward on another arrow (although it is acceptable to go forward on an arrow after *first* going backward), and
3. no path may include more than one double-headed curved arrow (Tate, 1992).

To illustrate this process, refer to Figure 8.3, which represents the same model as in Figure 8.2 but which now includes the path coefficients resulting from our regression analysis. If we wanted to obtain the reproduced correlation between z_1 and z_3 , the legitimate paths would be as follows:

Figure 8.3. Path Diagram for the Initial Model (Male Life Expectancy), Including Path Coefficients.



* Significant at the .05 level.

Path	Component
z_1 to z_3	p_{31}
z_1 to z_2 to z_3	$r_{12}p_{32}$

Therefore, the resulting equation for the reproduced correlation (symbolized by \hat{r}) between z_1 and z_3 is represented by the following:

$$\hat{r}_{13} = p_{31} + r_{12}p_{32}$$

Making the appropriate substitutions of path coefficients, we now have

$$\hat{r}_{13} = (-.395) + (-.621)(-.123) = -.319$$

As another example, let us consider the reproduced correlation between z_1 and z_5 . The legitimate paths are

<u>Path</u>	<u>Component</u>
z_1 to z_3 to z_5	$p_{31}p_{53}$
z_1 to z_2 to z_3 to z_5	$r_{12}p_{32}p_{53}$
z_1 to z_2 to z_4 to z_5	$r_{12}p_{42}p_{54}$

The resulting equation is obtained for \hat{r} between z_1 and z_5 :

$$\hat{r}_{15} = p_{31}p_{53} + r_{12}p_{32}p_{53} + r_{12}p_{42}p_{54}$$

Again, making the appropriate substitutions gives us the following:

$$\hat{r}_{15} = (-.395)(-.454) + (-.621)(-.123)(-.454) + (-.621)(-.596)(.573) = .357$$

Correlation decompositions such as these would be determined for all possible bivariate correlations in the model, with the exception of those between exogenous variables. The complete set of path decompositions and reproduced correlations for the model shown in Figure 8.3 is presented in Table 1. Reproduce these results in order to practice the identification of legitimate paths in a path model while adhering to the path tracing rules.

Notice that each path component in Table 1 includes an abbreviated label (i.e., D, I, S, or U). It is important to note the conceptual differences among the various types of path components—this is ultimately important when attempting to describe the direct, indirect, and total causal effects in a model (Tate, 1992). Causal effects are represented by paths consisting only of direct causal links—in other words, only straight arrows—that flow in only one direction. These causal effects may be *direct* (a causal path consisting of only one link; denoted “D” in Table 1) or *indirect* (consisting of two or more links; denoted “I” in Table 1). For instance, the \hat{r}_{15} decomposition shown above includes the indirect effect of z_1 on z_5 , mediated through z_3 ($p_{31}p_{53}$).

Any path components resulting from paths that have reversed causal direction at some point are called *spurious effects* (denoted “S” in Table 1), indicating that the relationship is caused by a common third factor (Tate, 1992). The paths may or may not include a double-headed curved arrow. For instance, in the decomposition of \hat{r}_{15} , the component $p_{31}r_{12}p_{42}p_{54}$ represents a spurious effect—in other words, portions of r_{35} are not due to *either* direct or indirect causal effects of z_3 on z_5 . Note that any path between two endogenous variables, which includes a curved arrow, will always represent a spurious effect (Tate, 1992).

Finally, in any model that contains more than one exogenous variable, as does Figure 8.3, the associated unexplained correlations among them will result in a degree of undeterminability with respect to the resolution of the direct and indirect effects of exogenous variables on endogenous variables (Tate, 1992). Because a model such as this does not explain the relationship among exogenous variables, we must recognize that this unanalyzed portion (denoted “U” in Table 1) may represent some degree of causal effect that has not been included in the model. In this situation, the total causal effect on an endogenous variable must be accompanied by a note specifying that there exists some uncertainty due to the unanalyzed component.

Table 1

Path Decompositions for the Initial Model (Male Life Expectancy) Shown in Figure 8.3

Reproduced Correlation	Path Decomposition
\hat{r}_{13}	$p_{31} + r_{12}p_{32}$ (D) (U)
\hat{r}_{14}	$r_{12}p_{42}$ (U)
\hat{r}_{15}	$p_{31}p_{53} + r_{12}p_{32}p_{53} + r_{12}p_{42}p_{54}$ (I) (U) (U)
\hat{r}_{23}	$p_{32} + r_{12}p_{31}$ (D) (U)
\hat{r}_{24}	p_{42} (D)
\hat{r}_{25}	$p_{32}p_{53} + p_{42}p_{54} + r_{12}p_{31}p_{53}$ (I) (I) (U)
\hat{r}_{34}	$p_{32}p_{42} + p_{31}r_{12}p_{42}$ (S) (S)
\hat{r}_{35}	$p_{53} + p_{32}p_{42}p_{54} + p_{31}r_{12}p_{42}p_{54}$ (D) (S) (S)
\hat{r}_{45}	$p_{54} + p_{42}p_{32}p_{53} + p_{42}r_{12}p_{31}p_{53}$ (D) (S) (S)

Once all of the reproduced correlations have been obtained for a path model (see Table 2), they are displayed adjacent to the observed correlations. Those reproduced correlations that have a difference greater than .05 from the empirical correlations are indicated with an asterisk (see Table 3). Any differences that are substantially larger than the .05 criterion indicate that the model is not consistent with the empirical data and revisions to the model are warranted prior to describing any of the causal effects. This method of testing for model fit is possible only when there are one or more missing paths in the model. If all possible paths are included, the reproduced correlations will *always* be exactly equivalent to the observed correlations (Tate, 1992)—by definition, the fit of the model will be perfect. *Recall that if one goal of any analysis is a parsimonious solution, we should always have some missing paths in a model.*

If it is determined that a model does not fit the data, consideration should be given to retaining included paths and incorporating excluded paths. This is accomplished by first testing all missing paths for each endogenous variable in the model (Tate, 1992). In our working example, we originally regressed z_4 on z_2 but chose to exclude the regression of z_4 on z_1 and z_3 . In order to test the missing paths for z_4 , we must regress z_4 on *all* of its direct causal determinants (z_1 , z_2 , and z_3). Similarly, we would then regress z_5 on z_1 , z_2 , z_3 , and z_4 . Support for adding any originally excluded paths is indicated by a significant path coefficient (β) in the computer output. Support for the original model is indicated by any nonsignificant path coefficients. Second, we would want to examine empirical support for all paths that we initially chose to include. This is also accomplished by examining the significance of each path coefficient—significance denotes that the model (at least, that particular coefficient) is supported by the data. If a path coefficient is not statistically

significant, one should consider dropping it from the model *unless there is strong theoretical support for its inclusion* (Tate, 1992).

Table 2

*Calculations of Reproduced Correlations for the Initial Model (Male Life Expectancy)
Shown in Figure 8.3*

$$\begin{aligned}\hat{r}_{13} &= p_{31} + r_{12}p_{32} \\ &= (-.394) + (-.626)(-.130) = \mathbf{-.313} \\ &\quad (\text{D}) \quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{14} &= r_{12}p_{42} \\ &= (-.626)(-.600) = \mathbf{.376} \\ &\quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{15} &= p_{31}p_{53} + r_{12}p_{32}p_{53} + r_{12}p_{42}p_{54} \\ &= (-.394)(-.454) + (-.626)(-.130)(-.454) + (-.626)(-.600)(.573) = \mathbf{.357} \\ &\quad (\text{I}) \quad (\text{U}) \quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{23} &= p_{32} + r_{12}p_{31} \\ &= (-.130) + (-.626)(-.394) = \mathbf{.117} \\ &\quad (\text{D}) \quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{24} &= p_{42} \\ &= (-.600) = \mathbf{-.600} \\ &\quad (\text{D})\end{aligned}$$

$$\begin{aligned}\hat{r}_{25} &= p_{32}p_{53} + p_{42}p_{54} + r_{12}p_{31}p_{53} \\ &= (-.130)(-.454) + (-.600)(.573) + (-.626)(-.394)(-.454) = \mathbf{-.397} \\ &\quad (\text{I}) \quad (\text{I}) \quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{34} &= p_{32}p_{42} + p_{31}r_{12}p_{42} \\ &= (-.130)(-.600) + (-.394)(-.626)(-.600) = \mathbf{-.070} \\ &\quad (\text{S}) \quad (\text{S})\end{aligned}$$

$$\begin{aligned}\hat{r}_{35} &= p_{53} + p_{32}p_{42}p_{54} + p_{31}r_{12}p_{42}p_{54} \\ &= (-.454) + (-.130)(-.600)(.573) + (-.394)(-.626)(-.600)(.573) = \mathbf{-.494} \\ &\quad (\text{D}) \quad (\text{S}) \quad (\text{S})\end{aligned}$$

$$\begin{aligned}\hat{r}_{45} &= p_{54} + p_{42}p_{32}p_{53} + p_{42}r_{12}p_{31}p_{53} \\ &= (.573) + (-.600)(-.130)(-.454) + (-.600)(-.626)(-.394)(-.454) = \mathbf{.605} \\ &\quad (\text{D}) \quad (\text{S}) \quad (\text{S})\end{aligned}$$

In assessing the fit of our model in Figure 8.3, it can be seen from Table 3 that 6 of the 10 reproduced correlations have differences greater (and substantially so) than .05. Upon examination of the significance tests for missing paths resulting from the supplemental regression runs as described in the previous paragraph, it was determined that several paths should be added—specifically, the paths from z_1 to z_4 (p_{41}), from z_2 to z_5 (p_{52}), and from z_3 to z_4 (p_{43}). In addition, because its beta coefficient was not significant, it was decided that the path from z_2 to z_3 (p_{32}) should be removed from the model. The resulting revised path diagram, including path coefficients, is presented in Figure 8.4.

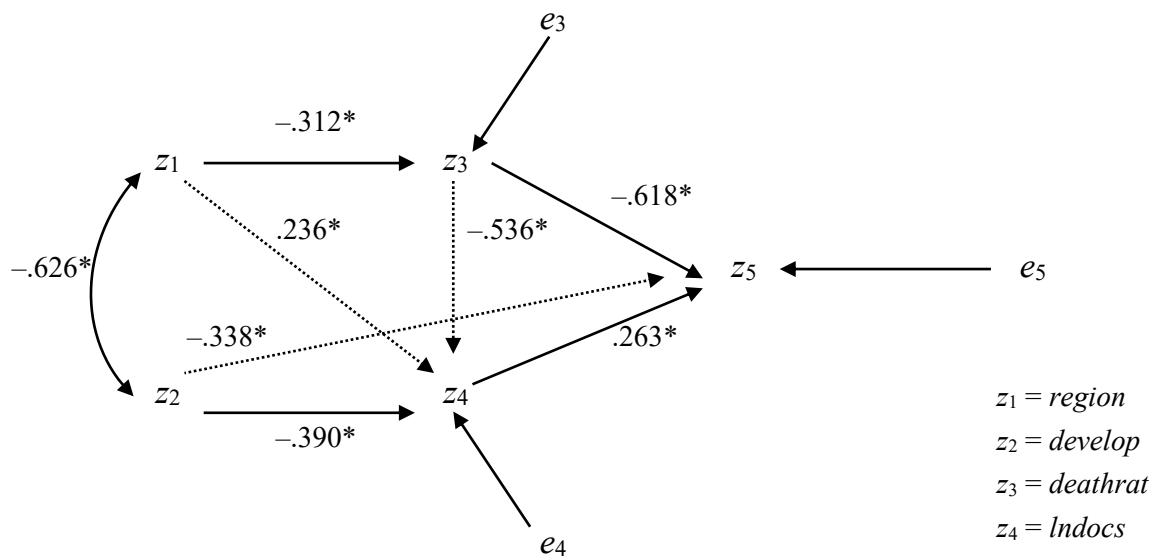
Table 3

Observed and Reproduced Correlations for the Initial Model (Figure 8.3)

	z_1	z_2	z_3	z_4	z_5
Observed Correlations					
z_1	1.000				
z_2	-.626	1.000			
z_3	-.312	.116	1.000		
z_4	.647	-.600	-.655	1.000	
z_5	.591	-.567	-.829	.870	1.000
Reproduced Correlations (Initial Model)					
z_1	1.000				
z_2	-.626	1.000			
z_3	-.313	.117	1.000		
z_4	.376*	-.600	-.070*	1.000	
z_5	.357*	-.397*	-.494*	.605*	1.000

*Difference between reproduced and observed correlation is greater than .05.

Figure 8.4. Path Diagram for the Revised Model (Male Life Expectancy), Including Path Coefficients.



* Significant at the .05 level.

Note. Revised paths are shown with dotted arrows.

Once a model has been revised, the fit should be reassessed. The path decompositions for our revised model are shown in Table 4. Calculation of the subsequent reproduced correlations is presented in Table 5. Reproduced correlations for the revised model are once again compared to the empirical correlations (see Table 6). This model obviously results in a much better fit than the initial model—only one of the reproduced correlations exceeds the .05 criterion. Had the fit *not* substantially improved, this process would continue until an adequate fit of the model to the empirical data had been achieved.

Table 4

Path Decompositions for the Revised Model (Male Life Expectancy) Shown in Figure 8.4

Reproduced Correlation	Path Decomposition				
\hat{r}_{13}	p_{31}				
	(D)				
\hat{r}_{14}	$p_{41} + r_{12}p_{42} + p_{31}p_{43}$				
	(D)	(U)	(I)		
\hat{r}_{15}	$p_{31}p_{53} + r_{12}p_{42}p_{54} + p_{41}p_{54} + p_{31}p_{43}p_{54} + r_{12}p_{52}$				
	(I)	(U)	(I)	(I)	(U)
\hat{r}_{23}	$r_{12}p_{31}$				
	(U)				
\hat{r}_{24}	$p_{42} + r_{12}p_{41} + r_{12}p_{31}p_{43}$				
	(D)	(U)	(U)		
\hat{r}_{25}	$p_{52} + p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{31}p_{43}p_{54} + r_{12}p_{41}p_{54}$				
	(D)	(I)	(U)	(U)	(U)
\hat{r}_{34}	$p_{43} + p_{31}p_{41} + p_{31}r_{12}p_{42}$				
	(D)	(S)	(S)		
\hat{r}_{35}	$p_{53} + p_{43}p_{54} + p_{31}p_{41}p_{54} + p_{31}r_{12}p_{52} + p_{31}r_{12}p_{42}p_{54}$				
	(D)	(I)	(S)	(S)	(S)
\hat{r}_{45}	$p_{54} + p_{43}p_{53} + p_{41}p_{31}p_{53} + p_{41}r_{12}p_{52} + p_{42}p_{52} + p_{42}r_{12}p_{31}p_{53} + p_{43}p_{31}r_{12}p_{52}$				
	(D)	(S)	(S)	(S)	(S)

Table 5

Calculations of Reproduced Correlations for the Revised Model (Male Life Expectancy) Shown in Figure 8.4

$$\hat{r}_{13} = p_{31}$$

$$= (-.312) = \mathbf{-.312}$$

(D)

$$\hat{r}_{14} = p_{41} + r_{12}p_{42} + p_{31}p_{43}$$

$$= (.236) + (-.626)(-.390) + (-.312)(-.536) = \mathbf{.647}$$

(D)

(U)

(I)

$$\hat{r}_{15} = p_{31}p_{53} + r_{12}p_{42}p_{54} + p_{41}p_{54} + p_{31}p_{43}p_{54} + r_{12}p_{52}$$

$$= (-.312)(-.618) + (-.626)(-.390)(.263) + (.236)(.263) + (-.312)(-.536)(.263) + (-.626)(-.338) = \mathbf{.575}$$

(I)

(U)

(I)

(I)

(U)

$$\hat{r}_{23} = r_{12}p_{31}$$

$$= (-.626)(-.312) = \mathbf{.195}$$

(U)

$$\hat{r}_{24} = p_{42} + r_{12}p_{41} + r_{12}p_{31}p_{43}$$

$$= (-.390) + (-.626)(.236) + (-.626)(-.312)(-.536) = \mathbf{-.642}$$

(D)

(U)

(U)

$$\hat{r}_{25} = p_{52} + p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{31}p_{43}p_{54} + r_{12}p_{41}p_{54}$$

$$= (-.338) + (-.390)(.263) + (-.626)(-.312)(-.618) + (-.626)(-.312)(-.536)(.263) + (-.626)(.236)(.263) = \mathbf{-.628}$$

(D)

(I)

(U)

(U)

(U)

$$\hat{r}_{34} = p_{43} + p_{31}p_{41} + p_{31}r_{12}p_{42}$$

$$= (-.536) + (-.312)(.236) + (-.312)(-.626)(-.390) = \mathbf{-.686}$$

(D)

(S)

(S)

$$\hat{r}_{35} = p_{53} + p_{43}p_{54} + p_{31}p_{41}p_{54} + p_{31}r_{12}p_{52} + p_{31}r_{12}p_{42}p_{54}$$

$$= (-.618) + (-.536)(.263) + (-.312)(.236)(.263) + (-.312)(-.626)(-.338) + (-.312)(-.626)(-.390)(.263) = \mathbf{-.864}$$

(D)

(I)

(S)

(S)

(S)

$$\hat{r}_{45} = p_{54} + p_{43}p_{53} + p_{41}p_{31}p_{53} + p_{41}r_{12}p_{52} + p_{42}p_{52} + p_{42}r_{12}p_{31}p_{53} + p_{43}p_{31}r_{12}p_{52}$$

$$= (.263) + (-.536)(-.618) + (.236)(-.312)(-.618) + (.236)(-.626)(-.338) + (-.390)(-.338) + (-.390)(-.626)(-.312)(-.618) + (-.536)(-.312)(-.626)(-.338) = \mathbf{.813}$$

(D)

(S)

(S)

(S)

(S)

At this point, we are satisfied with the fit of our model to the associated empirical data and can describe the causal effects of the variables and their correlations. A table that summarizes the causal effects of a model is typically presented in published research. Such a summary for our revised model is presented in Table 7. The reader should note that the indirect effects listed in the table are simply the sum of all indirect effects as identified in the path decompositions. The total effects consist of the sum of the direct and indirect effects.

Table 6

Observed and Reproduced Correlations for the Initial (Figure 8.3) and the Revised (Figure 8.4) Models (Male Life Expectancy)

	z_1	z_2	z_3	z_4	z_5
Observed Correlations					
z_1	1.000				
z_2	-.626	1.000			
z_3	-.312	.116	1.000		
z_4	.647	-.600	-.655	1.000	
z_5	.591	-.567	-.829	.870	1.000
Reproduced Correlations (Initial Model)					
z_1	1.000				
z_2	-.626	1.000			
z_3	-.313	.117	1.000		
z_4	.376*	-.600	-.070*	1.000	
z_5	.359*	-.397*	-.494*	.605*	1.000
Reproduced Correlations (Revised Model)					
z_1	1.000				
z_2	-.626	1.000			
z_3	-.312	.195*	1.000		
z_4	.647	-.642	-.686	1.000	
z_5	.575	-.628	-.864	.813	1.000

*Difference between reproduced and observed correlation is greater than .05.

Table 7

Summary of Causal Effects for Revised Model (Male Life Expectancy)

Outcome	Determinant	Causal Effects		
		Direct	Indirect	Total
Death rate ($R^2 = .101$)	Region	-.312*	—	-.312
	Developing status	—	—	— ⁺
Doctors ($R^2 = .738$)	Region	.236*	.167	.403 ⁺
	Developing status	-.390*	—	-.390 ⁺
	Death rate	-.536*	—	-.536 ⁺
Male life expectancy ($R^2 = .932$)	Region	—	.299	.299 ⁺
	Developing status	-.338*	-.103	-.441 ⁺
	Death rate	-.618*	-.141	-.759 ⁺
	Doctors	.263*	—	.263 ⁺

* Direct effect is significant at the .05 level.

⁺ Total effect may be incomplete due to unanalyzed components.

Interpretation of Results

Interpretation of the SPSS output for a path analysis is quite extensive because it requires several hand calculations. Once you have conducted all the regression analyses for the initial path model, the path (β) coefficients with the respective level of significance should be noted within the path model. These coefficients indicate the estimated change in the respective endogenous variables and are used to calculate the reproduced correlations through path decomposition. Reproduced correlations are then compared to the empirical correlations to test the model fit. If any reproduced correlations exhibit more than a .05 difference from the empirical correlations, the model is not consistent with the empirical data and should be revised. Once a consistent model has been generated, the specific causal effects for each endogenous variable are determined with respect to direct, indirect, and total effects. Utilizing the path decompositions is imperative in this process because both the direct and indirect effects are identified for each path. Note that a path may have several indirect effects. Consequently, the sum of these indirect values represents the overall indirect effect for the path. The total effect is also calculated by adding the direct and indirect effects for each path. Finally, R^2 is interpreted to indicate the amount of variance in each endogenous variable that is explained by its structural model.

Continuing with our example (now using *country-b.sav*) that seeks to investigate the causal effects among the variables region of the world (*region*), status as a developing nation (*develop*), number of deaths (*deathrat*), number of doctors (*docs*), and male life expectancy (*lifeexpm*), we screened data for missing cases and outliers. Outliers were identified by calculating Mahalanobis distance and conducting **Explore**. Figure 8.5 presents these results and indicates that cases 56 and 29 should be eliminated from further analysis because they exceed the chi-square criterion of 20.516 ($df = 5$). Therefore, all cases in which the Mahalanobis value exceeded the chi-square criterion were eliminated using **Select Cases, If MAH_1 \leq 20.516**. Variables were then evaluated for normality by creating a scatterplot matrix among the continuous variables (see Figure 8.6). Because the variable of doctors (*docs*) is curvilinearly related to male life expectancy, it was transformed to *Indocs* by taking its natural log. Next, multivariate normality and homoscedasticity were assessed by creating a residuals plot (see Figure 8.7). The fairly consistent spread of residuals indicates that the test assumptions are fulfilled. Prior to conducting the regression analyses, a correlation matrix (see Figure 8.8) was created because these empirical correlations will be needed later to test model fit.

Figure 8.5. Outliers Determined by Mahalanobis Distance.

Extreme Values			
		Case Number	Value
MAH_1	Highest	1	56
		2	24.06622
		3	20.82071
		4	20.00161
		5	19.16860
	Lowest	1	43
		2	16.08589
		3	.82904
		4	.86684
		5	.90204

Cases 56 and 29 exceed the $\chi^2(5) = 20.516$ criteria and should be eliminated from further analysis.

Figure 8.6. Scatterplot for Model (Male Life Expectancy) Variables.

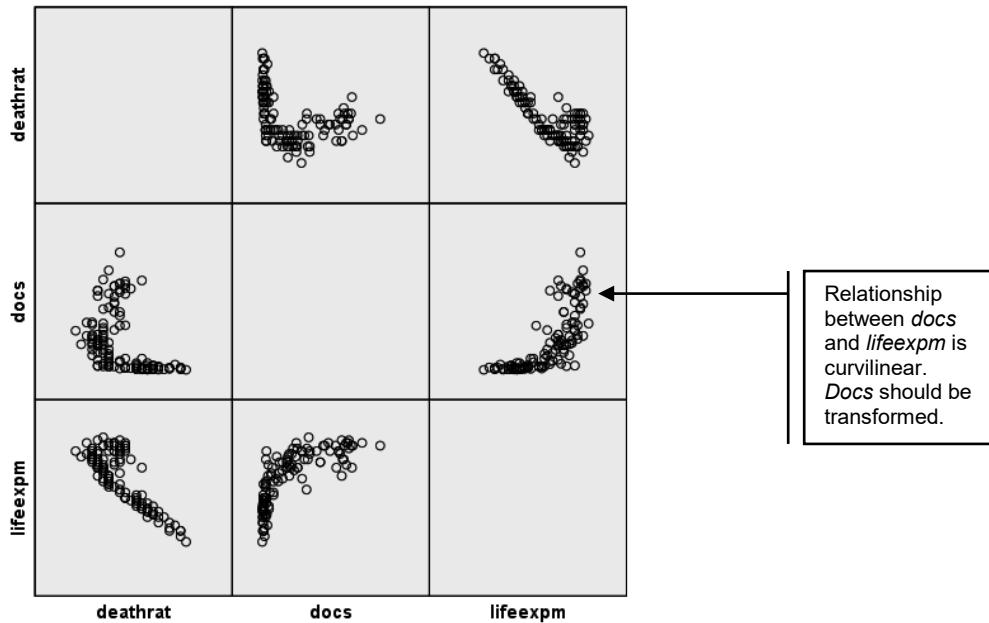


Figure 8.7. Residuals Plot for Model (Male Life Expectancy) Variables.

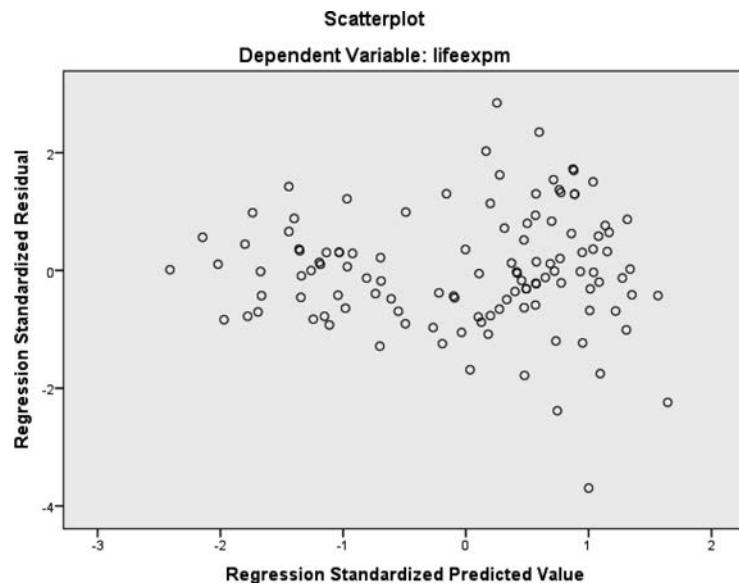


Figure 8.8. Correlation Matrix for Model (Male Life Expectancy) Variables.

		Correlations				
		region	develop	deathrat	Indocs	lifeexpm
region	Pearson Correlation	1	-.626**	-.312**	.647**	.591**
	Sig. (2-tailed)		.000	.001	.000	.000
	N	118	118	118	118	118
develop	Pearson Correlation	-.626**	1	.116	-.600**	-.567**
	Sig. (2-tailed)	.000		.211	.000	.000
	N	118	118	118	118	118
deathrat	Pearson Correlation	-.312**	.116	1	-.655**	-.829**
	Sig. (2-tailed)	.001	.211		.000	.000
	N	118	118	118	118	118
Indocs	Pearson Correlation	.647**	-.600**	-.655**	1	.870**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	118	118	118	118	118
lifeexpm	Pearson Correlation	.591**	-.567**	-.829**	.870**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	118	118	118	118	118

**. Correlation is significant at the 0.01 level (2-tailed).

Applying the initial model, the first series of regression analyses was conducted. Because this model has three endogenous variables, the following analyses were run: z_3 on z_1 and z_2 (see Figure 8.9), z_4 on z_2 (see Figure 8.10), and z_5 on z_3 and z_4 (see Figure 8.11). Prior to interpreting the path coefficients, one should review the tolerance statistic for each exogenous variable included in each regression analysis in order to determine if multicollinearity can be assumed. If tolerance is greater than .1, one may proceed with interpreting the path coefficients. For our example, tolerance statistics were all adequate. Path (beta) coefficients from the output were transferred to the path diagram of the initial model, as seen in Figure 8.3. All coefficients were significant with the exception of p_{32} . Reproduced correlations were then calculated through the path decompositions (see Table 2) and resulted in six of the reproduced correlations differing from the empirical correlations by more than .05 (see Table 3).

Figure 8.9. Regression Output for *deathrat* (z_3) on *region* (z_1) and *develop* (z_2).

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.328 ^a	.108	.092	4.286	

a. Predictors: (Constant), develop, region

b. Dependent Variable: deathrat

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	14.331	1.653				
	region	-.344	.099	-.394	-.3488	.001	.608
	develop	-1.392	1.204	-.130	-1.155	.250	1.644

a. Dependent Variable: deathrat

Figure 8.10. Regression Output for *Indocs* (z_4) on *develop* (z_2).

Model Summary ^b						
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate		
1	.600 ^a	.360	.354	1.25966		

a. Predictors: (Constant), develop
b. Dependent Variable: Indocs

Coefficients ^a						
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics
	B	Std. Error	Beta			Tolerance
1	(Constant) 3.202	.242		13.207	.000	1.000
	develop -2.229	.276	-.600	-8.076	.000	1.000

a. Dependent Variable: Indocs

Path coefficient is significant.

Figure 8.11. Regression Output for *lifeexpm* (z_5) on *deathrat* (z_3) and *Indocs* (z_4).

Model Summary ^b						
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate		
1	.935 ^a	.874	.872	3.553		

a. Predictors: (Constant), Indocs, deathrat
b. Dependent Variable: lifeexpm

Coefficients ^a						
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics
	B	Std. Error	Beta			Tolerance
1	(Constant) 67.029	1.342		49.959	.000	.571
	deathrat -1.002	.097	-.454	-10.370	.000	1.752
	Indocs 3.631	.277	.573	13.093	.000	.571

a. Dependent Variable: lifeexpm

Path coefficients are significant.

Because this initial model did not fit the empirical data, regression analyses were conducted on the following missing paths: z_4 on z_1 , z_2 , and z_3 (see Figure 8.12), and z_5 on z_1 , z_2 , z_3 , and z_4 (see Figure 8.13). Evaluation of the path coefficients and respective levels of significance indicated that only the following paths were significant: z_3 on z_1 ; z_4 on z_1 , z_2 , and z_3 ; and z_5 on z_2 , z_3 , and z_4 . Consequently, regression analyses were conducted again to include only those significant paths for each endogenous variable in order to obtain final path coefficients (nonsignificant paths were excluded, therefore changing the values of the significant paths). These analyses are presented in Figures 8.14 and 8.15. The revised model with respective path coefficients is displayed in Figure 8.4. Calculation of reproduced correlations through path decompositions (see Table 5) and subsequent comparison to the empirical correlations (see Table 6) indicate the revised model fits the empirical data. Utilizing calculations for the direct and indirect effects from Table 5, we summarize the causal effects of the revised model in Table 7. In addition, R^2 is noted for each endogenous variable within this summary table. R^2 can be found in the final (accepted) regression analyses for each endogenous variable (see Figures 8.12, 8.14, and 8.15). For instance, causal effects of *region*, *develop*, *deathrat*, and *Indocs* explain 93.2% ($R^2 = .932$) of variance in *lifeexpm*.

Figure 8.12. Regression Output of Missing Paths: *Indocs* (z_4) on *region* (z_1), *develop* (z_2), and *deathrat* (z_3).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.859 ^a	.738	.731	.81312

a. Predictors: (Constant), region, deathrat, develop
b. Dependent Variable: Indocs

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	3.904	.403		9.681	.000		
deathrat	-.187	.018	-.536	-10.563	.000	.892	1.121
develop	-1.450	.230	-.390	-6.310	.000	.601	1.663
region	.072	.020	.236	3.643	.000	.550	1.818

a. Dependent Variable: Indocs

Path coefficients are significant.

Figure 8.13. Regression Output of Missing Paths: *lifeexprm* (z_5) on *region* (z_1), *develop* (z_2), *deathrat* (z_3), and *Indocs* (z_4).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.966 ^a	.933	.930	2.624

a. Predictors: (Constant), Indocs, develop, region, deathrat
b. Dependent Variable: lifeexprm

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	79.189	1.757		45.083	.000		
region	.064	.067	.033	.960	.339	.493	2.030
develop	-7.677	.861	-.326	-8.913	.000	.446	2.244
deathrat	-1.366	.080	-.618	-17.011	.000	.451	2.218
Indocs	1.571	.302	.248	5.199	.000	.262	3.815

a. Dependent Variable: lifeexprm

Path coefficient of *lifeexprm* on *region* is NOT significant and should not be included.

Figure 8.14. Regression Output of Significant Paths: *deathrat* (z_3) on *region* (z_1).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.312 ^a	.098	.090	4.292	

a. Predictors: (Constant), region
b. Dependent Variable: *deathrat*

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	12.642	.773		16.345	.000		
region	-.273	.077	-.312	-3.540	.001	1.000	1.000

a. Dependent Variable: *deathrat*

Figure 8.15. Regression Output of Significant Paths: *lifeexprm* (z_5) on *develop* (z_2), *deathrat* (z_3), and *Indocs* (z_4).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate		
1	.965 ^a	.932	.930	2.623		

a. Predictors: (Constant), *Indocs*, *develop*, *deathrat*
b. Dependent Variable: *lifeexprm*

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics		Final path coefficients.
	B	Std. Error	Beta			Tolerance	VIF	
1 (Constant)	79.811	1.632		48.899	.000			
develop	-7.963	.808	-.338	-9.852	.000	.506	1.977	
deathrat	-1.364	.080	-.618	-17.002	.000	.451	2.217	
Indocs	1.665	.286	.263	5.823	.000	.293	3.417	

a. Dependent Variable: *lifeexprm*

Writing Up Results

Because the process of path analysis typically revises an initial model that was generated, the summary of results should first discuss the initial model. The path diagram with the path coefficients should be presented. Significant coefficients should be indicated with an asterisk. Within the results summary, you should state that reproduced correlations were calculated to check the model fit as well as indicate how many of the reproduced correlations were not consistent with the empirical correlations. This narrative should also include a description of how the revised model was derived (i.e., theory, logic, analysis of missing paths). You should then present the revised model in narrative and pictorial form (path diagram). Your summary should also discuss the significance of the revised model path coefficients. A table that compares the empirical and reproduced correlations for the initial and revised models should be presented. Because the revised model should be a good fit, indicate that reproduced correlations are consistent with empirical correlations in the narrative. A final component of a path analysis results section is a summary table of the causal effects for the revised model. This table should include the direct, indirect, and total effects for each endogenous variable. A flag is used to indicate total effects that may be incomplete due to

unanalyzed components. This table should also present R^2 for each endogenous variable. The amount of total effect (direct and indirect) for each endogenous variable should be discussed in the results summary, beginning with the endogenous variable of most interest. Figure 8.16 summarizes the components of the results narrative for path analysis as well as how it is supported by numerous tables and figures.

Figure 8.16. Steps for Presenting Path Analysis Results.

Results Narrative	Tables/Figures
1. Present initial model: variables and flow.	→ Summarize initial model in path diagram.
2. Describe any data elimination and/or transformation.	
3. Discuss significance of path coefficients.	→ Present path coefficients in path diagram.
4. Describe how reproduced correlations were not consistent with empirical correlations.	→ Create table that compares empirical correlations to reproduced correlations for the initial model.
5. Describe process of revising model.	
6. Present revised model: variables, flow, and significant path coefficients.	→ Summarize revised model in path diagram (including path coefficients).
7. Describe how reproduced correlations were consistent with empirical correlations.	→ Create table that compares empirical correlations to reproduced correlations for the revised model.
8. Discuss causal effects for each endogenous variable: total causal effects and R^2 .	→ Create table of causal effects (direct, indirect, and total) for each endogenous variable.

The following results narrative applies the output from Figures 8.3 through 8.15. Due to space constraints, we will utilize applicable figures and tables previously presented in the text.

A path analysis was conducted to determine the causal effects among the variables of region of the world (*region*, z_1), status as a developing nation (*develop*, z_2), number of deaths per 1,000 individuals (*deathrat*, z_3), number of doctors per 10,000 individuals (*docs*, z_4), and male life expectancy (*lifeexpr*, z_5). Prior to the analysis, two outliers were removed. In addition, the variable of *docs* was transformed by taking its natural log. The initial model, presented in Figure 8.3, was not consistent with the empirical data. More specifically, six of the reproduced correlations exceeded a difference of .05. Tests of the missing paths in the initial model indicated that three additional paths would significantly contribute to the model: *region* on *lndocs*, *develop* on *lifeexpr*, and *deathrat* on *lndocs*. In addition, the nonsignificant path of *develop* on *deathrat* was removed from the model. Thus, a revised model was generated and is presented in Figure 8.4. Recomputation of reproduced correlations for the revised model indicated consistency with the empirical correlations as only one reproduced correlation exceeded a difference of .05 (see Table 6). All path coefficients were significant at the .05 level. The direct, indirect, and total causal effects of the revised model are presented in

Table 7. The outcome of primary interest was male life expectancy. The determinant with the largest total causal effect was death rate (−.759). The remaining determinants of male life expectancy, as indicated by total causal effect, were number of doctors (.263), status as a developing nation (−.441), and region (.299). This model explained approximately 93% of variance in male life expectancy. The primary determinant of the number of doctors was death rate (−.536), with region (.403) and status as a developing nation (−.390) following. Approximately 74% of variance in the number of doctors was explained by the model. The primary determinant of death rate was region (−.312), which explained approximately 9.8% of variance in death rate.

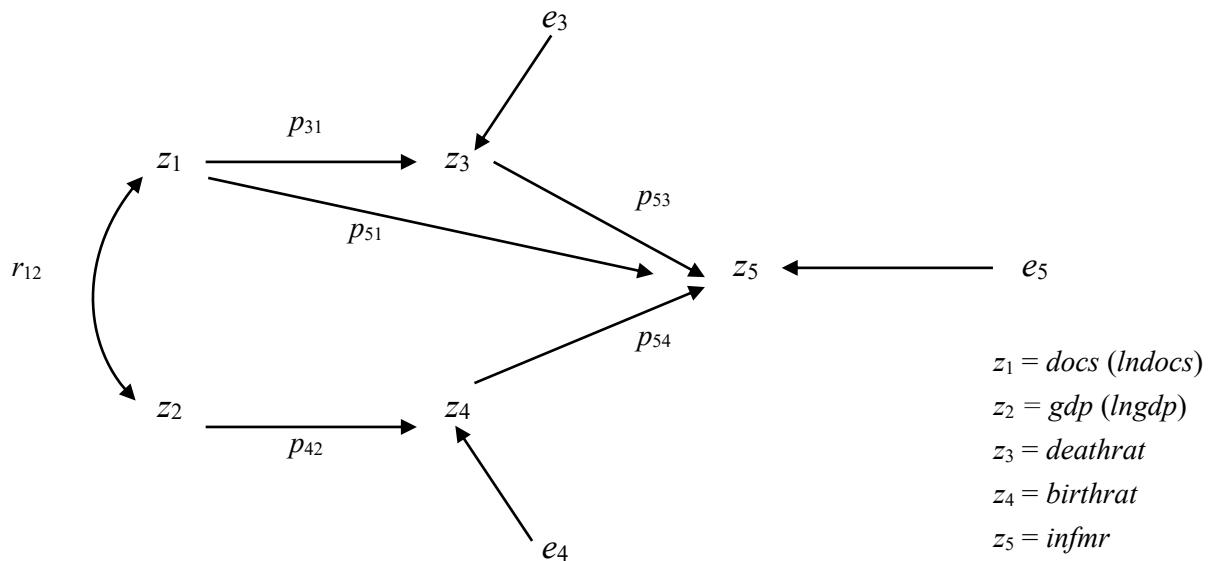
SECTION 8.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example that applies the entire process of conducting path analysis: development of model and research questions, data-screening methods, test methods, interpretation of output, and presentation of results. The SPSS data set of *country-b.sav* is utilized.

Problem

In this example, we are interested in developing a causal model for explaining infant mortality. More specifically, we will investigate the causal effects among the following variables: number of doctors per 10,000 individuals (*docs*, z_1), gross domestic product (*gdp*, z_2), death rate per 1,000 individuals (*deathrat*, z_3), birth rate per 1,000 (*birthrat*, z_4), and infant mortality per 1,000 live births (*infmr*, z_5). Utilizing logic and theory, we develop the path model shown in Figure 8.17.

Figure 8.17. Path Diagram for the Initial Model (Infant Mortality).



Specific research questions generated include the following:

1. Is our model—which describes the causal effects among the variables number of doctors, gross domestic product, death rate per 1,000, birth rate per 1,000, and infant mortality per 1,000 live births—consistent with our observed correlations among these variables?
2. If our model is consistent, what are the estimated direct, indirect, and total causal effects among the variables?

Methods, SPSS “How To,” Output, and Interpretation

Because path analysis requires a great deal of interpretation throughout the process of conducting the analysis, we have combined discussion of methods, output, and interpretation in this section. Once the path model was generated, all model variables were screened for missing data outliers and tested for assumptions. Identification of outliers was done by conducting a preliminary **Regression** to calculate Mahalanobis distance. The **Explore** procedure was completed to determine if any cases exceeded the chi-square criterion of 20.516 ($df = 5$). No outliers were found (see Figure 8.18). Test assumptions were assessed by creating a scatterplot matrix and a residuals plot. The scatterplot matrix indicated that the variables *docs* and *gdp* were curvilinear (see Figure 8.19). These variables were transformed by taking the natural log of each. The residuals plot was then created with the transformed variables and demonstrated fair dispersion (see Figure 8.20). With normality and homoscedasticity assumed, a correlation matrix was then created for all the model variables (see Figure 8.21).

Figure 8.18. Outliers Determined by Mahalanobis Distance.

Extreme Values			
		Case Number	Value
MAH_2	Highest	1	72
		2	101
		3	37
		4	91
		5	81
	Lowest	1	30
		2	50
		3	53
		4	45
		5	93

No cases exceed $\chi^2(5) = 20.516$.

Figure 8.19. Scatterplot for Model (Infant Mortality) Variables.

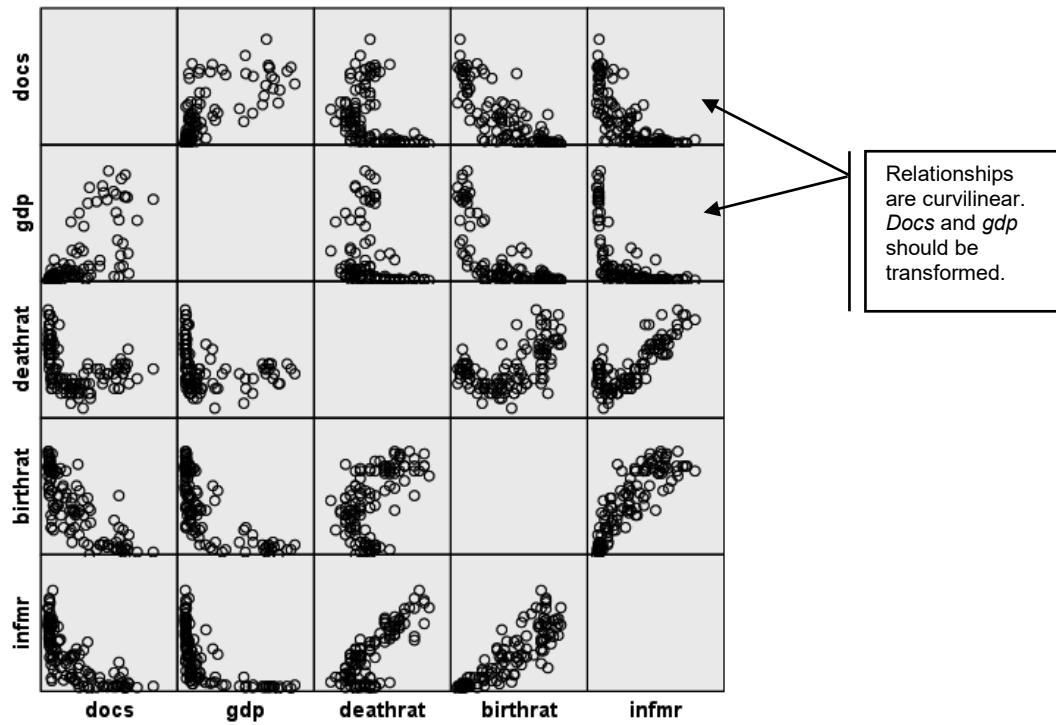


Figure 8.20. Residuals Plot for Model (Infant Mortality) Variables.

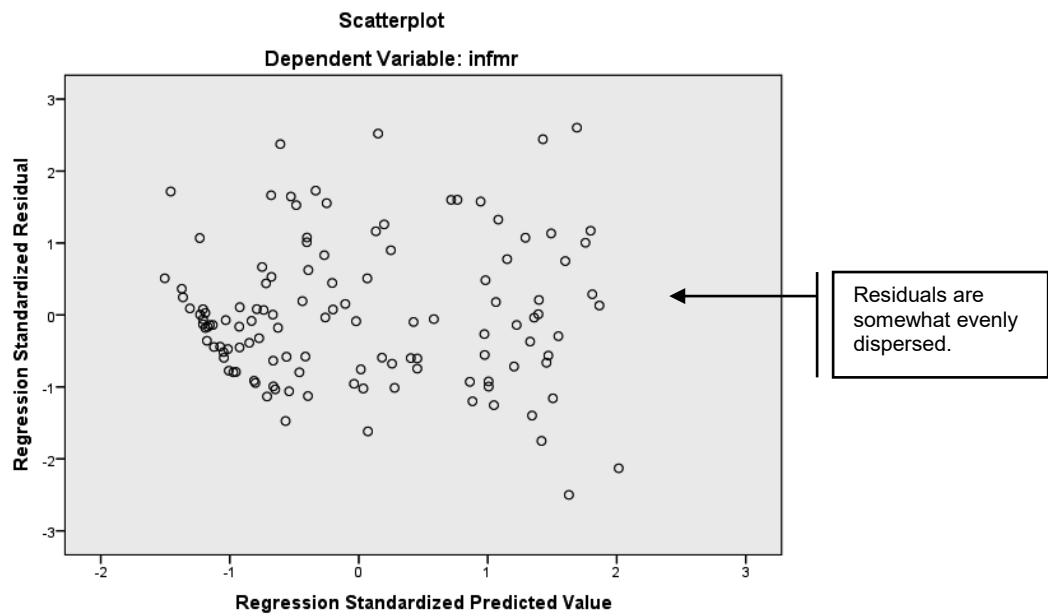


Figure 8.21. Correlation Matrix for Model (Infant Mortality Variables).

		Correlations				
		Indocs	lngdp	deathrat	birthrat	infmr
Indocs	Pearson Correlation	1	.824**	-.643**	-.821**	-.831**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	121	121	120	120	121
lngdp	Pearson Correlation	.824**	1	-.512**	-.803**	-.809**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	121	122	121	121	122
deathrat	Pearson Correlation	-.643**	-.512**	1	.568**	.780**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	120	121	121	121	121
birthrat	Pearson Correlation	-.821**	-.803**	.568**	1	.862**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	120	121	121	121	121
infmr	Pearson Correlation	-.831**	-.809**	.780**	.862**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	121	122	121	121	122

**. Correlation is significant at the 0.01 level (2-tailed).

Finally, the first set of **Regression** analyses was conducted for the three endogenous variables: z_3 on z_1 , z_4 on z_2 , and z_5 on z_1 , z_3 , and z_4 :

Analysis	Endogenous Variables	Exogenous Variables
1	deathrat (z_3)	Indocs (z_1)
2	birthrat (z_4)	lngdp (z_2)
3	infmr (z_5)	Indocs (z_1), deathrat (z_3), birthrat (z_4)

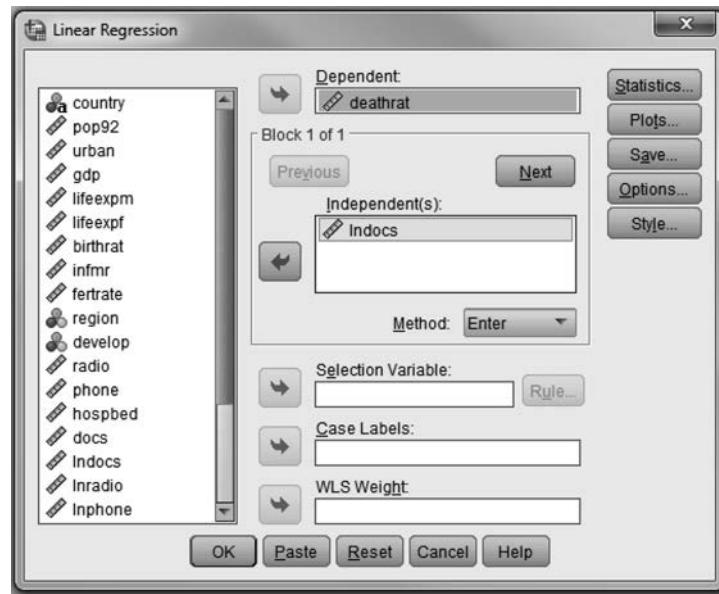
To open the Linear Regression dialog box, select the following:

Analyze
Regression
Linear

Linear Regression dialog box (see Figure 8.22)

Once in this box, click the first endogenous variable to be analyzed and move it to the **Dependent** box. For our example, the first endogenous variable is *deathrat*. Click each exogenous variable that has been identified as having a causal path to the specific endogenous variable and move to the **Independent(s)** box. For our initial model, we predicted only that *Indocs* would have a causal effect on *deathrat*. For method, select **Enter**. This is the default. Next, click **Statistics**.

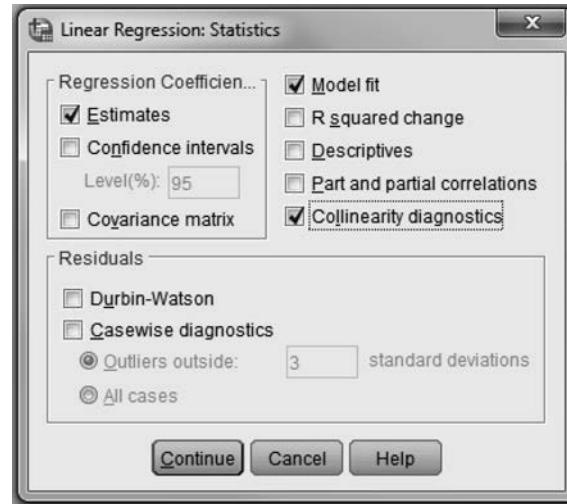
Figure 8.22. Linear Regression Dialog Box.



Linear Regression: Statistics dialog box (see Figure 8.23)

Within this box, select **Estimates** under Regression Coefficients, **Model fit**, and **Collinearity diagnostics**. Then click **Continue**. Click **OK**.

Figure 8.23. Linear Regression: Statistics Dialog Box.



Results for the first three analyses (z_3 on z_1 ; z_4 on z_2 ; and z_5 on z_1 , z_3 , and z_4) are presented in Figures 8.24 to 8.26. All tolerance statistics were greater than .1. Path coefficients can be seen in the path diagram (see Figure 8.27). Coefficients were then used to calculate the reproduced correlations through the path decompositions, which are displayed respectively in Tables 8 and 9. A comparison of the reproduced correlations to the empirical correlations shows that six of the reproduced correlations differ by more than .05 from the empirical correlations (see Table 10). Consequently, we concluded that our initial model was not consistent with the empirical data.

Figure 8.24. Regression Output for *deathrat* (z_3) on *Indocs* (z_1).

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.643 ^a	.414	.409	3.483	

a. Predictors: (Constant), Indocs

b. Dependent Variable: deathrat

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1 (Constant)	13.127	.439		29.875	.000		
Indocs	-1.874	.205	-.643	-9.131	.000	1.000	1.000

a. Dependent Variable: deathrat

Path coefficient is significant.

Figure 8.25. Regression Output for *birthrat* (z_4) on *lngdp* (z_2).

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.803 ^a	.645	.642	7.917	

a. Predictors: (Constant), lngdp

b. Dependent Variable: birthrat

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1 (Constant)	81.791	3.511		23.296	.000		
lngdp	-6.977	.475	-.803	-14.697	.000	1.000	1.000

a. Dependent Variable: birthrat

Path coefficient is significant.

Figure 8.26. Regression Output for *infmr* (z_5) on *Indocs* (z_1), *deathrat* (z_3), and *birthrat* (z_4).

Model Summary ^b					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.937 ^a	.879	.876	15.1959	

a. Predictors: (Constant), birthrat, deathrat, Indocs

b. Dependent Variable: infmr

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1 (Constant)	-28.326	9.337		-3.034	.003		
Indocs	-4.104	1.713	-.148	-2.395	.018	.273	3.662
deathrat	3.770	.402	.396	9.379	.000	.585	1.710
birthrat	1.720	.187	.521	9.202	.000	.326	3.067

a. Dependent Variable: infmr

All path coefficients are significant.

Table 8

Path Decompositions for the Initial Model (Infant Mortality) Shown in Figure 8.17

Reproduced Correlation	Path Decomposition
\hat{r}_{13}	P_{31} (D)
\hat{r}_{14}	$r_{12}p_{42}$ (U)
\hat{r}_{15}	$p_{51} + p_{31}p_{53} + r_{12}p_{42}p_{54}$ (D) (I) (U)
<hr/>	
\hat{r}_{23}	$r_{12}p_{31}$ (U)
\hat{r}_{24}	p_{42} (D)
\hat{r}_{25}	$p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{51}$ (I) (U) (U)
<hr/>	
\hat{r}_{34}	$p_{31}r_{12}p_{42}$ (S)
\hat{r}_{35}	$p_{53} + p_{31}p_{51} + p_{31}r_{12}p_{42}p_{54}$ (D) (S) (S)
<hr/>	
\hat{r}_{45}	$p_{54} + p_{42}r_{12}p_{51} + p_{42}r_{12}p_{31}p_{53}$ (D) (S) (S)

Table 9

Path Decompositions and Calculation of Reproduced Correlations for the Initial Model (Infant Mortality) Shown in Figure 8.17

$$\begin{aligned}\hat{r}_{13} &= p_{31} \\ &= (-.643) = \mathbf{-.643} \\ &\quad (\text{D})\end{aligned}$$

$$\begin{aligned}\hat{r}_{14} &= r_{12}p_{42} \\ &= (.824)(-.803) = \mathbf{-.662} \\ &\quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{15} &= p_{51} + p_{31}p_{53} + r_{12}p_{42}p_{54} \\ &= (-.148) + (-.643)(.396) + (.824)(-.803)(.521) = \mathbf{-.748} \\ &\quad (\text{D}) \quad (\text{I}) \quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{23} &= r_{12}p_{31} \\ &= (.824)(-.643) = \mathbf{-.530} \\ &\quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{24} &= p_{42} \\ &= (-.803) = \mathbf{-.803} \\ &\quad (\text{D})\end{aligned}$$

$$\begin{aligned}\hat{r}_{25} &= p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{51} \\ &= (-.803)(.521) + (.824)(-.643)(.396) + (.824)(-.148) = \mathbf{-.750} \\ &\quad (\text{I}) \quad (\text{U}) \quad (\text{U})\end{aligned}$$

$$\begin{aligned}\hat{r}_{34} &= p_{31}r_{12}p_{42} \\ &= (-.643)(.824)(-.803) = \mathbf{.425} \\ &\quad (\text{S})\end{aligned}$$

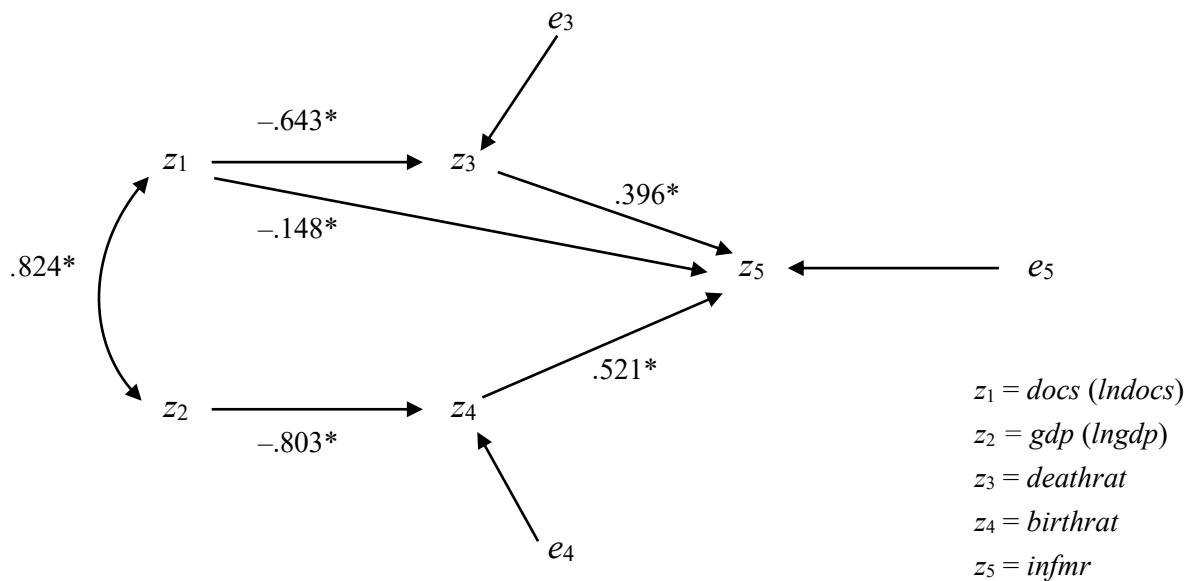
$$\begin{aligned}\hat{r}_{35} &= p_{53} + p_{31}p_{51} + p_{31}r_{12}p_{42}p_{54} \\ &= (.396) + (-.643)(-.148) + (-.643)(.824)(-.803)(.521) = \mathbf{.713} \\ &\quad (\text{D}) \quad (\text{S}) \quad (\text{S})\end{aligned}$$

$$\begin{aligned}\hat{r}_{45} &= p_{54} + p_{42}r_{12}p_{51} + p_{42}r_{12}p_{31}p_{53} \\ &= (.521) + (-.803)(.824)(-.148) + (-.803)(.824)(-.643)(.396) = \mathbf{.787} \\ &\quad (\text{D}) \quad (\text{S}) \quad (\text{S})\end{aligned}$$

Table 10*Empirical and Reproduced Correlations for the Initial Model (Figure 8.17)*

	z_1	z_2	z_3	z_4	z_5
Observed Correlations					
z_1	1.000				
z_2	.824	1.000			
z_3	-.643	-.512	1.000		
z_4	-.821	-.803	.568	1.000	
z_5	-.831	-.809	.780	.862	1.000
Reproduced Correlations (Initial Model)					
z_1	1.000				
z_2	.824	1.000			
z_3	-.643	-.530	1.000		
z_4	-.662*	-.803	.425*	1.000	
z_5	-.748*	-.750*	.713*	.787*	1.000

*Difference between reproduced and observed correlation is greater than .05.

Figure 8.27. Path Diagram for the Initial Model (Infant Mortality), Including Path Coefficients.

Because the initial model was not consistent with the empirical data, the following analyses using the same **Regression** steps were conducted to explore the significance of paths missing from the initial model:

Analysis	Endogenous Variables	Exogenous Variables
4	<i>deathrat</i> (z_3)	<i>Indocs</i> (z_1), <i>Ingdp</i> (z_2), <i>birthrat</i> (z_4)
5	<i>birthrat</i> (z_4)	<i>Indocs</i> (z_1), <i>Ingdp</i> (z_2), <i>deathrat</i> (z_3)
6	<i>infmr</i> (z_5)	<i>Indocs</i> (z_1), <i>Ingdp</i> (z_2), <i>deathrat</i> (z_3), <i>birthrat</i> (z_4)

Analysis results of missing paths, which essentially include all possible paths for each endogenous variable, are presented as follows: z_3 on z_1, z_2 , and z_4 (see Figure 8.28); z_4 on z_1, z_2 , and z_3 (see Figure 8.29); and z_5 on z_1, z_2, z_3 , and z_4 (see Figure 8.30). Analysis of missing paths for *deathrat* (z_3) reveals no additional paths that would contribute to the model. Evaluation of missing paths for *birthrat* (z_4) indicates that the path from *Indocs* (z_1) would significantly contribute to the model. Finally, analysis of missing paths for *infmr* (z_5) indicates two revisions: removal of the path from *Indocs* and the addition of the path from *Ingdp*.

Figure 8.28. Regression Output of Missing Paths: *deathrat* (z_3) on *Indocs* (z_1), *Ingdp* (z_2), and *birthrat* (z_4).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.647 ^a	.419	.404	3.499

a. Predictors: (Constant), *birthrat*, *Ingdp*, *Indocs*

b. Dependent Variable: *deathrat*

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
	9.503	3.689		2.576	.011		
1	(Constant)						
	Indocs	-1.899	.412	-.652	-4.605	.000	.250
	Ingdp	.346	.404	.116	.856	.394	.272
	birthrat	.037	.047	.107	.792	.430	.277
							3.682
							3.616

a. Dependent Variable: *deathrat*

Path coefficients for *Ingdp* on *deathrat* and *birthrat* on *deathrat* are NOT significant and should not be included.

Figure 8.29. Regression Output of Missing Paths: *birthrat* (z_4) on *Indocs* (z_1), *lngdp* (z_2), and *deathrat* (z_3).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.851 ^a	.725	.718	6.934

a. Predictors: (Constant), deathrat, lngdp, Indocs
b. Dependent Variable: birthrat

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	60.082	5.040		11.921	.000	
	Indocs	-3.851	.814	-.459	-4.732	.000	.252
	lngdp	-3.425	.738	-.399	-4.639	.000	.320
	deathrat	.145	.184	.050	.792	.430	.584

a. Dependent Variable: birthrat

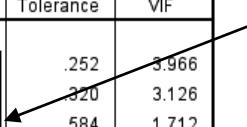
Path coefficient for *deathrat* on *birthrat* is NOT significant and should not be included.
 

Figure 8.30. Regression Output of Missing Paths: *infmr* (z_5) on *Indocs* (z_1), *lngdp* (z_2), *deathrat* (z_3), and *birthrat* (z_4).

Model Summary

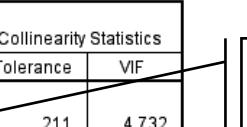
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.946 ^a	.895	.891	14.2021

a. Predictors: (Constant), birthrat, deathrat, lngdp, Indocs

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	25.204	15.399		1.637	.104	
	Indocs	-.452	1.820	-.016	-.248	.805	.211
	lngdp	-6.946	1.646	-.245	-4.219	.000	.270
	deathrat	3.896	.377	.410	10.338	.000	.581
	birthrat	1.402	.190	.425	7.374	.000	.275

a. Dependent Variable: infmr

Path coefficient for *infmr* on *Indocs* is NOT significant and should not be included.
 

The previous analyses revealed that only some of the paths were significant. Therefore, the following analyses were conducted both to determine path coefficients for those paths that were significant and to develop our revised model:

Analysis	Endogenous Variables	Exogenous Variables
7	<i>Birthrat</i> (z_4)	<i>Indocs</i> (z_1), <i>lngdp</i> (z_2)
8	<i>infmr</i> (z_5)	<i>lngdp</i> (z_2), <i>deathrat</i> (z_3), <i>birthrat</i> (z_4)

Because the very first analysis produced the path coefficient of *deathrat* on *Indocs*, this analysis did not need to be repeated. Note that this example required eight regression analyses to create an appropriate path model—this is quite common in path analysis. Regression analysis results are presented as

follows: z_4 on z_1 and z_2 (see Figure 8.31), and z_5 on z_2 , z_3 , and z_4 (see Figure 8.32). Because paths did not change for z_3 on z_1 , the results from the original analysis may be used (see Figure 8.24).

Figure 8.31. Regression Output for Significant Paths: *birthrat* (z_4) on *Indocs* (z_1) and *Ingdp* (z_2).

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.851 ^a	.723	.719	6.923	

a. Predictors: (Constant), Ingdp, Indocs

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	61.796	4.545		13.597	.000
(Constant)					
Indocs	-4.149	.720	-.494	-.5761	.000
Ingdp	-3.393	.736	-.396	-4.610	.000

a. Dependent Variable: birthrat

Amount of variance in birth rate accounted for by model.

Final path coefficients for revised model.

Figure 8.32. Regression Output for Significant Paths: *infmr* (z_5) on *Ingdp* (z_2), *deathrat* (z_3), and *birthrat* (z_4).

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.945 ^a	.893	.890	14.3681	

a. Predictors: (Constant), birthrat, deathrat, Ingdp

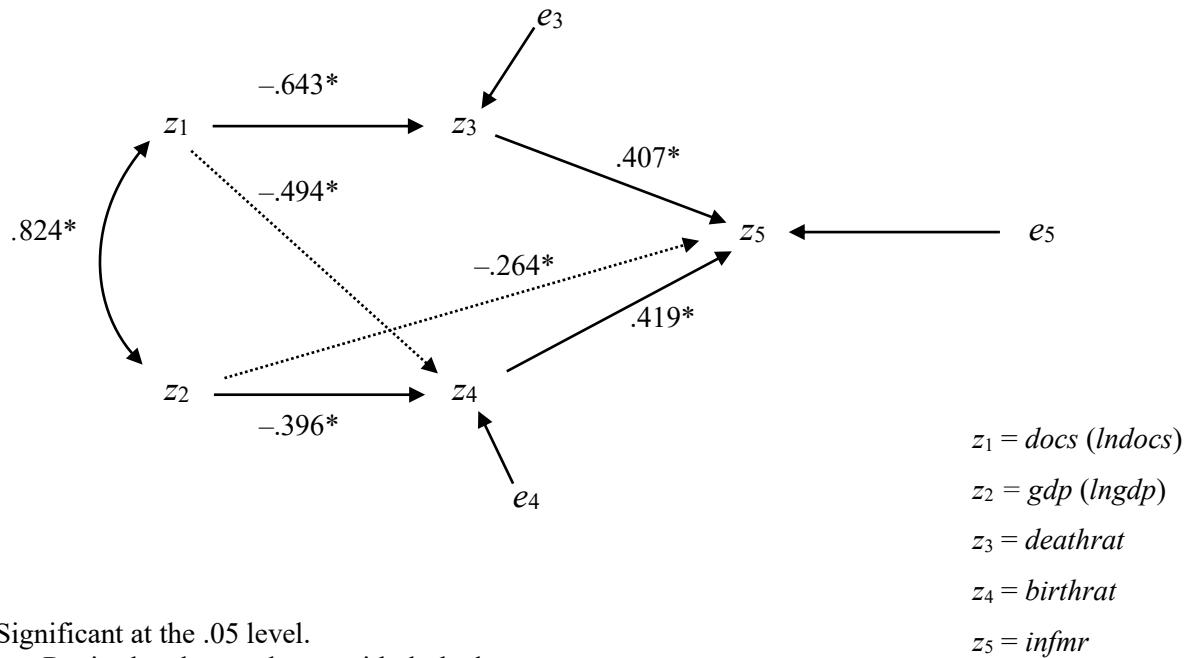
Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	30.347	15.347		1.977	.050
(Constant)					
Ingdp	-7.512	1.455	-.264	-5.163	.000
deathrat	3.792	.344	.407	11.012	.000
birthrat	1.374	.175	.419	7.864	.000

Amount of variance in infant mortality accounted for by model.

Final path coefficients for revised model.

The revised model is presented in Figure 8.33. Reproduced correlations were calculated, as defined by the path decompositions (Tables 11 and 12), and were compared to the empirical correlations (see Table 13). Only one reproduced correlation exceeded the criterion of a .05 difference. Thus, we concluded that the revised model is consistent with empirical data. The final step was to calculate the direct, indirect, and total effects for each endogenous variable. These causal effects are presented in Table 14.

Figure 8.33. Path Diagram for the Revised Model (Infant Mortality), Including Path Coefficients.



* Significant at the .05 level.

Note. Revised paths are shown with dashed arrows.

Table 11

Path Decompositions for the Revised Model (Infant Mortality) Shown in Figure 8.33

Reproduced Correlation	Path Decomposition
\hat{r}_{13}	p_{31} (D)
\hat{r}_{14}	$p_{41} + r_{12}p_{42}$ (D) (U)
\hat{r}_{15}	$p_{31}p_{53} + p_{41}p_{54} + r_{12}p_{52} + r_{12}p_{42}p_{54}$ (I) (I) (U) (U)
\hat{r}_{23}	$r_{12} p_{31}$ (U)
\hat{r}_{24}	$p_{42} + r_{12}p_{41}$ (D) (U)
\hat{r}_{25}	$p_{52} + p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{41}p_{54}$ (D) (I) (U) (U)
\hat{r}_{34}	$p_{31}p_{41} + p_{31}r_{12}p_{42}$ (S) (S)
\hat{r}_{35}	$p_{53} + p_{31}p_{41}p_{54} + p_{31}r_{12}p_{52} + p_{31}r_{12}p_{42}p_{54}$ (D) (S) (S) (S)
\hat{r}_{45}	$p_{54} + p_{42}p_{52} + p_{42}r_{12}p_{31}p_{53} + p_{41}p_{31}p_{53}$ (D) (S) (S) (S)

Table 12

Path Decompositions and Calculation of Reproduced Correlations for the Revised Model Shown in Figure 8.33

$$\hat{r}_{13} = p_{31} \\ = (-.643) = \mathbf{-.643} \\ \text{(D)}$$

$$\hat{r}_{14} = p_{41} + r_{12}p_{42} \\ = (-.494) + (.824)(-.396) = \mathbf{-.820} \\ \text{(D)} \quad \text{(U)}$$

$$\hat{r}_{15} = p_{31}p_{53} + p_{41}p_{54} + r_{12}p_{52} + r_{12}p_{42}p_{54} \\ = (-.643)(.407) + (-.494)(.419) + (.824)(-.264) + (.824)(-.396)(.419) = \mathbf{-.824} \\ \text{(I)} \quad \text{(I)} \quad \text{(U)} \quad \text{(U)}$$

$$\hat{r}_{23} = r_{12}p_{31} \\ = (.824)(-.643) = \mathbf{-.530} \\ \text{(U)}$$

$$\hat{r}_{24} = p_{42} + r_{12}p_{41} \\ = (-.396) + (.824)(-.494) = \mathbf{-.803} \\ \text{(D)} \quad \text{(U)}$$

$$\hat{r}_{25} = p_{52} + p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{41}p_{54} \\ = (-.264) + (-.396)(.419) + (.824)(-.643)(.407) + (.824)(-.494)(.419) = \mathbf{-.817} \\ \text{(D)} \quad \text{(I)} \quad \text{(U)} \quad \text{(U)}$$

$$\hat{r}_{34} = p_{31}p_{41} + p_{31}r_{12}p_{42} \\ = (-.643)(-.494) + (-.643)(.824)(-.396) = \mathbf{.528} \\ \text{(S)} \quad \text{(S)}$$

$$\hat{r}_{35} = p_{53} + p_{31}p_{41}p_{54} + p_{31}r_{12}p_{52} + p_{31}r_{12}p_{42}p_{54} \\ = (.407) + (-.643)(-.494)(.419) + (-.643)(.824)(-.264) + (-.643)(.824)(-.396)(.419) = \mathbf{.768} \\ \text{(D)} \quad \text{(S)} \quad \text{(S)} \quad \text{(S)}$$

$$\hat{r}_{45} = p_{54} + p_{42}p_{52} + p_{42}r_{12}p_{31}p_{53} + p_{41}p_{31}p_{53} \\ = (.419) + (-.396)(-.264) + (-.396)(.824)(-.643)(.407) + (-.494)(-.643)(.407) = \mathbf{.738} \\ \text{(D)} \quad \text{(S)} \quad \text{(S)} \quad \text{(S)}$$

Table 13

Empirical and Reproduced Correlations for the Initial Model (Figure 8.27) and the Revised Model (Figure 8.33)

	z_1	z_2	z_3	z_4	z_5
Observed Correlations					
z_1	1.000				
z_2	.824	1.000			
z_3	-.643	-.512	1.000		
z_4	-.821	-.803	.568	1.000	
z_5	-.831	-.809	.780	.862	1.000
Reproduced Correlations (Initial Model)					
z_1	1.000				
z_2	.824	1.000			
z_3	-.643	-.530	1.000		
z_4	-.662*	-.803	.425*	1.000	
z_5	-.748*	-.750*	.713*	.787*	1.000
Reproduced Correlations (Revised Model)					
z_1	1.000				
z_2	.824	1.000			
z_3	-.643	-.530	1.000		
z_4	-.820	-.803	.528	1.000	
z_5	-.824	-.817	.768	.738*	1.000

*Difference between reproduced and observed correlation is greater than .05.

Table 14

Summary of Causal Effects for Revised Model (Infant Mortality) Shown in Figures 8.31 to 8.33

Outcome	Determinant	Causal Effects		
		Direct	Indirect	Total
Death rate ($R^2 = .414$)	Doctors	-.643	—	-.643
	GDP	—	—	— ⁺
Birth rate ($R^2 = .723$)	Doctors	-.494	—	-.494
	GDP	-.396	—	-.396
	Death rate	—	—	— ⁺
Infant mortality ($R^2 = .893$)	Doctors	—	-.469	-.469
	GDP	-.264	-.166	-.430
	Death rate	.407	—	.407
	Birth rate	.419	—	.419

Presentation of Results

The following summary of results applies the output from Figures 8.18 through 8.33. Note that due to space constraints, we have referenced appropriate figures and tables that were previously presented in the text.

A path analysis was conducted to determine the causal effects among the variables of number of doctors per 10,000 individuals (*docs*, z_1), gross domestic product (*gdp*, z_2), number of deaths per 1,000 individuals (*deathrat*, z_3), birth rate per 1,000 individuals (*birthrat*, z_4), and infant mortality per 1,000 live births (*infmr*, z_5). Prior to the analysis, the variables of *docs* and *gdp* were transformed by taking the natural log. The initial model, presented in Figure 8.17, was not consistent with the empirical data. More specifically, six of the reproduced correlations exceeded a difference of .05. Tests of the missing paths in the initial model indicated that two additional paths would significantly contribute to the model: *birthrat* on *docs* and *infmr* on *gdp*. In addition, the nonsignificant path of *infmr* on *docs* was removed from the model. Thus, a revised model was generated and is presented in Figure 8.33. Computation of reproduced correlations for the revised model indicated consistency with the empirical correlations as only one reproduced correlation exceeded a difference of .05 (see Table 13). All path coefficients were significant at the .05 level. The direct, indirect, and total causal effects of the revised model are presented in Table 14. The outcome of primary interest was infant mortality. The determinant with the largest total causal effect was number of doctors (−.469). The remaining determinants of infant mortality, as indicated by total causal effect, were gross domestic product (−.430), birth rate (.419), and death rate (.407). This model explained approximately 89.3% of variance in infant mortality. The primary determinant of the birth rate was number of doctors (−.494), followed by gross domestic product (−.396). Approximately 72.3% of variance in the birth rate was explained by the model. The primary determinant of death rate was the number of doctors (−.643), which explained approximately 41.4% of variance in death rate.

SUMMARY

Path analysis allows the researcher to determine causal effects among numerous variables. This technique is not exploratory in nature. Rather, the researcher is testing the legitimacy of a causal model that has been based upon logic, theory, and/or experience. This causal model is depicted in a path diagram, in which effects between variables are represented by arrows. A straight line with a single arrowhead represents a direct effect, while a curved line with two arrowheads represents the bivariate correlation between two variables. An indirect effect occurs when a variable intervenes between the effect of two variables. Although a path model seeks to explain the causal determinants (i.e., IVs or, as referred to in path analysis, exogenous variables) of one variable (i.e., the DV or the endogenous variable), a model may examine several endogenous variables due to indirect effects.

Once a causal model has been developed, numerous regression analyses are conducted to determine path (beta) coefficients. To test the model fit, one must calculate the reproduced correlations for each path. Reproduced correlations are calculated through the development and application of path decompositions. If several reproduced correlations differ from empirical correlations by more than .05, then the model is not consistent with the empirical data. To revise the model, one examines missing paths within the model. Utilizing only paths that are significant, the model is revised and once again tested by comparing the empirical and reproduced correlations. Once a model has very few reproduced correlations that significantly differ from the empirical data, the model is said to be consistent with empirical data. Figure 8.34 provides a checklist for conducting path analysis.

KEYWORDS

- causal modeling
- causal paths
- disturbance term
- endogenous variable
- exogenous variable
- indirect effect
- intervening variable
- latent variable
- path analysis
- path coefficients
- path decomposition
- path diagram
- path tracing
- reproduced correlations
- spurious effects
- structural coefficients
- structural equation
- structural equation modeling

Figure 8.34. Checklist for Conducting Path Analysis.

I. Develop Path Model

- a. Create path diagram.
- b. Develop path decompositions.

II. Screen Data

- a. Missing Data?
- b. Multivariate Outliers?
 - Run preliminary Regression to calculate Mahalanobis distance.
 1. **Analyze...Regression...Linear.**
 - Identify a variable that serves as a case number and move to **Dependent** box.
 - Identify all appropriate quantitative variables and move to **Independent(s)** box.
 2. **Save.**
 - Check **Mahalanobis** under Distances.
 3. **Continue, OK.**
 4. Determine chi-square (χ^2) critical value at $p < .001$.
 - Conduct **Explore** to test outliers for Mahalanobis chi-square (χ^2).
 1. **Analyze...Descriptive statistics...Explore.**
 - Move **MAH_1** to **Dependent** box.
 - Leave **Factor** box empty.
 2. **Statistics.**
 - Check **Outliers**.
 3. **Continue, OK.**
 - Delete outliers for participants when χ^2 exceeds critical χ^2 at $p < .001$.
- c. Linearity, Normality, Homoscedasticity?
 - Create Scatterplot Matrix of all model variables.
 - If scatterplot shapes not close to elliptical shapes → reevaluate univariate normality and consider transformations.
 - Run Normality Plots with Tests within **Explore**.
 - Run preliminary Regression to create residual plot.
 1. **Analyze...Regression...Linear.**
 - Move primary endogenous variable (DV) to **Dependent** box.
 - Move exogenous variables (IVs) to **Independent(s)** box.
 2. **Plot.**
 - Select **ZRESID** for y-axis.
 - Select **ZPRED** for x-axis.
 3. **Continue, OK.**
 - If residuals are clustered at the top, bottom, left, or right area in plot → reevaluate univariate normality and consider transformations. Check to ensure that very few reproduced correlations differ from empirical correlations by more than .05.

III. Conduct Multiple Regression Analyses for Path Analysis

- a. Run Regression using **Linear Regression** for each endogenous variable.
 1. **Analyze...Regression...Linear.**
 - Move endogenous variable (DV) to **Dependent** box.
 - Move exogenous variables (IVs) to **Independent(s)** box.
 - Select **Enter**.
 2. **Statistics.**
 - Check **Model fit and Collinearity diagnostics**.
 - Check **Estimates** under **Regression Coefficients**.
 3. **Continue, OK.**
- b. Interpret tolerance.
- c. If tolerance for each exogenous variable is greater than .1, interpret path (beta) coefficient for each path.

Figure 8.34 continues on the next page.

Figure 8.34. Checklist for Conducting Path Analysis. (continued)

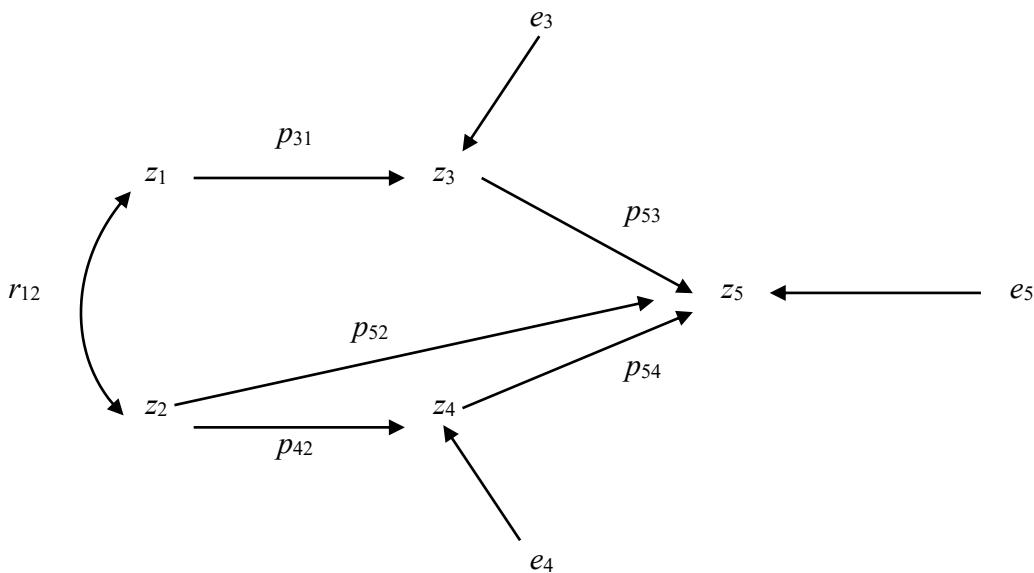
- d. Transfer path coefficients to path diagram.
- e. Calculate reproduced correlation coefficients through path decompositions.
- f. Compare reproduced correlations to empirical correlations.
- g. If only a few reproduced correlations differ from the empirical correlations by more than .05, your model is fairly consistent with the empirical data. Proceed with Step IV.
- h. If several reproduced correlations differ from the empirical correlations by more than .05, evaluate missing paths to determine if other paths may significantly contribute to the model. Analyze missing paths by following the steps beginning with III.a.
- i. Once significant paths have been determined, conduct regression analysis using only the significant paths. Analyze significant paths by following the steps beginning with III.a.

IV. Summarize Results

- a. Describe path model.
- b. Present path diagram.
- c. Describe any data elimination or transformation.
- d. Present path coefficients in path diagram.
- e. Describe how reproduced correlations were not consistent with empirical correlations.
- f. Describe process for revising model.
- g. Present revised path diagram with path coefficients.
- h. Describe how reproduced correlations were consistent with empirical correlations.
- i. Present table that compares empirical correlations to reproduced correlations for both the initial and the revised models.
- j. Discuss causal effects for each endogenous variable: total causal effects and R^2 .
- k. Present table of causal effects (direct, indirect, and total) for each endogenous variable.

Exercises for Chapter 8

The following exercises utilize the path model depicted below as well as the data set *country-c.sav*. Specifically, the variables of *Indocs* (z_1), *lngdp* (z_2), *deathrat* (z_3), *birthrat* (z_4), and *lifeexpf* (z_5) will be utilized.



1. Determine the path decompositions for the model. Be sure to label which are direct (D), indirect (I), unanalyzed (U), and spurious (S).
2. Identify the regression analyses necessary for testing this initial model.
3. Create a correlation matrix that includes all model variables. Conduct the regression analyses identified in Question 2. What are the following path coefficients?
 - a. $r_{12} =$
 - b. $p_{31} =$
 - c. $p_{42} =$
 - d. $p_{52} =$
 - e. $p_{53} =$
 - f. $p_{54} =$
4. Applying the path decompositions from Question 1, calculate the reproduced correlations.
5. Which reproduced correlations differ from the empirical correlations by more than .05?
6. Is this model consistent with empirical data? If not, what would you recommend to revise the model?

CHAPTER 9

FACTOR ANALYSIS

STUDENT LEARNING OBJECTIVES

After studying Chapter 9, students will be able to:

1. Explain what factor loadings actually measure.
2. Describe what is represented by communalities.
3. Differentiate between factor analysis and principal components analysis.
4. Define eigenvalue its relationship to a scree plot.
5. Describe how model fit is assessed in principal components analysis.
6. Reconcile the concepts of assessment of model fit and parsimony in a principal components analysis.
7. Discuss the process of rotation in terms of its benefit to an overall principal components analysis.
8. Distinguish between orthogonal and oblique rotations.
9. Develop research questions appropriate for principal components analysis.
10. Follow the appropriate SPSS guidelines provided to reduce the number of original variables in a data set by means of a principal components analysis.

In this chapter, we shift the focus of our attention to the third group of advanced/multivariate statistical techniques discussed in this text. Specifically, we present a discussion of a procedure known as *factor analysis*, which is used to describe the underlying structure that explains a set of variables. It is a technique—similar to correlation and regression—that capitalizes on shared variability. Factor analysis has many practical applications within social science settings.

SECTION 9.1 PRACTICAL VIEW

Purpose

Generally speaking, factor analysis is a procedure used to determine the extent to which *measurement overlap* (Williams, 1992)—that is, shared variance—exists among a set of variables. Its underlying purpose is to determine if measures for different variables are, in fact, measuring something in common. The mathematical procedure essentially takes the variance, as defined by the intercorrelations among a set of measures, and attempts to allocate it in terms of a smaller number of underlying hypothetical variables (Williams, 1992). These underlying hypothetical—and unobservable—variables are called factors. Factor analysis, then, is essentially a process by which the number of variables is reduced by determining which

variables cluster together, and factors are the groupings of variables that are measuring some common entity or construct.

The main set of results obtained from a factor analysis consists of ***factor loadings***. A factor loading is interpreted as the Pearson correlation coefficient of an original variable with a factor. Like correlations, loadings range in value from -1.00 (representing a perfect negative association with the factor) through 0 to $+1.00$ (indicating perfect positive association). Variables typically will have loadings on all factors but will usually have high loadings on only one factor (Aron, Aron, & Coups, 2006).

Another index provided in the results of a factor analysis is the list of ***communalities (h^2)*** for each variable. Communalities represent the proportion of variability for a given variable that is explained by the factors (Agresti & Finlay, 2009) and allows the researcher to examine how individual variables reflect the sources of variability (Williams, 1992). Communalities may also be interpreted as the squared multiple correlation of the variable as predicted from the combination of factors, or as the sum of squared loadings across all factors for that variable.

The process by which the factors are determined from a larger set of variables is called ***extraction***. There are actually several types of factor-extraction techniques, although the most commonly used empirical approaches are principal components analysis and factor analysis (Stevens, 2001). In both principal components analysis and factor analysis, linear combinations (the factors) of original variables are produced, and a small number of these combinations typically account for the majority of the variability within the set of intercorrelations among the original variables.

In ***principal components analysis***, all sources of variability—unique, shared, and error variability—are analyzed for each observed variable. However, in ***factor analysis***, only ***shared*** variability is analyzed—both unique and error variability are ignored. This is based on the belief that unique and error variance serve only to confuse the picture of the underlying structure of a set of variables (Tabachnick & Fidell, 2007). In other words, principal components analysis analyzes variance. Factor analysis analyzes covariance. Principal components analysis is usually the preferred method of factor extraction, especially when the focus of an analysis searching for underlying structure is truly exploratory, which is typically the case. Its goal is to extract the maximum variance from a data set, resulting in a few orthogonal (uncorrelated) components. When principal components analysis is used for extraction, the resulting linear combinations are often referred to as ***components***, as opposed to ***factors***. For the remainder of this chapter, we will limit our discussion to principal components analysis.

Because principal components analysis is an exploratory procedure, the first—and probably most important—decision to be made by the researcher is how many components or factors to retain and, thus, interpret. The most widely accepted criterion was developed in 1960 by Kaiser, and it has become known as ***Kaiser's rule***. The rule states that only those components whose eigenvalues are greater than 1 should be retained. An ***eigenvalue*** is defined as the amount of total variance explained by each factor, with the total amount of variability in the analysis equal to the number of original variables in the analysis (i.e., each variable contributes one unit of variability to the total amount due to the fact that the variance has been standardized).

A second, graphical method for determining the number of components is called the ***scree test*** and involves the examination of a scree plot. A ***scree plot*** is a graph of the magnitude of each eigenvalue (vertical axis) plotted against its ordinal numbers (horizontal axis). In order to determine the appropriate number

It should be noted that the term ***factor analysis*** is commonly used to represent the general process of variable reduction, regardless of the actual method of extraction utilized. For a detailed description of the various additional extraction techniques available, including maximum likelihood, unweighted least squares, generalized least squares, image factoring, and alpha factoring, refer to Tabachnick and Fidell, 2007.

of components to retain and interpret, one should look for the “knee,” or bend, in the line. A typical scree plot will show the first one or two eigenvalues to be relatively large in magnitude, with the magnitude of successive eigenvalues dropping off rather drastically. At some point, the line will appear to level off. This is indicative of the fact that these successive eigenvalues are relatively small and, for the most part, of equal size. The recommendation is to retain all components with eigenvalues in the sharp descent of the line *before* the first one where the leveling effect occurs (Stevens, 2001). If you are curious about the origin of the name for this type of plot, *scree* is formally defined as the rock debris located at the bottom of a cliff—an image one could envision in an actual scree plot.

A third criterion used to determine the number of factors to keep in a factor or principal components analysis is to retain and interpret as many factors as will account for a certain amount of total variance. A general rule of thumb is to retain the factors that account for at least 70% of the total variability (Stevens, 2001), although there may be situations where the researcher will desire an even greater amount of variability to be accounted for by the components. However, this may not always be feasible. For instance, assume we wanted to reduce the number of variables in an analysis containing 20 original variables. If it takes 15 components (or factors) to achieve the 70% criterion, we have not gained much with respect to variable reduction and some underlying structure. Realize that in this situation, some factors will undoubtedly be variable-specific (i.e., only one variable will load on a given factor). Therefore, we have not uncovered any underlying structure for the *combination* of original variables.

A final criterion for retaining components is the assessment of model fit. Recall that in Chapter 8 we discussed the assessment of model fit for a path analysis model. The assessment of model fit involved the computation of the reproduced correlations (i.e., those that would occur assuming the model represents reality) and comparing them to the original, observed correlations. If the number of correlations that are reasonably close (again, within 0.05 of each other) is small, it can be assumed that the model is consistent with the empirical data. One advantage of all factor-analytic procedures over path analysis is that computer analysis programs—including SPSS—will calculate the reproduced correlations for you, and they will even provide a percentage of the total that exceeds the cutoff value of 0.05.

With four different criteria to evaluate, how can one be sure of the number of components to retain and interpret, especially if examination of the four criteria results in different decisions regarding the number of components? Stevens (2001) offers several suggestions when faced with this often-occurring dilemma. He states that Kaiser’s rule has been shown to be quite accurate when the number of original variables is < 30 and the communalities are $> .70$ or when $N > 250$ and the mean communality is $\geq .60$ (p. 370). In other situations, use of the scree test with an $N > 250$ will provide fairly accurate results, provided that most of the communalities are somewhat large (i.e., $> .30$). Our recommendation is to examine all four criteria for alternative factor solutions and weigh them against the overall goal of any multivariate analysis—parsimony. It is our belief that the principle of parsimony is more important in factor or principal components analysis than in any other analysis technique.

Let us examine these various criteria for deciding how many components to keep through the development of an example (based upon the *country-b.sav* data set used in Chapter 8) that we will submit to a principal components analysis. Assume we wanted to determine what, if any, underlying structure exists for measures on 10 variables, consisting of

- male life expectancy (*lifeexpm*),
- female life expectancy (*lifeexpf*),
- births per 1,000 individuals (*birthrat*),
- infant mortality rate (*infmr*),
- fertility rate per woman (*fertrate*),

- natural log of doctors per 10,000 individuals (*lndocs*),
- natural log of radios per 100 individuals (*lnradio*),
- natural log of telephones per 100 individuals (*lnphone*),
- natural log of gross domestic product (*lngdp*), and
- natural log of hospital beds per 10,000 individuals (*lnbeds*).

We would first examine the number of eigenvalues greater than 1.00. The table of total variance accounted for in the initial factor solution for these 10 variables is shown in Figure 9.1. With an eigenvalue equal to 8.161, only the first component has an eigenvalue that exceeds the criterion value of 1.00. The second component (.590) does not even approach the criterion value. In addition, the first component accounts for nearly 82% of the total variability in the original variables, while the second component only accounts for about 6%.

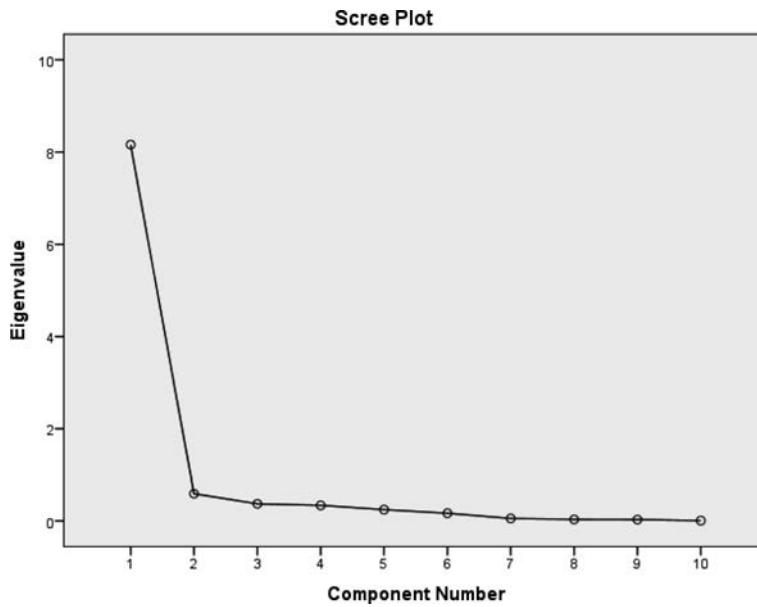
Figure 9.1. Initial Eigenvalues and Percentage of Variance Explained by Each Component.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.161	81.608	81.608	8.161	81.608	81.608
2	.590	5.902	87.510			
3	.372	3.717	91.227			
4	.338	3.384	94.611			
5	.246	2.460	97.071			
6	.168	1.677	98.748			
7	.055	.549	99.297			
8	.033	.332	99.629			
9	.030	.298	99.927			
10	.007	.073	100.000			

Extraction Method: Principal Component Analysis.

Figure 9.2. Scree Plot.



Number of Components	% Variance Accounted For	Number of Residuals > 0.05
1	81.6%	18 (40%)
2	5.9%	11 (24%)

Examination of the scree plot for this solution (see Figure 9.2) provides us with similar results. The first component is much larger than subsequent components in terms of eigenvalue magnitude. Eigenvalues of successive components drop off quite drastically. Clearly, the line begins to level off at the second component. Based on this plot, it appears that we should retain and interpret only the first component.

Although we do not provide the output here, the reproduced correlation matrix indicates that 18 (40%) of the residuals (i.e., differences between observed and reproduced correlations) have absolute values greater than 0.05. Because this indicates that the one-component model does not fit the empirical correlations very well, we might want to investigate the possibility of a two-component model. In this example, a two-component model is generated by reducing the eigenvalue threshold from its default of 1 to .5 in setting the extraction criteria of the factor analysis, thus forcing Component 2 (eigenvalue .590) to be included.

When attempting a revised model with a different number of factors, the values for the initial eigenvalues, percent of variance explained, and the appearance of the scree plot will not change—these are based solely on the original correlations. The only substantive difference will be noticed in the numbers of residuals that exceed the criterion value. In the two-component model, only 11 (24%) of the residuals exceed our .05 criterion—a substantial improvement over our previous percentage, equal to 40%. We now compare the two possible models side-by-side in order to determine which we will interpret.

Based on the variance explained and the scree plot, it appears that one component should be interpreted; however, this may be an oversimplification of the reduction of our original data. Furthermore, the addition of a second component certainly improved the fit of the model to our empirical data. For this latter reason, we will proceed with the interpretation of the two-component solution.

Before we attempt to interpret the components from the values of the loadings, it is imperative that we discuss the topic of factor (component) rotation. **Rotation** is a process by which a factor solution is made more interpretable without altering the underlying mathematical structure. Rotation is a complex mathematical procedure, and it is sometimes helpful to consider it from a geometric perspective. For the sake of simplicity, let us temporarily set aside the current example drawn from the *country-b.sav* data set and assume that we have four new variables, A through D, that we have submitted to a factor analysis. The analysis returned two components, and the associated hypothetical loadings are presented below:

Variable	Loading on Component 1	Loading on Component 2
A	.850	.120
B	.700	.210
C	-.250	.910
D	.210	-.750

If we were to plot these combinations of loadings in a scatterplot of Component 1 by Component 2, the result would appear as in Figure 9.3. Notice that the possible loadings on each component range from -1.00 to $+1.00$ and that the locations of the four variables in that geometric space have been plotted according to the combination of loadings on the two components in Part (a) of Figure 9.3—for instance, Variable A is located at $X = .850$, $Y = .120$. Although the points are generally near the lines, if we were able to rotate the axes, we would notice a better fit of the loadings to the actual components. In Part (b) of Figure 9.3, we now see how three of the four loadings line up nearly perfectly with the two components. This process alters the original values of the component loadings without changing their mathematical properties. This gives us the ability to name the components with greater ease because three of the variables correlate nearly perfectly with the two components, and the rotated factor loadings would change accordingly.

The researcher must decide whether to use an orthogonal or oblique rotation. **Orthogonal rotation** is a rotation of factors that results in factors being uncorrelated with each other. The resultant computer output is a *loading* matrix (i.e., a matrix of correlations between all observed variables and factors) where the size of the loading reflects the extent of the relationship between each observed variable and each factor. Because the goal of factor analysis is to obtain underlying factors that are uncorrelated (thereby representing some unique aspect of the underlying structure), it is recommended that orthogonal rotation be used instead of oblique. There are three types of orthogonal rotation procedures—varimax, quartimax, and equamax—which varimax is the most commonly used.

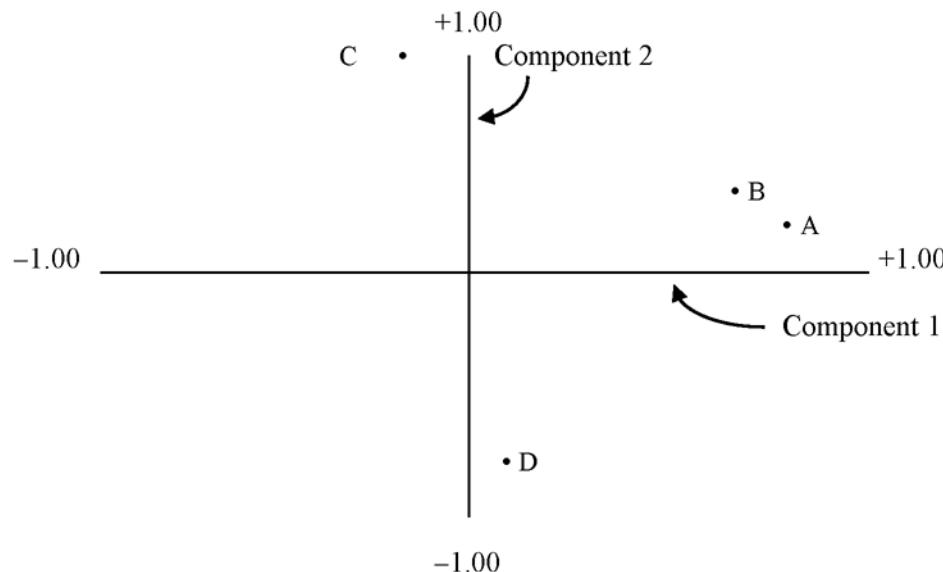
Oblique rotation results in factors being correlated with each other. Three matrices are produced from an oblique rotation: a **factor correlation matrix** (i.e., a matrix of correlations between all factors); a loading matrix that is separated into a *structure* matrix (i.e., correlations between factors and variables); and a *pattern* matrix (i.e., unique relationships with no overlapping among factors and each observed variable). The interpretation of factors is based on loadings found in the pattern matrix. One would use an oblique rotation only if there were some prior belief that the underlying factors were correlated. Several types of oblique rotations exist, including direct oblimin, direct quartimin, orthoblique, and promax. Direct oblimin is arguably the most frequently used form of oblique rotation.

Once we have rotated the initial solution, we are ready to *attempt* interpretation. We emphasize the *attempt* at interpretation because, by its very nature, interpretation of components or factors involves much

subjective decision making on the part of the researcher (Williams, 1992). The rotated component loadings for our working example that utilized varimax rotation are presented in Figure 9.4. Initially, one should notice that each variable has a loading on each component, although in most cases each has a high loading on only one component. Some of the variables in our example have loaded relatively high on both components, but we will assign a given variable to the component with the higher loading (as shown by the boxes) and attempt to interpret them in that fashion.

Figure 9.3. Illustration of Geometric Interpretation of Rotation of Components.

(a)



(b)

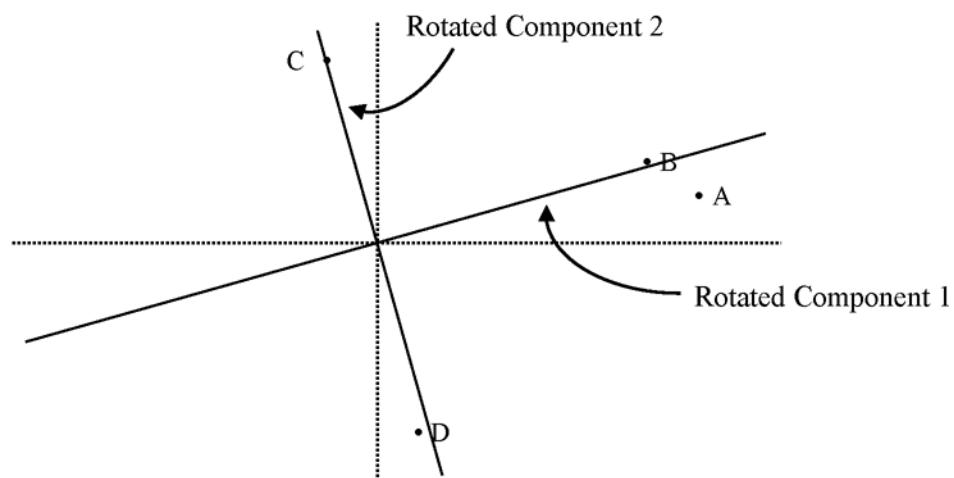


Figure 9.4. Component Loadings for the Rotated Solution.

	Component	
	1	2
fertrate	-.879	-.305
birthrat	-.858	-.393
lifeexpr	.846	.450
infmr	-.839	-.461
lifeexpf	.829	.509
Indocs	.763	.502
Inradio	.279	.879
Inbeds	.480	.719
Ingdp	.627	.688
Inphone	.675	.678

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

When interpreting or naming components, one should pay particular attention to two aspects—the size and direction of each loading. The name attached to each component should reflect the relative sizes and directions of the loadings. Notice that Component 1 contains both high positive and high negative loadings. This is referred to as a *bipolar* factor. In other words, the name we assign to Component 1 should *primarily* reflect the strong positive loadings for male life expectancy (*lifeexpr*), female life expectancy (*lifeexpf*), and subsequently the loading for *Indocs*. Due to negative loadings, the name should also reflect the *opposite* of fertility rate (*fertrate*), birth rate (*birthrat*), and infant mortality rate (*infmr*). Bipolar components or factors are usually more difficult to interpret. Because this component seems to address the *end* of an individual's lifespan as opposed to the beginning of that lifetime, and factoring in the number of docs (*Indocs*), we might consider attaching the label *Healthy Lifespan* to this component. The second component—on which all variables have positive loadings—addresses the numbers of radios, hospital beds, and phones, as well as the nation's gross domestic product. We might consider attaching the label *Economic Stature* to this component.

It is important to note that principal components analysis may be used as a variable reducing scheme for further analyses (Stevens, 2001). We have already examined the main application of the analysis—to determine empirically how many underlying constructs account for the majority of the variability among a set of variables. Principal components analysis may also be used as a precursor to multiple regression as a means of reducing the number of predictors, especially if the number of predictor variables is quite large relative to the number of participants. In addition, components analysis may be used to reduce the number of criterion variables in a multivariate analysis of variance. It is often recommended that a large number of DVs not be used in a MANOVA procedure. If you have a large number of DVs that you are interested in analyzing, reducing the overall number of variables through a principal components analysis would allow you to accomplish this task rather efficiently.

If a principal components analysis is to be followed by a subsequent analytic procedure, factor scores are often used. **Factor scores** are estimates of the scores participants would have received on each of the factors had they been measured directly (Tabachnick & Fidell, 2007). Many different procedures may be used to estimate factor scores, the most basic of which is to simply sum the values across all variables that load on a given factor or component. Alternatively, a mean could be calculated across all variables that would then represent the score on a factor. Factor scores could also be estimated or predicted through a regression analysis. Those values are then entered into the subsequent analysis as if they were raw variables.

Finally, there are two basic types of factor analytic procedures, based on their overall intended function: exploratory and confirmatory factor analyses. In **exploratory factor analysis**, the goal is to describe and summarize data by grouping together variables that are correlated. The variables included in the analysis may or may not have been chosen with these underlying structures in mind (Tabachnick & Fidell, 2007). Exploratory factor analysis usually occurs during the early stages of research, when it often proves useful to consolidate numerous variables.

Confirmatory factor analysis is much more advanced and sophisticated than exploratory factor analysis. It is often used to test a theory about latent (i.e., underlying, unobservable) processes that might occur among variables. A major difference between confirmatory and exploratory factor analyses is that in a confirmatory analysis, variables are painstakingly and specifically chosen in order to adequately represent the underlying processes (Tabachnick & Fidell, 2007). The main purpose of confirmatory factor analysis is to confirm—or disconfirm—some *a priori* theory. LISREL, as previously discussed in Chapter 8, is often used as the analytical computer program in such studies.

Sample Research Questions

Returning to the example based on the *country-b.sav* data set, we now specify the main research questions to be addressed by our principal components analysis. Using the 10 original variables, the appropriate research questions are as follows:

1. How many reliable and interpretable components are there among the following 10 variables: male life expectancy, female life expectancy, births per 1,000 individuals, infant mortality rate, fertility rate per woman, number of doctors per 10,000 individuals, number of radios per 100 individuals, number of telephones per 100 individuals, gross domestic product, and number of hospital beds per 10,000 individuals?
2. If reliable components are identified, how might we interpret those components?
3. How much variance in the original set of variables is accounted for by the components?

SECTION 9.2 ASSUMPTIONS AND LIMITATIONS

If principal components analysis and factor analysis are being used in a descriptive fashion as a method of summarizing the relationships among a large set of variables, assumptions regarding the distributions of variables in the population are really not in force and, therefore, do not need to be assessed (Tabachnick & Fidell, 2007). This is usually the case because, as previously mentioned, principal components and factor analyses are almost always exploratory and descriptive in nature. It should be noted, however, that if the variables are normally distributed, the resultant factor solution will be enhanced. To the extent to which normality fails, the solution is degraded—although it still may be worthwhile (Tabachnick & Fidell, 2007).

Previous versions of SPSS provided a statistical test of model fit—a value for a test statistic was provided and subsequently evaluated using a chi-square criterion. In situations where this test statistic was evaluated—in an inferential manner—and used to determine the number of factors or components, assessment of model assumptions takes on much greater importance. Because recent revisions of SPSS have omitted the chi-square test of model fit, this criterion can obviously no longer be used to determine the number of factors to interpret. Therefore, it is not necessary to test the assumptions of multivariate normality and linearity. However, we recommend that both of these assumptions be evaluated and any necessary transformations be made because *ensuring the quality of data will only improve the quality of the resulting factor or component solution.*

As a reminder, these two aforementioned assumptions are formally stated as follows:

1. All variables, as well as all linear combinations of variables, must be normally distributed (assumption of multivariate normality).
2. The relationships among all pairs of variables must be linear.

Factor analyses, in general, are subject to a potentially severe limitation. Recall that the bases for any underlying structure that is obtained from a factor analysis are the relationships among all original variables in the analysis. Correlation coefficients have a tendency to be less reliable when estimated from small samples. If unreliable—or, at least, *less reliable*—correlations exist among variables, and those variables are subjected to a factor analysis, the resultant factors will also not be very reliable. Tabachnick and Fidell (2007) offer the following guidelines for sample sizes and factor analyses:

Approximate Sample Size	Estimated Reliability
50	very poor
100	poor
200	fair
300	good
500	very good
1,000	excellent

As a general rule of thumb, they suggest that a data set include *at least* 300 cases for a factor analysis to return reliable factors. If a solution contains several high-loading variables ($> .80$), a smaller sample (e.g., $n = 150$) would be sufficient.

Stevens (2001) offered a different, although somewhat similar, set of recommendations based on the number of variables (with minimum/maximum loadings) per component (p. 384). Specifically, these recommendations are as follows:

1. Components with four or more loadings above .60 in absolute value (i.e., $|.60|$) are reliable, regardless of sample size.
2. Components with about 10 or more low loadings (i.e., $<|.40|$) are reliable as long as the sample size is greater than 150.
3. Components with only a few low loadings should not be interpreted unless the sample size is at least 300.

It should be noted that these recommendations constitute *general* guidelines, *not* specific criteria that *must* be met by the applied researcher. If researchers are planning a factor analysis with small sample sizes, it is recommended that they apply **Bartlett's sphericity test**. This procedure tests the null hypothesis that the variables in the population correlation matrix are uncorrelated. If one fails to reject this hypothesis, there is no reason to conduct a principal components analysis because the variables are already uncorrelated—that is, they have nothing in common (Stevens, 2001).

Methods of Testing Assumptions

Recall that the assessment of multivariate normality is not easily accomplished through the use of standard statistical software packages. The most efficient method of assessing multivariate normality is to assess univariate normality—remember that univariate normality is a necessary condition for multivariate normality. Normality among individual variables may be evaluated by examining the values for skewness and kurtosis, normal probability plots, and/or bivariate scatterplots.

The assumption of linearity is best tested through the inspection of bivariate scatterplots obtained for each pair of original variables. Recall that if a relationship is in fact linear, a general elliptical shape should be apparent in the scatterplot.

SECTION 9.3 PROCESS AND LOGIC

The Logic Behind Factor Analysis

The underlying, mathematical objective in principal components analysis is to obtain uncorrelated linear combinations of the original variables that account for as much of the total variance in the original variables as possible (Johnson & Wichern, 2008). These uncorrelated linear combinations are referred to as the **principal components**. The logic behind principal components analysis involves the partitioning of this total variance by initially finding the first principal component (Stevens, 2001). The **first principal component** is the linear combination that accounts for the *maximum* amount of variance and is defined by the following equation:

$$PC_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1p}x_p \quad (\text{Equation 9.1})$$

where PC_1 is the first principal component, x_i refers to the measure on the original variable, and a_{11} refers to the weight assigned to a given variable for the first principal component (the first subscript following the a identifies the specific principal component, and the second subscript identifies the original variable)—for instance, the term $a_{11}x_1$ refers to the product of the weight for variable 1 on PC_1 and the original value for an individual on variable 1. The subscript p is equal to the total number of original variables. This linear combination, then, accounts for the maximum amount of variance within the original set of variables—the variance of the first principal component is equal to the largest eigenvalue (i.e., the eigenvalue for the first component).

The analytic operation then proceeds to find the second linear combination—*uncorrelated* with the first linear combination—that accounts for the next largest amount of variance (after that which has been attributed to the first component has been removed). The resulting equation would be as follows:

$$PC_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2p}x_p \quad (\text{Equation 9.2})$$

It is important to note that the extracted principal components are not related. In other words,

$$r_{PC_1 \bullet PC_2} = 0$$

The third principal component is constructed so that it is uncorrelated with the first two and accounts for the next largest amount of variance in the system of original variables, after the two largest amounts have been removed. This process continues until all variance has been accounted for by the extracted principal components.

Interpretation of Results

The process of interpreting factor analysis results focuses on the determination of the number of factors to retain. As mentioned earlier, there are several methods/criteria to utilize in this process:

1. Eigenvalue—Components with eigenvalues greater than 1 should be retained. This criterion is fairly reliable when the number of variables is < 30 and communalities are $> .70$ or the number of individuals is > 250 and the mean communality is $\geq .60$.
2. Variance—Retain components that account for at least 70% of total variability.
3. Scree Plot—Retain all components within the sharp descent, before eigenvalues level off. This criterion is fairly reliable when the number of individuals is > 250 and communalities are $> .30$.
4. Residuals—Retain the components generated by the model if only a few residuals (the difference between the empirical and the reproduced correlations) exceed .05. If several reproduced correlations differ, you may want to include more components.

Because the sample size and number of variables can impact the number of factors generated in the analysis as well as the assessment of these four criteria, we recommend utilizing all four. Another reason to examine all four criteria is that within an exploratory factor analysis, the eigenvalue is the default criterion for determining the number of factors, which can lead to an inaccurate number of factors retained. For instance, if an analysis determines that only two components have eigenvalues greater than 1, the model generated will include only those two components. However, the researcher may examine the other three criteria and determine that one more component should be included. In such an instance, the analysis would have to be conducted again to override the eigenvalue criteria so that three components instead of two would be generated.

Once criteria have been evaluated and you have determined the appropriate number of components to retain (the reader should note that this decision may lead to further analysis in order to include the appropriate number of components), the nature of each component must be assessed in order to interpret/name it. This is done by noting positive and negative loadings, ordering variables with respect to loading strength, and examining the content of variables that composes each component.

Although this interpretation process has been somewhat applied to our initial example, we will describe this process in more depth in conjunction with the output. Our example, which applies the *country-b.sav* data set, seeks to determine what, if any, underlying structure exists for measures on the following 10 variables: male life expectancy (*lifeexpm*), female life expectancy (*lifeexpf*), births per 1,000 individuals (*birthrat*), infant mortality rate (*infmr*), fertility rate per woman (*fertrate*), natural log of doctors per 10,000 individuals (*lndocs*), natural log of radios per 100 individuals (*lnradio*), natural log of telephones per 100 individuals (*lnphone*), natural log of gross domestic product (*lngdp*), and natural log of hospital beds per 10,000 individuals (*lnbeds*). Data were first screened for missing data and outliers. No outliers were found when utilizing Mahalanobis distance. Univariate linearity and normality were analyzed

by creating a scatterplot matrix (see Figure 9.5). The elliptical shapes indicate normality and linearity. Note that the following variables were previously transformed variables by taking the natural log: number of doctors per 10,000 individuals, number of radios per 100 individuals, number of telephones per 100 individuals, gross domestic product, and number of hospital beds per 10,000 individuals. A factor analysis was then conducted using **Dimension Reduction**, which utilized the eigenvalue criterion and varimax rotation. Applying the four methods of interpretation, we first examined the eigenvalues in the table of total variance (see Figure 9.6). Only one component had an eigenvalue greater than 1. However, the eigenvalue criterion is only reliable when the number of variables is less than 30 and communalities are greater than .70. Figure 9.7 presents the communalities and indicates that two variables have values less than .70. Consequently, the application of the eigenvalue criterion is questionable. The next criterion to assess is variance, displayed in Figure 9.6. The first component accounts for nearly 82% of the total variance in the original variables, whereas the second component accounts for only 5.9%. Note that because only one component was retained, the factor solution was not rotated. The scree plot was then assessed and indicates that the eigenvalues after the first component drop off drastically (see Figure 9.2). These last two methods imply that only the first component should be retained. However, evaluation of residuals (differences between the reproduced correlations and the original correlations) indicate that two components should be investigated. Figure 9.8 presents the reproduced correlations as well as the residuals. Of the 45 residuals, 18 (40%) exceed the 0.05 criterion. Because two of the four criteria are in question, we will investigate retaining two components to improve model fit.

Figure 9.5. Scatterplot Matrix of Variables.

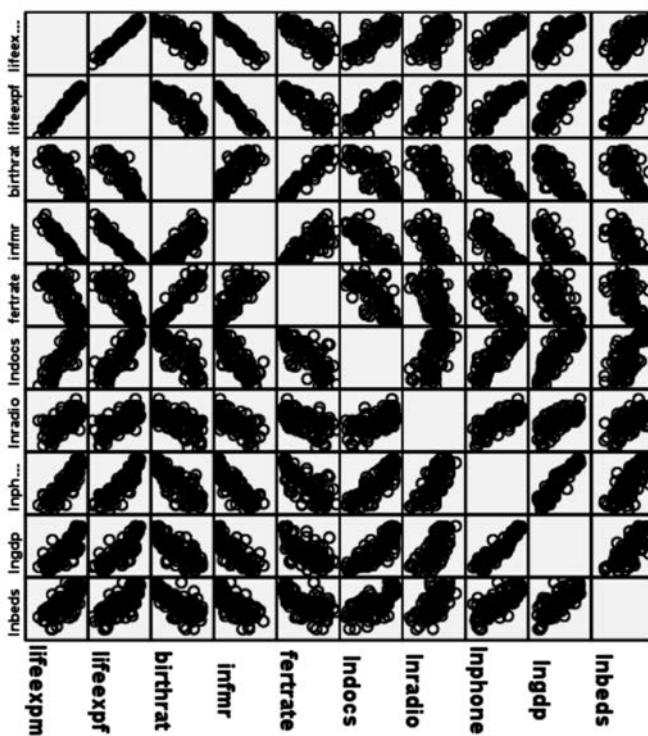


Figure 9.6. Table of Total Variance for One-Component Solution.

Component	Total Variance Explained					
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.161	81.608	81.608	8.161	81.608	81.608
2	.590	5.902	87.510			
3	.372	3.717	91.227			
4	.338	3.384	94.611			
5	.246	2.460	97.071			
6	.168	1.677	98.748			
7	.055	.549	99.297			
8	.033	.332	99.629			
9	.030	.298	99.927			
10	.007	.073	100.000			

Extraction Method: Principal Component Analysis.

Figure 9.7. Communalities for One-Component Solution.

	Initial	Extraction
lifeexprm	1.000	.894
lifeexpf	1.000	.936
birthrat	1.000	.846
infrm	1.000	.894
fertrate	1.000	.779
Indocs	1.000	.830
Inradio	1.000	.573
Inphone	1.000	.899
Indgp	1.000	.839
Inbeds	1.000	.670

Extraction Method: Principal Component Analysis.

Eigenvalue criterion is questionable because two communalities are less than .70.

Figure 9.8. Reproduced Correlations and Residuals for One-Component Solution.

Reproduced Correlations											
	lifeexprm	lifeexpf	birthrat	infmr	fertrate	Indocs	Inradio	Inphone	Ingdp	Inbeds	
Reproduced Correlation	.894 ^a	.915	-.870	-.894	-.834	.861	.716	.896	.866	.774	
	.915	.936 ^a	-.890	-.915	-.854	.881	.732	.917	.886	.792	
	-.870	-.890	.846 ^a	.870	.812	-.838	-.696	-.872	-.843	-.753	
	-.894	-.915	.870	.894 ^a	.835	-.862	-.716	-.897	-.866	-.774	
	-.834	-.854	.812	.835	.779 ^a	-.804	-.668	-.837	-.809	-.723	
	.861	.881	-.838	-.862	-.804	.830 ^a	.690	.864	.835	.746	
	.716	.732	-.696	-.716	-.668	.690	.573 ^a	.718	.694	.620	
	.896	.917	-.872	-.897	-.837	.864	.718	.899 ^a	.869	.776	
	.866	.886	-.843	-.866	-.809	.835	.694	.869	.839 ^a	.750	
	.774	.792	-.753	-.774	-.723	.746	.620	.776	.750	.670 ^a	
Residual ^b	lifeexprm		.073	.037	-.066	.026	.015	-.074	-.019	-.048	-.088
	lifeexpf	.073		.031	-.057	.022	-.003	-.043	-.020	-.045	-.051
	birthrat	.037		.031		-.017	.146	.021	.073	.056	.035
	infmr	.066	-.057	.017			-.004	.030	.063	.034	.055
	fertrate	.026	.022	.146	-.004			.037	.096	.063	.062
	Indocs	.015	-.003	.021	.030	.037			-.064	.001	-.010
	Inradio	-.074	-.043	.073	.063	.096	-.064			.021	.016
	Inphone	-.019	-.020	.056	.034	.063	.001	.021		.073	-.004
	Ingdp	-.048	-.045	.035	.055	.069	-.010	.016	.073		.017
	Inbeds	.088	-.051	.037	.052	.062	-.037	.029	-.004		.017

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 18 (40.0%) nonredundant residuals with absolute values greater than 0.05.

A factor analysis was conducted again, this time lowering the eigenvalue criteria to indicate that two factors should be retained. Varimax rotation was also applied. Because we have lowered the eigenvalue criteria and found both the scree plot and variance criteria to indicate retaining only one component, we will immediately move to the fourth criterion—assessment of residuals (see Figure 9.9). This time, only 11 residuals were greater than 0.05. Consequently, the model has been improved. Because two components were retained, the model was rotated to improve model fit. The table of total variance displays the amount of variance for each component before and after rotation (see Figure 9.10). One should note that factor rotation does not affect the total variance accounted for by the model but does change how the variance is distributed among the retained components. Prior to rotation, the first component accounted for 81.61% and the second for only 5.9%. However, once rotated, the first component accounted for 53.54% and the second for 33.97%. Figure 9.11 displays how variables were loaded into the components after rotation, with coefficients sorted by size. Assessment of component loadings is necessary to name each component. Component 1 was composed of both negative and positive loadings, which somewhat complicates matters. Positive loadings included the variables of male life expectancy, female life expectancy, and the number of doctors. Negative loadings included fertility rate, birth rate, and infant mortality rate. One should also note the variables with the highest loadings were fertility rate (−.879), followed by birth rate (−.858). Because these variables all seemed to relate to the health of one's life, this component will be named *Healthy Lifespan*. Component 2 included all positive loadings and addressed the number of radios, the number of hospital beds, gross domestic product, and the number of phones, respectively. This component will be labeled *Economic Stature*.

Figure 9.9. Reproduced Correlations and Residuals for Two-Component Solution.

Reproduced Correlations											
	lifeexprm	lifeexprf	birthrat	infrm	fertrate	Indocs	Inradio	Inphone	Ingdp	Inbeds	
Reproduced Correlation	.919 ^a	.931	-.903	-.917	-.881	.872	.632	.876	.840	.730	
	.931	.946 ^a	-.911	-.930	-.884	.888	.679	.904	.870	.764	
birthrat	-.903	-.911	.890 ^a	.900	.873	-.852	-.585	-.845	-.808	-.694	
infrm	-.917	-.930	.900	.915 ^a	.877	-.871	-.639	-.878	-.843	-.734	
fertrate	-.881	-.884	.873	.877	.865 ^a	-.824	-.513	-.800	-.761	-.641	
Indocs	.872	.888	-.852	-.871	-.824	.834 ^a	.654	.855	.824	.727	
Inradio	.632	.679	-.585	-.639	-.513	.654	.851 ^a	.785	.780	.767	
Inphone	.876	.904	-.845	-.878	-.800	.855	.785	.915 ^a	.890	.812	
Ingdp	.840	.870	-.808	-.843	-.781	.824	.780	.890	.866 ^a	.796	
Inbeds	.730	.764	-.694	-.734	-.641	.727	.767	.812	.796	.748 ^a	
Residual ^b	lifeexprm		.057	.070	-.043	.073	.005	.010	.001	-.022	-.044
	lifeexprf		.057		.053	-.042	.051	-.010	.011	-.007	-.028
	birthrat		.070	.053		-.047	.084	.035	-.038	.029	.001
	infrm		-.043	-.042	-.047		-.046	.039	-.014	.016	.011
	fertrate		.073	.051	.084		.056	-.059	.026	.020	-.020
	Indocs		.005	-.010	.035	.039		-.029	.010	.001	-.018
	Inradio		.010	.011	-.038	-.014	-.059		-.046	-.071	-.118
	Inphone		.001	-.007	.029	.016	.026	.010	-.046	.052	-.039
	Ingdp		-.022	-.028	.001	.031	.020	.001	-.071	.052	
	Inbeds		-.044	-.023	-.022	.011	-.020	-.018	-.118	-.039	-.029

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 11 (24.0%) nonredundant residuals with absolute values greater than 0.05.

Figure 9.10. Table of Total Variance for Two-Component Solution.

Component	Total Variance Explained								
	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.161	81.608	81.608	8.161	81.608	81.608	5.354	53.542	53.542
2	.590	5.902	87.510	.590	5.902	87.510	3.397	33.968	87.510
3	.372	3.717	91.227						
4	.338	3.384	94.611						
5	.246	2.460	97.071						
6	.168	1.677	98.748						
7	.055	.549	99.297						
8	.033	.332	99.629						
9	.030	.298	99.927						
10	.007	.073	100.000						

Extraction Method: Principal Component Analysis.

Figure 9.11. Factor Loadings for Rotated Components Sorted by Size.

	Component	
	1	2
fertrate	-.879	-.305
birthrat	-.858	-.393
lifeexprm	.846	.450
infmr	-.839	-.461
lifeexpf	.829	.509
Indocs	.763	.502
Inradio	.279	.879
Inbeds	.480	.719
Ingdp	.627	.688
Inphone	.675	.678

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Loadings for Component 1.

Loadings for Component 2.

Writing Up Results

Once again, the results narrative should always describe the elimination of participants and transformation of variables. The summary should then describe the type of factor analysis conducted and indicate if any rotation method was utilized. The interpretation process is then presented in conjunction with the results. In general, it is necessary to indicate only the number of factors retained and the criteria that led to that decision. The results of the final solution are then presented. One should summarize *each component* by presenting the following: the percentage of variance, the number and names of variables loaded into the component, and the component loadings. This is often displayed in table format, depending on the number of components and variables. Finally, the researcher indicates the names of components. The following summary applies the output presented in Figures 9.5 through 9.11.

Factor analysis was conducted to determine what, if any, underlying structure exists for measures on the following 10 variables: male life expectancy (*lifeexprm*), female life expectancy (*lifeexpf*), births per 1,000 individuals (*birthrat*), infant mortality rate (*infmr*), fertility rate per woman (*fertrate*), doctors per 10,000 individuals (*docs*), radios per 100 individuals (*radio*), telephones per 100 individuals (*phone*), gross domestic product (*gdp*), and hospital beds per 10,000 individuals (*beds*). Prior to the analysis, evaluation of linearity and normality led to the natural log transformation of *docs*, *radios*, *phones*, *gdp*, and *beds*. Principal components analysis was conducted utilizing a varimax rotation. The initial analysis retained only one component. Four criteria were used to determine the appropriate number of components to retain: eigenvalue, variance, scree plot, and residuals. Criteria indicated that retaining two components should be investigated. Thus, principal components analysis was conducted to retain two components and apply the varimax rotation. Inclusion of two components increased the model fit as it decreased the number of residuals exceeding the 0.05 criterion.

After rotation, the first component accounted for 53.54% and the second for 33.97%. Component 1 included items with both negative and positive loadings. Positive loadings included the variables of male life expectancy, female life expectancy, and the number of doctors. Negative loadings included fertility rate, birth rate, and infant mortality. Items with the highest loadings were

fertility rate and birth rate. Component 1 was named *Healthy Lifespan*. Component 2 included the number of radios, the number of hospital beds, gross domestic product, and the number of phones, respectively. This component was labeled *Economic Stature*. (See Table 1.)

Table 1
Component Loadings

	Loading
Component 1: Healthy Lifespan	
Fertility rate	-.879
Birth rate per 1,000 individuals	-.858
Male life expectancy	.846
Infant mortality	-.839
Female life expectancy	.829
Number of doctors per 10,000 individuals	.763
Component 2: Economic Stature	
Number of radios per 100 individuals	.879
Number of hospital beds per 10,000 individuals	.719
Gross domestic product	.688
Number of phones per 100 individuals	.678

SECTION 9.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example of the process of factor analysis. This process includes the development of research questions, data screening, test methods, interpretation of output, and presentation of results. The example that follows utilizes the data set *schools-b.sav* from the website that accompanies this book (see p. *xiii*).

Problem

We are interested in determining what, if any, underlying structures exist for measures on the following 12 variables: % graduating in 1993 (*grad93*); % graduating in 1994 (*grad94*); average ACT score in 1993 (*act93*); average ACT score in 1994 (*act94*); 10th grade average score in 1993 for math (*math93*), reading (*read93*), and science (*scienc93*); % meeting or exceeding state standards in 1994 for math (*math94me*), reading (*read94me*), and science (*sci94me*); and % limited English proficiency in 1993 (*lep93*) and 1994 (*lep94*).

Methods, SPSS “How To,” Output, and Interpretation

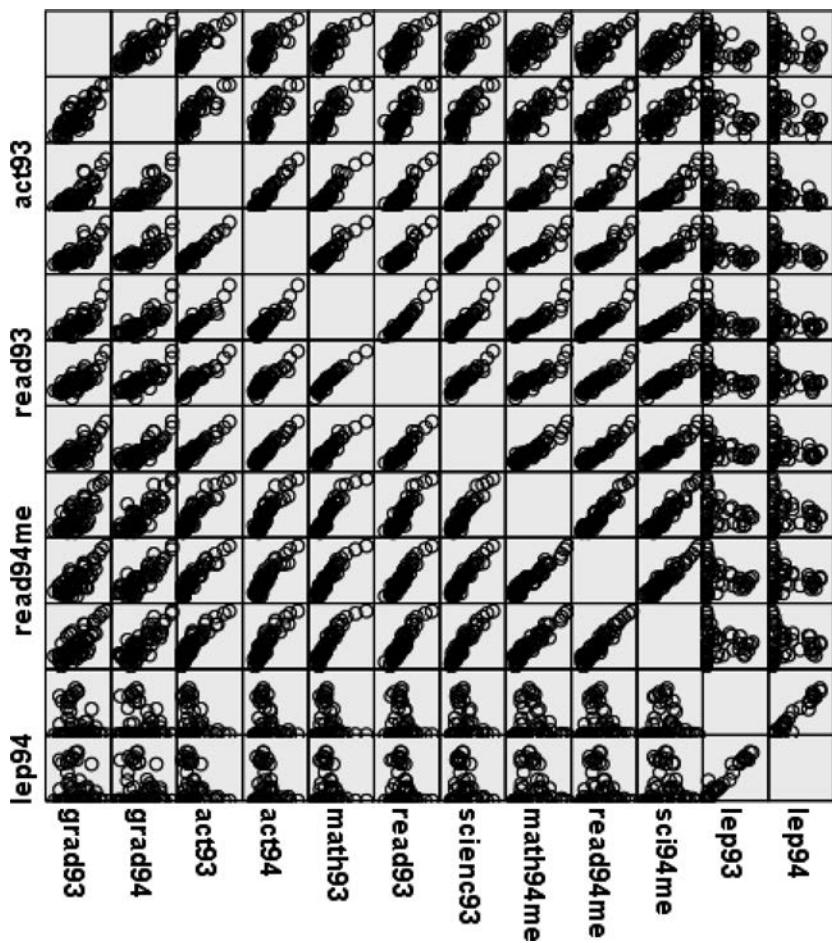
Because factor analysis requires a great deal of interpretation throughout the process of analysis, we have combined the discussion of methods, output, and interpretation in this section. Data were evaluated to screen for outliers and assess normality and linearity. Using Mahalanobis distance, two outliers (cases 37 and 64) were found (see Figure 9.12). Therefore, all cases in which the Mahalanobis value exceeded the chi-square criterion needed to be eliminated using **Select Cases, If MAH_1 \leq 32.909**. A scatterplot matrix revealed fairly normal distributions and linear relationship among variables (see Figure 9.13).

Figure 9.12. Outliers for Mahalanobis Distance.

Extreme Values		
		Case Number
MAH_1	Highest	Value
	1	37 40.30603
	2	64 38.07910
	3	46 23.28122
	4	6 23.09387
	5	39 22.68936
	Lowest	16 2.35371
	2	27 3.40542
	3	19 3.47370
	4	61 3.95236
	5	22 4.13483

Outliers
exceeding the
 $\chi^2(12) = 32.909$
at $p = .001$.

Figure 9.13. Scatterplot Matrix.



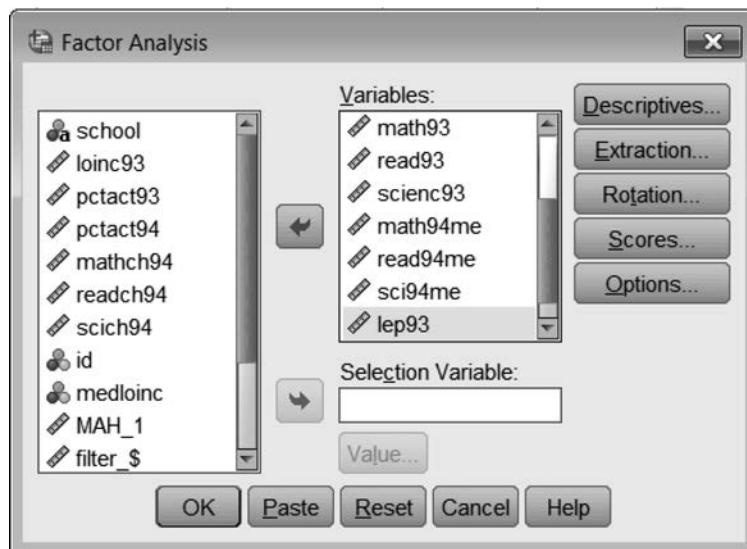
A factor analysis was then conducted using **Dimension Reduction**, which utilized the eigenvalue criteria and varimax rotation. To open the **Factor Analysis** dialog box, select the following:

Analyze
Dimension Reduction
Factor

Factor Analysis dialog box (Figure 9.14)

Select each variable to be included in the analysis and move each to the **Variables** box. Then click **Descriptives**.

Figure 9.14. Factor Analysis Dialog Box.



Factor Analysis: Descriptives dialog box (Figure 9.15)

Several descriptive statistics are provided in this dialog box. Under **Statistics**, two options are provided: **Univariate descriptives** and **Initial solution**. **Univariate descriptives** presents the means and standard deviations for each variable analyzed. **Initial solution** is selected by default and will present the initial communalities, eigenvalues, and percent accounted for by each factor. For our example, we utilized only **Initial solution**. Under **Correlation Matrix**, the following options are frequently used:

Coefficients—Presents original correlation coefficients of variables.

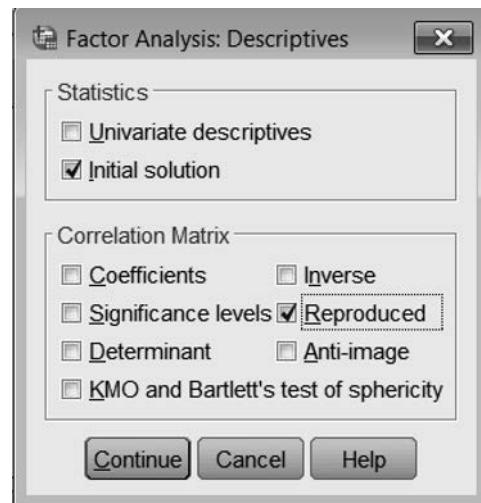
Significance levels—Indicates p values of each correlation coefficient.

KMO and Bartlett's test of sphericity—Tests both multivariate normality and sampling adequacy.

Reproduced—Presents reproduced correlation coefficients and residuals (difference between original and reproduced coefficients).

Our example utilized only the **Reproduced** option. After selecting the descriptive options, click **Continue**. Click **Extraction**.

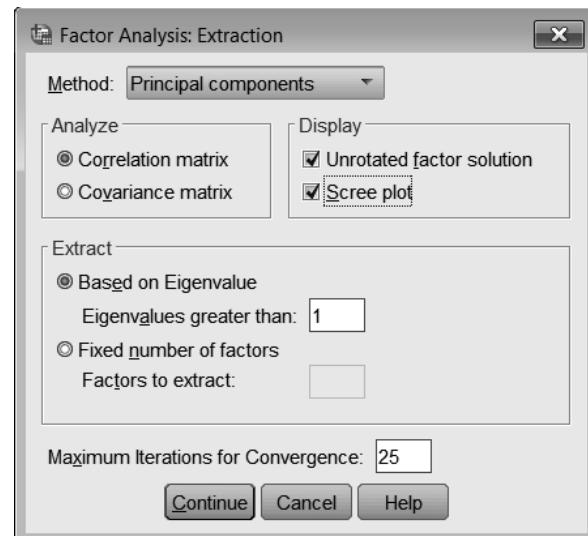
Figure 9.15. Factor Analysis: Descriptives Dialog Box.



Factor Analysis: Extraction dialog box (Figure 9.16)

For Method, select **Principal components**. Under Analyze, check **Correlation matrix**. Under Display, select **Unrotated factor solution** and **Scree plot**. Under Extract, the eigenvalue criterion of 1 is the default. Utilize the default unless a previous analysis indicated that more components should be retained, in which case you would indicate the number of factors. Next, click **Continue**. Click **Rotation**.

Figure 9.16. Factor Analysis: Extraction Dialog Box.



Factor Analysis: Rotation dialog box (Figure 9.17)

Select the rotation method you prefer. The rotation methods available are described as follows:

Varimax—Orthogonal method that minimizes factor complexity by maximizing variance for each factor.

Direct Oblimin—Oblique method that simplifies factors by minimizing cross products of loadings.

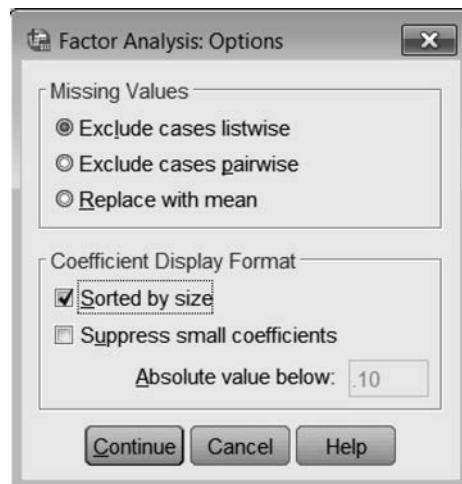
Quartimax—Orthogonal method that minimizes factor complexity by maximizing variance loadings on each variable.

Equamax—Orthogonal method that combines both Varimax and Quartimax procedures.

Promax—Oblique method that rotates orthogonal factors to oblique positions.

Our example utilized **Varimax**. If a rotation method is indicated, check **Rotated solution** under **Display**. Click **Continue**, then **Scores**.

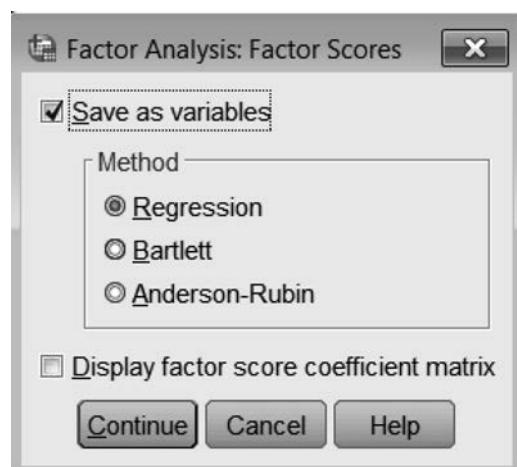
Figure 9.17. Factor Analysis: Rotation Dialog Box.



Factor Analysis: Factor Scores dialog box (Figure 9.18)

If you will be using the generated factors in future analyses, you will need to save factor scores. To do so, check **Save as variables** and utilize the default method of **Regression**. Click **Continue**, then **Options**.

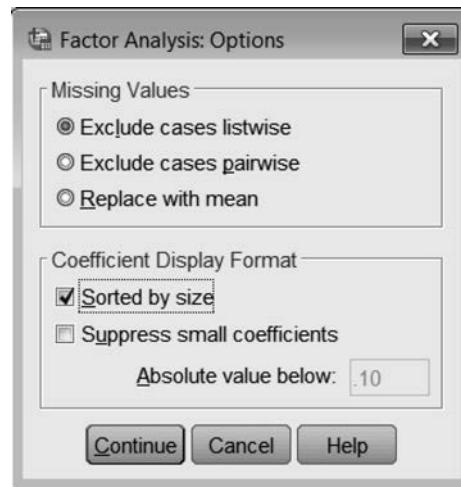
Figure 9.18. Factor Analysis: Factor Scores Dialog Box.



Factor Analysis: Options dialog box (Figure 9.19)

Under **Coefficient Display Format**, check **Sorted by size**. This will help in reading the Component Matrix. Click **Continue**. Click **OK**.

Figure 9.19. Factor Analysis: Options Dialog Box.



The output generated from this analysis is presented in Figures 9.20 to 9.23. Applying the four methods of interpretation, we first examine the eigenvalues in the table of total variance (see Figure 9.20). Two components were retained because they have eigenvalues greater than 1. In this example, the application of the eigenvalue criterion seems appropriate because the number of variables is less than 30 and all communalities are greater than .70 (see Figure 9.21). Evaluation of variance is done by referring back to Figure 9.20. After rotation, the first component accounts for 74.73% of the total variance in the original variables, while the second component accounts for 17.01%. The scree plot was then assessed and indicates that the eigenvalues after three components level off (see Figure 9.22). Evaluation of residuals indicates that only five residuals are greater than 0.05 (see Figure 9.23). Although the scree plot suggests that the inclusion of the third component may improve the model, the residuals reveal that any model improvement would be minimal. Consequently, two components were retained.

Figure 9.20. Table of Total Variance for Two-Component Solution.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.969	74.745	74.745	8.969	74.745	74.745	8.968	74.733	74.733
2	2.040	17.000	91.746	2.040	17.000	91.746	2.042	17.013	91.746
3	.407	3.390	95.136						
4	.205	1.705	96.841						
5	.136	1.132	97.973						
6	.073	.604	98.577						
7	.056	.470	99.047						
8	.047	.388	99.435						
9	.028	.232	99.667						
10	.022	.183	99.850						
11	.012	.100	99.950						
12	.006	.050	100.000						

Extraction Method: Principal Component Analysis.

Figure 9.21. Communalities.

	Initial	Extraction
grad93	1.000	.701
grad94	1.000	.800
act93	1.000	.936
act94	1.000	.905
math93	1.000	.964
read93	1.000	.932
scienc93	1.000	.948
math94me	1.000	.944
read94me	1.000	.950
sci94me	1.000	.945
lep93	1.000	.993
lep94	1.000	.991

Extraction Method: Principal Component Analysis.

Eigenvalue criterion is acceptable because all communalities are greater than .70, and less than 30 factors are analyzed.

Figure 9.22. Scree Plot.

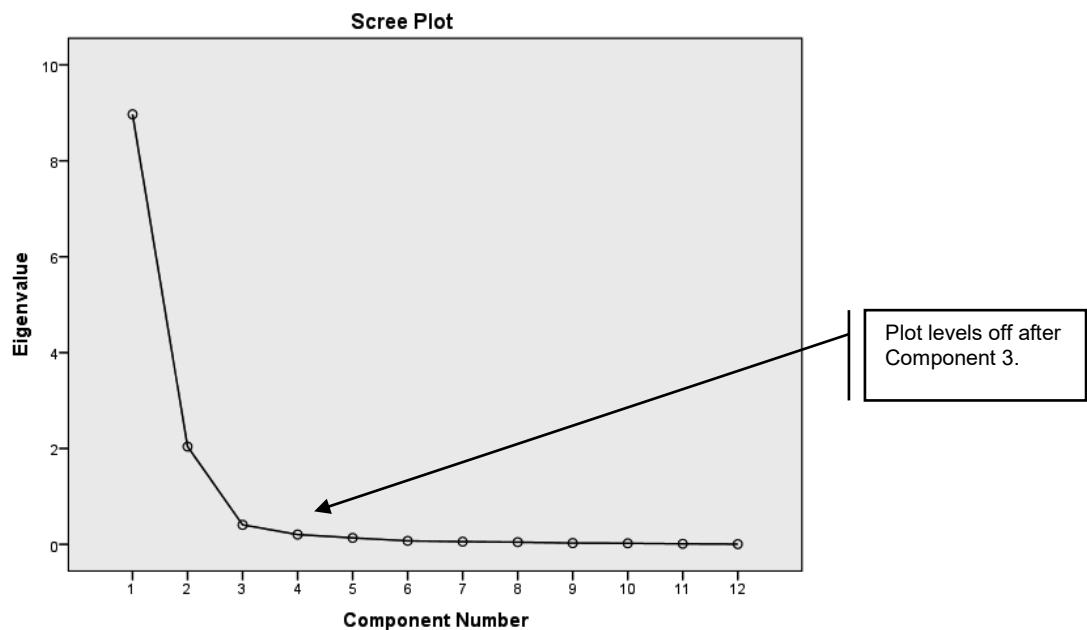


Figure 9.23. Reproduced Correlations and Residuals.

Reproduced Correlations														
	grad93	grad94	act93	act94	math93	read93	scienc93	math94me	read94me	sci94me	lep93	lep94		
Reproduced Correlation	.701 ^a	.742	.804	.783	.806	.796	.794	.780	.792	.778	-.182	-.172		
grad94		.800 ^a	.865	.850	.876	.863	.867	.859	.867	.858	-.075	-.065		
act93	.804		.865	.936 ^a	.919	.947	.933	.937	.927	.936	.926	-.097	-.085	
act94	.783	.850		.919	.905 ^a	.934	.918	.925	.919	.926	.919	-.032	-.020	
math93	.806	.876	.947		.934	.964 ^a	.947	.955	.950	.956	.950	-.022	-.010	
read93	.796	.863	.933	.918		.947	.932 ^a	.939	.932	.939	.931	-.042	-.030	
scienc93	.794	.867	.937	.925	.955		.939	.948 ^a	.944	.949	.944	.006	.018	
math94me	.780	.859	.927	.919	.950	.932		.944	.944 ^a	.946	.944	.068	.079	
read94me	.792	.867	.936	.926	.956	.939	.949		.946	.950 ^a	.946	.024	.036	
sci94me	.778	.858	.926	.919	.950	.931	.944	.946		.946	.945 ^a	.076	.087	
lep93	-.182	-.075	-.097	-.032	-.022	-.042	.006	.068	.024	.076	.993 ^a	.992		
lep94	-.172	-.065	-.085	-.020	-.010	-.030	.018	.079	.036	.087	.992	.991 ^a		
Residual ^b														
grad93		.064		-.024	-.054	-.032	-.015	-.060	-.047	-.029	-.050	.029	.034	
grad94		.064			-.040	-.069	-.023	-.024	-.045	-.003	-.010	-.027	.004	.018
act93	-.024		-.040			.048	-.003	-.022	-.015	-.011	-.022	-.011	.004	-.002
act94	.054		.069				-.010	-.023	-.037	-.010	-.016	-.009	.002	-.010
math93	-.032		.023					.021	-.007	.011	-.004	.004	-.005	.003
read93	-.015		.024					.022	-.006	-.018	.006	.009	-.005	.001
scienc93	.060		.045					.037	-.006	-.006	-.000	.007	-.004	-.008
math94me	.047		.003					.011	-.018	-.006	.013	.007	-.006	-.009
read94me	-.029		.010					.016	-.004	.006	.013	.005	-.003	-.008
sci94me	.050		.027					.011	-.009	.004	.009	.005	-.010	-.007
lep93	.029		.004					.002	-.005	-.005	-.006	-.003	-.010	.000
lep94	.034		.018					.010	-.003	.001	-.008	-.009	-.007	.000

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 5 (7.0%) nonredundant residuals with absolute values greater than 0.05.

The next step was to interpret each component. Figure 9.24 presents the factor loadings for the rotated components. Component 1 consisted of 10 of the 12 variables: *scienc93*, *read94me*, *math93*, *sci94me*, *read93*, *math94me*, *act93*, *act94*, *grad94*, and *grad93*. These variables had positive loadings and addressed *Academic Achievement*. The second component included the remaining two variables of percent of limited English proficiency in 1994 (*lep94*) and 1993 (*lep93*). Both variables had positive loadings. Component 2 was named *Limited English Proficiency*.

Figure 9.24. Factor Loadings for Rotated Components.

Rotated Component Matrix ^a														
	Component													
	1	2												
math93	.982	-.005												
read94me	.974	.041												
scienc93	.973	.023												
math94me	.968	.084												
sci94me	.968	.093												
read93	.965	-.026												
act93	.964	-.081												
act94	.951	-.016												
grad94	.892	-.060												
grad93	.820	-.169												
lep93	-.017	.996												
lep94	-.005	.996												

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Presentation of Results

The following summary applies the output from Figures 9.20 through 9.24.

Factor analysis was conducted to determine what, if any, underlying structures exist for measures on the following 12 variables: % graduating in 1993 (*grad93*) and in 1994 (*grad94*); average ACT score in 1993 (*act93*) and in 1994 (*act94*); 10th grade average score in 1993 for math (*math93*), reading (*read93*), and science (*scienc93*); % meeting or exceeding state standards in 1994 for math (*math94me*), reading (*read94me*), and science (*sci94me*); and % limited English proficiency in 1993 (*lep93*) and in 1994 (*lep94*). Prior to analysis, two outliers were eliminated. Principal components analysis was conducted utilizing a varimax rotation. The analysis produced a two-component solution, which was evaluated with the following criteria: eigenvalue, variance, scree plot, and residuals. Criteria indicated a two-component solution was appropriate.

After rotation, the first component accounted for 74.73% of the total variance in the original variables, while the second component accounted for 17.01%. Table 2 presents the loadings for each component. Component 1 consisted of 10 of the 12 variables: *math93*, *read94me*, *scienc93*, *math94me*, *sci94me*, *read93*, *act93*, *act94*, *grad94*, and *grad93*. These variables had positive loadings and addressed *Academic Achievement*. The second component included the remaining two variables of % of limited English proficiency in 1993 (*lep93*) and in 1994 (*lep94*). Both variables had positive loadings. Component 2 was labeled *Limited English Proficiency*. (See Table 2.)

Table 2
Component Loadings

	Loading
Component 1: Academic Achievement	
10th grade average math score (1993)	.982
% meeting/exceeding reading standards (1994)	.974
10th grade average science score (1993)	.973
% meeting/exceeding math standards (1994)	.968
% meeting/exceeding science standards (1994)	.968
10th grade average reading score (1993)	.965
Average ACT score (1993)	.964
Average ACT score (1994)	.951
% graduating (1994)	.892
% graduating (1993)	.820
Component 2: Limited English Proficiency	
% Limited English Proficiency (1993)	.996
% Limited English Proficiency (1994)	.996

SUMMARY

Factor analysis is a technique used to identify factors that explain common variance among variables. This statistical method is often used to reduce data by grouping variables that measure a common construct. Principal components analysis is one of the most commonly used methods of extraction because this method will evaluate all sources of variability for each variable. Factors or components can also be rotated to make the components more interpretable. Orthogonal rotation methods (i.e., varimax, quartimax, equamax) result in uncorrelated factors and are the most frequently used methods. Oblique rotation methods (i.e., oblimin, promax, orthoblique) result in factors being correlated with each other.

Because principal components analysis is typically exploratory, the researcher must determine the appropriate number of components to retain. Four criteria are used in this decision-making process:

1. Eigenvalue—Components with eigenvalues greater than 1 should be retained. This criterion is fairly reliable when the number of variables is < 30 and communalities are $> .70$ or the number of individuals is > 250 and the mean communality is $\geq .60$.
2. Variance—Retain components that account for at least 70% total variability.
3. Scree Plot—Retain all components within the sharp descent, before eigenvalues level off. This criterion is fairly reliable when the number of individuals is > 250 and communalities are $> .30$.
4. Residuals—Retain the components generated by the model if only a few residuals (the difference between the empirical and the reproduced correlations) exceed .05. If several reproduced correlations differ, you may want to include more components.

Once the appropriate number of components to retain has been determined, the researcher must then interpret/name the components by evaluating the types of variables included in each factor, the strength of factor loadings, and the direction of factor loadings. Figure 9.25 provides a checklist for conducting factor analysis.

KEYWORDS

- Bartlett's sphericity test
- communalities (h_i)
- confirmatory factor analysis
- eigenvalue
- exploratory factor analysis
- extraction
- factor analysis
- factor correlation matrix
- factor loadings
- factor scores
- first principal component
- oblique rotation
- orthogonal rotation
- principal components
- principal components analysis
- rotation
- scree plot

Figure 9.25. Checklist for Conducting Factor Analysis.

I. Screen Data

- a. Missing Data?
 - b. Multivariate Outliers?
 - Run preliminary Regression to calculate Mahalanobis distance.
 1. **Analyze... Regression... Linear.**
 - Identify a variable that serves as a case number and move to **Dependent Variable** box.
 - Identify all appropriate quantitative variables and move to **Independent(s)** box.
 2. **Save.**
 - Check **Mahalanobis** under **Distances**.
 3. **Continue, OK.**
 4. Determine chi-square (χ^2) critical value at $p < .001$.
 - Conduct **Explore** to test outliers for Mahalanobis chi-square (χ^2).
 1. **Analyze... Descriptive statistics... Explore.**
 - Move **MAH_1** to **Dependent Variable** box.
 - Leave **Factor** box empty.
 2. **Statistics.**
 - Check **Outliers**.
 3. **Continue, OK.**
 - Delete outliers for participants when χ^2 exceeds critical χ^2 at $p < .001$.
 - c. Linearity and Normality?
 - Create Scatterplot Matrix of all model variables.
 - If scatterplot shapes not close to elliptical shapes → reevaluate univariate normality and consider transformations.

II. Conduct Factor Analysis

- a. Run Factor Analysis using **Dimension Reduction**.
 1. **Analyze... Dimension Reduction... Factor.**
 - Move each studied variable to the **Variables** box.
 2. **Descriptives.**
 - Check: **Initial solution** and **Reproduced**.
 - **Continue.**
 3. **Extraction.**
 - Check **Correlation matrix**, **Unrotated factor solution**, **Scree plot**, and **Based on Eigenvalue**.
 - **Continue.**
 4. **Rotation.**
 - Check **Varimax** and **Rotated solution**.
 - **Continue.**
 5. **Scores.**
 - Check **Save as variables** and **Regression**.
 - **Continue.**
 6. **Options.**
 - Check **Sorted by size**.
 7. **Continue, OK.**
- b. Determine appropriate number of components to retain.
 1. Eigenvalue—Components with eigenvalues greater than 1 should be retained. This criterion is fairly reliable when the number of variables is < 30 and communalities are $> .70$ or the number of individuals is > 250 and the mean communality is $\geq .60$.
 2. Variance—Retain components that account for at least 70% total variability.
 3. Scree Plot—Retain all components within the sharp descent, before eigenvalues level off. This criterion is fairly reliable when the number of individuals is > 250 and communalities are $> .30$.
 4. Residuals—Retain the components generated by the model if only a few residuals (the difference between the empirical and the reproduced correlations) exceed 0.05. If several reproduced correlations differ, you may want to include more components.

Figure 9.25. Checklist for Conducting Factor Analysis. (continued)

- c. Conduct factor analysis again if more components should be retained.
- d. Interpret components.
 1. Evaluate the types of variables loaded into each component.
 2. Note the strength and direction of loadings.
 3. Label component accordingly.

III. Summarize Results

- a. Describe any data elimination or transformation.
- b. Describe the initial model.
- c. Describe the criteria used to determine the number of components to retain.
- d. Summarize the components generated by narrating the variables loaded into each component, the strength and direction of loadings, the component labels, and the percentage of variance.
- e. Create a table that summarizes each component (report component loadings).
- f. Draw conclusions.

Exercises for Chapter 9

The following exercises seek to determine what underlying structure exists among the following variables in *profile-a.sav*: highest degree earned (*degree*), hours worked per week (*hrs1*), job satisfaction (*satjob*), years of education (*educ*), hours per day watching TV (*tvhours*), general happiness (*happy*), degree to which life is exciting (*life*), and degree to which the lot of the average person is getting worse (*anomia5*).

1. The following output was generated for the initial analysis. Varimax rotation was utilized.

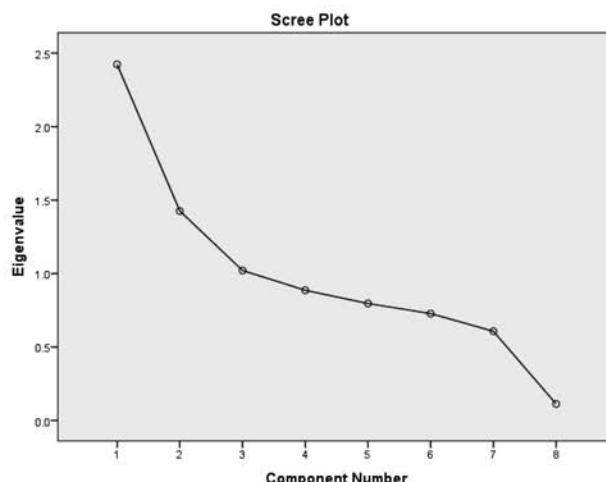
Communalities		
	Initial	Extraction
degree	1.000	.933
hrs1	1.000	.602
satjob	1.000	.447
educ	1.000	.939
tvhours	1.000	.556
happy	1.000	.576
life	1.000	.500
anomia5	1.000	.317

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.423	30.293	30.293	2.423	30.293	30.293	1.879	23.488	23.488
2	1.426	17.822	48.115	1.426	17.822	48.115	1.734	21.676	45.165
3	1.021	12.760	60.875	1.021	12.760	60.875	1.257	15.710	60.875
4	.886	11.077	71.952						
5	.796	9.955	81.907						
6	.728	9.094	91.001						
7	.607	7.589	98.590						
8	.113	1.410	100.000						

Extraction Method: Principal Component Analysis.



Reproduced Correlations

	degree	hrs1	satjob	educ	tvhours	happy	life	anomia5
Reproduced Correlation	degree	.933 ^a	.176	-.039	.935	-.239	-.119	.230
	hrs1	.176	.602 ^a	-.239	.194	-.576	-.077	.141
	satjob	-.039	-.239	.447 ^a	-.062	.214	.469	-.436
	educ	.935	.194	-.062	.939 ^a	-.255	-.142	.252
	tvhours	-.239	-.576	.214	-.255	.556 ^a	.066	-.136
	happy	-.119	-.077	.469	-.142	.066	.576 ^a	-.526
	life	.230	.141	-.436	.252	-.136	-.526	.500 ^a
	anomia5	.118	-.049	-.297	.131	.047	-.412	.371
Residual ^b	degree		.004	-.068	-.050	.032	-.004	-.034
	hrs1		.004		.104	.011	.361	-.046
	satjob		-.068		.104		-.037	.151
	educ		-.050		.011		.026	-.029
	tvhours		.032		.361		-.014	-.099
	happy		-.004		-.031		.014	.159
	life		-.034		.151		-.012	-.217
	anomia5		-.037		.112		.177	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 12 (42.0%) nonredundant residuals with absolute values greater than 0.05.

- a. Assess the eigenvalue criterion. How many components were retained? Is the eigenvalue appropriate, considering the number of factors and the communalities?

- b. Assess the variance explained by the retained components. What is the total variability explained by the model? Is this amount adequate?
 - c. Assess the scree plot. At which component does the plot begin to level off?
 - d. Assess the residuals. How many residuals exceed the 0.05 criterion?
 - e. Having applied the four criteria, do you believe the number of components retained in this analysis is appropriate? If not, what is your recommendation?
2. Assume that you believe four components should be retained from the analysis in the previous exercise. Conduct a factor analysis with varimax rotation (be sure to retain four components).
- a. Evaluate each of the four criteria. Has the model fit improved? Explain.
 - b. Provide two alternatives for improving the model.

CHAPTER 10

DISCRIMINANT ANALYSIS

STUDENT LEARNING OBJECTIVES

After studying Chapter 10, students will be able to:

1. Describe how discriminant analysis is often seen as the reverse of MANOVA.
2. Discuss similarities and differences between discriminant analysis, multiple regression, and principal components analysis.
3. Explain what a discriminant function is.
4. Describe what is being measured by a canonical correlation.
5. Summarize the logic behind the procedures in a discriminant analysis.
6. Define prior probability.
7. Develop research questions appropriate for discriminant analysis.
8. Design and conduct a discriminant analysis to classify subjects into a specific number of groups by following the appropriate SPSS guidelines provided.

In Chapters 10 and 11, we give attention to the fourth, and final, group of advanced/multivariate statistical techniques. Recall from Chapter 2 that the techniques presented in these chapters have as their primary purpose the prediction of group membership—sometimes referred to as *classification*—for individuals in the sample. This is accomplished by identifying specific IVs that serve as the best predictors of membership in a particular group, as defined by the DV. In this chapter, we present discussion of discriminant analysis, which seeks to identify a combination of IVs, measured at the interval level, that best predicts membership in a particular group, as measured by a categorical DV.

SECTION 10.1 PRACTICAL VIEW

Purpose

Discriminant analysis has two basic purposes: (1) to describe major differences among groups following a MANOVA analysis and (2) to classify participants into groups based on a combination of measures (Stevens, 2001). Discriminant analysis was originally developed in the 1930s with the second purpose in mind—to classify participants into one of two clearly defined groups. The technique was later expanded in order to classify participants into any number of groups (Pedhazur, 1982). However, of late, its main purpose has focused on the description of group differences (Tatsuoka, 1988). Both purposes are described briefly in the ensuing paragraphs.

If the discriminant analysis is being used to further describe the differences among two or more groups—often referred to as *descriptive discriminant analysis* (Stevens, 2001)—the analysis essentially

involves breaking down the “between” association from a MANOVA into its additive portions. This is accomplished through the identification of uncorrelated linear combinations of the original variables. These uncorrelated linear combinations are called **discriminant functions**. Groups are subsequently differentiated along one, two, or possibly several dimensions. There exists some carryover here from factor analysis (see Chapter 9) in that the researcher is responsible for providing a meaningful name for the various discriminant functions, identified as being statistically significant.

In contrast, if the discriminant analysis is being used for purposes of prediction or classification, discriminant functions are again obtained, but their interpretation changes slightly. Instead of trying to describe dimensions on which groups differ, the goal is to determine dimensions that serve as the basis for reliably—and accurately—classifying participants into groups. Because we have discussed several analytic techniques for investigating the existence of differences among groups, it is our intent in this chapter to focus on the use of discriminant analysis for classification purposes.

Regardless of the purpose for which it is being used, discriminant analysis is often seen as the reverse of MANOVA. You will recall that in MANOVA, the researcher takes two or more groups and compares their scores on a combination of DVs in an attempt to discover whether or not there exist significant group differences. However, in discriminant analysis, this process is reversed (Sprinthall, 2007). In MANOVA, the IVs are the grouping variables and the DVs are the predictors; whereas, in discriminant analysis, the DVs serve as the grouping variables and the IVs are the predictors (Tabachnick & Fidell, 2007). If the goal of the analysis is to describe group differences, the researcher would determine the number of dimensions (i.e., the *discriminant functions*) that maximize the differences among the groups in question. In contrast, if prediction is the goal of analysis, the researcher might use discriminant function scores in order to predict from which group participants came. This procedure can then be used to predict membership in a particular group for new or future participants from the same population.

Students often have difficulty envisioning practical uses for discriminant analysis, especially those used for classification purposes. We offer several examples from different fields of study where classifying participants may be of interest. The following is a brief list of possible examples:

- The dean of a college of education wants to determine a process by which she can identify those students who are likely to succeed as educators and those who are not.
- Based on a comprehensive set of psychological measures, a psychologist wants to classify patients into one of several categories.
- A special educator wants to reliably classify exceptional students as learning disabled or mentally handicapped.
- A credit card company wants to determine a method of accurately predicting an individual’s level of risk (e.g., high, moderate, low) as a potential credit customer.
- A psychiatrist wants to classify patients with respect to their level of anxiety (e.g., low and high) in social settings.

Now, let us develop our own working example, to which we will refer in this chapter.¹ Suppose that we wanted to determine a nation’s status in terms of its level of development based on several measures. Specifically, those measures consist of

- percentage of the population living in urban areas (*urban*),
- gross domestic product per capita (*gdp*),
- male life expectancy (*lifeexpm*),

¹ Data set *country-d.sav* was used for this example.

- female life expectancy (*lifeexpf*), and
- infant mortality rate per 1,000 live births (*infmr*).

Linear combinations of these five variables, which would allow us to predict a country's status as a developing nation (*develop*)—either (0) developed nation or (1) developing nation—will be determined and subsequently discussed.

The interpretation of results obtained from a discriminant analysis is fairly straightforward because they tend to parallel results that we have seen in previous analysis techniques. The main result obtained from a discriminant analysis is the summary of the discriminant functions. These functions are similar to factors or components in factor analysis—in the case where there is more than one, they represent uncorrelated linear combinations of the IVs. These combinations basically consist of regression equations—raw scores on each original variable are multiplied by assigned weights and then summed together in order to obtain a *discriminant score*, which is analogous to a factor score (discriminant functions will be discussed in greater detail in Section 10.3). The analysis returns several indices, many of which we have seen in previous chapters. An eigenvalue and percentage of variance explained are provided for each discriminant function. These values are interpreted in similar fashion to their analogous counterparts in a factor or principal components analysis. A value called the ***canonical correlation*** is also reported. This value is equivalent to the correlation between the discriminant scores and the levels of the dependent variable. A high value for this correlation indicates a function that discriminates well between participants; in other words, it will likely perform well in terms of classifying participants into groups (levels of the DV). It is important to realize that when the canonical correlation is less than perfect (i.e., not equal to 1.0), some degree of error will be reflected in the assignment of individual participants to groups (Williams, 1992). This is an important point that we will address for a moment.

In addition, we are provided with a test of the significance of each of the discriminant functions—specifically, we are provided a value for Wilks' Lambda (Λ), similar to that supplied in a MANOVA output. The significance of each discriminant function is tested using a chi-square criterion—statistical significance indicates that the function discriminates well based on levels of the DV.

Another portion of the results vital to the analysis involves the actual coefficients for each discriminant function. Similar to regression coefficients in multiple regression, these coefficients serve as the weights assigned by the computer to the various original variables in the analysis. Both unstandardized and standardized coefficients are provided for interpretation by the researcher. Unstandardized coefficients are the basis for the calculation of discriminant scores, a point that will be discussed further in Section 10.3. However, because they represent various measurement scales inherent in the original variables, they cannot be used to assess the relative contributions of individual variables to the discriminant function(s) (Williams, 1992). The standardized coefficients must be used for this purpose.

These standardized coefficients, along with coefficients presented in a structure matrix, are similar to factor loadings in factor analysis. The sizes of the standardized coefficients and the function loadings indicate the degree of relationship between each variable and the discriminant function. Recall that the relative sizes and directions of these coefficients are used to attach labels to the functions for greater ease in interpretation, especially interpretation of the resultant discriminant scores. As in factor analysis, the meaning of the function is inferred from the researcher, from the pattern of correlations, or loadings, between the function and the predictor IVs (Tabachnick & Fidell, 2007). Figure 10.1 presents the ***standardized*** and ***unstandardized discriminant function coefficients***, as well as structure coefficients, for our working example. Notice that the resulting single function is a bipolar discriminant function. The steps for data screening, transformation, and analysis will be discussed fully in Section 10.3.

Also, one should notice that—based on the absolute values of the coefficients—the order of importance of the predictors differs when comparing the standardized and structure coefficients. *It should be noted that examination of these two indices may provide different results.* In the matrix of standardized coefficients, the most important predictor is female life expectancy (1.780), followed by male life expectancy (−1.500), gross domestic product (1.203), percent urban (−.548), and infant mortality (.181). However, with respect to the correlation or structure coefficients, the most important predictor is percent urban (.904), followed by gross domestic product (.637), male life expectancy (−.616), female life expectancy (.575), and infant mortality (.453).

Figure 10.1. Unstandardized (a), Standardized (b), and Structure Coefficients (c) for “Classification as Developing Nation” Example.

(a) Unstandardized Canonical Discriminant Function Coefficients

	Function
	1
Percent urban 1992	−.025
GDP per capita	1.135
Male life expectancy 1992	−.180
Female life expectancy 1992	.196
Infant mortality rate 1992 (per 1,000 live births)	.005
(Constant)	−9.175

(b) Standardized Canonical Discriminant Function Coefficients

	Function
	1
Percent urban 1992	−.548
GDP per capita	1.203
Male life expectancy 1992	−1.500
Female life expectancy 1992	1.780
Infant mortality rate 1992 (per 1,000 live births)	.181

(c) Structure Coefficients

	Function
	1
Percent urban 1992	.904
GDP per capita	.637
Male life expectancy 1992	−.616
Female life expectancy 1992	.575
Infant mortality rate 1992 (per 1,000 live births)	.453

As previously mentioned, a nonperfect canonical correlation will result in some cases being classified incorrectly—this is, of course, almost inevitable. This situation is analogous to large residuals in multiple regression—recall that large residual values indicated that, for some individuals in the sample, the prediction equation did not accurately predict the value on the DV. Fortunately, discriminant analysis by way of statistical software programs will provide an assessment of the adequacy of classification. There are several means of accomplishing this assessment, depending on the analysis software being used by the researcher. In SPSS, the unstandardized coefficients are used to calculate the classification of cases in the original sample into DV groups. The number of correct classifications, also known as the **hit rate** (Stevens, 2001), is then compared to the actual group membership of participants in the original sample (Williams, 1992). A table showing the *actual* group membership and *predicted* group membership is presented. This table includes the percentage of correct classifications, based on the equation resulting from the unstandardized coefficients. Figure 10.2 shows the classification results for our working example. Although there is no rule of thumb regarding an acceptable rate of correct classifications, one would certainly hope to achieve a high percentage. In our example, the single discriminant function resulted in nearly 89.3% of the cases being classified correctly.

Figure 10.2. Assessment of Adequacy of Classification Results for “Classification as Developing Nation” Example.

		Classification Results ^{b,c}			Total
		Predicted Group Membership			
Original	develop	0	1		
		Count	23	5	28
		1	8	85	93
		%	82.1	17.9	100.0
		1	8.6	91.4	100.0
Cross-validated ^a	develop	Count	23	5	28
		1	8	85	93
		%	82.1	17.9	100.0
		1	8.6	91.4	100.0

- a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
- b. 89.3% of original grouped cases correctly classified.
- c. 89.3% of cross-validated grouped cases correctly classified.

Stevens (2001) suggested that assessing the accuracy of the hit rates with the *same* sample that was used to develop the discriminant function equation results in an unrealistic and misleading assessment. He suggests assessing the accuracy through an *external* classification analysis. In this manner, the data used to verify the classification are not used in the construction of the function itself. This can be accomplished in one of two ways: (1) If the sample is large enough, it may be split in half such that one-half is used to construct the function and the other half is used to assess its accuracy; or (2) through the use of the *jackknife* procedure, in which each participant is classified based on a classification statistic derived from the remaining ($n - 1$) participants. This second procedure is appropriate for small to moderate sample sizes. For more information on these two assessment procedures, refer to Stevens (2001) and Tabachnick and Fidell (2007).

Stevens (2001) further recommended that another important consideration not be overlooked when assessing a classification procedure. Obviously, it is important that the hit rate be high—in other words, we

should have mainly correct classifications. The additional consideration is the cost of misclassification—financially, morally, or ethically, and so on. To the researcher, it may initially look good to have a hit rate of approximately 90% to 95%, but what about the handful of cases that are classified incorrectly? There also exist ramifications of the misclassification of an individual from Group A, for instance, into Group B, resulting in greater cost than misclassifying someone in the other direction. For instance, assume we wanted to classify participants as either low-risk or high-risk in terms of developing heart disease based on family history and personal habits. Obviously, labeling a person low-risk when, in fact, he is high-risk is much more costly than identifying him as high-risk when he is actually low-risk. It is important to realize that these are issues that cannot be explained through statistical analyses but, rather, must be assessed and explicated by the knowledge and intuition of the researcher.

Similar to multiple regression, there are several approaches to conducting a discriminant analysis (Tabachnick & Fidell, 2007). In *standard*, or *direct*, discriminant analysis, each predictor (IV) is entered into the equation simultaneously and assigned only its unique association with the groups, as defined by the DV. In *sequential*, or *hierarchical*, discriminant analysis, the predictor IVs are entered into the analysis in an order specified by the researcher. The improvement in classification accuracy is assessed following the addition of each new predictor variable to the analysis. This procedure can be used to establish a priority among the predictor IVs and/or to reduce the number of possible predictors when a larger set is initially identified. Finally, in *stepwise*, or *statistical*, discriminant analysis, the order of entry of predictor variables is determined by statistical criteria. A reduced set of predictors will be obtained by retaining only statistically significant predictors.

Sample Research Questions

Again, based on our working example, let us now proceed to the specification of a series of possible research questions for our analysis:

1. Can status as a developing nation (i.e., *developed* or *developing*) be reliably predicted from knowledge of percent of population living in urban areas, gross domestic product, male and female life expectancy, and infant mortality rate?
2. If developing nation status can be predicted reliably, along how many dimensions do the two groups differ? How can those dimensions be interpreted?
3. Given the obtained classification functions, how adequate is the classification (in other words, what proportion of cases is classified correctly)?

SECTION 10.2 ASSUMPTIONS AND LIMITATIONS

Because the analytical procedures involved in discriminant analysis are so similar to those involved in MANOVA, the assumptions are basically the same. There are, of course, some adjustments to the assumptions, necessitated by the classification situation. The assumptions for discriminant analysis when being used for classification are as follows:

1. The observations on the predictor variables must be randomly sampled and independent of one another.
2. The sampling distribution of any linear combination of predictors is normal (multivariate normality).
3. The population covariance matrices for the predictor variables in each group must be equal (the homogeneity of covariance matrices assumption or the assumption of homoscedasticity).
4. The relationships among all pairs of predictors within each group must be linear.

An additional, and potentially serious, limitation of discriminant analysis is that it can be sensitive to sample size; therefore, consideration should be given to this issue. Stevens (2001) states that “unless sample size is large, relative to the number of variables, both the standardized coefficients and the correlations are very unstable” (p. 277). Further, he states that unless the ratio of total sample size to the number of variables (i.e., $\frac{N}{p}$) is quite large—approximately 20 to 1—one should use extreme caution in interpreting the results. For instance, if five variables are used in a discriminant analysis, there should be a minimum of 100 participants in order for the researcher to have confidence both in interpreting the discriminant function and in expecting to see the same function appear with another sample from the same population.

Methods of Testing Assumptions

Multivariate normality in discriminant analysis implies that the sampling distributions of the linear combinations of predictor variables are normally distributed (Tabachnick & Fidell, 2007). There exists no test of the normality of all linear combinations of sampling distributions of means of predictors; however, bivariate scatterplots are often used to examine univariate normality. It is important to note that discriminant analysis is robust to violations of multivariate normality, provided the violation is caused by skewness rather than by outliers (Tabachnick & Fidell, 2007). If outliers are present, it is imperative that the researcher transform or eliminate them prior to proceeding with the discriminant analysis. Multivariate outliers can be identified using Mahalanobis distance within **Regression**. You will recall from Chapter 6 that, as a conservative suggestion, robustness is expected with samples of 20 cases in the smallest group (on the DV), even if there are five or fewer predictors. In situations where multivariate normality is subject, an alternative classification procedure exists (Stevens, 2001). This technique is called *logistic regression* and is the topic of discussion in Chapter 11.

Discriminant analysis, when used for classification, is not robust to violations of the assumption of homogeneity of variance-covariance matrices (Tabachnick & Fidell, 2007). Cases tend to be overclassified into groups with a greater amount of dispersion, resulting in greater error as evidenced by a lower hit rate. Checking for univariate normality is a good starting point for assessing possible violations of homoscedasticity. A more applicable test for homoscedasticity in a discriminant analysis involves the examination of scatterplots of scores on the first two discriminant functions produced separately for each group. Obviously, this test is possible only in solutions resulting in more than one discriminant function. Rough equality in the overall size of the scatterplots provides evidence of homogeneity of variance-covariance matrices (Tabachnick & Fidell, 2007).

Linearity is, of course, best assessed through inspection of bivariate scatterplots. If both variables in the pair of predictors for each group on the DV are normally distributed and linearly related, the shape of the scatterplot should be elliptical. If one of the variables is not normally distributed, the relationship will not be linear and the scatterplot between the two variables will not appear oval shaped. Violations of the assumption of linearity are often seen as less serious than violations of other assumptions in that they tend to lead to reduced power as opposed to an inflated Type I error rate (Tabachnick & Fidell, 2007).

SECTION 10.3 PROCESS AND LOGIC

The Logic Behind Discriminant Analysis

Discriminant analysis is a *mathematical maximization* procedure. The goal of the procedure is to find uncorrelated linear combinations of the original (predictor) variables that maximize the between-to-within association, as measured by the sum-of-squares and cross-products (SSCP) matrices (Stevens, 2001). These uncorrelated linear combinations are referred to as the *discriminant functions*. The logic behind discriminant analysis involves finding the function with the largest eigenvalue—this results in maximum discrimination among groups (Stevens, 2001). The **first discriminant function** is the linear combination that maximizes the between-to-within association and is defined by the following equation:

$$DF_1 = a_{10}x_0 + a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1p}x_p \quad (\text{Equation 10.1})$$

where DF_1 is the first discriminant function, x_1 refers to the measure on an original predictor variable, and a_{11} refers to the weight (coefficient) assigned to a given variable for the first discriminant function (the first subscript following the a identifies the specific discriminant function, and the second subscript identifies the original variable)—for example, the term $a_{11}x_1$ refers to the product of the weight for variable 1 on DF_1 and the original value for an individual on variable 1. The term $a_{10}x_0$ represents the constant within the equation. The subscript p is equal to the total number of original predictor variables. This linear combination, then, provides for the maximum separation among groups.

This equation is used to obtain a discriminant function score for each participant. Using the unstandardized coefficients as presented in Figure 10.1, we can make the appropriate substitutions and arrive at the actual equation for our data and resulting discriminant function:

$$DF_1 = (-9.175) + (-.025)x_1 + (1.135)x_2 + (-.180)x_3 + (.196)x_4 + (.005)x_5$$

where each x represents an actual score on a predictor variable (specifically, x_1 = percent urban, x_2 = gross domestic product, x_3 = male life expectancy, x_4 = female life expectancy, and x_5 = infant mortality rate).

The analytic procedure then proceeds to find the second linear combination—*uncorrelated* with the first linear combination—that serves as the next best separator of the groups. The resulting equation would be:

$$DF_2 = a_{20}x_0 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2p}x_p \quad (\text{Equation 10.2})$$

As in factor analysis, it is again important to note that the constructed discriminant functions are not related. In other words:

$$r_{DF_1 \bullet DF_2} = 0$$

The third discriminant function is constructed so that it is uncorrelated with the first two and so that it serves as the third-best separator of the groups. This process continues until the maximum possible number of discriminant functions has been obtained. If k is the number of groups and p is the number of predictor variables, the maximum possible number of discriminant functions will be the smaller of p and $(k - 1)$ (Stevens, 2001). Thus, in a situation with three groups and 10 variables, the analysis would construct two discriminant functions. In any situation with only two groups, only one discriminant function would be constructed.

The actual classification of participants occurs in the following manner for a situation involving two groups (additional groups simply involve more of the same calculations). Once a discriminant function

has been constructed, the location of each group on the discriminant function (\bar{y}_k) is determined by multiplying the vectors of means for each participant on all predictor variables (\bar{x}_1) by the vector of unstandardized coefficients (a'):

$$\bar{y}_1 = a' \bar{x}_1 \quad (\text{Equation 10.3})$$

and

$$\bar{y}_2 = a' \bar{x}_2 \quad (\text{Equation 10.4})$$

The midpoint between the two groups on the discriminant function is then calculated as follows:

$$m = \frac{(\bar{y}_1 + \bar{y}_2)}{2} \quad (\text{Equation 10.5})$$

Participants are then classified—based on their individual discriminant function score, z_i —into one of the two groups based on the following *decision rule*:

- If $z_i \geq m$, then classify the participant into Group 1
- If $z_i < m$, then classify the participant into Group 2

Finally, it is important to note that within statistical analysis programs, it is possible to specify the prior probabilities of being classified into a specific group. **Prior probability** is essentially defined as the probability of membership in group k prior to collection of the data. The default option in the programs is to assign equal a priori probabilities. For instance, if we have two possible groups in which to classify participants, we would assume that they have an equal (50-50) chance of being classified in either group. The researcher may decide, based on *extensive* content knowledge, to assign different a priori probabilities. However, it is critical to note that this can have a substantial effect on the classification function; therefore, caution has been strongly recommended in using anything but equal prior probabilities (Stevens, 2001).

Interpretation of Results

Discriminant analysis output typically has four parts: (1) preliminary statistics that describe group differences and covariances, (2) significance tests and strength of relationship statistics for each discriminant function, (3) discriminant function coefficients, and (4) group classification. Interpretation of these four parts will be discussed subsequently.

Discriminant analysis produces a series of preliminary statistics that assist in interpreting the overall analysis results: a table of means and standard deviations for each IV by group, ANOVA analysis testing for group differences among the IVs, covariance matrices, and Box's test. Examination of group means and standard deviations and the ANOVA results is helpful in determining how groups differ within each IV. The ANOVA results, presented in the table of Tests of Equality of Group Means, includes Wilks' Lambda, F test, degrees of freedom, and p values for each IV. Group differences are usually significant. If they are not, the functions generated will not be very accurate in classifying individuals. The Box's M test is also included in preliminary statistics and is an indicator of significant differences in the covariance matrices among groups. A significant F test ($p < .001$) indicates that group covariances are not equal. Failure of the homogeneity of covariance assumption may limit the interpretation of results. However, one should keep in mind that the Box's test is highly sensitive to nonnormal distributions and therefore should be interpreted with caution.

The next section of output to interpret is the significance tests and strength of relationship statistics for each discriminant function, presented in the Eigenvalues and Wilks' Lambda Tables (see Figure 10.7). The Eigenvalue Table displays the eigenvalue, percentage of variance, and canonical correlation for each discriminant function. The canonical correlation represents the correlation between the discriminant function and the levels of the DV. By squaring the canonical correlation, we calculate the effect size (η^2) of the function, which indicates the percentage of variability in the function explained by the different levels in the DV.

The Wilks' Lambda Table provides chi-square tests of significance for each function. Statistics displayed include Wilks' Lambda, chi-square, degrees of freedom, and level of significance. Essentially, these statistics represent the degree to which there are significant group differences in the IVs after the effects of the previous function(s) have been removed. These significance tests help in determining the number of functions to interpret. For instance, if an analysis generated three functions, of which the first two functions were significant, only the first two functions would be interpreted.

Once the number of functions to interpret has been determined, each function can then be interpreted/named by examining the variables that are most related to it. Two tables are utilized for this process—Standardized Canonical Discriminant Function Coefficients and the Structure Matrix. The table of Standardized Canonical Discriminant Function Coefficients presents the standardized discriminant function coefficients, which represent the degree to which each variable contributes to each function. The Structure Matrix presents the correlation coefficients between the variables and functions.

The next step in interpreting discriminant analysis results is assessing the accuracy of the functions in classifying participants in the appropriate groups. The table of Classification results displays the original and predicted frequency and percentage of participants within each group. The final step in this interpretation process is to determine the extent to which group differences support the functions generated. This is done by reviewing group means for each function as presented in the table of Functions at Group Centroids.

Continuing with our first example (which used *country-d.sav*) that seeks to predict the status of a developing nation (*develop*) from the knowledge of percentage of population living in urban areas (*urban*), gross domestic product (*gdp*), male (*lifeexpm*) and female (*lifeexpf*) life expectancy, and infant mortality (*infmr*), we screened data for outliers. Utilizing Mahalanobis distance, one outlier (case number 83) was removed as it exceeded the chi-square critical value $\chi^2(5) = 20.515$ at $p = .001$ (see Figure 10.3). Therefore, all cases in which the Mahalanobis value exceeded the chi-square criterion were eliminated using **Select Cases: If MAH_1 \leq 20.515**. Assessment of normality and linearity was conducted by evaluating bivariate scatterplots of the IVs. Figure 10.4 indicates that the variable of *gdp* is not normally distributed and does not linearly relate with the other variables. Consequently, the transformed (natural log) version of this variable was utilized (*lngdp*). A discriminant analysis was then conducted using **Classify** (using range values 0,1). The Enter method was applied. Figure 10.5 reveals significant ($p < .001$) group differences for each IV. Although the Box's test indicates that homogeneity of covariance cannot be assumed, our interpretation will continue because the Box's test is highly sensitive to non-normality (refer to Figure 10.6). Figure 10.7 presents the Eigenvalues and Wilks' Lambda tables. Because the DV consisted of two levels, only one function could be generated. The canonical correlation ($r = .752$) indicates that the function is highly related to the levels in the DV. Squaring this value produces the effect size, which reveals that 56.5% ($\eta^2 = .752^2 = .566$) of function variance is accounted for by the DV. The overall Wilks' Lambda was significant, $\Lambda = .434$, $\chi^2(5, N = 121) = 97.21$, $p < .001$, and indicates that the function of predictors significantly differentiated between countries being classified as developed and developing. Evaluation of the standardized discriminant function coefficients reveals that female life expectancy (1.780) had the highest loading, followed by male life expectancy (-1.500), gross domestic product (1.203), percent urban (-.548), and

infant mortality (.181) (see Figure 10.8a). In contrast, variable correlations with the function (Figure 10.8b) indicate that gross national product (.904) has the strongest relationship, followed by female life expectancy (.637), infant mortality (−.616), male life expectancy (.575), and percent urban (.453). These differences in function and correlation coefficients make it somewhat difficult to interpret the function, especially because only one function was generated. However, the three variables that appear to be the most consistent in the function are gross domestic product, female life expectancy, and male life expectancy. With these variables in mind, we have named the function, *Life and Economic Well-Being*.

The next step in interpreting discriminant analysis is evaluating the accuracy of the function in classifying participants in the appropriate groups. Figure 10.9 presents the classification results, which are based upon group sizes. In addition, the cross-validation procedure was utilized to double-check the accuracy of classification. Initial classification indicates that 89.3% of the cases were correctly classified. Cross-validation supported this level of accuracy (89.3%). Initial classification results revealed that 82.1% of the *developed* countries were correctly classified, while 91.4% of the *developing* countries were correctly classified. The means of the discriminant functions are consistent with these results (see Figure 10.10). Developed countries had a function mean of 2.06, while developing countries had a mean of −.621. These results suggest that countries with high life expectancy and gross domestic product are likely to be classified as *developed*.

Figure 10.3. Mahalanobis Distance (Example 1).

Extreme Values				
		Case Number	Value	
MAH_1	Highest	1	83	27.55905
		2	122	17.05638
		3	37	15.69150
		4	81	13.50276
		5	119	11.95556
	Lowest	1	22	.61813
		2	24	.63039
		3	50	.78780
		4	29	1.04046
		5	30	1.15813

Only case 83 exceeds the chi-square critical value.

Figure 10.4. Scatterplot of Predictor Variables (Example 1).

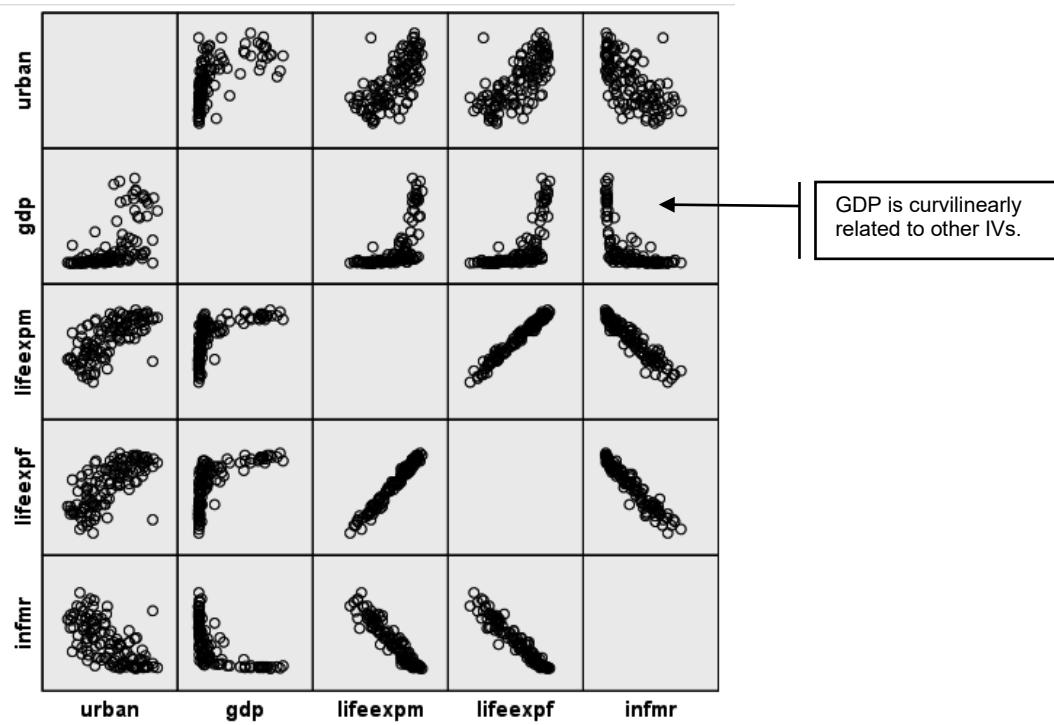


Figure 10.5. (a) Group Statistics; (b) ANOVA Summary Table (Example 1).

(a)

		Group Statistics			
		Mean	Std. Deviation	Valid N (listwise)	
develop				Unweighted	Weighted
Developed country	urban	68.6893	16.92881	28	28.000
	lifeexpm	71.8929	4.27169	28	28.000
	lifeexpf	78.4286	4.40899	28	28.000
	infmr	13.0857	13.87923	28	28.000
	Ingdp	9.2418	.82739	28	28.000
Developing country	urban	42.2839	22.95437	93	93.000
	lifeexpm	59.0215	9.20182	93	93.000
	lifeexpf	62.8602	10.07104	93	93.000
	infmr	71.4301	39.39637	93	93.000
	Ingdp	6.6702	1.11821	93	93.000
Total	urban	48.3942	24.36152	121	121.000
	lifeexpm	62.0000	9.93646	121	121.000
	lifeexpf	66.4628	11.20717	121	121.000
	infmr	57.9289	42.93888	121	121.000
	Ingdp	7.2653	1.51612	121	121.000

(b)

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
urban	.789	31.765	1	119	.000
lifeexpm	.699	51.225	1	119	.000
lifeexpf	.654	62.978	1	119	.000
infmr	.669	58.906	1	119	.000
Ingdp	.484	126.838	1	119	.000

All predictor variables show significant group differences.

Figure 10.6. Box's M Test (Example 1).

Test Results	
Box's M	71.966
F	Approx. 4.454
df1	15
df2	10051.412
Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Significance indicates that homogeneity of covariance cannot be assumed.

Figure 10.7. (a) Eigenvalues Table; (b) Wilks' Lambda Table (Example 1).

(a)

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.303 ^a	100.0	100.0	.752

a. First 1 canonical discriminant functions were used in the analysis.

Canonical correlation of function with predictor variables. Square this to calculate effect size.

(b)

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.434	97.205	5	.000

Figure 10.8. (a) Standardized Discriminant Function Coefficients Table; (b) Correlation Coefficient Table (Example 1).

(a)

Standardized Canonical Discriminant Function Coefficients

	Function
	1
urban	-.548
lifeexprm	-1.500
lifeexpf	1.780
infmr	.181
Ingdp	1.203

(b)

Structure Matrix

	Function
	1
Ingdp	.904
lifeexpf	.637
infmr	-.616
lifeexprm	.575
urban	.453

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

Figure 10.9. Classification Results (Example 1).

			Predicted Group Membership		Total
		Developed country		Developing country	
Original	Count	Developed country	23	5	28
		Developing country	8	85	93
	%	Developed country	82.1	17.9	100.0
Cross-validated ^b	Count	Developed country	23	5	28
		Developing country	8	85	93
	%	Developed country	82.1	17.9	100.0
		Developing country	8.6	91.4	100.0

a. 89.3% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 89.3% of cross-validated grouped cases correctly classified.

Percentage of cases correctly classified for original and cross-validated results by group.

Percentage of cases correctly classified for original and cross-validated results for the total sample.

Figure 10.10. Discriminant Function Means (Example 1).

develop	Function 1
Developed country	2.063
Developing country	-.621

Unstandardized canonical discriminant functions evaluated at group means

Writing Up Results

After stating any data transformations and/or outlier elimination, one should report the significance test results for each function. These results should include Wilks' Lambda, chi-square with degrees of freedom and number of participants, and p value. Applying these results, you should then indicate the number of functions you chose to interpret. It is then necessary to present the standardized function coefficients and the correlation coefficients. These coefficients are typically displayed in a table, while the narrative describes the primary variables associated with each function. The researcher should then indicate how the functions were labeled; the previous statistics should support this process. Classification results are presented and should include the percentage of accuracy for each group and the entire sample. Finally, function means are presented in support of the functions generated. The following results statement applies the results from Figures 10.3 through 10.10.

A discriminant analysis was conducted to determine whether five variables—urban population, female life expectancy, male life expectancy, infant mortality, and gross domestic product—could predict the development (developed vs. developing) status for a country. Prior to analysis, one outlier was eliminated. Due to nonnormality, the variable of gross domestic product was transformed by taking its natural log. One function was generated and was significant [$\Lambda = .434$, $\chi^2(5, N = 121)$

$= 97.21, p < .001$], indicating that the function of predictors significantly differentiated between countries being classified as developed and developing. Development status was found to account for 56.5% of function variance. Correlation coefficients and standardized function coefficients (see Table 1) revealed that the variables of gross domestic product, male life expectancy, and female life expectancy were most associated with the function. Based upon these results, the function was labeled *Life and Economic Well-Being*. Original classification results revealed that 82.1% of the *developed* countries were correctly classified, while 91.4% of the *developing* countries were correctly classified. For the overall sample, 89.3% were correctly classified. Cross-validation derived 89.3% accuracy for the total sample. The means of the discriminant functions are consistent with these results. Developed countries had a function mean of 2.06, while developing countries had a mean of $-.621$. These results suggest that countries with high life expectancy and gross domestic product are likely to be classified as *developed*.

Table 1

Correlation Coefficients and Standardized Function Coefficients for Life and Economic Well-Being

	Correlation Coefficients With Discriminant Function	Standardized Function Coefficients
Female life expectancy	.637	1.780
Male life expectancy	.575	-1.500
Gross domestic product	.904	1.203
Urban population	.453	-.548
Infant mortality	-.616	.181

SECTION 10.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example of a stepwise discriminant analysis. We will develop the research question, screen data, conduct data analysis, interpret output, and summarize the results. For this chapter's second example, SPSS data set *profile-c.sav* is utilized.

Problem

We are interested in predicting one's life perspective (*life*) with the following seven variables: age (*age*), hours worked per week (*hrs1*), years of education (*educ*), income (*rincom91*), number of siblings (*sibs*), hours of TV viewing per week (*tvhours*), and hours worked per week by spouse (*sphrs1*). The following research question is generated to address this scenario: Can life perspective be reliably predicted from knowledge of an individual's age, hours worked per week, years of education, income, number of siblings, hours of TV viewing per week, and hours worked per week by spouse? The DV is *life* and has three levels of response: dull, routine, and exciting. The predictor variables are the seven IVs. Because this scenario utilizes a variable that requires participants to be married (hours worked per week by spouse, *sphrs1*), nearly two-thirds of the respondents are not included because they are not married.

Methods and SPSS "How To"

Prior to conducting the discriminant analysis, data were screened for outliers, normality, and linearity. The variable of *rincom91* was transformed to eliminate cases greater than 22. This new variable was named *rincom2*. Mahalanobis distance was calculated with the seven predictor variables, which generated the variable *MAH_1*. Using the data set *profile-c.sav* (which includes all transformed variables) we can now

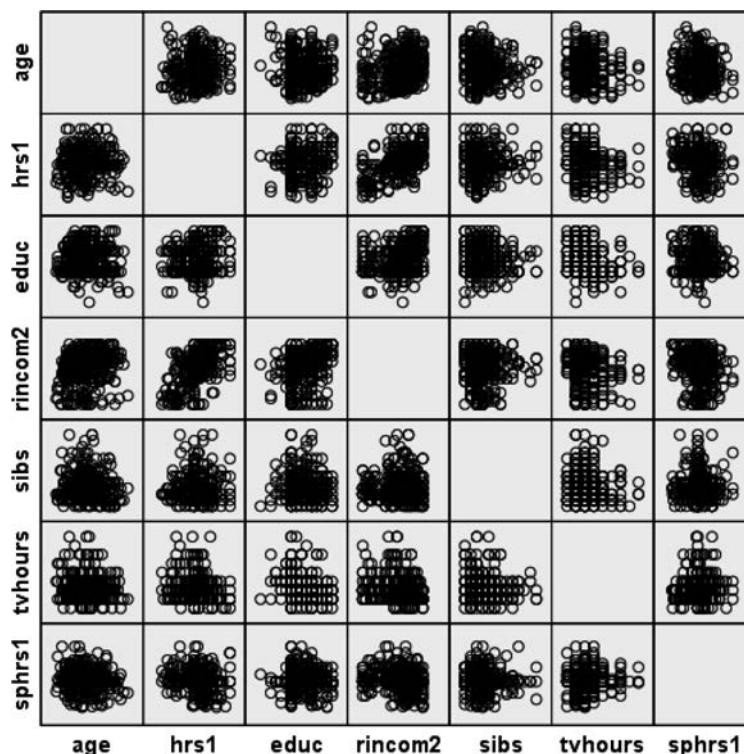
run **Explore** using *MAH_1*, which reveals numerous outliers (see Figure 10.11). Therefore, all cases in which the Mahalanobis value exceeded the chi-square criterion [$p = .001, \chi^2(7) = 24.322$] need to be eliminated using **Select Cases: If *MAH_1* ≤ 24.322** . A bivariate scatterplot was then created to evaluate normality and linearity (see Figure 10.12). In general, plots are fairly elliptical, indicating normality and linearity.

Figure 10.11. Mahalanobis Distance (Example 2).

Extreme Values			
		Case Number	Value
MAH_1	Highest	1	406
		2	121
		3	729
		4	926
		5	1420
	Lowest	1	.50833
		2	.66873
		3	1.03521
		4	1.05548
		5	1.24961

Because several cases exceed the χ^2 critical value, cases greater than $\chi^2(7) = 24.322$ were eliminated.

Figure 10.12. Scatterplot Matrix of Predictor Variables (Example 2).



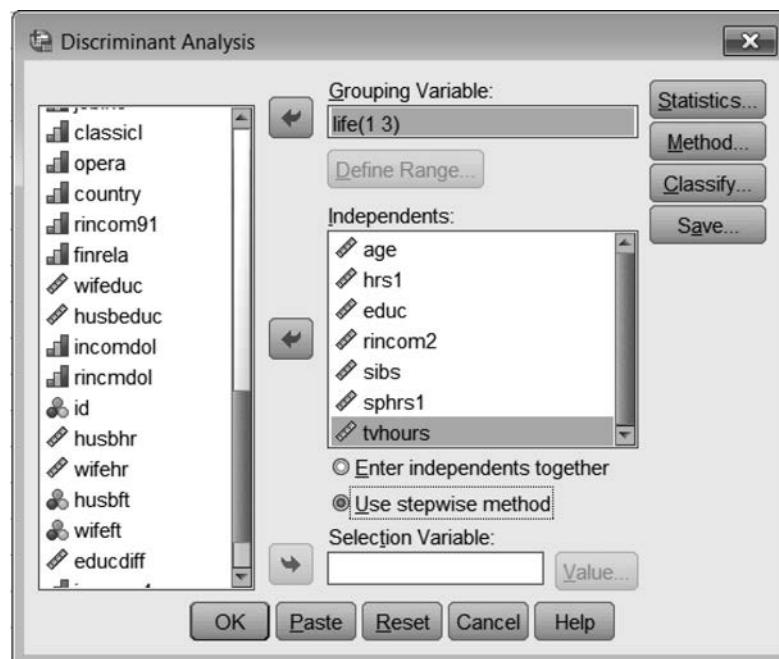
A stepwise discriminant analysis is conducted next using **Classify**. To open the **Discriminant Analysis** dialog box shown in Figure 10.13, select the following:

Analyze
 Classify
 Discriminant

Discriminant Analysis dialog box (Figure 10.13)

Click the DV and move it to the **Grouping Variable** box. Because *life* has three values ranging from 1 to 3, click **Define Range**; specify a minimum value of 1 and a maximum value of 3, and then click **Continue**. Click each IV and move it to the **Independents** box. Select the method by either checking **Enter independents together** or **Use stepwise method**. For our second example, we chose the stepwise. When the stepwise option is selected, the option to select **Bootstrap** is no longer available. Click **Statistics**.

Figure 10.13. Discriminant Analysis Dialog Box.



Discriminant Analysis: Statistics dialog box (see Figure 10.14)

Under **Descriptives**, check all the options: **Means**, **Univariate ANOVAs**, and **Box's M**. No other options were selected within this dialog box; however, a description of each option is as follows.

Fisher's—Canonical function coefficients that maximize discrimination between levels of the DV.

Unstandardized—Unstandardized function coefficients based on the variable raw scores.

Within-groups correlation—Correlation matrix of the IVs at each level of the DV.

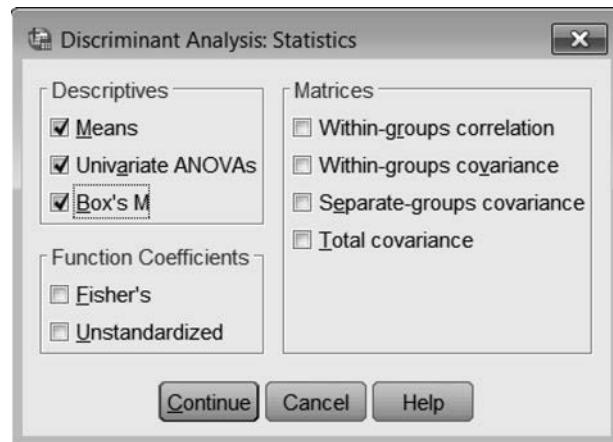
Within-groups covariance—Covariance matrix of the IVs at each level of the DV.

Separate-groups covariance—Same as within-groups covariance, but a separate matrix is produced for each level of the DV.

Total covariance—Covariance matrix for the entire sample.

After checking the **Statistics** options, click **Continue**, then **Method**.

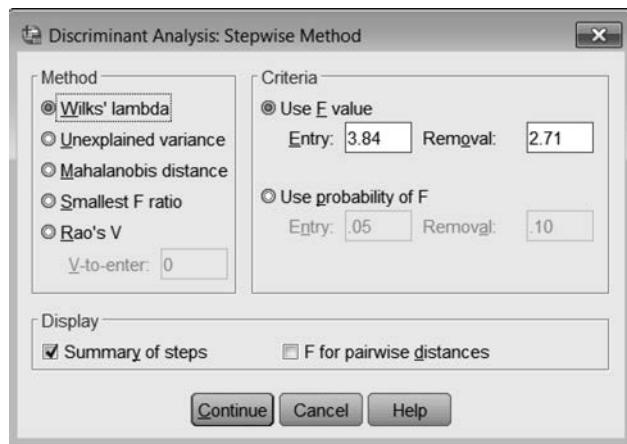
Figure 10.14. Discriminant Analysis: Statistics Dialog Box.



Discriminant Analysis: Stepwise Method dialog box (see Figure 10.15)

This dialog box will appear only if you have selected the **Stepwise Method**. Under **Method**, check **Wilks' lambda**. Under **Display**, check **Summary of steps**. Under **Criteria**, check **Use F value** and use the default values. Click **Continue**, then **Classify**.

Figure 10.15. Discriminant Analysis: Stepwise Method Dialog Box.



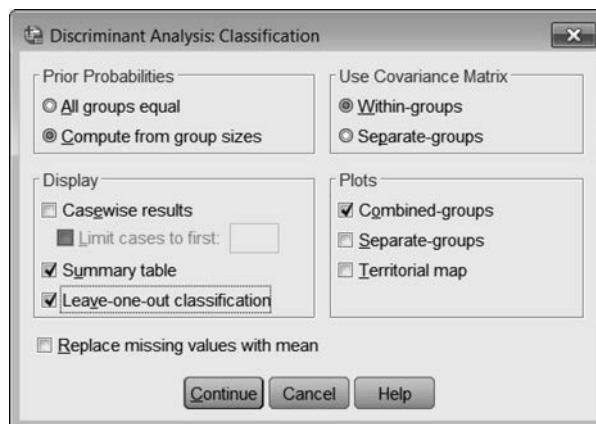
Discriminant Analysis: Classification dialog box (see Figure 10.16)

If group proportions in the population are fairly equal (within a 10% difference), then select **All groups equal** under **Prior Probabilities**. If groups are unequal, check **Compute from group sizes**; this is the most commonly used option and the one we chose in the example. Under **Display**, select **Summary table**, which will display the number and percentage of correct and incor-

rect classifications for each group. Also select **Leave-one-out classification**, which will re-classify each case based on the functions of all other cases excluding that case. Under **Use Covariance Matrix**, select the default **Within-groups**.

Depending upon the number of levels in the DV, you may select one or more of the options under **Plots**. The **Combined-groups** plot will create a histogram (for two groups) or a scatterplot (for three or more groups). This option is often helpful when the DV has three or more levels. The **Separate-groups** plot creates a separate plot for each group. The **Territorial map** charts centroids and boundaries and is used only when the DV has three or more levels. For our example, we selected the **Combined-groups** plot. However, the plot was not presented because only one function was interpreted. After you have selected the desired plots, click **Continue**, then **Save**.

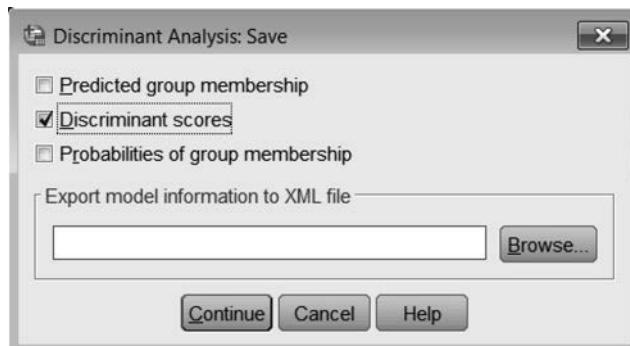
Figure 10.16. Discriminant Analysis: Classification Dialog Box.



Discriminant Analysis: Save dialog box (Figure 10.17)

This dialog box provides options for saving specific results as variables for future analysis. Three options are available: **Predicted group membership**, **Discriminant scores**, and **Probabilities of group membership**. It is common to save **Discriminant scores**. Once you have selected the type(s) of variables to save, click **Continue**, then **OK**.

Figure 10.17. Discriminant Analysis: Save Dialog Box.



Output and Interpretation of Results

Figure 10.18 presents the group means for each of the predictor variables. ANOVA results for group differences are presented in Figure 10.19. Group differences are significant ($p < .05$) only for the variables of *educ* and *rincom2*. The Box's M Test indicates equality of covariances (see Figure 10.20). Because a stepping procedure was utilized in this analysis, Figure 10.21 presents the variables entered at each step. Only two variables, *educ* and *rincom2*, were entered into the model. The Eigenvalues table (Figure 10.22) reveals that two functions were generated, with the DV accounting for only 8.3% ($\eta^2 = .288^2 = .083$) of the variance in the first function. The Wilks' Lambda table (Figure 10.23) indicates that the first function is significant [$\Lambda = .905$, $\chi^2(4, N = 236) = 23.18, p < .001$]. However, the second function is not significant [$\Lambda = .987$, $\chi^2(1, N = 236) = 3.11, p = .078$]. Consequently, only one function will be interpreted. The standardized function coefficients and correlation coefficients presented in Figure 10.24 indicate that *educ* has the highest relationship with the function. Although the variables of education level and income are somewhat different, we will refer to this function as *Work Success*. Group means for the function (see Figure 10.25) indicate that those who perceive life as dull had a function mean of -1.042 , those who perceive life as routine had a mean of $-.238$, and those who perceive life as exciting had a mean of $.262$. These results suggest that individuals who perceive life as dull have the least amount of education and the lowest income. Classification results (see Figure 10.26) reveal that the original grouped cases were classified with only 58.5% overall accuracy. Accuracy by each group was 0% for dull, 56.5% for routine, and 63.1% for exciting. The cross-validated results supported original accuracy levels with 58.1% correctly classified overall.

Figure 10.18. Group Statistics (Example 2).

		Group Statistics			
life		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Dull	age	37.8333	14.63443	6	6.000
	hrs1	42.8333	15.28943	6	6.000
	educ	12.8333	1.32916	6	6.000
	rincom2	7.1667	6.24233	6	6.000
	sibs	3.6667	1.63299	6	6.000
	tvhours	2.0000	2.28035	6	6.000
	sphrs1	44.1667	16.25320	6	6.000
Routine	age	40.8519	9.80350	108	108.000
	hrs1	42.3519	13.93195	108	108.000
	educ	13.5926	2.22575	108	108.000
	rincom2	12.7407	5.46932	108	108.000
	sibs	3.3056	1.92115	108	108.000
	tvhours	2.1574	1.18528	108	108.000
	sphrs1	41.0833	14.78198	108	108.000
Exciting	age	41.0902	10.03301	122	122.000
	hrs1	44.1475	13.24884	122	122.000
	educ	14.8197	2.75446	122	122.000
	rincom2	14.3279	5.12721	122	122.000
	sibs	2.9508	2.37707	122	122.000
	tvhours	2.0574	1.41596	122	122.000
	sphrs1	41.3033	14.31285	122	122.000
Total	age	40.8983	10.02031	236	236.000
	hrs1	43.2924	13.58391	236	236.000
	educ	14.2076	2.57221	236	236.000
	rincom2	13.4195	5.44273	236	236.000
	sibs	3.1314	2.16444	236	236.000
	tvhours	2.1017	1.33617	236	236.000
	sphrs1	41.2754	14.51970	236	236.000

Figure 10.19. ANOVA Summary Table (Example 2).

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
age	.997	.302	2	233	.739
hrs1	.996	.502	2	233	.606
educ	.937	7.827	2	233	.001
rincom2	.945	6.820	2	233	.001
sibs	.992	.957	2	233	.385
tvhours	.998	.177	2	233	.838
sphrs1	.999	.128	2	233	.880

Only *educ* and *rincom2* show significant group differences.

Figure 10.20. Box's M Test (Example 2).

Test Results	
Box's M	12.437
F	Approx. 1.902
df1	6
df2	1074.518
Sig.	.078

Tests null hypothesis of equal population covariance matrices.

NOT significant. Indicates that homogeneity of covariance can be assumed.

Figure 10.21. Summary Table of Steps (Example 2).

Step	Entered	Variables Entered/Removed ^{a,b,c,d}							
		Wilks' Lambda						Exact F	
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	educ	.937	1	2	233.000	7.827	2	233.000	.001
2	rincom2	.905	2	2	233.000	5.928	4	464.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

- a. Maximum number of steps is 14.
- b. Minimum partial F to enter is 3.84.
- c. Maximum partial F to remove is 2.71.
- d. F level, tolerance, or VIN insufficient for further computation.

Figure 10.22. Eigenvalues Table (Example 2).

Eigenvalues					
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation	
1	.090 ^a	87.0	87.0	.288	
2	.013 ^a	13.0	100.0	.115	

Canonical correlation of function with predictor variables. Square this to calculate effect size.

- a. First 2 canonical discriminant functions were used in the analysis.

Figure 10.23. Wilks' Lambda Table (Example 2).

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.905	23.176	4	.000
2	.987	3.105	1	.078

Indicates that first function is significant, while the second function is not.

Figure 10.24. (a) Standardized Discriminant Function Coefficients Table; (b) Correlation Coefficient Table (Example 2).

(a)

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
educ	.671	-.801
rincom2	.572	.875

(b)

Structure Matrix

	Function	
	1	2
educ	.837 ^a	-.547
rincom2	.767 ^a	.642
hrs1 ^b	.407 ^a	.393
tvhours ^b	-.222 ^a	.041
sphrs1 ^b	-.112 ^a	-.083
age ^b	.064	.159 ^a
sibs ^b	-.046	.050 ^a

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

^a. Largest absolute correlation between each variable and any discriminant function

^b. This variable not used in the analysis.

Figure 10.25. Discriminant Function Means.

		Function	
		1	2
life			
Dull		-1.042	-.589
Routine		-.238	.085
Exciting		.262	-.047

Unstandardized canonical
discriminant functions evaluated
at group means

Figure 10.26. Classification Results (Example 2).

			Classification Results ^{a,c}				Percentage of cases correctly classified for original and cross-validated results by group.	
			Predicted Group Membership			Total		
life			Dull	Routine	Exciting			
Original	Count	Dull	0	6	0	6	Percentage of cases correctly classified for original and cross-validated results by group.	
		Routine	0	61	47	108		
		Exciting	0	45	77	122		
		Ungrouped cases	0	54	60	114		
	%	Dull	.0	100.0	.0	100.0		
		Routine	.0	56.5	43.5	100.0		
		Exciting	.0	36.9	63.1	100.0		
		Ungrouped cases	.0	47.4	52.6	100.0		
Cross-validated ^b	Count	Dull	0	6	0	6	Percentage of cases correctly classified for original and cross-validated results for total sample.	
		Routine	0	61	47	108		
		Exciting	0	46	76	122		
	%	Dull	.0	100.0	.0	100.0		
		Routine	.0	56.5	43.5	100.0		
		Exciting	.0	37.7	62.3	100.0		

a. 58.5% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 58.1% of cross-validated grouped cases correctly classified.

Presentation of Results

The following report summarizes the stepwise discriminant analysis results.

A stepwise discriminant analysis was conducted to determine the ability of the seven variables—age, hours worked per week, years of education, income, number of siblings, number of hours spent watching TV per day, and hours worked per week by spouse—to predict one's life perspective (dull, routine, exciting) among married respondents. Prior to analysis, the variable of income was transformed by eliminating participants exceeding 22. In addition, several outliers were eliminated. The analysis generated two functions. However, only Function 1 was significant [$\Lambda = .905$, $\chi^2(4, N = 236) = 23.18, p < .001$], with only 8.3% of the function variability explained by life perspective. Two variables were entered into the function: years of education and income, respectively. The variables of age, hours worked per week, number of siblings, number of hours spent watching TV per day, and hours worked per week by spouse were excluded. Table 2 presents the correlation coefficients and standardized function coefficients. The function was labeled *Work Success*. Classification results revealed that the original grouped cases were classified with only 58.5% overall accuracy. Accuracy by each group was 0% for dull, 56.5% for routine, and 63.1% for exciting. The cross-validated results supported original accuracy levels with 58.1% correctly classified overall. Group means for the function indicated that those who perceive life as dull had a function mean of -1.042 , those who perceive life as routine had a mean of $-.238$, and those who perceive life as exciting had a mean of $.262$. These results suggest that individuals who perceive life as dull have the least amount of education and the lowest income.

Table 2

Correlation Coefficients and Standardized Function Coefficients for Work Success

	Correlation Coefficients With Discriminant Function	Standardized Function Coefficients
Education	.837	.671
Income	.767	.572

SUMMARY

Discriminant analysis is often used to predict group membership based on observed characteristics (predictor variables). The technique generates discriminant function(s) derived from linear combinations of the predictor variables that best discriminate between/among the groups. Prior to conducting discriminant analysis, data should be screened for missing data and outliers. Data should also be tested for normality, linearity, and homogeneity of covariance. The discriminant analysis procedure in SPSS generates four parts: (1) preliminary statistics that describe group differences and covariances, (2) significance tests and strength of relationship statistics for each discriminant function, (3) discriminant function coefficients, and (4) group classification. Preliminary statistics present group means and standard deviations for each IV, ANOVA results for group differences (Wilks' Lambda, F test, degrees of freedom, and p values for each IV), and Box's M test (test for homogeneity of covariance). Significance tests and strength of relationship statistics are presented in the Eigenvalues and Wilks' Lambda tables. The Eigenvalue table displays the eigenvalue, percentage of variance, and canonical correlation for each discriminant function. The Wilks' Lambda table provides chi-square tests of significance for each function (Wilks' Lambda, chi-square, degrees of freedom, and p value). Discriminant function coefficients are presented in the tables of Standardized Canonical Discriminant Function Coefficients (which represent the degree to which each variable contributes to each

function) and the Structure Matrix (which displays the correlation coefficients between the variables and functions). Assessment of these coefficients assists in labeling the function(s). Group classification results are utilized to assess the accuracy of the functions in classifying participants in the appropriate groups and are presented in a table that displays the original and predicted frequency and percentage of participants within each group. The final step in this interpretation process is to determine the extent to which group differences support the functions generated by reviewing group means for each function as presented in the table of Functions at Group Centroids. Figure 10.27 presents a checklist for conducting discriminant analysis.

KEYWORDS

- canonical correlation
- classification
- discriminant functions
- first discriminant function
- hit rate
- prior probability
- standardized discriminant function coefficient
- unstandardized discriminant function coefficient

Figure 10.27. Checklist for Conducting Discriminant Analysis.

I. Screen Data

- Missing Data?
- Multivariate Outliers?
 - Run preliminary Regression to calculate Mahalanobis distance.
 - Analyze... Regression... Linear.**
 - Identify a variable that serves as a case number and move to **Dependent Variable** box.
 - Identify all appropriate quantitative variables and move to **Independent(s)** box.
 - Save.**
 - Check **Mahalanobis** under Distances.
 - Continue, OK.**
 - Determine chi-square (χ^2) critical value at $p < .001$.
 - Conduct **Explore** to test outliers for Mahalanobis chi-square (χ^2).
 - Analyze... Descriptive statistics... Explore.**
 - Move **MAH_1** to **Dependent Variable** box.
 - Leave **Factor** box empty.
 - Statistics.**
 - Check **Outliers**.
 - Continue, OK.**
 - Delete outliers for participants when χ^2 exceeds critical χ^2 at $p < .001$.
- Linearity and Normality?
 - Create Scatterplot Matrix of all IVs.
 - Scatterplot shapes are not close to elliptical shapes → reevaluate univariate normality and consider transformations.

II. Conduct Discriminant Analysis

- Run Discriminant Analysis using **Classify**.
 - Analyze... Classify... Discriminant.**
 - Move the DV to the **Grouping Variable** box.
 - Define Range.**
 - Indicate the lowest and highest group value. **Continue.**
 - Move IVs to the **Independent(s)** box.
 - Enter or Stepwise.**
 - If **Stepwise** was selected, **Method**.
 - Check **Wilks' Lambda**, **Use F value**, and **Summary of steps**.
 - Continue.**
 - Statistics.**
 - Check: **Means**, **Univariate ANOVAs**, and **Box's M**.
 - Continue.**
 - Classify.**
 - Check **Compute from group sizes**, **Summary table**, **Leave-one-out classification**, **Within-groups Matrix**, and **Combined-groups plot**.
 - Continue.**
 - Save.**
 - Check **Discriminant scores**.
 - Continue, OK.**
- Interpret Box's M Test. If significant at $p < .001$, continue to interpret results, but proceed with caution.

III. Summarize Results

- Describe any data elimination or transformation.
- Report the significance test results for each function (Wilks' Lambda, chi-square, df , N , and p value).
- Indicate the number of functions you chose to interpret.
- Present the standardized function coefficients and the correlation coefficients in narration and table.
- Indicate how functions were labeled.
- Present classification results, including the percentage of accuracy for each group and the entire sample.
- Present function means in support of the functions generated.
- Draw conclusions.

Exercises for Chapter 10

This exercise utilizes the data set *schools-a.sav*, which can be downloaded from this website:

www.routledge.com/9781138289734

Conduct a stepwise discriminant analysis with the following variables:

IVs—*grad94, act94, pctact94, math93, math94me, read93, read94me, scienc93, sci94me*
DV—*medloinc*

1. Develop a research question.
2. Which predictor variables show significant group differences?
3. Can homogeneity of covariance be assumed?
4. Which variables were entered into the model?
5. How many functions were generated? Why?
6. Is (are) the function(s) significant? Explain.
7. Calculate the effect size for each function.
8. Evaluating the function coefficients and correlation coefficients, what would you label the function(s)?
9. How accurate is (are) the function(s) in predicting income level for each group and the total sample?
10. What is your conclusion regarding these results?

CHAPTER 11

LOGISTIC REGRESSION

STUDENT LEARNING OBJECTIVES

After studying Chapter 11, students will be able to:

1. Differentiate between multiple regression and logistic regression.
2. Describe the nature of the value of the criterion variable being predicted in a logistic regression analysis.
3. Describe the three main components when interpreting the results of a logistic regression.
4. Discuss what is being measured by a Wald statistic.
5. Explain problems that may be associated with situations where there are too few cases in relation to the number of predictor variables in a logistic regression.
6. Describe what is meant by *odds* in a logistic regression and how odds are calculated.
7. Develop research questions appropriate for a logistic regression analysis.
8. Design and conduct a logistic regression in order to classify subjects into a specific number of groups by following the appropriate SPSS guidelines provided.

In this chapter, we present a discussion of logistic regression—an alternative to discriminant analysis—and also discuss multiple regression for use in certain situations. Logistic regression has the same basic purpose as discriminant analysis—the classification of individuals into groups. It is, in some ways, more flexible and versatile than discriminant analysis, although mathematically, it can be quite a bit more cumbersome. In this chapter, we also examine how logistic regression seeks to identify a combination of IVs—which are limited in few, if any, ways—that best predicts membership in a particular group, as measured by a categorical DV.

SECTION 11.1 PRACTICAL VIEW

Purpose

Logistic regression is basically an extension of multiple regression in situations where the DV is not a continuous or quantitative variable (George & Mallery, 2000). In other words, the DV is categorical (or discrete) and may have as few as two values. For instance, in a logistic regression application, these categories might include values such as membership or nonmembership in a group, completion or noncompletion of an academic program, passing or failing to pass a course, survival or failure of a business, and so on.

Due to the nature of the categorical DV in logistic regression, this procedure is also sometimes used as an alternative to discriminant analysis. Because the goal is to predict values on a DV that is categorical,

we are essentially attempting to predict membership in one of two or more groups. The reader should likely see the similarity between this procedure and that discussed in the previous chapter for discriminant analysis (i.e., the classification, or prediction, of participants into groups). Although logistic regression may be used to predict values on a DV of two or more categories, our discussion will focus on binary logistic regression, in which the DV is dichotomous.

The basic concepts that are fundamental to multiple regression analysis—namely that several variables are regressed onto another variable using one of several selection processes—are the same for logistic regression analysis (George & Mallory, 2000), although the meaning of the resultant regression equation is considerably different. As you read in Chapter 7, a standard regression equation is composed of the sum of the products of weights and actual values on several predictor variables (IVs) in order to predict the values on the criterion variable (DV). In contrast, the value that is being predicted in logistic regression is actually a *probability*, which ranges from 0 to 1. More precisely, logistic regression specifies the probabilities of the particular outcomes (e.g., *pass* and *fail*) for each participant or case involved. In other words, logistic regression analysis produces a regression equation that accurately predicts the probability of whether an individual will fall into one category (e.g., *pass*) or the other (e.g., *fail*) (Tate, 1992).

Although logistic regression is fairly similar to both multiple regression and discriminant analysis, it does provide several distinct advantages over both techniques (Tabachnick & Fidell, 2007). Unlike discriminant analysis and multiple regression, logistic regression requires that no assumptions about the distributions of the predictor variables (IVs) need to be made by the researcher. In other words, the predictors do not have to be normally distributed, linearly related, or have equal variances within each group. This fact alone makes logistic regression much more flexible than the other two techniques. In addition, logistic regression cannot produce negative predictive probabilities, as can happen when applying multiple regression to situations involving dichotomous outcomes (Tate, 1992). In logistic regression, all probability values will be positive and will range from 0 to 1 (Tabachnick & Fidell, 2007). Another advantage is that logistic regression has the capacity to analyze predictor variables of all types—continuous, discrete, and dichotomous. Finally, logistic regression can be especially useful when the distribution of data on the criterion variable (DV) is expected or known to be *nonlinear* with one or more of the predictor variables (IVs). Logistic regression is able to produce nonlinear models, which again adds to its overall flexibility.

Let us develop working Example 1, to which we will refer in this chapter. (This example is based on the data set *country-e.sav*, and it includes transformations and screening steps that have already been performed as described later in this chapter.) Suppose we wanted to determine a country's status in terms of its level of development based on several measures. Specifically, those measures consist of

- population (1992) in millions (*pop92*),
- percentage of the population living in urban areas (*urban*),
- gross domestic product per capita (*gdp*),
- death rate per 1,000 individuals (*deathrat*),
- number of radios per 100 individuals (*radio*),
- number of hospital beds per 10,000 individuals (*hospbed*), and
- number of doctors per 10,000 (*docs*).

Combinations of these seven variables that would accurately predict the probability of a country's status as a developing country (*develop*)—either (0) developed country or (1) developing country—will be determined as a result of a logistic regression analysis. Following preliminary screening, our sample analysis was conducted using a forward method of entry for the seven predictors. A full discussion of the logic of this analysis is presented in Section 11.3.

The results obtained from a logistic regression analysis are somewhat different from those that we have seen from previous analysis techniques. There are basically three main output components to interpret. First, the resulting model is evaluated using goodness-of-fit tests. The table showing the results of chi-square goodness-of-fit tests for our working example is presented in Figure 11.1. Notice that the model resulted in the inclusion of three variables from the original seven predictors—*gdp* (entered at Step 1), *hospbed* (entered at Step 2), and *urban* (entered at Step 3). At each step, this test essentially compares the actual values for cases on the DV with the predicted values on the DV. All steps resulted in significance values $< .001$, indicating that these three variables are significant and important predictors of the DV *develop*. Also included in this table are the percentages of correct classification—based on the model—at each step (i.e., based on the addition of each variable). In Step 1, the model with only *gdp* correctly classified 92% of the cases. In Step 2, the model with *gdp* and *hospbed* correctly classified 92%. In Step 3, the model of *gdp*, *hospbed*, and *urban* correctly classified 95.5% of the cases.

Figure 11.1. Goodness-of-fit Indices for Example 1 (using Option: Display: At last step).

Step Summary ^{a,b}								
Step	Improvement			Model			Correct Class %	Variable
	Chi-square	df	Sig.	Chi-square	df	Sig.		
1	71.154	1	.000	71.154	1	.000	92.0%	IN: <i>gdp</i>
2	15.134	1	.000	86.289	2	.000	92.0%	IN: <i>hospbed</i>
3	9.875	1	.002	96.164	3	.000	95.5%	IN: <i>urban</i>

a. No more variables can be deleted from or added to the current model.
b. End block: 1

Three variables were entered into the model.

Final model significantly predicts group membership.

Figure 11.2 presents several indices for each step and the *overall* model fit. Smaller values on the first measure, labeled -2 Log likelihood, indicate that the model fits the data better. A perfect model has a value for this measure equal to 0 (George & Mallery, 2000). The second and third measures, Cox & Snell R Square and Nagelkerke R Square, are essentially estimates of R^2 indicating the proportion of variability in the DV that may be accounted for by all predictor variables included in the equation.

Figure 11.2. Model Summary for Example 1 (using Option: Display: At each step).

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	50.221 ^a	.470	.711
2	35.086 ^b	.537	.812
3	25.211 ^c	.576	.871

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.
b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.
c. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

Final model fit indices indicate fairly good fit.

The second component is a classification table for the DV. The classification table for *develop* is presented in Figure 11.3. The classification table compares the predicted values for the DV, based on the logistic regression model, with the actual observed values from the data. The predicted values are obtained by computing the probability for a particular case (this computation will be discussed in Section 11.3) and

classifying it into one of the two possible categories based on that probability. If the calculated probability is less than .50, the case is classified into the first value on the DV—in our example, the first category is *developed country* (coded 0).

The third and final component to be interpreted is the table of coefficients for variables included in the model. The coefficients for our working example are shown in Figure 11.4. These coefficients are interpreted in similar fashion to coefficients resulting from a multiple regression. The values labeled *B* are the regression coefficients or weights for each variable used in the equation. The significance of each predictor is tested not with a *t* test, as in multiple regression, but with a measure known as the *Wald statistic* and the associated significance value. The value *R* is the partial correlation coefficient between each predictor variable and the DV, holding constant all other predictors in the equation. Finally, *Exp(B)* provides an alternative method of interpreting the regression coefficients. The meaning of this coefficient will be explained further in Section 11.3.

Figure 11.3. Classification Table for Example 1 (using Option: Display: At each step).

Observed			Predicted		
			develop		Percentage Correct
			Developed country	Developing country	
Step 1	develop	Developed country	19	7	73.1
		Developing country	2	84	97.7
	Overall Percentage				92.0
Step 2	develop	Developed country	21	5	80.8
		Developing country	4	82	95.3
	Overall Percentage				92.0
Step 3	develop	Developed country	23	3	88.5
		Developing country	2	84	97.7
	Overall Percentage				95.5

a. The cutvalue is .500

Model is extremely accurate in classifying participants.

Figure 11.4. Regression Coefficients for Example 1 (using Option: Display: At each step).

Variables in the Equation						
	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	gdp	.000	.000	24.279	1	.000
	Constant	3.473	.551	39.679	1	.000
Step 2 ^b	gdp	.000	.000	14.396	1	.000
	hosptbed	-.048	.014	11.811	1	.001
Step 3 ^c	Constant	5.681	1.153	24.271	1	.000
	gdp	-.001	.000	6.604	1	.010
	hosptbed	-.083	.027	9.411	1	.002
	urban	.135	.067	4.078	1	.043
	Constant	3.047	1.252	5.927	1	.015

Odds ratios are fairly small.

a. Variable(s) entered on step 1: gdp.

b. Variable(s) entered on step 2: hosptbed.

c. Variable(s) entered on step 3: urban.

Sample Research Questions

The goal of logistic regression analysis is to correctly predict the category of outcome for individual cases. Further, attempts are made to reduce the number of predictors (in order to achieve parsimony) while maintaining a strong level of prediction. Based on our working example, let us now proceed to the specification of a series of possible research questions for our analysis:

1. Can status as a developing country (i.e., developed or developing) be correctly predicted from knowledge of population; percent of population living in urban areas; gross domestic product; death rate; and number of radios, hospital beds, and doctors?
2. If developing country status can be predicted correctly, which variables are central in the prediction of that status? Does the inclusion of a particular variable increase or decrease the probability of the specific outcome?
3. How good is the model at classifying cases for which the outcome is unknown? In other words, how many developing countries are classified correctly? How many developed countries are classified correctly?

SECTION 11.2 ASSUMPTIONS AND LIMITATIONS

As mentioned earlier, logistic regression does not require the adherence to any assumptions about the distributions of predictor variables (Tabachnick & Fidell, 2007). However, if distributional assumptions are met, discriminant analysis may be a stronger analysis technique. Thus, the researcher may want to opt for this procedure.

There are, however, several important issues related to the use of logistic regression. First, there is the issue of the ratio of cases to variables included in the analysis. Several problems may occur if too few cases relative to the number of predictor variables exist in the data. Logistic regression may produce extremely large parameter estimates and standard errors, especially in situations where combinations of discrete variables result in too many cells with no cases. If this situation occurs, the researcher is advised to collapse categories of the discrete variables, delete any offending categories (if patterns are evident), or simply delete any discrete variable if it is not important to the analysis (Tabachnick & Fidell, 2007). Another option open to the researcher is to increase the number of cases in the hope that this will fill in some of the empty cells.

Second, logistic regression relies on a goodness-of-fit test as a means of assessing the fit of the model to the data. You may recall from an earlier course in statistics that a goodness-of-fit test includes values for the expected frequencies for each cell in the data matrix formed by combinations of discrete variables. If any of the cells have expected frequencies that are too small (typically, $f_e < 5$), the analysis may have little power (Tabachnick & Fidell, 2007). All pairs of discrete variables should be evaluated to ensure that all cells have expected frequencies greater than 1 and that no more than 20% have frequencies less than 5. If either of these conditions fails, the researcher should consider accepting a lower level of power for the analysis, collapsing categories for variables with more than two levels, or deleting discrete variables so as to reduce the total number of cells (Tabachnick & Fidell, 2007).

Third, as with all varieties of multiple regression, logistic regression is sensitive to high correlations among predictor variables. This condition results in multicollinearity among predictor variables, as discussed in Chapter 7. If multicollinearity is present among variables in the analysis, one is advised to delete one or more of the redundant variables from the model in order to eliminate the multicollinear relationships (Tabachnick & Fidell, 2007).

Finally, extreme values on predictor variables should be examined carefully. As with multiple regression, resultant logistic regression models are very sensitive to outliers. A case that is actually in one outcome category may show a high probability for being in another category. Multiple cases such as this will result in a model with extremely poor fit. Standardized residuals should be examined in order to detect outliers. Any identified outliers—those cases with values $> |3|$ —should be addressed using standard methods (i.e., deletion from the sample).

SECTION 11.3 PROCESS AND LOGIC

The Logic Behind Logistic Regression

Mathematically speaking, logistic regression is based on probabilities, odds, and the logarithm of the odds (George & Mallery, 2000). Probabilities are simply the number of outcomes of a specific type expressed as a proportion of the total number of possible outcomes. For instance, if we roll a single die, the probability of rolling a three would be 1 out of 6—there is only one “3” on a die and there are six possible outcomes. This ratio could also be expressed as a proportion (.167) or a percentage (16.7%). In a logistic regression application, *odds* are defined as the ratio of the probability that an event will occur divided by the probability that the event will not occur. In other words,

$$Odds = \frac{p(X)}{1 - p(X)} \quad (\text{Equation 11.1})$$

where $p(X)$ is the probability of event X occurring and $1 - p(X)$ is the probability of event X not occurring. Therefore, the odds of rolling a 3 on a die are

$$Odds_{3} = \frac{p("3")}{1 - p("3")} = \frac{.167}{.833} = .200$$

It is important to keep in mind that probabilities will always have values that range from 0 to 1, but odds may be greater than 1. Applying the concept of odds to our working logistic regression example of classification as a developing country would give us the following equation:

$$Odds_{developing} = \frac{p(developing)}{1 - p(developing)}$$

The effect of an IV on a dichotomous outcome is usually represented by an odds ratio. The *odds ratio*—symbolized by ψ or $\text{Exp}(B)$ —is defined as a ratio of the odds of being classified in one category (i.e., $Y = 0$ or $Y = 1$) of the DV for two different values of the IV (Tate, 1992). For instance, we would be interested in the odds ratio for being classified as a “developing country” ($Y = 1$) for a given increase in the value of the score on the combination of the three significant predictors of *develop*—namely *urban*, *gdp*, and *hosbed*.

The ultimate model obtained by a logistic regression analysis is a nonlinear function (Tate, 1992). A key concept in logistic regression is known as a logit. A *logit* is the natural logarithm of the odds—an operation that most pocket calculators will perform. Again extending our simplified example, the logit for our odds of rolling a 3 would be

$$\ln(.200) = -1.609$$

Specifically, in logistic regression, \hat{Y} is the probability of having one outcome or another based on a nonlinear model resulting from the best linear combination of predictors. We can combine the ideas of probabilities, odds, and logits into one equation:

$$\hat{Y}_i = \frac{e^u}{1 + e^u} \quad (\text{Equation 11.2})$$

where \hat{Y}_i is the estimated probability that the i^{th} case is in one of the categories of the DV, and e is a constant equal to 2.718, raised to the power u , where u is the usual regression equation:

$$u = B_0 + B_1X_1 + B_2X_2 + \dots + B_kX_k \quad (\text{Equation 11.3})$$

The linear regression equation (u) is then the natural log of the probability of being in one group divided by the probability of being in the other group (Tabachnick & Fidell, 2007). The linear regression equation creates the logit or log of the odds:

$$\ln\left(\frac{\hat{Y}}{1 - \hat{Y}}\right) = B_0 + B_1X_1 + B_2X_2 + \dots + B_kX_k \quad (\text{Equation 11.4})$$

We tend to agree with George and Mallory (2000), who state in their text, “This equation is probably not very intuitive to most people...it takes a lot of experience before interpreting logistic regression equations becomes intuitive.” For most researchers, focus should more appropriately be directed at assessing the fit of the model, as well as its overall predictive accuracy.

Interpretation of Results

The output for logistic regression can be divided into three parts: the statistics for overall model fit, a classification table, and the summary of model variables. Although these components were introduced briefly in Section 11.1, a more detailed description will be presented here. The output for logistic regression looks considerably different from previous statistical methods because the output is presented in text and not in pivot tables. One should also keep in mind that the output can vary depending upon the stepping method utilized in the procedure. If a stepping method is applied, you have the option of presenting the output for each step or limiting the output to the last step. When output for each step is selected, the three components will be displayed for each step. Due to space constraints, our discussion of output and its subsequent interpretation will primarily be limited to output from the final step.

Several statistics for the overall model are presented in the first component of logistic regression output. The $-2 \text{ Log Likelihood}$ provides an index of model fit. A perfect model would have a $-2 \text{ Log Likelihood}$ of 0. Consequently, the lower this value, the better the model fits the data. This value actually represents the sum of the probabilities associated with the predicted and actual outcomes for each case (Tabachnick & Fidell, 2007). The next two values, Cox & Snell R Square and Nagelkerke R Square, represent two different estimates of the amount of variance in the DV accounted for by the model. Chi-square statistics with levels of significance are also computed for the model, block, and step. Chi-square for the model represents the difference between the constant-only model and the model generated. When using a stepwise method, the model generated will include only selected predictors. In contrast, the enter method generates a model with all the IVs included. Consequently, this comparison varies depending on the stepping method utilized. In general, a significant model chi-square indicates that the generated model is significantly better in predicting participant membership than the constant-only model. However, note that a large sample size increases the likelihood of finding significance when a poor-fitting model may have been generated. Chi-

square is also calculated for each step if a stepping method has been utilized. This value indicates the degree of model improvement when adding a selected predictor, or in other words, it represents the comparison between the model generated from the previous step to the current step.

The second component of output to interpret is the classification table. This table applies the generated regression model to predicting group membership. These predictions are then compared to the actual participant values. The percentage of participants correctly classified is calculated and serves as another indicator of model fit.

The third component of output is the summary of model variables. This summary presents several statistics—*B*, *S.E.*, *Wald*, *df*, *Sig.*, *R*, *Exp(B)*—for each variable included in the model as well as the constant. *B*, as in multiple regression, represents the unstandardized regression coefficient and represents the effect the IV has on the DV. *S.E.* is the standard error of *B*. *Wald* is a measure of significance for *B* and represents the significance of each variable in its ability to contribute to the model. Because several sources indicate that the *Wald* statistic is quite conservative (Tabachnick & Fidell, 2007), a more liberal significance level (i.e., $p < .05$ or $p < .1$) should be applied when interpreting this value. Degrees of freedom (*df*) and level of significance (*Sig.*) are also reported for the *Wald* statistic within the summary table. The partial correlation (*R*) of each IV with the DV (independent from the other model variables) is also presented. The final value presented in the summary table is *Exp(B)*, which is the calculated odds ratio for each variable. The odds ratio represents the increase (or decrease if *Exp(B)* is less than 1) in odds of being classified in a category when the predictor variable increases by 1.

Applying this process to our original example that utilizes *country-e.sav*, we sought to investigate which IVs (population, percentage of population living in urban areas, gross domestic product, death rate; and numbers of radios, hospital beds, and doctors) are predictors of status as a developing country (i.e., *developed* or *developing*). Because our investigation is exploratory in nature, we utilized the forward stepping method, such that only IVs that significantly predict the DV will be included in the model. Data were first screened for missing data and outliers. A preliminary multiple **Regression** was conducted to calculate Mahalanobis distance (to identify outliers) and examine multicollinearity among the seven predictors. Figure 11.5 presents the tolerance statistics for the seven predictors using *sequence* as the DV. Tolerance for all variables is greater than .1, indicating that multicollinearity is not a problem. The **Explore** procedure was then conducted to identify outliers (see Figure 11.6). Therefore, all cases in which the Mahalanobis value exceeded the chi-square criterion were eliminated using **Select Cases: If MAH_1 \leq 24.322**.

Figure 11.5. Tolerance Statistics for Example 1.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	48.105	12.241		3.930	.000		
urban	.116	.167	.080	.696	.488	.382	2.620
gdp	.002	.001	.366	3.048	.003	.352	2.840
deathrat	-.755	.674	-.094	-1.121	.265	.717	1.395
radio	-.197	.117	-.167	-1.683	.095	.517	1.934
hospbcd	.029	.129	.026	.226	.822	.368	2.719
docs	1.077	.449	.339	2.399	.018	.254	3.944
pop92	.019	.019	.073	.999	.320	.941	1.062

a. Dependent Variable: *sequence*

Tolerance for all variables exceeds .1. Multicollinearity is not a problem.

Figure 11.6. Outliers for Mahalanobis Distance (Example 1).

Extreme Values		
		Case Number
MAH_1	Highest	1
		2
		3
		4
		5
	Lowest	1
		2
		3
		4
		5

Eliminate cases that exceed $\chi^2(7) = 24.322$ at $p = .001$.



Binary Logistic Regression was then performed using the **Forward:LR** method. The three output components were presented in Figures 11.1 through 11.4. Figure 11.1 indicates that the three variables, *gdp*, *hospbed*, and *urban*, were entered into the overall model, which correctly classified 95.5% of the cases (see Figure 11.3). Figure 11.4 presents the summary of statistics for the model variables. Odds ratios, $Exp(B)$ or e^B , indicated that as the variable *urban* increases by 1, participants are 1.145 times more likely to be classified as *developing*. The odds ratios for *gdp* and *hospbed* were both below 1, indicating that as *gdp* ($e^B = .999$) and *hospbed* ($e^B = .920$) increase by 1, the odds of being classified as developing decrease by the respective ratio.

Writing Up Results

The results summary should always describe how variables have been transformed or deleted. The results for the overall model are reported within the narrative by first identifying the predictors entered into the model and addressing the following goodness-of-fit indices: -2 Log Likelihood and Model Chi-Square with degree of freedom and level of significance. The accuracy of classification should also be reported in the narrative. Finally, the regression coefficients for model variables should be presented in table and narrative format. The table should include *B*, *Wald*, *df*, level of significance, and odds ratio. The following results statement applies the example presented in Figures 11.1 through 11.4.

Forward logistic regression was conducted to determine which independent variables (population; percentage of population living in urban areas; gross domestic product; death rate; and numbers of radios, hospital beds, and doctors) were predictors of status as a developing country (developed or developing). Data screening led to the elimination of three outliers. Regression results indicated that the overall model of three predictors (*gdp*, *hospbed*, and *urban*) was statistically reliable in distinguishing between developed and developing countries [-2 Log Likelihood = 25.211, $\chi^2(2) = 96.164$, $p < .0001$]. The model correctly classified 95.5% of the cases. Regression coefficients are presented in Table 1. Wald statistics indicated that all variables significantly predict country development. However, odds ratios for these variables indicate little change in the likelihood of country development.

Table 1*Regression Coefficients*

	<i>B</i>	<i>Wald</i>	<i>df</i>	<i>p</i>	Odds Ratio
Urban	.135	4.08	1	.043	1.145
GDP	-.001	6.60	1	.010	.999
Hospital beds	-.083	9.41	1	.002	.920
Constant	3.047	5.93	1	.015	

SECTION 11.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example of the process of conducting logistic regression. This process includes the development of research questions, data-screening methods, analysis methods, interpretation of output, and presentation of results. The example utilizes the data set *profile-d.sav* from the website that accompanies this book (see p. *xiii*).

Problem

In the previous chapter on discriminant analysis, the second example investigated the ability of six IVs (age, hours worked per week, years of education, income, number of siblings, and number of hours spent watching TV per day) to predict one's life perspective (dull, routine, or exciting). For this example, we will utilize the same scenario. However, the DV will be recoded as dichotomous to fulfill the requirement of binary logistic regression. Because six IVs are being investigated, the forward stepping method will be applied. The following research question is generated to address this scenario:

Can life (*life*) perspective (dull, routine, or exciting) be reliably predicted from the knowledge of an individual's age (*age*), hours worked per week (*hrs1*), years of education (*educ*), income (*rincom91*), number of siblings (*sibs*), and number of hours spent watching TV per day (*tvhours*)?

Methods and SPSS "How To"

Prior to analysis, the variable of (*life*) was recoded as dichotomous (*life2*) and the following transformations were applied: 0 = missing, 1–2 = 0, 3 = 1, 8–9 = missing. Data were screened for missing data and outliers. A preliminary multiple **Linear Regression** was conducted to calculate Mahalanobis distance and to evaluate multicollinearity among the six continuous predictors. The table of regression coefficients (see Figure 11.7) indicates that multicollinearity was not violated because tolerance statistics for all six IVs were greater than .1. **Explore** was then conducted to determine which cases exceeded the chi-square criteria of $\chi^2(6) = 22.458$ at $p = .001$ (see Figure 11.8). Therefore, all cases in which the Mahalanobis value exceeds the chi-square criterion need to be eliminated using **Select Cases: If MAH_1 ≤ 22.458** .

Figure 11.7. Tolerance Statistics for Example 2.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	1026.574	113.922		9.011	.000		
age	-1.618	1.304	-.045	-1.241	.215	.893	1.119
hrs1	2.897	1.169	.096	2.478	.013	.763	1.310
educ	-12.739	5.841	-.081	-2.181	.029	.824	1.213
rincom91	-11.601	3.403	-.143	-3.409	.001	.655	1.527
sibs	-10.188	5.648	-.063	-1.804	.072	.947	1.056
tvhours	3.277	8.660	.013	.378	.705	.918	1.089

a. Dependent Variable: id

Tolerance for all variables exceeds .1. Multicollinearity is not a problem.

Figure 11.8. Outliers for Mahalanobis Distance (Example 2).

Extreme Values			
		Case Number	Value
MAH_1	Highest	1	466 126.80049
		2	1360 114.05523
		3	406 56.53054
		4	50 54.06832
		5	121 42.50807
	Lowest	1	649 .25234
		2	561 .26159
		3	734 .30765
		4	1032 .45789
		5	266 .48138

Eliminate cases that exceed $\chi^2(6) = 22.458$ at $p = .001$.

Binary Logistic Regression was then conducted using the Forward: LR method. To conduct **Binary Logistic Regression**, select the following menus:

Analyze
Regression
Binary Logistic

Logistic Regression dialog box (see Figure 11.9)

Once in this dialog box, identify the DV (*life2*) and move it to the **Dependent** box. Identify the six IVs and move each to the **Covariates** box. Next, select the desired regression method. SPSS provides seven different methods, five of which are described as follows:

Enter—Enters all the IVs at once into the model, regardless of significant contribution. This method is useful if you have previously tested the IVs and want all of them to be entered.

Forward: LR—One of the most common methods. Enters IVs one at a time. The likelihood-ratio is used to determine variable selection.

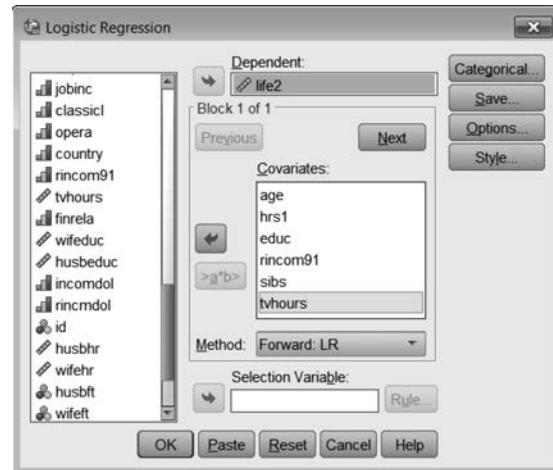
Forward: Wald—Enters IVs one at a time. The *Wald* statistic is used to determine variable selection.

Backward: LR—All IVs are entered at once, then variables are removed one at a time. The likelihood-ratio is used to determine variable removal.

Backward: Wald—All IVs are entered at once, then variables are removed one at a time. The *Wald* statistic is used to determine variable removal.

For our example, we selected **Forward: LR**. Next, click **Categorical**.

Figure 11.9. Logistic Regression Dialog Box.



Logistic Regression: Define Categorical Variables dialog box (see Figure 11.10)

By default in logistic regression, SPSS treats any numerical variable as continuous. Consequently, when an IV is categorical, you need to specify how SPSS should address it. Once in this dialog box, identify any categorical variables and move them to the **Categorical Covariates** box. Then, under **Change Contrast**, select the method of contrast. SPSS provides several contrast methods. The **Indicator** method is the default and is the most common. Three of the contrasting methods are described as follows:

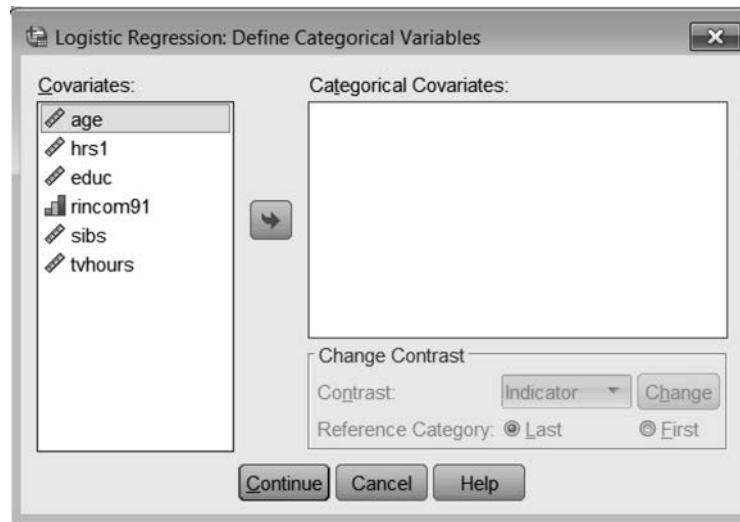
Indicator—Indicates the presence or absence of group membership. This is the default.

Simple—Each category of the IV (except the reference category) is compared to the reference category.

Deviation—Each category of the IV (except the reference category) is compared to the overall effect.

If you have selected one of these contrasting methods, you can identify a specific category to be used as the **Reference Category**. Two options are available: **Last** category (the default) or **First** category. If you have selected any options other than defaults for contrasting methods or reference category, you must then click **Change**. Because all numerical variables in the present example are continuous, we will exit the **Logistic Regression: Define Categorical Variables** window without making any changes. Click **Cancel**, then **Options**.

Figure 11.10. Logistic Regression: Define Categorical Variables Dialog Box.



Logistic Regression: Options dialog box (see Figure 11.11)

SPSS provides several options within logistic regression. Commonly used options are described as follows:

Classification plots—Graph of actual and predicted values for the DV.

Correlations of estimates—Correlation matrix of parameter estimates for model variables.

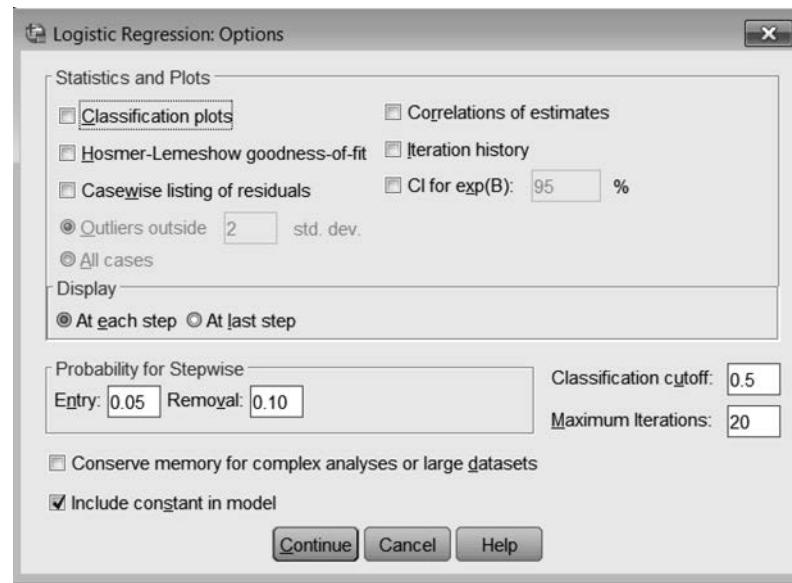
Iteration history—Presents coefficients and log likelihood at each iteration.

CI for exp(B)—Calculates the confidence intervals for the odds ratios of each model variable.

You can indicate the level of probability associated with this interval. The default is 95%.

For our example, we will not select any of the above options. A **Display** option is also available if a stepping method has been utilized. Your choice will affect how tables are generated in the SPSS output. If you select **At last step**, a step summary (see Figure 11.12) will be created. In order to display the tables shown in Figures 11.13 through 11.15, you will need to run the analysis again, selecting instead, **At each step**, as we did in Figure 11.11. Other options available are the probability for stepwise, classification cutoff, the maximum number of iterations, and inclusion of constant in the model. We maintained the defaults for these options. However, there may be times when it is necessary to increase the maximum number of iterations in order to generate a complete model. The reader should note that SPSS also provides options for saving variables. By clicking **Save**, you can save predicted values, residuals, and so on, as new variables. For our example, we will skip this step. Once you have selected the appropriate options, click **Continue**, then **OK**.

Figure 11.11. Logistic Regression: Options Dialog Box.



Output and Interpretation of Results

The three components of output are presented in Figures 11.12 through 11.15. The statistics for overall model fit are presented in Figure 11.12 and indicate that only two variables were entered into the model: *educ* and *rincom91*. Model fit statistics as seen in Figure 11.13 are extremely large and reveal a poor-fitting model [$-2 \text{ Log likelihood} = 748.595$]. The generated model was significantly different from the constant-only model [$\chi^2(1) = 33.098, p < .0001$]. Figure 11.14 presents the classification table and indicates that the model correctly classified only 59.6% of participants. The summary of model variables is displayed in Figure 11.15. Odds ratios for the *educ* ($e^B = 1.161$) and *rincom91* ($e^B = 1.040$) revealed little increase in the likelihood of perceiving life as exciting when the predictors increase by 1.

Figure 11.12. Goodness-of-fit Indices for Example 2 (Display: At last step selected).

Step	Improvement			Model			Correct Class %	Variable
	Chi-square	df	Sig.	Chi-square	df	Sig.		
1	27.898	1	.000	27.898	1	.000	59.0%	IN: educ
2	5.200	1	.023	33.098	2	.000	59.6%	IN: rincom91

a. No more variables can be deleted from or added to the current model.
b. End block: 1

Two variables were entered into the model.

Model significantly predicts group membership.

Figure 11.13. Model Summary for Example 2 (Display: At each step selected).

Model Summary				
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square	
1	753.795 ^a	.048	.064	
2	748.595 ^b	.057	.076	

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Model fit indices are extremely large. Fit is questionable.

Figure 11.14. Classification Table for Example 2 (Display: At each step selected).

Observed		Predicted			
		life2		Percentage Correct	
		.00	1.00		
Step 1	life2	.00	159	118	57.4
		1.00	113	174	60.6
	Overall Percentage				59.0
Step 2	life2	.00	170	107	61.4
		1.00	121	166	57.8
	Overall Percentage				59.6

a. The cut value is .500

Model is fairly accurate in classifying participants.

Figure 11.15. Regression Coefficients for Example 2 (Display: At each step selected).

Variables in the Equation						
	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	educ	.173	.034	25.935	1	.000
	Constant	-2.378	.481	24.434	1	.000
Step 2 ^b	educ	.149	.036	17.548	1	.000
	rincom91	.039	.017	5.148	1	.023
	Constant	-2.574	.493	27.289	1	.000

a. Variable(s) entered on step 1: educ.

b. Variable(s) entered on step 2: rincom91.

Odds ratios are fairly small.

Presentation of Results

Forward logistic regression was conducted to determine which independent variables (age, hours worked per week, years of education, income, number of siblings, and number of hours spent watching TV) are predictors of life perspective (dull, routine, or exciting). Data screening led to the elimination of several outliers. Regression results indicated that the overall model fit of two predictors (education and income) was questionable ($-2 \text{ Log likelihood} = 748.595$) but was statistically reliable in distinguishing between life perspective [$\chi^2(1) = 33.098, p < .0001$]. The model correctly classified only 59.6% of the cases. Regression coefficients are presented in Table 2. *Wald* statistics indicated that education and income significantly predict life perspective. However, odds ratios for these variables indicated little change in the likelihood of perceiving life as exciting.

Table 2

Regression Coefficients

	<i>B</i>	<i>Wald</i>	<i>df</i>	<i>p</i>	Odds Ratio
Education	.149	17.55	1	< .0001	1.161
Income	.039	5.15	1	.023	1.040
Constant	-2.574	27.29	1	< .0001	

SUMMARY

Logistic regression tests the ability of a model or group of variables to predict group membership as defined by some categorical DV. In binary logistic regression, the DV must be dichotomous, but the IVs may be categorical or continuous. Logistic regression actually predicts the probability of membership occurring, which varies from 0 to 1. A variety of methods can be used to test and develop different models (Enter, Forward: LR, Backward: Wald, etc.). Although logistic regression requires fulfillment of few test assumptions, data should be screened for outliers and multicollinearity. Logistic regressions output includes three parts: statistics for overall model fit, classification table, and summary of model variables. Statistics for overall model fit provide several indices of model fit: $-2 \text{ Log likelihood}$, Cox & Snell R Square, Nagelkerke R Square, and Model Chi-Square. The classification table presents the percentage of cases correctly classified with the generated model. The summary of model variables provides several variable statistics that indicate variable contribution to the model: *B*, *Wald*, *df*, level of significance, and odds ratio. A good-fitting model will typically have: fairly low values for $-2 \text{ Log likelihood}$, significant model chi-square, and variables with odds ratios greater than 1. Figure 11.16 provides a checklist for conducting binary logistic regression.

KEYWORDS

- logit
- odds
- odds ratio
- Wald statistic

Figure 11.16. Checklist for Conducting Binary Logistic Regression.

I. Screen Data

- a. Missing Data?
- b. Multivariate Outliers and Multicollinearity?
 - Run preliminary Linear Regression.
 1. **Analyze... Regression... Linear.**
 - Identify a variable that serves as a case number and move to **Dependent Variable** box.
 - Identify all appropriate quantitative variables and move to **Independent(s)** box.
 2. **Statistics.**
 - Check **Collinearity diagnostics**.
 - Continue.**
 3. **Save.**
 - Check **Mahalanobis** under **Distances**.
 4. **Continue, OK.**
 5. Determine chi-square (χ^2) critical value at $p < .001$.
 - Conduct **Explore** to test outliers for Mahalanobis chi-square (χ^2).
 1. **Analyze... Descriptive statistics... Explore.**
 - Move **MAH_1** to **Dependent List** box.
 - Leave **Factor** box empty.
 2. **Statistics.**
 - Check **Outliers**.
 3. **Continue, OK.**
 - Delete outliers for participants when χ^2 exceeds critical χ^2 at $p < .001$.

II. Conduct Logistic Regression

- a. Run Binary Logistic Regression using **Regression**.
 1. **Analyze... Regression... Binary Logistic....**
 - Move the DV to the **Dependent** box.
 - Move IVs to the **Covariates** box.
 - Select Method.
 2. **Categorical** (if any IVs are categorical).
 - Move any categorical IVs to the **Categorical Covariates** box.
 - Select **Contrast Method** and **Reference Category**.
 - Continue.**
 3. **Options.**
 - Check appropriate options.
 4. **Continue, OK.**

III. Summarize Results

- a. Describe any data elimination or transformation.
- b. Describe the model generated ($-2 \text{ Log Likelihood}$, Goodness-of-Fit, Model chi-square with df and p value).
- c. Report the accuracy of classification.
- d. Present the regression coefficients for model variables in table format.
- e. Report odds ratios for model variables.
- f. Draw conclusions.

Exercises for Chapter 11

This exercise utilizes the SPSS data set *profile-e.sav*, which can be downloaded from this website:

www.routledge.com/9781138289734

Conduct a **Forward: LR** logistic regression analysis with the following variables:

IV—*age, educ, hrsl, sibs, rincom91, life2* (categorical)
DV—*satjob2*

Note: The variable *life2* is categorical such that dull = 1, routine/exciting = 2, and all other values are system missing.

1. Develop a research question for the preceding scenario.
2. Conduct a preliminary **Linear Regression** to identify outliers and evaluate multicollinearity among the five continuous variables. Complete the following:
 - a. Using the Chi-Square table in Appendix B near the end of this book, identify the critical value at $p < .001$ for identifying outliers. Use **Explore** to determine if there are outliers. Which cases should be eliminated?
 - b. Is multicollinearity a problem among the five continuous variables?

3. Conduct **Binary Logistic Regression** using the **Forward: LR** method.

IV—*age, educ, hrsl, sibs, rincom91, life2* (categorical; last is the reference category)
DV—*satjob2*

Note: Make sure that any outliers identified in Exercise 2.a. are removed from data before running the logistic regression. Also, designating *life2* as a categorical covariate with the last category as the reference, essentially makes routine/exciting = 0 and dull = 1, so interpret the results accordingly.

- a. Which variables were entered into the model?
- b. To what degree does the model fit the data? Explain.
- c. Is the generated model significantly different from the constant-only model?

- d. How accurate is the model in predicting job satisfaction?
- e. What are the odds ratios for the model variables? Explain.

Appendix A: SPSS Data Sets (www.routledge.com/9781138289734)

career-a.sav	Used in Chapters 3, 5, and 6		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*) ¹	MEASURE	VALUES
age	Age of respondent	Scale	
educ	(*)Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
sex	Respondent's sex	Nominal	1 = male; 2 = female
degree	Respondent's highest degree	Ordinal	0 = less than HS; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
income4	Total family income in quartiles	Ordinal	1 = \$24,999 or less; 2 = \$25,000 to 39,999; 3 = \$40,000 to 59,999; 4 = \$60,000 or more
rincom91	Respondent's income	Scale	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+; 22 = refused; 98 = don't know; 99 = no answer
hrs1	(*)Number of hours worked last week	Scale	98 = don't know; 99 = no answer; -1 = not applicable
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unempl, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
jobinc	Importance of high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
bothft	Both spouses work full-time	Nominal	0 = no; 1 = yes
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
husbhr	Hrs worked last week by husband	Scale	
wifeft	Wife employed full-time	Nominal	0 = no; 1 = yes
wifehr	Hrs worked last week by wife	Scale	
id		Nominal	
agecat4	4 categories of age	Ordinal	1 = 18–29; 2 = 30–39; 3 = 40–49; 4 = 50+

End of Data Set

¹ Entries preceded by an asterisk (*) for all data sets in this appendix are *descriptions* added to the appendix for reference only. They do not appear in the data set file or in the figures as *labels*, as do all entries *not* marked with an asterisk.

career-b.sav	Used in Chapter 3		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	Age of respondent	Scale	
educ	(*)Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
sex	Respondent's sex	Nominal	1 = male; 2 = female
degree	Respondent's highest degree	Ordinal	0 = less than HS; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
income4	Total family income in quartiles	Ordinal	1 = \$24,999 or less; 2 = \$25,000 to 39,999; 3 = \$40,000 to 59,999; 4 = \$60,000 or more
rincom91	Respondent's income	Scale	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+; 22 = refused; 98 = don't know; 99 = no answer
hrs1	(*)Number of hours worked last week	Scale	98 = don't know; 99 = no answer; -1 = not applicable
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part time; 3 = temp not working; 4 = unempl, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
jobinc	Importance of high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
bothft	Both spouses work full-time	Nominal	0 = no; 1 = yes
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
husbhr	Hrs worked last week by husband	Scale	
wifeft	Wife employed full-time	Nominal	0 = no; 1 = yes
wifehr	Hrs worked last week by wife	Scale	
id		Nominal	
agecat4	4 categories of age	Ordinal	1 = 18–29; 2 = 30–39; 3 = 40–49; 4 = 50+
rincom2		Scale	
rincom3		Scale	
rincom4		Scale	
age2	Square root of age	Scale	

End of Data Set

career-c.sav	Used in Chapter 3		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	Age of respondent	Scale	
educ	(*)Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
sex	Respondent's sex	Nominal	1 = male; 2 = female
degree	Respondent's highest degree	Ordinal	0 = less than HS; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
income4	Total family income in quartiles	Ordinal	1 = \$24,999 or less; 2 = \$25,000 to 39,999; 3 = \$40,000 to 59,999; 4 = \$60,000 or more
rincom91	Respondent's income	Scale	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+; 22 = refused; 98 = don't know; 99 = no answer
hrs1	(*)Number of hours worked last week	Scale	98 = don't know; 99 = no answer; -1 = not applicable
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unempl, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
jobinc	Importance of high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
bothft	Both spouses work full-time	Nominal	0 = no; 1 = yes
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
husbhr	Hrs worked last week by husband	Scale	
wifeft	Wife employed full-time	Nominal	0 = no; 1 = yes
wifehr	Hrs worked last week by wife	Scale	
id		Nominal	
agecat4	4 categories of age	Ordinal	1 = 18–29; 2 = 30–39; 3 = 40–49; 4 = 50+
rincom2	(*)rincom91 transformed to eliminate coding problem	Scale	22 through highest = system missing
rincom3	(*)Transformed to eliminate outliers	Scale	3 or less = 4
rincom4	(*)Transformed to increase normality	Scale	rincom4 = $\text{SQRT}(22 - \text{rincom3})$
MAH_1	Mahalanobis distance ([*]for rincom4, age, educ, hrs1)	Scale	
filter_\$	Transformed to increase normality	Scale	
age2	(*)rincom91 transformed to eliminate coding problem	Scale	age2 = $\text{SQRT}(\text{age})$

End of Data Set

career-d.sav	Used in Chapter 4		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	Age of respondent	Scale	
educ	(*)Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
sex	Respondent's sex	Nominal	1 = male; 2 = female
degree	Respondent's highest degree	Ordinal	0 = less than HS; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
income4	Total family income in quartiles	Ordinal	1 = \$24,999 or less; 2 = \$25,000 to 39,999; 3 = \$40,000 to 59,999; 4 = \$60,000 or more
rincom91	Respondent's income	Scale	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+; 22 = refused; 98 = don't know; 99 = no answer
hrs1	(*)Number of hours worked last week	Scale	98 = don't know; 99 = no answer; -1 = not applicable
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unempl, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
jobinc	Importance of high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
bothft	Both spouses work full-time	Nominal	0 = no; 1 = yes
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
husbhr	Hrs worked last week by husband	Scale	
wifeft	Wife employed full-time	Nominal	0 = no; 1 = yes
wifehr	Hrs worked last week by wife	Scale	
id		Nominal	
agecat4	4 categories of age	Ordinal	1 = 18–29; 2 = 30–39; 3 = 40–49; 4 = 50+
rincom2	(*)rincom91 transformed to eliminate coding problem	Scale	22 through highest = system missing
rincom3	(*)Transformed to eliminate outliers	Scale	3 or less = system missing

End of Data Set

career-e.sav	Used in Chapter 5		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	Age of respondent	Scale	
educ	(*)Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
sex	Respondent's sex	Nominal	1 = male; 2 = female
degree	Respondent's highest degree	Ordinal	0 = less than HS; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
income4	Total family income in quartiles	Ordinal	1 = \$24,999 or less; 2 = \$25,000 to 39,999; 3 = \$40,000 to 59,999; 4 = \$60,000 or more
rincom91	Respondent's income	Scale	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+; 22 = refused; 98 = don't know; 99 = no answer
hrs1	(*)Number of hours worked last week	Scale	98 = don't know; 99 = no answer; -1 = not applicable
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unempl, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
jobinc	Importance of high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
bothft	Both spouses work full-time	Nominal	0 = no; 1 = yes
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
husbhr	Hrs worked last week by husband	Scale	
wifeft	Wife employed full time	Nominal	0 = no; 1 = yes
wifehr	Hrs worked last week by wife	Scale	
id		Nominal	
agecat4	4 categories of age	Ordinal	1 = 18–29; 2 = 30–39; 3 = 40–49; 4 = 50+
rincom2	(*)rincom91 transformed to eliminate coding problem	Scale	22 through highest = system missing
rincom3		Scale	
hrs2	(*)Transformed to eliminate outliers	Scale	16 or less = 17; 80 or greater = 79

End of Data Set

career-f.sav	Used in Chapter 6		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	Age of respondent	Scale	
educ	(*)Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
sex	Respondent's sex	Nominal	1 = male; 2 = female
degree	Respondent's highest degree	Ordinal	0 = less than HS; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
income4	Total family income in quartiles	Ordinal	1 = \$24,999 or less; 2 = \$25,000 to 39,999; 3 = \$40,000 to 59,999; 4 = \$60,000 or more
rincom91	Respondent's income	Scale	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+; 22 = refused; 98 = don't know; 99 = no answer
hrs1	(*)Number of hours worked last week	Scale	98 = don't know; 99 = no answer; -1 = not applicable
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unempl, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
jobinc	Importance of high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
bothft	Both spouses work full-time	Nominal	0 = no; 1 = yes
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
husbhr	Hrs worked last week by husband	Scale	
wifeft	Wife employed full-time	Nominal	0 = no; 1 = yes
wifehr	Hrs worked last week by wife	Scale	
id		Nominal	
agecat4	4 categories of age	Ordinal	1 = 18–29; 2 = 30–39; 3 = 40–49; 4 = 50+
rincom2	(*)rincom91 transformed to eliminate coding problem	Scale	22 through highest = system missing
hrs2	(*)Transformed to eliminate outliers	Scale	16 or less = 17; 80 or greater = 79
educ2	(*)Transformed to eliminate outliers	Scale	6 or less = system missing
End of Data Set			

country-a.sav	Used in Chapters 7 and 8		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
country	country	Nominal	
pop92	pop92	Scale	
urban	urban	Scale	
gdp	gdp	Scale	
lifeexpm	lifeexpm	Scale	
lifeexpf	lifeexpf	Scale	
birthrat	birthrat	Scale	
deathrat	deathrat	Scale	
infmr	infmr	Scale	
fertrate	fertrate	Scale	
region	region	Nominal	1 = Eastern Africa; 2 = Middle Africa; 3 = Northern Africa; 4 = Southern Africa; 5 = Western Africa; 6 = Caribbean; 7 = Central America; 8 = South America; 9 = North America; 10 = Eastern Asia; 11 = Southeast Asia; 12 = Southern Asia; 13 = Western Asia; 14 = Eastern Europe; 15 = Northern Europe; 16 = Southern Europe; 17 = Western Europe; 18 = Oceania; 19 = USSR
develop	develop	Nominal	0 = developed country; 1 = developing country
radio	radio	Scale	
phone	phone	Scale	
hospbed	hospbed	Scale	
docs	docs	Scale	
lndocs	lndocs	Scale	
lnradio	lnradio	Scale	
lnphone	lnphone	Scale	
lngdp	lngdp	Scale	
sequence	sequence	Nominal	
lnbeds	lnbeds	Scale	
MAH_1	Mahalanobis distance ([*]for urban, gdp, hospbed, docs, radio, phone, lifeexpf)	Scale	
MAH_2	Mahalanobis distance ([*]for urban, gdp, hospbed, docs, radio, phone, lifeexpm)	Scale	
End of Data Set			

country-b.sav	Used in Chapters 8 and 9		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
country	country	Nominal	
pop92	pop92	Scale	
urban	urban	Scale	
gdp	gdp	Scale	
lifeexpm	lifeexpm	Scale	
lifeexpf	lifeexpf	Scale	
birthrat	birthrat	Scale	
deathrat	deathrat	Scale	
infmr	infmr	Scale	
fertrate	fertrate	Scale	
region	region	Nominal	1 = Eastern Africa; 2 = Middle Africa; 3 = Northern Africa; 4 = Southern Africa; 5 = Western Africa; 6 = Caribbean; 7 = Central America; 8 = South America; 9 = North America; 10 = Eastern Asia; 11 = Southeast Asia; 12 = Southern Asia; 13 = Western Asia; 14 = Eastern Europe; 15 = Northern Europe; 16 = Southern Europe; 17 = Western Europe; 18 = Oceania; 19 = USSR
develop	develop	Nominal	0 = developed country; 1 = developing country
radio	radio	Scale	
phone	phone	Scale	
hospbed	hospbed	Scale	
docs	docs	Scale	
lndocs	lndocs	Scale	
lnradio	lnradio	Scale	
lnphone	lnphone	Scale	
lngdp	lngdp	Scale	
sequence	sequence	Nominal	
lnbeds	lnbeds	Scale	
MAH_1	(*)Mahalanobis distance for region, develop, deathrat, docs, lifeexpm	Scale	
filter_\$	MAH_1 <= 20.516 (filter)	Scale	
MAH_2	(*)Mahalanobis distance for docs, gdp, deathrat, birthrat, infmr	Scale	
End of Data Set			

country-c.sav	Used in Chapter 8		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
country	country	Nominal	
pop92	pop92	Scale	
urban	urban	Scale	
gdp	gdp	Scale	
lifeexpm	lifeexpm	Scale	
lifeexpf	lifeexpf	Scale	
birthrat	birthrat	Scale	
deathrat	deathrat	Scale	
infmr	infmr	Scale	
fertrate	fertrate	Scale	
region	region	Nominal	1 = Eastern Africa; 2 = Middle Africa; 3 = Northern Africa; 4 = Southern Africa; 5 = Western Africa; 6 = Caribbean; 7 = Central America; 8 = South America; 9 = North America; 10 = Eastern Asia; 11 = Southeast Asia; 12 = Southern Asia; 13 = Western Asia; 14 = Eastern Europe; 15 = Northern Europe; 16 = Southern Europe; 17 = Western Europe; 18 = Oceania; 19 = USSR
develop	develop	Nominal	0 = developed country; 1 = developing country
radio	radio	Scale	
phone	phone	Scale	
hospbed	hospbed	Scale	
docs	docs	Scale	
lndocs	lndocs	Scale	
lnradio	lnradio	Scale	
lnphone	lnphone	Scale	
lngdp	lngdp	Scale	
sequence	sequence	Nominal	
lnbeds	lnbeds	Scale	
End of Data Set			

country-d.sav	Used in Chapter 10		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
country	country	Nominal	
pop92	pop92	Scale	
urban	urban	Scale	
gdp	gdp	Scale	
lifeexpm	lifeexpm	Scale	
lifeexpf	lifeexpf	Scale	
birthrat	birthrat	Scale	
deathrat	deathrat	Scale	
infmr	infmr	Scale	
fertrate	fertrate	Scale	
region	region	Nominal	1 = Eastern Africa; 2 = Middle Africa; 3 = Northern Africa; 4 = Southern Africa; 5 = Western Africa; 6 = Caribbean; 7 = Central America; 8 = South America; 9 = North America; 10 = Eastern Asia; 11 = Southeast Asia; 12 = Southern Asia; 13 = Western Asia; 14 = Eastern Europe; 15 = Northern Europe; 16 = Southern Europe; 17 = Western Europe; 18 = Oceania; 19 = USSR
develop	develop	Nominal	0 = developed country; 1 = developing country
radio	radio	Scale	
phone	phone	Scale	
hospbed	hospbed	Scale	
docs	docs	Scale	
lndocs	lndocs	Scale	
lnradio	lnradio	Scale	
lnphone	lnphone	Scale	
lngdp	lngdp	Scale	
sequence	sequence	Nominal	
lnbeds	lnbeds	Scale	
MAH_1	Mahalanobis distance ([*]for develop, urban, gdp, lifeexpm, lifeexpf, infmr)	Scale	
filter_ \$	MAH_1 <= 20.515 (filter)	Scale	
End of Data Set			

country-e.sav	Used in Chapter 11		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
country	country	Nominal	
pop92	Population, 1992, in millions	Scale	
urban	Percentage urban, 1992	Scale	
gdp	GDP per capita	Scale	
lifeexpm	Male life expectancy 1992	Scale	
lifeexpf	Female life expectancy 1992	Scale	
birthrat	Births per 1000 population, 1992	Scale	
deathrat	Deaths per 1000 individuals, 1992	Scale	
infmr	Infant mortality rate 1992 (per 1000 live births)	Scale	
fertrate	Fertility rate per woman, 1990	Scale	
region	Region of the world	Nominal	1 = Eastern Africa; 2 = Middle Africa; 3 = Northern Africa; 4 = Southern Africa; 5 = Western Africa; 6 = Caribbean; 7 = Central America; 8 = South America; 9 = North America; 10 = Eastern Asia; 11 = Southeast Asia; 12 = Southern Asia; 13 = Western Asia; 14 = Eastern Europe; 15 = Northern Europe; 16 = Southern Europe; 17 = Western Europe; 18 = Oceania; 19 = USSR
develop	develop	Nominal	0 = developed country; 1 = developing country
radio	Radios per 100 individuals	Scale	
phone	Phones per 100 individuals	Scale	
hospbed	Hospital beds per 10,000 individuals	Scale	
docs	Doctors per 10,000 individuals	Scale	
lndocs	Natural log of doctors per 10,000	Scale	
lnradio	Natural log of radios per 100 individuals	Scale	
lnphone	Natural log of phones per 100 individuals	Scale	
lngdp	Natural log of GDP	Scale	
sequence	Arbitrary sequence number	Nominal	
lnbeds	Natural log hospital beds/10,000	Scale	
MAH_1	Mahalanobis distance ([*]for urban, gdp, deathrat, radio, hospbed, docs, pop92)	Scale	
filter_ \$	MAH_1 <= 24.322 (filter)	Scale	
End of Data Set			

profile-a.sav	Used in Chapters 4, 7, and 9		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	Age of respondent	Scale	
sex	Respondent's sex	Nominal	1 = male; 2 = female
educ	Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
income91	Total family income	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = 12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unemployed/laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
richwork	If rich, continue or stop working	Nominal	0 = not applicable; 1 = continue working; 2 = stop working; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
life	Is life exciting or dull	Ordinal	0 = not applicable; 1 = dull; 2 = routine; 3 = exciting; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
hrs1	Number of hours worked last week	Scale	–1 = not applicable; 98 = don't know; 99 = no answer
degree	Respondent's highest degree	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
anomia5	Lot of average man getting worse	Nominal	0 = not applicable; 1 = agree; 2 = disagree; 8 = don't know; 9 = no answer
degree2	College degree	Nominal	0 = no college degree; 1 = college degree; 7 = not applicable; 8 = don't know; 9 = no answer
maeduc	Highest year of school completed, mother	Scale	97 = not applicable; 98 = don't know; 99 = no answer
paeduc	Highest year of school completed, father	Scale	97 = not applicable; 98 = don't know; 99 = no answer
macolleg	Mother a college grad	Nominal	0 = no; 1 = yes
pacolleg	Father a college grad	Nominal	0 = no; 1 = yes
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
marital	Marital status	Nominal	1 = married; 2 = widowed; 3 = divorced; 4 = separated; 5 = never married; 9 = no answer
agewed	Age when first married	Scale	0 = not applicable; 98 = don't know; 99 = no answer
spwrksta	Spouse labor force status	Nominal	0 = not applicable; 1 = working full time; 2 = working part time; 3 = temp not working; 4 = unemployed, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
sphrs1	Number of hrs spouse worked last week	Scale	–1 = not applicable; 98 = don't know; 99 = no answer

profile-a.sav	Used in Chapters 4, 7, and 9		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
sibs	Number of brothers and sisters	Scale	-1 = not applicable; 98 = don't know; 99 = no answer
zodiac	Respondent's astrological sign	Nominal	0 = not applicable; 1 = Aries; 2 = Taurus; 3 = Gemini; 4 = Cancer; 5 = Leo; 6 = Virgo; 7 = Libra; 8 = Scorpio; 9 = Sagittarius; 10 = Capricorn; 11 = Aquarius; 12 = Pisces; 98 = don't know; 99 = no answer
speduc	Highest year school completed, spouse	Scale	97 = not applicable; 98 = don't know; 99 = no answer
spdeg	Spouse's highest degree	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
partyid	Political party affiliation	Ordinal	0 = strong Democrat; 1 = not strong Democrat; 2 = Ind, near Dem; 3 = Independent; 4 = Ind, near Rep; 5 = not strong Republican; 6 = strong Republican; 7 = other party; 8 = don't know; 9 = no answer
vote92	Did Respondent vote in 1992 election	Nominal	0 = not applicable; 1 = voted; 2 = did not vote; 3 = not eligible; 4 = refused; 8 = don't know; 9 = no answer
pres92	Vote for Clinton, Bush, Perot	Nominal	0 = not applicable; 1 = Clinton; 2 = Bush; 3 = Perot; 4 = other; 6 = no pres vote; 8 = don't know; 9 = no answer
postlife	Belief in life after death	Nominal	0 = not applicable; 1 = yes; 2 = no; 8 = don't know; 9 = no answer
happy	General happiness	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
hapmar	Happiness of marriage	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
jobinc	How important is a high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
classic1	Like or dislike classical music	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
opera	Like or dislike opera	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
country	Like or dislike country western music	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
rincom91	Respondent's income	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000-2999; 3 = \$3000-3999; 4 = \$4000-4999; 5 = \$5000-5999; 6 = \$6000-6999; 7 = \$7000-7999; 8 = \$8000-9999; 9 = \$10000-12499; 10 = \$12500-14999; 11 = \$15000-17499; 12 = \$17500-19999; 13 = \$20000-22499; 14 = \$22500-24999; 15 = \$25000-29999; 16 = \$30000-34999; 17 = \$35000-39999; 18 = \$40000-49999; 19 = \$50000-59999; 20 = \$60000-74999; 21 = \$75000+
tvhours	Hours per day watching TV	Scale	
finrela	Opinion of family income	Ordinal	1 = far below average; 2 = below average; 3 = average; 4 = above average; 5 = far above average; 8 = don't know; 9 = no answer
wifeduc		Scale	
husbeduc		Scale	
incomdol	Family income recoded to dollars	Ordinal	

profile-a.sav	Used in Chapters 4, 7, and 9		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
rincmdol	Respondent's income recoded to dollars	Ordinal	
id	Arbitrary ID numbers	Nominal	
husbhr	Hrs worked last week by husband	Scale	
wifehr	Hrs worked last week by wife	Scale	
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
wifeft	Wife employed full-time	Nominal	0 = no; 1 = yes
educdiff		Scale	
income4	Total family income in quartiles	Ordinal	
End of Data Set			

profile-b.sav	Used in Chapter 7		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	Age of respondent	Scale	
sex	Respondent's sex	Nominal	1= male; 2 = female
educ	Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
income91	Total family income	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6= \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–\$29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unemployed/laid off; 5 = retired ; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
richwork	If rich, continue or stop working	Nominal	0 = not applicable; 1 = continue working; 2 = stop working; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied ; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
life	Is life exciting or dull	Ordinal	0 = not applicable; 1 = dull; 2 = routine; 3 = exciting; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
hrs1	Number of hours worked last week	Scale	–1 = not applicable; 98 = don't know; 99 = no answer
degree	Respondent's highest degree	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate ; 7 = not applicable; 8 = don't know; 9 = no answer
anomia5	Lot of average man getting worse	Nominal	0 = not applicable; 1 = agree ; 2 = disagree; 8 = don't know; 9 = no answer
degree2	College degree	Nominal	0 = no college degree; 1 = college degree; 7 = not applicable; 8 = don't know; 9 = no answer
maeduc	Highest year of school completed, mother	Scale	97 = not applicable; 98 = don't know; 99 = no answer
paeduc	Highest year of school completed, father	Scale	97 = not applicable; 98 = don't know; 99 = no answer
macolleg	Mother a college grad	Nominal	0 = no; 1 = yes
pacolleg	Father a college grad	Nominal	0 = no; 1 = yes
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
marital	Marital status	Nominal	1 = married; 2 = widowed; 3 = divorced; 4 = separated; 5 = never married; 9 = no answer
agewed	Age when first married	Scale	0 = not applicable; 98 = don't know; 99 = no answer
spwrksta	Spouse labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part time; 3 = temp not working; 4 = unemployed, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
sphrs1	Number of hrs spouse worked last week	Scale	–1 = not applicable; 98 = don't know; 99 = no answer

profile-b.sav	Used in Chapter 7		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
sibs	Number of brothers and sisters	Scale	-1 = not applicable; 98 = don't know; 99 = no answer
zodiac	Respondent's astrological sign	Nominal	0 = not applicable; 1 = Aries; 2 = Taurus; 3 = Gemini; 4 = Cancer; 5 = Leo; 6 = Virgo; 7 = Libra; 8 = Scorpio; 9 = Sagittarius; 10 = Capricorn; 11 = Aquarius; 12 = Pisces; 98 = don't know; 99 = no answer
speduc	Highest year school completed, spouse	Scale	97 = not applicable; 98 = don't know; 99 = no answer
spdeg	Spouse's highest degree	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
partyid	Political party affiliation	Ordinal	0 = strong Democrat; 1 = not str Democrat; 2 = Ind, near Dem; 3 = Independent; 4 = Ind, near Rep; 5 = not str Republican; 6 = strong Republican; 7 = other party; 8 = don't know 9 = no answer
vote92	Did Respondent vote in 1992 election	Nominal	0 = not applicable; 1 = voted; 2 = did not vote; 3 = not eligible; 4 = refused; 8 = don't know; 9 = no answer
pres92	Vote for Clinton, Bush, Perot	Nominal	0 = not applicable; 1 = Clinton; 2 = Bush; 3 = Perot; 4 = other; 6 = no pres vote; 8 = don't know; 9 = no answer
postlife	Belief in life after death	Nominal	0 = not applicable; 1 = yes; 2 = no; 8 = don't know; 9 = no answer
happy	General happiness	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
hapmar	Happiness of marriage	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
jobinc	How important is a high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
classic1	Like or dislike classical music	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
opera	Like or dislike opera	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
country	Like or dislike country western music	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
rincom91	Respondent's income	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000-2999; 3 = \$3000-3999; 4 = \$4000-4999; 5 = \$5000-5999; 6 = \$6000-6999; 7 = \$7000-7999; 8 = \$8000-9999; 9 = \$10000-12499; 10 = \$12500-14999; 11 = \$15000-17499; 12 = \$17500-19999; 13 = \$20000-22499; 14 = \$22500-24999; 15 = \$25000-29999; 16 = \$30000-34999; 17 = \$35000-39999; 18 = \$40000-49999; 19 = \$50000-59999; 20 = \$60000-74999; 21 = \$75000+
tvhours	Hours per day watching TV	Scale	
finrela	Opinion of family income	Ordinal	1 = far below average; 2 = below average; 3 = average; 4 = above average; 5 = far above average; 8 = don't know; 9 = no answer
wifeduc		Scale	
husbeduc		Scale	
incomdol	Family income recoded to dollars	Ordinal	

profile-b.sav	Used in Chapter 7		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
rincmdol	Respondent's income recoded to dollars	Ordinal	
id	Arbitrary ID numbers	Nominal	
husbhr	Hrs worked last week by husband	Scale	
wifehr	Hrs worked last week by wife	Scale	
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
wifeft	Wife employed full-time	Nominal	0 = no; 1 = yes
educdiff		Scale	
income4	Total family income in quartiles	Ordinal	
MAH_1	Mahalanobis distance	Scale	
filter_\$	MAH_1 <= 22.458 (filter)	Scale	
MAH_2	Mahalanobis distance	Scale	
MAH_3	Mahalanobis distance	Scale	
End of Data Set			

profile-c.sav	Used in Chapter 10		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	age	Scale	
sex	sex	Nominal	1 = male; 2 = female
educ	educ	Scale	97 = not applicable; 98 = don't know; 99 = no answer
income91	income91	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+
wrkstat	wrkstat	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unemployed/laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
richwork	richwork	Nominal	0 = not applicable; 1 = continue working; 2 = stop working; 8 = don't know; 9 = no answer
satjob	satjob	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
life	life	Ordinal	0 = not applicable; 1 = dull; 2 = routine; 3 = exciting; 8 = don't know; 9 = no answer
impjob	impjob	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
hrs1	hrs1	Scale	-1 = not applicable; 98 = don't know; 99 = no answer
degree	degree	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
anomia5	anomia5	Nominal	0 = not applicable; 1 = agree; 2 = disagree; 8 = don't know; 9 = no answer
degree2	degree2	Nominal	0 = no college degree; 1 = college degree; 7 = not applicable; 8 = don't know; 9 = no answer
maeduc	maeduc	Scale	97 = not applicable; 98 = don't know; 99 = no answer
paeduc	paeduc	Scale	97 = not applicable; 98 = don't know; 99 = no answer
macolleg	macolleg	Nominal	0 = no; 1 = yes
pacolleg	pacolleg	Nominal	0 = no; 1 = yes
satjob2	satjob2	Nominal	1 = very satisfied; 2 = not very satisfied
marital	marital	Nominal	1 = married; 2 = widowed; 3 = divorced; 4 = separated; 5 = never married; 9 = no answer
agewed	agewed	Scale	0 = not applicable; 98 = don't know; 99 = no answer
spwrksta	spwrksta	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unemployed, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
sphrs1	sphrs1	Scale	-1 = not applicable; 98 = don't know; 99 = no answer
sibs	sibs	Scale	-1 = not applicable; 98 = don't know; 99 = no answer

profile-c.sav	Used in Chapter 10		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
zodiac	zodiac	Nominal	0 = not applicable; 1 = Aries; 2 = Taurus; 3 = Gemini; 4 = Cancer; 5 = Leo; 6 = Virgo; 7 = Libra; 8 = Scorpio; 9 = Sagittarius; 10 = Capricorn; 11 = Aquarius; 12 = Pisces; 98 = don't know; 99 = no answer
speduc	speduc	Scale	97 = not applicable; 98 = don't know; 99 = no answer
spdeg	spdeg	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
partyid	partyid	Ordinal	0 = strong Democrat; 1 = not strong Democrat; 2 = Ind, near Dem; 3 = Independent; 4 = Ind, near Rep; 5 = not strong Republican; 6 = strong Republican; 7 = other party; 8 = don't know; 9 = no answer
vote92	vote92	Nominal	0 = not applicable; 1 = voted; 2 = did not vote; 3 = not eligible; 4 = refused; 8 = don't know; 9 = no answer
pres92	pres92	Nominal	0 = not applicable; 1 = Clinton; 2 = Bush; 3 = Perot; 4 = other; 6 = no pres vote; 8 = don't know; 9 = no answer
postlife	postlife	Nominal	0 = not applicable; 1 = yes; 2 = no; 8 = don't know; 9 = no answer
happy	happy	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
hapmar	hapmar	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
jobinc	jobinc	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
classic1	classic1	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
opera	opera	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
country	country	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
rincom91	rincom91	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+
tvhours	tvhours	Scale	
finrela	finrela	Ordinal	1 = far below average; 2 = below average; 3 = average; 4 = above average; 5 = far above average; 8 = don't know; 9 = no answer
wifeduc	wifeduc	Scale	
husbeduc	husbeduc	Scale	
incomdol	incomdol	Ordinal	
rincmdol	rincmdol	Ordinal	
id	id	Nominal	

profile-c.sav	Used in Chapter 10		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
husbhr	husbhr	Scale	
wifehr	wifehr	Scale	
husbft	husbft	Nominal	0 = no; 1 = yes
wifeft	wifeft	Nominal	0 = no; 1 = yes
educdiff	educdiff	Scale	
income4	income4	Ordinal	
rincom2	rincom2 (*rincom91 transformed to eliminate coding problem	Scale	
MAH_1	Mahalanobis distance ([*]for age, hrs1, educ, rincom2, sibs, tvhours, sphrs1)	Scale	
filter_\$	MAH_1 <= 24.322 (filter)	Scale	
End of Data Set			

profile-d.sav	Used in Chapter 11		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	age	Scale	
sex	sex	Nominal	1 = male; 2 = female
educ	educ	Scale	97 = not applicable; 98 = don't know; 99 = no answer
income91	income91	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+
wrkstat	wrkstat	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unemployed/laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = NA
richwork	richwork	Nominal	0 = not applicable; 1 = continue working; 2 = stop working; 8 = don't know; 9 = no answer
satjob	satjob	Ordinal	0 = not applicable; 1 = very satisfied ; 2 = mod satisfied; 3 = a little dissatisfied ; 4 = very dissatisfied; 8 = don't know; 9 = no answer
life	life	Ordinal	0 = not applicable; 1 = dull; 2 = routine; 3 = exciting; 8 = don't know; 9 = no answer
impjob	impjob	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
hrs1	hrs1	Scale	-1= not applicable; 98 = don't know; 99 = no answer
degree	degree	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate ; 7 = not applicable; 8 = don't know; 9 = no answer
anomia5	anomia5	Nominal	0 = not applicable; 1 = agree; 2 = disagree; 8 = don't know; 9 = no answer
degree2	degree2	Nominal	0 = no college degree; 1 = college degree; 7 = not applicable; 8 = don't know; 9 = no answer
maeduc	maeduc	Scale	97 = not applicable; 98 = don't know; 99 = no answer
paeduc	paeduc	Scale	97 = not applicable; 98 = don't know; 99 = no answer
macolleg	macolleg	Nominal	0 = no; 1 = yes
pacolleg	pacolleg	Nominal	0 = no; 1 = yes
satjob2	satjob2	Nominal	1 = very satisfied; 2 = not very satisfied
marital	marital	Nominal	1 = married; 2 = widowed; 3 = divorced; 4 = separated; 5 = never married; 9 = no answer
agewed	agewed	Scale	0 = not applicable; 98 = don't know; 99 = no answer
spwrksta	spwrksta	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unemployed, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
sphrs1	sphrs1	Scale	-1 = not applicable; 98 = don't know; 99 = no answer
sibs	sibs	Scale	-1 = not applicable; 98 = don't know; 99 = no answer

profile-d.sav	Used in Chapter 11		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
zodiac	zodiac	Nominal	0 = not applicable; 1 = Aries; 2 = Taurus; 3 = Gemini; 4 = Cancer; 5 = Leo; 6 = Virgo; 7 = Libra; 8 = Scorpio; 9 = Sagittarius; 10 = Capricorn; 11 = Aquarius; 12 = Pisces; 98 = don't know; 99 = no answer
speduc	speduc	Scale	97 = not applicable; 98 = don't know; 99 = no answer
spdeg	spdeg	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
partyid	partyid	Ordinal	0 = strong Democrat; 1 = not str Democrat; 2 = Ind, near Dem; 3 = Independent; 4 = Ind, near Rep; 5 = not str Republican; 6 = strong Republican; 7 = other party; 8 = don't know; 9 = no answer
vote92	vote92	Nominal	0 = not applicable; 1 = voted; 2 = did not vote; 3 = not eligible; 4 = refused; 8 = don't know; 9 = no answer
pres92	pres92	Nominal	0 = not applicable; 1 = Clinton; 2 = Bush; 3 = Perot; 4 = other; 6 = no pres vote; 8 = don't know; 9 = no answer
postlife	postlife	Nominal	0 = not applicable; 1 = yes; 2 = no; 8 = don't know; 9 = no answer
happy	happy	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
hapmar	hapmar	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
jobinc	jobinc	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
classic1	classic1	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
opera	opera	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
country	country	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
rincom91	rincom91	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+
tvhours	tvhours	Scale	
finrela	finrela	Ordinal	1 = far below average; 2 = below average; 3 = average; 4 = above average; 5 = far above average; 8 = don't know; 9 = no answer
wifeduc	wifeduc	Scale	
husbeduc	husbeduc	Scale	
incomdol	incomdol	Ordinal	
rincmdol	rincmdol	Ordinal	
id	id	Nominal	

profile-d.sav	Used in Chapter 11		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
husbhr	husbhr	Scale	
wifehr	wifehr	Scale	
husbft	husbft	Nominal	0 = no; 1 = yes
wifeft	wifeft	Nominal	0 = no; 1 = yes
educdiff	educdiff	Scale	
income4	income4	Ordinal	
life2		Scale	
MAH_1	Mahalanobis distance ([*]for age, hrs1, educ, rincom91, sibs, tvhours)	Scale	
filter_	\$ MAH_1 <= 22.458 (filter)	Scale	
End of Data Set			

profile-e.sav	Used in Chapter 11		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
age	Age of respondent	Scale	
sex	Respondent's sex	Nominal	1 = male; 2 = female
educ	Highest year of school completed	Scale	97 = not applicable; 98 = don't know; 99 = no answer
income91	Total family income	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000–2999; 3 = \$3000–3999; 4 = \$4000–4999; 5 = \$5000–5999; 6 = \$6000–6999; 7 = \$7000–7999; 8 = \$8000–9999; 9 = \$10000–12499; 10 = \$12500–14999; 11 = \$15000–17499; 12 = \$17500–19999; 13 = \$20000–22499; 14 = \$22500–24999; 15 = \$25000–29999; 16 = \$30000–34999; 17 = \$35000–39999; 18 = \$40000–49999; 19 = \$50000–59999; 20 = \$60000–74999; 21 = \$75000+
wrkstat	Labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unemployed/laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = NA
richwork	If rich, continue or stop working	Nominal	0 = not applicable; 1 = continue working; 2 = stop working; 8 = don't know; 9 = no answer
satjob	Job satisfaction	Ordinal	0 = not applicable; 1 = very satisfied; 2 = mod satisfied; 3 = a little dissatisfied; 4 = very dissatisfied; 8 = don't know; 9 = no answer
life	Is life exciting or dull	Ordinal	0 = not applicable; 1 = dull; 2 = routine; 3 = exciting; 8 = don't know; 9 = no answer
impjob	Importance to Respondent of having a fulfilling job	Ordinal	0 = not applicable; 1 = one of the most important; 2 = very important; 3 = somewhat important; 4 = not too important; 5 = not at all important; 8 = don't know; 9 = no answer
hrs1	Number of hours worked last week	Scale	–1 = not applicable; 98 = don't know; 99 = no answer
degree	Respondent's highest degree	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
anomia5	Lot of average man getting worse	Nominal	0 = not applicable; 1 = agree; 2 = disagree; 8 = don't know; 9 = no answer
degree2	College degree	Nominal	0 = no college degree; 1 = college degree; 7 = not applicable; 8 = don't know; 9 = no answer
maeduc	Highest year of school completed, mother	Scale	97 = not applicable; 98 = don't know; 99 = no answer
paeduc	Highest year of school completed, father	Scale	97 = not applicable; 98 = don't know; 99 = no answer
macolleg	Mother a college grad	Nominal	0 = no; 1 = yes
pacolleg	Father a college grad	Nominal	0 = no; 1 = yes
satjob2	Job satisfaction	Nominal	1 = very satisfied; 2 = not very satisfied
marital	Marital status	Nominal	1 = married; 2 = widowed; 3 = divorced; 4 = separated; 5 = never married; 9 = NA
agewed	Age when first married	Scale	0 = not applicable; 98 = don't know; 99 = no answer
spwrksta	Spouse labor force status	Nominal	0 = not applicable; 1 = working full-time; 2 = working part-time; 3 = temp not working; 4 = unemployed, laid off; 5 = retired; 6 = school; 7 = keeping house; 8 = other; 9 = no answer
sphrs1	Number of hrs spouse worked last week	Scale	–1 = not applicable; 98 = don't know; 99 = no answer

profile-e.sav	Used in Chapter 11		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
sibs	Number of brothers and sisters	Scale	-1 = not applicable; 98 = don't know; 99 = no answer
zodiac	Respondent's astrological sign	Nominal	0 = not applicable; 1 = Aries; 2 = Taurus; 3 = Gemini; 4 = Cancer; 5 = Leo; 6 = Virgo; 7 = Libra; 8 = Scorpio; 9 = Sagittarius; 10 = Capricorn; 11 = Aquarius; 12 = Pisces; 98 = don't know; 99 = no answer
speduc	Highest year school completed, spouse	Scale	97 = not applicable; 98 = don't know; 99 = no answer
spdeg	Spouses highest degree	Ordinal	0 = less than high school; 1 = high school; 2 = junior college; 3 = bachelor's; 4 = graduate; 7 = not applicable; 8 = don't know; 9 = no answer
partyid	Political party affiliation	Ordinal	0 = strong Democrat; 1 = not strong Democrat; 2 = Ind, near Dem; 3 = Independent; 4 = Ind, near Rep; 5 = not strong Republican; 6 = strong Republican; 7 = other party; 8 = don't know; 9 = no answer
vote92	Did Respondent vote in 1992 election	Nominal	0 = not applicable; 1 = voted; 2 = did not vote; 3 = not eligible; 4 = refused; 8 = don't know; 9 = no answer
pres92	Vote for Clinton, Bush, Perot	Nominal	0 = not applicable; 1 = Clinton; 2 = Bush; 3 = Perot; 4 = other; 6 = no pres vote; 8 = don't know; 9 = no answer
postlife	Belief in life after death	Nominal	0 = not applicable; 1 = yes; 2 = no; 8 = don't know; 9 = no answer
happy	General happiness	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
hapmar	Happiness of marriage	Ordinal	0 = not applicable; 1 = very happy; 2 = pretty happy; 3 = not too happy; 8 = don't know; 9 = no answer
jobinc	How important is a high income	Ordinal	0 = not applicable; 1 = most impt; 2 = second; 3 = third; 4 = fourth; 5 = fifth; 8 = don't know; 9 = no answer
classic1	Like or dislike classical music	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
opera	Like or dislike opera	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
country	Like or dislike country western music	Ordinal	0 = not applicable; 1 = like very much; 2 = like it; 3 = mixed feelings; 4 = dislike it; 5 = dislike it very much; 8 = don't know much about it; 9 = no answer
rincom91	Respondent's income	Ordinal	0 = not applicable; 1 = LT \$1000; 2 = \$1000-2999; 3 = \$3000-3999; 4 = \$4000-4999; 5 = \$5000-5999; 6 = \$6000-6999; 7 = \$7000-7999; 8 = \$8000-9999; 9 = \$10000-12499; 10 = \$12500-14999; 11 = \$15000-17499; 12 = \$17500-19999; 13 = \$20000-22499; 14 = \$22500-24999; 15 = \$25000-29999; 16 = \$30000-34999; 17 = \$35000-39999; 18 = \$40000-49999; 19 = \$50000-59999; 20 = \$60000-74999; 21 = \$75000+
tvhours	Hours per day watching TV	Scale	
finrela	Opinion of family income	Ordinal	1 = far below average; 2 = below average; 3 = average; 4 = above average; 5 = far above average; 8 = don't know; 9 = no answer
wifeduc		Scale	
husbeduc		Scale	
incomdol	Family income recoded to dollars	Ordinal	

profile-e.sav	Used in Chapter 11		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
rincmdol	Respondent's income recoded to dollars	Ordinal	
id	Arbitrary ID numbers	Nominal	
husbhr	Hrs worked last week by husband	Scale	
wifehr	Hrs worked last week by wife	Scale	
husbft	Husband employed full-time	Nominal	0 = no; 1 = yes
wifeft	Wife employed full-time	Nominal	0 = no; 1 = yes
educdiff		Scale	
income4	Total family income in quartiles	Ordinal	
MAH_1	Mahalanobis distance ([*]for age, hrs1, educ, rincom91, sibs, tvhours)	Scale	
filter_\$	Mah_1 <= 20.515 (filter)	Scale	
life2		Scale	
End of Data Set			

salary-a.sav	Used in Chapter 4		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
id	Employee code	Nominal	
salbeg	Beginning salary	Scale	
sex	Sex of employee	Nominal	0 = male; 1 = female
time	Job seniority	Scale	
age	Age of employee	Scale	
salnow	Current salary	Scale	
edlevel	Educational level	Scale	
work	Work experience	Scale	
jobcat	Employment category	Nominal	1 = clerical; 2 = office trainee; 3 = security officer; 4 = college trainee; 5 = exempt employee; 6 = MBA trainee; 7 = technical
minority	Minority classification	Nominal	0 = white; 1 = nonwhite
sexrace	Sex & race classification	Nominal	1 = white males; 2 = minority males; 3 = white females; 4 = minority females
End of Data Set			

salary-b.sav	Used in Chapter 4		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
id	Employee code	Nominal	
salbeg	Beginning salary	Scale	
sex	Sex of employee	Nominal	0 = male; 1 = female
time	Job seniority	Scale	
age	Age of employee	Scale	
salnow	Current salary	Scale	
edlevel	Educational level	Scale	
work	Work experience	Scale	
jobcat	Employment category	Nominal	1 = clerical; 2 = office trainee; 3 = security officer; 4 = college trainee; 5 = exempt employee; 6 = MBA trainee; 7 = technical
minority	Minority classification	Nominal	0 = white; 1 = nonwhite
sexrace	Sex & race classification	Nominal	1 = white males; 2 = minority males; 3 = white females; 4 = minority females
salnow2		Scale	
salnow3		Scale	
End of Data Set			

schools-a.sav	Used in Chapters 3 and 10		
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
school	School name	Nominal	
loinc93	Percentage low income	Scale	
lep93	Percentage limited English proficiency	Scale	
lep94	% limited English proficiency 1994	Scale	
grad93	% graduating 1993	Scale	
grad94	% graduating 1994	Scale	
act94	Average ACT score 1994	Scale	
act93	Average ACT score 1993	Scale	
pctact93	Percentage taking ACT 1993	Scale	
pctact94	Percentage taking ACT 1994	Scale	
math93	10th grade average math score	Scale	
math94me	% meet or exceed state standards	Scale	
mathch94	Change in % meet/exceed 94–93	Scale	
read93	10th Grade Average Reading Score	Scale	
read94me	% meet or exceed state standards	Scale	
readch94	Change in % meet/exceed 94–93	Scale	
scienc93	11th grade average science score	Scale	
sci94me	% meet or exceed state standard	Scale	
scich94	Change in % meet/exc 94–93	Scale	
id		Nominal	
medloinc	Above or below median loinc	Nominal	
End of Data Set			

Used in Chapter 9			
VARIABLE NAME	VARIABLE LABEL (OR DESCRIPTION*)	MEASURE	VALUES
school	School name	Nominal	
loinc93	Percent low income	Scale	
lep93		Scale	
lep94		Scale	
grad93		Scale	
grad94		Scale	
act94		Scale	
act93		Scale	
pctact93	Percent taking ACT 1993	Scale	
pctact94	Percent taking ACT 1994	Scale	
math93		Scale	
math94me		Scale	
mathch94		Scale	
read93		Scale	
read94me		Scale	
readch94		Scale	
scienc93		Scale	
sci94me		Scale	
scich94	Change in % meet/exc 94-93	Scale	
id		Nominal	
medloinc	Above or below median loinc	Nominal	
MAH_1	Mahalanobis distance ([*]for grad93, grad94, act93, act94, math93, read93, scienc93, math94me, read94me, sci94me, lep93, lep94)	Scale	
filter_\$	MAH_1 <= 32.909 (filter)	Scale	
FAC1_1	REGR factor score 1 for analysis 1	Scale	
FAC2_1	REGR factor score 2 for analysis 1	Scale	
FAC1_2	REGR factor score 1 for analysis 2	Scale	
FAC2_2	REGR factor score 2 for analysis 2	Scale	
End of Data Set			

Appendix B: The Chi-Square Distribution

df	Proportion in critical region					
	0.10	0.05	0.025	0.01	0.005	0.001
1	2.71	3.84	5.02	6.63	7.88	10.828
2	4.61	5.99	7.38	9.21	10.60	13.816
3	6.25	7.81	9.35	11.34	12.84	16.266
4	7.78	9.49	11.14	13.28	14.86	18.467
5	9.24	11.07	12.83	15.09	16.75	20.515
6	10.64	12.59	14.45	16.81	18.55	22.458
7	12.02	14.07	16.01	18.48	20.28	24.322
8	13.36	15.51	17.53	20.09	21.96	26.125
9	14.68	16.92	19.02	21.67	23.59	27.877
10	15.99	18.31	20.48	23.21	25.19	29.588
11	17.28	19.68	21.92	24.72	26.76	31.264
12	18.55	21.03	23.34	26.22	28.30	32.909
13	19.81	22.36	24.74	27.69	29.82	34.528
14	21.06	23.68	26.12	29.14	31.32	36.123
15	22.31	25.00	27.49	30.58	32.80	37.697
16	23.54	26.30	28.85	32.00	34.27	39.252
17	24.77	27.59	30.19	33.41	35.72	40.790
18	25.99	28.87	31.53	34.81	37.16	42.312
19	27.20	30.14	32.85	36.19	38.58	43.820
20	28.41	31.41	34.17	37.57	40.00	45.315
21	29.62	32.67	35.48	38.93	41.40	46.797
22	30.81	33.92	36.78	40.29	42.80	48.268
23	32.01	35.17	38.08	41.64	44.18	49.728
24	33.20	36.42	39.36	42.98	45.56	51.179
25	34.38	37.65	40.65	44.31	46.93	52.620
26	35.56	38.89	41.92	45.64	48.29	54.052
27	36.74	40.11	43.19	46.96	49.64	55.476
28	37.92	41.34	44.46	48.28	50.99	56.892
29	39.09	42.56	45.72	49.59	52.34	58.302
30	40.26	43.77	46.98	50.89	53.67	59.703
40	51.81	55.76	59.34	63.69	66.77	73.402
50	63.17	67.50	71.42	76.15	79.49	86.661
60	74.40	79.08	83.30	88.38	91.95	99.607
70	85.53	90.53	95.02	100.42	104.22	112.317
80	96.58	101.88	106.63	112.33	116.32	124.839
90	107.56	113.14	118.14	124.12	128.30	137.208
100	118.50	124.34	129.56	135.81	140.17	149.449

Glossary

adjusted squared multiple correlation (R^2_{adj}) — the unbiased estimate of R^2 ; usual estimate of R^2 is positively biased

alpha (α) level — in hypothesis testing, the pre-established probability of being incorrect; also known as the *level of significance*

alternative hypothesis (H_1) — prediction that one method or group is expected to be better than the other; in other words, that the two group means are not equal and therefore represent a true difference in the population

backward deletion (in multiple regression) — form of statistical regression where order of entry of variables into solution is based entirely on statistical criteria; equation starts out with all IVs in the solution, and then variables are deleted one at a time (if they do not contribute significantly to the regression solution)

Bartlett's sphericity test — procedure that tests the null hypothesis that the variables in the population correlation matrix are uncorrelated; used for factor analysis with small samples

beta coefficients — see *regression coefficients*

beta weights — see *regression coefficients*

between-groups variability — the term for the numerator in the calculation of an F ratio

Box's M test for equality of variance-covariance matrices — the statistical test of homoscedasticity in multivariate situations

causal modeling — statistical technique that, using regression analysis, examines patterns of intercorrelations among variables in order to determine if they fit the researcher's underlying theory of which variables are causing which other variables

classification — statistical analysis where prediction of group membership is the primary purpose

coefficient of determination — see *squared multiple correlation*

communality (h_i) — amount of variance in each variable accounted for by the factors; equal to the squared multiple correlation of the variable as predicted from the factors; also equal to the sum of squared loadings for a variable across all factors; provided for each variable

concomitant variable — an accompanying variable not central to an analysis; also referred to as an *extraneous variable*

confirmatory factor analysis — more advanced than exploratory factor analysis; used to test a theory about latent (i.e., underlying, unobservable) processes

continuous variable — measured on a scale that changes smoothly over possible values rather than in steps; also referred to as *interval* or *quantitative*

control variable — a variable whose effect on a DV is removed; also referred to as a variable whose effect has been “partialed out”

correlation matrix (observed) — a square, symmetrical matrix; each row (and each column) represents a different variable; located at each intersection of a row and column is the bivariate correlation between the two variables

covariate — IV used as the basis for the adjustment of DV scores (as a statistical control mechanism) prior to examining main effects and interactions between IVs with respect to some DV

cross-validation — a procedure in which the predictive power of a regression equation is typically assessed by using the equation derived from one sample to predict Y values in a second sample (called the *cross-validation sample*)

data matrix — organization of raw scores, or data, where rows represent cases (participants) and columns represent variables

data transformations — the application of mathematical procedures to data in order to make them appear more normal

deflated correlations — occur when variables have a restricted range in the sampling of cases or when there exist very uneven splits in the cases in categories of discrete or dichotomous variables

determinants — a sort of generalized variance for an entire set of variables; used in the calculation of sum-of-squares and cross-products (SSCP) matrices

dichotomous variable — a discrete variable with only two possible values

discrepancy — measure of the impact of cases on a solution; measures the extent to which a case is in line with the others

discriminant function (or variate) — maximization of the linear combination of IVs used to discriminate among groups

disordinal interaction — interaction between variables when the lines plotted on the graph cross within the values of the graph

disturbance term — the residual term in causal modeling

effect size — the size of the treatment effect the researcher wishes to detect with respect to a given level of power; denoted as *ES* or partial η^2

eigenvalue (in factor analysis) — amount of total variance explained by each factor

endogenous variable — the variable being explained by a causal model; that is, having its variance explained by other variables included in the model; also referred to as the *dependent variable*

error variability — the term for the denominator in the calculation of an *F* ratio; also known as *within-groups variability*

eta squared (η^2) — a measure of strength of association (also known as *effect size*); statistical significance measures whether or not there is an association between IVs and DVs, but eta squared measures *how much* association there is; shows the proportion of variance in the DV that is attributable to the effect (IV)

exogenous variable — a variable not being explained by a causal model, whose variance is accounted for by other variables outside the model; also referred to as an *independent variable*

experimentwise alpha level — an accumulated error rate, across multiple group comparisons

exploratory factor analysis — goal is to describe and summarize data by grouping together variables that are correlated; variables may or may not have been chosen with these underlying structures in mind

extraction — the process by which the underlying factors from a larger set of variables are determined

F ratio — test statistic for ANOVA

factor — independent variable, in ANOVA terminology

factor analysis — a mathematical model is created, resulting in the estimation of *factors*; contrast with *principal components analysis*

factor extraction — the specific mathematical procedure by which factors are determined

factor loading — the Pearson correlation of an original variable with a factor

factor matrix — provides correlation coefficients between each IV and each factor in the solution; values can also be interpreted as that amount that each IV contributes to each factor

factor scores — estimates of the scores participants would have received on each of the factors had they been measured directly

factorial analysis — statistical analysis that involves more than one IV

first discriminant function — the linear combination of variables that maximizes the between-to-within association between variables in a discriminant analysis procedure

first principal component — the initial linear combination of IVs; accounts for the largest amount of total variance; equal to largest eigenvalue for the solution

fixed effects — in ANOVA, levels of each IV are selected based on interest in testing for significance of the particular IV

forward selection (in multiple regression) — form of statistical regression where order of entry of variables into solution is based entirely on statistical criteria; equation starts out empty and variables are entered one at a time (if they are statistically significant); once entered into the equation, the variable stays

group similarity (in classification) — represented by the classical probability of the data given group membership; that is, $P(X_i/G_k)$

heteroscedasticity — the violation of the assumption of homoscedasticity

hierarchical multiple regression — see *sequential analysis*

hit rate — the number of correct classifications in a discriminant function analysis

homogeneity of regression — in ANCOVA, an assumption that states that the regression slopes for a covariate are homogeneous

homoscedasticity — assumption that the variability in scores for one continuous variable is roughly the same at all values of another continuous variable

honest correlations — accurate bivariate correlations between pairs of variables

hypothesis testing — statistical tests of predictions made about a sample

indirect effect — an effect that occurs when a variable affects an endogenous variable through its effect on some other variable

inflated correlations — correlations between composite variables that become inflated when variables are reused in making up the composite variables

influence — product of leverage and discrepancy; assesses change in regression coefficients when a case is deleted; cases with influence scores greater than 1.00 are possible outliers with great influence on the solution

interaction between factors — occurs when the effect of one factor depends on different levels of the other factor

inverse criterion — the smaller the value, the more evidence for treatment effects or group differences

Kolmogorov-Smirnov statistic — tests the null hypothesis that the population is normally distributed

kurtosis — degree of peakedness of a distribution; equal to zero when a distribution is normal

least-squares solution — the regression model with coefficient values that minimize the sum of squared residuals

leptokurtosis — condition when values for kurtosis are positive, indicating that the distribution is too peaked with long, thin tails

level of significance — in hypothesis testing, the pre-established probability of being incorrect; also known as the *alpha level*

Levene's test — a statistical test of the homogeneity of variances

leverage — measure of the impact of cases on a solution; closely related to Mahalanobis distance; cases with high leverage are far away from other cases but essentially on the same line

likelihood ratio test — test of overall logistic relationship; analogous to F test of ΔR^2 in multiple regression

linearity — assumption that there is a straight-line relationship between two variables

logit — the natural logarithm of the odds

Mahalanobis distance — statistical measure of an outlier; distance of a case from the centroid of the remaining cases where the centroid is the point created by the means of all the variables

main effects — in ANOVA, any differences produced by either factor, independent of the other

matrix algebra — an extension of scalar algebra where mathematical operations are performed on an ordered array of numerical values

mean — the arithmetic average of a set of scores

median — the score in the distribution that divides the upper 50% of scores from the lower 50%

mode — the most frequently occurring score in a distribution

multicollinearity — problem created when IVs are very highly correlated ($r \geq .90$) with each other

multiple correlation (R) — correlation between the observed and predicted values of the DV

multivariate analysis — statistical analysis that involves more than one dependent variable

multivariate outlier — cases with an unusual pattern of scores; values on individual variables may look reasonable, but the combinations of two variables produce values that look unusual or discrepant

normality — assumption that all variables and linear combinations of variables are normally distributed

null hypothesis (H_0) — predicts that the only differences existing between groups are chance differences that represent only random sampling error

oblique rotation — rotation of factors resulting in factors being correlated with each other and producing several matrices: a *factor correlation* matrix (i.e., a matrix of correlations between all factors); a *loading* matrix separated into a *structure* matrix (i.e., correlations between factors and variables); and a *pattern* matrix (i.e., unique relationships with no overlapping among factors between each factor and each observed variable) upon which interpretation of factors is obtained

odds (in logistic regression) — ratio of the probability of outcome $Y = 1$ (e.g., program completion) to the probability of outcome $Y = 2$ (e.g., program noncompletion)

odds ratio (in logistic regression) — the ratio of odds of $Y = 1$, for instance, for two different values of the IV; symbolized by Ψ

one-way ANOVA — analysis of variance design that studies the effect that one factor has on one dependent variable

ordinal interaction — interaction between variables when the lines plotted on the graph do not cross within the values of the graph

orthogonal rotation — rotation of factors resulting in factors being uncorrelated with each other; result is a *loading* matrix (i.e., a matrix of correlations between all observed variables and factors) where the size of the loading reflects the extent of the relationship between each observed variable and each factor; interpretation of factors is obtained from the loading matrix

orthogonality — perfect nonassociation between variables

outlier — case with extreme values on one variable or on a combination of variables so that it distorts resulting statistics or unduly influence solutions or models; would result in an excessively large residual

pairwise comparisons — following ANOVA, tests that enable the researcher to compare individual treatments two at a time

partial correlation — a measure of the relationship between an IV and the DV, holding constant (or partialing out the effect of) all other IVs

partial out — another term for controlling for the effect of a particular variable

partial regression coefficient — see *unstandardized regression coefficient*

path analysis — a method of analyzing correlations among a set of variables in order to examine the pattern of causal relationships; usually depicted with a path diagram including arrows showing the direction of causation among variables

path coefficient — the standardized regression coefficients associated with causal paths in a causal model; see also *structural coefficient*

path decomposition — process that results in the reproduced correlations in a path analysis; the reproduced correlations are equal to the product of all coefficients in a given path; see also *path tracing*

path diagram — a pictorial representation of a causal model, showing the direction of causation and associated path coefficients

path tracing — process that results in the reproduced correlations in a path analysis; the reproduced correlations are equal to the product of all coefficients in a given path; see also *path decomposition*

Pearson r — the appropriate measure of correlation when variables are expressed as scores

percentile rank — a value that indicates the percentage of scores that fall at or below a given score

platykurtosis — condition when values for kurtosis are negative, indicating that the distribution is too flat, with many cases in the tails

power — the probability of rejecting H_0 when H_0 is in fact false; equal to $1 - \beta$

principal components analysis — most common for extracting factors in factor analysis; original variables are transformed into a new set of linear combinations by extracting the maximum variance from the data set with each component; results in *components*; contrast with *factor analysis*

prior probability (in classification) — represented by the probability of membership in group k prior to collection of the data (i.e., P_k)

probability level — in significance testing, the probability of being incorrect in drawing a conclusion about the population

quartile deviation — one-half of the difference between the 3rd quartile (i.e., the 75th percentile) and the 1st quartile (i.e., the 25th percentile)

random effects — in ANOVA, it may be desirable to generalize to a population only certain *levels* of the IV (as opposed to the entire IV), so levels of the IV are randomly selected from the population

range — the difference between the highest score and the lowest score in the distribution

regression coefficients — the values for the constants in a regression equation; also function as the weights attached to each IV; also known as *beta coefficients* or *beta weights*

regression line — the prediction line, as defined by the best-fitting line, through a series of points in a scatterplot

reproduced correlation matrix — the correlation matrix produced from the factor solution or from the theoretical model in path analysis

research hypothesis — see *alternative hypothesis*

residual correlation matrix — the difference between the observed and reproduced correlation matrices

residuals — portions of scores not accounted for by the analysis; also a measure of the difference between the obtained and predicted values on the DV, therefore referred to as *prediction error*

robustness — the degree to which a statistical test is still appropriate to apply when some of its assumptions are not met

rotation (of factors) — process by which the solution of a factor analysis is made more interpretable without altering the underlying mathematical structure

sampling error — the expected, chance variation among sample means

scree plot — a graph of the magnitude of each eigenvalue (vertical axis) plotted against their ordinal numbers (horizontal axis)

sequential analysis — researcher assigns priority for the entry of variables into the equation (solution); first variable to be entered is assigned both its unique variance and any overlapping variance it has with other IVs; upon entry, lower-priority variables are then assigned their unique variance and any remaining overlapping variance

significance tests — procedures used to determine if the difference between sample means is substantial enough to justify concluding that a *true* difference exists in the population as well

singularity — problem created when variables are redundant (i.e., one of the variables is a combination of two or more of the other variables)

skewness — degree of symmetry of a distribution; equal to zero when distribution is normal

Spearman rho — the appropriate measure of correlation when variables are expressed as ranks

spurious effect — causal effect between two variables that is largely due to a common third variable

squared multiple correlation (R^2) — the proportion of variability in the DV explained by the model (i.e., the combination of IVs); also referred to as the *coefficient of determination*

standard analysis — all variables are entered into the solution simultaneously; overlapping variance (i.e., variance in the DV explained by more than one IV) is ignored when assessing the contributions of individual IVs to the variability in the DV

standard deviation — the average distance of scores away from the mean

standard error of the mean — the amount by which one can expect sample means to differ if other samples from the same population are used; indicates how well a sample represents the population from which it was selected

standard score — a transformed score derived from the manipulation of a raw score that expresses how far away from the mean a given score is located, usually reported in standard deviation units

standardized discriminant function coefficient — represents the weight applied to each IV for a given discriminant function; in *z*-score form

standardized regression coefficient (β) — regression coefficient expressed in *z*-score form; interpreted as the amount of change in the DV associated with a one standard deviation unit change in that IV, with all other IVs held constant

statistical regression — order of entry of variables into solution is based entirely on statistical criteria; the meaning or interpretation of the inclusion of specific variables is not relevant

stepwise selection (in multiple regression) — form of statistical regression where order of entry of variables into solution is based entirely on statistical criteria; combination of forward and backward regression; equation starts out empty and variables are entered one at a time (if they are statistically significant), but they may also be removed (if they no longer contribute significantly to the solution); results in a better solution when compared to forward and backward procedures

structural coefficient — the standardized regression coefficients associated with causal paths in a causal model; see also *path coefficient*

structural equation — an assumed causal model, when stated as an equation; typically stated in its standardized form

structural equation modeling — sophisticated version of path analysis incorporating unobservable, unmeasurable (latent) variables into the path model

sum-of-squares and cross-products matrix — precursor to the variance-covariance matrix where the deviations have not yet been averaged

systematic bias — predictable variability that results from intact groups that differ systematically on several variables; is best addressed through means of random assignment of participants to groups, but can be addressed by using a covariance analysis

tolerance — a measure of collinearity among IVs, ranging from 0 (indicating multicollinearity) to 1 (indicating independence among IVs)

two-way analysis of variance — an ANOVA design, consisting of two IVs and one DV

Type I error — error made when it is concluded that a null hypothesis is false even though it is actually true; probability of this type of error is defined as α

Type II error — error made when a null hypothesis is actually false but it is concluded to be true; probability of this type of error is defined as β

univariate analysis — statistical analysis that involves only one dependent variable

univariate outlier — cases with very large standardized scores on a single variable

unstandardized regression coefficient (B) — the raw score weight associated with a specific IV in a regression equation; interpreted as the amount of change in the DV associated with a one-unit change in that IV, with all other IVs held constant

variance-covariance matrix — used when scores are measured along a continuous scale; a square, symmetrical matrix where the elements on the main diagonal are the variances for each variable and the elements on the off-diagonals are the covariances between pairs of variables

variance inflation factor (VIF) — a measure of the extent to which there exist multicollinear relationships for given predictor IVs

Wald statistic — the statistical test of each predictor in logistic regression analysis

within-groups variability — the term for the denominator in the calculation of an F ratio; also known as *error variability*

z-score — a standard score that indicates the distance away from the mean a score is in terms of standard deviation units; calculated by subtracting the mean from the raw score and then dividing the value by the standard deviation

References

- Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Aron, A., Aron, E. N., & Coups, E. (2006). *Statistics for psychology* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Aron, A., Aron, E. N., & Coups, E. (2008). *Statistics for the behavioral and social sciences: A brief course* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Asher, H. B. (1983). *Causal modeling*. Sage University Paper series on Quantitative Application in the Social Sciences, series no. 07-003. Beverly Hills and London: Sage.
- Brewer, J. K. (1978). *Everything you always wanted to know about statistics, but didn't know how to ask*. Dubuque, IA: Kendall/Hunt.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- George, D., & Mallory, P. (2000). *SPSS for Windows step-by-step: A simple guide and reference* (2nd ed.). Boston, MA: Allyn & Bacon.
- Gravetter, F. J., & Wallnau, L. B. (2008). *Essentials of statistics for the behavioral sciences* (6th ed.). Belmont, CA: Wadsworth.
- Harris, M. B. (1998). *Basic statistics for behavioral science research* (2nd ed.). Boston, MA: Allyn & Bacon.
- Huitema, B. (1980). *The analysis of covariance and alternatives*. New York, NY: Wiley.
- Johnson, R. A., & Wichern, D. W. (2008). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurements*, 20, 141–151.
- Kennedy, J. J., & Bush, A. J. (1985). *An introduction to the design and analysis of experiments in behavioral research*. Lanham, MD: University Press of America.
- Levin, J., & Fox, J. A. (2006). *Elementary statistics in social research* (10th ed.). Boston, MA: Allyn & Bacon.

- Long, J. S. (1983). *Covariance structure models: An introduction to LISREL*. Sage University Paper series on Quantitative Application in the Social Sciences, series no. 07-034. Beverly Hills and London: Sage.
- Newman, I., Newman, C., Brown, R., & McNeely, S. (2005). *Conceptual statistics for beginners* (3rd ed.). Lanham, MD: University Press of America.
- Norusis, M. J. (1998). *SPSS 8.0: Guide to data analysis*. Upper Saddle River, NJ: Prentice Hall.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). Fort Worth, TX: Holt, Rinehart, and Winston.
- Sprinthall, R. C. (2007). *Basic statistical analysis* (8th ed.). Boston, MA: Allyn & Bacon.
- Stevens, J. (2001). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.
- Tate, R. (1992). *General linear model applications*. Unpublished manuscript, Florida State University.
- Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York, NY: Macmillan.
- Thompson, B. (1998, April). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Vogt, W. P. (2005). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (3rd ed.). Thousand Oaks, CA: Sage.
- Williams, F. (1992). *Reasoning with statistics: How to read quantitative research* (4th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.

SUBJECT INDEX

A

- alternative hypothesis • 10–11, 72–76
 - analysis of covariance • 2, 99–107, 111, 121–122, 125, 145, 147–149, 153, 164
 - one-way • 15
 - testing assumptions • 104, 108
 - analysis of variance
 - one-way • 13, 15, 21, 71, 72, 74–75, 79–80, 99
 - two-way • 71, 74–75, 77–78, 80, 84, 92–93, 101, 107
 - ANCOVA (see analysis of covariance)
 - ANOVA (see analysis of variance)
-

B

- backward deletion • 175
 - Bartlett's sphericity test • 257
 - best-fitting line • 104, 170
 - beta coefficients • 172
 - beta weights • 172, 182, 184, 193, 196
 - between-groups variability • 72, 73, 80, 107
 - bivariate correlation • 13–14, 175, 182, 188, 205, 207, 212, 241
 - bivariate normality • 33
 - Bonferroni-type adjustment • 128, 146
 - Box's M test • 36, 61–62
-

C

- canonical correlation • 279, 281, 283, 288, 303
- causal modeling • 203–205, 207, 210, 226, 241
- central tendency, measures of • 7
- centroid • 31, 169–170, 288, 298, 304
- chi-square statistic • 31, 53–54, 59, 79, 256, 281, 288, 293, 303, 309, 313, 322
- coefficient of determination • 173
- communalities • 248, 249, 258–260, 266, 269–270, 273–274
- concomitant variable • 99–100, 145
- confirmatory factor analysis • 255
- correlation
 - bivariate • 13, 14, 21–24, 175, 182, 188, 205, 207, 212, 241
 - canonical • 279–281, 283, 288, 303
 - multiple • 146, 172, 176, 181–182, 197, 210, 248
 - partial • 181–184, 188, 193, 310, 314
 - reproduced • 207, 210, 211–219, 221–222, 224–225, 230, 237, 241, 249, 251, 258–259, 261, 266, 273
- correlation matrix • 4, 174, 188, 193, 219, 221, 227, 251–252, 257, 266, 296, 319

- covariate • 15–24, 85, 99–108, 111–113, 116–119, 120–122, 145–150, 153–161, 164
 - covariates
 - number of • 99
 - cross-validation • 176, 289, 294
-

D

- data matrix • 4, 130, 311
 - data screening • 27, 81, 108, 111, 113, 121, 132, 154, 179, 188, 197, 264, 315, 322
 - data transformations • 32, 33–34, 130, 135, 293
 - Decision-Making Tree for Statistical Tests • 13, 20–21
 - descriptive statistics • 7, 28, 38–40, 45, 55, 59, 81, 89, 110, 117, 119, 139, 158, 188, 193, 255, 266
 - determinant • 131, 204–206, 209–210, 213, 226, 241
 - direct causal effect • 204–207, 212
 - direct oblimin • 252, 268
 - direct quartimin • 252
 - discriminant analysis • 17, 21, 279–281, 284–289, 293–294, 303–304, 307–308, 311, 316
 - descriptive • 279
 - direct • 284
 - discriminant function • 279–283, 285–288, 294, 303
 - hierarchical • 284
 - logic behind • 286
 - sequential • 284
 - standard • 284
 - statistical • 284
 - stepwise • 284
 - discriminant score • 281
 - distribution of sample means • 10
 - disturbance term • 205
-

E

- effect size • 11, 77, 81, 84, 89–90, 92, 107–108, 111, 117, 119, 122, 131, 135, 136, 139, 140, 144, 149, 150, 153, 158, 159, 162, 288
- eigenvalue • 248–251, 257–259, 261, 263, 266–267, 269, 272–273, 276, 281, 286, 288, 299, 303
- endogenous variable • 203, 205–207, 209, 212, 213, 219, 221–222, 224–225, 229, 235–237, 241
- equamax • 252, 268, 273
- error variance • 72, 82, 100–102, 145
- eta squared • 77, 81, 89–90, 107, 131, 149
- exogenous variable • 205, 209, 212, 221, 229, 241
- experimentwise alpha level • 74
- exploratory factor analysis • 255, 258
- extraction • 248, 251, 258, 266–267, 273

F

- F* ratio • 71–72, 73, 76, 80–81, 84, 90, 94, 105, 107–108, 119, 122, 132, 135–136, 149–150, 153, 162, 164
factor • 9, 11, 13, 18, 21, 40–41, 53, 59, 61, 71, 74–78, 80–81, 84–85, 87, 89–90, 92–94, 102, 107–108, 111–112, 116–119, 122, 132, 139–140, 146, 149–150, 153, 155–158, 162, 164, 174, 212, 247–252, 254–261, 263–264, 266–269, 271–273, 280–281, 286
factor analysis • 18, 21, 102, 247–248, 251–252, 255–259, 261, 263–264, 266–269, 272–273, 280–281, 286
confirmatory • 255
exploratory • 248, 255, 258, 273
factor correlation matrix • 252
factor loadings • 248, 252, 271, 273, 281
factor scores • 255, 268
factorial designs • 74, 93
factors • 18, 73, 76–77, 81, 84–85, 87, 89–90, 94, 101, 107–108, 111–112, 116, 118, 136, 149, 154, 156, 159, 162, 164, 176, 209, 247–249, 251–252, 254–256, 258, 261, 263, 267–268, 273, 281
first discriminant function • 286
first principal component • 257
fixed effect • 93
forward selection • 175

H

- heteroscedasticity • 36, 57, 130, 179
hierarchical multiple regression • 175
hit rate • 283–285
homogeneity of covariance matrices • 130, 284
homogeneity of regression • 104, 105–107, 108, 116, 118–119, 122, 147–150, 154, 156, 164, 284
homogeneity of regression planes
assumption of • 148
homogeneity of regression slopes
assumption of • 105, 107, 148
homoscedasticity • 28, 32, 35–36, 51, 55, 57–59, 130, 148, 178–179, 183, 190, 194, 197, 210, 219, 227
assumption of • 36, 130, 148, 179
hypothesis
alternative • 10, 72, 74–75
null • 10–12, 33, 36, 72–76, 79, 105, 127–128, 145–146, 148, 257
hypothesis testing • 10–11, 71

I

- indirect causal effect • 204–205, 212
inferential statistics • 7, 9–10, 74

- interaction effect • 75–76, 100–101, 103, 148
disordinal • 76–77
ordinal • 8–9, 76–77, 248
inverse criterion • 128

J

- jackknife procedure • 283

K

- Kolmogorov-Smirnov statistic • 33, 45, 48, 104, 148, 178
kurtosis • 32–33, 45, 48, 79, 104, 148, 178, 257
leptokurtosis • 32
platykurtosis • 32

L

- latent variable • 207–208
latent variable modeling • 207
least squares solution • 171
leptokurtosis • 32
level of significance • 61, 81, 182, 188, 192, 197, 219, 288, 314–315, 322
Levene's test • 35–36, 51–52, 59, 79, 81, 90, 104, 107–108, 118, 148
linearity • 32, 34, 36, 45, 51, 55–57, 59–60, 62, 104, 108, 113, 126, 130, 132, 137, 148, 150, 154, 162, 178–179, 183, 188–190, 194, 197, 210, 230, 256–258, 263–264, 285, 288, 295, 303
LISREL • 208, 255
loading matrix • 252
Log Likelihood • 309, 313, 315, 319–320, 322
logistic regression • 18, 21, 285, 307–309, 311–313, 315–320, 322
logit • 312–313

M

- Mahalanobis distance • 31, 52–54, 59, 176, 183, 189, 195, 219, 258, 264, 285, 288, 314, 316
main effects • 75–77, 81, 84, 88, 90, 94, 100–101, 108, 112, 119, 132, 135, 136, 140, 149, 153–154, 159, 161–162, 164
MANCOVA (see multivariate analysis of covariance)
MANOVA (see multivariate analysis of variance)
factorial • 16–17, 21, 126, 162, 164
matrix algebra • 4
mean • 7–11, 15, 28–30, 32, 38–39, 71–72, 75–78, 90, 103, 126–130, 136, 145, 146–147, 177, 182, 195, 208, 249, 255, 258, 273, 289, 294, 299, 303
measurement overlap • 247

median • 7, 8, 31, 38–39
missing data • 28–29, 37–39, 52, 59, 62, 81, 108, 122, 132, 137, 162, 164, 183, 189, 197, 227, 258, 303, 314, 316
mode • 7–8
model cross-validation • 176
multicollinearity • 173–174, 179, 183, 221, 311, 314, 316, 322
multiple comparisons • 74, 140
multiple correlation • 146, 172, 176, 181–182, 197, 210, 248
multiple regression • 2, 14, 21, 34, 39, 169, 172–178, 181–183, 188–189, 197, 203–204, 207–210, 254, 281, 284, 307–308, 310, 311–312, 314
hierarchical • 175
multivariate • 176
sequential • 175
standard • 175, 197
stepwise • 175–177
multivariate analysis of covariance • 17, 101, 145, 147–149, 153, 164
factorial • 17
multivariate analysis of variance • 16, 125–129, 136, 254
factorial • 17
multivariate multiple regression • 176
multivariate statistics • 2–5, 125

N

nested design • 93, 101
normal probability plot • 32, 194, 257
normal Q-Q plot • 32, 45, 48, 55, 59, 104, 148
normality • 28, 32–33, 36, 41, 45, 48, 55–60, 62, 74, 79–81, 85–86, 113, 122, 130, 132, 137, 148, 162, 164, 178–179, 183, 190, 197, 210, 219, 227, 255–259, 263–264, 266, 284–285, 288, 293–295, 303
multivariate • 33, 36, 55, 59–60, 130, 183, 190, 219, 256–257, 266, 284–285
univariate • 32–33, 55–56, 130, 183, 190, 257, 285
null hypothesis • 10–12, 33, 36, 71–76, 79, 105, 127–128, 145–146, 148, 257

O

oblique rotation • 252, 273
odds • 18, 312–315
odds ratio • 312, 314–316, 319–320, 322
orthoblique • 252, 273
orthogonal rotation • 252, 273
orthogonality • 5, 173
outliers • 28–32, 36, 38–41, 44–45, 52–54, 58, 59, 62, 81, 85, 92, 108, 113, 121, 122, 130, 132, 136–137, 144, 153–154, 161–162, 164, 176, 179, 183, 189,

195, 197, 219, 225, 227, 258, 264, 272, 285, 288, 294–295, 303, 312, 314–316, 322
multivariate • 30–31, 52–53, 59, 285
univariate • 30–31, 39

P

pairwise comparisons • 74
partial correlation • 181–184, 188, 193, 310, 314
partial regression coefficient • 182
path analysis • 14–15, 21, 203–210, 219, 224–227, 236, 241, 249
path coefficient • 14, 204–205, 207, 209–211, 213, 215, 221–222, 224–225, 230, 236, 241
path decomposition • 210, 212, 217, 219, 221–222, 237, 241
path diagram • 14, 205, 207–208, 211, 215, 221, 224, 230, 241
path tracing • 210–212
pattern matrix • 252
Pearson correlation • 14, 132, 137, 150, 154, 170, 172, 182, 248
percentile rank • 8
platykurtosis • 32
post hoc tests • 15–16, 74, 77, 83–84, 89–90, 92, 94, 128–129, 132, 140, 144, 162
power • 11–12, 127, 174, 176, 285, 311, 313
principal components • 18, 21, 248–249, 254–255, 257–258, 263, 267, 272–273
first • 257
principal components analysis • 18, 248–249, 254–255, 257, 263, 272–277
prior probability • 287
probability level • 10–11
promax • 252, 268, 273

Q

quartile deviation • 8
quartimax • 252, 268, 273

R

random effect • 71, 93
range • 8, 28, 31, 37, 44, 174, 248, 252, 288, 296
regression • 14, 18, 21, 29, 34, 39, 52–53, 59, 103–105, 107–108, 116, 118–119, 122, 147–150, 154, 156, 164, 169–179, 181–183, 188–190, 192–197, 203–204, 207–211, 213, 215, 219, 221–222, 227, 236, 241, 247, 254–255, 281, 283–285, 307–319, 322
logistic • 18, 21, 285, 307–309, 311–313, 314–318, 322

multiple • 2, 14, 21, 34, 39, 169, 172–178, 181–183, 188, 189, 197, 203–204, 207–210, 254, 281, 307, 308, 310–312, 314
simple linear • 169, 172
regression coefficients • 172, 174, 181–182, 193, 197, 204, 230, 281, 310, 315, 322
partial • 182
unstandardized • 182, 188, 197, 314
regression line • 103–104, 170–171
relationship, measures of • 9, 34
relative position, measures of • 8
reproduced correlation • 207, 210–213, 215–217, 219, 221–222, 224–225, 230, 237, 241, 249, 251, 258–259, 273
research designs • 2, 3, 74, 93
research hypothesis (see alternative hypothesis)
residuals • 34, 57, 104, 171–172, 177, 179, 183, 190, 193–195, 208–210, 219, 227, 251, 258–259, 261, 263, 266, 269, 272–273, 283, 312, 319
robustness • 32, 74, 79, 130, 132, 137, 285
rotation • 252–253, 259, 261, 263, 266–269, 272–273
direct oblimin • 252
direct quartimin • 252
oblique • 252, 268, 273
orthoblique • 252, 273
orthogonal • 252, 273
promax • 252, 273
varimax • 252–253, 259, 261, 263, 266, 272

S

sampling error • 10–11
scree plot • 247–249, 251, 258–259, 261, 263, 269, 272
sequential analysis • 6–7
sequential multiple regression • 14, 175
significance tests • 10, 32, 104, 148, 210, 215, 287–288, 293, 303
simple linear regression • 169
skewness • 32–33, 45, 48, 55, 79, 81, 86, 104, 130, 148, 178, 257, 285
Spearman rho • 9
spurious effect • 212
standard analysis • 7
standard deviation • 8–11, 28, 30, 79, 84, 89, 119, 135–136, 181–182, 188, 205, 266, 287, 303
standard error • 10, 311, 314
standard multiple regression • 175, 197
standard score • 8
statistical multiple regression • 175
statistics
descriptive • 7, 28, 38–40, 45, 55, 59, 89, 117, 188, 266
inferential • 7, 9–10, 74
stepwise multiple regression • 175
stepwise selection • 175
structural coefficient • 204

structural equation • 204–205, 207–210
structural equation modeling • 204, 207–208
structure matrix • 252, 281, 288
sum of squares
partitioning of • 80
sum-of-squares and cross-products (SSCP) matrix • 131, 146, 148–149, 286
systematic bias • 100

T

T-score • 8–9
t test • 10, 15, 21, 28, 51, 59, 71–72, 74, 79, 310
Table of Statistical Tests • 13, 18–19, 21
tolerance • 174, 183, 193, 196–197, 221, 230, 314, 316
Type I error • 11–12, 74, 127–128, 146, 285
Type II error • 11–12

U

univariate statistics • 3–4
unstandardized regression coefficient • 182, 188, 197, 314

V

variability, measures of • 8
variables • 3
concomitant • 99
dichotomous • 3, 18, 28, 30, 308, 312, 316, 322
endogenous • 205–207, 209, 212–213, 219, 221–222, 224–225, 229, 235–237, 241
exogenous • 205–206, 209, 212, 221, 229, 241
latent • 208
variance-covariance matrix • 4, 36, 61, 130, 132–133, 137, 139–140, 149–150, 154, 156, 162, 164, 285, 296–297
variance inflation factor • 174
varimax • 252–253, 259, 261, 263, 266–268, 267–273

W

Wald statistic • 310, 314–315, 317–318, 322
Wilks' Lambda (Λ) • 128, 131–133, 136, 140, 144, 146, 149–150, 153, 156, 159, 161–132, 164, 281, 287–288, 293, 297, 299, 303
within-groups variability • 73, 80, 100, 107

Z

z-score • 8–9, 30–31, 182, 204