Hubert Gatignon

Statistical Analysis of Management Data

Third Edition

Springer

Statistical Analysis of Management Data

Hubert Gatignon

Statistical Analysis of Management Data

Third Edition

Springer

Hubert Gatignon
INSEAD
Fontainebleau Cedex, France

Statistical Analysis of Management Data. 1st Edition. Kluwer Academic Publishers, 2003
Statistical Analysis of Management Data. 2nd Edition. Springer Science+Business Media, LLC, 2010

ISBN 978-1-4614-8593-3 ISBN 978-1-4614-8594-0 (eBook)
DOI 10.1007/978-1-4614-8594-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013945080

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my daughters, Aline and Valérie

Preface

Preface to First Edition

I am very indebted to a number of people without whom I would not have envisioned this book. First, Paul Green helped me tremendously in the preparation of the first doctoral seminar I taught at the Wharton School. The orientations and objectives set for that book reflect those he had for the seminar on data analysis which he used to teach before I did. A second individual, Lee Cooper at UCLA, was determinant in the approach I used for teaching statistics. As my first teacher of multivariate statistics, the exercise of having to program all the methods in APL taught me the benefits of such an approach for the complete understanding of this material. Finally, I owe a debt to all the doctoral students in the various fields of management, both at Wharton and INSEAD, who have, by their questions and feedback, helped me develop this approach. I hope it will benefit future students in learning these statistical tools, which are basic to academic research in the field of management especially. Special thanks go to Bruce Hardie who helped me put together some of the databases and to Frédéric Dalsace who carefully identified sections that needed further explanation and editing. Also, my research assistant at INSEAD, Gueram Sargsyan, was instrumental in preparing the examples used in this manual to illustrate the various methods.

Preface to Second Edition

This second edition reflects a slight evolution in the methods for analysis of data for research in the field of management and in related fields in the social sciences. In particular, it places a greater emphasis on measurement models. This new version includes a separate chapter on confirmatory factor analysis, with new sections on second order factor analytic models and multiple group factor analysis. A new, separate section on analysis of covariance structure discusses multigroup problems

that are particularly useful for testing moderating effects. Some fundamental multivariate methods such as canonical correlation analysis and cluster analysis have also been added. Canonical correlation analysis is useful because it helps better understand other methodologies already covered in the first version of this book. Cluster analysis remains a classic method used across fields and in applied research.

The philosophy of the book remains identical to that of its original version, which I have put in practice continuously in teaching this material in my doctoral classes. The objectives articulated in Chap. 1 have guided the writing of the first edition of this book but also of this new edition.

In addition to all the individuals I am endebted to and who have been identified in the first edition of this book, I would like to express my thanks to the cohorts of students since then. The continuous feedback has helped select the new material covered in this book with the objective to improve the understanding of the material. Finally, I would like to thank my assistant of fifteen years, Georgette Duprat whose commitment to detail never fails.

Preface to Third Edition

The methods for analyzing data are evolving rapidly as are the software packages that are available. On the one hand, this software, combined with more sophisticated hardware, is increasingly user-friendly. On the other hand, the theories that are being empirically tested and the large databases that have become more easily available require more complex statistical methodologies. While preserving the original objective to provide foundations for the analysis of such data, this third edition develops further those methodologies that are particularly well suited to data analysis in the social sciences. This explains the extensive new chapter on the analysis of mediation and moderation effects. For each of these methods, this edition also contains illustrations of analysis using STATA. I have also introduced XLSTAT as an alternative to multidimensional scaling because of its flexibility and ease of use as Excel macros. I would like to thank especially all my students at INSEAD who have provided feedback on the drafts of these chapters. Particular thanks go to Kathy Sheram who has advised me in editing the third edition of this book. Her professionalism and precision allowed me to communicate more clearly. This is particularly important for social scientists who may not have a technical background. Kathy contributed immensely to presenting the complex material of this book with concision, precision, and clarity.

Contents

Chapter 1
Introduction

This introduction presents important insights into the basic learning philosophy that underpins the presentation style of the statistical methods and techniques explored in this book. It discusses the types of measurements that are available to researchers and how these measurements often determine what statistical methods may be used to analyze particular data. Indeed, as this first chapter describes, the nature of the measurement scales has determined the structure of this book.

1.1 Overview

This book covers multivariate statistical analyses that are important for researchers in all fields of management whether finance, production, accounting, marketing, strategy, technology, or human resources management. Although multivariate statistical techniques such as those described in this book play key roles in fundamental disciplines of the social sciences (e.g., economics and econometrics or psychology and psychometrics), the methodologies particularly relevant to and typically used in management research are the central focus of this study.

This book is especially designed to provide doctoral students with a theoretical knowledge of the basic concepts underlying the most important multivariate techniques and with an overview of actual applications in various fields. The book addresses both the underlying mathematics and *problems of application*. As such, a reasonable level of competence in both statistics and mathematics is needed. This book is not intended as a first introduction to statistics and statistical analysis. Instead, it assumes that the student is familiar with basic univariate statistical techniques. The book presents the techniques in a fundamental way but in a format accessible to students in a doctoral program, as well as to practicing academicians and data analysts. With this in mind, the reader may wish to review some basic statistics and matrix algebra such as those provided in the following books:

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_1, © Springer Science+Business Media New York 2014

Green, Paul E. (1978), *Mathematical Tools for Applied Multivariate Analysis*, New York, NY: Academic Press, [Chapters 2–4].
Maddala, Gangadharrao S. (1977), *Econometrics*, New York, NY: McGraw Hill, Inc. [Appendix A].

This book offers a clear, succinct exposition of each technique, with emphasis on when it is appropriate to use each technique and how to do so. The focus is on the essential aspects that a working researcher will encounter, in short, on using multivariate analysis appropriately through an understanding of the foundations of the methods to gain valid and fruitful insights into management problems. This book presents methodologies for analyzing primary or secondary data typically used by academics as well as analysts in management research and provides an opportunity for the researcher to gain hands-on experience with such methods.

1.2 Objectives

The main objectives of this book are:

1. To develop the student's knowledge of the technical details of various techniques for analyzing data.
2. To expose students to applications and hands-on use of various computer programs: This experience will enable students to carry out statistical analyses of their own data. Commonly available software is used throughout the book as much as possible, across methodologies, to avoid having to learn multiple systems, each with its own specific data manipulations and commands. In particular, most analyses are demonstrated with SAS and STATA. However, several additional statistical packages are used when particularly adapted to specific types of analysis, e.g., LIMDEP, LISREL, or XLSTAT.

1.2.1 Develop the Student's Knowledge of the Technical Details of Various Techniques for Analyzing Data

The first objective is to prepare the researcher with the basic technical knowledge required to understand the methods, to be able to use them appropriately, to know their limitations, and to access more advanced material about them. This requires a thorough understanding of the fundamental properties of the techniques. "Basic" knowledge means the book will not go into the more advanced issues of the methodologies. Understanding of such issues should be acquired later through specialized, more advanced study on the specific topics. The objective of this book is to provide enough detail for what is the minimum knowledge expected from a doctoral candidate in management studies or an academic researcher in management.

1.2.2 Expose the Student to Applications and Hands-On Use of Various Computer Programs for Carrying Out Statistical Analyses of Data

While the basic statistical methods corresponding to the various types of analysis are necessary, they are not sufficient to do research. The use of any method requires the knowledge of the statistical software corresponding to these analyses. It is indispensable that students learn both the statistical theory *and the practice* of using these methods *at the same time*. A very effective, albeit time-consuming, way to ensure that the intricacies of a technique are mastered is by programming the software oneself. A quicker way is to ensure that the use of the software coincides with the learning of the method by associating application examples with the abstract knowledge of the method and by analyzing data oneself using these methods.

Consequently, in this book each chapter contains four sections. The first section presents the methods from a theoretical point of view with the various properties of the method. The second section shows an example of an analysis with instructions on how to use a particular software program appropriate for that analysis. The third section gives an assignment so that students can actually practice the method of analysis. The data sets for these assignments are described in Appendix C (Chap. 14) and can be downloaded from the Web page of Hubert Gatignon at http://www.insead.edu/facultyresearch/faculty/profiles/hgatignon. Finally, the fourth section consists of a list of reference articles that use such techniques appropriately and serve as templates. Selected readings could have been reprinted in this book for each application; however, few articles illustrate all the facets of the techniques. Offering a range of articles allows students to choose the applications that correspond best to their interests. By accessing multiple articles in the area of interest, students enrich their learning. All these articles illustrating the particular multivariate techniques used in empirical analysis are drawn from the major research journals in the field of management.

1.3 Types of Scales

Data used in management research are obtained from existing sources (secondary data) such as data published by Ward for automobile sales in the USA or from vendors who collect data, such as panel data. Data are also collected for the explicit purpose of the study (primary data): survey data, scanner data, or panels.

In addition to this variety of data sources, differences in the type of data that are collected can be critical for their analysis. Some data are continuous measures, for example, the age of a person, with an absolute starting point at birth or the distance between two points. Some commonly used data do not have such an absolute starting point, for example, temperature. Yet in both cases, i.e., temperatures and

distances, multiple units of measurement exist throughout the world. These differences in the type of data are critical because the appropriateness of data analysis methods varies depending on the type of data at hand. In fact, very often the data may have to be collected in a certain way in order to be able to test hypotheses using the appropriate methodology. Failure to collect the appropriate type of data would prevent performing the test.

In this first chapter, we discuss the different types of scales that can be found in measuring variables used in management research.

1.3.1 Definition of Different Types of Scales

Scales are quantitative measures of a particular construct, usually not observed directly. Four basic types of scales can categorize management measurements:

- Ratio
- Interval
- Rank order or ordinal
- Categorical or nominal

1.3.2 The Impact of the Type of Scale on Statistical Analysis

The nature of analysis depends in particular on the scale of the variable(s). Table 1.1 summarizes the most frequently used statistics that are permissible according to the scale type. The order of the scales in the first column of Table 1.1 (from the top with "nominal" to the bottom with "ratio") is hierarchical in the sense that statistics that are permissible for a scale (a row of Table 1.1) are also permissible for the scale(s) below it. For example, a median is a legitimate statistic for an ordinal-scale variable but is also legitimate for an interval or a ratio scale. The reverse is not true; for example, a mean is not legitimate for an ordinal scale.

1.4 Topics Covered

This book presents the major methods of analysis that have been used in the recent management research literature. A survey of the leading journals in the various fields of management was conducted to identify these methods. This survey revealed interesting observations.

It is striking that the majority of the analyses involve the estimation of a single equation or of several equations independent of one another. Analyses involving a system of equations represent a very small percentage of the analyses performed in these articles. This appears at first glance surprising given the complexity of

Table 1.1 Scales of measurement and their properties

Scale	Mathematical group structure	Permissible statistics	Typical examples
Nominal	Permutation group $y = f(x)$ [f(x) means any one-to-one correspondence]	• Frequency distribution • Mode	• Numbering of brands • Assignment of numbers to types of products or models • Gender of consumers • Organization types
Ordinal	Isotonic group $y = f(x)$ [f(x) means any increasing monotonic function]	• Median • Percentiles • Order (Spearman) correlations • Sign test	• Order of entry • Rank order of preferences
Interval	General linear group $y = a + bx$ $b > 0$	• Mean • Average deviation • Standard deviation • Product–moment correlation • t test • F test	• Likert scale items (agree–disagree) • Semantic scale items (ratings on opposite adjectives)
Ratio	Similarity group $y = cx$ $c > 0$	• Geometric mean • Coefficient of variation	• Sales • Market share • Advertising expenditures

Adapted from Stevens (1962), p. 25, Stevens (1959), p. 27, and Green and Tull (1970), p.181

management phenomena. Possibly some of the simultaneous relationships analyzed are reflected in methodologies that explicitly consider measurement errors; these techniques appear to have grown in recent years. This is why the methodologies used for measurement modeling receive special attention in this book. Factor analysis is a fundamental method found in a significant proportion of the studies, typically to verify the unidimensionality of the constructs measured. The more advanced aspects such as second-order factor analysis and multiple-group factor analysis have gained popularity and are also discussed. Choice modeling has been an important topic, especially in marketing but also in the other fields of management, with studies estimating probit or logit models. A still very small percentage of articles use these models for ordered choice data (i.e., where the data reflect only the order in which brands are ranked from best to worst). Analysis of proximity data concerns few studies but cluster analysis and multidimensional scaling remain favorite methods for practice analysts.

Based on these survey results, the topics listed below were selected. They have been classified according to the type of key variable(s) that is of primary interest in the analysis. Indeed, as we discuss in Chap. 2 the nature of the criterion (also called dependent or endogenous) variable(s) determines the type of statistical analysis that may be performed. Consequently, the first issue that we address concerns the nature and properties of variables and the process of generating scales with the appropriate statistical procedures, followed by discussions of the various statistical methods of data analysis.

Introduction to multivariate statistics and tests about means

- Multivariate analysis of variance

Multiple item measures

- Reliability alpha
- Principle component analysis
- Exploratory factor analysis
- Confirmatory factor analysis
- Second-order factor analysis
- Multi-group factor analysis

Canonical correlation analysis

Single-equation econometrics

- Ordinary least squares
- Generalized least squares
- Tests of homogeneity of coefficients: pooling tests

System of equations econometrics

- Seemingly unrelated regression
- Two-stage least squares
- Three-stage least squares

Categorical dependent variables

- Discriminant analysis
- Quantal choice models: logit

Rank-ordered data

- Conjoint analysis
- Ordered probit

Analysis of covariance structure—Structural equation models

- LISREL

Testing mediation and moderation effects

Analysis of similarity data

- Cluster analysis
- Multidimensional scaling

A new chapter (Chap. 11) has been added in this third edition of ***Statistical Analysis of Management Data*** to reflect the increased use of mediation and moderation analysis in management research. This chapter covers the various techniques that are adapted to test theories that involve such processes.

1.5 Pedagogy

Three key learning outcomes are necessary in order to achieve the objectives of this book:

1. Having sufficient knowledge of statistical theory to be able to understand the methodologies, when they are applicable and when they are not appropriate.
2. Being able to perform such analyses using the proper statistical software.
3. Understanding how these methodologies have been applied in management research.

This book differs from others in that it is the only text on multivariate statistics or data analysis that addresses the specific needs of doctoral education. The three outcomes outlined above are weighted differently. This book emphasizes the first outcome by providing the mathematical and statistical analyses necessary to fully understand the given methodologies. This is in contrast to other books that prefer primarily or exclusively a verbal description of the method.

This book favors the understanding of the rationale for modeling choices, issues, and problems. While the verbal description of a method may be more easily accessible to a wider audience, it is often more difficult to follow the rationale, which is based on mathematics. For example, it is difficult to understand the problem of multicollinearity without understanding the effect on the determinant of the covariance matrix that needs to be inverted. The learning that results from verbal presentation tends, therefore, to be more mechanical.

This book also differs in that, instead of choosing only a few articles to illustrate the applications of the methods, as would be found in a book of readings (sometimes with short introductions), a broad list of application readings is provided. These readings tend to be relatively easy to access, especially with services available through the Internet. They cover a large cross section of examples and a history of the literature in this domain.

Finally, the examples of analyses are relatively self-explanatory and, although some explanations of the statistical software used are provided with each example, this book does not intend to replace the instruction manuals of those particular software packages. The reader is referred to those packages for details.

In summary, this book puts the emphasis on understanding the statistical methodology while providing enough information for the reader to develop skills in performing the analyses and in understanding how to apply them to management research problems.

More specifically, the learning of this material involves two parts: the learning of the statistical theory behind the technique and the learning of how to use the technique. Although there may be different ways to combine these two experiences, we recommend that students (1) learn the theory by reading the sections where the methodologies are presented and discussed, (2) study an actual example of the statistical software package (e.g., SAS, STATA, LIMDEP, LISREL, and other specialized packages) that is used to apply the methodology, (3) apply the technique

themselves using the data sets available from the Web page of Hubert Gatignon at http://www.insead.edu/facultyresearch/faculty/profiles/hgatignon, and finally, (4) explore application issues as illustrated by applications found in prior research and listed at the end of each chapter.

In addition to the books and articles listed in each chapter, the following books are highly recommended to further develop the student's skills in various methods of data analysis. Each of these books is more specialized and covers only a subset of the methods presented in this book. However, they are indispensable complements for students wishing to become proficient in the techniques used in research.

Bibliography

Green, P. E., & Tull, D. S. (1970). *Research for marketing decisions*. Englewood Cliffs, NJ: Prentice-Hall.

Greene, W. H. (1993). *Econometric analysis*. New York: MacMillan.

Hanssens, D. M., Parsons, L. J., & Shultz, R. L. (1990). *Market response models: econometric and time series analysis*. Norwell: Kluwer.

Judge, G. G., Griffiths, W. E., Carter Hill, R., Lutkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics*. New York, NY: Wiley.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories*. New York, NY: Wiley.

Stevens, S. S. (1962). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York, NY: Wiley.

Chapter 2
Multivariate Normal Distribution

In this chapter, we define the univariate and multivariate normal distribution density functions and then we discuss the tests of differences of means for multiple variables simultaneously across groups.

2.1 Univariate Normal Distribution

To review, in the case of a single random variable, the probability distribution or the density function of that variable x is represented by Eq. (2.1):

$$\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \tag{2.1}$$

2.2 Bivariate Normal Distribution

The bivariate distribution represents the joint distribution of two random variables. The two random variables x_1 and x_2 are related to each other in the sense that they are not independent of each other. This dependence is reflected by the correlation ρ between the two variables x_1 and x_2. The density function for the two variables jointly is

$$\Phi(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} \right] \right\} \tag{2.2}$$

This function can be represented graphically as in Fig. 2.1.

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_2, © Springer Science+Business Media New York 2014

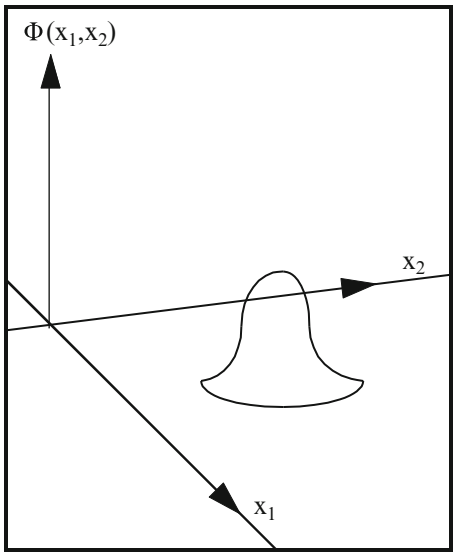


Fig. 2.1 The bivariate normal distribution

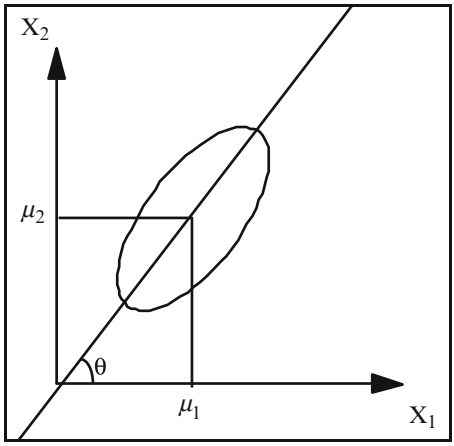


Fig. 2.2 The locus of points of the bivariate normal distribution at a given density level

The *isodensity contour* is defined as the set of points for which the values of x_1 and x_2 give the same value for the density function Φ. This contour is given by Eq. (2.3) for a fixed value of C, which defines a constant probability:

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}=C \tag{2.3}$$

Equation (2.3) defines an ellipse with centroid (μ_1,μ_2). This ellipse is the locus of points representing the combinations of the values of x_1 and x_2 with the same probability, as defined by the constant C (Fig. 2.2).

For various values of C, we get a family of concentric ellipses (at a different cut, i.e., cross section of the density surface with planes at various elevations) (see Fig. 2.3).

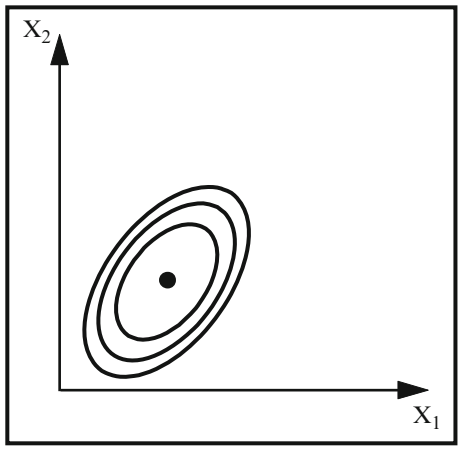


Fig. 2.3 Concentric ellipses at various density levels

The angle θ depends only on the values of σ_1, σ_2, and ρ. The higher the correlation between x_1 and x_2, the steeper the line going through the origin with angle θ, i.e., the bigger the angle.

2.3 Generalization to Multivariate Case

Let us represent the bivariate distribution in matrix algebra notation in order to derive the generalized format for more than two random variables.

The covariance matrix of (x_1, $x2$) can be written as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{2.4}$$

The determinant of the matrix $\mathbf{\Sigma}$ is

$$|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2\left(1-\rho^2\right) \tag{2.5}$$

Equation (2.3) can now be re-written as

$$C = [x_1-\mu_1, x_2-\mu_2]\mathbf{\Sigma}^{-1}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \tag{2.6}$$

where

$$\mathbf{\Sigma}^{-1} = 1/\left[\sigma_1^2\sigma_2^2\left(1-\rho^2\right)\right]\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} = \frac{1}{1-\rho^2}\begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \tag{2.7}$$

Note that $\boldsymbol{\Sigma}^{-1} = |\boldsymbol{\Sigma}|^{-1} \times$ matrix of cofactors.
Let

$$\mathbf{X} = \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

Then $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} = \chi^2$, which is a quadratic form of the variables $\mathbf{x}$ and is, therefore, a chi-square variate.

Also, because $|\boldsymbol{\Sigma}| = {\sigma_1}^2{\sigma_2}^2(1 - \rho^2), |\boldsymbol{\Sigma}|^{1/2} = \sigma_1\sigma_2\sqrt{(1-\rho^2)}$, and consequently,

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} = (2\pi)^{-1}|\boldsymbol{\Sigma}|^{-1/2} \tag{2.8}$$

The bivariate distribution function can now be expressed in matrix notation as

$$\Phi(x_1, x_2) = (2\pi)^{-1}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}} \tag{2.9}$$

Now, more generally with p random variables (x_1, x_2, . . ., x_p), let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

The density function is

$$\Phi(\mathbf{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{\left[-\frac{1}{2}(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right]} \tag{2.10}$$

For a fixed value of the density Φ, an ellipsoid is described. Let $\mathbf{X} = \mathbf{x} - \mu$. The inequality $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} \leq \chi^2$ defines any point within the ellipsoid.

2.4 Tests About Means

2.4.1 Sampling Distribution of Sample Centroids

2.4.1.1 Univariate Distribution

A random variable is normally distributed with mean μ and variance σ^2:

$$x \sim N(\mu, \sigma^2) \tag{2.11}$$

After n independent draws, the mean is randomly distributed with mean μ and variance σ^2/n:

$$\overline{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{2.12}$$

2.4.1.2 Multivariate Distribution

In the multivariate case with p random variables, where $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$, $\mathbf{x}$ is normally distributed following the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$:

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.13}$$

The mean vector for the sample of size n is denoted by

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix}$$

This sample mean vector is normally distributed with a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/n$:

$$\overline{\mathbf{x}} \sim N\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{n}\right) \tag{2.14}$$

2.4.2 Significance Test: One-Sample Problem

2.4.2.1 Univariate Test

The univariate test is illustrated in the following example. Let us test the hypothesis that the mean is 150 (i.e., $\mu_\circ = 150$) with the following information:

$$\sigma^2 = 256; \;\; n = 64; \;\; \overline{x} = 154$$

Then, the z score can be computed:

$$z = \frac{154 - 150}{\sqrt{\frac{256}{64}}} = \frac{4}{\frac{16}{8}} = 2$$

At $\alpha = 0.05$ (95% confidence interval), $z = 1.96$, as obtained from a normal distribution table. Therefore, the hypothesis is rejected. The confidence interval is

$$\left[154 - 1.96 \times \frac{16}{8},\ \ 154 + 1.96 \times \frac{16}{8}\right] = [150.08,\ \ 157.92]$$

This interval excludes 150. The hypothesis that $\mu_\circ = 150$ is rejected. If the variance σ had been unknown, the t statistic would have been used:

$$t = \frac{\overline{x} - \mu_\circ}{s/\sqrt{n}} \tag{2.15}$$

where s is the observed sample standard deviation.

2.4.2.2 Multivariate Test with Known Σ

Let us take an example with two random variables:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 25 & 10 \\ 10 & 16 \end{bmatrix} \quad n = 36$$

$$\overline{\mathbf{x}} = \begin{bmatrix} 20.3 \\ 12.6 \end{bmatrix}$$

The hypothesis is now about the mean values stated in terms of the two variables jointly:

$$H: \quad \boldsymbol{\mu}_\circ = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$$

At the alpha level of 0.05, the value of the density function can be written as in Eq. (2.16), which follows a chi-square distribution at the specified significance level α:

$$n(\boldsymbol{\mu}_\circ - \overline{\mathbf{x}})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_\circ - \overline{\mathbf{x}}) \sim \chi_p^2(\alpha) \tag{2.16}$$

Computing the value of the statistics,

$$|\boldsymbol{\Sigma}| = 25 \times 16 - 10 \times 10 = 300$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{300} \begin{bmatrix} 16 & -10 \\ -10 & 25 \end{bmatrix}$$

$$\chi^2 = 36 \times \frac{1}{300} (20 - 20.3,\ \ 15 - 12.6) \begin{bmatrix} 16 & -10 \\ -10 & 25 \end{bmatrix} \begin{bmatrix} 20 - 20.3 \\ 15 - 12.6 \end{bmatrix} = 15.72$$

The critical value at an alpha value of 0.05 with 2 degrees of freedom is provided by tables:

$$\chi^2_{p=2}(\alpha = 0.05) = 5.991$$

The observed value is greater than the critical value. Therefore, the hypothesis that $\boldsymbol{\mu} = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$ is rejected.

2.4.2.3 Multivariate Test with Unknown Σ

Just as in the univariate case, $\boldsymbol{\Sigma}$ is replaced with the sample value $\mathbf{S}/(n-1)$, where $\mathbf{S}$ is the sums-of-squares-and-cross-products (SSCP) matrix, which provides an unbiased estimate of the covariance matrix. The following statistics are then used to test the hypothesis:

$$\text{Hotelling}: \quad T^2 = n(n-1)(\overline{\mathbf{x}} - \boldsymbol{\mu}_\circ)' \mathbf{S}^{-1} (\overline{\mathbf{x}} - \boldsymbol{\mu}_\circ) \tag{2.17}$$

where if

$$\underset{n \times p}{\mathbf{X}^d} = \begin{bmatrix} x_{11} - \overline{x}_1 & x_{21} - \overline{x}_2 & \dots \\ x_{12} - \overline{x}_1 & x_{22} - \overline{x}_2 & \dots \\ \vdots & \vdots & \\ x_{1n} - \overline{x}_1 & x_{2n} - \overline{x}_2 & \dots \end{bmatrix}$$

then

$$\mathbf{S} = \mathbf{X}^{d\prime} \mathbf{X}^d$$

Hotelling showed that

$$\frac{n-p}{(n-1)p} T^2 \sim F^p_{n-p} \tag{2.18}$$

Replacing T^2 by its expression given in Eq. (2.17) leads to

$$\frac{n(n-p)}{p} (\overline{\mathbf{x}} - \boldsymbol{\mu}_\circ)' \mathbf{S}^{-1} (\overline{\mathbf{x}} - \boldsymbol{\mu}_\circ) \sim F^p_{n-p} \tag{2.19}$$

Consequently, the test is performed by computing the expression in Eq. (2.19) and by comparing its value with the critical value obtained in an F table with p and n-p degrees of freedom.

2.4.3 Significance Test: Two-Sample Problem

2.4.3.1 Univariate Test

Let us define $\overline{x}_1$ and $\overline{x}_2$ as the means of a variable on two unrelated samples. The test for the significance of the difference between the two means is given by

$$t = \frac{(\overline{x}_1 - \overline{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad or \quad t^2 = \frac{(\overline{x}_1 - \overline{x}_2)^2}{s^2\left(\frac{n_1 + n_2}{n_1 n_2}\right)} \tag{2.20}$$

where

$$s = \frac{\sqrt{(n_1 - 1)\frac{\sum\limits_i x_{1i}^2}{n_1 - 1} + (n_2 - 1)\frac{\sum\limits_i x_{2i}^2}{n_2 - 1}}}{(n_1 - 1) + (n_2 - 1)} = \sqrt{\frac{\sum\limits_i x_{1i}^2 + \sum\limits_i x_{2i}^2}{n_1 + n_2 - 2}} \tag{2.21}$$

s^2 is the pooled within-groups variance. It is an estimate of the assumed common variance $\boldsymbol{\sigma}^2$ of the two populations.

2.4.3.2 Multivariate Test

Let $\overline{\mathbf{x}}^{(1)}$ be the mean vector in sample 1 $= \begin{bmatrix} \overline{x}_1^{(1)} \\ \overline{x}_2^{(1)} \\ \vdots \\ \overline{x}_p^{(1)} \end{bmatrix}$ and similarly for sample 2.

We need to test the significance of the difference between $\overline{\mathbf{x}}^{(1)}$ and $\overline{\mathbf{x}}^{(2)}$. We will consider first the case where the covariance matrix, which is assumed to be the same in the two samples, is known. Then we will consider the case where an estimate of the covariance matrix needs to be used.

Σ Is Known (The Same in the Two Samples)

In this case, the difference between the two group means is normally distributed with a multivariate normal distribution:

$$\left(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}\right) \sim N\left(\mu_1 - \mu_2,\ \boldsymbol{\Sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \tag{2.22}$$

The computations for testing the significance of the differences are similar to those in Sect. 2.4.2.2 using the chi-square test.

Σ Is Unknown

If the covariance matrix is not known, it is estimated using the covariance matrices within each group but pooled.

Let $\mathbf{W}$ be the within-groups SSCP matrix. This matrix is computed from the matrix of deviations from the means on all p variables for each of n_k observations (individuals). For each group k,

$$\underset{n_k \times p}{\mathbf{X}^{d(k)}} = \begin{bmatrix} x_{11}^{(k)} - \overline{x}_1^{(k)} & x_{21}^{(k)} - \overline{x}_2^{(k)} & \dots \\ x_{12}^{(k)} - \overline{x}_1^{(k)} & x_{22}^{(k)} - \overline{x}_2^{(k)} & \dots \\ \vdots & \vdots & \\ x_{1n_k}^{(k)} - \overline{x}_1^{(k)} & x_{2n_k}^{(k)} - \overline{x}_2^{(k)} & \dots \end{bmatrix} \tag{2.23}$$

For each of the two groups (each k), the SSCP matrix can be derived:

$$\underset{p \times p}{\mathbf{S}_k} = \underset{p \times n_k}{\mathbf{X}^{\prime d(k)}} \underset{n_k \times p}{\mathbf{X}^{d(k)}} \tag{2.24}$$

The pooled SSCP matrix for the more general case of K groups is

$$\underset{p \times p}{\mathbf{W}} = \sum_{k=1}^{K} \underset{p \times p}{\mathbf{S}_k} \tag{2.25}$$

In the case of two groups, K is simply equal to 2.

Then, we can apply Hotelling's T, just as in Sect. 2.4.2.3, where the proper degrees of freedom depending on the number of observations in each group (n_k) are applied:

$$T^2 = \left(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}\right)' \left[\frac{\mathbf{W}}{n_1 + n_2 - 2} \frac{n_1 + n_2}{n_1 n_2}\right]^{-1} \left(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}\right) \tag{2.26}$$

$$= \frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2} \left(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}\right)' \mathbf{W}^{-1} \left(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}\right) \tag{2.27}$$

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \sim F^p_{n_1 + n_2 - p - 1} \tag{2.28}$$

2.4.4 Significance Test: K-Sample Problem

As in the case of two samples, the null hypothesis is that the mean vectors across the K groups are the same and the alternative hypothesis is that they are different.

Let us define Wilk's likelihood-ratio criterion:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \tag{2.29}$$

where $\mathbf{T}$ = total SSCP matrix and $\mathbf{W}$ = within-groups SSCP matrix.

$\mathbf{W}$ is defined as in Eq. (2.25). The total SSCP matrix is the sums of squares and cross products applied to the deviations from the grand means (i.e., the overall mean across the total sample with the observations of all the groups for each variable). Therefore, let the mean centered data for group k be noted as

$$\underset{n_k \times p}{\mathbf{X}^{d^*(k)}} = \begin{bmatrix} x_{11}^{(k)} - \overline{x}_1 & x_{21}^{(k)} - \overline{x}_2 & \dots \\ x_{12}^{(k)} - \overline{x}_1 & x_{22}^{(k)} - \overline{x}_2 & \dots \\ \vdots & \vdots & \\ x_{1n_k}^{(k)} - \overline{x}_1 & x_{2n_k}^{(k)} - \overline{x}_2 & \dots \end{bmatrix} \tag{2.30}$$

where $\overline{x}_j$ is the overall mean of the j's variate.

We create a new data matrix that comprises the centered data for each of the groups, stacked one upon the other:

$$\underset{n \times p}{\mathbf{X}^{d^*}} = \begin{bmatrix} \mathbf{X}^{d^*(1)} \\ \mathbf{X}^{d^*(2)} \\ \vdots \\ \mathbf{X}^{d^*(K)} \end{bmatrix} \tag{2.31}$$

The total SSCP matrix $\mathbf{T}$ is then defined as

$$\underset{p \times p}{\mathbf{T}} = \underset{p \times n}{\mathbf{X}^{d^{*\prime}}} \; \underset{n \times p}{\mathbf{X}^{d^*}} \tag{2.32}$$

Intuitively, if we reduce the space to a single variate so that we are only dealing with variances and no covariances, Wilk's lambda (Λ) is the ratio of the pooled within-groups variance to the total variance. If the group means are the same, the variances are equal and the ratio equals one. As the group means differ, the total variance becomes larger than the pooled within-groups variance. Consequently, the ratio Λ becomes smaller. Because of the existence of more than one variate, which implies more than one variance and covariances, the within-SSCP and total-SSCP matrices need to be reduced to a scalar in order to derive a scalar ratio. This is the role of the determinants. However, the interpretation remains the same as for the univariate case.

It should be noted that Wilk's Λ can be expressed as a function of the eigenvalues of $\mathbf{W^{-1}B}$ where $\mathbf{B}$ is the between-group covariance matrix (eigenvalues are explained in the next chapter). From the definition of Λ in Eq. (2.29), it follows that

$$\frac{1}{\Lambda} = \frac{|\mathbf{T}|}{|\mathbf{W}|} = \left|\mathbf{W}^{-1}\mathbf{T}\right| = \left|\mathbf{W}^{-1}(\mathbf{W}+\mathbf{B})\right| = \left|\mathrm{I} + \mathbf{W}^{-1}\mathbf{B}\right| = \prod_{i=1}^{K}(1+\lambda_i) \tag{2.33}$$

and consequently,

$$\Lambda = \frac{1}{\prod_{i=1}^{K}(1+\lambda_i)} = \prod_{i=1}^{K}\frac{1}{(1+\lambda_i)} \tag{2.34}$$

Also, it follows that

$$\mathrm{Ln}\Lambda = \mathrm{Ln}\frac{1}{\prod_{i=1}^{K}(1+\lambda_i)} = -\sum_{i=1}^{K}\mathrm{Ln}(1+\lambda_i) \tag{2.35}$$

When Wilk's Λ approaches 1, we showed that it means that the difference in means is negligible. This is the case when Ln Λ approaches 0. However, when Λ approaches 0, it means that the difference is large. Therefore, a large value of − LnΛ is an indication of the significance of the difference between the means.

Based on Wilk's Λ, we present two statistical tests: Bartlett's V and Rao's R.

Let N = total sample size across samples, p = number of variables, and K = number of groups (number of samples).

Bartlett's V is approximately distributed as a chi-square when $N - 1 - (p + K)/2$ is large:

$$V = -[N - 1 - (p + K)/2]\text{Ln}\Lambda \sim \chi^2_{p(K-1)} \tag{2.36}$$

Bartlett's V is relatively easy to calculate and can be used when $N - 1 - (p + K)/2$ is large.

Another test, Rao's R, can be applied; it is distributed approximately as an F variate. It is calculated as follows:

$$R = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{wt - p(K-1)/2 + 1}{p(K-1)} \approx F^{\nu_1 = p(K-1)}_{\nu_2 = wt - p(K-1)/2 + 1} \tag{2.37}$$

where

$$w = N - 1 - (p + K)/2$$

$$t = \sqrt{\frac{p^2(K-1)^2 - 4}{p^2 + (K-1)^2 - 5}}$$

The parameter t is set to 1 if either the numerator or the denominator of this last expression equals 0. The F statistic is exact when there are only one or two variables (p) or when the number of groups (K) equals 2 or 3.

A significant chi-square for Bartlett's test or a significant F test for Rao's test indicates significant differences in the group means.

2.5 Examples

2.5.1 Test of the Difference Between Two Mean Vectors: One-Sample Problem

In this example, the file "MKT_DATA" contains data about the market share of a brand over seven periods, as well as the percentage of distribution coverage and the price of the brand. These data correspond to one market, Norway. The question is whether or not the market share, distribution coverage, and prices are similar or

Table 2.1 Data example for the analysis of three variables

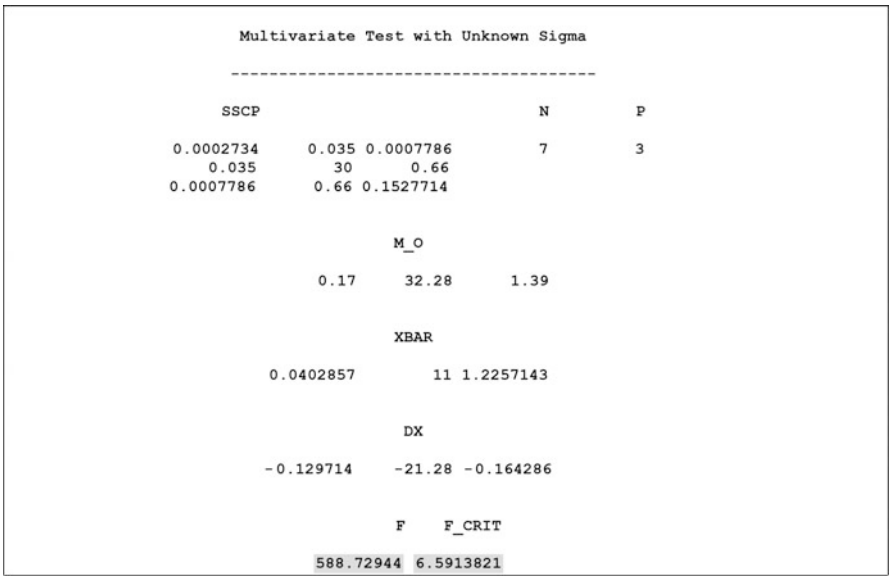
PERIOD	M_SHARE	DIST	PRICE
1	0.038	11	0.98
2	0.044	11	1.08
3	0.039	9	1.13
4	0.03	9	1.31
5	0.036	14	1.36
6	0.051	14	1.38
7	0.044	9	1.34

```
/* ************ Example2-1.sas ************** */
OPTIONS LS=80;
DATA work;
INFILE
"C:\SAMD\Chapter2\Examples\Mkt_Data.csv"
dlm = ',' firstobs=2;
INPUT PERIOD M_SHARE DIST PRICE;
data work;
        set work (drop = period) ;
run;
/* Multivariate Test with Unknown Sigma */
proc iml;
print " Multivariate Test with Unknown Sigma " ;
print "---------------------------------------" ;
use work;                /* Specifying the matrix with raw market data for Norway */
read all var {M_Share Dist Price} into Mkt_Data;
start SSCP;                        /* SUBROUTINE for calculation of the SSCP matrix */
        n=nrow(x);                      /* Number of rows */
        mean=x[+,]/n;                   /* Column means   */
        x=x-repeat(mean,n,1);           /* Variances      */
        sscp = x`*x;                    /* SSCP matrix    */
finish sscp;                            /* END SUBROUTINE */
x=Mkt_Data;                             /* Definition of the data matrix */
p=ncol(Mkt_Data);
run sscp;                               /* Execution of the SUBROUTINE   */
print SSCP  n p;

Xbar = mean;                            /* Definition of the mean vector */
m_o = { 0.17 32.28 1.39 };              /* Myu zero: the mean vector for Europe */

dX = Xbar - m_o;                        /* Matrix of deviations */
dXt = dX`;                              /* Calculation of the transpose of dX */

print m_o;
print Xbar;
print dX;

sscp_1 = inv(sscp);                  /* Calculation of the inverse of SSCP matrix */

T_sq = n*(n-1)*dX*sscp_1*dXt;           /* Calculation of the T_square    */
F    = T_sq*(n-p)/((n-1)*p);            /* Calculation of the F statistic */

Df_num = p;
Df_den = n-p ;
F_crit = finv(.95,df_num,df_den);     /* Critical F for .05 for df_num, df_den */
Print F F_crit;
quit;
```

Fig. 2.4 SAS input to perform the test of a mean vector (examp2-1.sas)

different from the data of that same brand for the rest of Europe, i.e., with values of market share, distribution coverage, and price, respectively, of 0.17, 32.28, and 1.39. The data are shown in Table 2.1.

The SAS file showing the SAS code needed to compute the necessary statistics is shown in Fig. 2.4. The first lines correspond to the basic SAS commands to read the



```
                Multivariate Test with Unknown Sigma

                ---------------------------------------

                SSCP                              N         P

              0.0002734      0.035 0.0007786      7         3
                  0.035         30      0.66
              0.0007786       0.66 0.1527714


                                M_O

                          0.17     32.28      1.39


                                XBAR

                       0.0402857        11 1.2257143


                                 DX

                      -0.129714    -21.28 -0.164286


                                 F     F_CRIT

                           588.72944  6.5913821
```

Fig. 2.5 SAS output of analysis defined in Fig. 2.4 (examp2-1.lst)

data from the file. Here, the data file was saved as a text file from Microsoft Excel. Consequently, the values in the file corresponding to different data points are separated by commas. This is indicated as the delimiter ("dlm"). Also, the data (first observation) start on line 2 because the first line is used for the names of the variables (as illustrated in Table 2.1). The variable PERIOD is dropped so that only the three variables needed for the analysis are kept in the SAS working data set. The IML procedure is used to perform matrix algebra computations.

This file could easily be used for the analysis of different databases. Obviously, it would be necessary to adapt some of the commands, especially the file name and path and the variables. Within the IML subroutine, only two items would need to be changed: (1) the variables used for the analysis and (2) the values for the null hypothesis (m_o).

The results are printed in the output file shown in Fig. 2.5.

The critical F statistic with 3 and 4 degrees of freedom at the 0.05 confidence level is 6.591, while the computed value is 588.7, indicating that the hypothesis of no difference is rejected.

2.5.2 *Test of the Difference Between Several Mean Vectors: K-Sample Problem*

The next example considers similar data for three different countries (Belgium, France, and the United Kingdom) for seven periods, as shown in Table 2.2. The question is whether or not the mean vectors are the same for the three countries.

Table 2.2 Data example for three variables in three countries (groups)

CNTRYNO	CNTRY	PERIOD	M_SHARE	DIST	PRICE
1	BELG	1	0.223	61	1.53
1	BELG	2	0.22	69	1.53
1	BELG	3	0.227	69	1.58
1	BELG	4	0.212	67	1.58
1	BELG	5	0.172	64	1.58
1	BELG	6	0.168	64	1.53
1	BELG	7	0.179	62	1.69
2	FRAN	1	0.038	11	0.98
2	FRAN	2	0.044	11	1.08
2	FRAN	3	0.039	9	1.13
2	FRAN	4	0.03	9	1.31
2	FRAN	5	0.036	14	1.36
2	FRAN	6	0.051	14	1.38
2	FRAN	7	0.044	9	1.34
3	UKIN	1	0.031	3	1.43
3	UKIN	2	0.038	3	1.43
3	UKIN	3	0.042	3	1.3
3	UKIN	4	0.037	3	1.43
3	UKIN	5	0.031	13	1.36
3	UKIN	6	0.031	14	1.49
3	UKIN	7	0.036	14	1.56

We first present an analysis that shows the matrix computations following precisely the equations presented in Sect. 2.4.4. These involve the same matrix manipulations in SAS as in the prior example, using the IML procedure in SAS. Then we present the MANOVA analysis proposed by SAS using the GLM procedure. The reader who wishes to skip the detailed calculations can go directly to the SAS GLM procedure that is illustrated in Fig. 2.8.

The SAS file that derived the computations for the test statistics is shown in Fig. 2.6.

The results are shown in the SAS output in Fig. 2.7.

These results indicate that the Bartlett's V statistic of 82.54 is larger than the critical chi-square with 6 degrees of freedom at the 0.05 confidence level $\left(\chi_{(df=6,\ \alpha=0.05)}^{2} = 12.59\right)$. Consequently, the hypothesis that the mean vectors are the same is rejected. The same conclusion can be derived from Rao's R statistic with its value of 55.10, which is larger than the corresponding F value with 6 and 32 degrees of freedom $\left(F_{\nu_2=32}^{\nu_1=6}(\alpha=0.05)=2.399\right)$.

The first lines of SAS commands in Fig. 2.8 read the data file in the same manner as in the prior examples. However, the code that follows is much simpler because the procedure automatically performs the MANOVA tests. For that analysis, the general procedure of the general linear model is called with the command "proc glm". The class statement indicates that the variable that follows (here CNTRY) is a discrete (nominal scaled) variable. This is the variable used to determine the K groups. K is calculated automatically according to the different values contained

```
****************** Examp2-2.sas ******************  */
OPTIONS LS=80;
DATA work;
INFILE
"C:\SAMD\CHAPTER2\EXAMPLES\Mkt_Dt_K.csv"
dlm = ',' firstobs=2;
INPUT CNTRYNO CNTRY $ PERIOD M_SHARE DIST PRICE;
data work;
        set work (drop = cntry period) ;
proc print;
proc freq;
tables cntryno / out = Nk_out (keep = count);
run;
/* Significance Test: K-Sample Problem */
proc iml;
reset center;
print " Multivariate Significance Test: K-Sample Problem " ;
print "-----------------------------------------------" ;
use work ;       /* Specifying the matrix with raw data */
read all var { CNTRYNO M_SHARE DIST PRICE}  into Mkt_Data;
use Nk_out;
read all var {count} into Nk_new;
                    /* Number of observations within each group            */
n_tot = nrow(Mkt_Data);
K=max(Mkt_Data[,1]);            /* Number of groups (samples) */
p=ncol(Mkt_Data)-1;             /* Number of variables */
print n_tot "    " K "    "  p;
start SSCP;                     /* SUBROUTINE for calculation of the SSCP matrix */
        n=nrow(x);
        mean=x[+,]/n;           /* Column means (mean vector) */
        x=x-repeat(mean,n,1);   /* Matrix of variances */
        SSCP = x`*x;            /* SSCP matrix */
print i "       "  mean;
finish SSCP;                    /* END SUBROUTINE */
S = J(p,p,0);                   /* Definition of a p x p square matrix with zeros */
do i = 1 to K;
if i = 1 then a = 1;
else
a=1+(i-1)*nk_new[i-1];
b=a+nk_new[i]-1;
x = Mkt_Data[a:b,2:4];
run SSCP;                       /* Execution of the SUBROUTINE for each group */
S = S + SSCP;                   /* Accumulation of the sum of SSCP matrices */
end;                            /* in order to calculate W (within-the-groups SSCP) */
W = S; DetW = Det(W);
print W "      " DetW;
x=Mkt_Data[,2:4];               /* Definition of the data matrix (dropping the first column:
CNTRYNO) */
run SSCP;                       /* Execution of the SUBROUTINE for total data */
T=SSCP;
DetT = Det(T);
print T "      " DetT;
Lmbd = Det(W) / Det(T);
m = n_tot-1-(p+K) / 2;
reset noname fw=5 nocenter;
print "Lambda =" Lmbd [format=10.6];
print "m =" m [format=2.0]
                "  Use Bartlett's V for large m's and Rao's R otherwise " ;
V = -m*Log(Lmbd);
s = sqrt((p*p*(K-1)**2-4)/(p*p+(K-1)**2-5));
R = (1-Lmbd**(1/s))*(m*s-p*(K-1)/2 + 1)/(Lmbd**(1/s)*p*(K-1));
Df_num = p*(K-1);                       Df_den = m*s-Df_num/2 +1 ;
Chi_crit = CINV(0.95,Df_num);                F_crit = finv(.95,df_num,df_den);
print "Bartlett's V = " V [format=9.6] "   DF =" DF_num [format=2.0]  ;
print "     Chi_crit =" Chi_crit [format=9.6];
print "Rao's R ="  R [format=9.6]
                    "   DF_NUM =" Df_num [format=2.0]
                    "   DF_DEN ="  Df_den [format=2.0] ;
print "     F_crit =" F_crit [format=9.6];
quit;
```

Fig. 2.6 SAS input to perform a test of difference in mean vectors across *K* groups (examp2-2.sas)

in the variable. On the left side of the equal sign, the model statement shows the list of the variates for which the means will be compared. On the right side is the group variable. The GLM procedure is in fact a regression where the dependent variable is regressed on the dummy variables that are automatically created by SAS (different dummy variables are created for each of the values of the grouping variable).

```
                Multivariate Significance Test: K-Sample Problem
               ---------------------------------------------------
                         N_TOT              K              P
                            21              3              3

                         I            MEAN
                         1       0.2001429 65.142857 1.5742857

                         I            MEAN
                         2       0.0402857        11 1.2257143

                         I            MEAN
                         3       0.0351429 7.5714286 1.4285714

                         W                                  DETW
                   0.0044351 0.2002857 -0.002814        0.246783
                   0.2002857 288.57143 1.8214286
                   -0.002814 1.8214286 0.2144286

                         I            MEAN
                         4       0.0918571 27.904762 1.4095238

                         T                                  DETT
                   0.1276486 42.601714 0.1808686       31.691145
                   42.601714  14889.81 63.809048
                   0.1808686 63.809048 0.6434952

Lambda =   0.007787

m = 17   Use Bartlett's V for large m's and Rao's R otherwise

Bartlett's V =  82.539814    DF =  6
     Chi_crit = 12.591587

Rao's R = 55.104665    DF_NUM =  6    DF_DEN = 32
     F_crit =  2.399080
```

Fig. 2.7 SAS output of test of difference across *K* groups (examp2-2.lst)

```
/*  ****************** Examp2-3-Manovasas.sas ******************  */

OPTIONS LS=80;
DATA work;
INFILE
""C:\SAMD\CHAPTER2\EXAMPLES\Mkt_Dt_K.csv"
dlm = ',' firstobs=2;
INPUT CNTRYNO CNTRY $ PERIOD M_SHARE DIST PRICE;


/* Chapter 2, IV.4 Significance Test: K-Sample Problem */
proc glm;
        class CNTRY;
        model M_SHARE DIST PRICE=CNTRY /nouni;
        manova h = CNTRY/ printe;
run;

quit;
```

Fig. 2.8 SAS input for MANOVA test of mean differences across *K* groups (examp2-3.sas)

The optional parameter "nouni" after the slash indicates that the univariate tests should not be performed (and consequently their corresponding output will not be shown). Finally, the last line of code is necessary to indicate that the MANOVA test concerns the differences across the grouping variable CNTRY.

The output shown in Fig. 2.9 provides the same information as shown in Fig. 2.7. Wilk's Λ has the same value of 0.007787. Several other tests are provided, and they

```
                              The GLM Procedure

                          Class Level Information

                    Class          Levels    Values

                    CNTRY               3    BELG FRAN UKIN


                    Number of Observations Read          21
                    Number of Observations Used          21

                       Multivariate Analysis of Variance

                            E = Error SSCP Matrix

                           M_SHARE              DIST             PRICE

         M_SHARE       0.0044351429       0.2002857143      -0.002814286
         DIST          0.2002857143       288.57142857       1.8214285714
         PRICE         -0.002814286       1.8214285714       0.2144285714


   Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

              DF = 18        M_SHARE           DIST          PRICE

              M_SHARE       1.000000       0.177039      -0.091258
                                             0.4684         0.7102

              DIST          0.177039       1.000000       0.231550
                              0.4684                        0.3402

              PRICE        -0.091258       0.231550       1.000000
                              0.7102         0.3402

           Characteristic Roots and Vectors of: E Inverse * H, where
                        H = Type III SSCP Matrix for CNTRY
                              E = Error SSCP Matrix

   Characteristic               Characteristic Vector  V'EV=1
           Root    Percent          M_SHARE            DIST           PRICE

     67.2013787      98.70        7.5885004       0.0457830       0.0045113
      0.8829099       1.30        3.7773797      -0.0204742       2.2231712
      0.0000000       0.00      -12.8623871       0.0361429       0.2847771


                  MANOVA Test Criteria and F Approximations for
                    the Hypothesis of No Overall CNTRY Effect
                        H = Type III SSCP Matrix for CNTRY
                              E = Error SSCP Matrix

                                S=2    M=0    N=7

Statistic                       Value    F Value    Num DF    Den DF    Pr > F

Wilks' Lambda                0.00778713      55.10         6        32    <.0001
Pillai's Trace               1.45424468      15.10         6        34    <.0001
Hotelling-Lawley Trace      68.08428858     176.86         6    19.652    <.0001
Roy's Greatest Root         67.20137868     380.81         3        17    <.0001

          NOTE: F Statistic for Roy's Greatest Root is an upper bound.
                  NOTE: F Statistic for Wilks' Lambda is exact.
```

Fig. 2.9 SAS output for MANOVA test of mean differences across K groups (examp2-3.lst)

all lead to the same conclusion that the differences in means are significant. In addition to the expression of Wilk's Λ as a function of the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$, three other measures are provided in the SAS output.

Pillai's trace is defined as $\sum_{i=1}^{K} \frac{\lambda_i}{1+\lambda_i}$.

```
insheet using
"/users/gatignon/documents/WORK_STATA/SAMD/Chapter2_MANOVA/Mkt_Dt_K.csv", clear
* manova test
manova m_share dist price = cntryno
mat list e(E)
mat list e(H_m)
mat list e(eigvals_m)
mat list e(aux_m)
```

Fig. 2.10 STATA input for MANOVA test of mean differences across K groups (examp2-3.do)

Hotelling–Lawley trace is simply the sum of the eigenvalues: $\sum_{i=1}^{K} \lambda_i$.

Roy's greatest root is the ratio $\frac{\lambda_{\max}}{1+\lambda_{\max}}$.

These tests tend to be consistent but the numbers are different. As noted in the SAS output, Roy's greatest root is an upper bound to the statistic.

Similar output is provided by STATA. Figure 2.10 shows the input for requesting MANOVA analysis in STATA.

Figure 2.11 presents the results of the analysis. It includes the within- and the between-SSCP matrices. The command "mat list e(E)" is used to print the within-SSCP matrix and "mat list e(H_m)" the between-SSCP matrix. The largest root is read from the eigenvector computed by "e(eigenvals_m)." Finally, the command "mat list e(aux_m)" lists the parameters m, s, and n that are used for the F values corresponding to the various statistics shown in the output. These parameters are defined as follows:

$$s = \min(K-1, p) \tag{2.38}$$

$$m = (|K-1-p| - 1)/2 \tag{2.39}$$

$$n = (N-K-p-1)/2 \tag{2.40}$$

where

N = total number of observations across groups;
K = number of groups;
p = number of variables.

For example, an approximate F statistic for Pillai's trace V with $s(2m + s + 1)$ and $s(2n + s + 1)$ degrees of freedom is

$$F = \frac{(2n+s+1)V}{(2m+s+1)(s-V)} \tag{2.41}$$

2.6 Assignment

In order to practice with these analyses, you will need to use the databases INDUP and PANEL described in Appendix C. These databases provide market share and marketing mix variables for a number of brands competing in five market segments. You can test the following hypotheses:

```
. insheet using
"/users/gatignon/documents/WORK_STATA/SAMD/Chapter2_MANOVA/Mkt_Dt_K.csv", clear
(6 vars, 21 obs)

. * manova test
. manova m_share dist price = cntryno

                           Number of obs =      21

                           W = Wilks' lambda      L = Lawley-Hotelling trace
                           P = Pillai's trace     R = Roy's largest root

      Source |  Statistic     df   F(df1,    df2) =   F   Prob>F
-------------+-----------------------------------------------------
     cntryno | W   0.0078      2     6.0    32.0    55.10 0.0000 e
             | P   1.4542            6.0    34.0    15.10 0.0000 a
             | L  68.0843            6.0    30.0   170.21 0.0000 a
             | R  67.2014            3.0    17.0   380.81 0.0000 u
             |-----------------------------------------------------
    Residual |                18
-------------+-----------------------------------------------------
       Total |                20
-------------------------------------------------------------------
                           e = exact, a = approximate, u = upper bound on F

. mat list e(E)

symmetric e(E)[3,3]
             m_share        dist       price
m_share   .00443514
   dist   .20028564   288.57143
  price  -.00281429   1.8214294   .21442857

. mat list e(H_m)

symmetric e(H_m)[3,3]
             m_share       dist      price
m_share  .12321343
   dist  42.401429  14601.238
  price  .18368288  61.987627  .42906667

. mat list e(eigvals_m)

e(eigvals_m)[1,2]
              c1          c2
r1  67.201384  .88290968

. mat list e(aux_m)

e(aux_m)[3,1]
    value
s       2
m       0
n       7
```

Fig. 2.11 STATA output for MANOVA test of mean differences across *K* groups (examp2-3.log)

1. The market behavioral responses of a given brand (e.g., awareness, perceptions, or purchase intentions) are different across segments.
2. The marketing strategy (i.e., the values of the marketing mix variables) of selected brands is different (perhaps corresponding to different strategic groups).

Figure 2.12 shows how to read the data within an SAS file and how to create new files with a subset of the data saved in a format that can be read easily using the examples provided throughout this chapter. Using the model described in the examples above, adapt these examples to the database to perform tests of differences across groups.

The commands to merge the INDUP and PANEL data sets in STATA are shown in Fig. 2.13.

```
/*************************************************************************
 Assign2.sas
 Creation of additional data files for Chapter2 assignments.
*************************************************************************/
option ls=120 ;
/*-------------------------------------------------------------------------
 Creating the dataset PANEL by reading data from c:\...\panel.csv
-------------------------------------------------------------------------*/
data panel;
  infile 'C:\SAMD\Chapter2\Assignments\panel.csv' firstobs=2   dlm = ',' ;
  input period    segment  segsize  ideal1-ideal3
         brand $  adv_pct   aware    intent shop1-shop3
         perc1-perc3 dev1-dev3 share ;
run;
proc sort data=panel;
        by period brand;
run;
/*-------------------------------------------------------------------------
 Creating the dataset INDUP by reading data from c:\...\indup.csv
-------------------------------------------------------------------------*/
data indup;
  infile 'C:\SAMD\Chapter2\Assignments\indup.csv' firstobs=2   dlm = ',' ;
  input  period  firm  brand $ price  advert
         char1-char5 salmen1-salmen3
         cost dist1-dist3 usales  dsales  ushare dshare  adshare  relprice ;
run;
proc sort data =indup;
        by period brand;
run;
/*-------------------------------------------------------------------------
 Merging PANEL and INDUP into ECON
-------------------------------------------------------------------------*/
data econ;
        merge panel indup;
        by period brand;
if segment<5 then delete;
run;
proc means noprint;
var intent share ;
output out = econmean mean=IntMean ShrMean;
run;
/*-------------------------------------------------------------------------
 Writing EconMean to a CSV file (easily opened by Excel)
-------------------------------------------------------------------------*/
data _NULL_;
  set EconMean (keep = IntMean ShrMean);
  by IntMean ;
  TAB = ',' ;
  FN = "C:\SAMD\CHAPTER2\ASSIGNMENTS\Mean1grp.CSV";
  file PLOTFILE filevar=FN;
  if ( FIRST.IntMean ) then
  do;
    put "IntMean" TAB "ShrMean" ;
  end;
    put  IntMean  TAB  ShrMean  ;
run;
/*-------------------------------------------------------------------------
 Creating a new dataset EconNew with selected variables from ECON
-------------------------------------------------------------------------*/
data EconNew;
set  Econ   ;
keep segment period brand intent share ;
where brand = 'salt';
run;
proc sort ;
by Brand Segment Period ;
run;
/*-------------------------------------------------------------------------
 Writing EconNew to a CSV file (easily opened by Excel)
-------------------------------------------------------------------------*/
data _NULL_;
  set EconNew;
  by BRAND Segment ;
  TAB = ',' ;
  FN = "C:\SAMD\CHAPTER2\ASSIGNMENTS\DatKgrp.CSV";
  file PLOTFILE filevar=FN;
  if ( FIRST.Brand ) then
  do;

      put "SEGMENT" TAB "BRAND" TAB "PERIOD" TAB "INTENT" TAB "SHARE" ;
    end;
      put  SEGMENT  TAB  BRAND  TAB   PERIOD TAB  Intent  TAB  Share  ;
  run;
```

Fig. 2.12 Example of SAS file for reading data sets INDUP and PANEL and creating new data files (assign2.sas)

```
insheet using "/users/fblgatignon/Documents/WORK_STATA/SAMD/panel.csv", clear
merge m:m period brand using "/users/fblgatignon/Documents/WORK_STATA/SAMD/indup.dta"
keep if segment ==5
drop if period ==0
regress awareness adshare
manova dolshare adshare relprice = firm
```

Fig. 2.13 Example of STATA file for reading and merging data sets INDUP and PANEL (MergeIndup_Panel_Mac.do)

Bibliography

Basic Technical Readings

Tatsuoka, M. M. (1971). *Multivariate analysis: techniques for educational and psychological research*. New York, NY: Wiley.

Application Readings

Cool, K., & Dierickx, I. (1993). Rivalry, strategic groups and firm profitability. *Strategic Management Journal, 14*, 47–59.

Kilduff, M., Angelmar, R., & Mehra, A. (2000). Top management-team diversity and firm performance: Examining the role of cognitions. *Organization Science, 11*(1), 21–34.

Long, R. G., Bowers, W. P., Barnett, T., et al. (1998). Research productivity of graduates in management: Effects of academic origin and academic affiliation. *Academy of Management Journal, 41*(6), 704–771.

Chapter 3
Reliability Alpha, Principal Component Analysis, and Exploratory Factor Analysis

In this chapter, we discuss the issues involved in building measures or scales. We focus the chapter on two types of analyses: (1) the measurement of reliability with Cronbach's alpha and (2) the verification of unidimensionality using factor analysis. We concentrate on exploratory factor analysis (EFA) and we only introduce the notion of confirmatory factor analysis. In the next chapter, we develop in detail the confirmatory factor analytic model and examine the measures of convergent and discriminant validity.

3.1 Notions of Measurement Theory

3.1.1 Definition of a Measure

If T is the true score of a construct and e represents the error associated with the measurement, the measure X is expressed as

$$X = T + e \tag{3.1}$$

This relationship can be represented graphically as in Fig. 3.1 where the observed variable or measure is shown in a box and the unobserved true score or construct is distinguished by a circle. The measurement error term is represented by the letter e. The directions of the arrows represent the "causal" directionality of the relationships. The heads of both arrows point towards the measure X because both the true construct and the measurement error are determinants of what is being observed.

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_3, © Springer Science+Business Media New York 2014

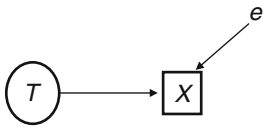


Fig. 3.1 Representation of simple measurement model

In addition, we assume that $E[e] = 0$ and $\text{Cov}[e,T] = 0$.

3.1.2 Parallel Measurements

Measures Y_1 and Y_2 are parallel if they meet the following characteristics:

$$Y_1 = T + e_1 \tag{3.2}$$

$$Y_2 = T + e_2 \tag{3.3}$$

$$\text{E}[e_1] = \text{E}[e_2] = 0 \tag{3.4}$$

$$\text{V}[e_1] = \text{V}[e_2] = \sigma_e^2 \tag{3.5}$$

$$\rho(e_1, e_2) = 0 \tag{3.6}$$

3.1.3 Reliability

The reliability of a measure is the squared correlation between the measure and the true score: $\rho^2(X,T)$, also noted as ${\rho_{XT}}^2$. It is also the ratio of the true score variance to the measured variance:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \tag{3.7}$$

This can be demonstrated as follows:

$$\begin{aligned}
\sigma(X,T) &= \text{E}[(X - \text{E}[X])(T - \text{E}[T])] \\
&= \text{E}[XT - \text{E}[X]T + \text{E}[X]\text{E}[T] - X\text{E}[T]] \\
&= \text{E}[XT] - \text{E}[X]\text{E}[T] + \text{E}[X]\text{E}[T] - \text{E}[X]\text{E}[T] \\
&= \text{E}[XT] - \text{E}[X]\text{E}[T] \\
&= \text{E}[(T+e)T] - \text{E}[T+e]\text{E}[T] \\
&= \text{E}\left[T^2 + eT\right] - (\text{E}[T])^2 \\
&= \text{E}\left[T^2\right] - (\text{E}[T])^2 \\
&= \text{E}\left[(T - \text{E}[T])^2\right]
\end{aligned} \tag{3.8}$$

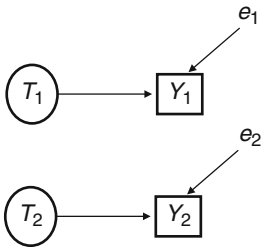


Fig. 3.2 A graphical representation of measures

This last equality can be shown as follows:

$$(T - \mathrm{E}[T])^2 = T^2 + (\mathrm{E}[T])^2 - 2T\mathrm{E}[T] \tag{3.9}$$

$$= T^2 + (\mathrm{E}[T])^2 - 2(\mathrm{E}[T])^2 \tag{3.10}$$

$$= T^2 - (\mathrm{E}[T])^2 \tag{3.11}$$

but $\mathrm{E}[(T - \mathrm{E}[T])^2] = {\sigma_T}^2$, which is the numerator of the reliability expression.

Let us now express the correlation between the true score and the measure:

$$\rho_{XT} = \frac{\sigma(X,T)}{\sigma(X)\sigma(T)} = \frac{\sigma_T^2}{\sigma_X \sigma_T} = \frac{\sigma_T}{\sigma_X} \tag{3.12}$$

$$\Rightarrow \rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \tag{3.13}$$

Therefore, the reliability can be expressed as the proportion of the observed score variance that is the true score variance. The problem with the definition and formulae above is that the variance of the true score is not known since the true score is not observed. This explains the need to use multiple measures and to form scales.

3.1.4 Composite Scales

A composite scale is built from using multiple items or components measuring the constructs. This can be represented graphically as in Fig. 3.2. Note that by convention, circles represent unobserved constructs and squares identify observable variables or measures.

The unweighted composite scale is the sum of the two items:

$$X = Y_1 + Y_2 \tag{3.14}$$

3.1.4.1 Reliability of a Two-Component Scale

In this section, we show that the reliability of a composite scale has a lower bound, which is the coefficient alpha. The two components of the scale are

$$Y_1 = T_1 + e_1 \tag{3.15}$$

$$Y_2 = T_2 + e_2 \tag{3.16}$$

The composite scale corresponds to a formative index:

$$\begin{aligned} X = Y_1 + Y_2 &= T_1 + T_2 + e_1 + e_2 \\ &= \underbrace{\quad}_{T} + \underbrace{\quad}_{e} \end{aligned} \tag{3.17}$$

Although, a priori, T_1 and T_2 appear as different true scores, we will see that they must be positively correlated, and we will show the impact of that correlation on the reliability of the scale. As a consequence, it is best to think of these scores as corresponding to different items of a single construct.

Computation of Coefficient α

From Eq. (3.17), the composite scale is defined as

$$X = Y_1 + Y_2 \tag{3.18}$$

$$T = T_1 + T_2 \tag{3.19}$$

$$\sigma_T^2 = \sigma^2(T_1) + \sigma^2(T_2) + 2\sigma(T_1, T_2) \tag{3.20}$$

However, because

$$[\sigma(T_1) - \sigma(T_2)]^2 \geq 0 \tag{3.21}$$

(equality if the test is parallel), then it follows that

$$\sigma^2(T_1) + \sigma^2(T_2) \geq 2\sigma(T_1, T_2) \tag{3.22}$$

This last inequality results from developing the left side of the inequality in Eq. (3.21):

$$[\sigma(T_1) - \sigma(T_2)]^2 = [\sigma(T_1)]^2 + [\sigma(T_2)]^2 - 2[\sigma(T_1)\sigma(T_2)] \tag{3.23}$$

Given a positive correlation between T_1 and T_2 and $\rho(T_1,T_2) < 1$,

$$\sigma(T_1, T_2) = \rho(T_1, T_2)\sigma(T_1)\sigma(T_2) \leq \sigma(T_1)\sigma(T_2) \tag{3.24}$$

It follows that

$$[\sigma(T_1)]^2 + [\sigma(T_2)]^2 - 2[\sigma(T_1)\sigma(T_2)] \leq [\sigma(T_1)]^2 + [\sigma(T_2)]^2 - 2[\sigma(T_1, T_2)] \quad (3.25)$$

The left side of the inequality above being positive, a fortiori, the right side is also positive. This is the conclusion in Eq. (3.22).

It should be noted that this property is only interesting for cases where the items (components) are positively correlated. Indeed, in the case of a negative correlation, the inequality is dominated by the fact that the left side is greater or equal to zero.

Therefore, in cases of positively correlated items, bringing together Eqs. (3.20) and (3.22) leads to

$$\sigma_T^2 \geq 4\sigma(T_1, T_2) \quad (3.26)$$

Consequently, the reliability has a lower bound, which is given by

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \geq \frac{4\sigma(T_1, T_2)}{\sigma_X^2} \quad (3.27)$$

But

$$\begin{aligned} \sigma(Y_1, Y_2) &= \mathrm{E}[(T_1 + e_1)(T_2 + e_2)] \\ &= \mathrm{E}[T_1 T_2] \\ &= \sigma(T_1, T_2) \end{aligned} \quad (3.28)$$

Therefore,

$$\rho_{XT}^2 \geq \frac{4\sigma(Y_1, Y_2)}{\sigma_X^2} \quad (3.29)$$

Since

$$\sigma_X^2 = \mathrm{E}\left[(Y_1 + Y_2)^2\right] = \mathrm{E}\left[Y_1^2\right] + \mathrm{E}\left[Y_2^2\right] + \mathrm{E}[2Y_1 Y_2] \quad (3.30)$$

$$= \sigma^2(Y_1) + \sigma^2(Y_2) + 2\sigma(Y_1, Y_2) \quad (3.31)$$

it follows that

$$2\sigma(Y_1, Y_2) = \sigma_X^2 - \sigma^2(Y_1) - \sigma^2(Y_2) \quad (3.32)$$

and, therefore,

$$\rho_{XT}^2 \geq 2\left[\frac{\sigma_X^2 - \sigma^2(Y_1) - \sigma^2(Y_2)}{\sigma_X^2}\right] = 2\left[1 - \frac{\sigma^2(Y_1) + \sigma^2(Y_2)}{\sigma_X^2}\right] \quad (3.33)$$

This demonstrates that there is a lower bound to the reliability. If this lower bound is high enough, it means that the actual reliability is even higher, and therefore the scale is reliable. It is also clear from Eq. (3.33) that as the (positive) correlation between the two items or components increases, the portion that is subtracted from one decreases so that coefficient alpha increases. If the correlation is zero, then coefficient alpha is zero.

3.1.4.2 Generalization to Composite Measurement with K Components

For a scale formed from K components or items,

$$X = \sum_{k=1}^{K} Y_k \tag{3.34}$$

The reliability coefficient alpha is a generalized form of the above calculation:

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum_{k=1}^{K} \sigma^2(Y_k)}{\sigma_X^2}\right] \tag{3.35}$$

α is a lower bound estimate of the reliability of the composite scale X that is of ${\rho_{XT}}^2$.

3.2 Exploratory Factor Analysis

Factor analysis can be viewed as a method to discover or confirm the structure of a covariance matrix. However, in the case of EFA, the analysis attempts to discover the underlying unobserved factor structure. In the case of confirmatory factor analysis, a measurement model is specified and tested against the observed covariance matrix.

EFA is a special type of rotation. Consequently, rotations are first reviewed in the general context of space geometry.

3.2.1 Axis Rotation

Let us consider Fig. 3.3, which shows a set of orthogonal axes X_1 and X_2. The vector Y_1 shows an angle θ relative to X_1. Similarly, the vector Y_2 forms an angle θ with X_2.

The problem consists in expressing the transformation that occurs when going from the coordinates in the original axes to the new axes. The derivation of such a transformation can be explained with a more detailed representation as in Fig. 3.4.

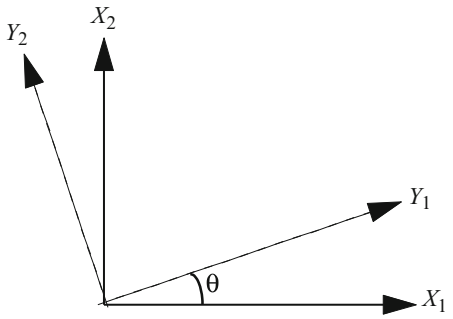


Fig. 3.3 Axis rotation

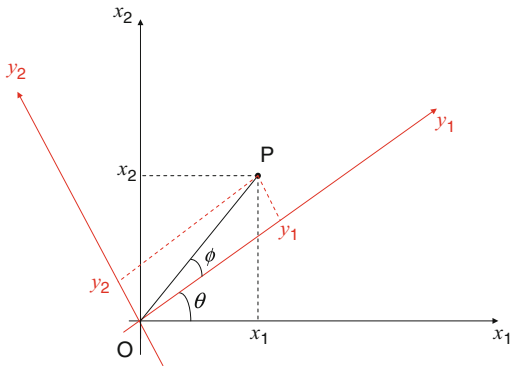


Fig. 3.4 The geometry of axis rotation

Let us define OP = r.
Applying the basic definitions of sines and cosines, we have

$$\cos\phi = \frac{y_1}{r} \tag{3.36}$$

and

$$\sin\phi = \frac{y_2}{r} \tag{3.37}$$

It follows that

$$y_1 = r \cdot \cos\phi \tag{3.38}$$

and

$$y_2 = r \cdot \sin\phi \tag{3.39}$$

Furthermore,

$$\cos(\phi + \theta) = \frac{x_1}{r} \tag{3.40}$$

and

$$\sin(\phi + \theta) = \frac{x_2}{r} \tag{3.41}$$

which leads to

$$x_1 = r \cdot \cos(\phi + \theta) \tag{3.42}$$

and

$$x_2 = r \cdot \sin(\phi + \theta) \tag{3.43}$$

Using the trigonometric rule that $\cos(\alpha + \beta) = \cos\alpha \cdot \cos\beta - \sin\alpha \cdot \sin\beta$,

$$\begin{aligned} x_1 &= r \cdot \cos(\phi + \theta) = r(\cos\phi\cos\theta - \sin\phi\sin\theta) \\ &= r\cos\phi\cos\theta - r\sin\phi\sin\theta \end{aligned} \tag{3.44}$$

However, using Eqs. (3.38) and (3.39), Eq. (3.44) becomes

$$x_1 = y_1\cos\theta - y_2\sin\theta \tag{3.45}$$

Similarly, using the rule on the sine of the sum of two angles, i.e.,

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$$

equation (3.43) becomes

$$\begin{aligned} x_2 &= r \cdot \sin(\phi + \theta) = r(\sin\phi\cos\theta + \cos\phi\sin\theta) \\ &= r\sin\phi\cos\theta + r\cos\phi\sin\theta \end{aligned} \tag{3.46}$$

which, again using Eqs. (3.38) and (3.39), leads to

$$x_2 = y_2\cos\theta + y_1\sin\theta \tag{3.47}$$

Equations (3.45) and (3.47) form a system of two equations with two unknowns.

To solve that system, let us multiply the right and left sides of Eq. (3.45) by $\cos\theta$ and both sides of Eq. (3.47) by $\sin\theta$. This leads to the system of equations

$$\begin{cases} x_1\cos\theta = y_1\cos^2\theta - y_2\sin\theta\cos\theta \\ x_2\sin\theta = y_1\sin^2\theta - y_2\sin\theta\cos\theta \end{cases} \tag{3.48}$$

Taking the sum of each side of the two equations, this gives

$$x_1\cos\theta + x_2\sin\theta = y_1\left(\cos^2\theta + \sin^2\theta\right) \tag{3.49}$$

However, because $\cos^2\alpha + \sin^2\alpha = 1$, it follows that Eq. (3.49) is simply

$$y_1 = x_1 \cos\theta + x_2 \sin\theta \tag{3.50}$$

We apply the same procedure to derive y_2.

Let us multiply the right and left sides of Eq. (3.45) by $\sin\theta$ and both sides of Eq. (3.47) by $(-\cos\theta)$. This leads to the system of equations

$$\begin{cases} x_1 \sin\theta = y_1 \sin\theta\cos\theta - y_2 \sin^2\theta \\ -x_2 \cos\theta = -y_1 \sin\theta\cos\theta - y_2 \cos^2\theta \end{cases} \tag{3.51}$$

Taking the sum of each side of the equations leads to

$$x_1 \sin\theta - x_2 \cos\theta = -y_2\left(\sin^2\theta + \cos^2\theta\right) \tag{3.52}$$

This is more simply

$$y_2 = -x_1 \sin\theta + x_2 \cos\theta \tag{3.53}$$

Therefore, Eqs. (3.50) and (3.53) provide the formulae for a rotation transformation of axes, which in matrix notation gives

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{3.54}$$

The weights represented in Eq. (3.54) correspond, therefore, to an orthogonal rotation and the constraints of orthogonality are respected. Indeed,

$$(\cos\theta)^2 + (\sin\theta)^2 = 1 \tag{3.55}$$

$$(-\sin\theta)^2 + (\cos\theta)^2 = 1 \tag{3.56}$$

$$(\cos\theta)(-\sin\theta) + (\sin\theta)(\cos\theta) = 0 \tag{3.57}$$

These constraints can be expressed in matrix notations as

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.58}$$

This corresponds to the constraint expressed more generally in Eq. (3.60).

Therefore, the rotation corresponds to a special linear transformation of $\mathbf{x}$ to $\mathbf{y}$. If $\mathbf{x}$ is a p-dimensional vector and $\mathbf{V}$ is a square matrix of size p by p (which represents the linear weights applied to vector $\mathbf{x}$), then $\mathbf{y}$, the linear transformation of $\mathbf{x}$, is also with dimension p. However, orthogonality conditions must be met so that $\mathbf{V}$ cannot be just any matrix. Therefore, the rotation can be expressed in the equations

$$\underset{px1}{\mathbf{y}} = \underset{pxp}{\mathbf{V}'} \underset{px1}{\mathbf{x}} \tag{3.59}$$

$$\text{s.t.} \mathbf{V}'\mathbf{V} = \mathbf{I} \tag{3.60}$$

so that conditions for orthogonal rotation are met.

3.2.2 *Variance-Maximizing Rotations (Eigenvalues and Eigenvectors)*

The advantage of an orthogonal rotation is that it enables the same points to be represented in a space using different axes but without affecting the covariance matrix, which remains unchanged. The idea is to find a specific rotation or linear transformation that will maximize the variance of the linear transformations.

3.2.2.1 The Objective

The objective is, therefore, to find the linear transformation of a vector that maximizes the variance of the transformed variable (of the linear combination), i.e., to find the weights v′ such that if for one observation (assumed to be mean centered) the transformation is

$$\underset{1\times1}{y_i} = \underset{1\times p}{\mathbf{v}'} \underset{p\times1}{\mathbf{x}_i}$$

and for all N observations

$$\underset{1\times N}{\mathbf{y}'} = \underset{1\times p}{\mathbf{v}'} \underset{p\times N}{\mathbf{X}'}$$

then the variance of the transformed variable which is proportional to

$$\underset{1\times1}{\mathbf{y}'\mathbf{y}} = \sum_{i=1}^{N} y_i^2 = \mathbf{v}'\mathbf{X}'\mathbf{X}\mathbf{v} = \underset{1\times p}{\mathbf{v}'} \underset{p\times p}{\mathbf{S}} \underset{p\times1}{\mathbf{v}}$$

is maximized.

In other words, the problem is

$$\text{Find } \mathbf{V} \,\big|\, \text{Max.} \mathbf{y}'\mathbf{y} \tag{3.61}$$

$$\text{s.t.} \quad \underset{1\times1}{\mathbf{v}'\mathbf{v}} = \sum_{j=1}^{p} v_j^2 = 1 \tag{3.62}$$

By replacing $\mathbf{y}$ with its expression as a linear combination of $\mathbf{X}$, the problem becomes equivalent to

$$\text{Max } \mathbf{v}'\mathbf{S}\mathbf{v} \tag{3.63}$$

$$\text{s.t.} \quad \mathbf{v}'\mathbf{v} = 1 \tag{3.64}$$

This can be resolved by maximizing the Lagrangian $\mathbf{L}$:

$$\text{Max} \quad \mathbf{L} = \mathbf{v}'\mathbf{S}\mathbf{v} - \lambda\left(\mathbf{v}'\mathbf{v} - 1\right) \tag{3.65}$$

Using the derivative rule $\partial\ \mathbf{x}'\mathbf{A}\mathbf{x}/\partial\ \mathbf{x} = 2\mathbf{A}\mathbf{x}$

$$\frac{\partial \mathbf{L}}{\partial \mathbf{v}} = 2\mathbf{S}\mathbf{v} - 2\lambda\mathbf{v} = 0 \tag{3.66}$$

$$= \underset{p\times p}{\left(\mathbf{S} - \lambda\mathbf{I}\right)}\ \underset{p\times 1}{\mathbf{v}} = \underset{p\times 1}{0} \tag{3.67}$$

Solving these equations provides the eigenvalues and eigenvectors. First we show how to derive the eigenvalues. Then, we will proceed with the calculation of the eigenvectors.

Finding the Eigenvalues

We need to resolve the following system of equations for $\mathbf{v}$ and λ:

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{v} = 0 \tag{3.68}$$

A trivial solution is $\mathbf{v} = 0$. Pre-multiplying by $(\mathbf{S} - \lambda\mathbf{I})^{-1}$

$$\mathbf{v} = (\mathbf{S} - \lambda\mathbf{I})^{-1}0 = 0 \tag{3.69}$$

This also implies that, for a nontrivial solution to exist, $(\mathbf{S} - \lambda\mathbf{I})$ must not have an inverse because if it does, $\mathbf{v} = 0$ and it gives a trivial solution.

Therefore, a condition for a nontrivial solution to Eq. (3.68) to exist is that the determinant is zero because the operation shown in Eq. (3.69) cannot then be performed:

$$|\mathbf{S} - \lambda\mathbf{I}| = 0 \tag{3.70}$$

Equation (3.70) results in a polynomial in λ of degree p which therefore has p roots. Following is an example. Let us assume that the covariance matrix is

$$\mathbf{S} = \begin{bmatrix} 16.81 & .88 \\ .88 & 6.64 \end{bmatrix}$$

Then

$$|\mathbf{S} - \lambda\mathbf{I}| = \begin{vmatrix} 16.81 - \lambda & .88 \\ .88 & 6.64 - \lambda \end{vmatrix} = \lambda^2 - 23.45\lambda + 110.844 = 0 \tag{3.71}$$

Resolving this second-degree equation gives the two roots:

$$\begin{cases} \lambda_1 = 16.8856 \\ \lambda_2 = 6.5644 \end{cases} \tag{3.72}$$

They are the eigenvalues.

Finding the Eigenvectors

Knowing the eigenvalues, the eigenvectors can now be easily computed. For each eigenvalue, there are p equations with p unknowns

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{v} = 0 \tag{3.73}$$

subject to normality, i.e., $\mathbf{v}'\mathbf{v} = 1$.

The p unknowns are then straightforward to estimate.

3.2.2.2 Properties of Eigenvalues and Eigenvectors

Two properties of eigenvectors and eigenvalues are indispensable in order to understand the implications of this rotation:

1. $\mathbf{V}'\mathbf{V} = \mathbf{I}$, and therefore : $\mathbf{V}' = \mathbf{V}^{-1}$ (3.74)
2. $\mathbf{V}'\mathbf{S}\mathbf{V} = \mathbf{\Lambda}$, where $\mathbf{\Lambda}_{p\times p} = \text{diag}\{\lambda_i\}$ (3.75)

It is important to understand the proof of this last property because it shows how the covariance matrix can be reconstituted with the knowledge of eigenvectors and eigenvalues.

From the first-order derivative of the Lagrangian ($\partial\ \mathbf{L}/\partial\ \mathbf{v} = 2\mathbf{S}\mathbf{v} - s\lambda\mathbf{v} = 0$), and putting all eigenvectors together

$$\underset{p\times p}{\mathbf{S}}\ \underset{p\times p}{\mathbf{V}} = \underset{p\times p}{\mathbf{V}}\ \underset{p\times p}{\mathbf{\Lambda}} \tag{3.76}$$

Pre-multiplying each side by $\mathbf{V}'$ gives

$$\mathbf{V}'\mathbf{S}\mathbf{V} = \underbrace{\mathbf{V}'\mathbf{V}}_{\mathbf{I}}\Lambda = \Lambda \tag{3.77}$$

Furthermore, a third property is that the eigenvalue is the variance of the linearly transformed variable $\mathbf{y}$. From Eq. (3.73), pre-multiplying the left side by $\mathbf{v}'$, one obtains for eigenvalue i and eigenvector i

$$\mathbf{v}'_i(\mathbf{S} - \lambda_i\mathbf{I})\mathbf{v}_i = 0 \tag{3.78}$$

or

$$\mathbf{v}'_i\mathbf{S}\mathbf{v}_i = \lambda_i\mathbf{v}'_i\mathbf{v}_i \tag{3.79}$$

However, the left side of Eq. (3.79) is the variance of the transformed variable y_i:

$$\mathbf{v}'_i\mathbf{S}\mathbf{v}_i = \mathbf{v}'_i\mathbf{X}'\mathbf{X}\mathbf{v}_i = \mathbf{y}'_i\mathbf{y}_i = \lambda_i \tag{3.80}$$

Therefore, the eigenvalue represents the variance of the new variable formed as a linear combination of the original variables.

In addition, considering the equality $\mathbf{\Lambda} = \mathbf{V}'\mathbf{S}\mathbf{V}$ in Eq. (3.77)

$$tr(\mathbf{\Lambda}) = tr\left(\mathbf{V}'\mathbf{S}\mathbf{V}\right) = tr\left(\mathbf{V}'\mathbf{V}\mathbf{S}\right) = tr(\mathbf{S}) \tag{3.81}$$

This means that the total variance in $\mathbf{X}$ as measured by the sum of the variances of all the $\mathbf{x}$s is equal to the sum of the eigenvalues.

It should be clear that if the variables $\mathbf{x}$ are normalized, the $\mathbf{S}$ matrix is the correlation matrix $\mathbf{R}$. The trace of $\mathbf{R}$ (i.e., the sum of the diagonal terms) is equal to the number of variables p. It then follows from the equality in Eq. (3.81) that the sum of the eigenvalues of a correlation matrix is equal to the number of variables p.

Furthermore, considering only the rth largest values of the eigenvalues, these first r linear combinations explain a percentage of the total variance in $\mathbf{X}$. This percentage is

$$\frac{\sum_{k=1}^{r}\lambda_k}{\sum_{k=1}^{p}\lambda_k} \times 100 \tag{3.82}$$

3.2.3 *Principal Component Analysis*

The problem in principal component analysis (PCA) is just what has been described in the prior section. It consists in finding the linear combination that maximizes the variance of the linear combinations of a set of variables (the first linear combination, then the second given that it should be perpendicular to the first, etc.) and reconstituting the covariance matrix $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$. Therefore, the problem is identical to finding the eigenvalues and eigenvectors of the covariance matrix.

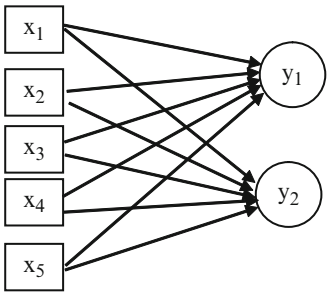


Fig. 3.5 A graphical representation of the principal component model

3.2.3.1 PCA: A Data Reduction Method

In PCA, new variables **y** are constructed as exact linear combinations of the original variables. This is represented graphically in Fig. 3.5, using the same convention for the representation of observed and unobserved variables with boxes and circles, respectively.

Furthermore, it is a data reduction method in the sense that the covariance matrix can be approximated with a number of dimensions smaller than p, the number of original variables. Indeed, from Eq. (3.77)

$$\mathbf{V}\mathbf{V}'\mathbf{S}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda} \tag{3.83}$$

$$\mathbf{S}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda} \tag{3.84}$$

$$\mathbf{S}\mathbf{V}\mathbf{V}' = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}' \tag{3.85}$$

$$\mathbf{S} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}' \tag{3.86}$$

Let **V*** include the eigenvectors corresponding to the r largest eigenvalues and **Λ*** include the r largest eigenvalues:

$$\underset{p\times p}{\mathbf{S}^*} = \underset{p\times r}{\mathbf{V}^*}\ \underset{r\times r}{\boldsymbol{\Lambda}^*}\ \underset{r\times p}{\mathbf{V}^{*'}} \tag{3.87}$$

Therefore, it can be seen from Eq. (3.87) that replacing the small eigenvalues by zero should not affect the ability to reconstitute the variance–covariance matrix **S** (**S*** should approximate **S**). Consequently, r data points are needed for each i instead of the original p variables.

3.2.3.2 Principal Component Loadings

The correlation between a single variable $\mathbf{x}_i$ and the composite variable $\mathbf{y}_k$ corresponding to the k's eigenvalue is called a loading. Let us consider the normalized data matrix $\underset{N\times p}{\widetilde{\mathbf{X}}}$. The principal component variables **Y** are such that

$$\underset{N\times p}{\mathbf{Y}} = \underset{N\times p}{\widetilde{\mathbf{X}}}\ \underset{p\times p}{\mathbf{V}} \tag{3.88}$$

where the weights $\mathbf{V}$ are the eigenvectors such that

$$\mathbf{R} = \frac{1}{N}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \mathbf{V}\Lambda\mathbf{V}' \tag{3.89}$$

The cross products of $\mathbf{Y}$ are given by

$$\frac{1}{N}\mathbf{Y}'\mathbf{Y} = \frac{1}{N}\mathbf{V}'\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{V} = \mathbf{V}'\mathbf{R}\mathbf{V} = \mathbf{V}'\mathbf{V}\Lambda\mathbf{V}'\mathbf{V} = \Lambda \tag{3.90}$$

Consequently, $\mathbf{Y}$ is normalized by post-multiplying $\mathbf{Y}$ by $\mathbf{\Lambda}^{-\frac{1}{2}}$. Let us write the normalized $\mathbf{Y}$s as

$$\widetilde{\mathbf{Y}} = \mathbf{Y}\Lambda^{-\frac{1}{2}} \tag{3.91}$$

The correlation between X and Y is

$$\underset{p\times p}{Cor(\mathbf{X},\mathbf{Y})} = \frac{1}{N}\underset{p\times N}{\widetilde{\mathbf{X}}'}\ \underset{N\times p}{\widetilde{\mathbf{Y}}} = \frac{1}{N}\widetilde{\mathbf{X}}'\mathbf{Y}\Lambda^{-\frac{1}{2}} = \frac{1}{N}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{V}\Lambda^{-\frac{1}{2}} \tag{3.92}$$

$$= \mathbf{R}\mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'\mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}} \tag{3.93}$$

Consequently, the loadings are given by

$$\underset{p\times p}{\mathbf{L}} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}} \tag{3.94}$$

3.2.3.3 PCA Versus Exploratory Factor Analysis

Two points can be made that distinguish PCA from factor analysis:

1. The new variables $\mathbf{y}$ are determined exactly by the p $\mathbf{x}$ variables. There is no noise introduced and, therefore, no measurement error as discussed in Sect. 3.1 on measurement theory is represented. Factor analysis introduces this notion of measurement error.
2. The new unobserved variables $\mathbf{y}$ are built by putting together the original p variables. Therefore, $\mathbf{y}$ is constructed from the original $\mathbf{x}$ variables in an index. This is represented graphically in Fig. 3.5. As opposed to this formative index, in factor analysis the observed $\mathbf{x}$ variables are reflections of the various unobserved variables or constructs.

This last distinction between reflective indicators and constitutive indices is developed in the next section.

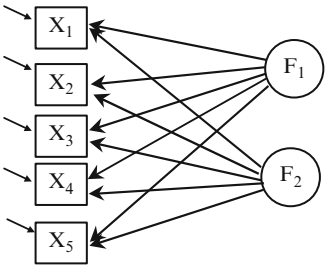


Fig. 3.6 A graphical representation of the exploratory factor analytic model

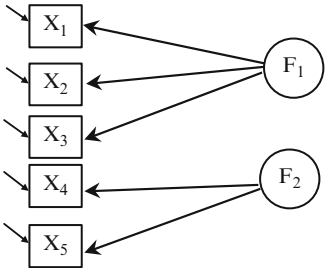


Fig. 3.7 A graphical representation of multiple measures with a confirmatory factor structure

3.2.4 Exploratory Factor Analysis

Now that we have explained the difference between PCA and factor analysis, we need to distinguish between two different types of factor analysis: EFA and confirmatory factor analysis. The basic difference lies in the fact that in confirmatory factor analysis, a structure is proposed in which the observed, measurable variables reflect only specific unobserved constructs while exploratory factor analysis allows all measurable variables to reflect each factor (where reflection implies a causal direction from the construct to the measure). These two types of factor analysis can easily be distinguished by the differences in their graphical representation. We examine the differences analytically in this chapter and the next.

EFA is graphically represented in Fig. 3.6 in an example with two unobserved constructs and five observed variables or measures.

The unobserved constructs are represented with circles while the measures are represented by squares. The arrows on the left side coming into the measured variable boxes indicate the random measurement errors.

Although the fundamental difference between the exploratory factor analytic model and the confirmatory factor analytic model is presented in the next chapter, it can be helpful to compare these models here. The basic distinction is that, in confirmatory factor analysis, only some measures reflect specific, individual unobserved constructs, as shown in Fig. 3.7.

EFA can be characterized by the fact that it is data driven, as opposed to confirmatory analysis, which represents a strong theory of measurement. The purpose of EFA is, in fact, to find or discover patterns that may help understand

the nature of the unobserved variables. Consequently, it is a method that, based on patterns of correlations among variables, inductively brings insights into the underlying factors. Considering Fig. 3.5, the weights assigned to each arrow linking each factor to each observed variable indicate the extent to which each variable reflects each factor. This can be shown analytically.

3.2.4.1 The Exploratory Factor Analysis Model

As discussed above, each observed variable is a function of all the factors underlying the structure. These variables also contain a measurement error term. For example, for two observed variables and two factors

$$X_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \varepsilon_1 \tag{3.95}$$

$$X_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \varepsilon_2 \tag{3.96}$$

where

$$\begin{aligned} &\sigma_1^2 = \mathrm{V}[\varepsilon_1]; \quad \sigma_2^2 = \mathrm{V}[\varepsilon_2] \\ &\mathrm{V}[F_1] = \mathrm{V}[F_2] = 1 \end{aligned} \tag{3.97}$$

The variances are equal to 1 because such standardization does not impose additional constraints while it allows identification. This in a sense simply determines the units of measure of the unobserved construct.

Let us now consider the consequences that these equations impose on the structure of the covariance matrix of the observed variables:

$$\mathrm{V}[X_1] = \lambda_{11}^2 + \lambda_{12}^2 + \sigma_1^2 \tag{3.98}$$

Using the property that the factors are orthogonal (uncorrelated, with a variance of 1),

$$Cov[X_1, X_2] = \mathrm{E}[(\lambda_{11}F_1 + \lambda_{12}F_2 + \varepsilon_1)(\lambda_{21}F_1 + \lambda_{22}F_2 + \varepsilon_2)] \tag{3.99}$$

$$= \lambda_{11}\lambda_{21}\mathrm{E}\left[F_1^2\right] + \lambda_{12}\lambda_{22}\mathrm{E}\left[F_2^2\right] + \mathrm{E}[\varepsilon_1\varepsilon_2] \tag{3.100}$$

$$= \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22} \tag{3.101}$$

These equalities follow from the fact that

$$Cov[F_1, F_2] = 0 \tag{3.102}$$

$$\mathrm{E}[\varepsilon_1\varepsilon_2] = 0 \tag{3.103}$$

$$\mathrm{V}[F_1] = \mathrm{V}[F_2] = 1 \tag{3.104}$$

Therefore, the variances in the covariance matrix are composed of two components—commonalities and unique components:

$$\mathrm{V}[X_1] = \underbrace{\lambda_{11}^2 + \lambda_{12}^2}_{c_1^2} + \sigma_1^2 \tag{3.105}$$

In Eq. (3.105) ${c_1}^2$ represents the proportion of variance explained by the common factors while ${\sigma_1}^2$ represents the unique variance.

The commonalities are our main concern because the error variance or the unique variances do not contain information about the data structure. This demonstrates that the noise or the measurement error needs to be removed, although measurement error only affects the variances (the diagonal of the covariance matrix) and not the covariances.

More generally, we can represent the data structure as

$$\mathbf{\Sigma} = \mathbf{C} + \mathbf{U} \tag{3.106}$$

where $\mathbf{U} = \mathrm{diag}\{\mathrm{u}\}$.

$\mathbf{C}$ is the matrix of common variances and covariances, and $\mathbf{U}$ is the matrix of unique variances. In EFA, the objective is to reduce the dimensionality of the $\mathbf{C}$ matrix to understand better the underlying factors driving this structure pattern.

Four steps are involved in EFA: (1) estimating commonalities, (2) extracting the initial factors, (3) determining the number of factors, and (4) rotating to a terminal solution. We discuss each step in turn and then we derive the factor loadings and the factor scores.

3.2.4.2 Estimating Commonalities

In this first step, we need to remove the unique component of the variance so that the variance is explained only by the common factors. In a typical EFA, the diagonal elements of $\mathbf{C}$ are specified as the squared multiple correlations of each variable with the remainder of the variables in the set (i.e., the percentage of explained variance obtained in regressing variable j on the $(p - 1)$ other variables). $\mathbf{U}$ (a diagonal matrix) contains the residual variances from these regressions.

3.2.4.3 Extracting Initial Factors

The initial factors are obtained by performing a PCA on $\mathbf{C}$:

$$\underset{p\times p}{\mathbf{C}} = \underset{p\times p}{\mathbf{V}} \quad \underset{p\times p}{\mathbf{\Lambda}} \quad \underset{p\times p}{\mathbf{V}'} \tag{3.107}$$

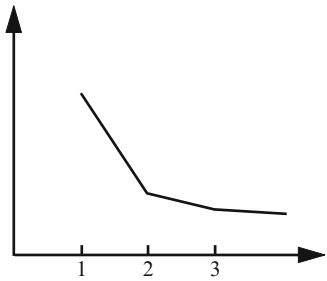


Fig. 3.8 The elbow rule

3.2.4.4 Determining the Number of Factors

The issue is to find the number of factors $r < p$ that are necessary to represent the covariance structure. Following from the properties of eigenvalues and eigenvectors:

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' \tag{3.108}$$

Let V* include the eigenvectors corresponding to the r largest eigenvalues and Λ^* include the r largest eigenvalues:

$$\underset{p\times p}{\mathbf{C}^*} = \underset{p\times r}{\mathbf{V}^*} \quad \underset{r\times r}{\mathbf{\Lambda}^*} \quad \underset{r\times p}{\mathbf{V}^{*'}} \tag{3.109}$$

The problem is to find r in order to account for most of the covariance matrix **C**.

Two rules are typically used to decide how many factors to retain.

1. $\lambda > 1$ (Kaiser's rule): Eliminate values less than 1. The rationale for this rule is that each factor should account for at least the variance of a single variable. However, this value is somewhat arbitrary.
2. The elbow rule based on the Scree plot: The Scree plot consists in plotting the eigenvalues in the order of their decreasing size. The elbow rule corresponds to finding the point on the Scree plot where the plotted line makes an elbow, as shown in Fig. 3.8. The elbow in the curve is due to the sharp decrease in the eigenvalues followed by smaller differences of the successive eigenvalues. Note that it may not always be easy to identify the exact point of the elbow.

None of these methods should be used blindly, even though the rule of the eigenvalue greater than 1 is the default option on most statistical analysis software packages, including SAS. Indeed, the interpretation of the factors is an important criterion for making sense out of the covariance structure.

3.2.4.5 Rotation to Terminal Solution

The reason why we perform a rotation at this stage, using only the retained factors, is to find factors that are more easily interpretable.

The most commonly used method is the VARIMAX rotation method. With this method, the rotation searches to give the maximum variance of the *squared* loadings for each factor (in order to avoid problems due to negative loadings). This results in obtaining extreme loadings (very high or very low).

3.2.4.6 Factor Loadings

If we consider the standardized correlation matrix of the **x** variables, which we write as **R**, Eq. (3.106) becomes

$$\mathbf{R} = \widetilde{\mathbf{C}} + \widetilde{\mathbf{U}} \tag{3.110}$$

The principal decomposition of $\widetilde{\mathbf{C}}$ leads to

$$\widetilde{\mathbf{C}} = \widetilde{\mathbf{V}}\,\widetilde{\Lambda}\,\widetilde{\mathbf{V}}' \tag{3.111}$$

However,

$$\widetilde{\mathbf{C}} = \widetilde{\mathbf{V}}\,\widetilde{\Lambda}\,\widetilde{\mathbf{V}}' = \widetilde{\mathbf{V}}\,\widetilde{\Lambda}^{\frac{1}{2}}\widetilde{\Lambda}^{\frac{1}{2}}\widetilde{\mathbf{V}}' = \mathbf{L}\mathbf{L}'$$

where

$$\underset{p\times p}{\mathbf{L}} = \widetilde{\mathbf{V}}\,\widetilde{\Lambda}^{\frac{1}{2}} \tag{3.112}$$

L is the matrix of factor loadings, similar to the formulation developed for PCA in Eq. (3.94), with the difference that it corresponded then to the principal decomposition of the common variance matrix rather than the full correlation matrix. The factor loadings are the correlations between the **x** variables and the factors.

3.2.4.7 Factor Scores

The factor scores provide the coordinates of the N observations on the (reduced number of) factors. The values of the **x** variables are combined in a linear fashion to form the factor scores $\widetilde{\mathbf{Y}}$:

$$\underset{N\times p}{\widetilde{\mathbf{Y}}} = \underset{N\times p}{\widetilde{\mathbf{X}}}\ \underset{p\times p}{\mathbf{B}} \tag{3.113}$$

where **B** is a matrix of the weights to apply. The problem consists in finding the weights that need to be applied. If we pre-multiply each side of Eq. (3.113) by $\frac{1}{N}\mathbf{X}'$, we obtain

$$\frac{1}{N}\widetilde{\mathbf{X}}'\widetilde{\mathbf{Y}} = \frac{1}{N}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{B} = \mathbf{RB} \tag{3.114}$$

Noticing that $\frac{1}{N}\widetilde{\mathbf{X}}'\widetilde{\mathbf{Y}} = \mathbf{L}$ from Eqs. (3.92), (3.93), and (3.94), it follows that

$$\mathbf{L} = \mathbf{RB} \tag{3.115}$$

Consequently,

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{L} \tag{3.116}$$

Therefore,

$$\underset{N\times p}{\widetilde{\mathbf{Y}}} = \underset{N\times p}{\widetilde{\mathbf{X}}}\ \underset{p\times p}{\mathbf{R}^{-1}}\ \underset{p\times p}{\mathbf{L}} \tag{3.117}$$

3.3 Application Examples

Figure 3.9 illustrates how to compute the means and the correlation matrix for a list of variables in SAS. The output is shown in Fig. 3.10. A factor analysis on the same list of variables is requested in Fig. 3.11 using the SAS procedure "Factor." The results are shown in Fig. 3.12. This factor analysis of the perception of innovations on nine characteristics is summarized by two factors with eigenvalues greater than 1 (the default option in SAS); these two factors explain 89.69% of the variance. The rotated factor pattern shows that Factor 1 is reflected by variables IT1, IT3, IT4, IT6, and IT7, while Factor 2 is reflected by variables IT5, IT8, and IT9. Variable IT2 does not discriminate well between the two factors, as it loads simultaneously

```
/*   examp3-1.sas
     computes means and correlation matrix
*/
option ls=120;
data data1;
infile 'c:\SAMD\Chapter3\Examples\product.dat';
input prod rad it1 it2 it3 it4 it5 it6 it7 it8 it9;
if it1=9 then it1=.;
if it2=9 then it2=.;
if it3=9 then it3=.;
if it4=9 then it4=.;
if it5=9 then it5=.;
if it6=9 then it6=.;
if it7=9 then it7=.;
if it8=9 then it8=.;
if it9=9 then it9=.;

proc means;
     var it1 it2 it3 it4 it5 it6 it7 it8 it9;
run;
proc corr;
     var it1 it2 it3 it4 it5 it6 it7 it8 it9;
run;
```

Fig. 3.9 SAS input file example for computing means and correlation matrix (examp3-1.sas)

```
Variable   N          Mean       Std Dev       Minimum       Maximum
                     ---------------------------------------------------------------
                     IT1        13     2.9230769     1.8009969     1.0000000     6.0000000
                     IT2        12     4.9166667     0.9962049     3.0000000     6.0000000
                     IT3        13     3.0000000     1.7320508     1.0000000     6.0000000
                     IT4        12     3.3333333     2.0150946     1.0000000     6.0000000
                     IT5        11     3.1818182     1.6624188     1.0000000     6.0000000
                     IT6        12     3.7500000     1.6583124     1.0000000     6.0000000
                     IT7        13     3.6923077     1.7504578     1.0000000     6.0000000
                     IT8        13     4.2307692     1.4232502     1.0000000     6.0000000
                     IT9        13     4.3846154     1.7577666     1.0000000     6.0000000
                     ---------------------------------------------------------------


                                        Correlation Analysis
          9 'VAR' Variables:  IT1      IT2      IT3      IT4      IT5      IT6      IT7      IT8      IT9
                                         Simple Statistics
     Variable                 N             Mean          Std Dev              Sum          Minimum          Maximum
     IT1                     13         2.923077         1.800997         38.000000         1.000000         6.000000
     IT2                     12         4.916667         0.996205         59.000000         3.000000         6.000000
     IT3                     13         3.000000         1.732051         39.000000         1.000000         6.000000
     IT4                     12         3.333333         2.015095         40.000000         1.000000         6.000000
     IT5                     11         3.181818         1.662419         35.000000         1.000000         6.000000
     IT6                     12         3.750000         1.658312         45.000000         1.000000         6.000000
     IT7                     13         3.692308         1.750458         48.000000         1.000000         6.000000
     IT8                     13         4.230769         1.423250         55.000000         1.000000         6.000000
     IT9                     13         4.384615         1.757767         57.000000         1.000000         6.000000
```

Fig. 3.10 SAS output example for computation of means and correlations (examp3-1.lst)

Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / Number of Observations

	IT1	IT2	IT3	IT4	IT5	IT6	IT7	IT8	IT9
IT1	1.00000	-0.81024	0.98843	-0.80100	-0.66790	-0.85535	-0.90687	-0.67522	-0.41105
	0.0	0.0014	0.0001	0.0017	0.0247	0.0004	0.0001	0.0113	0.1629
	13	12	13	12	11	12	13	13	13
IT2	-0.81024	1.00000	-0.78272	0.80847	0.73283	0.63671	0.77374	0.86405	0.57608
	0.0014	0.0	0.0026	0.0026	0.0159	0.0352	0.0031	0.0003	0.0500
	12	12	12	11	10	11	12	12	12
IT3	0.98843	-0.78272	1.00000	-0.79341	-0.66790	-0.85385	-0.90703	-0.67609	-0.41057
	0.0001	0.0026	0.0	0.0021	0.0247	0.0004	0.0001	0.0112	0.1635
	13	12	13	12	11	12	13	13	13
IT4	-0.80100	0.80847	-0.79341	1.00000	0.51068	0.89776	0.87782	0.59463	0.50232
	0.0017	0.0026	0.0021	0.0	0.1315	0.0001	0.0002	0.0414	0.0961
	12	11	12	12	10	12	12	12	12
IT5	-0.66790	0.73283	-0.66790	0.51068	1.00000	0.35687	0.63479	0.74788	0.44725
	0.0247	0.0159	0.0247	0.1315	0.0	0.3114	0.0359	0.0081	0.1678
	11	10	11	10	11	10	11	11	11
IT6	-0.85535	0.63671	-0.85385	0.89776	0.35687	1.00000	0.88462	0.46717	0.31882
	0.0004	0.0352	0.0004	0.0001	0.3114	0.0	0.0001	0.1257	0.3125
	12	11	12	12	10	12	12	12	12
IT7	-0.90687	0.77374	-0.90703	0.87782	0.63479	0.88462	1.00000	0.59951	0.25834
	0.0001	0.0031	0.0001	0.0002	0.0359	0.0001	0.0	0.0303	0.3941
	13	12	13	12	11	12	13	13	13
IT8	-0.67522	0.86405	-0.67609	0.59463	0.74788	0.46717	0.59951	1.00000	0.72770
	0.0113	0.0003	0.0112	0.0414	0.0081	0.1257	0.0303	0.0	0.0048
	13	12	13	12	11	12	13	13	13
IT9	-0.41105	0.57608	-0.41057	0.50232	0.44725	0.31882	0.25834	0.72770	1.00000
	0.1629	0.0500	0.1635	0.0961	0.1678	0.3125	0.3941	0.0048	0.0
	13	12	13	12	11	12	13	13	13

Fig. 3.10 (continued)

```
/*   examp3-2.sas
     Factor analysis
*/
option ls=120;
data data1;
infile 'c:\SAMD\Chapter3\Examples\product.dat';
input prod rad it1 it2 it3 it4 it5 it6 it7 it8 it9;
if it1=9 then it1=.;
if it2=9 then it2=.;
if it3=9 then it3=.;
if it4=9 then it4=.;
if it5=9 then it5=.;
if it6=9 then it6=.;
if it7=9 then it7=.;
if it8=9 then it8=.;
if it9=9 then it9=.;

proc factor rotate=varimax;
     var it1 it2 it3 it4 it5 it6 it7 it8 it9;
run;
```

Fig. 3.11 SAS input file example for factor analysis (examp3-2.sas)

on both, although it loads slightly more on Factor 2. The reliability coefficients of the scales (corresponding to the two factors) are then calculated in Fig. 3.13 when the variables are first standardized. Those variables with negative loadings are reversed so that each component has the same direction (positive correlations). The results are listed in Fig. 3.14, which shows the reliability coefficient alpha for each scale and the improvements that could be obtained by deleting any single variable one at a time. Finally, Fig. 3.15 shows how to create a scale composed of these standardized variables. The new scales "tech" and "mkt" involve two SAS functions: (1) the "sum(var1, var2, etc...)" function takes the sum of each variable in the list of variables in parentheses following the function (omitting the missing variables) and (2) the "n(var1, var2, etc...)" function returns the number of non-missing items in the variable list. As an example, also in Fig. 3.15, these scales are then used to perform a single analysis of variance, using the SAS procedure "proc ANOVA." The corresponding output in Fig. 3.16 shows, for example, the means of the two scales (labeled Tech and MKT) for two levels of the variable RAD.

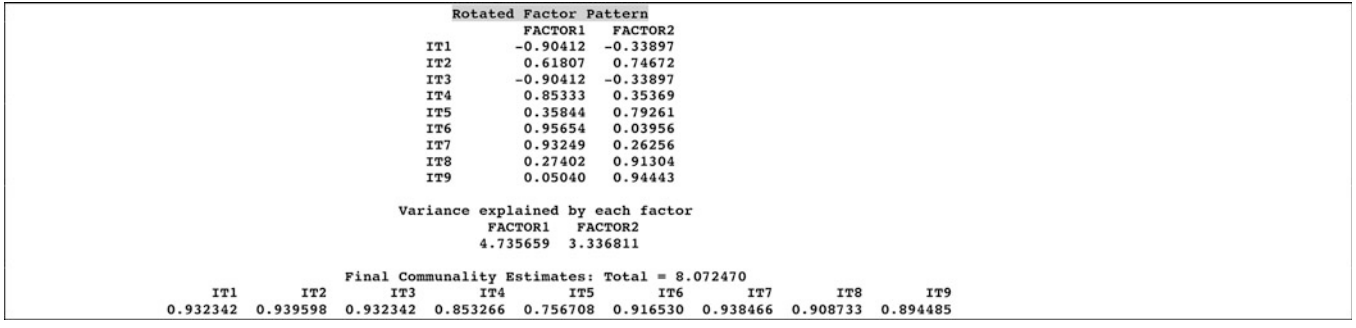
Using STATA, the input file corresponding to the same example is given in Fig. 3.17.

The "pca" procedure refers to PCA, as described earlier in this chapter. The list of variables to be analyzed simply follows the "pca" command. The commands are similar for EFA where "pca" is replaced with "factor." The option "mineigen(1)" indicates that only eigenvalues with a minimum of 1 will be retained. Note that the "pause" command is intended to give temporary control back to the researcher and thus provide the opportunity to save the graphs. To continue the execution of the do-file, just type "end" or "q" in the command zone. Alternatively, the graphs can be saved as files (STATA .gph files or other formats, such as .pdf files), using the commands shown in Fig. 3.18.

The command "graph save" saves the graph (with a .gph file extension) so that it can be read later using STATA. The command "graph export" is used for other

```
Initial Factor Method: Principal Components
                              Prior Communality Estimates: ONE
                        Eigenvalues of the Correlation Matrix:  Total = 9  Average = 1
                    1          2          3          4          5          6          7          8          9
Eigenvalue      6.3837     1.6888     0.4677     0.1936     0.1446     0.0792     0.0346     0.0078     0.0000
Difference      4.6949     1.2210     0.2742     0.0490     0.0654     0.0447     0.0267     0.0078
Proportion      0.7093     0.1876     0.0520     0.0215     0.0161     0.0088     0.0038     0.0009     0.0000
Cumulative      0.7093     0.8969     0.9489     0.9704     0.9865     0.9953     0.9991     1.0000     1.0000

                    2 factors will be retained by the MINEIGEN criterion.
                                      Factor Pattern
                                          FACTOR1    FACTOR2
                                  IT1     -0.92918    0.26260
                                  IT2      0.94032    0.23536
                                  IT3     -0.92918    0.26260
                                  IT4      0.89699   -0.22064
                                  IT5      0.75835    0.42615
                                  IT6      0.79402   -0.53485
                                  IT7      0.90676   -0.34096
                                  IT8      0.76170    0.57319
                                  IT9      0.60015    0.73096

                                 Variance explained by each factor
                                          FACTOR1    FACTOR2
                                         6.383698   1.688772
Initial Factor Method: Principal Components
                              Final Communality Estimates: Total = 8.072470
                IT1        IT2        IT3        IT4        IT5        IT6        IT7        IT8        IT9
           0.932342   0.939598   0.932342   0.853266   0.756708   0.916530   0.938466   0.908733   0.894485

Rotation Method: Varimax
                                  Orthogonal Transformation Matrix
                                              1          2
                                  1      0.80559    0.59247
                                  2     -0.59247    0.80559
```

Fig. 3.12 SAS output of factor analysis (examp3-2.lst)



Rotated Factor Pattern

	FACTOR1	FACTOR2
IT1	-0.90412	-0.33897
IT2	0.61807	0.74672
IT3	-0.90412	-0.33897
IT4	0.85333	0.35369
IT5	0.35844	0.79261
IT6	0.95654	0.03956
IT7	0.93249	0.26256
IT8	0.27402	0.91304
IT9	0.05040	0.94443

Variance explained by each factor

FACTOR1	FACTOR2
4.735659	3.336811

Final Communality Estimates: Total = 8.072470

IT1	IT2	IT3	IT4	IT5	IT6	IT7	IT8	IT9
0.932342	0.939598	0.932342	0.853266	0.756708	0.916530	0.938466	0.908733	0.894485

Fig. 3.12 (continued)

```
/*   examp3-3.sas
     Reliability Coefficient Alpha
*/
option ls=120;
data data1;
infile 'c:\SAMD\Chapter3\Examples\product.dat';
input prod rad it1 it2 it3 it4 it5 it6 it7 it8 it9;
if it1=9 then it1=.;
if it2=9 then it2=.;
if it3=9 then it3=.;
if it4=9 then it4=.;
if it5=9 then it5=.;
if it6=9 then it6=.;
if it7=9 then it7=.;
if it8=9 then it8=.;
if it9=9 then it9=.;
it1r=7-it1;
it3r=7-it3;

proc means;
     var it1r it2 it3r it4 it5 it6 it7 it8 it9;
     output out=results mean=m1r m2 m3r m4 m5 m6 m7 m8 m9
                        std=s1r s2 s3r s4 s5 s6 s7 s8 s9;
run;

data data2;
set data1;
if _n_=1 then set results;

it1rs=(it1r-m1r)/s1r;
it2s=(it2-m2)/s2;
it3rs=(it3r-m3r)/s3r;
it4s=(it4-m4)/s4;
it5s=(it5-m5)/s5;
it6s=(it6-m6)/s6;
it7s=(it7-m7)/s7;
it8s=(it8-m8)/s8;
it9s=(it9-m9)/s9;
run;
proc corr alpha;
  var it1rs it3rs it4s it6s it7s;
 run;
proc corr alpha;
  var it2s it5s it8s it9s;
run;
```

Fig. 3.13 SAS input file for reliability coefficient alpha (examp3-3.sas)

formats such as .pdf. The graph can then be imported into a document or a presentation. The "replace" option prevents an error message if the file already exists. The corresponding STATA output is shown in Fig. 3.19.

Figure 3.20 shows the Scree plot of eigenvalues that can help identify the number of relevant factors.

Finally, Fig. 3.21 can be used for the interpretation of the factors. As shown in the figure, the vector from the origin to the point representing a variable reflects the correlations between the variable and each of the factors. Consequently, the closer the vector is to a factor axis, the higher the correlation is between the variable and that factor.

Figure 3.17 also gives the commands to calculate the reliability coefficient alpha and to create unweighted composite scales. The simple command "alpha it1 it3 it4 it6 it7, generate(Tech) reverse(it1 it3) std" gives the instruction to compute the reliability coefficient alpha for the scale formed by the items that follow "alpha."

```
Variable   N          Mean       Std Dev       Minimum       Maximum
-----------------------------------------------------------------------
IT1R      13     4.0769231     1.8009969     1.0000000     6.0000000
IT2       12     4.9166667     0.9962049     3.0000000     6.0000000
IT3R      13     4.0000000     1.7320508     1.0000000     6.0000000
IT4       12     3.3333333     2.0150946     1.0000000     6.0000000
IT5       11     3.1818182     1.6624188     1.0000000     6.0000000
IT6       12     3.7500000     1.6583124     1.0000000     6.0000000
IT7       13     3.6923077     1.7504578     1.0000000     6.0000000
IT8       13     4.2307692     1.4232502     1.0000000     6.0000000
IT9       13     4.3846154     1.7577666     1.0000000     6.0000000
-----------------------------------------------------------------------

                        Correlation Analysis

    5 'VAR' Variables:  IT1RS    IT3RS    IT4S     IT6S     IT7S

                        Simple Statistics
Variable              N            Mean         Std Dev             Sum         Minimum         Maximum
IT1RS                13               0        1.000000               0       -1.708456        1.067785
IT3RS                13               0        1.000000               0       -1.732051        1.154701
IT4S                 12               0        1.000000               0       -1.157927        1.323346
IT6S                 12               0        1.000000               0       -1.658312        1.356801
IT7S                 13               0        1.000000               0       -1.538059        1.318336
```

Fig. 3.14 SAS output example of reliability coefficient alpha (examp3-3.lst)

Correlation Analysis

Cronbach Coefficient Alpha

for RAW variables : 0.975822
for STANDARDIZED variables: 0.972620

Deleted Variable	Raw Variables Correlation with Total	Raw Variables Alpha	Std. Variables Correlation with Total	Std. Variables Alpha
IT1RS	0.943833	0.967688	0.935001	0.963707
IT3RS	0.939013	0.968428	0.932076	0.964164
IT4S	0.888799	0.976050	0.876213	0.972786
IT6S	0.929027	0.969955	0.915387	0.966759
IT7S	0.944836	0.967534	0.943116	0.962438

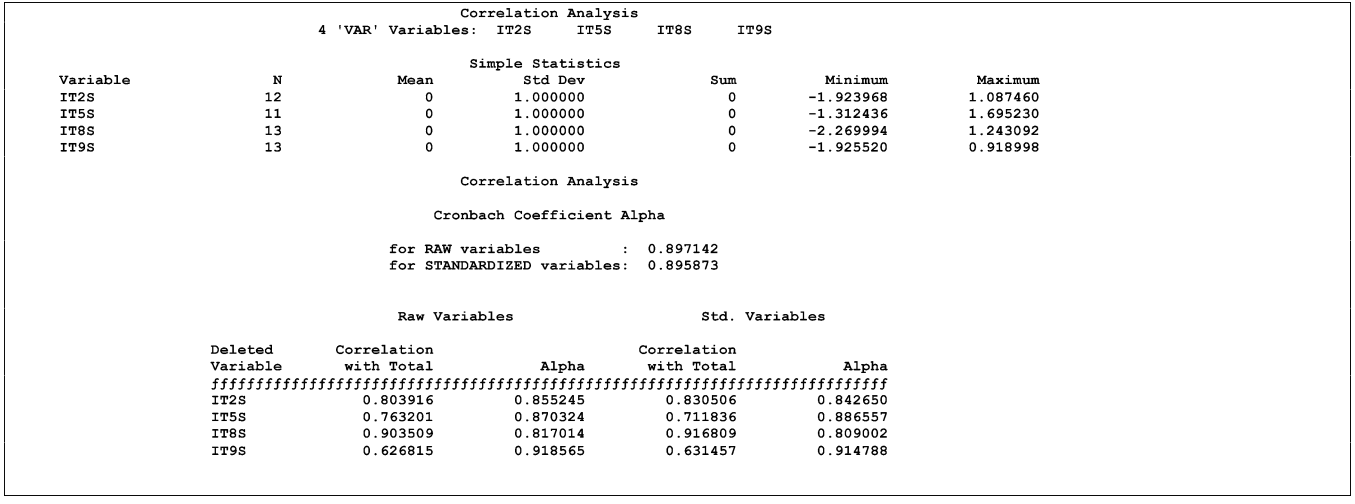
Fig. 3.14 (continued)

Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / Number of Observations

	IT1RS	IT3RS	IT4S	IT6S	IT7S
IT1RS	1.00000	0.98843	0.80100	0.85535	0.90687
	0.0	0.0001	0.0017	0.0004	0.0001
	13	13	12	12	13
IT3RS	0.98843	1.00000	0.79341	0.85385	0.90703
	0.0001	0.0	0.0021	0.0004	0.0001
	13	13	12	12	13
IT4S	0.80100	0.79341	1.00000	0.89776	0.87782
	0.0017	0.0021	0.0	0.0001	0.0002
	12	12	12	12	12
IT6S	0.85535	0.85385	0.89776	1.00000	0.88462
	0.0004	0.0004	0.0001	0.0	0.0001
	12	12	12	12	12
IT7S	0.90687	0.90703	0.87782	0.88462	1.00000
	0.0001	0.0001	0.0002	0.0001	0.0
	13	13	12	12	13

Fig. 3.14 (continued)



Correlation Analysis
4 'VAR' Variables: IT2S IT5S IT8S IT9S

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
IT2S	12	0	1.000000	0	-1.923968	1.087460
IT5S	11	0	1.000000	0	-1.312436	1.695230
IT8S	13	0	1.000000	0	-2.269994	1.243092
IT9S	13	0	1.000000	0	-1.925520	0.918998

Correlation Analysis

Cronbach Coefficient Alpha

for RAW variables : 0.897142
for STANDARDIZED variables: 0.895873

	Raw Variables		Std. Variables	
Deleted Variable	Correlation with Total	Alpha	Correlation with Total	Alpha
IT2S	0.803916	0.855245	0.830506	0.842650
IT5S	0.763201	0.870324	0.711836	0.886557
IT8S	0.903509	0.817014	0.916809	0.809002
IT9S	0.626815	0.918565	0.631457	0.914788

Fig. 3.14 (continued)

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / Number of Observations

	IT2S	IT5S	IT8S	IT9S
IT2S	1.00000	0.73283	0.86405	0.57608
	0.0	0.0159	0.0003	0.0500
	12	10	12	12
IT5S	0.73283	1.00000	0.74788	0.44725
	0.0159	0.0	0.0081	0.1678
	10	11	11	11
IT8S	0.86405	0.74788	1.00000	0.72770
	0.0003	0.0081	0.0	0.0048
	12	11	13	13
IT9S	0.57608	0.44725	0.72770	1.00000
	0.0500	0.1678	0.0048	0.0
	12	11	13	13

Fig. 3.14 (continued)

```
/*   examp3-4.sas
     Scales
*/
option ls=120;
data data1;
infile 'c:\SAMD\Chapter3\Examples\product.dat';
input prod rad it1 it2 it3 it4 it5 it6 it7 it8 it9;
if it1=9 then it1=.;
if it2=9 then it2=.;
if it3=9 then it3=.;
if it4=9 then it4=.;
if it5=9 then it5=.;
if it6=9 then it6=.;
if it7=9 then it7=.;
if it8=9 then it8=.;
if it9=9 then it9=.;
it1r=7-it1;
it3r=7-it3;
proc means;
     var it1r it2 it3r it4 it5 it6 it7 it8 it9;
     output out=results mean=m1r m2 m3r m4 m5 m6 m7 m8 m9
                        std=s1r s2 s3r s4 s5 s6 s7 s8 s9;
run;
data data2;
set data1;
if _n_=1 then set results;

it1rs=(it1r-m1r)/s1r;
it2s=(it2-m2)/s2;
it3rs=(it3r-m3r)/s3r;
it4s=(it4-m4)/s4;
it5s=(it5-m5)/s5;
it6s=(it6-m6)/s6;
it7s=(it7-m7)/s7;
it8s=(it8-m8)/s8;
it9s=(it9-m9)/s9;

tech=sum(it1rs,it3rs,it4s,it6s,it7s)/n(it1rs,it3rs,it4s,it6s,it7s);
mkt=sum(it2s,it5s,it8s,it9s)/n(it2s,it5s,it8s,it9s);
run;

proc anova;
  class rad;
  model tech mkt = rad;
  means rad;
run;
```

Fig. 3.15 SAS input file example for scale construction (examp3-4.sas)

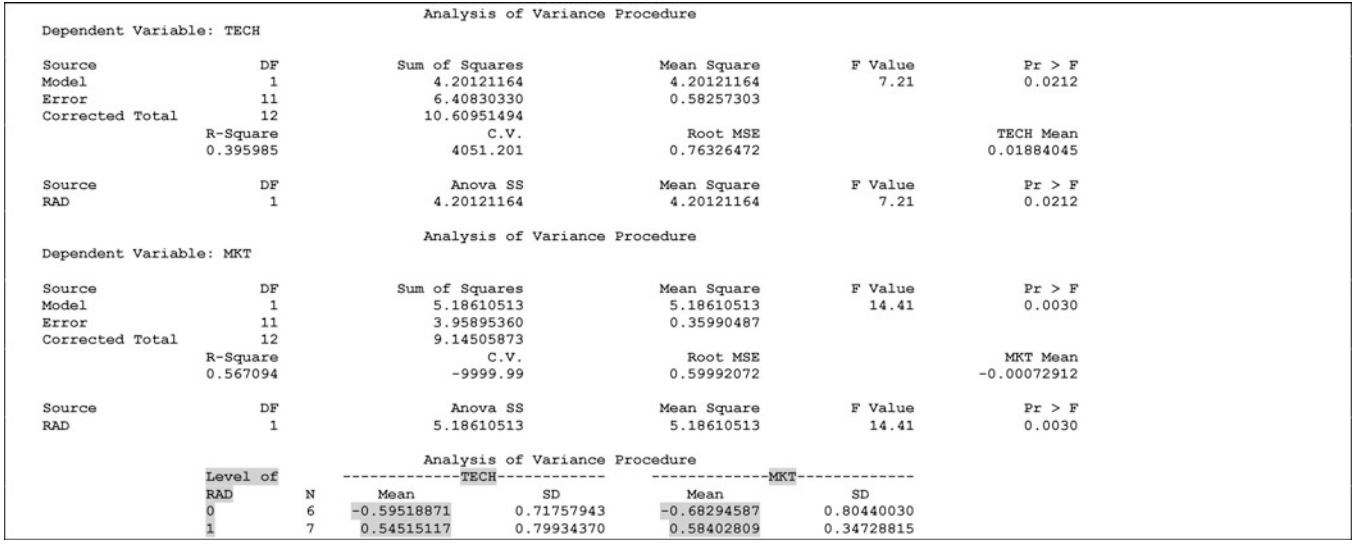
The scale "Tech" is then generated from the list of variables included in the list following the "alpha" command. Note that there is no need to generate separate variables that are reverse coded, as it is sufficient to list these items in parentheses after the word "reverse." The command "std" indicates that the standardized variables will be used as components of the unweighted composite scale.

Principal component scores and factor scores can be computed easily in STATA. Figure 3.22 lists the input to create new variables in the database containing the scores of the first four factors (as an example). The new variable names are score1, score2, score3, and score4. It is then easy to check that the correlation matrix of these variables is the identity matrix.

Similarly for factor analysis, the factor scores corresponding to the two factors obtained after rotation are obtained by the commands listed in Fig. 3.23.

```
Variable   N          Mean       Std Dev       Minimum       Maximum
------------------------------------------------------------------------
IT1R       13     4.0769231     1.8009969     1.0000000     6.0000000
IT2        12     4.9166667     0.9962049     3.0000000     6.0000000
IT3R       13     4.0000000     1.7320508     1.0000000     6.0000000
IT4        12     3.3333333     2.0150946     1.0000000     6.0000000
IT5        11     3.1818182     1.6624188     1.0000000     6.0000000
IT6        12     3.7500000     1.6583124     1.0000000     6.0000000
IT7        13     3.6923077     1.7504578     1.0000000     6.0000000
IT8        13     4.2307692     1.4232502     1.0000000     6.0000000
IT9        13     4.3846154     1.7577666     1.0000000     6.0000000
------------------------------------------------------------------------

                  Analysis of Variance Procedure
                     Class Level Information

                    Class    Levels    Values

                    RAD           2    0 1

                Number of observations in data set = 13
```

Fig. 3.16 SAS output example of scale construction and analysis of variance (examp3-4.lst)



Analysis of Variance Procedure

Dependent Variable: TECH

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.20121164	4.20121164	7.21	0.0212
Error	11	6.40830330	0.58257303		
Corrected Total	12	10.60951494			

R-Square	C.V.	Root MSE	TECH Mean
0.395985	4051.201	0.76326472	0.01884045

Source	DF	Anova SS	Mean Square	F Value	Pr > F
RAD	1	4.20121164	4.20121164	7.21	0.0212

Analysis of Variance Procedure

Dependent Variable: MKT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.18610513	5.18610513	14.41	0.0030
Error	11	3.95895360	0.35990487		
Corrected Total	12	9.14505873			

R-Square	C.V.	Root MSE	MKT Mean
0.567094	-9999.99	0.59992072	-0.00072912

Source	DF	Anova SS	Mean Square	F Value	Pr > F
RAD	1	5.18610513	5.18610513	14.41	0.0030

Analysis of Variance Procedure

Level of		TECH		MKT	
RAD	N	Mean	SD	Mean	SD
0	6	-0.59518871	0.71757943	-0.68294587	0.80440030
1	7	0.54515117	0.79934370	0.58402809	0.34728815

Fig. 3.16 (continued)

```
infile prod rad it1-it9 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter3_PCA-EFA/PRODUCT.DAT", clear
replace it1=. if it1==9
replace it2=. if it2==9
replace it3=. if it3==9
replace it4=. if it4==9
replace it5=. if it5==9
replace it6=. if it6==9
replace it7=. if it7==9
replace it8=. if it8==9
replace it9=. if it9==9
mean it1-it9
pwcorr it1-it9
pause on
pca it1- it9
factor it1-it9, mineigen(1)
rotate
screeplot
pause screeplot
loadingplot
pause loadplot
scoreplot
alpha it1 it3 it4 it6 it7, generate(Tech) reverse(it1 it3) std
alpha it2 it5 it8 it9, generate(Mkt) std
oneway Tech rad
oneway Mkt rad
mean Tech, over(rad)
mean Mkt, over (rad)
```

Fig. 3.17 STATA input file for principal component analysis, exploratory factor analysis, reliability coefficient alpha, scale construction, and analysis of variance example (Examp3-1.do)

```
screeplot
graph save "/Users/gatignon/Documents/WORK_STATA/SAMD/Chapter3_PCA-EFA/screeplot",
replace
graph export "/Users/gatignon/Documents/WORK_STATA/SAMD/Chapter3_PCA-
EFA/screeplot.pdf", replace
loadingplot
graph save "/Users/gatignon/Documents/WORK_STATA/SAMD/Chapter3_PCA-EFA/loadingplot",
replace
graph export "/Users/gatignon/Documents/WORK_STATA/SAMD/Chapter3_PCA-
EFA/loadingplot.pdf", replace
scoreplot
graph save "/Users/gatignon/Documents/WORK_STATA/SAMD/Chapter3_PCA-EFA/scoregplot",
replace
graph export "/Users/gatignon/Documents/WORK_STATA/SAMD/Chapter3_PCA-
EFA/scoreplot.pdf", replace
```

Fig. 3.18 STATA commands for saving graphs (Examp3-1B.do)

3.3.1 Assignment

The assignment consists in developing a composite scale, demonstrating its unidimensionality and computing its reliability. For that purpose, survey data are provided in the file SURVEY.ASC (Appendix C, Chap. 14). These data concern items about psychographic variables, which contain opinion, attitude, and lifestyle characteristics of individuals. A detailed description of the data is given in Appendix C. This type of data is useful for advertising and segmentation purposes.

In order to develop a scale, it may be useful to summarize the data using EFA on a wide range of variables. It is important, however, to make sure that only variables

possessing the properties necessary for the analysis are included. For example, because factor analysis is based on correlations, categorical or ordinal scale variables should be excluded from the analysis, since correlations are not permissible statistics with such scales. You need to interpret the factors, and you can concentrate on a subset of these factors to derive a single scale or multiple composite scales.

An alternative would be to reflect on the questions that seem interrelated and use them to develop a scale. This is in essence a mental factor analysis.

You need to demonstrate that each of the scales developed is unidimensional (through factor analysis) and that its reliability is sufficiently high.

Figure 3.24 lists the SAS file that can be used to read the data.

The commands to read the survey data with STATA are shown in Fig. 3.25. This defines how the data in the file "survey.asc" are formatted.

The data are then imported into STATA by executing the file "assign3_Mac.do" shown in Fig. 3.26.

```
. infile prod rad it1-it9 using "C:\DATA\WORK_STATA\SAMD\Chapter3_PCA-
EFA\PRODUCT.DAT", clear
(13 observations read)

. replace it1=. if it1==9
(0 real changes made)

. replace it2=. if it2==9
(1 real change made, 1 to missing)

. replace it3=. if it3==9
(0 real changes made)

. replace it4=. if it4==9
(1 real change made, 1 to missing)

. replace it5=. if it5==9
(2 real changes made, 2 to missing)

. replace it6=. if it6==9
(1 real change made, 1 to missing)

. replace it7=. if it7==9
(0 real changes made)

. replace it8=. if it8==9
(0 real changes made)

. replace it9=. if it9==9
(0 real changes made)

. mean it1-it9


Mean estimation                   Number of obs    =       9

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         it1 |          3   .5773503      1.668628    4.331372
         it2 |   4.777778   .3643021      3.937696     5.61786
```

Fig. 3.19 STATA output file for principal component analysis, exploratory factor analysis, reliability coefficient alpha, scale construction, and analysis of variance example (Examp3-1.log)

```
         it3 |          3   .5773503      1.668628    4.331372
         it4 |   3.222222   .7412036      1.513004    4.931441
         it5 |   3.111111   .5638273      1.810923    4.411299
         it6 |   3.666667   .5773503      2.335295    4.998039
         it7 |   3.444444   .6260793      2.000703    4.888186
         it8 |   4.222222   .4937886      3.083544    5.360901
         it9 |   4.666667   .5527708      3.391975    5.941358
-------------------------------------------------------------

. pwcorr it1-it9

             |      it1      it2      it3      it4      it5      it6      it7
-------------+---------------------------------------------------------------
         it1 |   1.0000
         it2 |  -0.8102   1.0000
         it3 |   0.9884  -0.7827   1.0000
         it4 |  -0.8010   0.8085  -0.7934   1.0000
         it5 |  -0.6679   0.7328  -0.6679   0.5107   1.0000
         it6 |  -0.8553   0.6367  -0.8539   0.8978   0.3569   1.0000
         it7 |  -0.9069   0.7737  -0.9070   0.8778   0.6348   0.8846   1.0000
         it8 |  -0.6752   0.8640  -0.6761   0.5946   0.7479   0.4672   0.5995
         it9 |  -0.4111   0.5761  -0.4106   0.5023   0.4473   0.3188   0.2583
             |      it8      it9
-------------+------------------
         it8 |   1.0000
         it9 |   0.7277   1.0000


. pca it1- it9

Principal components/correlation                  Number of obs    =          9
                                                  Number of comp.  =          8
                                                  Trace            =          9
    Rotation: (unrotated = principal)             Rho              =     1.0000

    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |       6.3837      4.69493             0.7093       0.7093
           Comp2 |      1.68877      1.22104             0.1876       0.8969
           Comp3 |       .467731       .274166            0.0520       0.9489
           Comp4 |       .193565     .0489504             0.0215       0.9704
           Comp5 |       .144614     .0653901             0.0161       0.9865
           Comp6 |     .0792242     .0446557             0.0088       0.9953
           Comp7 |     .0345684     .0267408             0.0038       0.9991
           Comp8 |    .00782764    .00782764             0.0009       1.0000
           Comp9 |            0            .             0.0000       1.0000
    --------------------------------------------------------------------------

Principal components (eigenvectors)

    ------------------------------------------------------------------------------------
------------
        Variable |    Comp1     Comp2     Comp3     Comp4     Comp5     Comp6
Comp7     Comp8
    -------------+----------------------------------------------------------------------
------------
           it1 |  -0.3678    0.2021    0.2732    0.0454    0.4702    0.0473   -
0.0116    0.1541
           it2 |   0.3722    0.1811    0.1012   -0.3829    0.1511   -0.2089   -
0.7673    0.1291
           it3 |  -0.3678    0.2021    0.2732    0.0454    0.4702    0.0473   -
0.0116    0.1541
           it4 |   0.3550   -0.1698    0.4729    0.3704    0.2614   -0.1856
0.0077   -0.6177
           it5 |   0.3001    0.3279   -0.6193    0.4249    0.4001    0.2658   -
0.0724   -0.0484
```

Fig. 3.19 (continued)

```
             it6 |   0.3143   -0.4116    0.3062    0.1510    0.0214    0.6497   -
0.0798    0.4316

             it7 |   0.3589   -0.2624   -0.1130   -0.1757    0.4024   -0.4708
0.4604    0.4040
             it8 |   0.3015    0.4411    0.1706   -0.5520    0.0498    0.3892
0.4119   -0.2391
             it9 |   0.2375    0.5625    0.3125    0.4161   -0.3765   -0.2270
0.1333    0.3812
    ------------------------------------------------------------------------------
------------


    ----------------------------
        Variable | Unexplained
    -------------+--------------
             it1 |            0
             it2 |            0
             it3 |            0
             it4 |            0
             it5 |            0
             it6 |            0
             it7 |            0
             it8 |            0
             it9 |            0
    ----------------------------

. factor it1-it9, mineigen(1)
(obs=9)
(collinear variables specified)

Factor analysis/correlation                       Number of obs    =         9
    Method: principal factors                     Retained factors =         2
    Rotation: (unrotated)                         Number of params =        17

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      6.33841      4.70974            0.7436       0.7436
        Factor2  |      1.62867      1.24160            0.1911       0.9346
        Factor3  |      0.38707      0.25467            0.0454       0.9800
        Factor4  |      0.13240      0.05566            0.0155       0.9956
        Factor5  |      0.07674      0.07094            0.0090       1.0046
        Factor6  |      0.00580      0.00580            0.0007       1.0053
        Factor7  |     -0.00000      0.01705           -0.0000       1.0053
        Factor8  |     -0.01705      0.01069           -0.0020       1.0033
        Factor9  |     -0.02773            .           -0.0033       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) =       . Prob>chi2 =      .

Factor loadings (pattern matrix) and unique variances

    -----------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
             it1 |  -0.9333    0.2576 |      0.0627
             it2 |   0.9358    0.2460 |      0.0637
             it3 |  -0.9333    0.2576 |      0.0627
             it4 |   0.8984   -0.2065 |      0.1503
             it5 |   0.7318    0.3805 |      0.3197
             it6 |   0.7946   -0.5209 |      0.0973
             it7 |   0.9064   -0.3299 |      0.0696
             it8 |   0.7551    0.5735 |      0.1009
             it9 |   0.5958    0.7341 |      0.1061
    -----------------------------------------------------

. rotate

Factor analysis/correlation                       Number of obs    =         9
    Method: principal factors                     Retained factors =         2
```

Fig. 3.19 (continued)

```
    Rotation: orthogonal varimax (Kaiser off)      Number of params =       17

    ------------------------------------------------------------------------------
         Factor  |     Variance   Difference         Proportion   Cumulative
    -------------+----------------------------------------------------------------
        Factor1  |      4.78896      1.61085            0.5618       0.5618
        Factor2  |      3.17812            .            0.3728       0.9346
    ------------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) =       . Prob>chi2 =      .

Rotated factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
              it1 |  -0.9122   -0.3243 |      0.0627
              it2 |   0.6254    0.7383 |      0.0637
              it3 |  -0.9122   -0.3243 |      0.0627
              it4 |   0.8543    0.3462 |      0.1503
              it5 |   0.3811    0.7314 |      0.3197
              it6 |   0.9497    0.0291 |      0.0973
              it7 |   0.9317    0.2497 |      0.0696
              it8 |   0.2896    0.9029 |      0.1009
              it9 |   0.0669    0.9431 |      0.1061
    -------------------------------------------------

Factor rotation matrix

    ---------------------------------
                 | Factor1  Factor2
    -------------+-------------------
         Factor1 |  0.8192   0.5736
         Factor2 | -0.5736   0.8192
    ---------------------------------

. screeplot

. loadingplot

. scoreplot

. alpha it1 it3 it4 it6 it7, generate(Tech) reverse(it1 it3) std


Test scale = mean(standardized items)

Reversed items: it1 it3

Average interitem correlation:      0.8780
Number of items in the scale:            5
Scale reliability coefficient:      0.9730


. alpha it2 it5 it8 it9, generate(Mkt) std

Test scale = mean(standardized items)


Average interitem correlation:      0.6843
Number of items in the scale:            4
Scale reliability coefficient:      0.8966
```

Fig. 3.19 (continued)

```
. oneway Tech rad

                        Analysis of Variance
    Source              SS         df       MS            F     Prob > F
------------------------------------------------------------------------
Between groups      4.20121221      1   4.20121221       7.21     0.0212
 Within groups      6.40830405     11   .582573096
------------------------------------------------------------------------
    Total           10.6095163     12   .884126356

Bartlett's test for equal variances:  chi2(1) =   0.0577  Prob>chi2 = 0.810


. oneway Mkt rad

                        Analysis of Variance
    Source              SS         df       MS            F     Prob > F
------------------------------------------------------------------------
Between groups      5.18610509      1   5.18610509     14.41     0.0030
 Within groups      3.95895361     11   .359904874
------------------------------------------------------------------------
    Total            9.1450587     12   .762088225

Bartlett's test for equal variances:  chi2(1) =   3.3214  Prob>chi2 = 0.068

. mean Tech, over(rad)

Mean estimation                     Number of obs    =      13

            0: rad = 0
            1: rad = 1
--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
Tech         |
           0 |  -.5951888   .2929506     -1.233473    .0430958
           1 |   .5451512   .3021235     -.1131194    1.203422
--------------------------------------------------------------

. mean Mkt, over (rad)

Mean estimation                     Number of obs    =      13

            0: rad = 0
            1: rad = 1
--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
Mkt          |
           0 |  -.6829459   .3283951     -1.398457    .0325655
           1 |   .5840281   .1312626      .2980315    .8700247
--------------------------------------------------------------
```

Fig. 3.19 (continued)

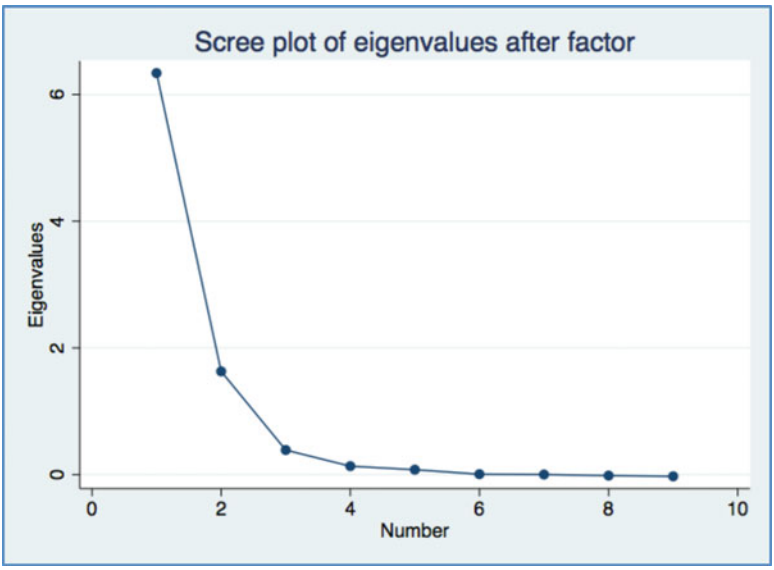


Fig. 3.20 STATA plot of eigenvalues

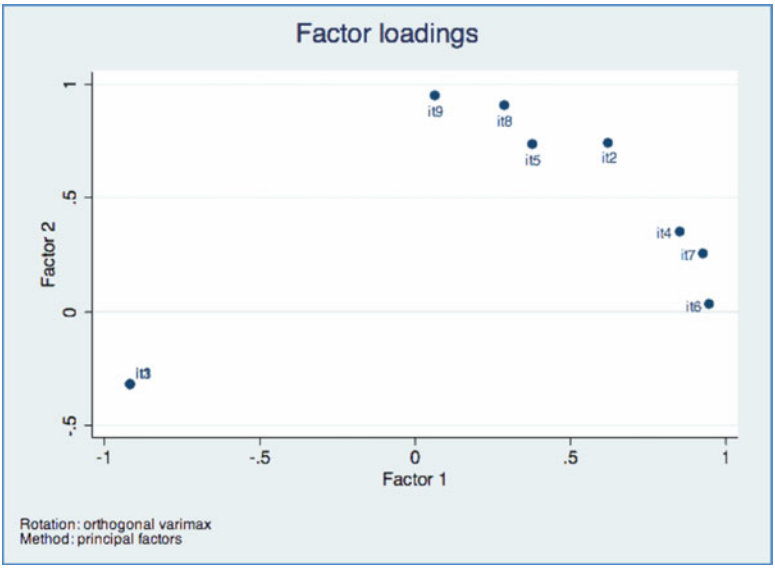


Fig. 3.21 STATA plot of factor loadings

```
pca it1- it9
predict score1-score4
```

Fig. 3.22 STATA commands to create variables for component scores

```
factor it1-it9, mineigen(1)
rotate
predict factor1 factor2
```

Fig. 3.23 STATA commands to create variables for factor scores

```
/*     Assign3.sas     */
filename survey 'c:\SAMD\Chapter3\Assignments\survey.asc';
data new;
infile survey firstobs=19;
input   (Age Marital Income Educatn HHSize Occuptn Location
        TryHair LatStyle DrssSmrt BlndsFun LookDif
        LookAttr GrocShp LikeBkng ClthFrsh WashHnds Sportng LikeClrs
        FeelAttr TooMchSx Social LikeMaid ServDnrs SaveRcps LikeKtch) (3.)
       #2 (LoveEat SpirtVal Mother ClascMsc Children Applianc ClsFamly
        LovFamly TalkChld Exercise LikeSelf CareSkin MedChckp
        EvngHome TripWrld HomeBody LondnPrs Comfort Ballet Parties
        WmnNtSmk BrghtFun Seasonng ColorTV SlppyPpl Smoke) (3.)
       #3 (Gasoline Headache Whiskey Bourbon FastFood Restrnts OutFrDnr
        OutFrLnc RentVide Catsup KnowSont PercvDif BrndLylt
        CatgMotv BrndMotv OwnSonit NecssSon OthrInfl DecsnTim
        RdWomen RdHomSrv RdFashn RdMenMag RdBusMag RdNewsMg
        RdGlMag) (3.)
       #4 (RdYouthM RdNwsppr WtchDay WtchEve WtchPrm
        WTchLate WtchWknd WtchCsby WtchFmTs WtchChrs WtchMoon
        WtchBoss WtchGrwP WtchMiaV WtchDns WtchGold WtchBowl) (3.);
proc freq;
tables OwnSonit*(Age Marital Income Educatn HHSize Occuptn);
run;
```

Fig. 3.24 SAS file to read SURVEY.ASC data file (assign3.sas)

```
infile dictionary using "/users/fblgatignon/Documents/WORK_STATA/SAMD/survey.asc" {
_first(19)
_lines(4)
_line(1)
        Age %3f
        Marital %3f
        Income %3f
        Educatn %3f
        HHSize %3f
        Occuptn %3f
        Location %3f
        TryHair %3f
        LatStyle %3f
        DrssSmrt %3f
        BlndsFun %3f
        LookDif %3f
        LookAttr %3f
        GrocShp %3f
        LikeBkng %3f
        ClthFrsh %3f
        WashHnds %3f
        Sportng %3f
        LikeClrs %3f
        FeelAttr %3f
        TooMchSx %3f
        Social %3f
        LikeMaid %3f
```

Fig. 3.25 STATA dictionary file to read SURVEY.ASC data file (assign3_Mac.dct)

```
        ServDnrs %3f
        SaveRcps %3f
        LikeKtch %3f
_line(2)
        LoveEat %3f
        SpirtVal %3f
        Mother %3f
        ClascMsc %3f
        Children %3f
        Applianc %3f
        ClsFamly %3f
        LovFamly %3f
        TalkChld %3f
        Exercise %3f
        LikeSelf %3f
        CareSkin %3f
        MedChckp %3f
        EvngHome %3f
        TripWrld %3f
        HomeBody %3f
        LondnPrs %3f
        Comfort %3f
        Ballet %3f
        Parties %3f
        WmnNtSmk %3f
        BrghtFun %3f
        Seasonng %3f
        ColorTV %3f
        SlppyPpl %3f
        Smoke %3f
_line(3)
        Gasoline %3f
        Headache %3f
        Whiskey %3f
        Bourbon %3f
        FastFood %3f
        Restrnts %3f
        OutFrDnr %3f
        OutFrLnc %3f
        RentVide %3f
        Catsup %3f
        KnowSont %3f
        PercvDif %3f
        BrndLylt %3f
        CatgMotv %3f
        BrndMotv %3f
        OwnSonit %3f
        NecssSon %3f
        OthrInfl %3f
        DecsnTim %3f
        RdWomen %3f
        RdHomSrv %3f
        RdFashn %3f
        RdMenMag %3f
        RdBusMag %3f
        RdNewsMg %3f
        RdGlMag %3f
_line(4)
        RdYouthM %3f
        RdNwsppr %3f
        WtchDay %3f
        WtchEve %3f
        WtchPrm %3f
        WTchLate %3f
        WtchWknd %3f
        WtchCsby %3f
        WtchFmTs %3f
        WtchChrs %3f
```

Fig. 3.25 (continued)

```
WtchMoon %3f
WtchBoss %3f
WtchGrwP %3f
WtchMiaV %3f
WtchDns %3f
WtchGold %3f
WtchBowl %3f
}
```

Fig. 3.25 (continued)

```
infile using "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter3_PCA-EFA
/Assign3_Mac"
```

Fig. 3.26 STATA do-file to read the file survey.asc

Bibliography

Basic Technical Readings

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314.

Cheryl, B. J., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(September), 199–218.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.

Diamanopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277.

Green, P. E. (1978). *Mathematical tools for applied multivariate analysis*. New York, NY: Academic [Chap. 5 and Chap. 6, Sect. 6.4].

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MS: Addison-Wesley [Chap. 4].

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw Hill.

Application Readings

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research, 34*(3), 347–356.

Anderson, E. (1985). The salesperson as outside agent or employee: A transaction cost analysis. *Marketing Science, 4*(Summer), 234–254.

Anderson, R., & Engledow, J. (1977). A factor analytic comparison of U.S. and German information seeker. *Journal of Consumer Research, 3*(4), 185–196.

Blackman, A. W. (1973). An innovation index based on factor analysis. *Technological Forecasting and Social Change, 4*, 301–316.

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*(February), 64–73.

Deshpande, R. (1982). The organizational context of market research use. *Journal of Marketing, 46*(4), 91–101.

Finn, A., & Kayandé, U. (1997). Reliability assessment and optimization of marketing measurement. *Journal of Marketing Research, 34*(2), 262–275.

Gilbert, F. W., & Warren, W. E. (1995). Psychographic constructs and demographic segments. *Psychology and Marketing, 12*(3), 223–237.

Green, S. G., Gavin, M. B., & Aiman-Smith, L. (1995). Assessing a multidimensional measure of radical technological innovation. *IEEE Transactions on Engineering Management, 42*(3), 203–214.

Murtha, T. P., Lenway, S. A., & Bagozzi, R. P. (1998). Global mind-sets and cognitive shift in a complex multinational corporation. *Strategic Management Journal, 19*, 97–114.

Perreault, W. D., Jr., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26*(May), 135–148.

Zaichowsky, J. L. (1985). Measuring the involvement construct. *Journal of Consumer Research, 12*(December), 341–352.

Chapter 4
Confirmatory Factor Analysis

As noted in Chap. 3, a measurement model of the type illustrated in Fig. 4.1 is assumed in confirmatory factor analysis (CFA).

The objective of a confirmatory analysis is to test if the data fit the measurement model.

4.1 Confirmatory Factor Analysis: A Strong Measurement Model

The graphical representation of the model shown in Fig. 4.1 can be expressed by the system of equations

$$\begin{cases} X_1 = \lambda_{11} F_1 + \mathrm{e}_1 \\ X_2 = \lambda_{21} F_1 + \mathrm{e}_2 \\ X_3 = \lambda_{31} F_1 + \mathrm{e}_3 \\ X_4 = \lambda_{42} F_2 + \mathrm{e}_4 \\ X_5 = \lambda_{52} F_2 + \mathrm{e}_5 \end{cases} \tag{4.1}$$

Let

$$\underset{5\times1}{\mathbf{x}} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix}; \quad \underset{2\times1}{\mathbf{F}} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}; \underset{5\times2}{\mathbf{\Lambda}} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \end{bmatrix}; \quad \underset{5\times1}{\mathbf{e}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

Equation (4.1) can be expressed in matrix notation as

$$\underset{5\times1}{\mathbf{x}} = \underset{5\times2}{\mathbf{\Lambda}} \underset{2\times1}{\mathbf{F}} + \underset{5\times1}{\mathbf{e}} \tag{4.2}$$

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_4, © Springer Science+Business Media New York 2014

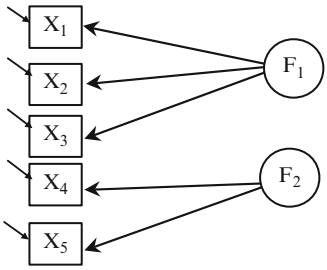


Fig. 4.1 A graphical representation of multiple measures with a confirmatory factor structure

with

$$\mathrm{E}\ [\mathbf{e}] = 0 \tag{4.3}$$

$$\mathrm{E}\ \left[\mathbf{e} \quad \mathbf{e}'\right] = \mathbf{D} = diag\{\delta_{ii}\} \tag{4.4}$$

$$\mathrm{E}\left[\mathbf{F}\mathbf{F}'\right] = \mathbf{\Phi} \tag{4.5}$$

If the factors are assumed to be independent,

$$\mathrm{E}\left[\mathbf{F}\mathbf{F}'\right] = \mathbf{I} \tag{4.6}$$

While we were referring to the specific model with five indicators in the expressions above, the matrix notation is general and can be used for representing a measurement model with q indicators and a factor matrix containing n unobserved factors:

$$\underset{q\times 1}{\mathbf{x}} = \underset{q\times n}{\mathbf{\Lambda}}\ \underset{n\times 1}{\mathbf{F}} + \underset{q\times 1}{\mathbf{e}} \tag{4.7}$$

The theoretical covariance matrix of $\mathbf{x}$ is given by

$$\mathbf{E}\left[\mathbf{x}\mathbf{x}'\right] = \mathrm{E}\left[(\mathbf{\Lambda}\mathbf{F} + \mathbf{e})(\mathbf{\Lambda}\mathbf{F} + \mathbf{e})'\right] \tag{4.8}$$

$$= \mathrm{E}\left[\mathbf{\Lambda}\mathbf{F}\mathbf{F}'\mathbf{\Lambda}' + \mathbf{e}\mathbf{e}'\right]$$

$$= \mathbf{\Lambda}\mathrm{E}\left[\mathbf{F}\mathbf{F}'\right]\mathbf{\Lambda}' + \mathrm{E}\left[\mathbf{e}\mathbf{e}'\right] \tag{4.9}$$

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{D} \tag{4.10}$$

Therefore, Eq. (4.10) expresses how the covariance matrix is structured, given the measurement model specification in Eq. (4.7). The structure is simplified in case of the independence of the factors:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{'} + \mathbf{D} \tag{4.11}$$

To facilitate comparison, especially between exploratory factor analysis (EFA) and CFA, the notation used above closely resembles the notation used in the previous chapter. However, we now introduce the notation found in LISREL because the software refers to specific variable names. In particular, Eq. (4.12) uses $\boldsymbol{\xi}$ for the vector of factors and $\boldsymbol{\delta}$ for the vector of measurement errors. Thus the measurement model is expressed as

$$\underset{q\times 1}{\mathbf{x}} = \underset{q\times n}{\boldsymbol{\Lambda}_x}\ \underset{n\times 1}{\boldsymbol{\xi}} + \underset{q\times 1}{\boldsymbol{\delta}} \tag{4.12}$$

with

$$\mathrm{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}^{'}\right] = \boldsymbol{\Phi} \tag{4.13}$$

and

$$\mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}^{'}\right] = \boldsymbol{\theta}_{\delta} \tag{4.14}$$

The methodology for estimating these parameters is presented in the next section.

4.2 Estimation

If the observed covariance matrix estimated from the sample is $\mathbf{S}$, we need to find the values of the lambdas (the elements of $\boldsymbol{\Lambda}$) and of the deltas (the elements of $\mathbf{D}$) that will reproduce a covariance matrix as similar as possible to the observed one. Maximum likelihood estimation is used to minimize $\mathbf{S} - \boldsymbol{\Sigma}$. The estimation consists in finding the parameters of the model that will replicate as closely as possible the observed covariance matrix in Eq. (4.10). For the maximum likelihood estimation, the comparison of the matrices $\mathbf{S}$ and $\boldsymbol{\Sigma}$ is made through the following expression:

$$F = \mathrm{Ln}|\boldsymbol{\Sigma}| + \mathrm{tr}\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\right) - \mathrm{Ln}|\mathbf{S}| - (q) \tag{4.15}$$

This expression follows directly from the maximization of the likelihood function. Indeed, based on the multivariate normal distribution of the data matrix $\underset{N\times q}{\mathbf{X}^{d}}$, which has been mean centered, the sampling distribution is

$$f(\mathbf{X}) = \prod_{i=1}^{N} (2\pi)^{-\frac{q}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{x}_i^{d'} \boldsymbol{\Sigma}^{-1} \mathbf{x}_i^{d} \right\} \tag{4.16}$$

which is also the likelihood

$$\boldsymbol{\ell} = \boldsymbol{\ell}\big(\text{parameters of } \boldsymbol{\Sigma}|\mathbf{X}\big) = \prod_{i=1}^{N} (2\pi)^{-\frac{q}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{x}_i^{d'} \boldsymbol{\Sigma}^{-1} \mathbf{x}_i^{d} \right\} \tag{4.17}$$

or

$$\begin{aligned}
\mathbf{L} &= Ln\boldsymbol{\ell} = \sum_{i=1}^{N} \left[-\frac{q}{2} Ln(2\pi) - \frac{1}{2} Ln|\boldsymbol{\Sigma}| - \frac{1}{2} \mathbf{x}_i^{d'} \boldsymbol{\Sigma}^{-1} \mathbf{x}_i^{d} \right] \\
&= -\frac{Nq}{2} Ln(2\pi) - \frac{N}{2} Ln|\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{N} \left(\mathbf{x}_i^{d'} \boldsymbol{\Sigma}^{-1} \mathbf{x}_i^{d} \right) \\
&= -\frac{N}{2} \left[qLn(2\pi) + Ln|\boldsymbol{\Sigma}| + \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}_i^{d'} \boldsymbol{\Sigma}^{-1} \mathbf{x}_i^{d} \right) \right] \\
&= -\frac{N}{2} \left[qLn(2\pi) + Ln|\boldsymbol{\Sigma}| + \frac{1}{N} tr\left(\mathbf{X}^{d'} \boldsymbol{\Sigma}^{-1} \mathbf{X}^{d} \right) \right] \\
&= -\frac{N}{2} \left[qLn(2\pi) + Ln|\boldsymbol{\Sigma}| + \frac{1}{N} tr\left(\mathbf{X}^{d'} \mathbf{X}^{d} \boldsymbol{\Sigma}^{-1} \right) \right]
\end{aligned} \tag{4.18}$$

$$\mathbf{L} = -\frac{N}{2} \left[qLn(2\pi) + Ln|\boldsymbol{\Sigma}| + tr\left(\mathbf{S}\boldsymbol{\Sigma}^{-1} \right) \right] \tag{4.19}$$

Therefore, given that the constant terms do not impact the function to maximize, the maximization of the likelihood function corresponds to minimizing the expression in Eq. (4.15). Note that the last terms of Eq. (4.15), i.e., $-\text{Ln}|\mathbf{S}| - (q)$, are constant terms.

The expression F is minimized by searching over the values for each of the parameters. If the observed variables $\mathbf{x}$ are distributed as a multivariate normal distribution, the parameter estimates that minimize Eq. (4.15) are the maximum likelihood estimates.

There are ½$(q)(q + 1)$ distinct elements that constitute the data; this comes from half of the symmetric matrix to which one needs to add back half of the diagonal in order to count the variances of the variables themselves (i.e., $\left[(q)\text{x}(q)/2 + \frac{q}{2}\right]$). Consequently, the number of degrees of freedom corresponds to the number of distinct data points as defined above minus the number of parameters in the model to estimate.

In the example shown in Fig. 4.1, ten parameters must be estimated:

$$5\ \lambda_{ij}'s + 5\ \delta_{ii}'s$$

These correspond to each of the arrows in the figure, i.e., the factor loadings and the variances of the measurement errors. There would be 11 parameters to estimate if the two factors were correlated.

4.2.1 Model Fit

The measure of the fit of the model to the data corresponds to the criterion that was minimized, i.e., a measure of the extent to which the model, given the best possible values of the parameters, can lead to a covariance matrix of the observed variables that is sufficiently similar to the actually observed covariance matrix. We first present and discuss the basic chi-square test of the fit of the model. We then introduce a number of measures of fit that are typically reported and that alleviate the problems inherent to the chi-square test. Finally, we discuss how modification indices can be used as diagnostics for model improvement.

4.2.1.1 Chi-Square Tests

Based on large-sample distribution theory, $\nu = (N-1)\hat{F}$ (where N is the sample size used to generate the covariance matrix of the observed variables and $\hat{F}$ is the minimum value of the expression F as defined by Eq. (4.15)) is distributed as a chi-square with the number of degrees of freedom corresponding to the number of data points minus the number of estimated parameters. If the value of v is significantly greater than zero, the model is rejected; this means that the theoretical model is unable to generate data with a covariance matrix close enough to the one obtained from the actual data.

The chi-square distribution of ν follows from the normal distribution assumption of the data. As discussed above, the likelihood function at its maximum value ($\mathbf{L}$) can be compared with $\mathbf{L}_0$, the likelihood of the full or saturated model with zero degrees of freedom. Such a saturated model reproduces the covariance matrix perfectly so that $\Sigma = \mathbf{S}$ and $tr(\mathbf{S}\boldsymbol{\Sigma}^{-1}) = tr(\mathbf{I}) = q$. Consequently,

$$\mathbf{L}_0 = -\frac{N}{2}[qLn(2\pi) + Ln|\mathbf{S}| + q] \tag{4.20}$$

The likelihood ratio test is

$$-2[\mathbf{L} - \mathbf{L}_0] \sim \chi^2_{df=[q(q+1)/2]-T} \tag{4.21}$$

where T is the number of parameters estimated.

Equation (4.21) results in the expression

$$N\left[\text{Ln}|\boldsymbol{\Sigma}| + \text{tr}\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\right) - \text{Ln}|\mathbf{S}| - (q)\right] \tag{4.22}$$

which is distributed as a chi-square with $[q(q + 1)/2] - T$ degrees of freedom.

It should be noted that it is possible to compare nested models. Indeed, the test of a restriction of a subset of the parameters implies the comparison of two of the measures of fit v, each distributed as a chi-square. Consequently, the difference between the value v_r of a restricted model and v_u of the unrestricted model follows a chi-square distribution with a number of degrees of freedom corresponding to the number of restrictions.

One problem with the expression v (or Eq. (4.22)) is that it contains N, the sample size. This means that as the sample size increases, it becomes less likely that the researcher will fail to reject the model. This is why several other measures of fit have been developed. They are discussed below. While this sample-size effect corresponds to the statistical power of a test consisting in rejecting a null hypothesis that a parameter is equal to zero, it is an issue in this context because the hypothesis for which the researcher would like to get support is the null hypothesis that there is no difference between the observed covariance matrix and the matrix that can be generated by the model. Failure to reject the hypothesis, and thus "accepting" the model, can, therefore, be due to the lack of power of the test. A small enough sample size can contribute to finding "fitting" models based on chi-square tests. It follows that it is more difficult to find fitting models when the sample size is large.

4.2.1.2 Other Goodness-of-Fit Measures

The LISREL output gives a goodness-of-fit index (GFI) that is a direct measure of the fit between the theoretical and observed covariance matrices following from the fit criterion of Eq. (4.15). This GFI is defined as

$$\text{GFI} = 1 - \frac{tr\left[\left(\hat{\Sigma}^{-1}\mathbf{S} - \mathbf{I}\right)^2\right]}{tr\left[\left(\hat{\Sigma}^{-1}\mathbf{S}\right)^2\right]} \tag{4.23}$$

From this equation, it is clear that if the estimated and the observed variances are identical, the numerator of the expression subtracted from 1 is 0 and, therefore, GFI = 1. To correct for the fact that the GFI is affected by the number of indicators, an adjusted goodness-of-fit index (AGFI) is also proposed. This measure of fit

corrects the GFI for the degrees of freedom, just as an adjusted R-squared would in a regression context:

$$\text{AGFI} = 1 - \left[\frac{(q)(q+1)}{(q)(q+1) - 2T}\right][1 - \text{GFI}] \tag{4.24}$$

where T is the number of estimated parameters.

As the number of estimated parameters increases, holding everything else constant, the adjusted GFI decreases.

A threshold value of 0.9 (for either the GFI or AGFI) has become a norm for the acceptability of the model fit (Bagozzi & Yi, 1988; Baumgartner & Homburg, 1996; Kuester, Homburg, & Robertson, 1999).

Another index that is often used to assess model fit is the root mean square error of approximation (RMSEA). It is defined as a function of the minimum fit function corrected by the degrees of freedom and the sample size:

$$\text{RMSEA} = \sqrt{\frac{\hat{F}_0}{d}} \tag{4.25}$$

where

$$\hat{F}_0 = \text{Max}\left\{\left(\hat{F} - [d/(N-1)]\right), 0\right\} \tag{4.26}$$

$$d = [q(q+1)/2] - T \tag{4.27}$$

A value of RMSEA smaller than 0.08 is considered to reflect reasonable errors of approximation, while a value of 0.05 indicates a close fit.

4.2.1.3 Modification Indices

The solution obtained for the parameter estimates uses the derivatives of the objective function relative to each parameter. This means that for a given solution, it is possible to know the direction in which a parameter should change in order to improve the fit and how steeply it should change. As a result, the modification indices indicate the expected gains in fit that would be obtained if a particular coefficient should become unconstrained (holding all other parameters fixed at their estimated value). Although not a substitute for the theory that leads to the model specification, this modification index can be useful in analyzing structural relationships and in particular in refining the correlational assumptions of random terms and for modeling control factors.

4.2.2 Test of Significance of Model Parameters

Because of the maximum likelihood properties of the estimates, which follow from the normal distribution assumption of the variables, the significance of each parameter can be tested using the standard t statistics formed by the ratio of the parameter estimate and its standard deviation.

4.2.3 Factor Scores

Similar to the process described in Chap. 3 for EFA, factor scores can be computed using the equation

$$\underset{N\times p}{\widetilde{\mathbf{Y}}} = \underset{N\times p}{\widetilde{\mathbf{X}}}\ \underset{p\times p}{\mathbf{R}^{-1}}\ \underset{p\times p}{\mathbf{L}} \tag{4.28}$$

In contrast to the case of EFA, however, zeros appear in the matrix of factor loadings. In addition, it should be noted that when multiple factors are analyzed simultaneously in a single CFA, the information contained in the correlations with all the variables is used to predict the scores. Therefore, it is not the case that only the variables loading into a factor are used to predict the factor scores. This can easily be seen from the fact that the matrix of "regression" weights $\mathbf{R}^{-1}\mathbf{L}$ uses all the information from the correlation matrix. Only a CFA per factor can provide factor scores determined solely by the items loading on that factor.

4.3 Summary Procedures for Scale Construction

Scale construction involves several procedures that are sequentially applied and that bring together the methods discussed in Chap. 3 with those presented in this chapter. These procedures include the following statistical analyses: EFA, CFA , and reliability coefficient alpha. The CFA technique can also be used to assess the discriminant and convergent validity of a scale. We now review these steps and the corresponding statistical analyses in turn.

4.3.1 Exploratory Factor Analysis

EFA can be performed separately for each hypothesized factor. This demonstrates the unidimensionality of each factor. One global factor analysis can also be performed in order to assess the degree of independence between the factors.

4.3.2 Confirmatory Factor Analysis

CFA can be used to assess the overall fit of the entire measurement model and to obtain the final estimates of the measurement model parameters. Although CFA is sometimes performed on the same sample as the EFA, it is preferable to use a new sample when it is possible to collect more data.

4.3.3 Reliability Coefficient Alpha

In cases where composite scales are developed, the reliability coefficient alpha is a measure of the reliability of the scales. Reliabilities of less than 0.7 for academic research and 0.9 for market research are typically not sufficient to warrant further analyses using these composite scales.

In addition, scale construction involves determining that the new scale developed is different (i.e., reflects and measures a construct that is different) from measures of other related constructs. This is a test of the scale's discriminant validity. It also involves a test of convergent validity, i.e., that this new measure relates to other, yet different, constructs.

4.3.4 Discriminant Validity

A construct must be different from other constructs (discriminant validity) but, at the same time, be mutually conceptually related (convergent validity). The discriminant validity of the constructs is assessed by comparing a measurement model where the correlation between the two constructs is estimated with a model where the correlation is constrained to be equal to one (thereby assuming a single-factor structure). The discriminant validity of the constructs is examined for each pair at a time. This procedure, proposed by Bagozzi, Yi, and Phillips (1991), indicates that, if the model where the correlation is not equal to one significantly improves the fit, then the two constructs are distinct from each other, although it is possible for them to be significantly correlated.

4.3.5 Convergent Validity

Convergent validity concerns the verification that some constructs thought to be conceptually and/or structurally related exhibit significant correlations among themselves. The convergent validity of the constructs is assessed by comparing a measurement model where the correlation between the two constructs is estimated

with a model where the correlation is constrained to be equal to zero. A significant improvement in fit indicates that the two constructs are indeed related, which confirms convergence validity. Combining the two tests (that the correlation is different from one and different from zero) demonstrates that the two constructs are different (discriminant validity), although related with a significantly different from zero correlation (convergent validity).

4.4 Second-Order Confirmatory Factor Analysis

In the second-order factor model, there are two levels of constructs. At the first level, constructs are measured through observable variables. These constructs are not independent and, in fact, their correlation is hypothesized to follow from the fact that they are themselves reflective of common second-order unobserved constructs of a higher conceptual level. This can be represented as in Fig. 4.2.

The relationships displayed in Fig. 4.2 can be expressed algebraically by the following equations:

$$\underset{p\times 1}{\mathbf{y}} = \underset{p\times m}{\boldsymbol{\Lambda}}\ \underset{m\times 1}{\boldsymbol{\eta}} + \underset{p\times 1}{\boldsymbol{\varepsilon}} \tag{4.29}$$

and

$$\underset{m\times 1}{\boldsymbol{\eta}} = \underset{m\times n}{\boldsymbol{\Gamma}}\ \underset{n\times 1}{\boldsymbol{\xi}} + \underset{m\times 1}{\boldsymbol{\zeta}} \tag{4.30}$$

Equation (4.29) expresses the first-order factor analytic model. The unobserved constructs $\boldsymbol{\eta}$ are the first-order factors; they are measured by the reflective items

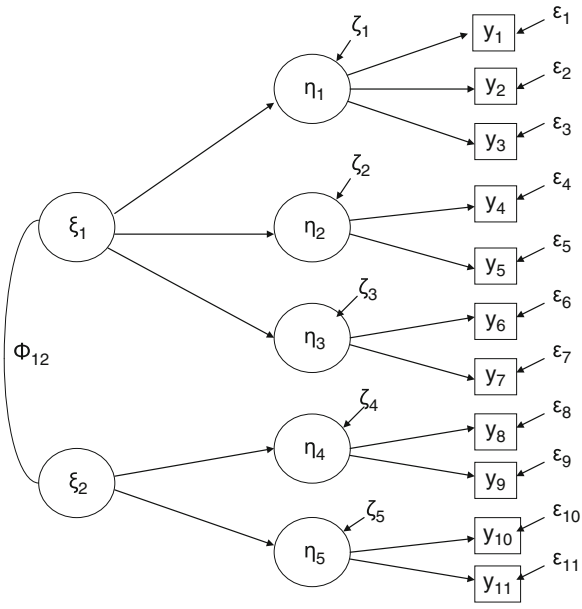


Fig. 4.2 Graphical representation of a second-order factor analytic model

represented by the variables $\mathbf{y}$. Equation (4.30) shows that the constructs $\boldsymbol{\eta}$ are derived from the second-order factors $\boldsymbol{\xi}$. The factor loadings corresponding, respectively, to the first-order and second-order factor models are the elements of matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$. Finally, the errors in measurement are represented by the vectors $\boldsymbol{\varepsilon}$ and $\boldsymbol{\zeta}$.

In addition to the structure expressed by these two equations, we use the following notation of the covariances:

$$\mathrm{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}'\right] = \underset{n\times n}{\boldsymbol{\Phi}} \tag{4.31}$$

$$\mathrm{E}\left[\boldsymbol{\zeta}\boldsymbol{\zeta}'\right] = \underset{m\times m}{\boldsymbol{\Psi}} \tag{4.32}$$

and

$$\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] = \underset{p\times p}{\boldsymbol{\Theta}_{\varepsilon}} \tag{4.33}$$

Furthermore, we assume that the elements of $\boldsymbol{\zeta}$ are uncorrelated to the elements of $\boldsymbol{\xi}$, and similarly that the elements of $\boldsymbol{\varepsilon}$ are uncorrelated to the elements of $\boldsymbol{\eta}$.

If the second-order factor model described by the equations above is correct, the covariance matrix of the observed variables $\mathbf{y}$ must have a particular structure. This structure is obtained as

$$\mathrm{E}\left[\mathbf{y}\mathbf{y}'\right] = \mathrm{E}\left[(\boldsymbol{\Lambda}\boldsymbol{\eta}+\boldsymbol{\varepsilon})(\boldsymbol{\Lambda}\boldsymbol{\eta}+\boldsymbol{\varepsilon})'\right] \tag{4.34}$$

If we develop

$$\mathrm{E}\left[\mathbf{y}\mathbf{y}'\right] = \boldsymbol{\Lambda}\mathrm{E}\left[\boldsymbol{\eta}\boldsymbol{\eta}'\right]\boldsymbol{\Lambda}' + \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \tag{4.35}$$

replacing $\boldsymbol{\eta}$ by its value expressed in Eq. (4.30)

$$\mathrm{E}\left[\mathbf{y}\mathbf{y}'\right] = \boldsymbol{\Lambda}\mathrm{E}\left[(\boldsymbol{\Gamma}\boldsymbol{\xi}+\boldsymbol{\zeta})(\boldsymbol{\Gamma}\boldsymbol{\xi}+\boldsymbol{\zeta})'\right]\boldsymbol{\Lambda}' + \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \tag{4.36}$$

$$\mathrm{E}\left[\mathbf{y}\mathbf{y}'\right] = \boldsymbol{\Lambda}\left(\boldsymbol{\Gamma}\mathrm{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}'\right]\boldsymbol{\Gamma}' + \mathrm{E}\left[\boldsymbol{\zeta}\boldsymbol{\zeta}'\right]\right)\boldsymbol{\Lambda}' + \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] \tag{4.37}$$

$$\mathrm{E}\left[\mathbf{y}\mathbf{y}'\right] = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\left(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}\right)\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_{\varepsilon} \tag{4.38}$$

where the elements on the right side of Eq. (4.38) are model parameters to be estimated such that their values combined in that matrix structure reproduce as closely as possible the observed covariance matrix $\mathbf{S}$ calculated from the sample data.

The estimation procedure follows the same principle as described above for the simple confirmatory factor analytic model. The number of parameters is, however, different.

How many parameters need to be estimated?

We typically define the covariance matrices $\mathbf{\Phi}$, $\boldsymbol{\psi}$, and $\mathbf{\Theta}_{\varepsilon}$ to be diagonal. Therefore, these correspond to $n + m + p$ parameters to be estimated, to which we would need to add the factor-loading parameters contained in matrices $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$. Taking the example in Fig. 4.2, $n = 2$, $m = 5$, and $p = 11$. One of the factor loadings for each first-order factor should be set to 1 to define the units of measurement of these factors. Consequently, $\mathbf{\Lambda}$ contains $11 - 5 = 6$ parameters to be estimated and $\mathbf{\Gamma}$ contains five parameters to be estimated. That gives a total of $2 + 5 + 11 + 6 + 5 = 29$ parameters to be estimated. Given that the sample data covariance matrix (an 11 by 11 matrix) contains $(11 \times 12)/2 = 66$ data points, the degrees of freedom are $66 - 29 = 37$.

The same measures of fit as described above for CFA are used to assess the appropriateness of the structure imposed on the data.

4.5 Multi-Group Confirmatory Factor Analysis

Multi-group CFA is appropriate for testing the homogeneity of measurement models across samples. It is particularly useful in the context of cross-national research where measurement instruments may vary due to cultural differences. This corresponds to the notion of measurement invariance. From that point of view, the model described by Eq. (4.2) must be expanded along two dimensions: (1) several sets of parameters must be estimated simultaneously for each of the groups and (2) some differences in the means of the unobserved constructs must be recognized between groups while they are ignored (assumed to be zero) in single-group CFA. These expansions are represented in Eqs. (4.39), (4.40), and (4.41). Equation (4.39) is identical to the single-group confirmatory factor analytic model.

The means of the factors are represented by the vector $\boldsymbol{\kappa}$ in Eq. (4.40), which contains n rows for the mean of each of the n factors. The vector $\boldsymbol{\tau}_x$ in Eq. (4.39) contains q rows for the scalar constant term of each of the q items:

$$\underset{q\times 1}{\mathbf{x}} = \underset{q\times 1}{\boldsymbol{\tau}_x} + \underset{q\times n}{\mathbf{\Lambda}_x}\ \underset{n\times 1}{\boldsymbol{\xi}} + \underset{q\times 1}{\boldsymbol{\delta}} \tag{4.39}$$

$$\mathrm{E}[\boldsymbol{\xi}] = \underset{n\times 1}{\boldsymbol{\kappa}} \tag{4.40}$$

$$\mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}'\right] = \underset{q\times q}{\mathbf{\Theta}_{\delta}} \tag{4.41}$$

Therefore, the means of the observed measures $\mathbf{x}$ are

$$\underset{q\times 1}{\boldsymbol{\mu}_x} = \mathrm{E}[\mathbf{x}] = \underset{q\times 1}{\boldsymbol{\tau}_x} + \underset{q\times n}{\boldsymbol{\Lambda}_x}\,\mathrm{E}\left[\underset{n\times 1}{\boldsymbol{\xi}}\right] = \underset{q\times 1}{\boldsymbol{\tau}_x} + \underset{p\times n}{\boldsymbol{\Lambda}_x}\,\underset{n\times 1}{\boldsymbol{\kappa}} \tag{4.42}$$

This model, with a mean structure such as is imposed in Eq. (4.42), can be estimated if we recognize that the log-likelihood function specified in Eq. (4.19) now contains not only the parameters that determine the covariance matrix $\boldsymbol{\Sigma}$ but also the expected values of the $\mathbf{x}$ variables, so that

$$\mathbf{S} = (\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)' \tag{4.43}$$

Consequently, when modeling the means in addition to the covariance structure, the objective function (the log likelihood) is

$$\mathbf{L} = -\frac{N}{2}\left[qLn(2\pi) + Ln|\boldsymbol{\Sigma}| + tr\left\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)'\boldsymbol{\Sigma}^{-1}\right\}\right] \tag{4.44}$$

We now add a notation to reflect that the model applies to group g with $g = 1,\ldots,G$:

$$\forall g = 1, \ldots G: \quad \underset{q\times 1}{\mathbf{x}^{(g)}} = \underset{q\times 1}{\boldsymbol{\tau}_x{}^{(g)}} + \underset{q\times n}{\boldsymbol{\Lambda}_x{}^{(g)}}\,\underset{n\times 1}{\boldsymbol{\xi}^{(g)}} + \underset{q\times 1}{\boldsymbol{\delta}^{(g)}} \tag{4.45}$$

and

$$\mathrm{E}\left[\boldsymbol{\xi}^{(g)}\right] = \boldsymbol{\kappa}^{(g)} \tag{4.46}$$

For identification, one of the groups must serve as a reference with the means of its factors centered at zero (the same requirement as for a single-group CFA). Usually group 1 serves as that reference, although in principle it can be any group:

$$\boldsymbol{\kappa}^{(1)} = 0 \tag{4.47}$$

It is also necessary to fix one factor loading for each factor in Λ_x to define the measurement unit of the unobserved constructs.

The estimation is again based on the maximum likelihood. The log likelihood is the sum of the log likelihoods for all the groups so that we now search for the values of the parameters that maximize

$$\mathbf{L} = -\frac{1}{2}\sum_{g=1}^{G} N^{(g)}\left[q^{(g)}Ln(2\pi) + Ln\left|\boldsymbol{\Sigma}^{(g)}\right| + tr\left\{\left(\mathbf{X}^{(g)} - \boldsymbol{\mu}_x{}^{(g)}\right)\left(\mathbf{X}^{(g)} - \boldsymbol{\mu}_x{}^{(g)}\right)'\boldsymbol{\Sigma}^{(g)^{-1}}\right\}\right] \tag{4.48}$$

It is then possible to impose equality constraints on the parameters to be estimated by defining them as invariant across groups. Different types of invariance can be imposed and tested.

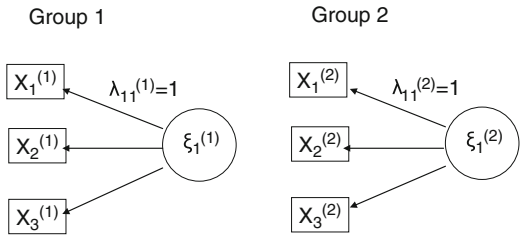


Fig. 4.3 Graphical representation of two-group confirmatory factor analysis

Metric invariance concerns the constraint of equality of factor loadings across groups:

$$\mathbf{\Lambda}_x^{(g)} = \mathbf{\Lambda}_x^{(g')} = \mathbf{\Lambda}_x \tag{4.49}$$

Scalar invariance restricts the scalar constants to be identical across groups:

$$\boldsymbol{\tau}_x^{(g)} = \boldsymbol{\tau}_x^{(g')} = \boldsymbol{\tau}_x \tag{4.50}$$

In order to illustrate the types of restrictions that need to be imposed, let us consider the example of two groups, as depicted in Fig. 4.3.

For the first item of the first group, the measurement model is

$$x_1^{(1)} = \tau_1 + \xi_1^{(1)} + \delta_1^{(1)} \tag{4.51}$$

with

$$\kappa_1^{(1)} = 0 \tag{4.52}$$

This means that the latent construct ${\xi_1}^{(1)}$ is measured in the units of ${x_1}^{(1)}$.

For identification, constraining τ_1 to be equal across groups is the same as estimating it in one group and fixing the value in the other groups to be equal across groups. For the first item of the second group, the measurement model is

$$x_1^{(2)} = \tau_1 + \xi_1^{(2)} + \delta_1^{(2)} \tag{4.53}$$

Even though the mean of ${\xi_1}^{(2)}$ can be different from ${\xi_1}^{(1)}$, the measurement units are fixed to be the units of ${x_1}^{(1)}$.

For the model to have different factor means $\boldsymbol{\kappa}$ that are meaningful, the following conditions must be met:

1. Metric invariance, i.e., the same factor loadings $\mathbf{\Lambda}_x$ across groups
2. Scalar invariance, i.e., the same constant for the scale of each item $\boldsymbol{\tau}_x$ across groups

These issues are particularly relevant in cross-cultural research where measurement instruments must be comparable across cultures/countries and especially when the factor means are of interest to the research.

Factor scores can also be computed as discussed earlier in Sect. 4.2.3. In the case of different means (which require scalar invariance properties), the factor scores are computed to reflect these differences in the distribution of the latent constructs.

4.6 Application Examples

We now present examples of the various methods discussed in this chapter using LISREL and STATA (with a quick example using AMOS). First we provide examples of CFA (Sect. 4.6.1). Next we give examples of discriminant validity tests (Sect. 4.6.2) and of convergent validity tests (Sect. 4.6.3) that demonstrate the estimation of a single-factor analytic structure and the estimation of a factor analytic structure with two correlated factors. Then we show examples of second-order factor analysis (Sect. 4.6.4), and finally, we illustrate multi-group factor analysis (Sect. 4.6.5).

4.6.1 Example of Confirmatory Factor Analysis

The example in Fig. 4.4 shows the input file in LISREL.

An exclamation mark indicates that what follows is a comment and is not part of the LISREL commands. Therefore, the first real input line in Fig. 4.4 starts with DA, which stands for data. On that line, NI indicates the number of input (observed) variables (six in this example), and MA = KM indicates the type of matrix to be modeled, KM for correlation or CV for covariance.

The second line of input is used to specify how to read the data. RA indicates that the raw data will be read (from which the correlation matrix will be automatically computed), and FI = *filename* indicates the name of the file containing those data, where *filename* is the Windows file name including the full path.

The third line, with LA, indicates that next come the labels of the indicator (input) variables. These are shown as Q5, Q7, etc., on the following line.

The next line specifies the model, as indicated by the code MO at the beginning of that line. NX indicates the number of indicators corresponding to the exogenous constructs (here, there are six). NK stands for the number of ksi constructs (we have a unique factor in this example). PH = ST indicates that the covariance matrix phi is specified here as a standardized matrix, i.e., a correlation matrix with 1s in the diagonal and 0s off-diagonal. The covariance matrix of the measurement model

```
!Examp4-1.spl
!Raw Data From File: Examp4-1.txt

DA NI=6 MA = KM XM = 9
RA FI=C:\SAMD\Chapter4\Examples\Examp4-1.txt
LA
Q5 Q7 Q8 Q12 Q13 Q14
MO NX = 6 NK = 1 PH = ST TD = SY
LK
FactorOne                                    !The First Factor
FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1) TD(3,2) TD(6,5)
Path Diagram
OU SE TV AD = 50 MI
```

Fig. 4.4 LISREL input example for confirmatory factor analytic model (examp4-1.spl)

error terms, theta delta, is specified as a symmetric matrix (TD = SY). A diagonal matrix (TD = DI) could have presented a simpler model where all covariances are zero. However, our example of a symmetric matrix illustrates how some of these covariance parameters can be estimated.

In the next line, LK stands for the label of the ksi constructs, although there is only one of them in this example. That label ("FactorOne") follows on the next line.

The following line starting with FR is the list of the parameters that are estimated where LX stands for lambda x and TD for theta delta. Each is followed by the row and column of the corresponding matrix, as defined by the model specification in Eqs. (4.2) and (4.4). In the standard factor analytic model, the measurement errors are typically uncorrelated and theta delta is just a diagonal matrix. Occasionally, a better fit is obtained if these correlations are estimated. Modification indices provide the information regarding the extent to which freeing these parameters can lead to a better fit. Nevertheless, caution should be exercised when letting these correlations take values other than zero. This is because correlated measurement errors mean that the items have something in common beyond what is already shared by all items reflecting a factor. It is especially critical to exercise caution when the error terms correspond to items that reflect different factors. Such a case would indicate that two items are used to measure different factors, although they also share common meanings through their residuals. This raises questions about the validity of such measures and about the appropriateness of the choice of items. In the example, the error correlations (identified by "TD(3,2) TD(6,5)") concern the same factor. They are estimated based on a preliminary analysis that indicated, based on the modification indices, that the fit would be improved if these were allowed to be different from zero.

The line "Path Diagram" indicates that a graphical representation of the model is requested.

The last line of the input file describes the output (OU) requested. SE means standard errors, TV their *t*-values, and MI the modification indices.

The input file in STATA is shown in Fig. 4.5.

```
infile Q5 Q7 Q8 Q12 Q13 Q14 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter4_CFA/Examp4-1.txt", clear
replace Q5=. if Q5==9
replace Q7=. if Q7==9
replace Q8=. if Q8==9
replace Q12=. if Q12==9
replace Q13=. if Q13==9
replace Q14=. if Q14==9
egen stQ5 = std(Q5)
egen stQ7 = std(Q7)
egen stQ8 = std(Q8)
egen stQ12 = std(Q12)
egen stQ13 = std(Q13)
egen stQ14 = std(Q14)
sem (FactorOne -> stQ5) (FactorOne -> stQ7) (FactorOne -> stQ8) (FactorOne -> stQ12)
(FactorOne -> stQ13) (FactorOne -> stQ14) ///
, cov(e.stQ7*e.stQ8) cov(e.stQ13*e.stQ14) ///
var(FactorOne@1) ///
nomeans latent(FactorOne )
estat gof, stats(all)
estat mindices, min(1)
estat framework, fitted
predict FactorScore, latent
```

Fig. 4.5 STATA input example for confirmatory factor analytic model (examp4-1_Mac.do)

The command "sem" signals input for structural equation models. Each relationship is represented by two variables separated by an arrow ("->" or "<-"), which indicates the causality or directionality. The variables are either observable measures (e.g., "stQ5") or latent variables (e.g., "FactorOne"). In this particular example, which is identical to the model described above to be estimated with LISREL (Fig. 4.4), the covariances to be estimated are indicated after a "," by means of the "cov(e.stQ7*e.stQ8)" option. The term "cov" stands for covariance and "e.var" for the error of the "var" variable. As we discussed when presenting the LISREL input, you would only request the estimation of these covariances of measurement errors ex post and if necessary based on the information provided in the modification indices.

Furthermore, "var(FactorOne@1)" indicates that the latent factor variance should be constrained to 1. The next three lines request statistics such as goodness-of-fit measures or modification indices. The last line ("predict FactorScore, latent") is optional. It computes the factor scores of the latent variable(s) and appends the scores in new variables in the data set. In this case, only one new variable name is given ("FactorScore") because the analysis specifies a single factor. If more than one factor were involved, the list of the names to be used would follow the "predict" command. The modified data set can then be saved as a ".dta" file for further analysis using separate do-files.

The LISREL output of such a model is given in Fig. 4.6 and the output from STATA follows.

```
                                L I S R E L  8.30
                                      BY

                             Karl G. Jöreskog & Dag Sörbom

                        This program is published exclusively by
                        Scientific Software International, Inc.
                           7383 N. Lincoln Avenue, Suite 100
                            Chicago, IL 60646-1704, U.S.A.
                    Phone: (800)247-6113, (847)675-0720, Fax: (847)675-2140
               Copyright by Scientific Software International, Inc., 1981-99
                 Use of this program is subject to the terms specified in the
                             Universal Copyright Convention.
                               Website: www.ssicentral.com

 The following lines were read from file C:\SAMD\CHAPTER8\EXAMPLES\EXAMP4-1.SPL:

 !Examp4-1.spl
 !Raw Data From File: Examp4-1.txt

 DA NI=6 MA = KM XM = 9
 RA FI=C:\SAMD\Chapter4\Examples\Examp4-1.txt
 LA
 Q5 Q7 Q8 Q12 Q13 Q14
 MO NX = 6 NK = 1 PH = ST TD = SY
 LK
 FactorOne      !The First Factor
 FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1) TD(3,2) TD(6,5)
 Path Diagram
 OU SE TV AD = 50 MI

 !Examp4-1.spl

                           Number of Input Variables  6
                           Number of Y - Variables    0
                           Number of X - Variables    6
                           Number of ETA - Variables  0
                           Number of KSI - Variables  1
                           Number of Observations   138

         Covariance Matrix to be Analyzed

                 Q5          Q7          Q8         Q12         Q13         Q14
            --------    --------    --------    --------    --------    --------
       Q5       1.00
       Q7       0.47        1.00
       Q8       0.58        0.75        1.00
      Q12       0.55        0.60        0.65        1.00
      Q13       0.44        0.40        0.51        0.50        1.00
      Q14       0.39        0.44        0.57        0.55        0.59        1.00


Parameter Specifications

         LAMBDA-X

            FactorOn
            --------
       Q5          1
       Q7          2
       Q8          3
      Q12          4
      Q13          5
      Q14          6

         THETA-DELTA

                 Q5          Q7          Q8         Q12         Q13         Q14
            --------    --------    --------    --------    --------    --------
       Q5          7
       Q7          0           8
       Q8          0           9          10
      Q12          0           0           0          11
      Q13          0           0           0           0          12
```

Fig. 4.6 LISREL for Windows output example for confirmatory factor analytic model (examp4-1.out)

```
Number of Iterations =  7

 LISREL Estimates (Maximum Likelihood)

        LAMBDA-X

           FactorOn
           --------
      Q5       0.68
              (0.08)
                8.45

      Q7       0.71
              (0.08)
                8.69

      Q8       0.83
              (0.08)
               11.01

     Q12       0.81
              (0.08)
               10.64

     Q13       0.62
              (0.08)
                7.46

     Q14       0.66
              (0.08)
                8.07


        PHI

           FactorOn
           --------
              1.00


        THETA-DELTA

                Q5         Q7         Q8        Q12        Q13        Q14
           --------   --------   --------   --------   --------   --------
      Q5       0.54
              (0.08)
                7.09

      Q7        - -       0.50
                         (0.08)
                           6.44

      Q8        - -       0.16       0.31
                        (0.06)     (0.06)
                           2.81       4.99

     Q12        - -        - -        - -       0.35
                                     (0.06)
                                       5.54

     Q13        - -        - -        - -        - -       0.62
                                                        (0.08)
                                                          7.36

     Q14        - -        - -        - -        - -       0.18       0.57
                                                        (0.06)     (0.08)
                                                          2.89       7.17


        Squared Multiple Correlations for X - Variables

                Q5         Q7         Q8        Q12        Q13        Q14
           --------   --------   --------   --------   --------   --------
              0.46       0.50       0.69       0.65       0.38       0.43
```

Fig. 4.6 (continued)

```
                    Goodness of Fit Statistics

                        Degrees of Freedom = 7
              Minimum Fit Function Chi-Square = 6.61 (P = 0.47)
    Normal Theory Weighted Least Squares Chi-Square = 6.27 (P = 0.51)
              Estimated Non-centrality Parameter (NCP) = 0.0
          90 Percent Confidence Interval for NCP = (0.0 ; 9.27)

                    Minimum Fit Function Value = 0.048
               Population Discrepancy Function Value (F0) = 0.0
          90 Percent Confidence Interval for F0 = (0.0 ; 0.068)
          Root Mean Square Error of Approximation (RMSEA) = 0.0
          90 Percent Confidence Interval for RMSEA = (0.0 ; 0.098)
            P-Value for Test of Close Fit (RMSEA < 0.05) = 0.71

              Expected Cross-Validation Index (ECVI) = 0.26
          90 Percent Confidence Interval for ECVI = (0.26 ; 0.32)
                    ECVI for Saturated Model = 0.31
                ECVI for Independence Model = 3.02

  Chi-Square for Independence Model with 15 Degrees of Freedom = 402.09
                        Independence AIC = 414.09
                            Model AIC = 34.27
                          Saturated AIC = 42.00
                        Independence CAIC = 437.65
                            Model CAIC = 89.26
                          Saturated CAIC = 124.47

                Root Mean Square Residual (RMR) = 0.020
                      Standardized RMR = 0.020
                Goodness of Fit Index (GFI) = 0.98
            Adjusted Goodness of Fit Index (AGFI) = 0.95
            Parsimony Goodness of Fit Index (PGFI) = 0.33

                    Normed Fit Index (NFI) = 0.98
                  Non-Normed Fit Index (NNFI) = 1.00
              Parsimony Normed Fit Index (PNFI) = 0.46
                 Comparative Fit Index (CFI) = 1.00
                 Incremental Fit Index (IFI) = 1.00
                  Relative Fit Index (RFI) = 0.96

                        Critical N (CN) = 383.87


Modification Indices and Expected Change

No Non-Zero Modification Indices for LAMBDA-X

No Non-Zero Modification Indices for PHI

        Modification Indices for THETA-DELTA

              Q5        Q7        Q8        Q12       Q13       Q14
        --------  --------  --------  --------  --------  --------
  Q5        - -
  Q7       0.50       - -
  Q8       1.00       - -       - -
 Q12       0.00      3.20      3.82       - -
 Q13       0.96      0.43      0.00      0.00       - -
 Q14       2.38      0.54      1.23      0.33       - -       - -

        Expected Change for THETA-DELTA

              Q5        Q7        Q8        Q12       Q13       Q14
        --------  --------  --------  --------  --------  --------
  Q5        - -
  Q7      -0.03       - -
  Q8       0.04       - -       - -
 Q12       0.00      0.08     -0.09       - -
 Q13       0.05     -0.03      0.00      0.00       - -
 Q14      -0.08     -0.03      0.04      0.03       - -       - -

Maximum Modification Index is    3.82 for Element ( 4, 3) of THETA-DELTA
```

Fig. 4.6 (continued)

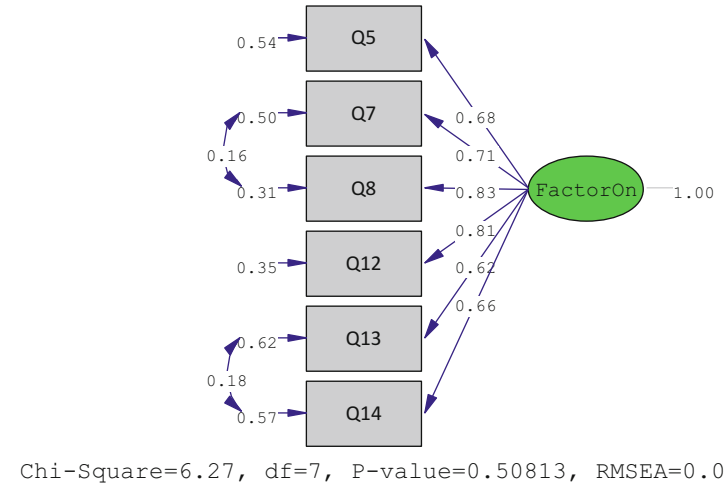


Fig. 4.7 Path diagram of confirmatory factor analytic model from LISREL (examp4-1.pth)

In the output, as shown in Fig. 4.6, after listing the commands described earlier according to the model specified in the corresponding input file, the observed covariance matrix (in this case a correlation matrix) to be modeled is printed.

The "Parameter Specifications" section indicates the list and number of parameters to be estimated, with a detail of all the matrices containing the parameters. The value zero indicates that the corresponding parameter is fixed and is not to be estimated. Unless specified otherwise, the default value of these fixed parameters is set to zero.

The number of iterations shows the number that was necessary to obtain convergence and the parameter estimates follow. Below each parameter estimate value, its standard error is shown in parentheses and the *t*-value below it.

Then follow the goodness-of-fit statistics, among which those described earlier can be found. The example run in Fig. 4.6 shows that the single-factor model represents well the observed correlation matrix since the chi-square is not statistically significant and the GFI is high with a value of 0.98 (highlighted in grey in the figure).

The modification indices are reasonably small, which indicates that freeing additional parameters would not lead to a big gain in fit.

The diagram of such a confirmatory factor analytic model is shown in Fig. 4.7.

The STATA output follows in Fig. 4.8.

4.6.2 *Example of Model to Test Discriminant Validity Between Two Constructs*

The following example (illustrated with LISREL and STATA) is typical of an analysis where the goal is to assess the validity of a construct. Figure 4.9 shows the input file to estimate a two-factor model (such analyses are usually performed two factors at a time because the modeling of all the factors at once typically involves

```
. infile Q5 Q7 Q8 Q12 Q13 Q14 using "C:\DATA\WORK_STATA\SAMD\Chapter4_CFA\Examp4-1.txt", clear
(146 observations read)

. replace Q5=. if Q5==9

(5 real changes made, 5 to missing)

. replace Q7=. if Q7==9

(6 real changes made, 6 to missing)

. replace Q8=. if Q8==9

(6 real changes made, 6 to missing)

. replace Q12=. if Q12==9

(5 real changes made, 5 to missing)

. replace Q13=. if Q13==9

(5 real changes made, 5 to missing)

. replace Q14=. if Q14==9

(6 real changes made, 6 to missing)

. egen stQ5 = std(Q5)

(5 missing values generated)

. egen stQ7 = std(Q7)

(6 missing values generated)

. egen stQ8 = std(Q8)

(6 missing values generated)

. egen stQ12 = std(Q12)

(5 missing values generated)

. egen stQ13 = std(Q13)

(5 missing values generated)

. egen stQ14 = std(Q14)

(6 missing values generated)

. sem (FactorOne -> stQ5) (FactorOne -> stQ7) (FactorOne -> stQ8) (FactorOne -> stQ12) (FactorOne

> -> stQ13) (FactorOne -> stQ14) ///

> , cov(e.stQ7*e.stQ8) cov(e.stQ13*e.stQ14) ///

> var(FactorOne@1) ///

> nomeans latent(FactorOne )

(8 observations with missing values excluded;

 specify option 'method(mlmv)' to use all observations)
```

Fig. 4.8 STATA output example for confirmatory factor analytic model (examp4-1.log)

```
Endogenous variables

Measurement:  stQ5 stQ7 stQ8 stQ12 stQ13 stQ14

Exogenous variables

Latent:       FactorOne

Fitting target model:

Iteration 0:   log likelihood = -989.65934
Iteration 1:   log likelihood = -979.43554
Iteration 2:   log likelihood = -976.06586
Iteration 3:   log likelihood = -976.01067
Iteration 4:   log likelihood = -976.01056
Iteration 5:   log likelihood = -976.01056


Structural equation model                       Number of obs      =       138

Estimation method  = ml

Log likelihood     = -976.01056

 ( 1)  [var(FactorOne)]_cons = 1

------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Measurement  |
  stQ5 <-    |
    FactorOne |   .6788884   .0800289     8.48   0.000     .5220346    .8357422
  -----------+----------------------------------------------------------------
  stQ7 <-    |
    FactorOne |   .7039627   .0806121     8.73   0.000     .5459659    .8619595
  -----------+----------------------------------------------------------------
  stQ8 <-    |
    FactorOne |   .8316671   .0755962    11.00   0.000     .6835012    .9798331
  -----------+----------------------------------------------------------------
  stQ12 <-   |
    FactorOne |   .8093582   .0762004    10.62   0.000     .6600081    .9587083
  -----------+----------------------------------------------------------------
  stQ13 <-   |
    FactorOne |   .6201457   .0829863     7.47   0.000     .4574955     .782796
  -----------+----------------------------------------------------------------
  stQ14 <-   |
    FactorOne |   .6586628    .081367     8.09   0.000     .4991864    .8181391
-------------+----------------------------------------------------------------
Variance     |
       e.stQ5 |   .5384067    .075745                      .4086583    .7093501
       e.stQ7 |   .4919239   .0759154                      .3635278    .6656689
       e.stQ8 |    .312071   .0635017                      .2094339    .4650073
      e.stQ12 |   .3499078   .0642353                      .2441698    .5014359
      e.stQ13 |   .6210431   .0843644                       .475874    .8104973
      e.stQ14 |   .5707536   .0794053                      .4345366    .7496714
   FactorOne |          1  (constrained)
-------------+----------------------------------------------------------------
Covariance   |
  e.stQ7     |
       e.stQ8 |   .1588179   .0563846     2.82   0.005     .0483063    .2693296
  -----------+----------------------------------------------------------------
  e.stQ13    |
      e.stQ14 |   .1802037   .0622552     2.89   0.004     .0581858    .3022217
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(7)   =      6.67, Prob > chi2 = 0.4642
```

Fig. 4.8 (continued)

```
. estat gof, stats(all)

------------------------------------------------------------------------------
Fit statistic        |      Value   Description
---------------------+--------------------------------------------------------
Likelihood ratio     |
          chi2_ms(7) |      6.668   model vs. saturated
            p > chi2 |      0.464
         chi2_bs(15) |    405.180   baseline vs. saturated
            p > chi2 |      0.000
---------------------+--------------------------------------------------------
Population error     |
               RMSEA |      0.000   Root mean squared error of approximation
 90% CI, lower bound |      0.000
         upper bound |      0.102
              pclose |      0.675   Probability RMSEA <= 0.05
---------------------+--------------------------------------------------------
Information criteria |
                 AIC |   1980.021   Akaike's information criterion
                 BIC |   2021.003   Bayesian information criterion
---------------------+--------------------------------------------------------
Baseline comparison  |
                 CFI |      1.000   Comparative fit index
                 TLI |      1.002   Tucker-Lewis index
---------------------+--------------------------------------------------------
Size of residuals    |
                SRMR |      0.020   Standardized root mean squared residual
                  CD |      0.861   Coefficient of determination
------------------------------------------------------------------------------

. estat mindices, min(1)

Modification indices

-------------------------------------------------------------
             |                                     Standard
             |         MI     df   P>MI        EPC        EPC
-------------+-----------------------------------------------
Covariance   |
  e.stQ5     |
     e.stQ8  |      1.006      1   0.32   .0450477   .1098983
    e.stQ14  |      2.402      1   0.12  -.0811566  -.1464011
  -----------+-----------------------------------------------
  e.stQ7     |
    e.stQ12  |      3.220      1   0.07   .0840234   .2025231
  -----------+-----------------------------------------------
  e.stQ8     |
    e.stQ12  |      3.853      1   0.05  -.0949834   -.287438
    e.stQ14  |      1.238      1   0.27    .044602   .1056827
-------------------------------------------------------------
EPC = expected parameter change

. estat framework, fitted

Endogenous variables on endogenous variables

               | observed
          Beta |      stQ5       stQ7       stQ8      stQ12      stQ13      stQ14
    -----------+------------------------------------------------------------------
    observed   |
           stQ5 |         0
           stQ7 |         0          0
           stQ8 |         0          0          0
          stQ12 |         0          0          0          0
          stQ13 |         0          0          0          0          0
          stQ14 |         0          0          0          0          0          0
    ------------------------------------------------------------------------------
```

Fig. 4.8 (continued)

```
Exogenous variables on endogenous variables

                | latent
          Gamma | FactorOne
    ------------+-----------
    observed    |
           stQ5 |  .6788884
           stQ7 |  .7039627
           stQ8 |  .8316671
          stQ12 |  .8093582
          stQ13 |  .6201457
          stQ14 |  .6586628
    ------------------------

Covariances of error variables

                | observed
            Psi |    e.stQ5     e.stQ7     e.stQ8    e.stQ12    e.stQ13    e.stQ14
    ------------+-------------------------------------------------------------------
    observed    |
         e.stQ5 |  .5384067
         e.stQ7 |          0   .4919239
         e.stQ8 |          0   .1588179    .312071
        e.stQ12 |          0          0          0   .3499078
        e.stQ13 |          0          0          0          0   .6210431
        e.stQ14 |          0          0          0          0   .1802037   .5707536
    ---------------------------------------------------------------------------------

Covariances of exogenous variables

                | latent
            Phi | FactorOne
    ------------+-----------
    latent      |
       FactorOne |         1
    ------------------------

Fitted covariances of observed and latent variables

                | observed                                                      | latent
          Sigma |      stQ5       stQ7       stQ8      stQ12      stQ13      stQ14 | FacorOne
    ------------+---------------------------------------------------------------+---------
    observed    |                                                               |
           stQ5 |  .9992962                                                     |
           stQ7 |  .4779121   .9874874                                          |
           stQ8 |  .5646092   .7442806   1.003741                               |
          stQ12 |  .5494639    .569758   .6731166   1.004969                    |
          stQ13 |  .4210097   .4365595   .5157548     .50192   1.005624         |
          stQ14 |  .4471585    .463674   .5477882   .5330941   .5886706    1.00459 |
    ------------+---------------------------------------------------------------+---------
    latent      |                                                               |
       FactorOne |  .6788884   .7039627   .8316671   .8093582   .6201457   .6586628 |        1
    ---------------------------------------------------------------------------------
```

Fig. 4.8 (continued)

```
!Examp4-2.spl
!Raw Data From File: Examp4-2.txt

DA NI=12 MA = KM XM = 9
RA FI=C:\SAMD\Chapter4\Examples\Examp4-2.txt
LA
Q5 Q7 Q8  Q12 Q13 Q14
Q6 Q9 Q10 Q11 Q17 Q18

MO NX = 12 NK = 2 PH = ST TD = SY      !CORR = Free
LK
FactorOne                                          !Competence Destroying
FactorTwo                                          !Competence Enhancing
FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1) C
   LX(7,2) LX(8,2) LX(9,2) LX(10,2) LX(11,2) LX(12,2) C
   TD(3,2) TD(6,5) TD(8,7) TD(10,8) TD(10,7)
Path Diagram
OU SE TV RS MR FS AD = 50 MI
```

Fig. 4.9 LISREL input for model with two factors (examp4-2.spl)

```
infile Q5 Q7 Q8 Q12 Q13 Q14 Q6 Q9 Q10 Q11 Q17 Q18 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter4_CFA/Examp4-2.txt", clear
replace Q5=. if Q5==9
replace Q7=. if Q7==9
replace Q8=. if Q8==9
replace Q12=. if Q12==9
replace Q13=. if Q13==9
replace Q14=. if Q14==9
replace Q6=. if Q6==9
replace Q9=. if Q9==9
replace Q10=. if Q10==9
replace Q11=. if Q11==9
replace Q17=. if Q17==9
replace Q18=. if Q18==9
egen stQ5 = std(Q5)
egen stQ7 = std(Q7)
egen stQ8 = std(Q8)
egen stQ12 = std(Q12)
egen stQ13 = std(Q13)
egen stQ14 = std(Q14)
egen stQ6 = std(Q6)
egen stQ9 = std(Q9)
egen stQ10 = std(Q10)
egen stQ11 = std(Q11)
egen stQ17 = std(Q17)
egen stQ18 = std(Q18)
sem (FactorOne -> stQ5 stQ7 stQ8 stQ12 stQ13 stQ14) ///
(FactorTwo -> stQ6 stQ9 stQ10 stQ11 stQ17 stQ18) ///
, cov(e.stQ7*e.stQ8) cov(e.stQ13*e.stQ14) ///
cov(e.stQ6*e.stQ11) cov(e.stQ9*e.stQ11) cov(e.stQ6*e.stQ9) ///
var(FactorOne@1) var(FactorTwo@1) ///
nomeans latent(FactorOne FactorTwo)
estat gof, stats(all)
estat mindices, min(1)
estat framework, fitted
```

Fig. 4.10 STATA input for model with two factors (examp4-2_Mac.do)

problems too big to obtain satisfactory fits). The commands are identical to those described earlier, except that now two constructs ("FactorOne" and "FactorTwo") are specified.

Commands in STATA that are equivalent to those illustrated above with LISREL are shown in Fig. 4.10.

In this case, each of the variances of the latent factors is set to 1 with the "var (FactorOne@1)" and "var(FactorTwo@1)" commands. In this example, you can also see that the measurement model is defined without repeating the relationships between the factor and each of the items, as was the case in the prior example. Thus here FactorOne is defined only once by using "(FactorOne -> stQ5 stQ7 stQ8 stQ12 stQ13 stQ14)" as a single command. If no specific instructions are given, then the correlation between the two latent variables is estimated.

The output is shown first for LISREL and then for STATA. The LISREL output corresponding to this two-factor confirmatory factor structure is shown in Fig. 4.11. The description of this output is similar to the one described above involving a single factor. The major difference is the estimate of the correlation between the two factors, which is shown to be −0.56 in this particular example. The diagram representing that factor analytic structure is shown in Fig. 4.12.

The STATA output is shown in Fig. 4.13, where the first lines corresponding to data recoding have been deleted.

```
                                   L I S R E L  8.30

                                        BY

                              Karl G. Jöreskog & Dag Sörbom

                        This program is published exclusively by
                        Scientific Software International, Inc.
                           7383 N. Lincoln Avenue, Suite 100
                            Chicago, IL 60646-1704, U.S.A.
                    Phone: (800)247-6113, (847)675-0720, Fax: (847)675-2140
                Copyright by Scientific Software International, Inc., 1981-99
                  Use of this program is subject to the terms specified in the
                             Universal Copyright Convention.
                              Website: www.ssicentral.com

 The following lines were read from file C:\SAMD\CHAPTER8\EXAMPLES\EXAMP4-2.SPL:

 !Examp4-2.spl
!Raw Data From File: Examp4-2.txt

 DA NI=12 MA = KM XM = 9
 RA FI=C:\SAMD\Chapter4\Examples\Examp4-2.txt
 LA
 Q5 Q7 Q8  Q12 Q13 Q14
 Q6 Q9 Q10 Q11 Q17 Q18

 MO NX = 12 NK = 2 PH = ST TD = SY  !CORR = Free
 LK
 FactorOne      !Competence Destroying
 FactorTwo     !Competence Enhancing
 FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1) LX(7,2) LX(8,2) LX(9,2) LX(10,2) C
 LX(11,2) LX(12,2) TD(3,2) TD(6,5) TD(8,7) TD(10,8) TD(10,7)
 Path Diagram
 OU SE TV RS MR FS AD = 50 MI


THETA-DELTA

               Q5         Q7         Q8        Q12        Q13        Q14
           --------   --------   --------   --------   --------   --------
      Q5         14
      Q7          0         15
      Q8          0         16         17
     Q12          0          0          0         18
     Q13          0          0          0          0         19
     Q14          0          0          0          0         20         21
      Q6          0          0          0          0          0          0
      Q9          0          0          0          0          0          0
     Q10          0          0          0          0          0          0
     Q11          0          0          0          0          0          0
     Q17          0          0          0          0          0          0
     Q18          0          0          0          0          0          0

        THETA-DELTA

               Q6         Q9        Q10        Q11        Q17        Q18
           --------   --------   --------   --------   --------   --------
      Q6         22
      Q9         23         24
     Q10          0          0         25
     Q11         26         27          0         28
     Q17          0          0          0          0         29
     Q18          0          0          0          0          0         30
```

Fig. 4.11 LISREL output for model with two factors (examp4-2.out)

```
Number of Iterations = 10

 LISREL Estimates (Maximum Likelihood)

         LAMBDA-X

            FactorOn   FactorTw
            --------   --------
       Q5       0.65        - -
              (0.08)
                7.92

       Q7       0.70        - -
              (0.08)
                8.59

       Q8       0.80        - -
              (0.08)
               10.35

      Q12       0.84        - -
              (0.08)
               11.06

      Q13       0.60        - -
              (0.08)
                7.14

      Q14       0.67        - -
              (0.08)
                8.18

       Q6        - -       0.57
                          (0.09)
                            6.22

       Q9        - -       0.56
                          (0.09)
                            6.12

      Q10        - -       0.65
                          (0.09)
                            7.48

      Q11        - -       0.62
                          (0.09)
                            6.99

      Q17        - -       0.69
                          (0.09)
                            8.01

      Q18        - -       0.69
                          (0.09)
                            8.01


         PHI

            FactorOn   FactorTw
            --------   --------
 FactorOn       1.00

 FactorTw      -0.56       1.00
              (0.08)
               -6.93
```

Fig. 4.11 (continued)

THETA-DELTA

	Q5	Q7	Q8	Q12	Q13	Q14
Q5	0.58 (0.08) 7.19					
Q7	- -	0.51 (0.08) 6.60				
Q8	- -	0.18 (0.06) 3.21	0.36 (0.06) 5.65			
Q12	- -	- -	- -	0.30 (0.06) 5.01		
Q13	- -	- -	- -	- -	0.64 (0.09) 7.35	
Q14	- -	- -	- -	- -	0.19 (0.06) 3.01	0.55 (0.08) 7.04
Q6	- -	- -	- -	- -	- -	- -
Q9	- -	- -	- -	- -	- -	- -
Q10	- -	- -	- -	- -	- -	- -
Q11	- -	- -	- -	- -	- -	- -
Q17	- -	- -	- -	- -	- -	- -
Q18	- -	- -	- -	- -	- -	- -

THETA-DELTA

	Q6	Q9	Q10	Q11	Q17	Q18
Q6	0.68 (0.10) 7.00					
Q9	0.25 (0.08) 3.27	0.69 (0.10) 7.04				
Q10	- -	- -	0.58 (0.09) 6.51			
Q11	0.23 (0.07) 3.13	0.35 (0.08) 4.48	- -	0.61 (0.09) 6.67		
Q17	- -	- -	- -	- -	0.52 (0.09) 6.13	
Q18	- -	- -	- -	- -	- -	0.52 (0.09) 6.12

Fig. 4.11 (continued)

```
        Squared Multiple Correlations for X - Variables

              Q5         Q7         Q8        Q12        Q13        Q14
        --------   --------   --------   --------   --------   --------
            0.42       0.49       0.64       0.70       0.36       0.45

        Squared Multiple Correlations for X - Variables

              Q6         Q9        Q10        Q11        Q17        Q18
        --------   --------   --------   --------   --------   --------
            0.32       0.31       0.42       0.39       0.48       0.48


                        Goodness of Fit Statistics

                          Degrees of Freedom = 48
              Minimum Fit Function Chi-Square = 54.78 (P = 0.23)
    Normal Theory Weighted Least Squares Chi-Square = 55.76 (P = 0.21)
             Estimated Non-centrality Parameter (NCP) = 7.76
          90 Percent Confidence Interval for NCP = (0.0 ; 30.50)

                     Minimum Fit Function Value = 0.41
               Population Discrepancy Function Value (F0) = 0.058
             90 Percent Confidence Interval for F0 = (0.0 ; 0.23)
           Root Mean Square Error of Approximation (RMSEA) = 0.035
           90 Percent Confidence Interval for RMSEA = (0.0 ; 0.069)
             P-Value for Test of Close Fit (RMSEA < 0.05) = 0.73

                Expected Cross-Validation Index (ECVI) = 0.87
           90 Percent Confidence Interval for ECVI = (0.81 ; 1.04)
                      ECVI for Saturated Model = 1.17
                    ECVI for Independence Model = 5.81

     Chi-Square for Independence Model with 66 Degrees of Freedom = 748.31
                           Independence AIC = 772.31
                               Model AIC = 115.76
                             Saturated AIC = 156.00
                           Independence CAIC = 819.08
                              Model CAIC = 232.69
                            Saturated CAIC = 460.03

                  Root Mean Square Residual (RMR) = 0.048
                        Standardized RMR = 0.048
                   Goodness of Fit Index (GFI) = 0.93
                Adjusted Goodness of Fit Index (AGFI) = 0.89
               Parsimony Goodness of Fit Index (PGFI) = 0.58

                       Normed Fit Index (NFI) = 0.93
                    Non-Normed Fit Index (NNFI) = 0.99
                 Parsimony Normed Fit Index (PNFI) = 0.67
                    Comparative Fit Index (CFI) = 0.99
                    Incremental Fit Index (IFI) = 0.99
                     Relative Fit Index (RFI) = 0.90

                          Critical N (CN) = 179.90


Fitted Covariance Matrix

              Q5         Q7         Q8        Q12        Q13        Q14
        --------   --------   --------   --------   --------   --------
      Q5      1.00
      Q7      0.46       1.00
      Q8      0.52       0.74       1.00
     Q12      0.54       0.59       0.67       1.00
     Q13      0.39       0.42       0.48       0.50       1.00
     Q14      0.44       0.47       0.53       0.56       0.59       1.00
      Q6     -0.21      -0.22      -0.25      -0.26      -0.19      -0.21
      Q9     -0.20      -0.22      -0.25      -0.26      -0.19      -0.21
     Q10     -0.24      -0.26      -0.29      -0.30      -0.22      -0.24
     Q11     -0.23      -0.24      -0.28      -0.29      -0.21      -0.23
     Q17     -0.25      -0.27      -0.31      -0.32      -0.23      -0.26
     Q18     -0.25      -0.27      -0.31      -0.32      -0.23      -0.26
```

Fig. 4.11 (continued)

```
         Fitted Covariance Matrix

              Q6        Q9       Q10       Q11       Q17       Q18
        --------  --------  --------  --------  --------  --------
Q6        1.00
Q9        0.56      1.00
Q10       0.37      0.36      1.00
Q11       0.58      0.70      0.40      1.00
Q17       0.39      0.38      0.45      0.43      1.00
Q18       0.39      0.38      0.45      0.43      0.48      1.00

         Fitted Residuals

              Q5        Q7        Q8       Q12       Q13       Q14
        --------  --------  --------  --------  --------  --------
Q5        0.00
Q7        0.00      0.00
Q8        0.05      0.00      0.00
Q12      -0.01      0.01     -0.03      0.00
Q13       0.04     -0.02      0.03     -0.01      0.00
Q14      -0.04     -0.03      0.04      0.00      0.00      0.00
Q6        0.07     -0.05      0.05     -0.09      0.13      0.02
Q9        0.03     -0.04      0.07     -0.12      0.11      0.10
Q10       0.11     -0.01      0.07     -0.09      0.03     -0.01
Q11      -0.04     -0.01      0.05     -0.07      0.02      0.04
Q17       0.06     -0.02     -0.01     -0.02     -0.03     -0.06
Q18       0.05      0.00      0.10     -0.08      0.13      0.04

         Fitted Residuals

              Q6        Q9       Q10       Q11       Q17       Q18
        --------  --------  --------  --------  --------  --------
Q6        0.00
Q9        0.00      0.00
Q10      -0.01     -0.03      0.00
Q11       0.00      0.00      0.00      0.00
Q17      -0.01      0.03     -0.01      0.00      0.00
Q18       0.01      0.00      0.02     -0.01      0.00      0.00

 Summary Statistics for Fitted Residuals

 Smallest Fitted Residual =   -0.12
   Median Fitted Residual =    0.00
  Largest Fitted Residual =    0.13

 …

Modification Indices and Expected Change

         Modification Indices for LAMBDA-X

            FactorOn   FactorTw
            --------   --------
Q5           - -        2.14
Q7           - -        1.44
Q8           - -        4.74
Q12          - -        9.41
Q13          - -        1.70
Q14          - -        0.09
Q6          0.00         - -
Q9          0.01         - -
Q10         0.00         - -
Q11         0.08         - -
Q17         0.11         - -
Q18         0.29         - -
```

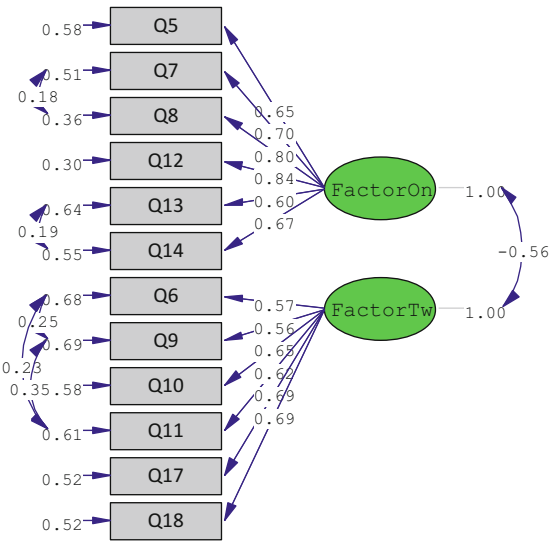
Fig. 4.11 (continued)

```
        Expected Change for LAMBDA-X

            FactorOn   FactorTw
            --------   --------
       Q5       - -       0.15
       Q7       - -      -0.10
       Q8       - -       0.17
      Q12       - -      -0.29
      Q13       - -       0.13
      Q14       - -      -0.03
       Q6      0.00        - -
       Q9      0.01        - -
      Q10      0.00        - -
      Q11     -0.02        - -
      Q17     -0.04        - -
      Q18      0.06        - -

No Non-Zero Modification Indices for PHI

        Modification Indices for THETA-DELTA

                Q5         Q7         Q8        Q12        Q13        Q14
            --------   --------   --------   --------   --------   --------
       Q5       - -
       Q7      0.48        - -
       Q8      3.12        - -        - -
      Q12      0.22       1.53       4.28        - -
      Q13      1.54       0.20       0.24       0.16        - -
      Q14      1.72       1.35       2.58       0.00        - -        - -
       Q6      1.47       1.03       0.39       1.11       3.57       0.95
       Q9      0.46       1.00       1.30       4.69       0.66       1.61
      Q10      2.69       0.25       1.55       2.84       0.00       0.13
      Q11      3.77       0.73       0.17       1.55       1.97       0.26
      Q17      0.53       0.36       2.02       2.86       1.18       0.89
      Q18      0.02       0.37       2.27       3.02       3.07       0.00

        Modification Indices for THETA-DELTA

                Q6         Q9        Q10        Q11        Q17        Q18
            --------   --------   --------   --------   --------   --------
       Q6       - -
       Q9       - -        - -
      Q10      0.00       0.86        - -
      Q11       - -        - -       0.39        - -
      Q17      0.13       0.73       0.09       0.15        - -
      Q18      0.15       0.01       0.33       0.17       0.01        - -

        Expected Change for THETA-DELTA

                Q5         Q7         Q8        Q12        Q13        Q14
            --------   --------   --------   --------   --------   --------
       Q5       - -
       Q7     -0.03        - -
       Q8      0.08        - -        - -
      Q12     -0.03       0.06      -0.10        - -
      Q13      0.07      -0.02       0.02      -0.02        - -
      Q14     -0.07      -0.05       0.06       0.00        - -        - -
       Q6      0.06      -0.04       0.02      -0.05       0.10      -0.05
       Q9      0.03      -0.04       0.04      -0.09       0.04       0.05
      Q10      0.09      -0.02       0.05      -0.08       0.00      -0.02
      Q11     -0.09       0.03      -0.01       0.05      -0.06       0.02
      Q17      0.04       0.03      -0.06       0.08      -0.06      -0.05
      Q18     -0.01      -0.03       0.06      -0.08       0.09       0.00

        Expected Change for THETA-DELTA

                Q6         Q9        Q10        Q11        Q17        Q18
            --------   --------   --------   --------   --------   --------
       Q6       - -
       Q9       - -        - -
      Q10      0.00      -0.05        - -
      Q11       - -        - -       0.03        - -
      Q17     -0.02       0.04      -0.02      -0.02        - -
      Q18      0.02       0.00       0.04      -0.02      -0.01        - -

Maximum Modification Index is    9.41 for Element ( 4, 2) of LAMBDA-X
```

Fig. 4.11 (continued)

```
Covariances

         X - KSI

                Q5          Q7          Q8          Q12         Q13         Q14
           --------    --------    --------    --------    --------    --------
 FactorOn       0.65        0.70        0.80        0.84        0.60        0.67
 FactorTw      -0.36       -0.39       -0.45       -0.47       -0.34       -0.37

         X - KSI

                Q6          Q9          Q10         Q11         Q17         Q18
           --------    --------    --------    --------    --------    --------
 FactorOn      -0.32       -0.31       -0.36       -0.35       -0.38       -0.38
 FactorTw       0.57        0.56        0.65        0.62        0.69        0.69

Factor Scores Regressions

         KSI

                Q5          Q7          Q8          Q12         Q13         Q14
           --------    --------    --------    --------    --------    --------
 FactorOn       0.15        0.10        0.25        0.37        0.09        0.13
 FactorTw      -0.03       -0.02       -0.04       -0.06       -0.01       -0.02

         KSI

                Q6          Q9          Q10         Q11         Q17         Q18
           --------    --------    --------    --------    --------    --------
 FactorOn      -0.01       -0.01       -0.03       -0.01       -0.03       -0.03
 FactorTw       0.11        0.06        0.24        0.14        0.28        0.28


          The Problem used    22936 Bytes (=  0.0% of Available Workspace)

                         Time used:     0.230 Seconds
```

Fig. 4.11 (continued)



Chi-Square=55.76, df=48, P-value=0.20619, RMSEA=0.035

Fig. 4.12 Path diagram for model with two factors from LISREL (examp4-2.pth)

```
Endogenous variables

Measurement:  stQ5 stQ7 stQ8 stQ12 stQ13 stQ14 stQ6 stQ9 stQ10 stQ11 stQ17 stQ18

Exogenous variables

Latent:       FactorOne FactorTwo

Fitting target model:

Iteration 0:   log likelihood = -1961.4632  (not concave)
Iteration 1:   log likelihood = -1946.4116
Iteration 2:   log likelihood = -1935.4528
Iteration 3:   log likelihood = -1929.9116
Iteration 4:   log likelihood = -1929.6055
Iteration 5:   log likelihood = -1929.6051
Iteration 6:   log likelihood = -1929.6051

Structural equation model                       Number of obs      =       134
Estimation method  = ml
Log likelihood     = -1929.6051

 ( 1)  [var(FactorOne)]_cons = 1
 ( 2)  [var(FactorTwo)]_cons = 1
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Measurement  |
  stQ5 <-    |
    FactorOne |    .652372   .0823235     7.92   0.000     .4910209     .813723
  -----------+----------------------------------------------------------------
  stQ7 <-    |
    FactorOne |   .7009202   .0813093     8.62   0.000     .5415568    .8602835
  -----------+----------------------------------------------------------------
  stQ8 <-    |
    FactorOne |   .7963326   .0778098    10.23   0.000     .6438282     .948837
  -----------+----------------------------------------------------------------
  stQ12 <-   |
    FactorOne |   .8310708   .0760511    10.93   0.000     .6820134    .9801281
  -----------+----------------------------------------------------------------
  stQ13 <-   |
    FactorOne |   .6059133   .0848436     7.14   0.000     .4396229    .7722038
  -----------+----------------------------------------------------------------
  stQ14 <-   |
    FactorOne |    .677898   .0827758     8.19   0.000     .5156604    .8401355
  -----------+----------------------------------------------------------------
  stQ6 <-    |
    FactorTwo |   .5622107   .0901373     6.24   0.000     .3855448    .7388765
  -----------+----------------------------------------------------------------
  stQ9 <-    |
    FactorTwo |   .5534423    .090282     6.13   0.000      .376493    .7303917
  -----------+----------------------------------------------------------------
  stQ10 <-   |
    FactorTwo |   .6445772   .0859819     7.50   0.000     .4760557    .8130987
  -----------+----------------------------------------------------------------
  stQ11 <-   |
    FactorTwo |   .6184136   .0881891     7.01   0.000     .4455662     .791261
  -----------+----------------------------------------------------------------
  stQ17 <-   |
    FactorTwo |   .6895885   .0859867     8.02   0.000     .5210576    .8581194
  -----------+----------------------------------------------------------------
  stQ18 <-   |
    FactorTwo |   .6897951   .0858138     8.04   0.000     .5216031     .857987
```

Fig. 4.13 STATA output for model with two factors (examp4-2.log)

```
-------------+------------------------------------------------------------------
Variance     |
       e.stQ5 |   .5763156    .080433                      .4383926    .7576307
       e.stQ7 |   .5042038   .0760896                      .3751034     .677737
       e.stQ8 |   .3598109   .0669065                      .2499161    .5180293
      e.stQ12 |   .2989374   .0634779                      .1971659    .4532405
      e.stQ13 |   .6442079   .0877125                      .4933216    .8412439
      e.stQ14 |   .5651523   .0803581                      .4276954    .7467865
       e.stQ6 |   .6732243    .095884                      .5092457    .8900046
       e.stQ9 |   .6785103   .0961955                      .5138989    .8958497
      e.stQ10 |   .5629018   .0863587                      .4167195    .7603639
      e.stQ11 |     .60936   .0911042                      .4545823    .8168371
      e.stQ17 |   .5250779   .0857337                      .3812775    .7231132
      e.stQ18 |   .5242229   .0853955                      .3809395    .7213998
    FactorOne |          1  (constrained)
    FactorTwo |          1  (constrained)
-------------+------------------------------------------------------------------
Covariance   |
  e.stQ7     |
       e.stQ8 |   .1823409   .0577028     3.16   0.002     .0692455    .2954363
  -----------+------------------------------------------------------------------
  e.stQ13    |
      e.stQ14 |   .1928592    .064171     3.01   0.003     .0670863     .318632
  -----------+------------------------------------------------------------------
  e.stQ6     |
       e.stQ9 |   .2443625   .0746056     3.28   0.001     .0981382    .3905869
      e.stQ11 |   .2288315   .0728893     3.14   0.002     .0859711    .3716919
  -----------+------------------------------------------------------------------
  e.stQ9     |
      e.stQ11 |   .3494844   .0777755     4.49   0.000     .1970473    .5019215
  -----------+------------------------------------------------------------------
  FactorOne  |
    FactorTwo |  -.5579635   .0820559    -6.80   0.000    -.7187901    -.397137
-------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(48)  =     55.23, Prob > chi2 = 0.2204
```

Fig. 4.13 (continued)

Figure 4.14 shows the input file in LISREL for a factor analytic structure where a single factor is assumed to be reflected by all the items.

Figure 4.15 is the corresponding LISREL output for the factor analytic structure where a single factor is assumed to be reflected by all the items.

In Fig. 4.15, the resulting chi-square ($\chi^2 = 126.75$) can be compared with the chi-square resulting from a model with a correlation between the two factors ($\chi^2 = 54.78$ in Fig. 4.11). The χ^2 difference ($126.75 - 54.78$) has 1 degree of freedom and its significance indicates that there are indeed two different constructs (factors). This demonstrates the discriminant validity of the constructs.

4.6.3 Example of Model to Assess the Convergent Validity of a Construct

Next, in order to assess the convergent validity, we need to compare the fit of a model with zero correlation between the factors with a model where the factors are correlated (as in Fig. 4.11 for LISREL or Fig. 4.13 for STATA). The input file in

```
. estat gof, stats(all)

------------------------------------------------------------------------------
Fit statistic        |      Value   Description
---------------------+--------------------------------------------------------
Likelihood ratio     |
         chi2_ms(48) |     55.227   model vs. saturated
            p > chi2 |      0.220
         chi2_bs(66) |    754.104   baseline vs. saturated
            p > chi2 |      0.000
---------------------+--------------------------------------------------------
Population error     |
                RMSEA |      0.034   Root mean squared error of approximation
 90% CI, lower bound |      0.000
         upper bound |      0.068
              pclose |      0.748   Probability RMSEA <= 0.05
---------------------+--------------------------------------------------------
Information criteria |
                 AIC |   3919.210   Akaike's information criterion
                 BIC |   4006.145   Bayesian information criterion
---------------------+--------------------------------------------------------
Baseline comparison  |
                 CFI |      0.989   Comparative fit index
                 TLI |      0.986   Tucker-Lewis index
---------------------+--------------------------------------------------------
Size of residuals    |
                SRMR |      0.048   Standardized root mean squared residual
                  CD |      0.960   Coefficient of determination
------------------------------------------------------------------------------

. estat mindices, min(1)

Modification indices

-------------------------------------------------------------
             |                                     Standard
             |         MI     df   P>MI        EPC        EPC
-------------+-----------------------------------------------
Measurement  |
  stQ5 <-    |
    FactorTwo |      2.158      1   0.14   .1533973   .1532514
  -----------+-----------------------------------------------
  stQ7 <-    |
    FactorTwo |      1.454      1   0.23   -.101899  -.1021294
  -----------+-----------------------------------------------
  stQ8 <-    |
    FactorTwo |      4.778      1   0.03   .1695152   .1700298
  -----------+-----------------------------------------------
  stQ12 <-   |
    FactorTwo |      9.485      1   0.00  -.2869304  -.2884319
  -----------+-----------------------------------------------
  stQ13 <-   |
    FactorTwo |      1.715      1   0.19    .130595   .1298608
-------------+-----------------------------------------------
Covariance   |
  e.stQ5     |
       e.stQ8 |      3.148      1   0.08   .0789308   .1733321
      e.stQ13 |      1.556      1   0.21   .0689573   .1131715
      e.stQ14 |      1.735      1   0.19  -.0702762  -.1231388
       e.stQ6 |      1.481      1   0.22   .0640008   .1027484
      e.stQ10 |      2.721      1   0.10   .0933285   .1638581
      e.stQ11 |      3.807      1   0.05  -.0882305  -.1488852
  -----------+-----------------------------------------------
  e.stQ7     |
      e.stQ12 |      1.544      1   0.21   .0566216   .1458442
      e.stQ14 |      1.358      1   0.24   -.049985  -.0936385
       e.stQ6 |      1.034      1   0.31  -.0434994  -.0746621
       e.stQ9 |      1.009      1   0.32  -.0383362  -.0655431
  -----------+-----------------------------------------------
  e.stQ8     |
      e.stQ12 |      4.313      1   0.04  -.0949785  -.2895998
```

Fig. 4.13 (continued)

```
      e.stQ14 |      2.598      1    0.11    .0646087    .1432752
       e.stQ9 |      1.304      1    0.25     .039059    .0790508
      e.stQ10 |      1.556      1    0.21    .0515188    .1144753
      e.stQ17 |      2.035      1    0.15   -.0581871   -.1338683
      e.stQ18 |      2.288      1    0.13    .0616638    .1419826
--------------+----------------------------------------------------
e.stQ12       |
      e.stQ6  |      1.116      1    0.29   -.0459271   -.1023761
      e.stQ9  |      4.727      1    0.03   -.0842564   -.1870831
      e.stQ10 |      2.865      1    0.09   -.0795793   -.1939965
      e.stQ11 |      1.559      1    0.21    .0466886    .1093915
      e.stQ17 |      2.874      1    0.09    .0788534    .1990298
      e.stQ18 |      3.039      1    0.08   -.0810514   -.2047445
--------------+----------------------------------------------------
e.stQ13       |
      e.stQ6  |      3.599      1    0.06    .0963776    .1463468
      e.stQ11 |      1.984      1    0.16   -.0615165   -.0981842
      e.stQ17 |      1.193      1    0.27   -.0588443   -.1011765
      e.stQ18 |      3.096      1    0.08    .0947413    .1630304
--------------+----------------------------------------------------
e.stQ14       |
      e.stQ9  |      1.622      1    0.20     .055032    .0888698
-------------------------------------------------------------------

EPC = expected parameter change

. estat framework, fitted

Endogenous variables on endogenous variables

             | observed
        Beta |      stQ5       stQ7       stQ8      stQ12      stQ13      stQ14       stQ6
-------------+-----------------------------------------------------------------------------
observed     |
        stQ5 |         0
        stQ7 |         0          0
        stQ8 |         0          0          0
       stQ12 |         0          0          0          0
       stQ13 |         0          0          0          0          0
       stQ14 |         0          0          0          0          0          0
        stQ6 |         0          0          0          0          0          0          0
        stQ9 |         0          0          0          0          0          0          0
       stQ10 |         0          0          0          0          0          0          0
       stQ11 |         0          0          0          0          0          0          0
       stQ17 |         0          0          0          0          0          0          0
       stQ18 |         0          0          0          0          0          0          0
-----------------------------------------------------------------------------------------

             | observed
        Beta |      stQ9      stQ10      stQ11      stQ17      stQ18
-------------+--------------------------------------------------------
observed     |
        stQ9 |         0
       stQ10 |         0          0
       stQ11 |         0          0          0
       stQ17 |         0          0          0          0
       stQ18 |         0          0          0          0          0
-------------------------------------------------------------------------
```

Fig. 4.13 (continued)

LISREL for a model with independent factors (zero correlation) is shown in Fig. 4.16.

The commands in STATA that correspond to the LISREL example in Fig. 4.16 are shown in Fig. 4.17.

The constraint that the covariance between the two latent factors is zero is represented by the "cov(FactorOne*FactorTwo@0)" commands. The LISREL output file for such a model with independent factors (zero correlation) is shown in Fig. 4.18.

```
Exogenous variables on endogenous variables

               | latent
         Gamma | FactorOne  FactorTwo
    -----------+----------------------
    observed   |
           stQ5 |   .652372          0
           stQ7 |  .7009202          0
           stQ8 |  .7963326          0
          stQ12 |  .8310708          0
          stQ13 |  .6059133          0
          stQ14 |   .677898          0
           stQ6 |          0   .5622107
           stQ9 |          0   .5534423
          stQ10 |          0   .6445772
          stQ11 |          0   .6184136
          stQ17 |          0   .6895885
          stQ18 |          0   .6897951
    -----------------------------------


Covariances of error variables

               | observed
           Psi |    e.stQ5     e.stQ7     e.stQ8    e.stQ12    e.stQ13    e.stQ14     e.stQ6
    -----------+-------------------------------------------------------------------------------
    observed   |
         e.stQ5 |  .5763156
         e.stQ7 |          0   .5042038
         e.stQ8 |          0   .1823409   .3598109
        e.stQ12 |          0          0          0   .2989374
        e.stQ13 |          0          0          0          0   .6442079
        e.stQ14 |          0          0          0          0   .1928592   .5651523
         e.stQ6 |          0          0          0          0          0          0   .6732243
         e.stQ9 |          0          0          0          0          0          0   .2443625
        e.stQ10 |          0          0          0          0          0          0          0
        e.stQ11 |          0          0          0          0          0          0   .2288315
        e.stQ17 |          0          0          0          0          0          0          0
        e.stQ18 |          0          0          0          0          0          0          0
    -------------------------------------------------------------------------------------------

               | observed
           Psi |    e.stQ9    e.stQ10    e.stQ11    e.stQ17    e.stQ18
    -----------+--------------------------------------------------------
    observed   |
         e.stQ9 |  .6785103
        e.stQ10 |          0   .5629018
        e.stQ11 |  .3494844          0     .60936
        e.stQ17 |          0          0          0   .5250779
        e.stQ18 |          0          0          0          0   .5242229
    ------------------------------------------------------------------------


Covariances of exogenous variables

               | latent
           Phi | FactorOne  FactorTwo
    -----------+----------------------
    latent     |
       FactorOne |          1
       FactorTwo | -.5579635          1
    -----------------------------------
```

Fig. 4.13 (continued)

```
 Fitted covariances of observed and latent variables

                | observed
        Sigma   |      stQ5       stQ7       stQ8      stQ12      stQ13      stQ14       stQ6
 ---------------+-------------------------------------------------------------------------------
 observed       |
          stQ5  |  1.001905
          stQ7  |  .4572607   .9954929
          stQ8  |  .5195051   .7405065   .9939565
         stQ12  |  .5421673   .5825143   .6618088   .9896161
         stQ13  |  .3952809   .4246969   .4825085   .5035568   1.011339
         stQ14  |  .4422416   .4751523   .5398323   .5633812   .6036066   1.024698
          stQ6  | -.2046445  -.2198738   -.249804  -.2607011  -.1900708  -.2126519   .9893052
          stQ9  | -.2014529  -.2164446   -.245908  -.2566352  -.1871064  -.2093353   .5555137
         stQ10  |  -.234626  -.2520864  -.2864015  -.2988951  -.2179171  -.2438064   .3623882
         stQ11  | -.2251024  -.2418541  -.2747764  -.2867628  -.2090717  -.2339102   .5765102
         stQ17  |   -.25101  -.2696897  -.3064011  -.3197671  -.2331344  -.2608316    .387694
         stQ18  | -.2510852  -.2697705  -.3064929  -.3198629  -.2332042  -.2609097   .3878101
 ---------------+-------------------------------------------------------------------------------
 latent         |
    FactorOne   |   .652372   .7009202   .7963326   .8310708   .6059133    .677898  -.3136931
    FactorTwo   | -.3639998  -.3910879  -.4443246  -.4637072  -.3380775  -.3782423   .5622107
 ---------------+-------------------------------------------------------------------------------
                | observed                                                 | latent
        Sigma   |      stQ9      stQ10      stQ11      stQ17      stQ18   | FactorOne  FactorTwo
 ---------------+----------------------------------------------------------+----------------------
 observed       |                                                          |
          stQ9  |  .9848087                                                |
         stQ10  |  .3567363   .9783816                                     |
         stQ11  |  .6917407   .3986153   .9917954                          |
         stQ17  |  .3816474    .444493   .4264509    1.00061               |
         stQ18  |  .3817618   .4446262   .4265786   .4756747    1.00004    |
 ---------------+----------------------------------------------------------+----------------------
 latent         |                                                          |
    FactorOne   | -.3088006  -.3596506  -.3450522  -.3847652  -.3848805    |         1
    FactorTwo   |  .5534423   .6445772   .6184136   .6895885   .6897951    | -.5579635          1
 ---------------+----------------------------------------------------------+----------------------
```

Fig. 4.13 (continued)

```
!Examp4-3.spl
!Raw Data From File: Examp4-2.txt

DA NI=12 MA = KM XM = 9
RA FI=C:\SAMD\Chapter4\Examples\Examp4-2.txt
LA
Q5 Q7 Q8  Q12 Q13 Q14
Q6 Q9 Q10 Q11 Q17 Q18

MO NX = 12 NK = 1 PH = ST TD = SY
LK
FactOne

FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1)   C
   LX(7,1) LX(8,1) LX(9,1) LX(10,1) LX(11,1) LX(12,1)        C
   TD(3,2) TD(6,5) TD(8,7) TD(10,8) TD(10,7)
Path Diagram
OU SE TV RS MR FS AD = 50 MI
```

Fig. 4.14 LISREL input for model with single factor (examp4-3.spl)

Still using the example in LISREL, the independent factor model has a chi-square of 84.34 (Fig. 4.18), which, when compared with the chi-square of the model estimating a correlation between the two constructs (Fig. 4.11), shows a chi-square difference of 29.56. This difference is significant (with 1 degree of freedom at the 0.05 level), and thus it indicates that the constructs are not independent. Therefore, the chi-square test supports the convergent validity of the two constructs.

```
                                L I S R E L  8.30

                                      BY

                           Karl G. Jöreskog & Dag Sörbom

                        This program is published exclusively by
                        Scientific Software International, Inc.
                           7383 N. Lincoln Avenue, Suite 100
                            Chicago, IL 60646-1704, U.S.A.
                Phone: (800)247-6113, (847)675-0720, Fax: (847)675-2140
              Copyright by Scientific Software International, Inc., 1981-99
                Use of this program is subject to the terms specified in the
                           Universal Copyright Convention.
                             Website: www.ssicentral.com

 The following lines were read from file C:\SAMD\CHAPTER8\EXAMPLES\EXAMP4-3.SPL:

 !Examp4-3.spl
 !Raw Data From File: Examp4-2.txt

 DA NI=12 MA = KM XM = 9
 RA FI=C:\SAMD\Chapter4\Examples\Examp4-2.txt
 LA
 Q5 Q7 Q8  Q12 Q13 Q14
 Q6 Q9 Q10 Q11 Q17 Q18

 MO NX = 12 NK = 1 PH = ST TD = SY
 LK
 FactOne      !Competence Destroying
 FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1) C
 LX(7,1) LX(8,1) LX(9,1) LX(10,1) LX(11,1) LX(12,1) C
 TD(3,2) TD(6,5) TD(8,7) TD(10,8) TD(10,7)
 Path Diagram
 OU SE TV RS MR FS AD = 50 MI

 !Examp4-3.spl
                           Number of Input Variables 12
                           Number of Y - Variables    0
                           Number of X - Variables   12
                           Number of ETA - Variables  0
                           Number of KSI - Variables  1
                           Number of Observations   134

Covariance Matrix to be Analyzed

               Q5         Q7         Q8        Q12        Q13        Q14
         --------   --------   --------   --------   --------   --------
       Q5       1.00
       Q7       0.46       1.00
       Q8       0.57       0.74       1.00
      Q12       0.53       0.60       0.64       1.00
      Q13       0.43       0.40       0.51       0.49       1.00
      Q14       0.40       0.44       0.58       0.56       0.59       1.00
       Q6      -0.13      -0.27      -0.20      -0.36      -0.06      -0.19
       Q9      -0.17      -0.26      -0.18      -0.38      -0.08      -0.11
      Q10      -0.13      -0.27      -0.22      -0.40      -0.19      -0.26
      Q11      -0.26      -0.25      -0.23      -0.36      -0.18      -0.19
      Q17      -0.19      -0.29      -0.32      -0.34      -0.26      -0.32
      Q18      -0.20      -0.27      -0.21      -0.40      -0.10      -0.22
```

Fig. 4.15 LISREL output of model with single factor (examp4-3.out)

Covariance Matrix to be Analyzed

	Q6	Q9	Q10	Q11	Q17	Q18
Q6	1.00					
Q9	0.56	1.00				
Q10	0.36	0.33	1.00			
Q11	0.58	0.70	0.41	1.00		
Q17	0.38	0.41	0.44	0.43	1.00	
Q18	0.40	0.38	0.47	0.42	0.47	1.00

Parameter Specifications

LAMBDA-X

	FactOne
Q5	1
Q7	2
Q8	3
Q12	4
Q13	5
Q14	6
Q6	7
Q9	8
Q10	9
Q11	10
Q17	11
Q18	12

THETA-DELTA

	Q5	Q7	Q8	Q12	Q13	Q14
Q5	13					
Q7	0	14				
Q8	0	15	16			
Q12	0	0	0	17		
Q13	0	0	0	0	18	
Q14	0	0	0	0	19	20
Q6	0	0	0	0	0	0
Q9	0	0	0	0	0	0
Q10	0	0	0	0	0	0
Q11	0	0	0	0	0	0
Q17	0	0	0	0	0	0
Q18	0	0	0	0	0	0

Fig. 4.15 (continued)

Instead of defining the variances of the unobserved constructs to unity, we would have obtained the same result if we had fixed one lambda to one for each construct. In that case, we would have estimated the variances of these constructs. Although we could have illustrated this model easily using LISREL or STATA following the principles described above, we use the input needed to run the model with AMOS in order to introduce its commands.

The input of the corresponding two-factor confirmatory factor model with AMOS is shown in Fig. 4.19.

In AMOS (Fig. 4.19), each equation for the measurement model can be represented with a variable on the left side of an equation and a linear combination of other variables on the right side. These equations correspond to the measurement

```
          THETA-DELTA

               Q6          Q9         Q10         Q11         Q17         Q18
          --------    --------    --------    --------    --------    --------
       Q6         21
       Q9         22          23
      Q10          0           0          24
      Q11         25          26           0          27
      Q17          0           0           0           0          28
      Q18          0           0           0           0           0          29


Number of Iterations = 18

 LISREL Estimates (Maximum Likelihood)

          LAMBDA-X

              FactOne
              --------
       Q5        0.61
                (0.08)
                 7.37

       Q7        0.68
                (0.08)
                 8.35

       Q8        0.75
                (0.08)
                 9.48

      Q12        0.85
                (0.07)
                11.50

      Q13        0.57
                (0.09)
                 6.66

      Q14        0.65
                (0.08)
                 7.85

       Q6       -0.40
                (0.09)
                -4.54

       Q9       -0.40
                (0.09)
                -4.50

      Q10       -0.46
                (0.09)
                -5.27

      Q11       -0.45
                (0.09)
                -5.08

      Q17       -0.48
                (0.09)
                -5.57

      Q18       -0.47
                (0.09)
                -5.34
```

Fig. 4.15 (continued)

```
PHI

          FactOne
          --------
             1.00


THETA-DELTA

              Q5         Q7         Q8         Q12        Q13        Q14
          --------   --------   --------   --------   --------   --------
Q5          0.62
           (0.08)
            7.41

Q7           - -       0.54
                      (0.08)
                       6.96

Q8           - -       0.24       0.44
                      (0.06)     (0.07)
                       4.00       6.51

Q12          - -        - -        - -       0.27
                                    (0.06)
                                     4.83

Q13          - -        - -        - -        - -       0.68
                                                       (0.09)
                                                        7.54

Q14          - -        - -        - -        - -       0.23       0.58
                                                       (0.07)     (0.08)
                                                        3.48       7.24

Q6           - -        - -        - -        - -        - -        - -

Q9           - -        - -        - -        - -        - -        - -

Q10          - -        - -        - -        - -        - -        - -

Q11          - -        - -        - -        - -        - -        - -

Q17          - -        - -        - -        - -        - -        - -

Q18          - -        - -        - -        - -        - -        - -


THETA-DELTA

              Q6         Q9         Q10        Q11        Q17        Q18
          --------   --------   --------   --------   --------   --------
Q6          0.84
           (0.11)
            7.91

Q9          0.40       0.84
           (0.08)     (0.11)
            4.80       7.92

Q10          - -        - -       0.79
                                   (0.10)
                                    7.83

Q11         0.40       0.52        - -       0.80
           (0.08)     (0.09)                (0.10)
            4.88       5.95                  7.85

Q17          - -        - -        - -        - -       0.77
                                                       (0.10)
                                                        7.78

Q18          - -        - -        - -        - -        - -       0.78
                                                                   (0.10)
                                                                    7.82
```

Fig. 4.15 (continued)

```
   Squared Multiple Correlations for X - Variables

         Q5         Q7         Q8        Q12        Q13        Q14
   --------   --------   --------   --------   --------   --------
       0.38       0.46       0.56       0.73       0.32       0.42

   Squared Multiple Correlations for X - Variables

         Q6         Q9        Q10        Q11        Q17        Q18
   --------   --------   --------   --------   --------   --------
       0.16       0.16       0.21       0.20       0.23       0.22


                    Goodness of Fit Statistics

                      Degrees of Freedom = 49
        Minimum Fit Function Chi-Square = 126.75 (P = 0.00)
Normal Theory Weighted Least Squares Chi-Square = 158.94 (P = 0.00)
          Estimated Non-centrality Parameter (NCP) = 109.94
     90 Percent Confidence Interval for NCP = (75.53 ; 151.95)

                Minimum Fit Function Value = 0.95
          Population Discrepancy Function Value (F0) = 0.83
        90 Percent Confidence Interval for F0 = (0.57 ; 1.14)
        Root Mean Square Error of Approximation (RMSEA) = 0.13
       90 Percent Confidence Interval for RMSEA = (0.11 ; 0.15)
         P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00

           Expected Cross-Validation Index (ECVI) = 1.63
      90 Percent Confidence Interval for ECVI = (1.37 ; 1.95)
                 ECVI for Saturated Model = 1.17
               ECVI for Independence Model = 5.81

Chi-Square for Independence Model with 66 Degrees of Freedom = 748.31
                     Independence AIC = 772.31
                         Model AIC = 216.94
                       Saturated AIC = 156.00
                     Independence CAIC = 819.08
                        Model CAIC = 329.97
                      Saturated CAIC = 460.03

             Root Mean Square Residual (RMR) = 0.10
                   Standardized RMR = 0.10
               Goodness of Fit Index (GFI) = 0.83
          Adjusted Goodness of Fit Index (AGFI) = 0.74
         Parsimony Goodness of Fit Index (PGFI) = 0.52

                  Normed Fit Index (NFI) = 0.83
                Non-Normed Fit Index (NNFI) = 0.85
           Parsimony Normed Fit Index (PNFI) = 0.62
              Comparative Fit Index (CFI) = 0.89
              Incremental Fit Index (IFI) = 0.89
               Relative Fit Index (RFI) = 0.77

                   Critical N (CN) = 79.62


   Fitted Covariance Matrix

         Q5         Q7         Q8        Q12        Q13        Q14
   --------   --------   --------   --------   --------   --------
 Q5      1.00
 Q7      0.42       1.00
 Q8      0.46       0.74       1.00
Q12      0.52       0.58       0.64       1.00
Q13      0.35       0.39       0.42       0.48       1.00
Q14      0.40       0.44       0.48       0.55       0.59       1.00
 Q6     -0.25      -0.27      -0.30      -0.34      -0.23      -0.26
 Q9     -0.25      -0.27      -0.30      -0.34      -0.23      -0.26
Q10     -0.28      -0.31      -0.34      -0.39      -0.26      -0.30
Q11     -0.27      -0.30      -0.33      -0.38      -0.25      -0.29
Q17     -0.30      -0.33      -0.36      -0.41      -0.27      -0.31
Q18     -0.29      -0.32      -0.35      -0.40      -0.26      -0.30

   Fitted Covariance Matrix
```

Fig. 4.15 (continued)

```
               Q6          Q9         Q10         Q11         Q17         Q18
         ---------   ---------   ---------   ---------   ---------   ---------
      Q6       1.00
      Q9       0.56        1.00
     Q10       0.19        0.18        1.00
     Q11       0.58        0.70        0.21        1.00
     Q17       0.20        0.19        0.22        0.22        1.00
     Q18       0.19        0.19        0.22        0.21        0.23        1.00

        Fitted Residuals

               Q5          Q7          Q8          Q12         Q13         Q14
         ---------   ---------   ---------   ---------   ---------   ---------
      Q5       0.00
      Q7       0.04        0.00
      Q8       0.11        0.00        0.00
     Q12       0.01        0.02        0.01        0.00
     Q13       0.09        0.01        0.08        0.01        0.00
     Q14       0.00        0.00        0.10        0.01        0.00        0.00
      Q6       0.11        0.00        0.10       -0.01        0.17        0.07
      Q9       0.08        0.02        0.12       -0.04        0.15        0.15
     Q10       0.16        0.05        0.13        0.00        0.07        0.04
     Q11       0.01        0.05        0.10        0.02        0.07        0.10
     Q17       0.11        0.04        0.04        0.07        0.01        0.00
     Q18       0.08        0.05        0.14        0.00        0.17        0.09

        Fitted Residuals

               Q6          Q9         Q10         Q11         Q17         Q18
         ---------   ---------   ---------   ---------   ---------   ---------
      Q6       0.00
      Q9       0.00        0.00
     Q10       0.17        0.14        0.00
     Q11       0.00        0.00        0.20        0.00
     Q17       0.19        0.22        0.22        0.21        0.00
     Q18       0.21        0.20        0.25        0.21        0.25        0.00

 Summary Statistics for Fitted Residuals

 Smallest Fitted Residual =   -0.04
   Median Fitted Residual =    0.05
  Largest Fitted Residual =    0.25

…

 Modification Indices and Expected Change

 No Non-Zero Modification Indices for LAMBDA-X

 No Non-Zero Modification Indices for PHI

        Modification Indices for THETA-DELTA

               Q5          Q7          Q8          Q12         Q13         Q14
         ---------   ---------   ---------   ---------   ---------   ---------
      Q5       - -
      Q7      0.27         - -
      Q8      7.37         - -         - -
     Q12      0.20        0.61        0.02         - -
     Q13      3.16        0.06        1.25        0.08         - -
     Q14      0.45        1.24        4.69        0.02         - -         - -
      Q6      3.20        0.84        1.33        0.21        4.75        0.83
      Q9      1.10        0.86        2.05        3.59        0.84        1.78
     Q10      7.77        0.17        6.68        0.02        0.99        0.11
     Q11      1.76        0.88        0.00        3.90        1.32        0.39
     Q17      3.93        0.16        0.37        6.64        0.05        0.02
     Q18      2.25        0.24        7.86        0.01        5.50        0.35
```

Fig. 4.15 (continued)

```
        Modification Indices for THETA-DELTA

              Q6        Q9       Q10       Q11       Q17       Q18
        --------  --------  --------  --------  --------  --------
Q6         - -
Q9         - -       - -
Q10       1.41      0.01       - -
Q11        - -       - -      3.21       - -
Q17       1.05      1.58     11.37      1.29       - -
Q18       2.51      0.55     15.04      1.36     14.90       - -

        Expected Change for THETA-DELTA

              Q5        Q7        Q8       Q12       Q13       Q14
        --------  --------  --------  --------  --------  --------
Q5         - -
Q7       -0.02       - -
Q8        0.12       - -       - -
Q12       0.02      0.03     -0.01       - -
Q13       0.10     -0.01      0.05      0.01       - -
Q14      -0.04     -0.05      0.09      0.01       - -       - -
Q6        0.10     -0.04      0.05     -0.02      0.11     -0.05
Q9        0.05     -0.04      0.05     -0.08      0.04      0.06
Q10       0.18     -0.02      0.13     -0.01      0.06      0.02
Q11      -0.06      0.04      0.00      0.08     -0.05      0.03
Q17       0.13      0.02      0.03      0.14      0.01     -0.01
Q18       0.10     -0.03      0.14      0.00      0.14      0.03

        Expected Change for THETA-DELTA

              Q6        Q9       Q10       Q11       Q17       Q18
        --------  --------  --------  --------  --------  --------
Q6         - -
Q9         - -       - -
Q10       0.07      0.00       - -
Q11        - -       - -      0.09       - -
Q17       0.06      0.07      0.24      0.06       - -
Q18       0.10      0.04      0.28      0.06      0.27       - -

Maximum Modification Index is   15.04 for Element (12, 9) of THETA-DELTA


Covariances

        X - KSI

              Q5        Q7        Q8       Q12       Q13       Q14
        --------  --------  --------  --------  --------  --------
FactOne       0.61      0.68      0.75      0.85      0.57      0.65

        X - KSI

              Q6        Q9       Q10       Q11       Q17       Q18
        --------  --------  --------  --------  --------  --------
FactOne     -0.40     -0.40     -0.46     -0.45     -0.48     -0.47
```

Fig. 4.15 (continued)

model as specified by Eq. (4.2). Inserting "(1)" before a variable on the right side indicates that the coefficient is fixed to that value and that the corresponding parameters will not be estimated. The program recognizes automatically which variables are observed and which are unobserved.

Correlations are indicated by "*variable1* <> *variable*", where *variable1* and *variable2* are the labels of observed variables or of hypothetical constructs. The output provides information similar to that which is available in LISREL or STATA.

```
 Factor Scores Regressions

         KSI

                  Q5          Q7          Q8         Q12         Q13         Q14
            --------    --------    --------    --------    --------    --------
  FactOne       0.13        0.09        0.17        0.40        0.07        0.12

         KSI

                  Q6          Q9         Q10         Q11         Q17         Q18
            --------    --------    --------    --------    --------    --------
  FactOne      -0.03       -0.02       -0.08       -0.04       -0.08       -0.08
```

Fig. 4.15 (continued)

```
!Examp4-4.spl
!Raw Data From File: Examp4-2.txt

DA NI=12 MA = KM XM = 9
RA FI=C:\SAMD\Chapter4\Examples\Examp4-2.txt
LA
Q5 Q7 Q8  Q12 Q13 Q14
Q6 Q9 Q10 Q11 Q17 Q18

MO NX = 12 NK = 2 PH = DI TD = SY
LK
FactOne                                          !Competence Destroying
FactTwo                                        !Competence Enhancing
FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1)
   LX(7,2) LX(8,2) LX(9,2) LX(10,2) LX(11,2) LX(12,2)
   TD(3,2) TD(6,5) TD(8,7) TD(10,8) TD(10,7)
Path Diagram
OU SE TV RS MR FS AD = 50 MI
```

Fig. 4.16 LISREL input for model with two independent factors (examp4-4.spl)

```
…
sem (FactorOne -> stQ5 stQ7 stQ8 stQ12 stQ13 stQ14) ///
(FactorTwo -> stQ6 stQ9 stQ10 stQ11 stQ17 stQ18) ///
, cov(e.stQ7*e.stQ8) cov(e.stQ13*e.stQ14) ///
cov(e.stQ6*e.stQ11) cov(e.stQ9*e.stQ11) cov(e.stQ6*e.stQ9) ///
var(FactorOne@1) var(FactorTwo@1) cov(FactorOne*FactorTwo@0) ///
nomeans latent(FactorOne FactorTwo)
…
```

Fig. 4.17 STATA input for model with two independent factors (examp4-4_Mac.do)

4.6.4 Example of Second-Order Factor Model

Next we present an example of second-order factor analysis using the same data as in the previous examples. Since two factors are correlated, we can test a model where these two factors reflect a single higher-order construct. Figure 4.20 shows the LISREL input file.

For the most part, the input file contains commands similar to the description of the input files of regular (first-order) CFA. It should be noted that the matrix to be

```
                                L I S R E L  8.30

                                       BY

                            Karl G. Jöreskog & Dag Sörbom

                        This program is published exclusively by
                        Scientific Software International, Inc.
                           7383 N. Lincoln Avenue, Suite 100
                            Chicago, IL 60646-1704, U.S.A.
              Phone: (800)247-6113, (847)675-0720, Fax: (847)675-2140
        Copyright by Scientific Software International, Inc., 1981-99
          Use of this program is subject to the terms specified in the
                        Universal Copyright Convention.
                          Website: www.ssicentral.com

The following lines were read from file C:\SAMD\CHAPTER8\EXAMPLES\EXAMP4-4.SPL:

!Examp4-4.spl

!Raw Data From File: Examp4-2.txt

DA NI=12 MA = KM XM = 9
RA FI=C:\SAMD\Chapter8\Examples\Examp4-2.txt
LA
Q5 Q7 Q8  Q12 Q13 Q14
Q6 Q9 Q10 Q11 Q17 Q18

MO NX = 12 NK = 2 PH = DI TD = SY  !CORR = 0
```

Fig. 4.18 LISREL output of model with two independent factors (examp4-4.out)

analyzed here is the covariance matrix, rather than the correlation matrix typically analyzed in single-group CFA. This is indicated in the data line ("DA" line) with the "MA = CM" command. In this particular example, the sample size is also provided on the data line ("NO = 145"). The difference is that in the model statement NX has been replaced by NY, the number of indicator variables for the elements of $\boldsymbol{\eta}$. NE corresponds to the number of first-order factors (the ηs). NK is set to 1 in this example because only one second-order factor is assumed. GA indicates that the elements of the $\boldsymbol{\Gamma}$ matrix will be fixed by default, although we will specify which elements to estimate in the "FREE" line below. The covariance matrix of the second-order factors is set to be standardized ("PH = ST"); although in our example this matrix is simply a scalar, the LISREL command sets the variance of the second-order factor to be fixed to unity for identification. Alternatively, one of the gamma parameters could be set to unity. This is the choice made in the STATA formulation of the same example shown in Fig. 4.21. The labels for the first-order factors are the same as in the earlier example of regular CFA, except that they now correspond to the ηs, which is why they are introduced by "LE" (*L*abel *E*tas). The label for the second-order factor ("new") follows the "LK" (*L*abel *K*sis) command.

One of the factor loadings for each first-order factor is fixed to 1 in order to define the unit of the factors to the units of that item. Finally, the parameters to be estimated are freed; they are the elements of the factor loading matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$.

The commands in STATA are straightforward, as shown in Fig. 4.21.

Although it is not strictly necessary to indicate the constraints for the unit loadings needed for identification (STATA generates the constraints

```
LK
FactOne      !Competence Destroying
FactTwo     !Competence Enhancing
FR LX(1,1) LX(2,1) LX(3,1) LX(4,1) LX(5,1) LX(6,1) C
LX(7,2) LX(8,2) LX(9,2) LX(10,2) LX(11,2) LX(12,2) C
TD(3,2) TD(6,5) TD(8,7) TD(10,8) TD(10,7)
Path Diagram
OU SE TV RS MR FS AD = 50 MI

!Examp4-4.spl


                        Number of Input Variables 12
                        Number of Y - Variables    0
                        Number of X - Variables   12
                        Number of ETA - Variables  0
                        Number of KSI - Variables  2
                        Number of Observations   134


        Covariance Matrix to be Analyzed

              Q5         Q7         Q8        Q12        Q13        Q14
        --------   --------   --------   --------   --------   --------
 Q5       1.00
 Q7       0.46       1.00
 Q8       0.57       0.74       1.00
Q12       0.53       0.60       0.64       1.00
Q13       0.43       0.40       0.51       0.49       1.00
Q14       0.40       0.44       0.58       0.56       0.59       1.00
 Q6      -0.13      -0.27      -0.20      -0.36      -0.06      -0.19
 Q9      -0.17      -0.26      -0.18      -0.38      -0.08      -0.11
Q10      -0.13      -0.27      -0.22      -0.40      -0.19      -0.26
Q11      -0.26      -0.25      -0.23      -0.36      -0.18      -0.19
Q17      -0.19      -0.29      -0.32      -0.34      -0.26      -0.32
Q18      -0.20      -0.27      -0.21      -0.40      -0.10      -0.22

        Covariance Matrix to be Analyzed

              Q6         Q9        Q10        Q11        Q17        Q18
        --------   --------   --------   --------   --------   --------
 Q6       1.00
 Q9       0.56       1.00
Q10       0.36       0.33       1.00
Q11       0.58       0.70       0.41       1.00
Q17       0.38       0.41       0.44       0.43       1.00
Q18       0.40       0.38       0.47       0.42       0.47       1.00
```

Fig. 4.18 (continued)

automatically), it is informative to be explicit about these constraints at the model specification stage.

The LISREL output corresponding to this second-order factor analysis is shown in Fig. 4.22.

The graphical representation of the results is shown in Fig. 4.23.

As seen in Fig. 4.22, the highly significant chi-square indicates that the second-order factor model has a poor fit. Nevertheless, the parameter estimates for the second-order factor loadings on the first-order factors correspond to what would be expected from the correlation pattern between these two constructs (a positive loading on FactorOne and a negative loading on FactorTwo).

```
Parameter Specifications

        LAMBDA-X

             FactOne    FactTwo
            --------   --------
       Q5          1          0
       Q7          2          0
       Q8          3          0
      Q12          4          0
      Q13          5          0
      Q14          6          0
       Q6          0          7
       Q9          0          8
      Q10          0          9
      Q11          0         10
      Q17          0         11
      Q18          0         12

        THETA-DELTA

             Q5         Q7         Q8        Q12        Q13        Q14
            --------   --------   --------   --------   --------   --------
       Q5         13
       Q7          0         14
       Q8          0         15         16
      Q12          0          0          0         17
      Q13          0          0          0          0         18
      Q14          0          0          0          0         19         20
       Q6          0          0          0          0          0          0
       Q9          0          0          0          0          0          0
      Q10          0          0          0          0          0          0
      Q11          0          0          0          0          0          0
      Q17          0          0          0          0          0          0
      Q18          0          0          0          0          0          0

        THETA-DELTA

             Q6         Q9        Q10        Q11        Q17        Q18
            --------   --------   --------   --------   --------   --------
       Q6         21
       Q9         22         23
      Q10          0          0         24
      Q11         25         26          0         27
      Q17          0          0          0          0         28
      Q18          0          0          0          0          0         29



 Number of Iterations = 29
```

Fig. 4.18 (continued)

4.6.5 Example of Multi-Group Factor Analysis

The example we use to illustrate the analysis of factors across groups concerns the subjective well-being of men in three different countries (USA, Austria, and Australia). There are five items to measure subjective well-being. We first illustrate this analysis using LISREL and then using STATA. Figure 4.24 lists the input for performing this analysis in LISREL.

We indicate that the data file contains raw data (rather than correlations or covariances) by specifying on the third line "RA =" followed by the full name of

```
LISREL Estimates (Maximum Likelihood)

        LAMBDA-X

            FactOne    FactTwo
           --------   --------
      Q5       0.67        - -
              (0.08)
                8.11

      Q7       0.71        - -
              (0.08)
                8.50

      Q8       0.83        - -
              (0.08)
               10.76

     Q12       0.80        - -
              (0.08)
               10.31

     Q13       0.62        - -
              (0.08)
                7.29

     Q14       0.67        - -
              (0.08)
                8.19

      Q6        - -       0.56
                         (0.09)
                           6.08

      Q9        - -       0.56
                         (0.09)
                           5.97

     Q10        - -       0.65
                         (0.09)
                           7.35

     Q11        - -       0.62
                         (0.09)
                           6.78

     Q17        - -       0.68
                         (0.09)
                           7.75

     Q18        - -       0.70
                         (0.09)
                           7.97


        PHI
        Note: This matrix is diagonal.

            FactOne    FactTwo
           --------   --------
              1.00       1.00
```

Fig. 4.18 (continued)

THETA-DELTA

	Q5	Q7	Q8	Q12	Q13	Q14
Q5	0.56 (0.08) 7.04					
Q7	- -	0.50 (0.08) 6.32				
Q8	- -	0.16 (0.06) 2.74	0.32 (0.06) 4.91			
Q12	- -	- -	- -	0.36 (0.07) 5.55		
Q13	- -	- -	- -	- -	0.62 (0.09) 7.22	
Q14	- -	- -	- -	- -	0.18 (0.06) 2.83	0.55 (0.08) 6.92
Q6	- -	- -	- -	- -	- -	- -
Q9	- -	- -	- -	- -	- -	- -
Q10	- -	- -	- -	- -	- -	- -
Q11	- -	- -	- -	- -	- -	- -
Q17	- -	- -	- -	- -	- -	- -
Q18	- -	- -	- -	- -	- -	- -

THETA-DELTA

	Q6	Q9	Q10	Q11	Q17	Q18
Q6	0.68 (0.10) 6.87					
Q9	0.25 (0.08) 3.21	0.69 (0.10) 6.91				
Q10	- -	- -	0.57 (0.09) 6.31			
Q11	0.24 (0.08) 3.08	0.36 (0.08) 4.40	- -	0.62 (0.10) 6.53		
Q17	- -	- -	- -	- -	0.53 (0.09) 5.96	
Q18	- -	- -	- -	- -	- -	0.51 (0.09) 5.74

Fig. 4.18 (continued)

```
        Squared Multiple Correlations for X - Variables

              Q5         Q7         Q8        Q12        Q13        Q14
        --------   --------   --------   --------   --------   --------
             0.44       0.50       0.68       0.64       0.38       0.45

        Squared Multiple Correlations for X - Variables

              Q6         Q9        Q10        Q11        Q17        Q18
        --------   --------   --------   --------   --------   --------
             0.32       0.31       0.43       0.38       0.47       0.49


                        Goodness of Fit Statistics

                          Degrees of Freedom = 49
              Minimum Fit Function Chi-Square = 84.34 (P = 0.0013)
    Normal Theory Weighted Least Squares Chi-Square = 77.75 (P = 0.0055)
              Estimated Non-centrality Parameter (NCP) = 28.75
          90 Percent Confidence Interval for NCP = (8.62 ; 56.81)

                    Minimum Fit Function Value = 0.63
              Population Discrepancy Function Value (F0) = 0.22
            90 Percent Confidence Interval for F0 = (0.065 ; 0.43)
          Root Mean Square Error of Approximation (RMSEA) = 0.066
        90 Percent Confidence Interval for RMSEA = (0.036 ; 0.093)
           P-Value for Test of Close Fit (RMSEA < 0.05) = 0.16

              Expected Cross-Validation Index (ECVI) = 1.02
          90 Percent Confidence Interval for ECVI = (0.87 ; 1.23)
                      ECVI for Saturated Model = 1.17
                    ECVI for Independence Model = 5.81

    Chi-Square for Independence Model with 66 Degrees of Freedom = 748.31
                          Independence AIC = 772.31
                              Model AIC = 135.75
                            Saturated AIC = 156.00
                          Independence CAIC = 819.08
                             Model CAIC = 248.79
                           Saturated CAIC = 460.03

                  Root Mean Square Residual (RMR) = 0.17
                        Standardized RMR = 0.17
                    Goodness of Fit Index (GFI) = 0.91
                Adjusted Goodness of Fit Index (AGFI) = 0.86
              Parsimony Goodness of Fit Index (PGFI) = 0.57

                       Normed Fit Index (NFI) = 0.89
                    Non-Normed Fit Index (NNFI) = 0.93
                Parsimony Normed Fit Index (PNFI) = 0.66
                   Comparative Fit Index (CFI) = 0.95
                   Incremental Fit Index (IFI) = 0.95
                    Relative Fit Index (RFI) = 0.85

                         Critical N (CN) = 119.15
```

Fig. 4.18 (continued)

```
          Fitted Covariance Matrix

               Q5       Q7       Q8       Q12      Q13      Q14
          -------- -------- -------- -------- -------- --------
       Q5       1.00
       Q7       0.47     1.00
       Q8       0.55     0.74     1.00
      Q12       0.53     0.56     0.66     1.00
      Q13       0.41     0.43     0.51     0.49     1.00
      Q14       0.45     0.48     0.56     0.54     0.59     1.00
       Q6        - -      - -      - -      - -      - -      - -
       Q9        - -      - -      - -      - -      - -      - -
      Q10        - -      - -      - -      - -      - -      - -
      Q11        - -      - -      - -      - -      - -      - -
      Q17        - -      - -      - -      - -      - -      - -
      Q18        - -      - -      - -      - -      - -      - -

          Fitted Covariance Matrix

               Q6       Q9       Q10      Q11      Q17      Q18
          -------- -------- -------- -------- -------- --------
       Q6       1.00
       Q9       0.56     1.00
      Q10       0.37     0.36     1.00
      Q11       0.58     0.70     0.40     1.00
      Q17       0.38     0.38     0.45     0.42     1.00
      Q18       0.39     0.39     0.46     0.43     0.48     1.00

          Fitted Residuals

               Q5       Q7       Q8       Q12      Q13      Q14
          -------- -------- -------- -------- -------- --------
       Q5       0.00
       Q7      -0.01     0.00
       Q8       0.02     0.00     0.00
      Q12       0.00     0.03    -0.02     0.00
      Q13       0.02    -0.03     0.00     0.00     0.00
      Q14      -0.05    -0.04     0.02     0.02     0.00     0.00
       Q6      -0.13    -0.27    -0.20    -0.36    -0.06    -0.19
       Q9      -0.17    -0.26    -0.18    -0.38    -0.08    -0.11
      Q10      -0.13    -0.27    -0.22    -0.40    -0.19    -0.26
      Q11      -0.26    -0.25    -0.23    -0.36    -0.18    -0.19
      Q17      -0.19    -0.29    -0.32    -0.34    -0.26    -0.32
      Q18      -0.20    -0.27    -0.21    -0.40    -0.10    -0.22

          Fitted Residuals

               Q6       Q9       Q10      Q11      Q17      Q18
          -------- -------- -------- -------- -------- --------
       Q6       0.00
       Q9       0.00     0.00
      Q10      -0.01    -0.03     0.00
      Q11       0.00     0.00     0.01     0.00
      Q17       0.00     0.03    -0.01     0.01     0.00
      Q18       0.01     0.00     0.01    -0.01    -0.01     0.00

 Summary Statistics for Fitted Residuals

 Smallest Fitted Residual =   -0.40
   Median Fitted Residual =   -0.03
  Largest Fitted Residual =    0.03

…
```

Fig. 4.18 (continued)

Modification Indices and Expected Change

Modification Indices for LAMBDA-X

	FactOne	FactTwo
Q5	- -	0.20
Q7	- -	2.35
Q8	- -	1.02
Q12	- -	15.73
Q13	- -	0.49
Q14	- -	0.89
Q6	0.20	- -
Q9	0.00	- -
Q10	1.35	- -
Q11	0.82	- -
Q17	3.42	- -
Q18	0.78	- -

Expected Change for LAMBDA-X

	FactOne	FactTwo
Q5	- -	0.03
Q7	- -	-0.10
Q8	- -	0.06
Q12	- -	-0.27
Q13	- -	0.05
Q14	- -	-0.07
Q6	-0.03	- -
Q9	0.00	- -
Q10	-0.09	- -
Q11	-0.06	- -
Q17	-0.14	- -
Q18	-0.07	- -

Modification Indices for PHI

	FactOne	FactTwo
FactOne	- -	
FactTwo	25.54	- -

Expected Change for PHI

	FactOne	FactTwo
FactOne	- -	
FactTwo	-0.54	- -

Modification Indices for THETA-DELTA

	Q5	Q7	Q8	Q12	Q13	Q14
Q5	- -					
Q7	0.42	- -				
Q8	0.99	- -	- -			
Q12	0.00	3.59	4.40	- -		
Q13	0.94	0.21	0.01	0.04	- -	
Q14	2.41	1.07	1.55	0.52	- -	- -
Q6	1.18	1.09	0.17	1.21	3.24	1.12
Q9	0.27	1.11	0.94	4.50	0.55	1.37
Q10	1.81	0.33	0.80	3.39	0.09	0.33
Q11	4.05	0.76	0.19	1.19	2.12	0.24
Q17	0.35	0.37	2.98	1.85	1.56	1.03
Q18	0.34	0.47	1.19	3.71	2.34	0.04

Fig. 4.18 (continued)

```
         Modification Indices for THETA-DELTA

                Q6          Q9         Q10         Q11         Q17         Q18
          --------    --------    --------    --------    --------    --------
       Q6       - -
       Q9       - -         - -
      Q10      0.00        0.94         - -
      Q11       - -         - -        0.52         - -
      Q17      0.10        0.78        0.03        0.06         - -
      Q18      0.12        0.00        0.16        0.19        0.05         - -

         Expected Change for THETA-DELTA

                Q5          Q7          Q8          Q12         Q13         Q14
          --------    --------    --------    --------    --------    --------
       Q5       - -
       Q7     -0.03         - -
       Q8      0.05         - -         - -
      Q12      0.00        0.09       -0.10         - -
      Q13      0.05       -0.02        0.00       -0.01         - -
      Q14     -0.08       -0.05        0.05        0.04         - -         - -
       Q6      0.06       -0.05        0.02       -0.05        0.09       -0.05
       Q9      0.02       -0.04        0.03       -0.09        0.03        0.05
      Q10      0.08       -0.03        0.04       -0.09       -0.02       -0.03
      Q11     -0.09        0.03       -0.01        0.04       -0.06        0.02
      Q17      0.03        0.03       -0.07        0.07       -0.07       -0.05
      Q18     -0.03       -0.03        0.04       -0.09        0.08       -0.01

         Expected Change for THETA-DELTA

                Q6          Q9         Q10         Q11         Q17         Q18
          --------    --------    --------    --------    --------    --------
       Q6       - -
       Q9       - -         - -
      Q10      0.00       -0.05         - -
      Q11       - -         - -        0.04         - -
      Q17     -0.02        0.05       -0.01       -0.01         - -
      Q18      0.02        0.00        0.03       -0.02       -0.02         - -

 Maximum Modification Index is   25.54 for Element ( 2, 1) of PHI


 Covariances

         X - KSI

                Q5          Q7          Q8          Q12         Q13         Q14
          --------    --------    --------    --------    --------    --------
  FactOne     0.67        0.71        0.83        0.80        0.62        0.67
  FactTwo      - -         - -         - -         - -         - -         - -

         X - KSI

                Q6          Q9         Q10         Q11         Q17         Q18
          --------    --------    --------    --------    --------    --------
  FactOne      - -         - -         - -         - -         - -         - -
  FactTwo     0.56        0.56        0.65        0.62        0.68        0.70


 Factor Scores Regressions

         KSI

                Q5          Q7          Q8          Q12         Q13         Q14
          --------    --------    --------    --------    --------    --------
  FactOne     0.17        0.10        0.32        0.31        0.10        0.14
  FactTwo      - -         - -         - -         - -         - -         - -
         KSI

                Q6          Q9         Q10         Q11         Q17         Q18
          --------    --------    --------    --------    --------    --------
  FactOne      - -         - -         - -         - -         - -         - -
  FactTwo     0.11        0.07        0.26        0.14        0.29        0.31
```

Fig. 4.18 (continued)

```
! FactorOne vs. FactorTwo in AMOS with non-zero Theta-Deltas
$Standardized
$Smc

$Structure
Q5  = ( 1 )    FactorOne + (1) eps5
Q7  =  FactorOne + (1) eps7
Q8  =  FactorOne + (1) eps8
Q12 =  FactorOne + (1) eps12
Q13 =  FactorOne + (1) eps13
Q14 =  FactorOne + (1) eps14

Q6  = ( 1 )    FactorTwo + (1) eps6
Q9  =  FactorTwo + (1) eps9
Q10 =  FactorTwo + (1) eps10
Q11 =  FactorTwo + (1) eps11
Q17 =  FactorTwo + (1) eps17
Q18 =  FactorTwo + (1) eps18

eps8  <> eps7
eps13 <> eps14
eps6  <> eps9
eps6  <> eps11
eps9  <> eps11

$Include = Examp4-5.amd
```

Fig. 4.19 AMOS input example for confirmatory factor analytic model with two factors (examp4-5.ami)

```
!Examp4-6.spl
!Raw Data From File: Examp4-2.txt
DA NI=12 MA = CM XM = 9 NO=145
RA FI=F:\WORK_STATA\SAMD\Chapter4_CFA\Examp4-2.txt
LA
Q5 Q7 Q8  Q12 Q13 Q14
Q6 Q9 Q10 Q11 Q17 Q18
MO NY = 12 NE = 2 NK = 1 GA = FI PH = ST
LE
FactorOne
FactorTwo
LK
SecdOrder
VA 1 LY 1 1 LY 7 2
FR LY(2,1) LY(3,1) LY(4,1) LY(5,1) LY(6,1) LY(8,2) LY(9,2) LY(10,2) C
LY(11,2) LY(12,2) GA(2,1) GA(1,1)
ST 1 ALL
Path Diagram
OU SS NS
```

Fig. 4.20 Input for second-order factor analysis using LISREL (examp4-6.spl)

```
…
sem (FactorOne -> Q5@1 Q7 Q8 Q12 Q13 Q14) ///
(FactorTwo -> Q6@1 Q9 Q10 Q11 Q17 Q18) ///
(FactorOne@1 FactorTwo <- SecondOrder) ///
, nomeans latent(FactorOne FactorTwo SecondOrder)
…
```

Fig. 4.21 Input for second-order factor analysis using STATA (examp4-6_Mac.do)

```
                                LISREL 8.80 (STUDENT EDITION)

                                          BY

                                Karl G. J^reskog & Dag S^rbom

                          This program is published exclusively by
                          Scientific Software International, Inc.
                             7383 N. Lincoln Avenue, Suite 100
                              Lincolnwood, IL 60712, U.S.A.
                    Phone: (800)247-6113, (847)675-0720, Fax: (847)675-2140
                  Copyright by Scientific Software International, Inc., 1981-2006
                    Use of this program is subject to the terms specified in the
                                Universal Copyright Convention.
                                  Website: www.ssicentral.com

 The following lines were read from file F:\WORK_STATA\SAMD\CHAPTER4_CFA\Examp4-6.spl:

 !Examp4-6.spl
 !Raw Data From File: Examp4-2.txt
 DA NI=12 MA = CM XM = 9 NO=145
 RA FI=F:\WORK_STATA\SAMD\Chapter4_CFA\Examp4-2.txt

                       ---------------------------------
                        EM Algorithm for missing Data:
                       ---------------------------------
```

Fig. 4.22 LISREL output for second-order factor analytic model (examp4-6.out)

the file, including the directory path. The first line indicates the label for the first group (country, in this case).

The second line indicates that the data contain five indicators ("NI = 5"), that there are three groups ("NG = 3"), that the number of observations in the first group is 226 ("NO = 226"), and that the covariance matrix is analyzed ("MA = CM").

The model line (which starts with "MO") indicates that there are five x indicators (observed items) and one factor ξ ("NK = 1"), and that tau is estimated ("TA = FR") but kappa is fixed ("KA = FI"). $\boldsymbol{\Theta}_{\delta}$ is specified as symmetric because we estimate some of the covariance terms that appear to be non-zero.

We label the factor as "SWB" for subjective well-being, below the line LK for Label Xsi. The lambda matrix is then specified with five rows of 1s and the first value is fixed to the value 1 (the line "FI LX 1 1" fixes the parameter and the line "VA 1 LX 1 1" sets it to the value 1). The diagonal elements of the measurement error covariance matrix are then freed so that these elements can be estimated (as well as one of the covariances).

Then the output line "OU MI" requests that the modification indices be included in the output.

Similar information is then entered in turn for the other two groups, except that some of the parameters do not need to be repeated.

The path diagram is requested through the "PD" command.

For this unconstrained analysis, the CFA is conducted separately country by country. The chi-square for the three countries is the sum of the chi-squares for each of the three groups.

```
          Percentage missing values=   0.60

     Note:
       The Covariances and/or Means to be analyzed are estimated
       by the EM procedure and are only used to obtain starting
       values for the FIML procedure

LA
Q5 Q7 Q8  Q12 Q13 Q14
Q6 Q9 Q10 Q11 Q17 Q18

MO NY = 12 NE = 2 NK = 1 GA = FI PH = ST
LE
FactorOne
FactorTwo
LK
SecdOrder
VA 1 LY 1 1 LY 7 2
FR LY(2,1) LY(3,1) LY(4,1) LY(5,1) LY(6,1) LY(8,2) LY(9,2) LY(10,2) C
LY(11,2) LY(12,2) GA(2,1) GA(1,1)
ST 1 ALL
Path Diagram
OU SS NS

!Examp4-6.spl

                        Number of Input Variables 12
                        Number of Y - Variables   12
                        Number of X - Variables    0
                        Number of ETA - Variables  2
                        Number of KSI - Variables  1
                        Number of Observations   140


      Covariance Matrix

              Q5         Q7         Q8        Q12        Q13        Q14
        --------   --------   --------   --------   --------   --------
Q5         4.25
Q7         1.84       3.78
Q8         2.20       2.66       3.46
Q12        2.08       2.24       2.28       3.58
Q13        1.61       1.40       1.70       1.72       3.22
Q14        1.47       1.60       1.92       1.88       1.93       3.35
Q6        -0.50      -0.85      -0.56      -1.15      -0.18      -0.71
Q9        -0.54      -0.75      -0.43      -1.02      -0.22      -0.32
Q10       -0.37      -0.74      -0.52      -1.07      -0.54      -0.72
Q11       -0.83      -0.74      -0.65      -1.05      -0.57      -0.60
Q17       -0.62      -0.94      -0.89      -1.02      -0.78      -0.94
Q18       -0.65      -0.76      -0.64      -1.26      -0.33      -0.74

      Covariance Matrix

              Q6         Q9        Q10        Q11        Q17        Q18
        --------   --------   --------   --------   --------   --------
Q6         3.04
Q9         1.71       2.56
Q10        1.09       0.86       2.36
Q11        1.74       1.89       1.12       2.90
Q17        1.24       1.10       1.14       1.19       2.73
Q18        1.04       1.08       1.26       1.28       1.34       2.82
```

Fig. 4.22 (continued)

Because no constraints are imposed, the construct means (one mean value for each country) cannot be estimated and the mean for each country is zero. Figure 4.25 gives the values of the parameters estimated by LISREL on a graphical representation of the model.

It is clear from Fig. 4.25 that the estimated loading parameters are country specific.

```
Parameter Specifications

        LAMBDA-Y

           FactorOn   FactorTw
           --------   --------
      Q5          0          0
      Q7          1          0
      Q8          2          0
     Q12          3          0
     Q13          4          0
     Q14          5          0
      Q6          0          0
      Q9          0          6
     Q10          0          7
     Q11          0          8
     Q17          0          9
     Q18          0         10

        GAMMA

           SecdOrde
           --------
FactorOn         11
FactorTw         12

        PSI

           FactorOn   FactorTw
           --------   --------
                 13         14

        THETA-EPS

                 Q5         Q7         Q8        Q12        Q13        Q14
           --------   --------   --------   --------   --------   --------
                 15         16         17         18         19         20

        THETA-EPS

                 Q6         Q9        Q10        Q11        Q17        Q18
           --------   --------   --------   --------   --------   --------
                 21         22         23         24         25         26


Number of Iterations = 24
```

Fig. 4.22 (continued)

The input corresponding to unconstrained multi-group analysis in STATA is shown in Fig. 4.26.

An alternative (also STATA) is shown in Fig. 4.27. Although the commands appear more complex, they have the advantage of being precise and explicit about the parameters to be estimated. The commands used in Fig. 4.27 are especially useful to know, as they are needed for the subsequent models that make parameter restrictions across groups.

The STATA output corresponding to the input of Fig. 4.26 is listed in Fig. 4.28.

In metric invariance, the factor loadings are constrained to be the same across groups. The scalar values tau can, however, vary across groups, which makes it impossible to assess different means for the construct across groups. Figure 4.29

```
LISREL Estimates (Maximum Likelihood)

        LAMBDA-Y

            FactorOn   FactorTw
            --------   --------
      Q5        1.00        - -
      Q7        1.14        - -
              (0.15)
                7.64
      Q8        1.21        - -
              (0.15)
                8.21
     Q12        1.12        - -
              (0.14)
                7.75
     Q13        0.84        - -
              (0.13)
                6.38
     Q14        0.92        - -
              (0.14)
                6.76
      Q6         - -        1.00
      Q9         - -        0.99
                          (0.12)
                            8.44
     Q10         - -        0.69
                          (0.11)
                            6.18
     Q11         - -        1.09
                          (0.13)
                            8.68
     Q17         - -        0.78
                          (0.12)
                            6.47
     Q18         - -        0.78
                          (0.12)
                            6.40

        GAMMA

            SecdOrde
            --------
FactorOn        0.83
              (0.26)
                3.22
FactorTw       -0.82
              (0.24)
               -3.48

        Covariance Matrix of ETA and KSI

            FactorOn   FactorTw   SecdOrde
            --------   --------   --------
FactorOn        1.76
FactorTw       -0.69       1.59
SecdOrde        0.83      -0.82       1.00
```

Fig. 4.22 (continued)

lists the LISREL input to run such a partially constrained model (the example in STATA follows in Fig. 4.31).

The input in Fig. 4.29 is identical to the unconstrained estimation, except for the statement concerning the factor loadings in the second and third groups. Indeed, for these two countries, the statement "LX = IN" indicates that these parameters must be constrained to be invariant, i.e., equal across groups. Figure 4.30 provides the LISREL output for this problem.

```
PHI

   SecdOrde
   --------
       1.00

PSI
Note: This matrix is diagonal.

   FactorOn   FactorTw
   --------   --------
       1.07       0.91
                 (0.39)
                   2.33

Squared Multiple Correlations for Structural Equations

   FactorOn   FactorTw
   --------   --------
       0.39       0.43

THETA-EPS

         Q5         Q7         Q8        Q12        Q13        Q14
   --------   --------   --------   --------   --------   --------
       2.46       1.49       0.88       1.33       1.97       1.85
     (0.32)     (0.22)     (0.17)     (0.20)     (0.26)     (0.25)
       7.61       6.74       5.34       6.60       7.70       7.50

THETA-EPS

         Q6         Q9        Q10        Q11        Q17        Q18
   --------   --------   --------   --------   --------   --------
       1.42       0.98       1.59       0.99       1.75       1.83
     (0.21)     (0.16)     (0.21)     (0.17)     (0.23)     (0.24)
       6.79       6.13       7.67       5.72       7.52       7.60

Squared Multiple Correlations for Y - Variables

         Q5         Q7         Q8        Q12        Q13        Q14
   --------   --------   --------   --------   --------   --------
       0.42       0.60       0.74       0.63       0.39       0.44

Squared Multiple Correlations for Y - Variables

         Q6         Q9        Q10        Q11        Q17        Q18
   --------   --------   --------   --------   --------   --------
       0.53       0.62       0.32       0.66       0.36       0.35


     Global Goodness of Fit Statistics, Missing Data Case


       -2ln(L) for the saturated model =        5842.057
       -2ln(L) for the fitted model    =        5958.628

                  Degrees of Freedom = 52
     Full Information ML Chi-Square  = 116.57 (P = 0.00)
   Root Mean Square Error of Approximation (RMSEA) = 0.094
  90 Percent Confidence Interval for RMSEA = (0.071 ; 0.12)
    P-Value for Test of Close Fit (RMSEA < 0.05) = 0.0013
```

Fig. 4.22 (continued)

Although the error variances vary across countries, the factor loadings are identical, i.e., invariant. As indicated in Fig. 4.28, the means of the unobserved factors are still zero for each group.

The STATA input for metric invariance is shown in Fig. 4.31. The STATA output is not shown since it gives the same results as that of LISREL.

```
Standardized Solution

        LAMBDA-Y

           FactorOn   FactorTw
           --------   --------
      Q5       1.33        - -
      Q7       1.51        - -
      Q8       1.60        - -
     Q12       1.49        - -
     Q13       1.11        - -
     Q14       1.22        - -
      Q6        - -       1.26
      Q9        - -       1.25
     Q10        - -       0.87
     Q11        - -       1.37
     Q17        - -       0.98
     Q18        - -       0.98

        GAMMA

           SecdOrde
           --------
FactorOn       0.63
FactorTw      -0.65

        Correlation Matrix of ETA and KSI

           FactorOn   FactorTw   SecdOrde
           --------   --------   --------
FactorOn       1.00
FactorTw      -0.41       1.00
SecdOrde       0.63      -0.65       1.00

        PSI
        Note: This matrix is diagonal.

           FactorOn   FactorTw
           --------   --------
               0.61       0.57
```

Fig. 4.22 (continued)

In the scalar invariance model, the factor loadings are equal, i.e., invariant across groups as in metric invariance. However, in addition, the scalars corresponding to the elements of tau are also invariant. We illustrate scalar invariance estimation in LISREL and then in STATA. In LISREL, tau is specified as invariant (i.e., equal across groups) by indicating "TX = IN" for the last two groups, as shown in Fig. 4.32.

The means are then shown in Fig. 4.33.

It can be seen from Fig. 4.33 that the means of the SWB factor in the USA and Austria are almost the same (zero for the USA and close to zero for Austria but slightly below as indicated by the negative sign before the 0.00). However, the mean is −0.58 for SWB in Australia, indicating an inferior perception of well-being in that country relative to the USA and Austria.

The input in STATA that is equivalent to the LISREL input (Fig. 4.32) is shown in Fig. 4.34. The STATA output is not shown, as it gives the same results as the LISREL output.

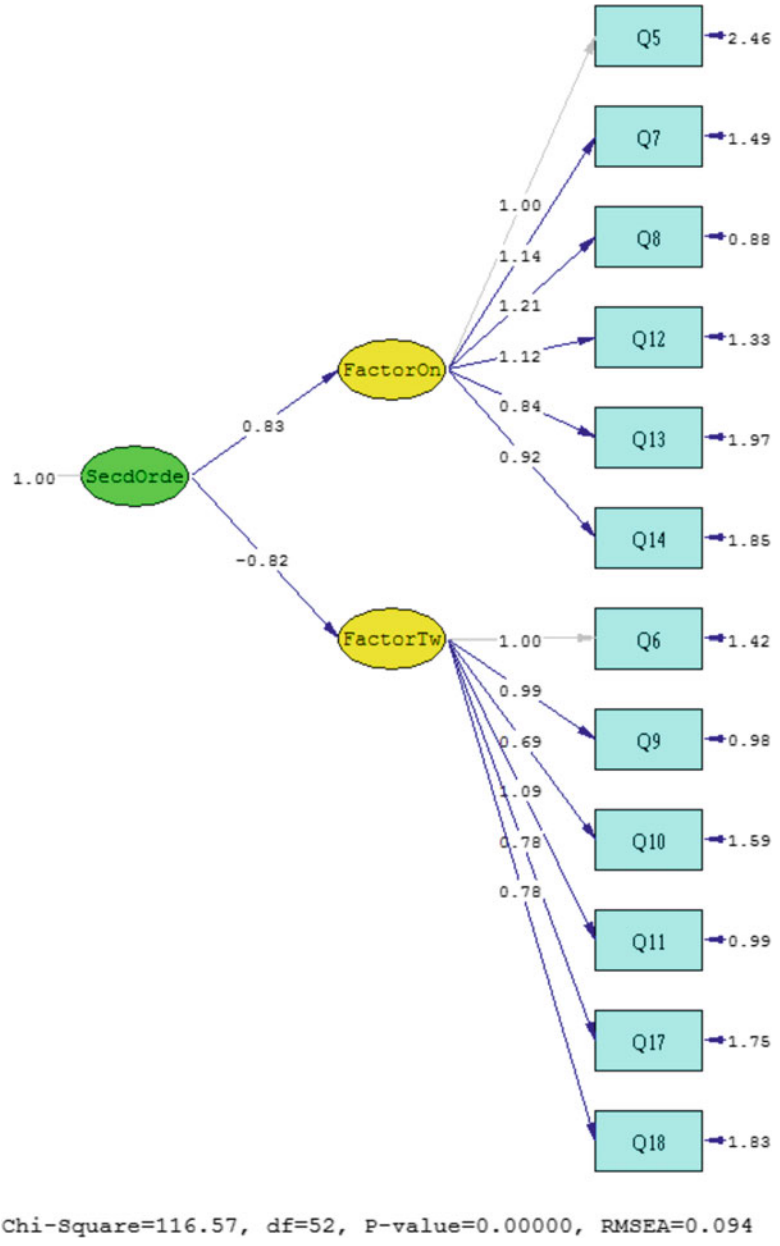


Fig. 4.23 Second-order factor analytic model in LISREL (examp4-6.pth)

```
USAM
DA NI=5 NG=3 NO=226 MA=CM
RA=C:\SAMD\CHAPTER4\EXAMPLES\usam.txt
MO NX=5 NK=1 TX=FR KA=FI TD=SY,FI
LK
SWB
PA LX
5(1)
FI LX 1 1
VA 1 LX 1 1
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5 TD 5 4
OU MI
AUSTRIAM
DA NO=63
RA=C:\SAMD\CHAPTER4\EXAMPLES\austriam.txt
MO LX=FR TX=FR KA=FI TD=SY,FI
LK
SWB
PA LX
5(1)
FI LX 1 1
VA 1 LX 1 1
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5 TD 5 1
OU
AUSTRALIAM
DA NO=56
RA=C:\SAMD\CHAPTER4\EXAMPLES\australiam.txt
MO LX=FR TX=FR KA=FI TD=SY,FI
LK
SWB
PA LX
5(1)
FI LX 1 1
VA 1 LX 1 1
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5
PD
OU
```

Fig. 4.24 Unconstrained CFA for subjective well-being of men in three countries-LISREL (examp4-7.ls8)

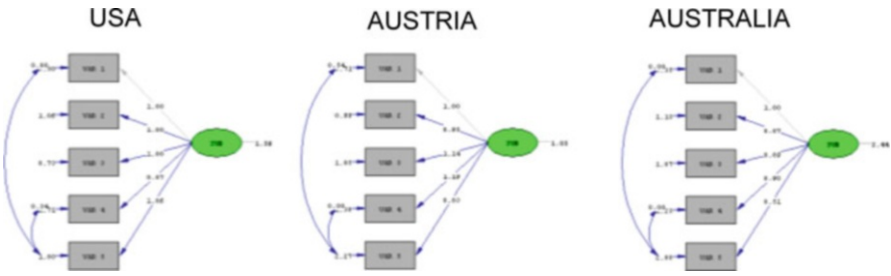


Fig. 4.25 Unconstrained estimates from LISREL (examp4-7.pth)

```
cd "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter4_CFA/"
use allcty, clear
*  Unconstrained model
sem (SWB -> var1 var2 var3 var4 var5) ///
, group(country) ginvariant(meanex) ///
cov(0:e.var1*e.var5) cov(0:e.var4*e.var5)
estat ggof
estat gof, stats(all)
estat mindices, min(1)
estat framework, fitted
```

Fig. 4.26 Unconstrained CFA for subjective well-being of men in three countries-STATA (examp4-7_Mac_Unconstrained.do)

```
cd "/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter4_CFA/"
use allcty, clear
* metric invariance model
sem (0:var1 <- SWB@a1) (0:var2 <- SWB@b1) (0:var3 <- SWB@c1) (0:var4 <- SWB@d1)
(0:var5 <- SWB@f1) ///
(1:var1 <- SWB@a2) (1:var2 <- SWB@b2) (1:var3 <- SWB@c2) (1:var4 <- SWB@d2) (1:var5 <-
SWB@f2) ///
(2:var1 <- SWB@a3) (2:var2 <- SWB@b3) (2:var3 <- SWB@c3) (2:var4 <- SWB@d3) (2:var5 <-
SWB@f3) ///
(0:var1 <- _cons@ca1) (0:var2 <- _cons@cb1) (0:var3 <- _cons@cc1) (0:var4 <-
_cons@cd1) (0:var5 <- _cons@cf1) ///
(1:var1 <- _cons@ca2) (1:var2 <- _cons@cb2) (1:var3 <- _cons@cc2) (1:var4 <-
_cons@cd2) (1:var5 <- _cons@cf2) ///
(2:var1 <- _cons@ca3) (2:var2 <- _cons@cb3) (2:var3 <- _cons@cc3) (2:var4 <-
_cons@cd3) (2:var5 <- _cons@cf3) ///
, group(country) ginvariant(meanex) ///
cov(0:e.var1*e.var5) cov(0:e.var4*e.var5)
estat ggof
estat gof, stats(all)
estat mindices, min(1)
estat framework, fitted
```

Fig. 4.27 Unconstrained CFA for subjective well-being of men in three countries-STATA alternative (examp4-7_Mac_Unconstrained_Alt.do)

```
Endogenous variables

Measurement:  var1 var2 var3 var4 var5

Exogenous variables

Latent:       SWB

Fitting target model:

Iteration 0:   log likelihood = -2902.6334
Iteration 1:   log likelihood = -2898.5623
Iteration 2:   log likelihood = -2898.1069
Iteration 3:   log likelihood = -2898.1066
Iteration 4:   log likelihood = -2898.1066

Structural equation model                       Number of obs      =       345
Grouping variable  = country                    Number of groups   =         3
Estimation method  = ml
Log likelihood     = -2898.1066

 ( 1)  [var1]0bn.country#c.SWB = 1
 ( 2)  [var1]1.country#c.SWB = 1
 ( 3)  [var1]2.country#c.SWB = 1
 ( 4)  [cov(e.var1,e.var5)]1.country = 0
 ( 5)  [cov(e.var4,e.var5)]1.country = 0
 ( 6)  [cov(e.var1,e.var5)]2.country = 0
 ( 7)  [cov(e.var4,e.var5)]2.country = 0
------------------------------------------------------------------------------
                |                 OIM
                |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
Measurement     |
  var1 <-       |
    SWB         |
            [*] |          1  (constrained)
    _cons       |
      USA       |   4.823009   .1019545    47.31   0.000     4.623182    5.022836
      AUSTRALIA |   4.160714   .2212799    18.80   0.000     3.727014    4.594415
```

Fig. 4.28 STATA output of unconstrained CFA for subjective well-being of men in three countries (examp4-7_Unconstrained.log)

```
    SWB          |
      USA        |  .9916609   .0821521    12.07   0.000     .8306458    1.152676
      AUSTRALIA  |  .8687131   .1141639     7.61   0.000     .6449559     1.09247
      AUSTRIA    |  .8041267   .1683849     4.78   0.000     .4740982    1.134155
    _cons        |
      USA        |  4.858407   .1072467    45.30   0.000     4.648207    5.068607
      AUSTRALIA  |  4.589286   .2283715    20.10   0.000     4.141686    5.036886
      AUSTRIA    |         5   .1738802    28.76   0.000     4.659201    5.340799
-----------------+------------------------------------------------------------------
var3 <-          |
    SWB          |
      USA        |  .9940045   .0770689    12.90   0.000     .8429523    1.145057
      AUSTRALIA  |  .8857529   .1124612     7.88   0.000     .6653331    1.106173
      AUSTRIA    |  1.013393   .1757259     5.77   0.000     .6689769     1.35781
    _cons        |
      USA        |  5.207965    .100631    51.75   0.000     5.010732    5.405198
      AUSTRALIA  |  4.535714   .2286332    19.84   0.000     4.087601    4.983827
      AUSTRIA    |  5.095238   .1940333    26.26   0.000      4.71494    5.475536
-----------------+------------------------------------------------------------------
var4 <-          |
    SWB          |
      USA        |   .863139   .0920428     9.38   0.000     .6827384     1.04354
      AUSTRALIA  |  .9015024   .1154084     7.81   0.000     .6753061    1.127699
      AUSTRIA    |  1.043834   .1875264     5.57   0.000     .6762893    1.411379
    _cons        |
      USA        |  4.743363   .1127333    42.08   0.000      4.52241    4.964316
      AUSTRALIA  |  4.053571    .237466    17.07   0.000     3.588147    4.518996
      AUSTRIA    |  4.746032   .2007397    23.64   0.000     4.352589    5.139474
-----------------+------------------------------------------------------------------
var5 <-          |
    SWB          |
      USA        |  1.064512   .1006879    10.57   0.000     .8671672    1.261856
      AUSTRALIA  |  .5146027   .1526025     3.37   0.001     .2155072    .8136981
      AUSTRIA    |  .9498816   .1853836     5.12   0.000     .5865365    1.313227
    _cons        |
      USA        |  4.070796   .1248546    32.60   0.000     3.826086    4.315507
      AUSTRALIA  |  4.071429   .2459586    16.55   0.000     3.589359    4.553499
      AUSTRIA    |  4.333333   .2164791    20.02   0.000     3.909042    4.757625
-----------------+------------------------------------------------------------------
Variance         |
  e.var1         |
      USA        |  .7790827   .1111476                      .5890432    1.030433
      AUSTRALIA  |  .3446865   .1640343                      .1356258    .8760043
      AUSTRIA    |   .500845   .1698938                      .2576105    .9737404
  e.var2         |
      USA        |  1.055376   .1273666                      .8330689    1.337006
      AUSTRALIA  |  1.111416   .2521222                      .7124975    1.733683
      AUSTRIA    |  1.105465   .2349426                      .7288538    1.676677
  e.var3         |
      USA        |  .7372576   .1046776                      .5581664    .9738115
      AUSTRALIA  |  1.046442   .2425484                      .6643857      1.6482
      AUSTRIA    |  1.102432   .2490009                      .7081011    1.716361
  e.var4         |
      USA        |  1.702435   .1814478                      1.381492    2.097939
      AUSTRALIA  |  1.209509   .2661402                      .7857979     1.86169
      AUSTRIA    |  1.191815    .275463                      .7576547    1.874762
  e.var5         |
      USA        |    1.7438   .2192126                      1.362989    2.231007
      AUSTRALIA  |  2.752901   .5306805                      1.886702    4.016778
      AUSTRIA    |  1.837064   .3756226                      1.230495    2.742638
  SWB            |
      USA        |  1.570122   .2246098                      1.186225     2.07826
      AUSTRALIA  |  2.397342   .5356739                      1.547153    3.714725
      AUSTRIA    |  1.236116   .3297212                      .7328418    2.085012
-----------------+------------------------------------------------------------------
Covariance       |
  e.var1         |
    e.var5       |
      USA        | -.0535675    .113635    -0.47   0.637    -.2762881    .1691531
      AUSTRALIA  |         0  (constrained)
      AUSTRIA    |         0  (constrained)
-----------------+------------------------------------------------------------------
  e.var4         |
    e.var5       |
      USA        |  .3199462   .1497472     2.14   0.033     .0264471    .6134453
```

Fig. 4.28 (continued)

```
  -------------+----------------------------------------------------------------
  e.var4       |
    e.var5     |
      USA      |   .3199462   .1497472     2.14   0.033     .0264471    .6134453
      AUSTRALIA|           0  (constrained)
      AUSTRIA  |           0  (constrained)
------------------------------------------------------------------------------
Note: [*] identifies parameter estimates constrained to be equal across groups.
LR test of model vs. saturated: chi2(13)  =     22.49, Prob > chi2 = 0.0483

. estat ggof

Group-level fit statistics
------------------------------------------------------------------------------
             |        N       SRMR         CD       chi2       df     p>chi2
-------------+----------------------------------------------------------------
     country |
           0 |      226      0.016      0.877      6.560        3      0.087
           1 |       56      0.036      0.924      5.713        5      0.335
           2 |       63      0.047      0.859     10.213        5      0.069
------------------------------------------------------------------------------

. estat gof, stats(all)

------------------------------------------------------------------------------
Fit statistic        |      Value   Description
---------------------+--------------------------------------------------------
Likelihood ratio     |
         chi2_ms(13) |     22.485   model vs. saturated
            p > chi2 |      0.048
         chi2_bs(30) |    806.952   baseline vs. saturated
            p > chi2 |      0.000
---------------------+--------------------------------------------------------
Population error     |
               RMSEA |      0.080   Root mean squared error of approximation
 90% CI, lower bound |      0.007
         upper bound |      0.134
---------------------+--------------------------------------------------------
Information criteria |
                 AIC |   5890.213   Akaike's information criterion
                 BIC |   6070.860   Bayesian information criterion
---------------------+--------------------------------------------------------
Baseline comparison  |
                 CFI |      0.988   Comparative fit index
                 TLI |      0.972   Tucker-Lewis index
---------------------+--------------------------------------------------------
Size of residuals    |
                SRMR |      0.036   Standardized root mean squared residual
                  CD |      0.885   Coefficient of determination
------------------------------------------------------------------------------
Note: pclose is not reported because of multiple groups.

. estat mindices, min(1)

Modification indices

-------------------------------------------------------------------
                 |                                        Standard
                 |         MI     df   P>MI         EPC         EPC
-----------------+-------------------------------------------------
Covariance       |
  e.var1         |
    e.var2       |
      AUSTRIA    |      2.515      1   0.11  -.2477366  -.3329401
    e.var3       |
      USA        |      2.586      1   0.11   .1810327   .2388667
    e.var4       |
      USA        |      4.838      1   0.03   -.251171  -.2180934
    e.var5       |
      AUSTRIA    |      8.606      1   0.00   .5644475   .5884502
  -------------+-----------------------------------------------
```

Fig. 4.28 (continued)

```
      USA          |      1.753      1   0.19   .1756686   .1294918
      AUSTRALIA    |      1.662      1   0.20  -.3306916  -.1890556
      AUSTRIA      |      2.871      1   0.09  -.3537234  -.2482157
  -----------------+----------------------------------------------
  e.var3           |
    e.var4         |
      USA          |      2.472      1   0.12   .1636384    .146063
    e.var5         |
      USA          |      1.753      1   0.19  -.1760838  -.1552964
  -----------------+----------------------------------------------
  e.var4           |
    e.var5         |
      AUSTRALIA    |      2.453      1   0.12   .4187452   .2294827
-------------------------------------------------------------------
EPC = expected parameter change

. estat framework, fitted

Group #1 (country=0; N=226) ----------------------------------------------

Endogenous variables on endogenous variables

                 | observed
           Beta  |      var1       var2       var3       var4       var5
    -------------+-------------------------------------------------------
    observed     |
            var1 |          0
            var2 |          0          0
            var3 |          0          0          0
            var4 |          0          0          0          0
            var5 |          0          0          0          0          0
    -------------------------------------------------------------------

Exogenous variables on endogenous variables

                 | latent
           Gamma |       SWB
    -------------+-----------
    observed     |
            var1 |          1
            var2 |  .9916609
            var3 |  .9940045
            var4 |   .863139
            var5 |  1.064512
    -------------------------

Covariances of error variables

                 | observed
             Psi |    e.var1     e.var2     e.var3     e.var4     e.var5
    -------------+-------------------------------------------------------
    observed     |
          e.var1 |  .7790827
          e.var2 |          0   1.055376
          e.var3 |          0          0  .7372576
          e.var4 |          0          0          0  1.702435
          e.var5 | -.0535675          0          0  .3199462     1.7438
    -------------------------------------------------------------------

Intercepts of endogenous variables

                 | observed
           alpha |      var1       var2       var3       var4       var5
    -------------+-------------------------------------------------------
           _cons |  4.823009   4.858407   5.207965   4.743363   4.070796
    -------------------------------------------------------------------

Covariances of exogenous variables

                 | latent
             Phi |       SWB
    -------------+-----------
    latent       |
             SWB |  1.570122
    -------------------------
```

Fig. 4.28 (continued)

```
Means of exogenous variables

               | latent
         kappa |      SWB
    -----------+----------
          mean |         0
    -------------------------

Fitted covariances of observed and latent variables

               | observed                                                  | latent
         Sigma |      var1       var2       var3       var4       var5    |      SWB
    -----------+-----------------------------------------------------------+----------
    observed   |                                                           |
          var1 |   2.349205                                                |
          var2 |   1.557029    2.59942                                     |
          var3 |   1.560709   1.547694   2.288609                          |
          var4 |   1.355234   1.343933   1.347109    2.87219               |
          var5 |   1.617846   1.657476   1.661393   1.762609    3.52304    |
    -----------+-----------------------------------------------------------+----------
    latent     |                                                           |
           SWB |   1.570122   1.557029   1.560709   1.355234   1.671414    |  1.570122
    ------------------------------------------------------------------------------

Fitted means of observed and latent variables

               | observed                                                  | latent
            mu |      var1       var2       var3       var4       var5    |      SWB
    -----------+-----------------------------------------------------------+----------
            mu |   4.823009   4.858407   5.207965   4.743363   4.070796    |         0
    ------------------------------------------------------------------------------

Group #2 (country=1; N=56) ------------------------------------------------

Endogenous variables on endogenous variables

               | observed
          Beta |      var1       var2       var3       var4       var5
    -----------+-----------------------------------------------------------
    observed   |
          var1 |          0
          var2 |          0          0
          var3 |          0          0          0
          var4 |          0          0          0          0
          var5 |          0          0          0          0          0
    ------------------------------------------------------------------------

Exogenous variables on endogenous variables

               | latent
         Gamma |      SWB
    -----------+----------
    observed   |
          var1 |          1
          var2 |   .8687131
          var3 |   .8857529
          var4 |   .9015024
          var5 |   .5146027
    -------------------------

Covariances of error variables

               | observed
           Psi |    e.var1     e.var2     e.var3     e.var4     e.var5
    -----------+-----------------------------------------------------------
    observed   |
        e.var1 |   .3446865
        e.var2 |          0   1.111416
        e.var3 |          0          0   1.046442
        e.var4 |          0          0          0   1.209509
        e.var5 |          0          0          0          0   2.752901
    ------------------------------------------------------------------------

Intercepts of endogenous variables
```

Fig. 4.28 (continued)

```
             | observed
       alpha |      var1       var2       var3       var4       var5
-------------+------------------------------------------------------
       _cons |  4.160714   4.589286   4.535714   4.053571   4.071429
------------------------------------------------------------------------------

Covariances of exogenous variables

             | latent
         Phi |       SWB
-------------+----------
  latent     |
         SWB |  2.397342
-------------------------

Means of exogenous variables

             | latent
       kappa |       SWB
-------------+----------
        mean |         0
-------------------------

Fitted covariances of observed and latent variables

             | observed                                                  | latent
       Sigma |      var1       var2       var3       var4       var5 |       SWB
-------------+-----------------------------------------------------------+----------
  observed   |                                                           |
        var1 |  2.742028                                                 |
        var2 |  2.082602   2.920599                                      |
        var3 |  2.123452   1.844671   2.927296                           |
        var4 |  2.161209   1.877471   1.914297   3.157844                |
        var5 |  1.233678   1.071713   1.092734   1.112164   3.387755     |
-------------+-----------------------------------------------------------+----------
  latent     |                                                           |
         SWB |  2.397342   2.082602   2.123452   2.161209   1.233678 |  2.397342
------------------------------------------------------------------------------

Fitted means of observed and latent variables

             | observed                                                  | latent
          mu |      var1       var2       var3       var4       var5 |       SWB
-------------+-----------------------------------------------------------+----------
          mu |  4.160714   4.589286   4.535714   4.053571   4.071429 |         0
------------------------------------------------------------------------------

Group #3 (country=2; N=63) ----------------------------------------------------

Endogenous variables on endogenous variables

             | observed
        Beta |      var1       var2       var3       var4       var5
-------------+----------------------------------------------------------
  observed   |
        var1 |         0
        var2 |         0          0
        var3 |         0          0          0
        var4 |         0          0          0          0
        var5 |         0          0          0          0          0
------------------------------------------------------------------------------

Exogenous variables on endogenous variables

             | latent
       Gamma |       SWB
-------------+----------
  observed   |
        var1 |         1
        var2 |  .8041267
        var3 |  1.013393
        var4 |  1.043834
        var5 |  .9498816
-------------------------
```

Fig. 4.28 (continued)

```
       Sigma |      var1       var2       var3       var4       var5 |       SWB
-------------+------------------------------------------------------+----------
observed     |                                                      |
        var1 |  1.736961                                            |
        var2 |  .9939942   1.904762                                 |
        var3 |  1.252672   1.007307   2.371882                      |
        var4 |  1.290301   1.037565   1.307582   2.538675           |
        var5 |  1.174164   .9441768    1.18989   1.225633   2.952381 |
-------------+------------------------------------------------------+----------
latent       |                                                      |
         SWB |  1.236116   .9939942   1.252672   1.290301   1.174164 |  1.236116
-------------------------------------------------------------------------------

Fitted means of observed and latent variables

                | observed                                             | latent
           mu |      var1       var2       var3       var4       var5 |       SWB
-------------+------------------------------------------------------+----------
           mu |  4.761905          5   5.095238   4.746032   4.333333 |         0
-------------------------------------------------------------------------------
```

Fig. 4.28 (continued)

```
USAM
DA NI=5 NG=3 NO=226 MA=KM
RA=C:\SAMD\CHAPTER4\EXAMPLES\usam.txt
MO NX=5 NK=1 TX=FR KA=FI TD=SY,FI
LK
SWB
PA LX
5(1)
FI LX 1 1
VA 1 LX 1 1
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5 TD 5 4
OU MI
AUSTRIAM
DA NO=63
RA=C:\SAMD\CHAPTER4\EXAMPLES\austriam.txt
MO LX=IN TX=FR KA=FI TD=SY,FI
LK
SWB
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5 TD 5 1
OU
AUSTRALIAM
DA NO=56
RA=C:\SAMD\CHAPTER4\EXAMPLES\australiam.txt
MO LX=IN TX=FR KA=FI TD=SY,FI
LK
SWB
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5
PD
OU
```

Fig. 4.29 LISREL input for metric invariance model of subjective well-being for three countries (examp4-8.ls8)

The full outputs are not listed, as they provide the same information as in the case of single-group CFA. The chi-square of each of these models can be compared because these are nested constrained models. The difference in chi-squares with the proper difference across models in the degrees of freedom is also chi-square distributed and can serve to test the extent of the loss in fit due to the imposition

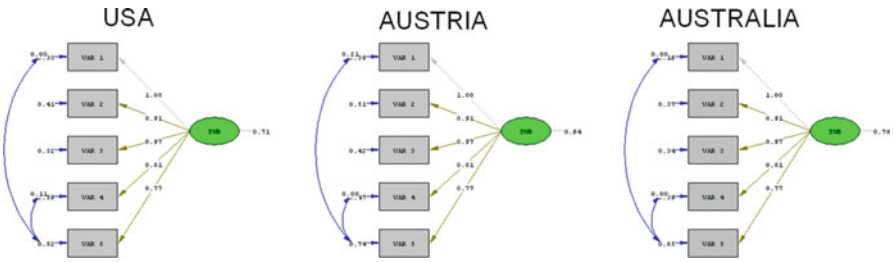


Fig. 4.30 Output from LISREL for metric invariance (examp4-8.pth)

```
cd "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter4_CFA/"
use allcty, clear
*  metric invariance model
sem (SWB -> var1 var2 var3 var4 var5) ///
, group(country) ginvariant(meanex mcoef) ///
cov(0:e.var1*e.var5) cov(0:e.var4*e.var5)
estat ggof
estat gof, stats(all)
estat mindices, min(1)
estat framework, fitted
```

Fig. 4.31 STATA input for metric invariance model of subjective well-being for three countries (examp4-7_MetricInvariance.do)

```
USAM
DA NI=5 NG=3 NO=226 MA=CM
RA=C:\SAMD\CHAPTER4\EXAMPLES\usam.txt
MO NX=5 NK=1 TX=FR KA=FI TD=SY,FI
LK
SWB
PA LX
5(1)
FI LX 1 1
VA 1 LX 1 1
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5 TD 5 4
OU MI
AUSTRIAM
DA NO=63
RA=C:\SAMD\CHAPTER4\EXAMPLES\austriam.txt
MO LX=IN TX=IN KA=FR TD=SY,FI
LK
SWB
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5 TD 5 1
OU
AUSTRALIAM
DA NO=56
RA=C:\SAMD\CHAPTER4\EXAMPLES\australiam.txt
MO LX=IN TX=IN KA=FR TD=SY,FI
LK
SWB
FR TD 1 1 TD 2 2 TD 3 3 TD 4 4 TD 5 5
PD
OU
```

Fig. 4.32 LISREL input for scalar invariance model (examp4-9.spl)

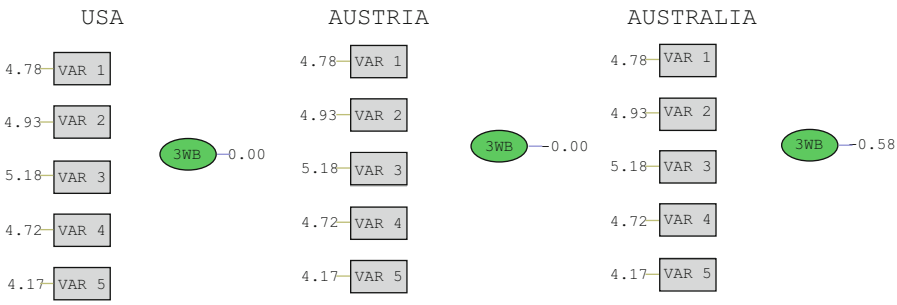


Fig. 4.33 Factor means with scalar invariance model (examp4-9.pth)

```
cd "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter4_CFA/"
use allcty, clear
* scalar invariance model
sem (SWB -> var1 var2 var3 var4 var5) ///
, group(country) ///
cov(0:e.var1*e.var5) cov(0:e.var4*e.var5)
estat ggof
estat gof, stats(all)
estat mindices, min(1)
estat framework, fitted
```

Fig. 4.34 STATA input for scalar invariance model (examp4-7_Mac_ScalarInvariance.do)

Table 4.1 Number of parameters and degrees of freedom of each model

Parameter	Unconstrained model	Metric invariance model	Scalar invariance model
Λ_x	4+4+4	4	4
Φ	1+1+1	1+1+1	1+1+1
Θ_δ	6+6+5	6+6+5	6+6+5
T	5+5+5	5+5+5	5
K	0	0	0+1+1
Number of parameters	47	39	31
Number of degrees of freedom	13	21	29
Chi-square	14.79	25.26	40.00

of the constraint. The constructs can only be compared across groups if the chi-square values are insignificant when imposing metric invariance first and scalar invariance next.

The graphical representations of the outputs of the three models under different constraints were discussed above. The corresponding basic statistics needed are (1) the number of data points to be reproduced, (2) the number of parameters to be estimated, and (3) the chi-square values for each of the models.

First, we calculate the number of data points available. For each country, there is a 5 × 5 covariance matrix, which provides 15 different data points, i.e., 45 for the three countries. In addition, there are five means for the five items for each country, i.e., 15 means. The total number of data points is, therefore, 45 + 15 = 60.

Next, we calculate the number of parameters to be estimated for each model. Table 4.1 provides the details.

In the unconstrained model, there are four lambdas to be estimated for each country (one loading must be fixed to unity to define the unit of measurement); this is indicated in the corresponding cell of the table by "4+4+4." In both the metric and scalar invariance models, only four lambdas are estimated since these lambdas are constrained to be equal across groups. The error term variances are five for each country, but for two countries a covariance term has also been estimated. This explains the "6+6+5," as no covariance is estimated for the third country.

When we subtract the number of parameters from the number of data points (i.e., 60), we obtain the degrees of freedom for each model.

Given the nested structure of these three models, it is possible to compare the extent to which imposing additional constraints makes the fit worse. When we compare the unrestricted model to the metric invariance constraint (same loadings across groups) model, the chi-square goes from 14.79 to 25.26 (a difference of 10.47), which is chi-square distributed with 8 degrees of freedom (21 − 13). The critical chi-square with 8 degrees of freedom at $\alpha = 0.05$ is 15.51. Consequently, we fail to reject this difference as significant. This supports the restriction that there is metric invariance.

Similarly, we can further evaluate the impact of the restriction that there is scalar invariance by comparing the chi-square of the metric invariance model with that of the scalar invariance model. The chi-square increases from 25.26 to 40.00 when we impose the constraint that the tau's are the same, even if we now can estimate the mean of the unobserved construct relative to one of the countries (USA) that serves as reference. The difference (40.00 − 25.26) = 14.74 is still not significant with 8 degrees of freedom (29 − 21) at $\alpha = 0.05$. We therefore infer scalar invariance, which allows us to interpret the means estimated under this scalar invariance model. Figure 4.33 shows an example of these means estimated with LISREL.

4.7 Assignment

Using the SURVEY data (Appendix C, Chap. 14), estimate the parameters of a measurement model corresponding to a CFA of two or three constructs. Include an analysis of convergent and discriminant validity.

Considering a categorical variable that distinguishes between respondents, define several groups of respondents (e.g., respondents of different ages). Then perform a multi-group analysis to test the invariance of the measurement model of your choice.

```
/*     Assign4.sas     */
filename survey 'C:\SAMD\CHAPTER4\Assignments\survey.asc';
data new;
infile survey firstobs=19;
input   (Age Marital Income Educatn HHSize Occuptn Location
         TryHair LatStyle DrssSmrt BlndsFun LookDif
         LookAttr GrocShp LikeBkng ClthFrsh WashHnds Sportng LikeClrs
         FeelAttr TooMchSx Social LikeMaid ServDnrs SaveRcps LikeKtch) (3.)
        #2 (LoveEat SpirtVal Mother ClascMsc Children Applianc ClsFamly
         LovFamly TalkChld Exercise LikeSelf CareSkin MedChckp
         EvngHome TripWrld HomeBody LondnPrs Comfort Ballet Parties
         WmnNtSmk BrghtFun Seasonng ColorTV SlppyPpl Smoke) (3.)
        #3 (Gasoline Headache Whiskey Bourbon FastFood Restrnts OutFrDnr
         OutFrLnc RentVide Catsup KnowSont PercvDif BrndLylt
         CatgMotv BrndMotv OwnSonit NecssSon OthrInfl DecsnTim
         RdWomen RdHomSrv RdFashn RdMenMag RdBusMag RdNewsMg
         RdGlMag) (3.)
        #4 (RdYouthM RdNwsppr WtchDay WtchEve WtchPrm
         WTchLate WtchWknd WtchCsby WtchFmTs WtchChrs WtchMoon
         WtchBoss WtchGrwP WtchMiaV WtchDns WtchGold WtchBowl) (3.);
data _NULL_;
set new;
TAB = ',';
FN = " C:\SAMD\CHAPTER4\Assignments\SURBSUB.CSV";
file PLOTFILE filevar=FN;
put TryHair TAB LatStyle TAB DrssSmrt TAB BlndsFun TAB LookDif;
run;
```

Fig. 4.35 SAS code example to create a new data file containing a subset of the full survey data to use with LISREL

```
use "/users/fblgatignon/Documents/WORK_STATA/SAMD/survey.dta", clear
* CFA of Survey Data
```

Fig. 4.36 STATA code example to read a new STATA data file containing the survey data

It is useful to first create a new data file that contains only the items relevant for your analysis. The SAS file listed in Fig. 4.35 shows an example of how to create such a new data file. This data file containing only the subset of relevant data can then be used with LISREL.

With STATA, the commands provided in the previous chapter can be used to read the data file created for this assignment. An alternative, once the data are saved as a STATA file ("survey.dta"), is to read the STATA file directly, as shown in Fig. 4.36.

Bibliography

Basic Technical Readings

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*(1), 74–94.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*(3), 421–458.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314.

Burke Jarvis, C., MacKenzie, S. B., & Podsakoff, P. M. (2003 September). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*, 199–218.

Diamanopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 25*(2), 186–192.

Application Readings

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research, 34*(3), 347–356.

Anderson, R. D., & Engledow, J. (1977). A factor analytic comparison of U.S. and German information seeker. *Journal of Consumer Research, 3*(4), 185–196.

Aricak, O. T., & Oakland, T. (2009). Multigroup confirmatory factor analysis for the teacher form, ages 5 to 21, of the adaptive behavior assessment system-II. *Journal of Psychoeducational Assessment, 28*(6), 578–584.

Bardo, J. W., & Hughey, J. B. (1979). A second-order factor analysis of community satisfaction in a Midwestern city. *Journal of Social Psychology, 109*(2), 231–235.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13*(2), 139–161.

Baumgartner, H., & Steenkamp, J.-B. E. M. (1998). Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. *Marketing Letters, 9*(1), 21–35.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). An extended paradigm for measurement analysis of marketing constructs applicable to panel data. *Journal of Marketing Research, 43*(3), 431–442.

Benson, J., & Bandalos, D. L. (1992). Second-order confirmatory factor analysis of the reactions to tests scale with cross-validation. *Multivariate Behavioral Research, 27*(3), 459–487.

Blackman, A. W. (1973). An innovation index based on factor analysis. *Technological Forecasting and Social Change, 4*, 301–316.

Chen, I. J., & Paulraj, A. (2004). Towards a theory of supply chain management: The constructs and measurements. *Journal of Operations Management, 22*, 119–150.

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the revised NEO personality inventory. *Journal of Personality and Social Psychology, 101*(5), 1068–1089.

Churchill, G. A., Jr. (1979 February). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*, 64–73.

De Jong, M. G., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007 August). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34*, 260–278.

Deshpande, R. (1982). The organizational context of market research use. *Journal of Marketing, 46*(4), 91–101.

Finn, A., & Kayandé, U. (1997). Reliability assessment and optimization of marketing measurement. *Journal of Marketing Research, 34*(2), 262–275.

Gilbert, F. W., & Warren, W. E. (1995). Psychographic constructs and demographic segments. *Psychology and Marketing, 12*(3), 223–237.

Green, S. G., Gavin, M. B., & Aiman-Smith, L. (1995). Assessing a multidimensional measure of radical technological innovation. *IEEE Transactions on Engineering Management, 42*(3), 203–214.

Jarvis, C. B., Mackenzie, S. B., Podsakoff, P. M., Mick, D. G., & Bearden, W. O. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(2), 199–218.

Kuester, S., Homburg, C., & Robertson, T. S. (1999). Retaliatory behavior to new product entry. *Journal of Marketing, 63*(4), 90.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73*(1), 107–117.

Murtha, T. P., Lenway, S. A., & Bagozzi, R. P. (1998). Global mind-sets and cognitive shift in a complex multinational corporation. *Strategic Management Journal, 19*, 97–114.

Myers, M. B., Calantone, R. J., Page, T. J., Jr., & Taylor, C. R. (2000). An application of multigroup causal models in assessing cross-cultural measurement equivalence. *Journal of International Marketing, 8*(4), 108–121.

Paulssen, M., & Bagozzi, R. P. (2006). Goal hierarchies as antecedents of market structure. *Psychology and Marketing, 23*(8), 689–709.

Rijkeboer, M. M., & Bergh, H. (2006). Multiple group confirmatory factor analysis of the young schema-questionnaire in a Dutch clinical versus non-clinical population. *Cognitive Therapy and Research, 30*(3), 263–278.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*(1), 51.

Rubera, G., Ordanini, A., & Griffith, D. A. (2011). Incorporating cultural values for understanding the influence of perceived product quality on intention to buy: An examination in Italy and the US. *Journal of International Business Studies, 42*(4), 459–476.

Rueda-Manzanares, A., Aragón-Correa, J. A., & Sharma, S. (2008). The influence of stakeholders on the environmental strategy of service firms: The moderating effects of complexity, uncertainty and munificence. *British Journal of Management, 19*(2), 185–203.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998 June). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78–90.

Chapter 5
Multiple Regression with a Single Dependent Variable

In this chapter we examine the principles that are basic to a proper understanding of the issues involved in the analysis of management data. The chapter cannot provide the depth of a specialized econometric book. It is, however, designed to provide the elements of econometric theory essential for a researcher to develop and evaluate regression models. Multiple regression is not a multivariate technique in the strictest sense because the focus of the analysis is a single dependent variable. Nevertheless, the multivariate normal distribution is involved in the distribution of the error term, which, combined with the fact that there are multiple independent or predictor variables, leads to considering simple multiple regression within the domain of multivariate data analysis techniques.

The first section of this chapter presents the basic linear model with inferences obtained through the estimation of the model parameters. The second section discusses the issue of heterogeneity of coefficients, an important aspect of data analysis, especially in the context of testing contingency theories. While many other econometric issues remain, such as autocorrelation or multicollinearity, the reader is referred to specialized books for these topics.

5.1 Statistical Inference: Least Squares and Maximum Likelihood

The linear model is first presented with its basic assumptions. Then, point estimates using the least squares criterion are derived, followed by the maximum likelihood estimation. Finally, the properties of these estimators are discussed.

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_5, © Springer Science+Business Media New York 2014

5.1.1 The Linear Statistical Model

The dependent variable y_t is modeled as a linear function of K independent variables:

$$\underset{T\times 1}{\mathbf{y}} = \underset{T\times K}{\mathbf{X}}\ \underset{K\times 1}{\boldsymbol{\beta}} + \underset{T\times 1}{\mathbf{e}} \tag{5.1}$$

where T = number of observations (for example T periods), $\mathbf{X}$ = matrix of K independent variables, $\boldsymbol{\beta}$ = vector of K weights applied to each independent variable k, $\mathbf{y}$ = vector of the dependent variable for $t = 1$ to T, and $\mathbf{e}$ = vector of residuals corresponding to a unique aspect of $\mathbf{y}$ that is not explained by $\mathbf{X}$.

It should be noted that $\mathbf{X}$ is given, fixed, observed data. $\mathbf{X}$ is, in fact, not only observable but is also measured without error (the case of measurement error is discussed in Chap. 10). We assume that $\mathbf{X}$ is correctly specified. This means that $\mathbf{X}$ contains the proper variables explaining the dependent variable with the proper functional form (i.e., some of the variables expressed in $\mathbf{X}$ may have been transformed, for example, by taking their logarithm). Finally, the first column of $\mathbf{X}$ is typically a vector where each element is 1. This means that the first element of the parameter vector $\boldsymbol{\beta}$ is a parameter that corresponds to a constant term that applies equally to each value of the dependent variable y_t from $t = 1$ to T.

5.1.1.1 Error Structure

Some assumptions are needed in order to make some statistical inferences. Not all the assumptions below are necessarily used. In fact, in Sect. 5.1.4.3, we identify which assumptions are necessary in order to be able to obtain the specific properties of the estimators. Because $\mathbf{y}$ and $\mathbf{X}$ are given data points and $\boldsymbol{\beta}$ is the parameter vector on which we want to make inferences, the assumptions can only be on the unobserved factor $\mathbf{e}$.

Assumption 1: Expected Value of Error Term

$$\mathrm{E}[\mathbf{e}] = 0 \tag{5.2}$$

Assumption 2: Covariance Matrix of Error Term

Homoscedasticity
Usually, each observation has an error term e_t independently and identically distributed with the same variance:

$$\mathrm{e_t} \sim iid \Rightarrow \mathrm{E}\left[\mathbf{e}\mathbf{e}'\right] = \sigma^2\mathbf{I}_\mathrm{T} \tag{5.3}$$

where $\mathbf{I}$ = identity matrix.

This means that the variances for each observation t are the same and that they are uncorrelated. The unknown parameters that need to be estimated are $\boldsymbol{\beta}$ and σ^2.

Heteroscedasticity
More generally

$$\mathrm{E}\left[\mathbf{e}\mathbf{e}'\right] = \sigma^2\boldsymbol{\Psi} = \boldsymbol{\Phi} \tag{5.4}$$

Note that $\boldsymbol{\Phi}$, a covariance matrix, is a symmetric matrix. Heteroscedasticity occurs, therefore, when $\boldsymbol{\Psi} \neq \mathbf{I}$. This occurs if either the diagonal elements of the matrix $\boldsymbol{\Psi}$ are not identical (each error term e_t has a different variance) or if its off-diagonal elements are different from zero.

Assumption 3: Normality of Distribution
The probability density function of the error vector can be written formally as per Eq. (5.5) for the case of homoscedasticity or Eq. (5.6) for the case of heteroscedasticity:

$$\mathbf{e} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \tag{5.5}$$

or

$$\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Phi}) \tag{5.6}$$

5.1.2 Point Estimation

Point estimates are inferences that can be made without the normality assumption of the distribution of the error term e. The problem can be defined as follows: to find a suitable function of the observed random variables y, given x, that will yield the "best" estimate of unknown parameters.

We will restrict $\boldsymbol{\beta}$ to the class that are linear functions of $\mathbf{y}$:

$$\underset{K\times 1}{\hat{\boldsymbol{\beta}}} = \underset{K\times T}{\mathbf{A}}\ \underset{T\times 1}{\mathbf{y}} \tag{5.7}$$

The elements $\{a_{kt}\}$ of the matrix $\mathbf{A}$ are scalars that weight each observation; $\mathbf{A}$ is a summarizing operator.

In order to solve the problem defined above, we need to (1) select a criterion, (2) determine the $\mathbf{A}$ matrix and consequently $\hat{\boldsymbol{\beta}}$, and (3) evaluate the sampling performance of the estimator. These three issues are discussed in the following sections.

5.1.2.1 Ordinary Least Squares Estimator

We now consider the case of homoscedasticity where

$$\boldsymbol{\Psi} = \mathbf{I}_{\mathrm{T}} \tag{5.8}$$

The criterion we use to estimate the "best" parameter is to minimize the sum of squares residuals:

$$\text{Min } l_1 = \underset{1\times T}{\mathbf{e}'}\ \underset{T\times 1}{\mathbf{e}} = (\mathbf{y} - \mathbf{X\beta})'(\mathbf{y} - \mathbf{X\beta}) \tag{5.9}$$

$$= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X\beta}, \tag{5.10}$$

noting that $\mathbf{y}'\mathbf{X\beta} = \boldsymbol{\beta}'\,\mathbf{X}'\mathbf{y}$ is a scalar.

We resolve the problem of finding the parameters that minimize this least squares quantity (l_1 in Eq. (5.9)) by taking the derivative relative to the parameter vector $\boldsymbol{\beta}$, setting it to zero, and solving that equation:

$$\frac{\partial l_1}{\partial \boldsymbol{\beta}} = 2\mathbf{X}'\mathbf{X\beta} - 2\mathbf{X}'\mathbf{y} = 0 \tag{5.11}$$

Note that the derivative in Eq. (5.11) is obtained by using the following matrix derivative rules also found in Appendix A (Chap. 14):

$$\frac{\partial \mathbf{a}'\mathbf{v}}{\partial \mathbf{v}} = \mathbf{a}, \text{ and}$$

$$\frac{\partial \mathbf{v}'\mathbf{A}\mathbf{v}}{\partial \mathbf{v}} = \left(\mathbf{A} + \mathbf{A}'\right)\mathbf{v}$$

and especially

$$\frac{\partial 2\mathbf{y}'\mathbf{X\beta}}{\partial \boldsymbol{\beta}} = 2\mathbf{X}'\mathbf{y} \tag{5.12}$$

Therefore, applying these rules to Eq. (5.10), we obtain

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{5.13}$$

This assumes that **X'X** can be inverted. If collinearity in the data exists, i.e., if a variable x_k is a linear combination of a subset of the other x variables, the inverse does not exist (the determinant is zero). In a less strict case, multicollinearity can occur if the determinant of **X'X** approaches zero. The matrix may still be invertible and an estimate of $\boldsymbol{\beta}$ will exist. We will briefly discuss the problem in the subsection Computation of Covariance Matrix of Sect. 5.1.4.2.

b is a linear function of y:

$$\mathbf{b} = \mathbf{A}\mathbf{y} \tag{5.14}$$

where

$$\mathbf{A} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \tag{5.15}$$

5.1.2.2 Generalized Least Squares or Aitken Estimator

In the general case of heteroscedasticity, the covariance matrix of the error term vector is positive definite symmetric:

$$\boldsymbol{\Psi} \neq \mathbf{I}_{\mathrm{T}} \tag{5.16}$$

The criterion is the quadratic form of the error terms weighted by the inverse of the covariance matrix. The rationale for that criterion is best understood in the case where $\boldsymbol{\Psi}$ is diagonal. In such a case, it can be easily seen that the observations with the largest variances are given smaller weights than the others.

The objective is then

$$\text{Min } l_2 = \mathbf{e}'\boldsymbol{\Psi}^{-1}\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Psi}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{5.17}$$

$$= \left(\mathbf{y}'\boldsymbol{\Psi}^{-1} - \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Psi}^{-1}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{5.18}$$

$$= \mathbf{y}'\boldsymbol{\Psi}^{-1}\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\boldsymbol{\beta} - \underset{1\times k}{\boldsymbol{\beta}'}\ \underset{k\times T}{\mathbf{X}'}\ \underset{T\times T}{\boldsymbol{\Psi}^{-1}}\ \underset{T\times 1}{\mathbf{y}} - \underset{1\times 1}{\mathbf{y}'\boldsymbol{\Psi}^{-1}\mathbf{X}\boldsymbol{\beta}} \tag{5.19}$$

$$= \mathbf{y}'\boldsymbol{\Psi}^{-1}\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}'\boldsymbol{\Psi}^{-1}\mathbf{X}\boldsymbol{\beta} \tag{5.20}$$

Minimizing of the quadratic expression in Eq. (5.20) is performed by solving the equation

$$\frac{\partial l_2}{\partial \boldsymbol{\beta}} = 2\left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)\boldsymbol{\beta} - 2\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y} = 0 \tag{5.21}$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{GLS} = \left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y} \tag{5.22}$$

Consequently, $\hat{\boldsymbol{\beta}}$ is still a linear function of y such as in Eq. (5.14), but with the linear weights given by

$$\mathbf{A} = \left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1} \tag{5.23}$$

5.1.3 Maximum Likelihood Estimation

So far, the estimators that we have derived are point estimates. They do not allow the researcher to perform statistical tests of significance on the parameter vector $\boldsymbol{\beta}$. In this section, we derive the maximum likelihood estimators, which leads to the presentation of the distributional properties of these estimators. The problem is to find the value of the parameter $\boldsymbol{\beta}$ that will maximize the probability of obtaining the observed sample.

The assumption that is needed to derive the maximum likelihood estimator is the normal distribution of the error term

$$\mathbf{e} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}_{\mathrm{T}}\right) \tag{5.24}$$

It is then possible to write the likelihood function, which for the homoscedastic case is

$$l_1\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right) = \left(2\pi\sigma^2\right)^{-\mathrm{T}/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \tag{5.25}$$

or for the case of heteroscedasticity

$$l_2\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right) = \left(2\pi\sigma^2\right)^{-\mathrm{T}/2} |\boldsymbol{\Psi}|^{-\mathrm{T}/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Psi}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \tag{5.26}$$

We can then maximize the likelihood or, equivalently, its logarithm:

$$\text{Max } l_1 \Leftrightarrow \text{Max Ln } l_1 \Leftrightarrow \text{Max}\left[-\frac{\mathrm{T}}{2}\mathrm{Ln}\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \tag{5.27}$$

which is equivalent to minimizing the negative of that expression, i.e.,

$$\text{Min}\left[\frac{\mathrm{T}}{2}\mathrm{Ln}\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \tag{5.28}$$

This can be done by solving the derivative of Eq. (5.28) relative to β:

$$\frac{\partial[-Ln(l_1)]}{\partial\boldsymbol{\beta}} = 0 \Rightarrow \tilde{\boldsymbol{\beta}}_1 = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{5.29}$$

which is simply the least squares estimator.

Similar computations lead to the maximum likelihood estimator in the case of heteroscedasticity, which is identical to the generalized least squares (GLS) estimator:

$$\tilde{\boldsymbol{\beta}}_2 = \left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y} \tag{5.30}$$

We can now compute the maximum likelihood estimator of the variance by finding the value of σ that maximizes the likelihood or minimizes the expression in Eq. (5.28):

$$\underset{\sigma}{Min}\left[\frac{\mathrm{T}}{2}\mathrm{Ln}2\pi + \mathrm{T}\mathrm{Ln}\sigma + \frac{1}{2}\sigma^{-2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \tag{5.31}$$

This is solved by setting the derivative relative to σ to zero:

$$\frac{\partial[-Ln(l_1)]}{\partial\sigma} = \frac{\mathrm{T}}{\sigma} + \frac{1}{2}\left(-2\sigma^{-3}\right)(\mathbf{y}-\mathbf{X\beta})'(\mathbf{y}-\mathbf{X\beta}) = 0 \tag{5.32}$$

This results in

$$\frac{\mathrm{T}}{\sigma} - \frac{1}{\sigma^3}(\mathbf{y}-\mathbf{X\beta})'(\mathbf{y}-\mathbf{X\beta}) = 0 \Rightarrow \frac{1}{\sigma^3}(\mathbf{y}-\mathbf{X\beta})'(\mathbf{y}-\mathbf{X\beta}) = \frac{\mathrm{T}}{\sigma} \tag{5.33}$$

which leads to the maximum likelihood estimator:

$$\widetilde{\sigma}^2 = \frac{1}{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}_1\right)'\left(\mathbf{y}-\mathbf{X}\widetilde{\boldsymbol{\beta}}_1\right) = \frac{1}{\mathrm{T}}\hat{\mathbf{e}}'\hat{\mathbf{e}} \tag{5.34}$$

where $\hat{\mathbf{e}}$ is the vector of residuals obtained when using the maximum likelihood estimator of $\boldsymbol{\beta}$ to predict $\mathbf{y}$.

The same computational approach can be applied for the heteroscedastic case.

5.1.4 Properties of Estimator

We have obtained estimators for the parameters $\boldsymbol{\beta}$ and σ. The next question is to determine how good these estimators are. Two criteria are important for evaluating these parameters. Unbiasedness refers to the fact that on average the parameters are correct, i.e., on average, we obtain the true parameter. The second criterion concerns the fact that the estimator should have the smallest possible variance.

5.1.4.1 Unbiasedness

Definition: An estimator is unbiased if its expected value is equal to the true parameter, i.e.,

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} \tag{5.35}$$

$\mathbf{b}$ and $\hat{\boldsymbol{\beta}}$, and, a fortiori, the maximum likelihood estimators $\widetilde{\boldsymbol{\beta}}_1$ and $\widetilde{\boldsymbol{\beta}}_2$, are linear functions of random vector $\mathbf{y}$**.** Consequently, they are also random vectors with the following mean:

$$\mathrm{E}[\mathbf{b}] = \mathrm{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\right] = \mathrm{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X\beta}+\mathbf{e})\right] \tag{5.36}$$

$$= \mathrm{E}\left[\underbrace{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{e}\right] \tag{5.37}$$

$$= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underbrace{\mathrm{E}[\mathbf{e}]}_{=\mathbf{0}} = \boldsymbol{\beta} \tag{5.38}$$

This proves the least squares estimator is unbiased. Similarly for the GLS estimator

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{E}\left[\left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{y}\right] = \boldsymbol{\beta} + \mathbf{E}\left[\left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{e}\right] = \boldsymbol{\beta} \tag{5.39}$$

This means that on average the GLS estimator is the true parameter and is thus unbiased.

5.1.4.2 Best Linear Estimator

How do the linear rules above compare with other linear unbiased rules in terms of the precision, i.e., in terms of the covariance matrix? We want an estimator that has the smallest variance possible. This means that we need to compute the covariance matrix of the estimator, and then we need to show that it has minimum variance.

Computation of Covariance Matrix

The covariance of the least squares estimator **b** is

$$\begin{aligned}
\underset{K\times K}{\boldsymbol{\Sigma}_{\mathbf{b}}} &= \mathrm{E}\left[(\mathbf{b} - \mathrm{E}[\mathbf{b}])(\mathbf{b} - \mathrm{E}[\mathbf{b}])'\right] \\
&= \mathrm{E}\left[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'\right] \\
&= \mathrm{E}\left[\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta}\right)\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta}\right)'\right] \\
&= \mathrm{E}\left[\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) - \boldsymbol{\beta}\right)\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) - \boldsymbol{\beta}\right)'\right] \\
&= \mathrm{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right] \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathrm{E}\left[\mathbf{e}\mathbf{e}'\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned} \tag{5.40}$$

Therefore,

$$\underset{K\times K}{\boldsymbol{\Sigma}_{\mathbf{b}}} = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{5.41}$$

In the case of multicollinearity, $(\mathbf{X}'\mathbf{X})^{-1}$ is very large (because the determinant is close to zero). This means that the variance of the estimator will be very large. Consequently, multicollinearity results in parameter estimates that are unstable.

Following similar calculations, the variance–covariance matrix of the GLS estimator $\hat{\boldsymbol{\beta}}$ is

$$\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}} = \mathrm{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right] = \mathrm{E}\left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{e}\mathbf{e}'\boldsymbol{\Psi}^{-1}\mathbf{X}\left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1} \tag{5.42}$$

$$\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}} = \sigma^2\left(\mathbf{X}'\boldsymbol{\Psi}^{-1}\mathbf{X}\right)^{-1} \tag{5.43}$$

Best Linear Unbiased Estimator

Out of the class of linear unbiased rules, the ordinary least squares (OLS) (or the GLS depending on the error term covariance structure) estimator is the best, i.e., provides minimum variance. We will provide the proof with the OLS estimator when $\boldsymbol{\Psi} = \mathbf{I}_{\mathrm{T}}$; however, the proof is similar for the GLS estimator when $\boldsymbol{\Psi} \neq \mathbf{I}_{\mathrm{T}}$.

The problem is equivalent to minimizing the variance of a linear combination of the K parameters for any linear combination.

Let $\underset{K\times 1}{\boldsymbol{\varphi}}$ be a vector of constants. The scalar θ is the linear combination of the regression parameters $\boldsymbol{\beta}$:

$$\underset{1\times 1}{\boldsymbol{\theta}} = \underset{1\times K}{\boldsymbol{\varphi}'}\ \underset{K\times 1}{\boldsymbol{\beta}}$$

The least squares estimator of θ is

$$\hat{\boldsymbol{\theta}}_{LS} = \boldsymbol{\varphi}'\mathbf{b} = \boldsymbol{\varphi}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{5.44}$$

The problem is therefore to determine if there exists another unbiased linear estimator that is better than the least squares estimator.

An alternative linear estimator would be written in a general way as

$$\underset{1\times 1}{\hat{\theta}} = \underset{1\times T}{\mathbf{A}'}\ \underset{T\times 1}{\mathbf{y}} + \underset{1\times 1}{\mathbf{a}} \tag{5.45}$$

$\hat{\theta}$ should be unbiased. This means that

$$\forall \boldsymbol{\beta}: \quad \mathrm{E}\left[\hat{\boldsymbol{\theta}}\right] = \boldsymbol{\varphi}'\boldsymbol{\beta} \tag{5.46}$$

By substitution of the expression of the estimator $\hat{\theta}$,

$$\mathrm{E}\left[\hat{\theta}\right] = \mathrm{E}\left[\mathbf{A}'\mathbf{y} + \mathbf{a}\right] = \mathbf{A}'\mathrm{E}[\mathbf{y}] + \mathbf{a} \tag{5.47}$$

$$= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{a} \tag{5.48}$$

For $\hat{\theta}$ to be unbiased, Eq. (5.46) must be verified, i.e.,

$$\boldsymbol{\varphi}'\boldsymbol{\beta} = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{a} \tag{5.49}$$

This can only be true if

$$\mathbf{a} = 0 \tag{5.50}$$

and

$$\boldsymbol{\varphi}' = \mathbf{A}'\mathbf{X} \tag{5.51}$$

What is the value of A that will minimize the variance of the estimator? The variance is

$$\mathrm{V}\left[\hat{\theta}\right] = \mathbf{A}'\mathrm{V}[\mathbf{y}]\mathbf{A} \tag{5.52}$$

However,

$$\begin{gathered} \underset{T\times 1}{\mathrm{V}\,[\mathbf{y}]} = \mathrm{V}[\mathbf{X}\boldsymbol{\beta} + \mathbf{e}] \\ = \mathrm{E}\left[((\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) - \mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}))((\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) - \mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}))'\right] \\ = \mathrm{E}\left[\mathbf{e}\mathbf{e}'\right] = \sigma^2\mathbf{I} \end{gathered} \tag{5.53}$$

Therefore,

$$\mathrm{V}\left[\hat{\theta}\right] = \sigma^2\mathbf{A}'\mathbf{A} \tag{5.54}$$

The problem now is to minimize $\mathrm{V}\left[\hat{\theta}\right]$ subject to the unbiasedness restrictions stated in Eqs. (5.50) and (5.51), i.e.,

$$\mathrm{Min}\ \sigma^2\mathbf{A}'\mathbf{A}$$

$$\text{s.t.}\boldsymbol{\varphi}' = \mathbf{A}'\mathbf{X}$$

This is a Lagrangian multiplier problem.
The Lagrangian is

$$\mathbf{L} = \sigma^2 \underset{1\times T}{\mathbf{A}'} \underset{T\times 1}{\mathbf{A}} + 2 \underset{1\times K}{\boldsymbol{\lambda}'} \left(\underset{K\times 1}{\boldsymbol{\varphi}} - \underset{K\times T}{\mathbf{X}'} \underset{T\times 1}{\mathbf{A}} \right) \tag{5.55}$$

$$\frac{\partial \mathbf{L}}{\partial \mathbf{A}} = 2\sigma^2 \mathbf{A}' - 2\boldsymbol{\lambda}' \mathbf{X}' = 0 \tag{5.56}$$

Therefore,

$$\sigma^2 \mathbf{A}' - \boldsymbol{\lambda}' \mathbf{X}' = \mathbf{0}$$

$$\sigma^2 \mathbf{A}' \mathbf{X} - \boldsymbol{\lambda}' \mathbf{X}' \mathbf{X} = \mathbf{0}$$
$$\boldsymbol{\lambda}' = \sigma^2 \mathbf{A}' \mathbf{X} \left(\mathbf{X}' \mathbf{X} \right)^{-1}$$

$$\boldsymbol{\lambda}' = \quad \sigma^2 \boldsymbol{\varphi}' \left(\mathbf{X}' \mathbf{X} \right)^{-1} \tag{5.57}$$

In addition,

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\lambda}} = \boldsymbol{\varphi}' - \mathbf{A}' \mathbf{X} = \mathbf{0} \tag{5.58}$$

Considering again the derivative relative to $\mathbf{A}$ given in Eq. (5.56), i.e.,

$$\frac{\partial \mathbf{L}}{\partial \mathbf{A}} = 2\sigma^2 \mathbf{A}' - 2\boldsymbol{\lambda}' \mathbf{X}'$$

replacing $\boldsymbol{\lambda}$ by the expression obtained in Eq. (5.57), we obtain

$$\frac{\partial \mathbf{L}}{\partial \mathbf{A}} = 2\sigma^2 \mathbf{A}' - 2\sigma^2 \boldsymbol{\varphi}' \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' = \mathbf{0} \tag{5.59}$$

and, therefore,

$$\mathbf{A}' = \boldsymbol{\varphi}' \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' \tag{5.60}$$

However,

$$\theta = \mathbf{A}' \mathbf{y}$$

Thus, the minimum variance linear unbiased estimator of $\boldsymbol{\varphi}'\boldsymbol{\beta}$ is obtained by replacing $\mathbf{A}'$ with the expression in Eq. (5.60):

$$\hat{\theta} = \boldsymbol{\varphi}' \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{y} \tag{5.61}$$

Table 5.1 Properties of estimators

Property	Assumption(s) needed
$\mathbf{E}[\mathbf{b}\|\mathbf{X}] = \beta$	No.1
$\mathbf{V}[\mathbf{b}\|\mathbf{X},s^2] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	No.1, 2
b is BLUE	No.1, 2
b is the MLE	No.3
$\mathbf{b} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$	No.3

which is the one obtained from the OLS estimator:

$$\hat{\theta} = \boldsymbol{\varphi}'\mathbf{b} \tag{5.62}$$

We have just shown that the OLS estimator has minimum variance.

5.1.4.3 Summary of Properties

Not all three assumptions discussed in Sect. 5.1.1 are needed for all the properties of the estimator. Unbiasedness only requires assumption no.1. The computation of the variance and the best linear unbiased estimator (BLUE) property of the estimator only involve assumptions no.1 and no.2, and do not require the normal distributional assumption of the error term. Statistical tests about the significance of the parameters can only be performed with assumption no.3 about the normal distribution of the error term. These properties are shown in Table 5.1.

5.1.5 R-Squared as a Measure of Fit

We first present the R-squared measure and its interpretation as a percentage of explained variance in the presence of homoscedasticity. We then discuss the issues that appear when the error term is heteroscedastic.

5.1.5.1 Normal Case of Homoscedasticity

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}} \tag{5.63}$$

Let $\overline{\mathbf{y}}$ be the $T \times 1$ vector containing T times the mean of y. Subtracting $\overline{\mathbf{y}}$ from each side of Eq. (5.63):

$$\mathbf{y} - \overline{\mathbf{y}} = \hat{\mathbf{y}} - \overline{\mathbf{y}} + \hat{\mathbf{e}} \tag{5.64}$$

Multiplying each side by its transpose:

$$\left(\mathbf{y}-\overline{\mathbf{y}}\right)'\left(\mathbf{y}-\overline{\mathbf{y}}\right)=\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}+\hat{\mathbf{e}}\right)'\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}+\hat{\mathbf{e}}\right) \tag{5.65}$$

$$=\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)'\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)+\hat{\mathbf{e}}'\hat{\mathbf{e}}+\hat{\mathbf{e}}'\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)+\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)'\hat{\mathbf{e}} \tag{5.66}$$

$$=\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)'\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)+\hat{\mathbf{e}}'\hat{\mathbf{e}}+2\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)'\hat{\mathbf{e}} \tag{5.67}$$

$$=\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)'\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)+\hat{\mathbf{e}}'\hat{\mathbf{e}}+2\left(\hat{\mathbf{y}}'\hat{\mathbf{e}}-\overline{\mathbf{y}}'\hat{\mathbf{e}}\right) \tag{5.68}$$

The last term in the equation is equal to 0 because

$$\hat{\mathbf{y}}'\hat{\mathbf{e}}=\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\hat{\mathbf{e}}=\hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\mathbf{e}}=\hat{\boldsymbol{\beta}}'\mathbf{X}'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)=\hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

$$=\hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\right)=\hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{y}\right)=0 \tag{5.69}$$

and $\overline{\mathbf{y}}'\hat{\mathbf{e}}=0$ because it is the mean of the error term, which is zero if the equation contains a constant term.

Therefore, the equality in Eq. (5.65) shows that the total sum of squares (*TSS*) is equal to the regression sum of squares (*RSS*) plus the error sum of squares (*ESS*):

$$\left(\mathbf{y}-\overline{\mathbf{y}}\right)'\left(\mathbf{y}-\overline{\mathbf{y}}\right)=\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)'\left(\hat{\mathbf{y}}-\overline{\mathbf{y}}\right)+\hat{\mathbf{e}}'\hat{\mathbf{e}} \tag{5.70}$$

$$TSS=RSS+ESS$$

Consequently, a measure of fit is the R^2:

$$R^2=1-\frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{\left(\mathbf{y}-\overline{\mathbf{y}}\right)'\left(\mathbf{y}-\overline{\mathbf{y}}\right)}=1-\frac{ESS}{TSS} \tag{5.71}$$

This measure can be interpreted as the proportion of explained variance because of Eq. (5.70). For the same reason,

$$R^2\in[0,1]$$

It should be noted that if Eq. (5.63) does not contain a constant term, the equality in Eq. (5.70) does not hold because $\overline{\mathbf{y}}'\hat{\mathbf{e}}\neq 0$. In such a case, the R^2 computed as in Eq. (5.71) cannot be interpreted as the percentage of explained variance.

5.1.5.2 Case with Non-scalar Error Covariance Matrix: $\mathbf{E}[\mathbf{ee}'] = \mathbf{\Phi} \neq \sigma^2\mathbf{I}$

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}} = \hat{\mathbf{y}} + \hat{\mathbf{e}} \tag{5.72}$$

where the appropriate estimator is the GLS estimator:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{y} \tag{5.73}$$

Considering again Eq. (5.64),

$$\mathbf{y} - \overline{\mathbf{y}} = \hat{\mathbf{y}} - \overline{\mathbf{y}} + \hat{\mathbf{e}} \tag{5.74}$$

Multiplying each side by its transpose:

$$\left(\mathbf{y} - \overline{\mathbf{y}}\right)'(\mathbf{y} - \overline{\mathbf{y}}) = \left(\hat{\mathbf{y}} - \overline{\mathbf{y}} + \hat{\mathbf{e}}\right)'\left(\hat{\mathbf{y}} - \overline{\mathbf{y}} + \hat{\mathbf{e}}\right) \tag{5.75}$$

$$= \left(\hat{\mathbf{y}} - \overline{\mathbf{y}}\right)'(\hat{\mathbf{y}} - \overline{\mathbf{y}}) + \hat{\mathbf{e}}'\hat{\mathbf{e}} + \hat{\mathbf{e}}'(\hat{\mathbf{y}} - \overline{\mathbf{y}}) + \left(\hat{\mathbf{y}} - \overline{\mathbf{y}}\right)'\hat{\mathbf{e}} \tag{5.76}$$

$$= \left(\hat{\mathbf{y}} - \overline{\mathbf{y}}\right)'(\hat{\mathbf{y}} - \overline{\mathbf{y}}) + \hat{\mathbf{e}}'\hat{\mathbf{e}} + 2\left(\hat{\mathbf{y}} - \overline{\mathbf{y}}\right)'\hat{\mathbf{e}} \tag{5.77}$$

$$= \left(\hat{\mathbf{y}} - \overline{\mathbf{y}}\right)'(\hat{\mathbf{y}} - \overline{\mathbf{y}}) + \hat{\mathbf{e}}'\hat{\mathbf{e}} + 2\left(\hat{\mathbf{y}}'\hat{\mathbf{e}} - \overline{\mathbf{y}}'\hat{\mathbf{e}}\right) \tag{5.78}$$

The problem this time is that the last term in the equation is not equal to 0, because

$$\hat{\mathbf{y}}'\hat{\mathbf{e}} = \left(\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\hat{\mathbf{e}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\mathbf{e}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}\right)$$
$$= \hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{y}\right) \neq 0 \tag{5.79}$$

Therefore,

$$R^2 = 1 - \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{(\mathbf{y} - \overline{\mathbf{y}})'(\mathbf{y} - \overline{\mathbf{y}})} \tag{5.80}$$

cannot be interpreted any longer as the proportion of explained variance because the equality in Eq. (5.70) is no longer true. For the same reason, $R^2 \notin [0,1]$. In fact, $R^2 \in [-\infty, 1]$.

5.2 Pooling Issues

The pooling issues refer to the ability to pool together subsets of data. Therefore, this concerns the extent to which data sets are homogeneous or are generated by the same data-generating function. This question can be addressed by testing whether or not the parameters of different subsets of data are the same. If the parameters are different, the objective may become, in a second stage, to develop models that contain variables explaining why these parameters differ. This would lead to varying parameter models that are outside the scope of this book, but which students may wish to explore in specialized manuals.

5.2.1 Linear Restrictions

Let us write a linear model for two sets of data with T_1 and T_2 observations, respectively:

$$\text{Data set1}: \underset{T_1\times 1}{\mathbf{y}_1} = \underset{T_1\times K}{\mathbf{X}_1}\boldsymbol{\beta}_1 + \mathbf{u}_1 \tag{5.81}$$

$$\text{Data set2}: \underset{T_2\times 1}{\mathbf{y}_2} = \underset{T_2\times K}{\mathbf{X}_2}\boldsymbol{\beta}_2 + \mathbf{u}_2 \tag{5.82}$$

where the **y**s and the **X**s represent the same variables in each subset of data. The subscripts in Eqs. (5.81) and (5.82) represent the two subsets of observations. For example, the dependent variable may be sales of a product and **X** may contain a vector of 1s for an intercept and the price of the product. The subscript can represent the country (in this case, countries 1 and 2). There would be T_1 time periods of observations in country 1 and T_2 periods in country 2.

Assembling the two data sets gives

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} + \begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} \tag{5.83}$$

or

$$\underset{T\times 1}{\mathbf{y}} = \underset{T\times 2K}{\widetilde{\mathbf{X}}}\ \underset{2K\times 1}{\boldsymbol{\beta}} + \underset{T\times 1}{\mathbf{u}} \tag{5.84}$$

where $T = T_1 + T_2$.

$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ can also be written as $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 = 0$ or

$$[\mathbf{1} \quad -\mathbf{1}]\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} = 0 \tag{5.85}$$

which can also be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{0} \tag{5.86}$$

where $\mathbf{R} = [\mathbf{1} \quad -\mathbf{1}]$.

This can be generalized to more than two subsets of data. Then the estimation can be done as for any linear restriction on the parameters as described in Sect. 5.2.1.1.

This linear restriction can also be represented by the model

$$\begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta} + \mathbf{u} \tag{5.87}$$

or

$$\underset{T\times 1}{\mathbf{y}} = \underset{T\times K}{\mathbf{X}}\ \underset{K\times 1}{\boldsymbol{\beta}} + \underset{T\times 1}{\mathbf{u}} \tag{5.88}$$

Let RRSS be the restricted residual sum of squares coming from Eq. (5.87) and URSS be the unrestricted residual sum of squares coming from Eq. (5.83) or obtained by summing up the residual sum of squares of each equation estimated separately. Each one follows a chi-square distribution:

$$\begin{aligned}\text{RRSS} &\sim \chi^2_{\nu=T_1+T_2-K}\\ \text{URSS} &\sim \chi^2_{\nu=T_1+T_2-2K}\end{aligned}$$

The test involves checking if the fit is made significantly worse by imposing the constraint on the parameters. Therefore, a test of the restriction that the coefficients from the two data sets are equal is given by the following F test, which compares the residual sum of squares after corrections for differences in degrees of freedom:

$$\frac{(\text{RRSS} - \text{URSS})/K}{\text{URSS}/(T_1 + T_2 - 2K)} \sim F^{\nu_1=K}_{\nu_2=T_1+T_2-2K} \tag{5.89}$$

This test requires that the number of observations in each set be greater than the number of parameters in order to have sufficient degrees of freedom. Otherwise, the unrestricted model cannot be estimated. If $T_2 < K$, it is still possible to test that the T_2 observations are generated by the same model as the one used for the T_1 observations.

The model is first estimated using only the T_1 observations from the first set of data, as in Eq. (5.81). The residual sum of squares for these T_1 observations is RSS_1. Then, the pooled model is estimated as in Eq. (5.87) to obtain the residual sum of squares RRSS.

The two residual sums of squares RSS_1 and RRSS have independent chi-square distributions, each with, respectively, $T_1 - K$ and $T_1 + T_2 - K$ degrees of freedom. The test of homogeneity of coefficients is therefore obtained from the significance of the difference between the two residual sums of squares:

$$\frac{(\text{RRSS} - \text{RSS}_1)/(T_1 + T_2 - K - (T_1 - K))}{\text{RSS}_1/(T_1 - K)}$$

Therefore, the test considers the F distribution:

$$\frac{(\text{RRSS} - \text{RSS}_1)/T_2}{\text{RSS}_1/(T_1 - K)} = F^{\nu_1=T_2}_{\nu_2=T_1-K} \tag{5.90}$$

5.2.1.1 Constrained Estimation

Any linear constraint on the parameters can be written as

$$\mathbf{R}\boldsymbol{\beta} - \mathbf{r} = 0 \tag{5.91}$$

Minimizing the sum of squares under the linear constraint consists in minimizing the Lagrangian:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - 2\lambda'(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}) \tag{5.92}$$

This leads to

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{R}'\lambda = 0 \tag{5.93}$$

Pre-multiplying by $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$:

$$\mathbf{R}\underbrace{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}}_{\mathbf{b}} - \underbrace{\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}_{\mathbf{r}} = \mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\lambda \tag{5.94}$$

with $\mathbf{R}\boldsymbol{\beta} - \mathbf{r} = 0$.

Therefore,

$$\lambda = \left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1}[\mathbf{R}\mathbf{b} - \mathbf{r}] \tag{5.95}$$

Replacing the value of λ into Eq. (5.93):

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{R}'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) = 0 \tag{5.96}$$

This develops into

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} - \mathbf{R}'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \tag{5.97}$$

and

$$\hat{\boldsymbol{\beta}}^{R} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \tag{5.98}$$

or

$$\hat{\boldsymbol{\beta}}^{R} = \mathbf{b} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \tag{5.99}$$

5.2.2 Pooling Tests and Dummy Variable Models

In this section we assume that there are multiple firms, individuals, or territories. There are T observations for each of these N firms, individuals, or territories. We can write the equation for a single observation y_{it}. The subscripts i and t indicate that the observations vary along two dimensions, for example individuals (i) and time (t). For example, if y_{it} represents sales in a district in a given month, then it can be expressed as a linear function of factors measured in this same district at the same time period:

$$y_{it} = \beta_{1i} + \sum_{k=2}^{K}\beta_k x_{kit} + e_{it} \tag{5.100}$$

β_{1i} represents the intercept for observation i. This can be expressed in terms of an individual difference from a mean value of the intercept across all observations:

$$\beta_{1i} = \overline{\beta} + \mu_i \tag{5.101}$$

which, when inserted into Eq. (5.100), gives

$$y_{it} = \overline{\beta}_1 + \mu_i + \sum_{k=2}^{K}\beta_k x_{kit} + e_{it} \tag{5.102}$$

Depending on the nature of the variable μ, the model is a dummy variable model or an error component model.

If μ_i is fixed, then it is a dummy variable or a covariance model. If μ_i is random, we would be dealing with an error component model. In this section, we consider the dummy variable model (i.e., μ_i is fixed).

Let us consider a model with constant slope coefficients and an intercept that varies over individuals. The dummy variable model can be represented for all the T observations in a given territory i as

$$\underset{T\times 1}{\mathbf{y}_i} = \left(\overline{\beta}_1 + \mu_i\right) \underset{T\times 1}{\mathbf{j}_{\mathrm{T}}} + \underset{T\times(K-1)}{\mathbf{X}_{si}} \underset{(K-1)\times 1}{\boldsymbol{\beta}_s} + \mathbf{e}_i \tag{5.103}$$

where

$$\begin{aligned} &\mathrm{E}[\mathbf{e}_i] = \mathbf{0} \\ &\mathrm{E}\left[\mathbf{e}_i \mathbf{e}_i^{'}\right] = \sigma_e^2 \mathbf{I}_T \\ &\mathrm{E}\left[\mathbf{e}_i \mathbf{e}_j^{'}\right] = \mathbf{0} \qquad \forall i \neq j \end{aligned}$$

This is identical to creating a dummy variable, where each observation d_{itm} is such that

$$\mathrm{d}_{itm} = 1 \quad \text{if} \quad i = m \text{ and 0 otherwise,}$$

where i and m represent indices for the cross sections.

Equations (5.100) or (5.102) can then be rewritten as

$$y_{it} = \sum_{m=1}^{N} \beta_{1m} d_{itm} + \sum_{k=2}^{K} \beta_k x_{kit} + e_{it} \tag{5.104}$$

We can then form a vector of dummy variables for each territory ($\mathbf{D}_1, \ldots \mathbf{D}_i, \ldots \mathbf{D}_{\mathrm{N}}$). Each of these dummy variables vector has T rows ($T \times 1$) where each row is a 1. Then the full data can be expressed as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_i \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} & \cdots & & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 & & & & \\ \vdots & & \ddots & & & \\ & & & \mathbf{D}_i & & \\ \vdots & & & & \ddots & \vdots \\ \mathbf{0} & \cdots & & & \cdots & \mathbf{D}_N \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1i} \\ \vdots \\ \beta_{1N} \end{bmatrix} + \mathbf{X}_s \boldsymbol{\beta}_s + \mathbf{e} \tag{5.105}$$

Let $\mathrm{PRSS}_{\mathrm{slopes}}$ denote the residual sum of squares obtained from the least squares estimation of Eq. (5.105). This indicates that the model is partially restricted (PR) on the slopes, which are assumed to be equal.

The model with equal intercepts and different slopes is

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_i \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_i \\ \vdots \\ \mathbf{D}_N \end{bmatrix} \beta_1 + \begin{bmatrix} \mathbf{X}_{s1} & \mathbf{0} & \cdots & & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{s2} & & & & \vdots \\ \vdots & & \ddots & & & \\ & & & \mathbf{X}_{si} & & \\ \vdots & & & & \ddots & \vdots \\ \mathbf{0} & \cdots & & & \cdots & \mathbf{X}_{sN} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_s^1 \\ \boldsymbol{\beta}_s^2 \\ \\ \boldsymbol{\beta}_s^i \\ \\ \boldsymbol{\beta}_s^N \end{bmatrix} + \mathbf{e} \tag{5.106}$$

Let $\text{PRSS}_{\text{intercept}}$ denote the residual sum of squares obtained from the least square estimation of Eq. (5.106). This indicates a partial restriction on the intercepts that are assumed to be equal.

With the complete restriction that the intercepts and the slopes are equal, the model is given by

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_i \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \boldsymbol{\beta} + \mathbf{e} \tag{5.107}$$

This results in the residual sum of squares CRSS.

Finally, the completely unrestricted model is one where slopes and intercepts are different. This model is estimated by running N separate regressions, one for each individual or territory. The completely unrestricted residual sum of squares is CUSS.

We now develop an example of these models with two groups.

Let $\text{d}_{1i} = 1$ if observation i belongs to group 1 and 0 otherwise, and $\text{d}_{2i} = 1$ if observation i belongs to group 2 and 0 otherwise. The model can be written as

$$y_{it} = d_{1i}\beta_{01} + d_{2i}\beta_{02} + x_{it}d_{1i}\beta_{11} + x_{it}d_{2i}\beta_{12} + u_{it} \tag{5.108}$$

The first two terms correspond to the dummy intercepts and the last two terms correspond to the dummy slopes (the interaction between the variable x and the group dummy variables).

Homogeneity (i.e., equality) of intercepts and/or slopes can be tested using F tests based on the comparison of restricted and unrestricted residual sum of squares. The next section discusses the strategies for such pooling tests. Note that in all cases, the homogeneity along the second dimension is assumed. For example, homogeneity across time periods is assumed and pooling tests are performed across sections (i.e., firms, territories, or individuals, for example).

5.2.3 Strategy for Pooling Tests

The two possible strategies are based on decomposing the tests of equality of the intercepts and of the slopes across sections. The two strategies differ according to the sequencing of the tests. The process follows the one depicted in Fig. 5.1.

The first test consists of an overall test of homogeneity of intercept and slopes. For that purpose, the residual sum of squares from the completely unrestricted model (CUSS) is compared to the partially restricted model where intercept and slopes are restricted to be the same (CRSS). A failure to reject this test indicates that the intercept and slopes are the same across all sections. No more tests are needed. In

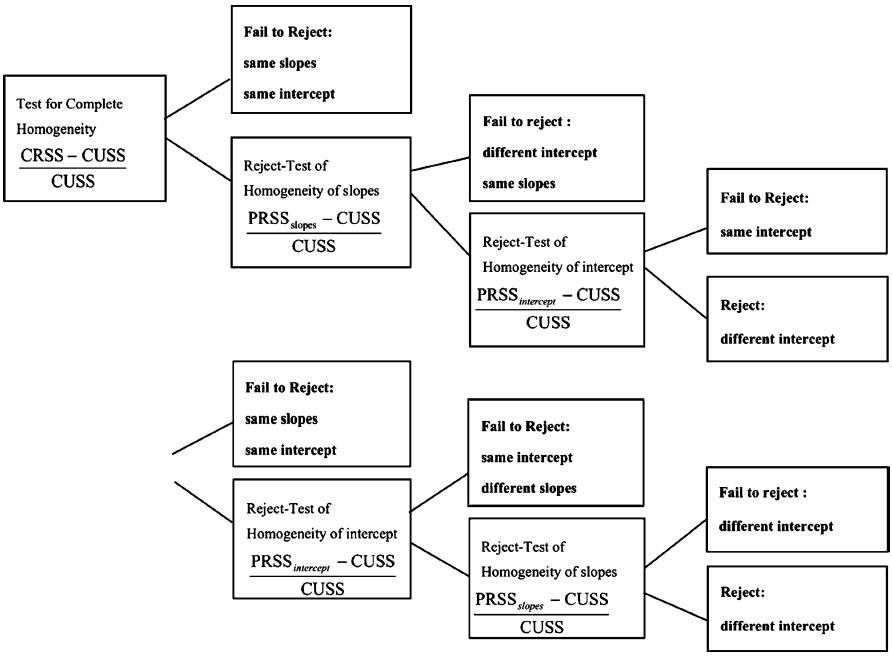


Fig. 5.1 Strategy for pooling tests

the case of rejection of the equality of intercepts and slopes, we now must perform a second test of whether the difference comes from the intercept only, the slope only, or both. A third test is then performed to check for the equality of the slopes.

For that purpose, we now compare the residual sum of squares from the completely unrestricted model (CUSS) with the residual sum of squares obtained from constraining the slopes to be equal ($PRSS_{slopes}$). A failure to reject the difference between these two models indicates that the slopes are equal. Because the slopes are equal but the full restriction leads to significant differences, we must conclude that the intercept is different across sections. If we reject the hypothesis of equal slopes, the slopes are different, in which case we must still determine if the intercept of the cross sections is the same or not.

Therefore, a third test is performed where we now compare the completely unrestricted residual sum of squares (CUSS) with the residual sum of squares of the model, with the restriction that the intercept is the same across sections ($PRSS_{intercept}$). A failure to reject the hypothesis indicates that the slopes are the only source of heterogeneity (the intercept is the same across sections). A rejection of the test indicates that both intercept and slopes are different across sections.

In this case, we began by checking the source of heterogeneity by restricting the slopes and checking if the slopes where statistically different or not across sections. Instead, we could have first restricted the intercept, i.e., tested for the homogeneity of the intercept. If the hypothesis had been rejected, we would then have tested for the homogeneity of slopes. This is the second line of tests shown in Fig. 5.1.

5.3 Examples of Linear Model Estimation with SAS and STATA

Let us consider an example where the data set consists of the market share of four brands during seven periods. This market share is predicted by two variables, the percentage of distribution outlets carrying the brand during each period and the price charged for each brand during the period.

Figure 5.2 shows an example of an SAS file to run a regression with such data. The data are first read: period (period), brand number (brandno), market share (ms), distribution (dist), and price (price). The variables are then transformed to obtain their logarithms so that the coefficients correspond to sensitivity parameters. Dummy variables for each brand except the first one (that serves as the base) are created. These will be used for estimating a model with a different intercept for each brand. They are also used to compute new variables created for distribution and price for each brand.

Three models are estimated as per the SAS file shown in Fig. 5.2. The SAS procedure REG is first called. Then a model statement indicates the model specification with the dependent variable on the left side of the equal sign and the list of independent variables on the right side. The first model statement is the completely unrestricted model where each brand has a different intercept and slopes. A second model statement is used for the completely restricted model (same intercept and slopes for all the brands). Finally, the third model statement corresponds to the partially restricted model where each brand has a different intercept but the same distribution and price parameters.

The corresponding input for STATA is given in Fig. 5.3.

The SAS output is shown in Fig. 5.4.

Similarly, the output in STATA is listed in Fig. 5.5.

```
OPTIONS LS=80;
DATA DATA1;
INFILE "c:\SAMD\Chapter5\Examples\Examp5.csv"  dlm = ',';
INPUT period brandno ms dist price;
if ms gt 0 then do;
  lms=log(ms);
  ldist=log(dist);
  lprice=log(price);
end;
else lms=.;
if brandno=2 then brand2=1; else brand2=0;
if brandno=3 then brand3=1; else brand3=0;
if brandno=4 then brand4=1; else brand4=0;
ldist2=ldist*brand2;
ldist3=ldist*brand3;
ldist4=ldist*brand4;
lprice2=lprice*brand2;
lprice3=lprice*brand3;
lprice4=lprice*brand4;

proc reg;
  model lms=brand2 brand3 brand4 ldist ldist2 ldist3 ldist4
            lprice lprice2 lprice3 lprice4;
  model lms=ldist lprice;
  model lms=brand2 brand3 brand4 ldist lprice;
run;
```

Fig. 5.2 Example of SAS input file for regression analysis (examp5.sas)

```
insheet period brandno ms dist price using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter5_MRA/Examp5.csv", clear
* Regression Example
sort brandno period
* generate new rariables in logarithms
generate lms = log(ms) if (ms>0 & !missing(ms))
generate ldist = log(dist) if (dist>0 & !missing(dist))
generate lprice = log(price) if (price>0 & !missing(price))
* generate new dummy variables for brands
generate brand2=0
replace brand2=1 if brandno ==2
generate brand3=0
replace brand3=1 if brandno ==3
generate brand4=0
replace brand4=1 if brandno ==4
* generate interaction variables
generate ldist2=ldist*brand2
generate ldist3=ldist*brand3
generate ldist4=ldist*brand4
generate lprice2=lprice*brand2
generate lprice3=lprice*brand3
generate lprice4=lprice*brand4
* regress
regress lms ldist lprice
regress lms brand2 brand3 brand4 ldist lprice
regress lms brand2 brand3 brand4 ldist lprice ldist2 lprice2 ldist3 lprice3 ldist4
lprice4
```

Fig. 5.3 Example of STATA input file for regression analysis (examp5_Mac.do)

From the output shown in Fig. 5.5, the residual sum of squares for the completely unrestricted model appears in the first model (i.e., CUSS = 0.14833). The degrees of freedom for this model are the number of observations (28, which follows from four brands having seven periods of data each) minus the number of parameters (12), that is, 16 degrees of freedom. The second model shows the completely restricted case where all intercepts are the same and the slopes are the same as well. There are three parameters estimated and the CRSS is 46.3733. The third model has a different intercept for each brand but the same slopes. Therefore, six parameters are estimated and the PRSS_{slopes} is 0.19812.

Tests of poolability can then be performed following the discussion in Sect. 5.2. The test for complete homogeneity is given by the statistic

$$\frac{(\text{CRSS} - \text{CUSS})/9}{\text{CUSS}/16} = \frac{(46.37733 - 0.14833)/9}{0.14833/16} = 554.07$$

Checking in the table for the F distribution with 9 and 16 degrees of freedom (Appendix B, Chap. 14), the difference is clearly significant and the hypothesis of complete homogeneity is clearly rejected.

We then proceed with testing for the homogeneity of slopes. We therefore compare the completely unrestricted model with the model where the slopes are restricted to be equal, which corresponds to the specification of the third model. There are six parameters and the residual sum of squares is 0.19812. The test is, therefore,

$$\frac{\left(\text{PRSS}_{slopes} - \text{CUSS}\right)/6}{\text{CUSS}/16} = \frac{(0.19812 - 0.14833)/6}{0.14833/16} = 0.895$$

Comparing this statistic with the critical value of F with 6 and 16 degrees of freedom, it is clear that the constraint does not imply a significantly worse fit. Consequently, we

```
Model: MODEL1
Dependent Variable: LMS

                          Analysis of Variance
                             Sum of          Mean
       Source          DF    Squares        Square      F Value       Prob>F
       Model           11   47.19807       4.29073      462.832       0.0001
       Error           16    0.14833       0.00927
       C Total         27   47.34640

           Root MSE       0.09628     R-square       0.9969
           Dep Mean      -2.17730     Adj R-sq       0.9947
           C.V.          -4.42217

                          Parameter Estimates
                       Parameter      Standard    T for H0:
     Variable  DF      Estimate         Error   Parameter=0    Prob > |T|
     INTERCEP   1     -1.676908    0.03642376       -46.039        0.0001
     BRAND2     1     -2.231837    0.05161904       -43.237        0.0001
     BRAND3     1     -1.014442    0.05151212       -19.693        0.0001
     BRAND4     1      1.264971    0.05150013        24.562        0.0001
     LDIST      1      0.955385    0.51824563         1.843        0.0839
     LDIST2     1      0.106274    0.55309599         0.192        0.8500
     LDIST3     1     -0.034930    0.75256037        -0.046        0.9636
     LDIST4     1      0.704706    1.64183978         0.429        0.6735
     LPRICE     1      0.248777    0.80524111         0.309        0.7613
     LPRICE2    1     -1.855944    0.92552212        -2.005        0.0622
     LPRICE3    1     -0.905538    1.19626264        -0.757        0.4601
     LPRICE4    1     -1.104439    1.12309972        -0.983        0.3401
```

Fig. 5.4 SAS output for regression analysis (examp5.lst)

```
Model: MODEL2
Dependent Variable: LMS

                          Analysis of Variance
                             Sum of          Mean
       Source          DF    Squares        Square      F Value       Prob>F
       Model            2    0.96907       0.48454        0.261       0.7722
       Error           25   46.37733       1.85509
       C Total         27   47.34640

           Root MSE       1.36202     R-square       0.0205
           Dep Mean      -2.17730     Adj R-sq      -0.0579
           C.V.         -62.55535

                          Parameter Estimates
                       Parameter      Standard    T for H0:
     Variable  DF      Estimate         Error   Parameter=0    Prob > |T|
     INTERCEP   1     -2.168119    0.25785160        -8.408        0.0001
     LDIST      1      1.724982    2.38741557         0.723        0.4767
     LPRICE     1     -1.191476    4.35217499        -0.274        0.7865
```

Fig. 5.4 (continued)

```
Model: MODEL3
Dependent Variable: LMS

                          Analysis of Variance
                                Sum of          Mean
      Source            DF      Squares        Square      F Value       Prob>F
      Model              5     47.14828       9.42966     1047.081       0.0001
      Error             22      0.19812       0.00901
      C Total           27     47.34640

          Root MSE       0.09490     R-square       0.9958
          Dep Mean      -2.17730     Adj R-sq       0.9949
          C.V.          -4.35852

                          Parameter Estimates
                       Parameter      Standard    T for H0:
     Variable  DF       Estimate         Error   Parameter=0    Prob > |T|
     INTERCEP   1     -1.678270    0.03587286       -46.784        0.0001
     BRAND2     1     -2.228246    0.05078652       -43.875        0.0001
     BRAND3     1     -1.012968    0.05072523       -19.970        0.0001
     BRAND4     1      1.265914    0.05072809        24.955        0.0001
     LDIST      1      1.057472    0.16668030         6.344        0.0001
     LPRICE     1     -0.939927    0.30325788        -3.099        0.0052
```

Fig. 5.4 (continued)

can conclude that the parameters of the distribution and price variables are homogeneous across the brands. However, each brand has a separate intercept.

Tests of equality of coefficients (or any linear combination of coefficients) can be easily performed in STATA using the "lincom" command. This is illustrated in Fig. 5.6.

The expression highlighted in grey in Fig. 5.6 represents the difference between the coefficient of the variable "brand2" (indicated by _b[brand2]) and the

```
. insheet period brandno ms dist price using
"C:\DATA\WORK_STATA\SAMD\Chapter5_MRA\Examp5.csv", clear
(5 vars, 91 obs)

. * Regression Example
. sort brandno period

. * generate new rariables in logarithms
. generate lms = log(ms) if (ms>0 & !missing(ms))
(63 missing values generated)

. generate ldist = log(dist) if (dist>0 & !missing(dist))
(56 missing values generated)

. generate lprice = log(price) if (price>0 & !missing(price))
(56 missing values generated)

. * generate new dummy variables for brands
. generate brand2=0

. replace brand2=1 if brandno ==2
(14 real changes made)

. generate brand3=0

. replace brand3=1 if brandno ==3
(14 real changes made)

. generate brand4=0

. replace brand4=1 if brandno ==4
(14 real changes made)

. * generate interaction variables
. generate ldist2=ldist*brand2
(56 missing values generated)

. generate ldist3=ldist*brand3
(56 missing values generated)

. generate ldist4=ldist*brand4
(56 missing values generated)

. generate lprice2=lprice*brand2
(56 missing values generated)

. generate lprice3=lprice*brand3
(56 missing values generated)

. generate lprice4=lprice*brand4
(56 missing values generated)

. * regress
. regress lms ldist lprice

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  2,    25) =    0.26
       Model |  .969070847     2  .484535423           Prob > F      =  0.7722
    Residual |  46.3773339    25  1.85509335           R-squared     =  0.0205
-------------+------------------------------           Adj R-squared = -0.0579
       Total |  47.3464047    27  1.75357054           Root MSE      =   1.362

------------------------------------------------------------------------------
         lms |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldist |    1.724982   2.387416     0.72   0.477    -3.191993    6.641957
      lprice |   -1.191472   4.352175    -0.27   0.787    -10.15494       7.772
       _cons |   -2.168119   .2578516    -8.41   0.000    -2.699175   -1.637064
```

Fig. 5.5 STATA output for regression analysis (examp5.log)

```
------------------------------------------------------------------------------

. regress lms brand2 brand3 brand4 ldist lprice

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  5,    22) = 1047.08
       Model |    47.14828     5    9.429656           Prob > F      =  0.0000
    Residual |  .198124694    22  .009005668           R-squared     =  0.9958
-------------+------------------------------           Adj R-squared =  0.9949
       Total |  47.3464047    27  1.75357054           Root MSE      =   .0949

------------------------------------------------------------------------------
         lms |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      brand2 |  -2.228246   .0507865   -43.87   0.000    -2.333571   -2.122922
      brand3 |  -1.012968   .0507252   -19.97   0.000    -1.118166   -.9077703
      brand4 |   1.265914   .0507281    24.95   0.000     1.16071    1.371117
       ldist |   1.057471   .1666804     6.34   0.000     .7117975    1.403145
      lprice |  -.9399269    .303258    -3.10   0.005    -1.568845   -.3110083
       _cons |   -1.67827   .0358729   -46.78   0.000    -1.752665   -1.603874
------------------------------------------------------------------------------

. regress lms brand2 brand3 brand4 ldist lprice ldist2 lprice2 ldist3 lprice3 ldist4
lprice4

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F( 11,    16) =  462.83
       Model |   47.198075    11  4.29073409           Prob > F      =  0.0000
    Residual |  .148329751    16  .009270609           R-squared     =  0.9969
-------------+------------------------------           Adj R-squared =  0.9947
       Total |  47.3464047    27  1.75357054           Root MSE      =  .09628

------------------------------------------------------------------------------
         lms |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      brand2 |  -2.231837   .0516191   -43.24   0.000    -2.341265    -2.12241
      brand3 |  -1.014442   .0515121   -19.69   0.000    -1.123643   -.9052411
      brand4 |   1.264971   .0515001    24.56   0.000     1.155796    1.374147
       ldist |   .9553852   .5182457     1.84   0.084    -.1432467    2.054017
      lprice |   .2487778    .805241     0.31   0.761    -1.458257    1.955812
      ldist2 |    .106274   .5530961     0.19   0.850    -1.066237    1.278785
     lprice2 |  -1.855945   .9255221    -2.01   0.062    -3.817964    .1060746
      ldist3 |  -.0349298   .7525606    -0.05   0.964    -1.630287    1.560427
     lprice3 |  -.9055384   1.196263    -0.76   0.460    -3.441502    1.630425
      ldist4 |   .7047077   1.641841     0.43   0.673    -2.775839    4.185254
     lprice4 |  -1.104441     1.1231    -0.98   0.340    -3.485306    1.276425
       _cons |  -1.676908   .0364238   -46.04   0.000    -1.754123   -1.599693
------------------------------------------------------------------------------
```

Fig. 5.5 (continued)

```
…
regress lms brand2 brand3 brand4 ldist lprice ldist2 lprice2 ldist3 lprice3 ldist4
lprice4
*test of equality of two coefficients
lincom _b[brand2] -_b[brand3]
```

Fig. 5.6 STATA input for test of equality of coefficients (examp6.do)

coefficient of the variable "brand3" (indicated by _b[brand3]). The results are shown in Fig. 5.7.

In this example, the brand-specific constants are different from each other for brand2 and brand3 because the difference between the two corresponding parameters is highly significant (coef. $= -1.217395$, $t = -23.58$). In the expression for the "lincom" command, we could have used any linear combination of the estimated parameters.

Note that, in the calculation of the standard error of such a linear expression, the covariances of the parameters involved must be used to compute the variance of the linear combination of normally distributed random variables. The standard output

```
( 1)  brand2 - brand3 = 0

------------------------------------------------------------------------------
         lms |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -1.217395   .0516202   -23.58   0.000    -1.326825   -1.107966
------------------------------------------------------------------------------
```

Fig. 5.7 STATA output of test of equality of coefficients (examp6.log)

```
…
regress lms ldist lprice
vce
```

Fig. 5.8 STATA input for requesting the covariance matrix of coefficients (examp7.do)

```
. regress lms ldist lprice

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  2,    25) =    0.26
       Model |  .969070847     2  .484535423           Prob > F      =  0.7722
    Residual |  46.3773339    25  1.85509335           R-squared     =  0.0205
-------------+------------------------------           Adj R-squared = -0.0579
       Total |  47.3464047    27  1.75357054           Root MSE      =   1.362

------------------------------------------------------------------------------
         lms |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldist |   1.724982   2.387416     0.72   0.477    -3.191993    6.641957
      lprice |  -1.191472   4.352175    -0.27   0.787    -10.15494      7.772
       _cons |  -2.168119   .2578516    -8.41   0.000    -2.699175   -1.637064
------------------------------------------------------------------------------

. vce

Covariance matrix of coefficients of regress model

        e(V) |      ldist      lprice       _cons
-------------+---------------------------------
       ldist |  5.6997535
      lprice | -3.6916991   18.941428
       _cons |  .03083197   .01341266   .06648745
```

Fig. 5.9 STATA output of covariances of coefficients (examp7.log)

in SAS or STATA does not list the covariances. We can, however, request them by the command "vce" in STATA, as shown in Fig. 5.8.

The results are shown in Fig. 5.9, where the covariance matrix appears following the regression results.

5.4 Assignment

Two data sets are available that contain information about a market in which multiple brands compete in an industry composed of five market segments. The full description of the data is given in Appendix C (Chap. 14).

```
                /* ----------------------------------*/
                /*    Example of                     */
                /*        (1)  Merging files for     */
                /*        (2)  regression analysis   */
                /*----------------------------------*/
option ls=80 ;
data panel;
  infile 'C:\SAMD\Chapter5\Assignments\panel.csv' firstobs=2   dlm = ',' ;

  input period  segment  segsize  ideal1-ideal3
        brand $  aware  intent shop1-shop3
        perc1-perc3 dev1-dev3 share ;

run;
proc sort data=panel;
        by period brand;
run;
/* proc print;
        title 'panel sorted';
run;
*/
data indup;
  infile 'C:\SAMD\Chapter4\Assignments\indup.csv' firstobs=2   dlm = ',' ;
  input period  firm  brand $ price  advert
        char1-char5 salmen1-salmen3
        cost dist1-dist3 usales  DSls1000  dsales  ushare dshare  adshare
        relprice ;
run;
proc sort data =indup;
        by period brand;
run;
/* proc print;
        title 'indup sorted';
run;
*/
data econ;
        merge panel indup;
        by period brand;

/* proc print;
        title 'merged data';
run;
*/
if segment<5 then delete;
run;
proc sort data=econ out=econ2;
        by brand period;
run;
data econ3;
set econ2;
lagaw =lag1(aware);
if period=0 then delete;
run;
/*proc print;
        var period segment brand aware lagaw;
run;*/
proc reg;
        model aware = lagaw adshare;
/*        by brand;*/
run;
```

Fig. 5.10 Example of SAS file for reading data sets INDUP.CSV and PANEL.CSV and for running regressions (assign5.sas)

The PANEL.CSV data set contains information at the segment level while the INDUP.CSV data set provides information at the industry level.

The file ASSIGN5.SAS in Fig. 5.10 is a SAS file that reads both data sets (INDUP.CSV and PANEL.CSV) and merges the two files.

The equivalent commands in STATA are shown in Fig. 5.11.

```
insheet using "/users/fblgatignon/Documents/WORK_STATA/SAMD/panel.csv", clear
merge m:m period brand using "/users/fblgatignon/Documents/WORK_STATA/SAMD/indup.dta"
keep if segment ==5
encode brand, generate(brandno)
tsset brandno period
drop if period ==0
regress awareness L.awareness adshare
```

Fig. 5.11 Example of STATA file for reading data sets INDUP.CSV and PANEL.CSV and for running regressions (assign5_Mac.do)

It should be noted that STATA registers the panel nature of the data with the commands "tsset brandno period" that define the structure of the time series. The lagged awareness does not need to be created, as it can be used with the command "L.awareness" directly in the regression model statement highlighted in grey in Fig. 5.11.

The assignment consists in developing a model using cross sections and time series data. For example, it is possible to model sales for each brand as a function of the price and the advertising for the brand, sales force size, etc.

Regardless of the model, you need to test whether the intercepts and slopes are homogenous. As another example, you may decide to model the awareness of each brand as a function of the awareness in the prior period and of the brand advertising of the current period. You may want to test if the process of awareness development is the same across brands.

Bibliography

Basic Technical Readings

Chow, G. C. (1960). Tests of equality between subsets of coefficients in two linear regression. *Econometrica, 28*, 591–605.

Fuller, W. A., & Battese, G. E. (1973). Transformation for estimation of linear models with nested error structure. *Journal of the American Statistical Association, 68*(343), 626–632.

Maddala, G. S. (1971). The use of variance component models in pooling cross section and time series data. *Econometrica, 39*(2), 341–358.

Mundlack, Y. (1978). On the pooling of time series and cross section data. *Econometrica, 46*, 69–85.

Nerlove, M. (1971). Further evidence on the estimation of dynamic economic relations from a time series of cross sections. *Econometrica, 39*(2), 359–382.

Application Readings

Bass, F. M., Cattin, P., & Wittink, D. R. (1978). Firm effects and industry effects in the analysis of market structure and profitability. *Journal of Marketing Research, 15*(1), 3–10.

Bass, F. M., & Leone, R. P. (1983). Temporal aggregation, the data interval bias, and empirical estimation of bimonthly relations from annual data. *Management Science, 29*(1), 1–11.

Bass, F. M., & Wittink, D. R. (1975). Pooling issues and methods in regression analysis with examples in marketing research. *Journal of Marketing Research, 12*(4), 414–425.

Bemmaor, A. C. (1984). Testing alternative econometric models on the existence of advertising threshold effect. *Journal of Marketing Research, 21*(3), 298–308.

Bowman, D., & Gatignon, H. (1996). Order of entry as a moderator of the effect of the marketing mix on market share. *Marketing Science, 15*(3), 222–242.

Gatignon, H. (1984). Competition as a moderator of the effect of advertising on sales. *Journal of Marketing Research, 21*(4), 387–398.

Gatignon, H., Eliashberg, J., & Robertson, T. S. (1989). Modeling multinational diffusion patterns: An efficient methodology. *Marketing Science, 8*(3), 231–247.

Gatignon, H., & Hanssens, D. M. (1987). Modeling marketing interactions with application to salesforce effectiveness. *Journal of Marketing Research, 24*(3), 247–257.

Gatignon, H., Robertson, T. S., & Fein, A. J. (1997). Incumbent defense strategies against new product entry. *International Journal of Research in Marketing, 14*(2), 163–176.

Gatignon, H., Weitz, B. A., & Bansal, P. (1989). Brand introduction strategies and competitive environments. *Journal of Marketing Research, 27*(4), 390–401.

Hatten, K. J., & Schendel, D. (1977). Heterogeneity within an industry: Firm conduct in the U.S. Brewing Industry, 1952–71. *Strategic Management Journal, 26*(2), 97–113.

Jacobson, R., & Aaker, D. A. (1985, Fall). Is market share all that it's cracked up to be? *Journal of Marketing, 49*, 11–22.

Johar, G. V., Jedidi, K., & Jacoby, J. (1997, September). A varying-parameter averaging model of on-line brand evaluations. *Journal of Consumer Research, 24*, 232–247.

Lambin, J.-J. (1970). Optimal allocation of competitive marketing efforts: An empirical study. *Journal of Business, 43*(4), 468–484.

Miller, C. E., Reardon, J., & McCorkle, D. E. (1999). The effects of competition on retail structure: An examination of intratype, intertype, and intercategory competition. *Journal of Marketing, 63*(4), 107–120.

Montgomery, D. B., & Silk, A. J. (1972). Estimating dynamic effects of market communications expenditures. *Management Science, 18*(10), B485–B501.

Naert, P., & Bultez, A. (1973, August). Logically consistent market share models. *Journal of Marketing Research, 10*, 334–340.

Parson, L. J. (1974). An econometric analysis of advertising, retail availability and sales of a new brand. *Management Science, 20*(6), 938–947.

Parson, L. J. (1975). The product life cycle and time varying advertising elasticities. *Journal of Marketing Research, 12*(3), 476–480.

Robinson, W. T. (1968). Marketing mix reactions to entry. *Marketing Science, 7*(4), 368–385.

Robinson, W. T. (1988). Sources of market pioneer advantages: The case of industrial goods industries. *Journal of Marketing Research, 25*(1), 87–94.

Robinson, W. T., & Fornell, C. (1985). Sources of market pioneer advantages in consumer goods industries. *Journal of Marketing Research, 22*(3), 305–317.

Steenkamp, J.-B. E. M., ter Hofstede, F., et al. (1999, April). A cross-national investigation into the individual and national cultural antecedents of consumer innovativeness. *Journal of Marketing, 63*, 55–69.

Urban, G. L., Carter, T., Gaskin, S., & Mucha, Z. (1986, June). Market share rewards to pioneering brands: An empirical analysis and strategic implications. *Management Science, 32*, 645–659.

Chapter 6
System of Equations

In this chapter we consider the case where several dependent variables are explained by linear relationships with other variables. Independent analysis of each relationship by ordinary least squares could result in incorrect statistical inferences either because the estimation is not efficient (a simultaneous consideration of all the explained variables may lead to more efficient estimators for the parameters) or may be biased in cases where the dependent variables influence each other.

In Sect. 6.1 we present a model of seemingly unrelated regression (SUR). In Sect. 6.2, we discuss the estimation of simultaneous relationships between dependent or endogenous variables. And in Sect. 6.3, we discuss the issue of identification when systems of equations are involved.

6.1 Seemingly Unrelated Regression

The case of SUR occurs when several dependent variables are expressed as a linear function of explanatory variables, leading to multiple equations with error terms that may not be independent of each other. Therefore, each equation appears unrelated to the other. However, they are in fact linked by the error terms, which leads to a disturbance-related set of equations. After first presenting the model, we derive the proper efficient estimator for the parameters, and then discuss the particular case when the predictor variables are the same in each equation.

6.1.1 Set of Equations with Contemporaneously Correlated Disturbances

Let us consider time series of M cross sections. Each cross section i presents T observations, usually over time, although t could represent individuals for which M characteristics are modeled. Therefore, for each cross section, the vector

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_6, © Springer Science+Business Media New York 2014

of dependent variables has T observations (the vector $\mathbf{y}_i$ is dimensioned $T \times 1$). In this equation for the ith cross section, there are K_i predictor variables. A priori, the variables explaining a dependent variable y_{it} are different for each cross section or variable i. Consequently, the matrix $\mathbf{X}_i$ contains T rows and K_i columns. The linear equation for each cross section can, therefore, be represented by Eq. (6.1):

$$\forall i = 1, \ldots M: \quad \underset{T\times 1}{\mathbf{y}_i} = \underset{T\times K_i}{\mathbf{X}_i}\ \underset{K_i\times 1}{\boldsymbol{\beta}_i} + \underset{T\times 1}{\mathbf{e}_i} \tag{6.1}$$

Stacking all the cross sections together, the model for all cross sections can be expressed as

$$\underset{MT\times 1}{\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_i\\ \vdots\\ \mathbf{y}_M\end{bmatrix}} = \underset{MT\times K}{\begin{bmatrix}\mathbf{X}_1 & & & & \\ & \mathbf{X}_2 & & 0 & \\ & & \ddots & & \\ & & & \mathbf{X}_i & \\ & & & & \ddots \\ & 0 & & & \mathbf{X}_M\end{bmatrix}} \underset{K\times 1}{\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \vdots\\ \boldsymbol{\beta}_i\\ \vdots\\ \boldsymbol{\beta}_M\end{bmatrix}} + \underset{MT\times 1}{\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \vdots\\ \mathbf{e}_i\\ \vdots\\ \mathbf{e}_M\end{bmatrix}} \tag{6.2}$$

where $K = \Sigma_{i=1}^{M} K_i$

This can be written more compactly as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{e} \tag{6.3}$$

The error terms have zero mean and their variances (σ_{ii}) vary for each equation. In addition, the covariance corresponding to the same time period t for each pair of cross sections is σ_{ij}. All other covariances are zero. This can be expressed for each cross-sectional vector of disturbances as

$$\forall i: \quad \mathrm{E}[\mathbf{e}_i] = \mathbf{0} \tag{6.4}$$

and

$$\forall i, j: \quad \mathrm{E}\left[\mathbf{e}_i \mathbf{e}_j^{'}\right] = \sigma_{ij}\mathbf{I}_\mathrm{T} \tag{6.5}$$

It may be useful to write the full expression for Eq. (6.5) for two cross sections i and j:

$$\mathrm{E}\left[\begin{pmatrix} e_{i1}\\ e_{i2}\\ \vdots\\ e_{it}\\ \vdots\\ e_{iT}\end{pmatrix}\begin{pmatrix} e_{j1} & e_{j2} & \cdots & e_{jt} & \cdots & e_{jT}\end{pmatrix}\right] = \begin{bmatrix}\sigma_{ij} & & & & & \\ & \sigma_{ij} & & & & \\ & & & & 0 & \\ & 0 & & \sigma_{ij} & & \\ & & & & \ddots & \\ & & & & & \sigma_{ij}\end{bmatrix} \tag{6.6}$$

The matrix for all time periods of all cross sections is expressed as

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 & \sigma_{12} & 0 & \cdots & 0 & & & \\ 0 & \sigma_{11} & \cdots & 0 & 0 & \sigma_{12} & \cdots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & & & \\ 0 & 0 & \cdots & \sigma_{11} & 0 & 0 & \cdots & \sigma_{12} & & & \\ \sigma_{12} & 0 & \cdots & 0 & \sigma_{22} & 0 & \cdots & 0 & & & \\ 0 & \sigma_{12} & \cdots & 0 & 0 & \sigma_{22} & \cdots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & & & \\ 0 & 0 & \cdots & \sigma_{12} & 0 & 0 & \cdots & \sigma_{22} & & & \\ \vdots & \vdots & & \vdots & & & & \vdots & \ddots & & \\ \vdots & \vdots & & \vdots & & & & \vdots & & \ddots & \\ \sigma_{1M} & 0 & \cdots & 0 & & & & & & & \ddots \\ 0 & \sigma_{1M} & \cdots & 0 & & & & & & & \\ \vdots & \vdots & & \vdots & & & & & & & \\ 0 & 0 & \cdots & \sigma_{1M} & \cdots & \cdots & & & & & \end{bmatrix} \tag{6.7}$$

Let $\boldsymbol{\Sigma}$ be the contemporaneous covariance matrix, i.e., the matrix where each cell represents the covariance of the error term of two equations (cross sections) for the same t:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & & \ddots & \vdots \\ \sigma_{1M} & \cdots & \cdots & \sigma_{MM} \end{bmatrix} \tag{6.8}$$

Consequently, using the Kronecker product, we can write the covariance matrix for the full set of cross sections and time series data in Eq. (6.7):

$$\mathrm{E}\left[\mathbf{e}\mathbf{e}^{'}\right] = \boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbf{I}_T \tag{6.9}$$

6.1.2 Estimation

The structure of the covariance matrix of the error term is characteristic of heteroscedasticity. Consequently, the generalized least squares estimator will be the best linear unbiased estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\mathbf{Z}^{'}\boldsymbol{\Omega}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^{'}\boldsymbol{\Omega}^{-1}\mathbf{y} \tag{6.10}$$

However, from Eq. (6.9) and using the property of the inverse of a Kronecker product of two matrices

$$(\boldsymbol{\Sigma} \otimes \mathbf{I})^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I} \tag{6.11}$$

and, therefore,

$$\hat{\boldsymbol{\beta}}_{GLS} = \left[\mathbf{Z}'\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}\right)\mathbf{Z}\right]^{-1}\mathbf{Z}'\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}\right)\mathbf{y} \tag{6.12}$$

This estimation only requires the inversion of an $M \times M$ matrix, the matrix of contemporaneous covariances.

The generalized least squares estimator is unbiased:

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}_{GLS}\right] = \boldsymbol{\beta} \tag{6.13}$$

Its variance–covariance matrix is

$$\mathrm{V}\left[\hat{\boldsymbol{\beta}}_{GLS}\right] = \left(\mathbf{Z}'\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}\right)\mathbf{Z}\right)^{-1} \tag{6.14}$$

In practice, however, the contemporaneous covariance matrix is unknown. If the matrix can be estimated by a consistent estimator, then the estimated generalized least squares (EGLS) estimator can be computed by replacing the contemporaneous covariance matrix in Eq. (6.12) by its estimated value.

$\boldsymbol{\Sigma}$ is estimated by following the three steps below:

Step 1: Ordinary least squares (OLS) are performed on each equation separately to obtain the parameters for each equation or cross section i:

$$\mathbf{b}_i = \left(\mathbf{X}'_i\mathbf{X}_i\right)^{-1}\mathbf{X}'_i\mathbf{y}_i \tag{6.15}$$

These OLS estimators are unbiased.

Step 2: The residuals are computed:

$$\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{X}_i\mathbf{b}_i \tag{6.16}$$

Step 3: The contemporaneous covariance matrix can then be computed:

$$\hat{\boldsymbol{\Sigma}} = \{\hat{\sigma}_{ij}\} = \left\{\frac{1}{T}\hat{\mathbf{e}}'_i\hat{\mathbf{e}}_j\right\} \tag{6.17}$$

Alternatively, the cross-product residuals can be divided by $T - K_\mathrm{i}$ instead of T.

The EGLS estimator is then found as

$$\hat{\boldsymbol{\beta}}_{EGLS} = \left[\mathbf{Z}'\left(\hat{\Sigma}^{-1} \otimes \mathbf{I}\right)\mathbf{Z}\right]^{-1}\mathbf{Z}'\left(\hat{\Sigma}^{-1} \otimes \mathbf{I}\right)\mathbf{y} \tag{6.18}$$

It is then possible to compute the new residuals obtained from the EGLS estimation and recalculate an updated covariance matrix to find a new EGLS estimate. This iterative procedure (ITSUR) converges to the maximum likelihood estimator.

6.1.3 Special Cases

There are two special cases where it can be demonstrated that the generalized least squares estimator obtained from the SUR is identical to the ordinary least squares estimator obtained one equation (cross section) over time. These two cases are when:

1. The independent variables in each equation are identical (i.e., same variables and same values):

$$\forall i, j: \quad \mathbf{X}_i = \mathbf{X}_j \tag{6.19}$$

2. The contemporaneous covariance matrix is diagonal, i.e., the errors across equations or cross sections are independent:

$$\boldsymbol{\Sigma} = \text{diag}\{\sigma_{ii}\} \tag{6.20}$$

Consequently, in both of these cases, there is no need to compute the covariance matrix.

6.2 A System of Simultaneous Equations

In this section we first describe the problem caused by simultaneity in estimating the parameters of the equations. We then present two estimation methods, two-stage least squares, and three-stage least squares, that provide proper estimators for these parameters.

6.2.1 The Problem

As in Sect. 6.1 for SUR, the problem here consists in estimating several equations, each corresponding to a variable to be explained by explanatory variables. The difference now, however, is that one of the variables that is explained by the model

can itself be an explanatory variable of another one, thereby creating an endogenous system. These variables are then called endogenous variables, and the variables that are not explained by the system are exogenous variables. Therefore, we need to estimate the parameters of a system of N linear equations, where there are T observations for each equation.

For one observation t:

$$\underset{N\times 1}{\mathbf{y}_t} \text{ is a vector of endogenous variables}$$

$$\underset{K\times 1}{\mathbf{x}_t} \text{ is a vector of all the exogenous variables in the system.}$$

For two equations (i.e., $N = 2$ for two endogenous variables) and two exogenous variables, we have the following system of equations:

$$\begin{cases} \gamma_{11}y_{1t} + \gamma_{12}y_{2t} = \beta_{11}x_{1t} + \beta_{12}x_{2t} + \varepsilon_{1t} \\ \gamma_{21}y_{1t} + \gamma_{22}y_{2t} = \beta_{21}x_{1t} + \beta_{22}x_{2t} + \varepsilon_{2t} \end{cases} \tag{6.21}$$

Or, in matrix notation:

$$\begin{pmatrix} y_{1t} & y_{2t} \end{pmatrix}\begin{pmatrix} \gamma_{11} & \gamma_{21} \\ \gamma_{12} & \gamma_{22} \end{pmatrix} = \begin{pmatrix} x_{1t} & x_{2t} \end{pmatrix}\begin{pmatrix} \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} & \varepsilon_{2t} \end{pmatrix} \tag{6.22}$$

Generally, the system of N equations for each t can, therefore, be expressed as

$$\underset{1\times N}{\mathbf{y}'_t}\ \underset{N\times N}{\mathbf{\Gamma}} = \underset{1\times K}{\mathbf{x}'_t}\ \underset{K\times N}{\mathbf{B}} + \underset{1\times N}{\boldsymbol{\varepsilon}'_t} \tag{6.23}$$

where the matrices $\mathbf{\Gamma}$ and B contain the parameters of all the equations.

In addition, the error terms have the following properties:

$$\forall t : \mathrm{E}\left[\underset{N\times 1}{\boldsymbol{\varepsilon}'_t}\right] = \underset{N\times 1}{\mathbf{0}} \tag{6.24}$$

and the contemporaneous covariance matrix is the symmetric matrix:

$$\forall t : \mathrm{E}\left[\underset{N\times N}{\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t}\right] = \underset{N\times N}{\mathbf{\Sigma}} \tag{6.25}$$

while the noncontemporaneous error terms are independent:

$$\forall t \neq j : \mathrm{E}\left[\underset{N\times N}{\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_j}\right] = \underset{N\times N}{\mathbf{0}} \tag{6.26}$$

The reduced form can be obtained by post-multiplying Eq. (6.23) by $\mathbf{\Gamma}^{-1}$, assuming the inverse exists:

$$\mathbf{y}'_t = \mathbf{x}'_t\mathbf{B}\mathbf{\Gamma}^{-1} + \boldsymbol{\varepsilon}'_t\mathbf{\Gamma}^{-1} \tag{6.27}$$

or

$$\underset{1\times N}{\mathbf{y}'_t} = \underset{1\times K}{\mathbf{x}'_t}\,\underset{K\times N}{\mathbf{\Pi}} + \underset{1\times N}{\mathbf{u}'_t} \tag{6.28}$$

where $\mathbf{\Pi} = \mathbf{B}\mathbf{\Gamma}^{-1}$ and $\mathbf{u}'_t = \boldsymbol{\varepsilon}'_t\mathbf{\Gamma}^{-1}$ (or $\mathbf{u}_t = (\mathbf{\Gamma}^{-1})'\boldsymbol{\varepsilon}_t$).

The elements of the matrix $\mathbf{\Pi}$ are the parameters of the reduced form of the system of equations.

The random term $\mathbf{u}_t$ is distributed with the following mean and covariance:

$$\forall t : \underset{N\times 1}{\mathrm{E}[\mathbf{u}_t]} = \mathbf{0} \tag{6.29}$$

$$\forall t : \quad \mathrm{E}\left[\mathbf{u}_t\mathbf{u}'_t\right] = \mathrm{E}\left[\left(\mathbf{\Gamma}^{-1}\right)'\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t\mathbf{\Gamma}^{-1}\right] = \left(\mathbf{\Gamma}^{-1}\right)'\mathbf{\Sigma}\mathbf{\Gamma}^{-1} \tag{6.30}$$

Equation (6.28) represents a straightforward set of equations similar to those discussed in Sect. 6.1. We can always get estimates $\hat{\mathbf{\Pi}}$. The issue is to determine whether or not we can go from $\hat{\mathbf{\Pi}}$ to $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Gamma}}$, i.e., is the knowledge about $\hat{\mathbf{\Pi}}$ sufficient to enable us to make inferences about the individual coefficients of $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Gamma}}$.

Let us write the entire model represented by Eq. (6.23) for the T observations ($t = 1, \ldots T$).

Let

$$\underset{T\times N}{\mathbf{Y}} = \begin{bmatrix} \mathbf{y}'_1 \\ \mathbf{y}'_2 \\ \vdots \\ \mathbf{y}'_t \\ \vdots \\ \mathbf{y}'_T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{21} & \cdots \\ y_{12} & y_{22} & \cdots \\ \vdots & \vdots & \\ y_{1t} & y_{2t} & \cdots \\ \vdots & \vdots & \\ y_{1T} & y_{2T} & \cdots \end{bmatrix}$$

and

$$\underset{T\times K}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_t \\ \vdots \\ \mathbf{x}'_T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & \cdots \\ x_{12} & x_{22} & \cdots \\ \vdots & \vdots & \\ x_{1t} & x_{2t} & \cdots \\ \vdots & \vdots & \\ x_{1T} & x_{2T} & \cdots \end{bmatrix}$$

Then, the system of equations is

$$\underset{T\times N}{\mathbf{Y}}\,\underset{N\times N}{\mathbf{\Gamma}} = \underset{T\times K}{\mathbf{X}}\,\underset{K\times N}{\mathbf{B}} + \underset{T\times N}{\mathbf{E}} \tag{6.31}$$

Similar to what was done above by post-multiplying by the inverse of $\mathbf{\Gamma}$:

$$\mathbf{Y} = \mathbf{X}\mathbf{B}\mathbf{\Gamma}^{-1} + \mathbf{E}\mathbf{\Gamma}^{-1} \tag{6.32}$$

or

$$\underset{T\times N}{\mathbf{Y}} = \underset{T\times K}{\mathbf{X}}\ \underset{K\times N}{\mathbf{\Pi}} + \underset{T\times N}{\mathbf{U}} \tag{6.33}$$

Because $\mathrm{E}[\mathbf{U}] = \mathbf{0}$, the ordinary least squares estimator of $\mathbf{\Pi}$ is unbiased:

$$\underset{K\times N}{\hat{\mathbf{\Pi}}} = \left(\underset{K\times T}{\mathbf{X}'}\ \underset{T\times K}{\mathbf{X}}\right)^{-1} \underset{K\times T}{\mathbf{X}'}\ \underset{T\times N}{\mathbf{Y}} \tag{6.34}$$

Therefore we can predict $\hat{\mathbf{Y}}$.

Why is this useful? Let us consider one equation ($i = 1$). Let $\mathbf{\Gamma} = [\mathbf{\Gamma}_1\, \mathbf{\Gamma}_2\ \ldots\ \mathbf{\Gamma}_N]$ and $\mathbf{B} = [\mathbf{B}_1\, \mathbf{B}_2\ \ldots\ \mathbf{B}_N]$.

Then, the first equation can be represented by

$$\underset{T\times N}{\mathbf{Y}}\ \underset{N\times 1}{\mathbf{\Gamma}_1} = \underset{T\times K}{\mathbf{X}}\ \underset{K\times 1}{\mathbf{B}_1} + \underset{T\times 1}{\mathbf{e}_1} \tag{6.35}$$

so that

$$\underset{T\times 1}{\mathbf{y}_1}\,\gamma_{11} + \mathbf{y}_2\gamma_{12} + \ldots \mathbf{y}_N\gamma_{1N} = \underset{T\times 1}{\mathbf{x}_1}\,\underset{1\times 1}{\beta_{11}} + \underset{T\times 1}{\mathbf{x}_2}\,\beta_{12} + \ldots \underset{T\times 1}{\mathbf{x}_K}\,\beta_{1K} + \mathbf{e}_1 \tag{6.36}$$

Let $\gamma_{11} = 1$

$$\mathbf{y}_1 = -\mathbf{y}_2\gamma_{12} \ldots - \mathbf{y}_N\gamma_{1N} + \mathbf{x}_1\beta_{11} + \mathbf{x}_2\beta_{12} + \ldots \mathbf{x}_K\beta_{1K} + \mathbf{e}_1 \tag{6.37}$$

or

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{e}_1 \tag{6.38}$$

Why are we unable to estimate the parameter vector $\boldsymbol{\alpha}$ using ordinary least squares?

The reason is that the estimator would be biased because $\mathbf{y}_n$ and $\mathbf{e}_1$ are correlated. This comes from the fact that $\mathbf{y}_n = \mathbf{Z}_n\,\boldsymbol{\alpha}_n + \mathbf{e}_n$ and that $\mathbf{e}_1$ and $\mathbf{e}_n$ are correlated due to $\mathbf{\Sigma}$. Indeed, for example, with two equations and one exogenous variable in each equation:

$$\begin{cases} \mathbf{y}_1 = -\mathbf{y}_2\gamma_{12} + \mathbf{x}_1\beta_{11} + \mathbf{e}_1 \\ \mathbf{y}_2 = -\mathbf{y}_1\gamma_{21} + \mathbf{x}_2\beta_{22} + \mathbf{e}_2 \end{cases} \tag{6.39}$$

The covariance matrix between $\mathbf{e}_1$ and $\mathbf{y}_2$ is

$$\mathrm{E}\Big[\big(\mathbf{e}_1 - \mathrm{E}[\mathbf{e}_1]\big)\big(\mathbf{y}_2 - \mathrm{E}\big[\mathbf{y}_2\big]\big)'\Big] \tag{6.40}$$

$$
\begin{aligned}
&= \mathrm{E}\left[\mathbf{e}_1\left(-\mathbf{y}_1\gamma_{21} + \mathbf{x}_2\beta_{22} + \mathbf{e}_2 - \mathrm{E}\left[-\mathbf{y}_1\gamma_{21} + \mathbf{x}_2\beta_{22} + \mathbf{e}_2\right]\right)'\right] \\
&= \mathrm{E}\left[\mathbf{e}_1\left(-\mathbf{y}_1\gamma_{21} + \mathbf{x}_2\beta_{22} + \mathbf{e}_2 - \mathbf{x}_2\beta_{22} + \gamma_{21}\mathrm{E}\left[\mathbf{y}_1\right]\right)'\right] \\
&= \mathrm{E}\left[\mathbf{e}_1\left(-\mathbf{y}_1\gamma_{21} + \mathbf{e}_2 + \gamma_{21}\mathrm{E}\left[\mathbf{y}_1\right]\right)'\right] \\
&= \mathrm{E}\left[\mathbf{e}_1\left(\mathbf{e}_2 - \gamma_{21}\left(\mathbf{y}_1 - \mathrm{E}\left[\mathbf{y}_1\right]\right)\right)'\right] \\
&= \mathrm{E}\left[\mathbf{e}_1\left(\mathbf{e}_2 - \gamma_{21}\mathbf{e}_1\right)'\right] \qquad = \mathrm{E}\left[\mathbf{e}_1\mathbf{e}'_2 - \gamma_{21}\mathbf{e}_1\mathbf{e}'_1\right] \\
&= \sigma_{12}\mathbf{I} - \gamma_{21}\sigma_{11}\mathbf{I} \neq \mathbf{0}
\end{aligned} \tag{6.41}
$$

What, then, can we do? We can predict $\hat{\mathbf{y}}_1$ from the reduced form, which is

$$
\underset{T\times 1}{\mathbf{y}_1} = \underset{T\times K}{\mathbf{X}}\,\underset{K\times 1}{\boldsymbol{\Pi}_1} + \mathbf{u}_{1T\times 1} \tag{6.42}
$$

This estimation is based on the ordinary least squares estimates of the $\boldsymbol{\Pi}$ parameters that are obtained by regressing $\mathbf{y}_1$ on the entire set of exogenous variables, as follows from Eq. (6.42). This means that all the variables found throughout all the equations in the system are included and not just the variables in the first equation of the system of equations. The OLS estimator is

$$
\hat{\boldsymbol{\Pi}}_1 = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}_1 \tag{6.43}
$$

Therefore, the predicted values of $\mathbf{y}_1$ are given by

$$
\hat{\mathbf{y}}_1 = \mathbf{X}\hat{\boldsymbol{\Pi}}_1 \tag{6.44}
$$

Note that $\hat{\mathbf{y}}_1$ is not correlated with $\mathbf{e}_1$, because the $\mathbf{X}$s are uncorrelated with $\mathbf{e}_1$ and that $\hat{\mathbf{y}}_2$ is not correlated with $\mathbf{e}_1$ because $\mathbf{e}_2$ has been removed. Therefore, we can replace $\mathbf{y}_2$ in Eq. (6.38) by its predicted value $\hat{\mathbf{y}}_2$.

6.2.2 Two-Stage Least Squares (2SLS)

The two-stage least squares estimation follows directly from the conclusion derived in the prior section. We can remove the bias introduced by the endogeneity of the dependent variables by regressing separately each endogenous variable on the full set of exogenous variables in a first stage; we can then use the estimated coefficients to predict each endogenous variable. In the second stage, each equation is estimated separately using the model as specified in each equation but replacing the actual

values of the endogenous variables specified on the right side of the equation by its predicted values as computed from the first stage. More specifically:

Stage 1: Using ordinary least squares, regress each $\mathbf{y}$ on all exogenous variables $\mathbf{X}$:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Pi} + \mathbf{U} \tag{6.45}$$

$$\Rightarrow \hat{\boldsymbol{\Pi}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \tag{6.46}$$

and compute the predicted endogenous variables Y:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\Pi}} \tag{6.47}$$

Stage 2: Using ordinary least squares, regress each $\mathbf{y}_n$ on the exogenous variables of the equation for $\mathbf{y}_n$ and on the predicted endogenous as well as exogenous variables specified in that equation:

$$\mathbf{y}_n = \hat{\mathbf{Z}}_n\boldsymbol{\alpha}_n + \mathbf{e}_n \tag{6.48}$$

The estimated parameters $\hat{\boldsymbol{\Gamma}}_n$ and $\hat{\mathbf{B}}_n$ are unbiased.

However, because of the nonzero covariances ($\Sigma \neq \text{diag}(\sigma_{nn})$), the estimation does not provide efficient estimators. As we discuss in the next section, the purpose, therefore, of the third stage in the three-stage least squares estimation method is to obtain efficient estimates, at least asymptotically.

6.2.3 Three-Stage Least Squares (3SLS)

The first two stages of 3SLS are identical to those described above for the two-stage least squares estimation. We now add the third stage:

Stage 3: (i) Compute the residuals for each equation from the estimated coefficients obtained in the second stage:

$$\hat{\mathbf{e}}_n = \mathbf{y}_n - \hat{\mathbf{Z}}_n\hat{\boldsymbol{\alpha}}_n \tag{6.49}$$

(ii) Estimate the contemporaneous covariance matrix $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1N} \\ \hat{\sigma}_{12} & \hat{\sigma}_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \\ \hat{\sigma}_{1\nu} & & & \\ \vdots & & & \\ \hat{\sigma}_{1N} & \cdots & \cdots & \hat{\sigma}_{NN} \end{bmatrix} \tag{6.50}$$

where

$$\hat{\sigma}_{in} = \frac{1}{T-K} \underset{1\times T}{\hat{\mathbf{e}}'_i} \underset{T\times 1}{\hat{\mathbf{e}}_n} \tag{6.51}$$

(iii) Compute the EGLS estimate similar to the SUR case using the following system of equations

$$\begin{cases} \mathbf{y}_1 = \hat{\mathbf{Z}}_1 \boldsymbol{\alpha}_1 + \mathbf{e}_1 \\ \mathbf{y}_2 = \hat{\mathbf{Z}}_2 \boldsymbol{\alpha}_2 + \mathbf{e}_2 \\ \quad \vdots \\ \mathbf{y}_N = \hat{\mathbf{Z}}_N \boldsymbol{\alpha}_N + \mathbf{e}_N \end{cases}$$

6.3 Simultaneity and Identification

Simultaneity in a system of equations introduces a complexity in the sense that it raises the question of how to distinguish between the effect of y_1 on y_2 and the effect of y_2 on y_1. In this section, we examine this problem and discuss methods of ensuring that it is possible to make such a distinction, i.e., that it is possible to identify the system of equations.

6.3.1 The Problem

The typical example used in economics to discuss the problem of identification concerns the interrelationship of supply and demand. While the supply and demand curves in the price–quantity map can be represented as in Fig. 6.1, we only observe P_t and Q_t.

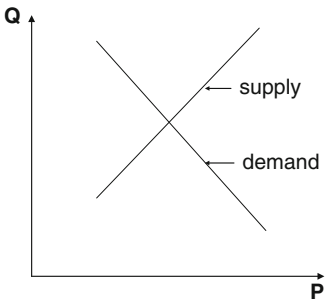


Fig. 6.1 Supply and demand curves

The question consists, therefore, in determining how we can differentiate empirically between these two curves.

We now provide a similar marketing example using sales and advertising expenditures to illustrate the problem. While sales are a function of advertising expenditures, advertising budgets are very often a reflection of the level of sales. This is an issue especially with cross-sectional data. Therefore, we are dealing with the following two functions:

$$\text{Equation 1}: \mathrm{S_t} = f(\mathrm{A_t}) \tag{6.52}$$

$$\text{Equation 2}: \mathrm{A_t} = g(\mathrm{S_t}) \tag{6.53}$$

The first equation is the market response function. The second equation is the marketing decision function.

Fortunately, in most circumstances sales are not driven solely by advertising. Similarly, the decision regarding the advertising budget is a complex one.

The solution to the identification problem resides in specifying additional (exogenous) variables that will help differentiate the two curves. It is important to note that these additional variables must be different across equations; otherwise, the problem remains.

6.3.2 Order and Rank Conditions

We now present two conditions that provide information regarding the identification of a system of equations. The second condition, known as the rank condition, guarantees identification but is complex to verify. The first condition, called the order condition, is simple to apply but does not guarantee identification. The order and the rank conditions are alternative ways of verifying the identification of a system of equations.

6.3.2.1 Order Condition

If an equation n is identified, then the number of excluded variables in that equation is equal to at least the number of equations minus one (i.e., $N - 1$). Therefore, checking for the order condition consists in making sure that each equation excludes on the right side at least $N - 1$ variables (exogenous or endogenous).

This order condition is necessary but not sufficient for the system of equations to be identified.

6.3.2.2 Rank Condition

The rank condition provides necessary and sufficient conditions for identification.

Recall the system of equations for a time period or cross section t:

$$\underset{1\times N}{\mathbf{y}'_t}\ \underset{N\times N}{\mathbf{\Gamma}} = \underset{1\times K}{\mathbf{x}'_t}\ \underset{K\times N}{\mathbf{B}} + \underset{1\times N}{\boldsymbol{\varepsilon}'_t} \tag{6.54}$$

We will use the example with two equations, which, for a time period t, can be written as

$$(y_{1t} \quad y_{2t})\begin{pmatrix}\gamma_{11} & \gamma_{21}\\ \gamma_{12} & \gamma_{22}\end{pmatrix} = (x_{1t} \quad x_{2t})\begin{pmatrix}\beta_{11} & \beta_{21}\\ \beta_{12} & \beta_{22}\end{pmatrix} + (\varepsilon_{1t} \quad \varepsilon_{2t}) \tag{6.55}$$

or

$$\begin{cases}\gamma_{11}y_{1t} + \gamma_{12}y_{2t} = \beta_{11}x_{1t} + \beta_{12}x_{2t} + \varepsilon_{1t}\\ \gamma_{21}y_{1t} + \gamma_{22}y_{2t} = \beta_{21}x_{1t} + \beta_{22}x_{2t} + \varepsilon_{2t}\end{cases} \tag{6.56}$$

It should be clear from Eq. (6.56) that the two equations are indistinguishable. More generally, from Eq. (6.54)

$$\mathbf{y}'_t\mathbf{\Gamma} - \mathbf{x}'_t\mathbf{B} = \boldsymbol{\varepsilon}'_t \tag{6.57}$$

or

$$\left(\mathbf{y}'_t\mathbf{x}'_t\right)\begin{pmatrix}\mathbf{\Gamma}\\ -\mathbf{B}\end{pmatrix} = \varepsilon'_t$$

Let

$$\mathbf{A} = \begin{pmatrix}\mathbf{\Gamma}\\ -\mathbf{B}\end{pmatrix} = [\boldsymbol{\alpha}_1 \ \ \boldsymbol{\alpha}_2 \ldots \ \ \boldsymbol{\alpha}_n \ldots \ \ \boldsymbol{\alpha}_N] \tag{6.58}$$

Using again the case of two equations expressed in Eq. (6.56),

$$\mathbf{A} = \begin{bmatrix}\gamma_{11} & \gamma_{21}\\ \gamma_{12} & \gamma_{22}\\ -\beta_{11} & -\beta_{21}\\ -\beta_{12} & -\beta_{22}\end{bmatrix} = [\boldsymbol{\alpha}_1 \quad \boldsymbol{\alpha}_2] \tag{6.59}$$

Let $\mathbf{r}_n$ be the row vector of zeros and ones, which when applied to the corresponding column vector $\boldsymbol{\alpha}_n$ defines a restriction imposed on Equation n.

For example, the restriction on Equation 1 that $\beta_{11} = 0$ can be expressed in a general way as $\mathbf{r}_1\,\boldsymbol{\alpha}_1 = 0$.

It follows that $\beta_{11} = 0$ by defining $\mathbf{r}_1 = (0\ 0\ 1\ 0)$

Indeed, we then have

$$\mathbf{r}_1 \ \boldsymbol{\alpha}_1 = (0 \quad 0 \quad 1 \quad 0) \begin{pmatrix} \gamma_{11} \\ \gamma_{12} \\ -\beta_{11} \\ -\beta_{12} \end{pmatrix} = \mathbf{0} \tag{6.60}$$

$$\Leftrightarrow \beta_{11} = 0$$

By post-multiplying the restriction vector $\mathbf{r}_n$ by the matrix $\mathbf{A}$, the rank condition for the equation n to be identified is that the rank of this matrix is at least equal to the number of equations minus one. The equation is just-identified if $\rho(\mathbf{r}_n\mathbf{A}) = N - 1$. If the rank is less than $N - 1$, the equation is under-identified. If the rank is greater than $N - 1$, the equation is over-identified. The equation must be just or over-identified to be able to obtain parameter estimates. For example,

$$\mathbf{r}_1 \ \mathbf{A} = (0 \quad 0 \quad 1 \quad 0) \begin{bmatrix} \gamma_{11} & \gamma_{21} \\ \gamma_{12} & \gamma_{22} \\ -\beta_{11} & -\beta_{21} \\ -\beta_{12} & -\beta_{22} \end{bmatrix} \tag{6.61}$$

$$= (-\beta_{11} \quad -\beta_{21}) = (0 \quad -\beta_{21}) \tag{6.62}$$

if $\beta_{21} \neq 0$, then $\rho(\mathbf{r}_1 \mathbf{A}) = 1$. Because $N - 1 = 1$ ($N = 2$), the first equation is just-identified.

6.4 Conclusion

In this chapter, we have presented the issues that arise when estimating models involving different cases of simultaneity of variables. The various cases presented in the above sections can be defined by the structure of two of the matrices defined earlier, i.e., $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$.

6.4.1 Structure of Γ Matrix

If the matrix $\boldsymbol{\Gamma}$ is diagonal, the system of equations is not simultaneous, except as expressed by the correlation of the error terms. In such a case, the model corresponds to the case of SURs. If the matrix $\boldsymbol{\Gamma}$ is not diagonal but triangular, this also results in a system that is not truly simultaneous. In such a case, one dependent variable may affect another but not the inverse. The system is then recursive. The various estimations that are appropriate for each of these cases are summarized in Fig. 6.2.

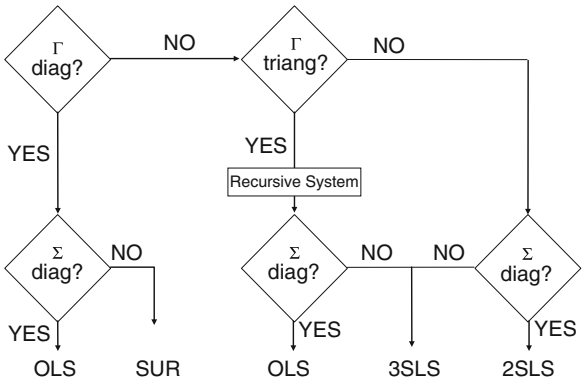


Fig. 6.2 Model specification and estimation methods (adapted from Parsons and Schultz 1976)

As shown in Fig. 6.2, the estimation method depends on the model specification as reflected in the matrix $\mathbf{\Gamma}$ discussed above and in the covariance structure of the error term $\mathbf{\Sigma}$.

6.4.2 Structure of Σ Matrix

When $\mathbf{\Gamma}$ is diagonal, the SUR estimator provides an efficient estimator if the covariance matrix $\mathbf{\Sigma}$ is not diagonal; otherwise (i.e., in the absence of correlated errors), each equation can be estimated separately by OLS since the results are identical to the SUR estimator.

If the $\mathbf{\Gamma}$ matrix is triangular, i.e., the case of a recursive system, OLS estimation of each equation separately provides unbiased parameter estimates. However, in the case where the covariance matrix $\mathbf{\Sigma}$ is not diagonal, the covariance structure must be taken into consideration and the EGLS obtained from the 3SLS procedure provides an efficient estimator. Note that in this case the second stage in 3SLS does not serve any purpose because there is no bias to correct for due to simultaneity (which only arises with non-recursive systems of equations). If $\mathbf{\Sigma}$ is diagonal, there is no need to proceed with multiple stage estimation and parameters can be estimated via OLS.

Finally, if the system of equations is simultaneous, i.e., $\mathbf{\Gamma}$ is neither diagonal nor triangular, the OLS estimators would be biased. Therefore, depending on whether $\mathbf{\Sigma}$ is diagonal or not, 2SLS or 3SLS should be used.

This points out the importance of knowing the structure of the covariance matrix $\mathbf{\Sigma}$. In most cases, it is an empirical question. Therefore, it is critical to estimate the covariance matrix, to report it, and to use the appropriate estimator. This means that a test must be performed to check the structure of the error term covariance matrix $\mathbf{\Sigma}$.

6.4.3 Test of Covariance Matrix

The test concerns the hypothesis that the correlation matrix of the error terms is the identity matrix (Morrison 1976):

$$\begin{cases} H_0 : \mathbf{R} = \mathbf{I} \\ H_1 : \mathbf{R} \neq \mathbf{I} \end{cases} \tag{6.63}$$

where $\mathbf{R}$ is the correlation matrix computed from the covariance matrix $\mathbf{\Sigma}$.

Two statistical tests of the identity structure of a correlation matrix are possible.

6.4.3.1 Bartlett's Test

The following function of the determinant of the correlation matrix follows a chi-square distribution with ν degrees of freedom:

$$-\left(T - 1 - \frac{2N+5}{6}\right)\mathrm{Ln}|\mathbf{R}| = {\chi_\nu}^2 \tag{6.64}$$

where T is the number of observations in each equation, N is the number of equations and $\nu = \frac{1}{2}N(N-1)$, i.e., the number of correlations in the correlation matrix.

6.4.3.2 Lawley's Approximation

The statistic as expressed in Eq. (6.64) can be approximated by

$$\left(T - 1 - \frac{2N+5}{6}\right)\sum_i \sum_{j>i} r_{ij}^2 = {\chi_\nu}^2 \tag{6.65}$$

where only the upper half of the correlations is considered in the summation.

6.4.4 Use of 3SLS Versus 2SLS

The EGLS estimator is only asymptotically more efficient than the OLS estimator. Consequently, in small samples, the property of the EGLS estimator is not clear. Therefore, when the sample size is small, it may be appropriate to report the 2SLS estimates instead of the 3SLS ones.

6.5 Examples of Estimation of Systems of Equations Using SAS and STATA

The three estimation methods presented in this chapter—SUR, 2SLS, and 3SLS—are now illustrated with examples using SAS and STATA.

6.5.1 Seemingly Unrelated Regression Example

In the following example, three characteristics of innovations developed by firms are modeled as a function of company factors and industry characteristics. We first present the analysis performed using SAS and then using STATA.

The SAS file is presented without going into the details of the substantive content of the model in order to focus on the technical aspects. Figure 6.3 shows that after reading the file containing the data, SAS standardizes the variables and builds the scales. The model is specified within the SAS *procedure SYSLIN* for systems of linear equations. The SUR statement following the PROC SYSLIN commands indicates that the parameters will be estimated using SUR. The dependent variables concern the relative advantage of the innovation, the radicalness of the innovation and its relative cost. The *model* statements for each equation specify the independent or predictor variables. Some variables are the same but others are different across equations.

The same model can also be estimated with iterative SUR (ITSUR). The only difference with the single iteration SUR in the SAS commands is that SUR is replaced with ITSUR (see Fig. 6.4).

We will take advantage of this SAS example to illustrate the use of STATA to read a data file that requires multiple lines per observation, in which case a dictionary complementary file is set up. Figure 6.5 lists the information on the variables to be read and their structure (i.e., on which line the information on each observation for each variable can be found).

The actual input file in STATA is listed in Fig. 6.6.

The example in Fig. 6.6 highlights the distinction between the "regress" command presented in the previous chapter for multiple regression analysis and SUR estimation. The commands highlighted in grey show that the procedure "sureg" is the STATA equivalent of "SUR" in SAS. Each equation is specified within parentheses with the list of variables used in that equation. The dependent variable is listed first within the parentheses. The last line "mat list e(Sigma)" instructs STATA to display the covariance matrix of error terms.

The output of the SUR estimation with SAS is shown in Fig. 6.7.

The output of the iterative method ITSUR estimation is shown in Fig. 6.8.

First, in both cases the OLS estimation is performed separately for each equation and the results are printed in the output.

```
/*    Examp6-1.sas     */
option ls=120;
data raw;
infile ' c:\SAMD\Chapter6\Examples\innov.asc ';
input #1 L1C1 L1C7 L1C10 L1C14 L1C19 L1C21 L1C23 L1C25 L1C27
       L1C29 L1C31 L1C33
      #2 L1C35 L1C37 L1C39 L1C41 L1C43 L1C45 L1C47 L1C49 L1C51
         . . .
/*----MISSING VALUES----*/
IF L1C7 =99 THEN L1C7=.;
IF L1C10=999 THEN L1C10=.;
IF L1C14=999 THEN L1C14=.;
         . . .
/*----reversal of items----*/
L1C21R=7-L1C21;
L1C23R=7-L1C23;
        . . .

/* Standardization of Variables*/
l1c41rs=l1c41r;
L1c45s=l1c45;
L1c53s=l1c53;
l1c55s=l1c55;
        . . .

proc standard mean=0 std=1 out=scale;
  var l1c41rs l1c45s l1c53s l1c55s l1c73s l1c61s l2c19s l1c69s
      l1c33rs l1c39s
        . . .
      L4C11s L4C67s L4C71s l1c59s l2c69s;

data data2;
set scale;
grow0=l1c14;
grow1=l2c7;

tech=sum(of L1C41Rs L1C45s L1C53s L1C55s L1C73s)/
     n(of L1C41Rs L1C45s L1C53s L1C55s L1C73s);

comp1=sum(of L1C59s L1C61s l2c19s)/
     n(of L1C59s L1C61s l2c19s);
        . . .

proc syslin sur;
     model dadv1 =  dto1 dres1;
     model dradic1 = dco11 dto1 dgrow01 ddemuncl dres1;
     model dcost1 = dto1 ic1 dgrow01 ddemuncl dres1;
run;
```

Fig. 6.3 Example of SAS input file for SUR estimation (examp6-1.sas)

```
proc syslin itsur;
     model dadv1 =  dto1 dres1;
     model dradic1 = dco11 dto1 dgrow01 ddemuncl dres1;
     model dcost1 = dto1 ic1 dgrow01 ddemuncl dres1;
run;
```

Fig. 6.4 Example of SAS input file for iterative SUR estimation (examp6-2.sas)

Then, the correlations from the residuals estimated from the OLS estimates are shown. A test should be performed to check that the correlation matrix is statistically significantly different from the identity matrix in order to detect whether it is useful to use the SUR estimator.

Finally, the SUR estimates are provided for each equation.

```
infile dictionary using "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter6-SUR-
2SLS-3SLS/innov.asc" {
_lines(14)
_line(1)
      L1C1 L1C7 L1C10 L1C14 L1C19 L1C21 L1C23 L1C25 L1C27 L1C29 L1C31 L1C33
_line(2)
   L1C35 L1C37 L1C39 L1C41 L1C43 L1C45 L1C47 L1C49 L1C51 L1C53 L1C55 L1C57 L1C59
_line(3)
    L1C61 L1C63 L1C65 L1C67 L1C69 L1C71 L1C73 L1C75 L1C77 L2C7 L2C12 L2C16 L2C19
_line(4)
   L2C21 L2C23 L2C25 L2C27 L2C29 L2C31 L2C33 L2C35 L2C37 L2C39 L2C41 L2C43 L2C45
_line(5)
   L2C47 L2C49 L2C51 L2C53 L2C55 L2C57 L2C59 L2C61 L2C63 L2C65 L2C67 L2C69 L2C71
_line(6)
    L2C73 L2C75 L2C77 L2C79 L3C7 L3C9 L3C11 L3C13 L3C15 L3C17 L3C19 L3C21 L3C23
_line(7)
   L3C25 L3C27 L3C29 L3C33 L3C37 L3C40 L3C43 L3C45 L3C47 L3C49 L3C51 L3C53 L3C55
_line(8)
    L3C57 L3C59 L3C61 L3C63 L3C65 L3C67 L3C69 L3C71 L3C73 L3C75 L3C77 L3C79 L4C7
_line(9)
    L4C9 L4C11 L4C13 L4C15 L4C17 L4C19 L4C21 L4C23 L4C25 L4C27 L4C29 L4C31 L4C33
_line(10)
   L4C35 L4C37 L4C39 L4C41 L4C43 L4C45 L4C47 L4C49 L4C51 L4C53 L4C55 L4C57 L4C59
_line(11)
    L4C61 L4C63 L4C65 L4C67 L4C69 L4C71 L4C73 L4C75 L4C77 L4C79 L5C7 L5C9 L5C11
_line(12)
   L5C13 L5C15 L5C17 L5C19 L5C21 L5C23 L5C25 L5C27 L5C29 L5C31 L5C33 L5C35 L5C37
_line(13)
   L5C39 L5C41 L5C43 L5C45 L5C47 L5C49 L5C51 L5C53 L5C55 L5C57 L5C59 L5C61 L5C63
_line(14)
   L5C65 L5C67 L5C69 L5C71
}
```

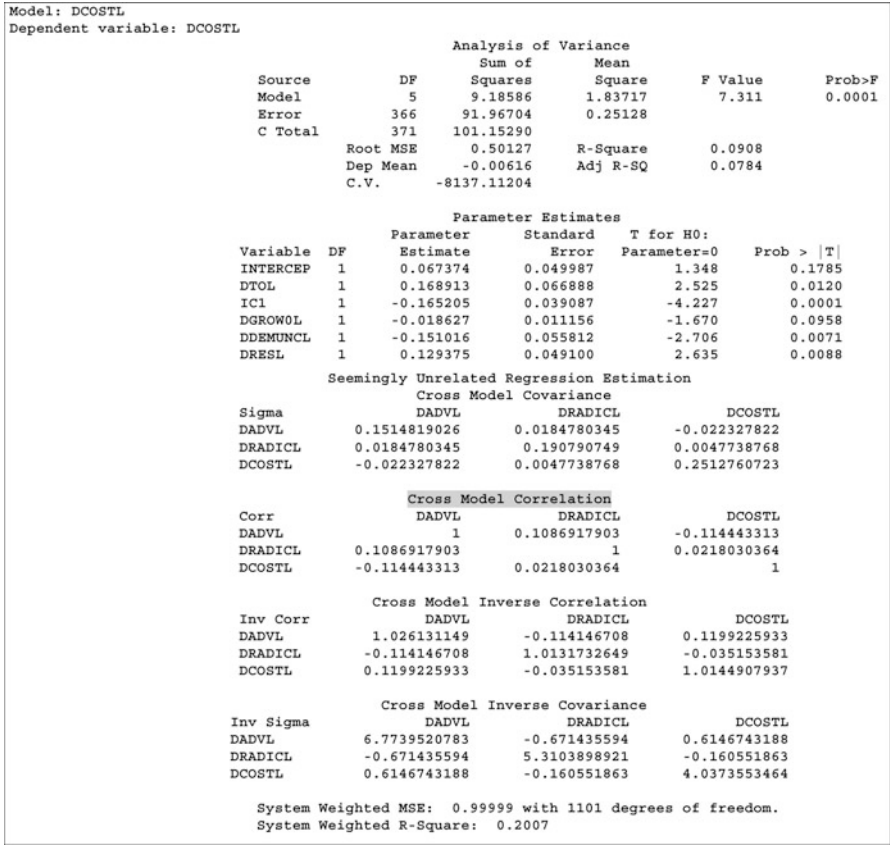
Fig. 6.5 Example of dictionary file to specify data structure and variable list in STATA (Techdic_Mac.dct)

```
infile using "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter6-SUR-2SLS-
3SLS/Techdic_Mac.dct", clear
* missing values
replace L1C7=. if L1C7 ==99
replace L1C10=. if L1C10==999
replace L1C14=. if L1C14==999
…
alpha L1C41 L1C45 L1C53 L1C55 L1C73, generate(tech) reverse(L1C41) std
alpha L1C59 L1C61 L2C19, generate(comp1) std
generate grow0=L1C14
generate grow1=L2C7
…
generate dco1l=log(dco1)
generate dcpo1l=log(dcpo1)
generate dto1l=log(dto1)
generate dto2l=log(dto2)
generate dto1=log(dto)
…
regress dadv1 dto1 dres1
regress dradic1 dco1l dto1 dgrow01 ddemunc1 dres1
regress dcost1 dto1 ic1 dgrow01 ddemunc1 dres1
sureg (dadv1 dto1 dres1) ///
(dradic1 dco1l dto1 dgrow01 ddemunc1 dres1) ///
(dcost1 dto1 ic1 dgrow01 ddemunc1 dres1)
mat list e(Sigma)
```

Fig. 6.6 Example of STATA input file for SUR estimation (examp6-1_Mac.do)

```
                                    SYSLIN Procedure
                                Ordinary Least Squares Estimation
Model: DADVL
Dependent variable: DADVL

                                    Analysis of Variance
                                       Sum of       Mean
      Source           DF            Squares      Square      F Value       Prob>F
      Model             2           16.80451     8.40225       55.467       0.0001
      Error           369           55.89682     0.15148
      C Total         371           72.70133
                Root MSE            0.38921     R-Square       0.2311
                Dep Mean           -0.02430     Adj R-SQ       0.2270
                C.V.           -1601.97200

                                    Parameter Estimates
                                  Parameter      Standard    T for H0:
      Variable  DF                Estimate         Error   Parameter=0    Prob > |T|
      INTERCEP   1               -0.018943      0.020187        -0.938        0.3487
      DTOL       1                0.272755      0.048234         5.655        0.0001
      DRESL      1                0.223258      0.037982         5.878        0.0001
Model: DRADICL
Dependent variable: DRADICL

                                    Analysis of Variance
                                       Sum of       Mean
      Source           DF            Squares      Square      F Value       Prob>F
      Model             5           26.08201     5.21640       27.341       0.0001
      Error           366           69.82941     0.19079
      C Total         371           95.91142
                Root MSE            0.43680     R-Square       0.2719
                Dep Mean           -0.00971     Adj R-SQ       0.2620
                C.V.           -4500.00823

                                    Parameter Estimates
                                  Parameter      Standard    T for H0:
      Variable  DF                Estimate         Error   Parameter=0    Prob > |T|
      INTERCEP   1               -0.103912      0.043629        -2.382        0.0177
      DCO1L      1               -0.082012      0.044723        -1.834        0.0675
      DTOL       1                0.611063      0.058287        10.484        0.0001
      DGROW0L    1                0.024730      0.009733         2.541        0.0115
      DDEMUNCL   1               -0.114126      0.048688        -2.344        0.0196
      DRESL      1               -0.066688      0.042855        -1.556        0.1205
```

Fig. 6.7 Example of SAS output file for SUR estimation (examp6-1.lst)



Model: DCOSTL
Dependent variable: DCOSTL

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	9.18586	1.83717	7.311	0.0001
Error	366	91.96704	0.25128		
C Total	371	101.15290			

Root MSE	0.50127	R-Square	0.0908
Dep Mean	-0.00616	Adj R-SQ	0.0784
C.V.	-8137.11204		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.067374	0.049987	1.348	0.1785
DTOL	1	0.168913	0.066888	2.525	0.0120
IC1	1	-0.165205	0.039087	-4.227	0.0001
DGROW0L	1	-0.018627	0.011156	-1.670	0.0958
DDEMUNCL	1	-0.151016	0.055812	-2.706	0.0071
DRESL	1	0.129375	0.049100	2.635	0.0088

Seemingly Unrelated Regression Estimation
Cross Model Covariance

Sigma	DADVL	DRADICL	DCOSTL
DADVL	0.1514819026	0.0184780345	-0.022327822
DRADICL	0.0184780345	0.190790749	0.0047738768
DCOSTL	-0.022327822	0.0047738768	0.2512760723

Cross Model Correlation

Corr	DADVL	DRADICL	DCOSTL
DADVL	1	0.1086917903	-0.114443313
DRADICL	0.1086917903	1	0.0218030364
DCOSTL	-0.114443313	0.0218030364	1

Cross Model Inverse Correlation

Inv Corr	DADVL	DRADICL	DCOSTL
DADVL	1.026131149	-0.114146708	0.1199225933
DRADICL	-0.114146708	1.0131732649	-0.035153581
DCOSTL	0.1199225933	-0.035153581	1.0144907937

Cross Model Inverse Covariance

Inv Sigma	DADVL	DRADICL	DCOSTL
DADVL	6.7739520783	-0.671435594	0.6146743188
DRADICL	-0.671435594	5.3103898921	-0.160551863
DCOSTL	0.6146743188	-0.160551863	4.0373553464

System Weighted MSE: 0.99999 with 1101 degrees of freedom.
System Weighted R-Square: 0.2007

Fig. 6.7 (continued)

```
Model: DADVL
Dependent variable: DADVL

                                  Parameter Estimates
                               Parameter       Standard    T for H0:
                Variable  DF    Estimate          Error   Parameter=0    Prob > |T|
                INTERCEP   1   -0.018943       0.020187        -0.938        0.3487
                DTOL       1    0.272755       0.048234         5.655        0.0001
                DRESL      1    0.223258       0.037982         5.878        0.0001

Model: DRADICL
Dependent variable: DRADICL

                                  Parameter Estimates
                               Parameter       Standard    T for H0:
                Variable  DF    Estimate          Error   Parameter=0    Prob > |T|
                INTERCEP   1   -0.105821       0.043440        -2.436        0.0153
                DCO1L      1   -0.084707       0.044444        -1.906        0.0574
                DTOL       1    0.612306       0.058237        10.514        0.0001
                DGROW0L    1    0.025224       0.009675         2.607        0.0095
                DDEMUNCL   1   -0.112170       0.048400        -2.318        0.0210
                DRESL      1   -0.066688       0.042852        -1.556        0.1205
Model: DCOSTL
Dependent variable: DCOSTL

                                  Parameter Estimates
                               Parameter       Standard    T for H0:
                Variable  DF    Estimate          Error   Parameter=0    Prob > |T|
                INTERCEP   1    0.069499       0.049748         1.397        0.1633
                DTOL       1    0.169196       0.066825         2.532        0.0118
                IC1        1   -0.165555       0.038818        -4.265        0.0001
                DGROW0L    1   -0.019188       0.011083        -1.731        0.0842
                DDEMUNCL   1   -0.153609       0.055445        -2.770        0.0059
                DRESL      1    0.129626       0.049097         2.640        0.0086
```

Fig. 6.7 (continued)

```
                    Iterative Seemingly Unrelated Regression Estimation
                                  Cross Model Covariance
 Sigma                DADVL           DRADICL           DCOSTL
 DADVL        0.1514819026      0.0185171321      -0.022348653
 DRADICL      0.0185171321      0.1907948382      0.0047481056
 DCOSTL       -0.022348653      0.0047481056      0.2512793046

                                 Cross Model Correlation
 Corr                 DADVL           DRADICL           DCOSTL
 DADVL                    1      0.1089206033      -0.114549349
 DRADICL      0.1089206033                 1      0.0216849635
 DCOSTL       -0.114549349      0.0216849635                 1

                           Cross Model Inverse Correlation
 Inv Corr             DADVL           DRADICL           DCOSTL
 DADVL          1.026207679      -0.114378043       0.120031705
 DRADICL       -0.114378043       1.0132186939      -0.035073541
 DCOSTL         0.120031705      -0.035073541       1.0145101221

                           Cross Model Inverse Covariance
 Inv Sigma            DADVL           DRADICL           DCOSTL
 DADVL          6.7744572875      -0.672789142       0.6152296236
 DRADICL       -0.672789142       5.3105141817      -0.160183559
 DCOSTL         0.6152296236      -0.160183559       4.0373803318

    System Weighted MSE:  1 with 1101 degrees of freedom.
    System Weighted R-Square:  0.2007

Model: DADVL
Dependent variable: DADVL

                                  Parameter Estimates
                                Parameter       Standard     T for H0:
Variable  DF                     Estimate          Error    Parameter=0    Prob > |T|
INTERCEP   1                   -0.018943       0.020187         -0.938        0.3487
DTOL       1                    0.272755       0.048234          5.655        0.0001
DRESL      1                    0.223258       0.037982          5.878        0.0001

Model: DRADICL
Dependent variable: DRADICL

                                  Parameter Estimates
                                Parameter       Standard     T for H0:
Variable  DF                     Estimate          Error    Parameter=0    Prob > |T|
INTERCEP   1                   -0.105825       0.043440         -2.436        0.0153
DC01L      1                   -0.084717       0.044443         -1.906        0.0574
DTOL       1                    0.612311       0.058237         10.514        0.0001
DGROW0L    1                    0.025226       0.009675          2.607        0.0095
DDEMUNCL   1                   -0.112166       0.048399         -2.318        0.0210
DRESL      1                   -0.066688       0.042852         -1.556        0.1205

Model: DCOSTL
Dependent variable: DCOSTL

                                  Parameter Estimates
                                Parameter       Standard     T for H0:
Variable  DF                     Estimate          Error    Parameter=0    Prob > |T|
INTERCEP   1                    0.069501       0.049748          1.397        0.1632
DTOL       1                    0.169194       0.066825          2.532        0.0118
IC1        1                   -0.165553       0.038818         -4.265        0.0001
DGROW0L    1                   -0.019189       0.011083         -1.731        0.0842
DDEMUNCL   1                   -0.153611       0.055445         -2.771        0.0059
DRESL      1                    0.129626       0.049098          2.640        0.0086
```

Fig. 6.8 Example of SAS output file for iterative ITSUR estimation (examp6-2.lst)

It can be seen from the output of the iterative SUR that the steps are identical. The estimates reported are those obtained at the last step when convergence is achieved.

The SUR output using STATA is shown in Fig. 6.9.

6.5.2 Two-Stage Least Squares Example

In the example here for two-stage least squares (as well as the example for three-stage least squares in the next section), we now specify some endogeneity in the system in that some variables on the left side of an equation can also be found on the right side of another equation. In the example shown in Fig. 6.10, the model

```
. sureg (dadvl dtol dresl) ///
> (dradicl dcoll dtol dgrow0l ddemuncl dresl) ///
> (dcostl dtol icl dgrow0l ddemuncl dresl)

Seemingly unrelated regression
------------------------------------------------------------------------------
Equation           Obs  Parms        RMSE    "R-sq"       chi2         P
------------------------------------------------------------------------------
dadvl              360      2    .3856953    0.2210     102.14    0.0000
dradicl            360      5    .4186645    0.2840     144.32    0.0000
dcostl             360      5    .4988242    0.0834      33.20    0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dadvl        |
        dtol |   .2715593   .0484446     5.61   0.000     .1766097     .366509
       dresl |   .2072549   .0380967     5.44   0.000     .1325868    .2819231
       _cons |  -.0117502    .020334    -0.58   0.563    -.0516042    .0281037
-------------+----------------------------------------------------------------
dradicl      |
       dcoll |  -.0974295   .0432495    -2.25   0.024    -.1821969   -.0126621
        dtol |   .6050885   .0565433    10.70   0.000     .4942657    .7159113
    dgrow0l |   .0343637   .0097226     3.53   0.000     .0153078    .0534197
   ddemuncl |  -.0944976   .0478577    -1.97   0.048    -.1882969   -.0006983
       dresl |  -.0726166   .0415731    -1.75   0.081    -.1540984    .0088651
       _cons |  -.1524931   .0443876    -3.44   0.001    -.2394912   -.0654951
-------------+----------------------------------------------------------------
dcostl       |
        dtol |   .1748069   .0672157     2.60   0.009     .0430665    .3065473
         icl |  -.1594804   .0390791    -4.08   0.000    -.236074    -.0828869
    dgrow0l |  -.0194596   .0115465    -1.69   0.092    -.0420902     .003171
   ddemuncl |  -.1342377   .0568209    -2.36   0.018    -.2456045   -.0228709
       dresl |   .1162543   .0494856     2.35   0.019     .0192644    .2132442
       _cons |   .0697575   .0526953     1.32   0.186    -.0335234    .1730385
. mat list e(Sigma)

symmetric e(Sigma)[3,3]
              dadvl     dradicl      dcostl
  dadvl    .14876085
dradicl    .01865257   .17526695
 dcostl   -.02493313  -.00039789   .24880884
```

Fig. 6.9 Example of STATA output file for SUR estimation (examp6-1_Mac.log)

```
/*    Examp6-3.sas */
       . . .

proc syslin 2SLS;
     endogenous dadvl dradicl dcostl;
     instruments dcoll dtol icl dresl dgrow0l ddemuncl;
     model dadvl =  dradicl dtol dresl;
     model dradicl = dcoll dtol dgrow0l ddemuncl dresl;
     model dcostl = dradicl dadvl dtol icl;
run;
```

Fig. 6.10 Example of SAS input file for two-stage least squares estimation (examp6-3.sas)

definition shows that the variable “dadvl” is a predicted variable and is also found in the equation to predict the “dcostl” variable.

The endogenous variables are identified by the command “ENDOGENOUS” followed by the names of these endogenous variables.

The statement “INSTRUMENTS” lists all the exogenous variables in the system. These variables will be used in the first stage of the estimation procedure to calculate the predicted values of the endogenous variables. These predicted values will then be used in the second stage of the estimation procedure.

The estimation method is simply indicated on the procedure line “proc syslin” by the “2SLS” command.

```
reg3 (dadvl dtol dresl) ///
(dradicl dcoll dtol dgrow01 ddemuncl dresl) ///
(dcostl dtol icl dgrow01 ddemuncl dresl), 2sls
  mat list e(Sigma)
```

Fig. 6.11 Example of STATA input file for two-stage least squares estimation (examp6-3_Mac.do)

```
                                    SYSLIN Procedure
                                Two-Stage Least Squares Estimation
Model: DADVL
Dependent variable: DADVL
                                        Analysis of Variance
                                           Sum of        Mean
                    Source          DF     Squares      Square      F Value       Prob>F
                    Model            3    16.84679     5.61560       35.317       0.0001
                    Error          368    58.51337     0.15900
                    C Total        371    72.70133
                              Root MSE      0.39875     R-Square       0.2236
                              Dep Mean     -0.02430     Adj R-SQ       0.2172
                              C.V.      -1641.26304

                                        Parameter Estimates
                                Parameter      Standard    T for H0:
                 Variable  DF    Estimate         Error   Parameter=0    Prob > |T|
                 INTERCEP   1    -0.019782      0.020746        -0.954        0.3410
                 DRADICL    1    -0.120184      0.233060        -0.516        0.6064
                 DTOL       1     0.342387      0.143788         2.381        0.0178
                 DRESL      1     0.214282      0.042630         5.027        0.0001
Model: DRADICL
Dependent variable: DRADICL
                                        Analysis of Variance
                                           Sum of        Mean
                    Source          DF     Squares      Square      F Value       Prob>F
                    Model            5    26.08201     5.21640       27.341       0.0001
                    Error          366    69.82941     0.19079
                    C Total        371    95.91142
                              Root MSE      0.43680     R-Square       0.2719
                              Dep Mean     -0.00971     Adj R-SQ       0.2620
                              C.V.      -4500.00823

                                        Parameter Estimates
                                Parameter      Standard    T for H0:
                 Variable  DF    Estimate         Error   Parameter=0    Prob > |T|
                 INTERCEP   1    -0.103912      0.043629        -2.382        0.0177
                 DCO1L      1    -0.082012      0.044723        -1.834        0.0675
                 DTOL       1     0.611063      0.058287        10.484        0.0001
                 DGROW0L    1     0.024730      0.009733         2.541        0.0115
                 DDEMUNCL   1    -0.114126      0.048688        -2.344        0.0196
                 DRESL      1    -0.066688      0.042855        -1.556        0.1205
Model: DCOSTL
Dependent variable: DCOSTL
                                        Analysis of Variance
                                           Sum of        Mean
                    Source          DF     Squares      Square      F Value       Prob>F
                    Model            4     7.24071     1.81018         4.530       0.0014
                    Error          367   146.66736     0.39964
                    C Total        371   101.15290
                              Root MSE      0.63217     R-Square       0.0470
                              Dep Mean     -0.00616     Adj R-SQ       0.0367
                              C.V.     -10261.91552

                                        Parameter Estimates
                                Parameter      Standard    T for H0:
                 Variable  DF    Estimate         Error   Parameter=0    Prob > |T|
                 INTERCEP   1     0.013729      0.033768         0.407        0.6846
                 DRADICL    1     0.410360      0.387537         1.059        0.2903
                 DADVL      1     0.698638      0.309916         2.254        0.0248
                 DTOL       1    -0.268182      0.300295        -0.893        0.3724
                 IC1        1    -0.158223      0.049659        -3.186        0.0016
```

Fig. 6.12 Example of SAS output file for two-stage least squares estimation (examp6-3.lst)

The STATA commands are given in Fig. 6.11, where the procedure "reg3" (highlighted in grey in the figure) corresponds to systems of equations (using three-stage least squares as the default, as discussed in the next section). However, the option "2SLS" shown at the end of the model specification (highlighted in grey in the figure) is selected to obtain a two-stage least squares estimation.

The output shown in Fig. 6.12 provides the estimates of the second stage for each equation.

The output for STATA is similarly provided in Fig. 6.13.

```
. reg3 (dadvl dtol dresl) ///
> (dradicl dcoll dtol dgrow01 ddemuncl dresl) ///
> (dcostl dtol icl dgrow01 ddemuncl dresl), 2sls

Two-stage least-squares regression
----------------------------------------------------------------------------
Equation           Obs  Parms        RMSE    "R-sq"     F-Stat        P
----------------------------------------------------------------------------
dadvl              360      2    .3873125    0.2210      50.64   0.0000
dradicl            360      5    .4221819    0.2841      28.09   0.0000
dcostl             360      5    .5030168    0.0834       6.45   0.0000
----------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dadvl        |
        dtol |   .2715593   .0486477     5.58   0.000     .1761031    .3670156
       dresl |   .2072549   .0382564     5.42   0.000     .1321884    .2823215
       _cons |  -.0117502   .0204193    -0.58   0.565    -.0518169    .0283164
-------------+----------------------------------------------------------------
dradicl      |
       dcoll |  -.0928441   .0439103    -2.11   0.035    -.1790047   -.0066835
        dtol |   .6032243   .0570729    10.57   0.000     .4912363    .7152124
     dgrow01 |   .0330821   .0098707     3.35   0.001     .0137138    .0524504
    ddemuncl |  -.0948834   .0485868    -1.95   0.051    -.1902201    .0004534
       dresl |  -.0727504   .0419269    -1.74   0.083    -.1550192    .0095184
       _cons |  -.1473469   .0449894    -3.28   0.001    -.2356249   -.0590689
-------------+----------------------------------------------------------------
dcostl       |
        dtol |   .1760208   .0678593     2.59   0.010     .0428678    .3091738
         icl |  -.1621388   .0397458    -4.08   0.000    -.2401277   -.0841498
     dgrow01 |  -.0178086   .0117429    -1.52   0.130    -.0408505    .0052333
    ddemuncl |  -.1333889   .0577877    -2.31   0.021    -.2467796   -.0199982
       dresl |   .1162032   .0499069     2.33   0.020     .0182762    .2141302
       _cons |   .0632322   .0534798     1.18   0.237    -.0417055      .16817
------------------------------------------------------------------------------
Endogenous variables:  dadvl dradicl dcostl
Exogenous variables:   dtol dresl dcoll dgrow01 ddemuncl icl
----------------------------------------------------------------------------

. mat list e(Sigma)

symmetric e(Sigma)[3,3]
                dadvl      dradicl      dcostl
  dadvl    .15001094
dradicl    .01888884    .17823758
 dcostl   -.02524896   -.00040464    .25302594
```

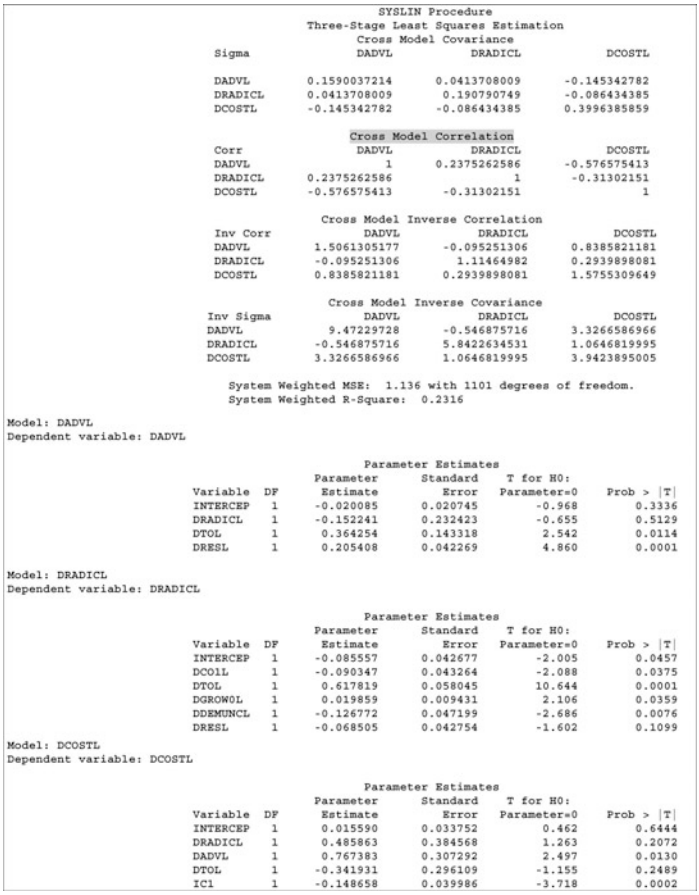
Fig. 6.13 Example of STATA output file for two-stage least squares estimation (examp6-3_Mac.log)

```
/*     Examp6-4.sas    */
      . . .
proc syslin 3SLS;
     endogenous dadvl dradicl dcostl;
     instruments dcoll dtol icl dresl dgrow01 ddemuncl;
     model dadvl =  dradicl dtol dresl;
     model dradicl = dcoll dtol dgrow01 ddemuncl dresl;
     model dcostl = dradicl dadvl dtol icl;
run;
```

Fig. 6.14 Example of SAS input file for three-stage least squares estimation (examp6-4.sas)

```
reg3 (dadvl dtol dresl) ///
(dradicl dcoll dtol dgrow01 ddemuncl dresl) ///
(dcostl dtol icl dgrow01 ddemuncl dresl)
mat list e(Sigma)
```

Fig. 6.15 Example of STATA input file for three-stage least squares estimation (examp6-4_Mac.do)



SYSLIN Procedure
Three-Stage Least Squares Estimation
Cross Model Covariance

Sigma	DADVL	DRADICL	DCOSTL
DADVL	0.1590037214	0.0413708009	-0.145342782
DRADICL	0.0413708009	0.190790749	-0.086434385
DCOSTL	-0.145342782	-0.086434385	0.3996385859

Cross Model Correlation

Corr	DADVL	DRADICL	DCOSTL
DADVL	1	0.2375262586	-0.576575413
DRADICL	0.2375262586	1	-0.31302151
DCOSTL	-0.576575413	-0.31302151	1

Cross Model Inverse Correlation

Inv Corr	DADVL	DRADICL	DCOSTL
DADVL	1.5061305177	-0.095251306	0.8385821181
DRADICL	-0.095251306	1.11464982	0.2939898081
DCOSTL	0.8385821181	0.2939898081	1.5755309649

Cross Model Inverse Covariance

Inv Sigma	DADVL	DRADICL	DCOSTL
DADVL	9.47229728	-0.546875716	3.3266586966
DRADICL	-0.546875716	5.8422634531	1.0646819995
DCOSTL	3.3266586966	1.0646819995	3.9423895005

System Weighted MSE: 1.136 with 1101 degrees of freedom.
System Weighted R-Square: 0.2316

Model: DADVL
Dependent variable: DADVL

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-0.020085	0.020745	-0.968	0.3336
DRADICL	1	-0.152241	0.232423	-0.655	0.5129
DTOL	1	0.364254	0.143318	2.542	0.0114
DRESL	1	0.205408	0.042269	4.860	0.0001

Model: DRADICL
Dependent variable: DRADICL

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-0.085557	0.042677	-2.005	0.0457
DCO1L	1	-0.090347	0.043264	-2.088	0.0375
DTOL	1	0.617819	0.058045	10.644	0.0001
DGROW0L	1	0.019859	0.009431	2.106	0.0359
DDEMUNCL	1	-0.126772	0.047199	-2.686	0.0076
DRESL	1	-0.068505	0.042754	-1.602	0.1099

Model: DCOSTL
Dependent variable: DCOSTL

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.015590	0.033752	0.462	0.6444
DRADICL	1	0.485863	0.384568	1.263	0.2072
DADVL	1	0.767383	0.307292	2.497	0.0130
DTOL	1	-0.341931	0.296109	-1.155	0.2489
IC1	1	-0.148658	0.039986	-3.718	0.0002

Fig. 6.16 Example of SAS output file for three-stage least squares estimation (examp6-4.lst)

6.5.3 Three-Stage Least Squares Example

Similar to the case of two-stage least squares, the estimation method is simply indicated on the procedure line "proc syslin" by the "3SLS" command, as shown in Fig. 6.14 (highlighted in grey). All other statements are identical to those for two-stage least squares.

Figure 6.15 gives the STATA input for three-stage least squares.

As noted in the previous section, the default of "reg3" is three-stage least squares. Consequently, there is no need to indicate "3SLS" as an option. The commands "mat list e(Sigma)" serve to obtain the covariance matrix of the error terms.

The SAS output for the 3SLS procedure first provides the estimates of the second stage for each equation. These estimates are not shown in Fig. 6.16 because they are identical to the SAS output shown in Fig. 6.12. In Fig. 6.16, however, the estimated

```
. reg3 (dadvl dtol dresl) ///
> (dradicl dco11 dtol dgrow01 ddemuncl dresl) ///
> (dcostl dtol ic1 dgrow01 ddemuncl dresl)

Three-stage least-squares regression
------------------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2         P
------------------------------------------------------------------------------
dadvl             360      2    .3856953    0.2210     102.14    0.0000
dradicl           360      5    .4186645    0.2840     144.32    0.0000
dcostl            360      5    .4988242    0.0834      33.20    0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dadvl        |
        dtol |   .2715593   .0484446     5.61   0.000     .1766097     .366509
       dresl |   .2072549   .0380967     5.44   0.000     .1325868    .2819231
       _cons |  -.0117502    .020334    -0.58   0.563    -.0516042    .0281037
-------------+----------------------------------------------------------------
dradicl      |
       dco11 |  -.0974295   .0432495    -2.25   0.024    -.1821969   -.0126621
        dtol |   .6050885   .0565433    10.70   0.000     .4942657    .7159113
     dgrow01 |   .0343637   .0097226     3.53   0.000     .0153078    .0534197
    ddemuncl |  -.0944976   .0478577    -1.97   0.048    -.1882969   -.0006983
       dresl |  -.0726166   .0415731    -1.75   0.081    -.1540984    .0088651
       _cons |  -.1524931   .0443876    -3.44   0.001    -.2394912   -.0654951
-------------+----------------------------------------------------------------
dcostl       |
        dtol |   .1748069   .0672157     2.60   0.009     .0430665    .3065473
         ic1 |  -.1594804   .0390791    -4.08   0.000     -.236074   -.0828869
     dgrow01 |  -.0194596   .0115465    -1.69   0.092    -.0420902     .003171
    ddemuncl |  -.1342377   .0568209    -2.36   0.018    -.2456045   -.0228709
       dresl |   .1162543   .0494856     2.35   0.019     .0192644    .2132442
       _cons |   .0697575   .0526953     1.32   0.186    -.0335234    .1730385
------------------------------------------------------------------------------
Endogenous variables:  dadvl dradicl dcostl
Exogenous variables:   dtol dresl dco11 dgrow01 ddemuncl ic1
------------------------------------------------------------------------------

. mat list e(Sigma)

symmetric e(Sigma)[3,3]
              dadvl     dradicl      dcostl
  dadvl    .14876085
dradicl    .01865257   .17526695
 dcostl   -.02493313  -.00039789   .24880884
```

Fig. 6.17 Example of STATA output file for three-stage least squares estimation (examp6-4.log)

correlation matrix of the error terms across equations is shown. A test of significance of the set of correlations can then be performed to know whether it can be useful to continue to the third stage. These third-stage EGLS estimates are then provided in the SAS output.

In Fig. 6.17, the STATA output immediately gives the estimated parameters of the third stage (i.e., without intermediary information). The covariance matrix of the error terms is displayed after the three-stage least squares estimates.

6.6 Assignment

The data found in the files INDUP.CSV and PANEL.CSV, which are described in Appendix C (Chap. 14), provide opportunities to apply the estimation of systems of equations discussed in this chapter. Chapter 5 describes how to read these data in SAS and STATA. The assignment consists simply in specifying a system of equations to be estimated via the proper estimation method. The modeling exercise should include (1) a system of seemingly unrelated equations or a recursive system of equations and (2) a model with simultaneous relationships.

Examples of such models can concern the following:

1. A model of the hierarchy of effects that consists of awareness, purchase intentions, and sales;
2. A model of the sales or market share for multiple segments or for multiple brands;
3. A model of a market response function and of the marketing decisions.

Proper justification for the estimation method used must be included (i.e., test of the covariance structure of the error terms).

Bibliography

Basic Technical Readings

Dhrymes, P. J. (1978). *Introductory econometrics*. New York, NY: Springer [Chap. 6].

Judge, G. G., Griffiths, W. E., Carter Hill, R., Lutkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics*. New York, NY: John Wiley and Sons [Chap. 14 and Chap. 15].

Morrison, D. F. (1976). *Multivariate statistical methods*. New York, NY: McGraw-Hill Book Company.

Parsons, L. J., & Schultz, R. L. (1976). *Marketing models and econometric research*. New York, NY: North Holland.

Theil, H. (1971). *Principles of econometrics*. New York, NY: John Wiley and Sons [Chap. 9 and Chap. 10].

Application Readings

Bass, F. M. (1969). A simultaneous equation regression study of advertising and sales of cigarettes. *Journal of Marketing Research, 6*(3), 291–300.

Bayus, B. L., & Putsis, W. P., Jr. (1999). Product proliferation: An empirical analysis of product line determinants and market outcomes. *Marketing Science, 18*(2), 137–153.

Beckwith, N. E. (1972). Multivariate analysis sales responses of competing brands to advertising. *Journal of Marketing Research, 9*(2), 168–176.

Cool, K., & Dierickx, I. (1993). Rivalry, strategic groups and firm profitability. *Strategic Management Journal, 14*, 47–59.

Cool, K., & Schendel, D. (1988). Performance differences among strategic group members. *Strategic Management Journal, 9*, 207–223.

Gatignon, H., & Xuereb, J.-M. (1997). Strategic orientation of the firm and new product performance. *Journal of Marketing Research, 34*(1), 77–90.

Lambin, J.-J., Naert, P., & Bultez, A. (1975). Optimal marketing behavior in oligopoly. *European Economic Review, 6*, 105–128.

Metwally, M. M. (1978). Escalation tendencies of advertising. *Oxford Bulletin of Economics and Statistics, 40*(2), 243–256.

Norton, J. A., & Bass, F. M. (1987). Diffusion and theory model of adoption and substitution for successive generations of high-technology products. *Management Science, 33*(9), 1069–1086.

Parker, P. M., & Roller, L.-H. (1997). Collusive conduct in duopolies: Multimarket contact and cross-ownership in the mobile telephone industry. *The RAND Journal of Economics, 28*(2), 304–322.

Reibstein, D., & Gatignon, H. (1984). Optimal product line pricing: The influence of elasticities and cross-elasticities. *Journal of Marketing Research, 21*(3), 259–267.

Schultz, R. L. (1971). Market measurement and planning with a simultaneous equation model. *Journal of Marketing Research, 8*(2), 153–164.

Wildt, A. (1974). Multifirm analysis of competitive decision variables. *Journal of Marketing Research, 11*(1), 50–62.

Chapter 7
Canonical Correlation Analysis

In canonical correlation analysis the objective is to relate a set of dependent or criterion variables to another set of independent or predictor variables. For example, we would like to establish the relationship between socioeconomic status and consumption by households. A set of characteristics determines socioeconomic status: education level, age, income, etc. Another set of variables measures consumption such as purchases of cars, luxury items, or food products.

7.1 The Method

In order to establish a relationship between these two sets of variables, we find two scalars, one defined as a linear combination of the dependent variables, and the other defined as a linear combination of the independent variables. The criterion used to judge the relationship between this set of independent variables with the set of dependent variables is simply the correlation between the two scalars. Canonical correlation analysis then consists in finding the weights to apply to the linear combinations of the independent and dependent variables that will maximize the correlation coefficient between those two linear combinations. The problem can be represented graphically as in Fig. 7.1.

In the figure, **z** and **w** represent two unobserved constructs that are correlated. The Xs are indicators that determine **z** and the Ys are indicators that determine **w**.

Formally, let $\underset{N\times p}{\mathbf{X}}$ be the matrix of p predictor variables (centered, i.e., taking the deviations from their means) on N observations and $\underset{N\times p}{\mathbf{Y}}$ be the matrix of q criterion variables (also centered) on the same N observations.

We will call z_i the scalar representing a linear combination of the independent variables for observation i. Therefore

$$\underset{1\times 1}{z_i} = \underset{1\times p}{\mathbf{x}'_i}\ \underset{p\times 1}{\mathbf{u}} \tag{7.1}$$

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_7,

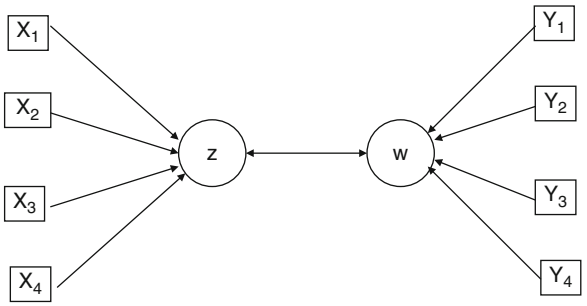


Fig. 7.1 Graphical representation of the canonical correlation model

Similarly, w_i is the scalar representing a linear combination of the dependent variables for observation i:

$$\underset{1\times 1}{w_i} = \underset{1\times q}{\mathbf{y}'_i}\ \underset{q\times 1}{\mathbf{v}} \tag{7.2}$$

The correlation between variables z and w is

$$r_{zw} = \frac{\sum_{i=1}^{N} z_i w_i}{\sqrt{\left(\sum_{i=1}^{N} z_i^2\right)\left(\sum_{i=1}^{N} w_i^2\right)}} \tag{7.3}$$

More compactly, for the N observations

$$\underset{N\times 1}{\mathbf{z}} = \underset{N\times p}{\mathbf{X}}\ \underset{p\times 1}{\mathbf{u}} \tag{7.4}$$

and

$$\underset{N\times 1}{\mathbf{w}} = \underset{N\times q}{\mathbf{Y}}\ \underset{q\times 1}{\mathbf{v}} \tag{7.5}$$

The problem consists in finding the vectors ($\mathbf{u}$, $\mathbf{v}$) so as to maximize the correlation between $\mathbf{z}$ and $\mathbf{w}$. In matrix notation, the correlation in Eq. (7.3) is

$$\boldsymbol{r}_{\boldsymbol{zw}} = \frac{\mathbf{z}'\mathbf{w}}{\sqrt{(\mathbf{z}'\mathbf{z})(\mathbf{w}'\mathbf{w})}} = \frac{\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v}}{\sqrt{(\mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u})(\mathbf{v}'\mathbf{Y}'\mathbf{Y}\mathbf{v})}} \tag{7.6}$$

Let $\mathbf{S}_{\mathrm{xy}} = \mathbf{X}'\mathbf{Y}$, $\mathbf{S}_{\mathrm{xx}} = \mathbf{X}'\mathbf{X}$, and $\mathbf{S}_{\mathrm{yy}} = \mathbf{Y}'\mathbf{Y}$. Then

$$r_{zw} = \frac{\mathbf{u}'\mathbf{S}_{\mathrm{xy}}\mathbf{v}}{\sqrt{(\mathbf{u}'\mathbf{S}_{\mathrm{xx}}\mathbf{u})(\mathbf{v}'\mathbf{S}_{\mathrm{yy}}\mathbf{v})}} \tag{7.7}$$

The latent variables z and w can be normalized without loss of generality and for determinacy, i.e.,

$$\mathbf{u}'\mathbf{S}_{xx}\mathbf{u} = \mathbf{v}'\mathbf{S}_{yy}\mathbf{v} = 1 \tag{7.8}$$

Therefore, the problem is to find $(\mathbf{u}, \mathbf{v})$ so as to maximize $\mathbf{u}'\mathbf{S}_{xy}\mathbf{v}$ subject to $\mathbf{u}'\mathbf{S}_{xx}\mathbf{u} = \mathbf{v}'\mathbf{S}_{yy}\mathbf{v} = 1$.

The Lagrangian is

$$\mathbf{L}(\mathbf{u}, \mathbf{v}) = \mathbf{u}'\mathbf{S}_{xy}\mathbf{v} - \frac{\lambda}{2}(\mathbf{u}'\mathbf{S}_{xx}\mathbf{u} - 1) - \frac{\mu}{2}(\mathbf{v}'\mathbf{S}_{yy}\mathbf{v} - 1) \tag{7.9}$$

The maximum of the Lagrangian can be obtained by setting the derivatives relative to $\mathbf{u}$ and $\mathbf{v}$ equal to zero:

$$\frac{\partial \mathbf{L}}{\partial \mathbf{u}} = \mathbf{S}_{xy}\mathbf{v} - \lambda\mathbf{S}_{xx}\mathbf{u} = 0 \tag{7.10}$$

and

$$\frac{\partial \mathbf{L}}{\partial \mathbf{v}} = \mathbf{u}'\mathbf{S}_{xy} - \mu\mathbf{v}'\mathbf{S}_{yy} = 0 \tag{7.11}$$

From Eqs. (7.10) and (7.11), it follows that

$$\mathbf{u}'\mathbf{S}_{xy}\mathbf{v} = \lambda\mathbf{u}'\mathbf{S}_{xx}\mathbf{u} \tag{7.12}$$

and

$$\mathbf{u}'\mathbf{S}_{xy}\mathbf{v} = \mu\mathbf{v}'\mathbf{S}_{yy}\mathbf{v} \tag{7.13}$$

Consequently,

$$\lambda\mathbf{u}'\mathbf{S}_{xx}\mathbf{u} = \mu\mathbf{v}'\mathbf{S}_{yy}\mathbf{v} \tag{7.14}$$

However, because the transformed linear combination variables are standardized with unit variance, the result is

$$\lambda = \mu \tag{7.15}$$

Therefore, from Eq. (7.10), replacing λ by μ

$$\mathbf{S}_{xy}\mathbf{v} = \mu\mathbf{S}_{xx}\mathbf{u} \tag{7.16}$$

and from Eq. (7.11), by taking its transpose

$$\mathbf{S}_{yx}\mathbf{u} = \mu\mathbf{S}_{yy}\mathbf{v} \tag{7.17}$$

Solving for $\mathbf{v}$ in Eq. (7.17) leads to

$$\mathbf{v} = \frac{1}{\mu}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{u} \tag{7.18}$$

Replacing the value of $\mathbf{v}$ expressed in Eq. (7.18) into Eq. (7.16):

$$\mathbf{S}_{xy}\left(\frac{1}{\mu}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{u}\right) = \mu\mathbf{S}_{xx}\mathbf{u} \tag{7.19}$$

Or, multiplying each side of the equation by $\mu\mathbf{S}_{xx}^{-1}$:

$$\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{u} = \mu^2\mathbf{S}_{xx}^{-1}\mathbf{S}_{xx}\mathbf{u} \tag{7.20}$$

Equation (7.20) results in solving for the equation

$$\left(\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx} - \mu^2\mathbf{I}\right)\mathbf{u} = 0 \tag{7.21}$$

which is resolved by finding the eigenvalues and eigenvectors of $\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}$.

The eigenvalue gives the maximum squared correlation r_{zw}. This is the percentage of variance in w explained by z.

Two additional notions can be helpful in understanding the relationships between the set of $\mathbf{x}$ and the set of $\mathbf{y}$ variables: canonical loadings and redundancy analysis.

7.1.1 Canonical Loadings

The canonical loadings are defined as the correlations between the original $\mathbf{x}$ and $\mathbf{y}$ variables and their corresponding canonical variate $\mathbf{z}$ and $\mathbf{w}$. For the x variables

$$\underset{p\times 1}{\rho_{xz}} = \frac{1}{N-1}\underset{p\times N}{\mathbf{X}'}\underset{N\times 1}{\mathbf{z}} = \frac{1}{N-1}\mathbf{X}'(\mathbf{X}\mathbf{u}) = \frac{1}{N-1}\mathbf{S}_{xx}\mathbf{u} \tag{7.22}$$

Similarly, for the $\mathbf{y}$ variables

$$\underset{q\times 1}{\rho_{yw}} = \frac{1}{N-1}\underset{q\times N}{\mathbf{Y}'}\underset{N\times 1}{\mathbf{w}} = \frac{1}{N-1}\mathbf{Y}'(\mathbf{Y}\mathbf{v}) = \frac{1}{N-1}\mathbf{S}_{yy}\mathbf{v} \tag{7.23}$$

7.1.2 Canonical Redundancy Analysis

Canonical redundancy measures how well the original variables **y** can be predicted from the canonical variables. It reflects the correlation between the **z** and the **y** variables. Redundancy is the product of the percentage variance in **w** explained by **z** and the percentage variance in **y** explained by **w**. The first component is the squared correlation μ^2. The second component is the sum of squares of the canonical loadings for **y**.

Therefore,

$$\text{Redundancy} = \mu^2 \frac{\rho'_{\text{yw}} \rho_{\text{yw}}}{q} \tag{7.24}$$

7.2 Testing the Significance of the Canonical Correlations

It is possible to test the significance of these eigenvalues directly. However, the output in SAS shows eigenvalues that are different, albeit equivalent, from these eigenvalues or canonical correlation coefficients. These eigenvalues are related to the solution to the equation

$$\left(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}\right)\mathbf{u} \tag{7.25}$$

Such an equation corresponds to Wilk's lambda in MANOVA (see Chap. 2) and to discriminant analysis discussed in Chap. 7. However, canonical correlation analysis differs from these two contexts because here we do not have the notions of between- and within-group variances due to the nonexistence of groups. These notions are generalized, however, to the concepts of total variance and error variance. Therefore, Λ is redefined as

$$\Lambda = \left|\frac{\mathbf{E}}{\mathbf{T}}\right| \tag{7.26}$$

where **T** is the total variance–covariance matrix and **E** is the residual variance–covariance matrix after removing the effects of each pair of canonical variable correlations. However, it should be noted here that the solution to Eq. (7.25) or (7.26) can be expressed as a function of the eigenvalues of Eq. (7.21):

$$\lambda_i = \frac{{\mu_i}^2}{1 - {\mu_i}^2} \tag{7.27}$$

where the $\mu_i{}^2$s are the solution to Eq. (7.21) and λ_i is the solution to

$$\left(\mathbf{E}^{-1}\mathbf{H} - \lambda\mathbf{I}\right)\mathbf{u} = 0 \tag{7.28}$$

From the generalized definition of Wilk's lambda $\Lambda = \left|\frac{\mathbf{E}}{\mathbf{T}}\right|$, it follows that

$$\frac{1}{\Lambda} = \left|\frac{\mathbf{T}}{\mathbf{E}}\right| = \left|\mathbf{E}^{-1}\mathbf{T}\right| = \left|\mathbf{E}^{-1}(\mathbf{H}+\mathbf{E})\right| = \left|\mathbf{E}^{-1}\mathbf{H}+\mathbf{I}\right| = \prod_i (\lambda_i + 1) \tag{7.29}$$

where $\mathbf{T} = \mathbf{H} + \mathbf{E}$ because of their independence. When we replace the λ_is by the μ_is using the equality in Eq. (7.27), Λ can be expressed as a function of the μ_is, i.e., the canonical correlations:

$$\Lambda = \prod_i \frac{1}{\lambda_i + 1} = \prod_i \frac{1}{1 + \frac{\mu_i{}^2}{1 - \mu_i{}^2}} = \prod_i \left(1 - \mu_i{}^2\right) \tag{7.30}$$

Based on this expression of Λ, either as a function of the λ_is or as a function of the μ_is, it is possible to compute Bartlett's V or Rao's R, as discussed in Chap. 2. The degrees of freedom are not expressed in terms of the number of groups K, since this notion of group does not fit the canonical correlation model concerned with continuous variables. Instead, the equivalent is the parameter $(q - 1)$, the number of variates on the left side, which corresponds to the number of dummy variables that would be required to determine K groups.

Bartlett's V is

$$V = -[N - 1 - (p + q - 1)/2]Ln\ \Lambda = \left[N - \frac{3}{2} - (p+q)/2\right] \sum_{i=1}^{q} Ln(1 + \lambda_i) \tag{7.31}$$

or equivalently

$$V = -[N - 1 - (p + q - 1)/2]Ln\ \Lambda = \left[N - \frac{3}{2} - (p+q)/2\right] \sum_{i=1}^{q} Ln\left(1 - \mu_i^2\right) \tag{7.32}$$

Bartlett's V is approximately distributed as a chi-square with pq degrees of freedom. Alternatively, Rao's R can be computed as shown in Chap. 2 for MANOVA, where K is replaced by $q - 1$:

$$R = \frac{1 - \Lambda^{\frac{1}{t}}}{\Lambda^{\frac{1}{t}}} \frac{wt - \frac{pq}{2} + 1}{pq} \tag{7.33}$$

where $w = N - \frac{3}{2} - \frac{p+q}{2}$ and $t = \sqrt{\frac{p^2q^2 - 4}{p^2 + q^2 - 5}}$.

R is distributed approximately as an F distribution with pq degrees of freedom in the numerator and $wt - \frac{pq}{2} + 1$ degrees of freedom in the denominator. This last test (Rao's R) is the one reported in the SAS output (rather than Bartlett's V).

These tests are joint tests of the significance of the q canonical correlations. However, each term in the sum containing the eigenvalues in Eq. (7.31) or (7.32) is distributed approximately as a chi-square with $p + q - (2i - 1)$ degrees of freedom where i is the ith eigenvalue from $i = 1$ to q.

Any subset of eigenvalues is the sum of that subset of terms in $Ln\ (1 - \mu_i^2)$. Consequently, one can test if the residual canonical correlations are significant, after having removed the first canonical variate, then the first two, and so on. For example, the joint test of all q canonical correlations is V as in Eq. (7.32) with pq degrees of freedom. The test of the first eigenvalue is

$$V_1 = \left[N - \frac{3}{2} - (p+q)/2\right] Ln\left(1 - \mu_1^2\right) \tag{7.34}$$

with $(p + q - 1)$ degrees of freedom.

Consequently, the joint test that the remaining canonical correlations $\mu_2, \mu_3, \mu_4, \ldots \mu_q$ are zero is obtained by subtracting V_1 from V. $V - V_1$ is approximately chi-square distributed and the number of degrees of freedom is the difference between the degrees of freedom of V and those of V_1, i.e., $pq - (p + q - 1)$. This can be continued until the last qth eigenvalue. The same computations as those detailed above with Bartlett's V can be performed with Rao's R.

7.3 Multiple Regression as a Special Case of Canonical Correlation Analysis

In the case of multiple regression analysis, the dependent variable is a single variate represented by the vector $\underset{N\times 1}{\mathbf{y}}$ for the N observations. Consequently, the vector $\mathbf{v}$ reduces to a single scalar, set to the value 1. It follows that $\mathbf{w} = \mathbf{y}$. The expression for the correlation between $\mathbf{x}$ and $\mathbf{w}$ in Eq. (7.7) becomes

$$r_{zw} = \frac{\mathbf{u}'\mathbf{X}'\mathbf{y}}{\sqrt{(\mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u})(\mathbf{y}'\mathbf{y})}} \tag{7.35}$$

However, because the transformed independent variables are standardized and the single dependent variable y can be standardized to unit variance without loss of generality, the problem is to maximize the correlation coefficient r_{zw} subject to the constraint $\mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u} = 1$. This is solved by maximizing the Lagrangian:

$$\mathbf{L} = \mathbf{u}'\mathbf{X}'\mathbf{y} - \frac{\lambda}{2}(\mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u} - 1) \tag{7.36}$$

$$\frac{\partial \mathbf{L}}{\partial \mathbf{u}} = \mathbf{X}'\mathbf{y} - \lambda \mathbf{X}'\mathbf{X}\mathbf{u} = 0 \tag{7.37}$$

Solving for $\mathbf{u}$ leads to the least square estimator presented in Chap. 4:

$$\mathbf{u} = \frac{1}{\lambda}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{7.38}$$

7.4 Examples

Figure 7.2 shows the SAS commands to run a canonical correlation analysis. The data concern a number of new products that are characterized by a number of innovation characteristics, all rated on 7-point Likert scales from 1 (disagree) to 7 (agree):

X1: This new product is hard to understand.
X2: This new product is not really easy to use.
X3: Using this new product is not very compatible with the way I do things.
. . .
X13: I feel positive about this new product.
X14: I really like this new product.
X15: I am favorably disposed towards this new product.

The SAS procedure "proc cancorr" runs the canonical correlation analysis. The X variables (see Fig. 7.1) are indicated in the list following the key word "VAR" and the Y variables (see Fig. 7.1) are listed after the key word "with." Titles can be inserted for the output in single quotes after the word "title."

The input for STATA is shown in Fig. 7.3.

```
/* ************ Examp7_1.sas ************** */
OPTIONS LS=120;
DATA work;
INFILE "F:\WORK_STATA\WORK_SAS\SASMVS\CanonicalCorr\NewProdSurvey.csv" firstobs=2
dlm=',';
INPUT x1-x16;
proc cancorr;
 var x1-x3;
 with x13-x15;
 title 'Example of Canonical Correlation Analysis';
run;
```

Fig. 7.2 Example of SAS code for canonical correlation analysis (examp7-1.sas)

```
insheet using "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter7-
CCA/NewProdSurvey.csv", clear
canon (x1 x2 x3) (x13 x14 x15)
mat list e(canload11)
mat list e(canload22)
estat loadings
canon , test (1 2 3)
```

Fig. 7.3 Example of STATA code for canonical correlation analysis (examp7-1.do)

The procedure "canon" is used in STATA with the X and the Y variables listed in their own sets of parentheses. The matrices "canload11" and "canload22" correspond to the canonical loadings of the X and Y variables, respectively. These canonical loadings can also be displayed using the command "estat loadings." In addition to the canonical loadings, the correlations between the X variables and the W canonical variates, as well as the correlations between the Y variables and the Z canonical variates, are displayed. The last line of commands in Fig. 7.3 concerns the test of the significance of the individual canonical correlations. The command "canon" (without arguments) repeats the output of the prior canonical analysis requested and the "test" option is followed by the canonical correlation numbers for which testing is requested: test (1) tests for all three canonical correlations, test (2) tests for the significance of canonical correlations 2 and 3 jointly, and so on.

Figure 7.4 lists the SAS output from running the canonical correlation analysis.

When the canonical correlations are listed, we see that one correlation coefficient of 0.35131 appears larger than the other two values. Therefore, we can concentrate on this larger value. These correlations correspond to the eigenvalues that give a solution to Eq. (7.21) (the canonical correlation is the square root of these eigenvalues).

The eigenvalues λ_i, which are the solution to Eq. (7.28), are those shown under the column "Eigenvalue" in the SAS output. For example, the first (highest) eigenvalue of 0.1408 is related to the first canonical correlation as

$$0.1408 = \frac{(0.3513)^2}{\left[1 - (0.3513)^2\right]} \tag{7.39}$$

Given the relationship between the λ_is and the μ_is, these eigenvalues provide the same information as the canonical correlations. The F test corresponding to Rao's R (highlighted in grey in Fig. 7.4) indicates that the set of canonical correlations (or eigenvalues) are jointly significantly different from zero ($F = 6.21$ with 9 and 959.04 degrees of freedom). Then, the next row in that part of the table shows that after removing the first canonical correlation, the remaining canonical correlations are jointly statistically insignificant at the 0.05 level ($F = 0.61$ with 4 and 790 degrees of freedom). Therefore, we can concentrate on the results concerning the first canonical variable.

Example of Canonical Correlation Analysis

The CANCORR Procedure
Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.353338	0.338908	0.043884	0.123424	0.1408	0.1346	0.9577	0.9577	0.87115229	6.21	9	959.04	<.0001
2	0.078499	.	0.049754	0.006162	0.0062	0.0062	0.0422	0.9998	0.99381316	0.61	4	790	0.6528
3	0.004988	.	0.050061	0.000025	0.0000		0.0002	1.0000	0.99997512	0.01	1	396	0.9210

Multivariate Statistics and F Approximations

S=3 M=-0.5 N=196

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.87115229	6.21	9	959.04	<.0001
Pillai's Trace	0.12961147	5.96	9	1188	<.0001
Hotelling-Lawley Trace	0.14702822	6.42	9	615.97	<.0001
Roy's Greatest Root	0.14080301	18.59	3	396	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Raw Canonical Coefficients for the VAR Variables

	V1	V2	V3
x1	-0.874143575	-2.034342152	0.188108428
x2	-0.043496381	1.1867328824	1.3931995736
x3	-0.068875313	1.1846748684	-1.490537759

Raw Canonical Coefficients for the WITH Variables

	W1	W2	W3
x13	0.4985992263	-0.40536551	-3.164990656
x14	0.3439635378	2.385102473	1.9915713971
x15	0.3498829102	-2.098530917	1.4048067891

Standardized Canonical Coefficients for the VAR Variables

	V1	V2	V3
x1	-0.9008	-2.0965	0.1939
x2	-0.0434	1.1852	1.3915
x3	-0.0721	1.2395	-1.5595

Fig. 7.4 Example of SAS output of canonical correlation analysis (examp7-1.out)

Standardized Canonical Coefficients for the WITH Variables

	W1	W2	W3
x13	0.4463	-0.3628	-2.8330
x14	0.2998	2.0786	1.7356
x15	0.2923	-1.7532	1.1737

Correlations Between the VAR Variables and Their Canonical Variables

	V1	V2	V3
x1	-0.9989	-0.0470	0.0088
x2	-0.8468	0.3966	0.3544
x3	-0.8797	0.3481	-0.3239

Correlations Between the WITH Variables and Their Canonical Variables

	W1	W2	W3
x13	0.9789	0.0278	-0.2023
x14	0.9629	0.2414	0.1208
x15	0.9390	-0.2899	0.1851

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables

	W1	W2	W3
x1	-0.3509	-0.0037	0.0000
x2	-0.2975	0.0311	0.0018
x3	-0.3091	0.0273	-0.0016

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables

	V1	V2	V3
x13	0.3439	0.0022	-0.0010
x14	0.3383	0.0189	0.0006
x15	0.3299	-0.0228	0.0009

Fig. 7.4 (continued)

The raw (highlighted in grey in Fig. 7.4) and the standardized eigenvectors are then listed in the SAS output. The raw values are subject to variations due to the scale units of each variate and should be interpreted accordingly. It should be noted that the canonical variables are normalized to unit variance as per Eq. (7.8), and consequently, the magnitude of the coefficients that are the elements of the eigenvectors **u** and **v** are affected as well by the unit of the variates. The first eigenvector indicates that innovations that are not complex and that are easy to

```
. insheet using "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter7-
CCA/NewProdSurvey.csv", clear
(16 vars, 400 obs)

. canon (x1 x2 x3) (x13 x14 x15)

Canonical correlation analysis                        Number of obs =     400

Raw coefficients for the first variable set

                  |        1         2         3
    --------------+------------------------------
              x1  |   -0.8741   -2.0343   -0.1881
              x2  |   -0.0435    1.1867   -1.3932
              x3  |   -0.0689    1.1847    1.4905
    ------------------------------------------------

Raw coefficients for the second variable set

                  |        1         2         3
    --------------+------------------------------
             x13  |    0.4986   -0.4054    3.1650
             x14  |    0.3440    2.3851   -1.9916
             x15  |    0.3499   -2.0985   -1.4048
    ------------------------------------------------

-------------------------------------------------------------------------------
Canonical correlations:
  0.3513  0.0785  0.0050

-------------------------------------------------------------------------------
Tests of significance of all canonical correlations

                         Statistic      df1      df2            F     Prob>F
         Wilks' lambda     .871152        9  959.043       6.2140     0.0000 a
        Pillai's trace     .129611        9     1188       5.9604     0.0000 a
Lawley-Hotelling trace     .147028        9     1178       6.4148     0.0000 a
    Roy's largest root     .140803        3      396      18.5860     0.0000 u
-------------------------------------------------------------------------------
                          e = exact, a = approximate, u = upper bound on F

. mat list e(canload11)

e(canload11)[3,3]
             1           2           3
x1  -.99885734  -.04696972  -.00882417
x2  -.84681442   .39663864  -.35437709
x3  -.87972518   .34805833   .32394291

. mat list e(canload22)

e(canload22)[3,3]
             1           2           3
x13   .97892406   .02777063   .20232766
x14   .96288434   .24138724  -.12077234
x15   .93898312  -.28993824  -.18505812

. estat loadings

Canonical loadings for variable list 1

                  |        1         2         3
    --------------+------------------------------
              x1  |   -0.9989   -0.0470   -0.0088
              x2  |   -0.8468    0.3966   -0.3544
              x3  |   -0.8797    0.3481    0.3239
    ------------------------------------------------

Canonical loadings for variable list 2

                  |        1         2         3
    --------------+------------------------------
             x13  |    0.9789    0.0278    0.2023
             x14  |    0.9629    0.2414   -0.1208
             x15  |    0.9390   -0.2899   -0.1851
```

Fig. 7.5 Example of STATA output of canonical correlation analysis (examp7-1.log)

```
-------------+------------------------------
         x13 |   0.9789    0.0278    0.2023
         x14 |   0.9629    0.2414   -0.1208
         x15 |   0.9390   -0.2899   -0.1851
-----------------------------------------------

Correlation between variable list 1 and canonical variates from list 2

             |        1         2         3
-------------+------------------------------
          x1 |  -0.3509   -0.0037   -0.0000
          x2 |  -0.2975    0.0311   -0.0018
          x3 |  -0.3091    0.0273    0.0016
-----------------------------------------------

Correlation between variable list 2 and canonical variates from list 1

             |        1         2         3
-------------+------------------------------
         x13 |   0.3439    0.0022    0.0010
         x14 |   0.3383    0.0189   -0.0006
         x15 |   0.3299   -0.0228   -0.0009
-----------------------------------------------


. canon , test (1 2 3)

…

--------------------------------------------------------------------------------
Canonical correlations:
  0.3513  0.0785  0.0050

--------------------------------------------------------------------------------
Tests of significance of all canonical correlations

                       Statistic      df1      df2            F     Prob>F
        Wilks' lambda     .871152        9  959.043       6.2140     0.0000 a
       Pillai's trace     .129611        9     1188       5.9604     0.0000 a
Lawley-Hotelling trace    .147028        9     1178       6.4148     0.0000 a
    Roy's largest root    .140803        3      396      18.5860     0.0000 u
--------------------------------------------------------------------------------
Test of significance of canonical correlations 1-3

                       Statistic      df1      df2            F     Prob>F
        Wilks' lambda     .871152        9  959.043       6.2140     0.0000 a
--------------------------------------------------------------------------------
Test of significance of canonical correlations 2-3

                       Statistic      df1      df2            F     Prob>F
        Wilks' lambda     .993813        4      790       0.6138     0.6528 e
--------------------------------------------------------------------------------
Test of significance of canonical correlation 3

                       Statistic      df1      df2            F     Prob>F
        Wilks' lambda     .999975        1      396       0.0099     0.9210 e
--------------------------------------------------------------------------------
                          e = exact, a = approximate, u = upper bound on F
```

Fig. 7.5 (continued)

understand (x1, x2, and x3) are associated with greater positive responses (x13, x14, and x15).

Then, the correlation of each variate to the canonical variables (composite variable v and then w) is contained in the last tables of Fig. 7.4. This allows us to assess the strength of the relationships that form a composite (unobserved) canonical variable and of the relationship of a variable to the other composite canonical variable.

The STATA output is shown in Fig. 7.5.

In the last section of the STATA output, the heading "Test of significance of canonical correlation 1–3" corresponds to the joint test of all the canonical correlations shown at the top of the output. The "Test of significance of canonical correlation 2–3" is the joint test of canonical correlations 2 through 3. Given that it is insignificant (F = 0.6138), we conclude that only the first canonical correlation is significant.

7.5 Assignment

Using the survey data described for the assignment in Chap. 3, associate certain types of consumer behaviors to their psychographic profiles. The sample SAS code file to read the data is shown in Fig. 3.16.

Bibliography

Application Readings

Gomez, L. F. (2009). Time to socialize: Organizational socialization structures and temporality. *Journal of Business Communication, 46*(2), 179–207.

Hosamane, M. D., & Alroaia, Y. V. (2009). Entrepreneurship and development of small-scale industries in Iran: Strategic management tools and business performance assessment. *The Icfai University Journal of Entrepreneurship Development, 6*(1), 27–40.

Hultink, E. J., Griffin, A., Robben, H. S. J., & Hart, S. (1998). In search of generic launch strategies for new products. *International Journal of Research in Marketing, 15*(3), 269–285.

Voss, M. D., Calantone, R. J., & Keller, S. B. (2005). Internal service quality: Determinants of distribution center performance. *International Journal of Physical Distribution & Logistics Management, 35*(3), 161–176.

Chapter 8
Categorical Dependent Variables

In this chapter, we consider statistical models to analyze variables where the numbering does not have any meaning and, in particular, where there is no relationship between one level of the variable and another level. In these cases, we are typically trying to establish whether it is possible to explain with other variables the level observed of the criterion variable. The chapter is divided into two parts. The first part presents discriminant analysis, which is a traditional method in multivariate statistical analysis. The second part introduces quantal choice statistical models. The models are described, as well as their estimation. Their measures of fit are also discussed.

8.1 Discriminant Analysis

If there is only one variable, the test (i.e., a measure) of the extent of differences across groups is the ratio of the sum of squares between groups to the sum of squares within groups corrected by the degrees of freedom of the numerator and the denominator:

$$\frac{SS_b(x)/(K-1)}{SS_w(x)/(N-K)} \tag{8.1}$$

where N is the sample size and K is the number of groups. This is simply the F test for the significance of differences across groups for one variable.

In presenting discriminant analysis, the discriminant criterion, which is the basis for understanding the methodology, is first introduced. Then the derivation and the explanation of the discriminant functions are provided. Finally issues of classification and measures of fit are discussed.

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_8, © Springer Science+Business Media New York 2014

8.1.1 The Discriminant Criterion

The objective in discriminant analysis is to determine a linear combination of a set of variables such that several group means (each group corresponding to a level of the dependent variable) will differ widely on this linear combination.

Let p = number of independent variables, N = number of observations, N_j = number of observations for group $j = 1 \ldots K$, and K = number of groups. Then, $\underset{i1\times p}{\mathbf{x}'}$ is the vector representing the values on p variables for one observation i, and $\underset{p\times 1}{\mathbf{v}}$ is the vector of weights to be attributed to each of the p variables to form a linear combination. The linear combination is given by Eq. (8.2):

$$\underset{1\times 1}{y_i} = \underset{1\times p}{\mathbf{x}'_i}\ \underset{p\times 1}{\mathbf{v}} = v_1 x_{i1} + v_2 x_{i2} + \ldots v_p x_{ip} \tag{8.2}$$

We will assume that $\mathbf{x}_i$ follows a multivariate normal distribution. It follows that each y_i is normally distributed.

The problem consists in finding $\mathbf{v}$ that is going to maximize the F-ratio for testing the significance of the overall difference among several group means on a *single variable* y.

This value F is given by the ratio of the between-group variance to the pooled within-group variance of the variable y:

$$F = \frac{SS_b(y)/(K-1)}{SS_w(y)/(N-K)} \tag{8.3}$$

where N = number of observations or individuals, K = number of groups, $SS_b(y)$ = between-group sum of squares, and $SS_w(y)$ = pooled within-group sum of squares.

In the case where there are only two groups ($K = 2$), it is the classic t test of a difference of two means. The problem, therefore, is to find the value of $\mathbf{v}$ that will maximize F.

The ratio $(K-1)/(N-K)$ is a constant; therefore,

$$\underset{\mathbf{v}}{\text{Max}}\ F \Leftrightarrow \underset{\mathbf{v}}{\text{Max}}\ \frac{SS_b(y)}{SS_w(y)} = \lambda$$

The pooled within-group sum of squares is the sum over the groups (j) of the squares of the deviations of variable y from their group mean:

$$SS_w(y) = \sum_{j=1}^{K} SS_j(y) \tag{8.4}$$

Let

$$\underset{N_j\times p}{\overline{\mathbf{X}}_j} = \left\{\overline{\mathbf{x}}'_j\right\} \tag{8.5}$$

where the mean vector for group j ($\overline{\mathbf{x}}'_j$) is repeated N_j times (i.e., N_j rows).

For each group j (where $j = 1, \ldots K$), we can write the vector of the values obtained from the linear combination of the variables. This vector has N_j elements corresponding to the number of observations in group j.

Let

$$\underset{N_j \times p}{\mathbf{X}_j^d} = \underset{N_j \times p}{\mathbf{X}_j} - \underset{N_j \times p}{\overline{\mathbf{X}}_j} \tag{8.6}$$

and

$$\forall j : \underset{N_j \times 1}{\mathbf{y}_j^d} = \underset{N_j \times p}{\left(\mathbf{X}_j - \overline{\mathbf{X}}_j\right)} \underset{p \times 1}{\mathbf{v}} = \mathbf{X}_j^d \mathbf{v} \tag{8.7}$$

Then,

$$SS_j(\mathbf{y}) = \mathbf{y}_j^{d'} \mathbf{y}_j^d = \mathbf{v}' \underset{p \times p}{\mathbf{X}_j^{d'} \mathbf{X}_j^d} \mathbf{v} = \mathbf{v}' \mathbf{S}_j \mathbf{v} \tag{8.8}$$

where $\mathbf{S}_j = \mathbf{X}_j^{d\ \prime}\ \mathbf{X}_j^d$. Therefore,

$$SS_w(\mathbf{y}) = \sum_{j=1}^{K} \mathbf{v}' \mathbf{S}_j \mathbf{v} = \mathbf{v}' \left(\sum_{j=1}^{K} \mathbf{S}_j \right) \mathbf{v} \tag{8.9}$$

Let

$$\mathbf{W} = \sum_{j=1}^{K} \mathbf{S}_j \tag{8.10}$$

Then,

$$SS_w(y) = \mathbf{v}' \mathbf{W} \mathbf{v} \tag{8.11}$$

Let

$$\underset{N \times p}{\overline{\mathbf{X}}} = \begin{bmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ \vdots \\ \overline{\mathbf{X}}_K \end{bmatrix}$$

$\underset{N \times p}{\overline{\overline{\mathbf{X}}}}$ = matrix composed of the vector of grand means (across all groups) repeated N times:

$$\mathbf{B} = \left(\overline{\mathbf{X}} - \overline{\overline{\mathbf{X}}} \right)' \left(\overline{\mathbf{X}} - \overline{\overline{\mathbf{X}}} \right) \tag{8.12}$$

Therefore,

$$SS_b(y) = \mathbf{v}'\mathbf{B}\mathbf{v} \tag{8.13}$$

and consequently,

$$\lambda = \frac{\mathbf{v}'\mathbf{B}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}} \tag{8.14}$$

We can maximize λ (the discriminant criterion) by taking the first derivative relative to $\mathbf{v}$ and setting it equal to 0 (we use the matrix derivation rule A.2 in Appendix A: $\partial\ \mathbf{v}'\mathbf{A}\mathbf{v}/\partial\ \mathbf{v} = 2\mathbf{A}\mathbf{v}$):

$$\frac{\partial \lambda}{\partial \underset{p\times 1}{\mathbf{v}}} = \frac{\left(\underset{1\times 1}{\mathbf{v}'\mathbf{W}\mathbf{v}}\right)\left(2\underset{p\times p}{\mathbf{B}}\ \underset{p\times 1}{\mathbf{v}}\right) - \left(\underset{1\times 1}{\mathbf{v}'\mathbf{B}\mathbf{v}}\right)\left(2\underset{p\times p}{\mathbf{W}}\ \underset{p\times 1}{\mathbf{v}}\right)}{\left(\underset{1\times 1}{\mathbf{v}'\mathbf{W}\mathbf{v}}\right)^2} = \mathbf{0} \tag{8.15}$$

From Eq. (8.14)

$$\mathbf{v}'\mathbf{B}\mathbf{v} = \lambda \mathbf{v}'\mathbf{W}\mathbf{v} \tag{8.16}$$

By substitution in Eq. (8.15)

$$\frac{\partial \lambda}{\partial \underset{p\times 1}{\mathbf{v}}} = \frac{(\mathbf{v}'\mathbf{W}\mathbf{v})(2\mathbf{B}\mathbf{v}) - \lambda(\mathbf{v}'\mathbf{W}\mathbf{v})(2\mathbf{W}\mathbf{v})}{(\mathbf{v}'\mathbf{W}\mathbf{v})^2} \tag{8.17}$$

and consequently,

$$\frac{\partial \lambda}{\partial \underset{p\times 1}{\mathbf{v}}} = 2\left[\frac{\mathbf{B}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}} - \frac{\lambda \mathbf{W}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}}\right] = \mathbf{0} \tag{8.18}$$

$$\frac{\mathbf{B}\mathbf{v} - \lambda \mathbf{W}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}} = \mathbf{0} \tag{8.19}$$

Equation (8.19) is true if

$$\mathbf{B}\mathbf{v} - \lambda \mathbf{W}\mathbf{v} = \mathbf{0} \tag{8.20}$$

or

$$(\mathbf{B} - \lambda \mathbf{W})\mathbf{v} = \mathbf{0} \tag{8.21}$$

which by premultiplying by $\mathbf{W}^{-1}$ gives

$$(\mathbf{W}^{-1}\mathbf{B} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0} \tag{8.22}$$

Therefore, the solution for λ is given by the eigenvalues of $\mathbf{W}^{-1}\ \mathbf{B}$, and the solution for $\mathbf{v}$ is given by the corresponding eigenvectors of $\mathbf{W}^{-1}\ \mathbf{B}$.

8.1.2 Discriminant Function

The matrix $\mathbf{W}^{-1}\,\mathbf{B}$ is not symmetric. In fact, there are $K-1$ linearly independent rows in $\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}$.

Consequently, the rank of $\mathbf{B}$ is $K-1$. $\mathbf{W}^{-1}$ is of full rank (p); if it were singular, it could not be inverted.

Therefore, the number of nonzero eigenvalues is the smaller of the rank of $\mathbf{W}^{-1}$ and of $\mathbf{B}$, which is usually $K-1$ (following from the fact that typically there are more variables than groups, i.e., $K-1<p$).

This means that discriminant analysis provides $K-1$ nonzero eigenvalues and $K-1$ discriminant functions.

The first discriminant function $\mathbf{v}_1$ has the largest discriminant criterion value λ_1 (eigenvalue), and each of the others has a *conditionally* maximal discriminant criterion value.

The centroids for each group j consist of the mean value of $\mathbf{y}$ for the group for each of the $K-1$ eigenvectors or discriminating functions:

$$\overline{y}_{1j},\overline{y}_{2j},\cdots\overline{y}_{rj},\cdots\overline{y}_{K-1,j} \tag{8.23}$$

where r represents the index for the rth eigenvalue and eigenvector:

$$\overline{y}_{rj}=\overline{\mathbf{x}}'_j\mathbf{v}_r \tag{8.24}$$

These are the dimensions along which one can find the largest differences across groups.

8.1.2.1 Special Case of K = 2

It is possible to estimate a multiple regression equation where the dependent variable is a dummy variable (0 for alternative 1 and 1 for the other alternative). Such a regression would yield weights for the independent variables that would be proportional to the discriminant weights. However, it is important to note that the t statistics should not be used. Indeed, the errors are not normally distributed with mean 0 and variance $\sigma^2\mathbf{I}$, as will be demonstrated in the sections below.

8.1.2.2 Testing the Significance of the Discriminant Solutions

Recalling that Wilk's lambda is the statistic we discussed when testing the significance of differences of means for multiple variates (MANOVA), we consider this statistic in the context of discriminant analysis. As indicated in Chap. 2,

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \tag{8.25}$$

Consequently, using rule (A.8) in Appendix A

$$\frac{1}{\Lambda} = \frac{|\mathbf{T}|}{|\mathbf{W}|} = \left|\mathbf{W}^{-1}\mathbf{T}\right| = \left|\mathbf{W}^{-1}(\mathbf{W}+\mathbf{B})\right| = \left|\left(\mathbf{I}+\mathbf{W}^{-1}\mathbf{B}\right)\right| \tag{8.26}$$

However, according to this rule, the inverse of Wilk's lambda can be expressed in terms of the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$:

$$\frac{1}{\Lambda} = \prod_{i=1}^{K-1}(1+\lambda_i) \tag{8.27}$$

Consequently,

$$\Lambda = \frac{1}{\prod_{i=1}^{K-1}(1+\lambda_i)} = \prod_{i=1}^{K-1}\frac{1}{(1+\lambda_i)} \tag{8.28}$$

The statistic used for MANOVA, Bartlett's V, can then be expressed in terms of the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$:

$$V = -[N-1-(p+K)/2]Ln\,\Lambda = [N-1-(p+K)/2]\sum_{i=1}^{K-1}Ln(1+\lambda_i) \tag{8.29}$$

Bartlett's V is distributed approximately as a chi-square with $p(K-1)$ degrees of freedom and, because each of the discriminant functions is uncorrelated, each element of the terms of the sum in Eq. (8.29) corresponding to the r's eigenvalue is distributed as a chi-square with degrees of freedom $p + K - 2r$. Let

$$V_r = [N-1-(p+K)/2]Ln(1+\lambda_r) \tag{8.30}$$

It is then feasible to test the significance of the residual discrimination after removing the first discriminant function by comparing the value of $V - V_1$. If this difference is still significant, it means that the remaining discriminant functions still have a discriminant power. The process continues by comparing $V - (V_1 + V_2)$ and then more generally $V - \left(\sum_{i=1}^{r} V_r\right)$ until this expression becomes insignificant.

8.1.3 Classification and Fit

8.1.3.1 Classification

The issue we need to address now concerns how to classify the observations.

A group prediction can be made, based on the value of the linear combination obtained from the first discriminant function:

$$\hat{y}_{1i} = \mathbf{x}'_i \hat{\mathbf{v}}_1 \tag{8.31}$$

The group prediction then depends on the value obtained in Eq. (8.31), relative to a critical value y_{1crit}, i.e., based on the sign of

$$\hat{y}_{1i} - \hat{y}_{1\text{crit}} \tag{8.32}$$

The rule can then be based on the distance from group means: assign observation i to the group to which it is closest (corrected for covariance). The midpoints are then used as the critical values.

For example, in the two-group case, there is a single eigenvector:

$$\mathbf{v} = \mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \tag{8.33}$$

$$y = \mathbf{x}'\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \tag{8.34}$$

$$\text{Group 1} : \overline{y}_1 = \overline{\mathbf{x}}'_1\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \tag{8.35}$$

$$\text{Group 2} : \overline{y}_2 = \overline{\mathbf{x}}'_2\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \tag{8.36}$$

The classification is based on the midpoint:

$$y_{\text{crit}} = \frac{1}{2}(\overline{y}_1 + \overline{y}_2) \quad \Rightarrow y_{\text{crit}} = \frac{1}{2}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)'\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \tag{8.37}$$

Then the classification rule is

$$\text{if } y_{1i} < y_{\text{crit}} \Rightarrow i \in \text{Group 1 else } i \in \text{Group 2},$$

which is equivalent to defining w as

$$w = y_i - y_{\text{crit}}$$

Then, if $w < 0$ then $i \in$ Group 1 or else $i \in$ Group 2. This is represented graphically in Fig. 8.1, where the vertical line represents the critical value appearing at the midpoint between the mean of each of the two groups $\overline{y}_1$ and $\overline{y}_2$.

As discussed above

$$y_i < y_{\text{crit}} \Rightarrow i \in \text{Group 1}$$

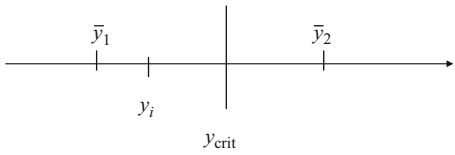


Fig. 8.1 Classification of observations

or equivalently

$$\mathrm{w} = y_i - y_{\text{crit}} < 0 \Rightarrow i \in \text{Group 1}$$

For more than two groups (i.e., $K > 2$), similar concepts apply.
Let

$$w_{jk}(i) = \underbrace{\mathbf{x}'_i \mathbf{W}^{-1}\left(\overline{\mathbf{x}}_j - \overline{\mathbf{x}}_k\right)}_{y_i} \underbrace{-\frac{1}{2}\left(\overline{\mathbf{x}}_j + \overline{\mathbf{x}}_k\right)' \mathbf{W}^{-1}\left(\overline{\mathbf{x}}_j - \overline{\mathbf{x}}_k\right)}_{y_{\text{crit}}} \tag{8.38}$$

The rule consists of assigning i to group j if $w_{jk}i) > 0$ for all $k \neq j$, which means that y_i is closer to k than to j.

For example, for three groups ($K = 3$), we can compute w_{12}, w_{13}, and w_{23} (note that $w_{21} = -w_{12}$). But, because $w_{23} = w_{13} - w_{12}$, we do not need w_{23}.

Then we can classify i as belonging to

Group 1: if $w_{12} > 0$ and $w_{13} > 0$
Group 2: if $w_{12} < 0$ and $w_{13} > w_{12}$
Group 3: if $w_{13} < 0$ and $w_{12} > w_{13}$

For more than two groups, a plot of the centroids $\overline{y}_j$ on the discriminant functions as axes can help to interpret them.

8.1.3.2 Measures of Fit

Fit measures are based on the ability of the discriminant functions to classify observations correctly. This information is contained in the classification table, as shown in Fig. 8.2.

Percent Correctly Classified

The classification table is a $K \times K$ matrix that indicates the number or the percentage of observations in each group that have been classified into that same group (therefore the correctly classified observations) or into another group (the incorrectly classified observations).

The diagonal cells in Fig. 8.2 represent the observations that are correctly classified. The percentage of correctly classified observations can easily be computed as

Fig. 8.2 Classification table

	Predicted					
Actual		1	2		K	
	1	n_{11}				
	2		n_{22}			
	K				n_{KK}	
						N

$$n_c = \frac{\left(\sum_j n_{jj}\right)}{N} \tag{8.39}$$

where n_{jj} = number of observations actually in category j and predicted to be in category j, and N = total number of observations.

This measure of fit presents two problems:

- It uses the same N individuals for discrimination and prediction. This leads to an upward bias in the probability of classifying the observations correctly. A solution is to use a split sample for prediction.
- If the sample is not distributed evenly across the groups (i.e., the observed proportions are different across groups), then by merely classifying all observations arbitrarily into the group with the highest proportion, one can get at least max $\{p_j\}$ classified correctly, where $\mathrm{p_j}$ is the actual proportion of observations in Group j.

Maximum Chance Criterion

This last value, i.e., max $\{p_j\}$, is defined as the maximum chance criterion. Because no model is required in order for us to arrive at such a rate of correct assignment to groups, we can use the maximum chance criterion as a minimum standard, and any model should be able to improve on this rate.

Percent Correctly Classified by Chance: The Proportional Chance Criterion

Assume two groups:

$$P(\text{correct}) = P(\text{correct}|\text{j} = 1).\, P(\text{j} = 1) + P(\text{correct}|\text{j} = 2).\, P(\text{j} = 2)$$

Let p_j be the observed proportion of observations actually in group j, as defined earlier, and α_j the proportion of observations classified in group j:

$$P\left(\begin{array}{c}\text{correct}\\ \text{by chance}\end{array}\right) = \sum_j p_j \alpha_j \tag{8.40}$$

Let us assume that the discriminant function is meaningful. Then we want to classify in the same proportion as the actual groups.

Under our decision rule, $\alpha_j = p_j$; therefore,

$$P\left(\begin{array}{c}\text{correct}\\ \text{by chance}\end{array}\right) = \sum_j p_j \alpha_j = \sum_j p_j^2 \tag{8.41}$$

Equation (8.41) provides the formula for the proportional chance criterion.

Tau Statistic

The tau statistic involves the same rationale but standardizes the information:

$$\tau = \frac{n_c - \Sigma_j p_j n_j}{N - \Sigma_j p_j n_j} = \frac{(n_c/N) - \Sigma_j p_j \alpha_j}{1 - \Sigma_j p_j \alpha_j} \tag{8.42}$$

where n_j = number of observations classified in group j, and n_c = number of correctly classified observations.

8.2 Quantal Choice Models

In this section, we introduce logit models of choice. Although we could also discuss probit models here, we do not because they follow the same rationale as for the logit model. We start by discussing the difficulties inherent in using the standard regression model with a categorical dependent variable, even a binomial one. Then we discuss methodologies that can be used to resolve some of those problems. We then present the logit model with two variants and explain the estimation of the logit model parameters. Finally, we present the various measures of fit.

8.2.1 The Difficulties of the Standard Regression Model with Categorical Dependent Variables

Let us assume the case of two groups. The variable representing the group assignment can take two values, 0 and 1:

$$y_i = \begin{cases} 0 \\ 1 \end{cases} \tag{8.43}$$

This group assignment is made on the basis of a linear model:

$$\forall i = 1, \ldots, N: \quad \underset{1\times 1}{y_i} = \underset{1\times p}{\mathbf{x}'_i} \; \underset{p\times 1}{\boldsymbol{\beta}} + \underset{1\times 1}{e_i} \tag{8.44}$$

Are the usual assumptions verified?

1. Is E $[e_i] = 0$?
   This would imply in this case that the error terms for each observation follow a specific random process. Indeed, from Eq. (8.44) it follows that

$$e_i = y_i - \mathbf{x}'_i\boldsymbol{\beta} \tag{8.45}$$

   Consequently, the following distribution for y_i would be required for the expectation of the error term to be zero, i.e., for E $[e_i] = 0$:

$$P(y_i = 0) = 1 - \mathbf{x}'_i\boldsymbol{\beta} \tag{8.46}$$

$$P(y_i = 1) = \mathbf{x}'_i \tag{8.47}$$

   However, this is not generally the case, in part because

$$\mathbf{x}'_i\boldsymbol{\beta} \notin [0, 1]$$

   Therefore, the distribution is impossible. Hence, $\hat{\beta}_{OLS}$ is biased.
2. Is $E[e_i^2] = \sigma^2$?
   The second assumption is the homoscedasticity of the error terms.
   e_i is distributed as a Bernoulli process:

$$\mathrm{V}[e_i] = \left(\mathbf{x}'_i\boldsymbol{\beta}\right)\left(1 - \mathbf{x}'_i\boldsymbol{\beta}\right) \tag{8.48}$$

   This implies that heteroscedasticity and consequently ordinary least squares are inefficient.
3. The range constraint problem: $\hat{y}_i \notin [0, 1]$.
   A third problem occurs because the predicted values of the predicted variable can be outside the range of the theoretical values, which are either 0 or 1.

8.2.2 Transformational Logit

8.2.2.1 Resolving the Efficiency Problem

We may be able to resolve the efficiency problem with the estimated generalized least squares estimator.

Let us assume that the data can be grouped into K groups:

$$j = 1, \ldots, K$$

$$n_j = \text{size of group } j$$

where the K groups correspond to "settings" of independent variables.

Let

$$z_j = \sum_{i|j} y_{ij} \tag{8.49}$$

where

$$y_{ij} \begin{cases} 0 \\ 1 \end{cases}$$

z_j is the number of 1s in group j:

$$p_j = \frac{z_j}{n_j} \tag{8.50}$$

The model for a given group is

$$p_j = \mathbf{X}_j \boldsymbol{\beta} + e_j \tag{8.51}$$

For the entire K groups, the proportions are represented by

$$\underset{K\times 1}{\mathbf{p}} = \underset{K\times p}{\mathbf{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{K\times 1}{\mathbf{e}} \tag{8.52}$$

In Eq. (8.51), the true proportion for group j is given by

$$P_j = \mathbf{X}_j \boldsymbol{\beta} \tag{8.53}$$

Therefore,

$$p_j = P_j + e_j \tag{8.54}$$

e_j follows a binomial distribution:

$$e_j \sim B\left(0, P_j\left(1 - P_j\right)/n_j\right) \tag{8.55}$$

The variance is obtained because z_j is such that

$$\mathrm{E}\left[z_j\right] = n_j P_j \tag{8.56}$$

$$\mathrm{V}[z_j] = n_j P_j(1 - P_j) \tag{8.57}$$

Therefore, dividing by n_j

$$\mathrm{E}[p_j] = \mathrm{E}\left[\frac{z_j}{n_j}\right] = P_j \tag{8.58}$$

$$\mathrm{V}[p_j] = \mathrm{V}\left[\frac{z_j}{n_j}\right] = \frac{1}{n_j^2}\mathrm{V}[z_j] = \frac{P_j(1 - P_j)}{n_j} \tag{8.59}$$

Consequently, the covariance of the error term in Eq. (8.52) is

$$\mathrm{E}\left[\mathbf{e}\mathbf{e}'\right] = \mathbf{\Phi} = diag\{P_j(1 - P_j)/n_j\} \tag{8.60}$$

The generalized least squares estimator would be

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Phi}^{-1}\mathbf{p} \tag{8.61}$$

But $\mathbf{\Phi}$ is unknown. It can be replaced by a consistent estimator to obtain the estimated generalized least squares estimator. Such an estimator of $\mathbf{\Phi}$ is

$$\hat{\mathbf{\Phi}} = diag\{\hat{p}_j(1 - \hat{p}_j)/n_j\} \tag{8.62}$$

where

$$\hat{\mathbf{p}} = \mathbf{X}\mathbf{b} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{p} \tag{8.63}$$

The ordinary least squares estimator $\mathbf{b}$ provides estimates for $\mathbf{p}$ that are consistent with the theoretical model specification. The estimated generalized least squares estimator is

$$\hat{\boldsymbol{\beta}}_{\mathrm{EGLS}} = \left(\mathbf{X}'\hat{\mathbf{\Phi}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{\Phi}}^{-1}\mathbf{p} \tag{8.64}$$

Several problems remain:

(i) There is no guarantee that the predicted probabilities $\hat{p}_j = \mathbf{X}_j\mathbf{b}$ are between 0 and 1: an empirical solution that has been recommended is to restrict the variance so that if $\hat{p}_i(1 - \hat{p}_i) \leq 0$, set $\hat{p} = 0.05$ or $\hat{p} = 0.98$.

(ii) Even then, there is no guarantee that $\hat{\hat{p}}$ based on $\hat{\boldsymbol{\beta}}_{\mathrm{EGLS}}$ is between 0 and 1. This points out the need to constrain the range of the elements of $\mathbf{p}$ to the interval [0,1].

8.2.2.2 Resolving the Range Constraint Problem

We can also solve the range constraint problem through the transformational logit.

Let

$$\mathbf{I}_j = \mathbf{x}'_j \boldsymbol{\beta} \tag{8.65}$$

$$P_j = \frac{1}{1 + e^{-\mathrm{I}_j}} \tag{8.65}$$

$$p_j = P_j + e_j = \frac{1}{1 + e^{-\mathrm{I}_j}} + e_j \tag{8.66}$$

It can be shown that

$$\mathrm{Ln}\frac{p_j}{1 - p_j} = \mathbf{x}'_j \boldsymbol{\beta} + \frac{e_j}{P_j(1 - P_j)} \tag{8.67}$$

Let

$$\mathrm{Ln}\frac{p_j}{1 - p_j} = v_j \text{ and } \frac{e_j}{P_j(1 - P_j)} = u_j$$

Then

$$v_j = \mathbf{x}'_j \beta + u_j \tag{8.68}$$

or for the full sample

$$\underset{K\times 1}{\mathbf{v}} = \underset{K\times p}{\mathbf{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{K\times 1}{\mathbf{u}} \tag{8.69}$$

$$\boldsymbol{\Phi} = \mathrm{E}\left[\underset{K\times K}{\mathbf{u}\mathbf{u}'}\right] = diag\left\{\mathrm{E}\left[u_j^2\right]\right\} \tag{8.70}$$

$$\mathrm{E}\left[u_j^2\right] = \mathrm{E}\left[\left(\frac{e_j}{P_j(1 - P_j)}\right)^2\right] = \frac{1}{P_j^2(1 - P_j)^2}\mathrm{V}\left[e_j\right] \tag{8.71}$$

$$= \frac{1}{P_j^2(1 - P_j)^2}\left[\frac{P_j(1 - P_j)}{n_j}\right] \tag{8.72}$$

$$= \frac{1}{n_j P_j(1 - P_j)} \tag{8.73}$$

Therefore, the generalized least squares estimator provides the minimum variance estimator:

$$\hat{\boldsymbol{\beta}}_{\mathbf{GLS}} = \left(\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{v} \tag{8.74}$$

where

$$\boldsymbol{\Phi} = diag\left\{\frac{1}{n_j P_j(1-P_j)}\right\} \tag{8.75}$$

But P_j is unknown. We can replace P_j by p_j in Eq. (8.75) and obtain the estimated generalized least squares estimator:

$$\hat{\boldsymbol{\beta}}_{\mathrm{EGLS}} = \left(\mathbf{X}'\hat{\boldsymbol{\Phi}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Phi}}^{-1}\mathbf{v} \tag{8.76}$$

In practice, let us define

$$\hat{\boldsymbol{\Phi}}^{-1/2} = diag\left\{[n_i p_i(1-p_i)]^{1/2}\right\} \tag{8.77}$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{EGLS}} = \left(\mathbf{X}'\hat{\boldsymbol{\Phi}}^{-1/2}\hat{\boldsymbol{\Phi}}^{-1/2}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Phi}}^{-1/2}\hat{\boldsymbol{\Phi}}^{-1/2}\mathbf{v} \tag{8.78}$$

Therefore, we can perform a transformation of the right and left sides of the equation and obtain the ordinary least squares of the transformed variables.

Let

$$\mathbf{v}^* = \hat{\boldsymbol{\Phi}}^{-1/2}\mathbf{v} \tag{8.79}$$

$$\mathbf{X}^* = \hat{\boldsymbol{\Phi}}^{-1/2}\mathbf{X} \tag{8.80}$$

and consequently,

$$\hat{\boldsymbol{\beta}}_{\mathrm{EGLS}} = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{v}^* \tag{8.81}$$

8.2.3 Conditional Logit Model

Let us consider an individual i facing a choice among K alternatives.

Let us define the variable y_{ij}:

$$\forall j = 1, \ldots, K : y_{ij} = \begin{cases} 1 & \text{if alternative } j \text{ is chosen} \\ 0 & \text{otherwise} \end{cases} \tag{8.82}$$

$$P_{ij} = P\left[y_{ij} = 1\right] \tag{8.83}$$

Only one alternative can be chosen; therefore

$$\sum_{j=1}^{K} y_{ij} = 1 \tag{8.84}$$

Also, probabilities sum to 1 by definition and therefore

$$\sum_{j=1}^{K} P_{ij} = 1 \tag{8.85}$$

Consequently, the likelihood function for an individual i is

$$\ell_i = \prod_{j=1}^{K} P_{ij}^{y_{ij}} \tag{8.86}$$

The likelihood function for all individuals is

$$\ell = \prod_{i=1}^{N} \prod_{j=1}^{K} P_{ij}^{y_{ij}} \tag{8.87}$$

For the multinomial logit model, if the unobserved utilities are a function of attributes and an error term that is distributed iid with the extreme value distribution (i.e., the cumulative distribution function is $F(\varepsilon_i < \varepsilon) = \exp(-e^{-\varepsilon})$), then the probability P_{ij} is defined as

$$P_{ij} = \frac{e^{u_{ij}}}{\sum_{k=1}^{K} e^{u_{ik}}} \tag{8.88}$$

where u_{ij} represents the utility associated with alternative j for individual i.

For the conditional logit model, two cases can be found depending on whether or not the explanatory variables determining the utility of the alternatives vary across alternatives. The first case (see Sect. 8.2.3.1) is when the variation in utilities of the alternatives comes from the differences in the explanatory variables but the marginal utilities are invariant. The second case (see Sect. 8.2.3.2) is when the source of variation in the utilities of the alternatives comes from the marginal utilities only.

8.2.3.1 Conditional Logit: Case 1

Different names are used for this first case of the conditional logit model. It is called Discrete Choice in LIMDEP and Alternative-Specific Conditional Logit in STATA.

The utility of an option varies because of different values of **x**s (e.g., attribute values of a brand):

$$P_{ij} = \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{\sum_{k=1}^{K} e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}} \tag{8.89}$$

For identification, we set $\mathbf{x}'_{i1} = \mathbf{0}$ or let us define

$$\mathbf{x}^{*'}_{ij} = \mathbf{x}'_{ij} - \mathbf{x}'_{i1} \tag{8.90}$$

This demonstrates that no constant term can be estimated in this model; a constant term would be indeterminate because the intercept disappears in Eq. (8.90).

The model parameters are estimated by maximum likelihood. The likelihood for individual i is

$$\boldsymbol{\ell}_i = \prod_{j=1}^{K} P_{ij}^{y_{ij}} \tag{8.91}$$

$$= \prod_{j=1}^{K} \left(\frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{\sum_{k=1}^{K} e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}} \right)^{y_{ij}} \tag{8.92}$$

For the N observations, the likelihood is

$$\boldsymbol{\ell} = \prod_{i=1}^{N} \boldsymbol{\ell}_i = \prod_{i=1}^{N} \prod_{j=1}^{K} \left(\frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{\sum_{k=1}^{K} e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}} \right)^{y_{ij}} \tag{8.93}$$

$$\mathbf{L} = \mathrm{Ln}\boldsymbol{\ell} = \sum_{i=1}^{N} \sum_{j=1}^{K} \mathrm{Ln} \left(\frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{\sum_{k=1}^{K} e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}} \right)^{y_{ij}} \tag{8.94}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \mathrm{Ln} \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{\sum_{k=1}^{K} e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}} \tag{8.95}$$

$$\mathbf{L} = \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \left(\mathbf{x}'_{ij}\boldsymbol{\beta} - \mathrm{Ln} \sum_{k=1}^{K} e^{\mathbf{x}'_{ik}\boldsymbol{\beta}} \right) \tag{8.96}$$

The optimization follows the iterative procedure described below.

Let t = iteration number. The gradient at iteration t is

$$\underset{p\times 1}{S[\boldsymbol{\beta}(t)]} = \left\{\frac{\partial \mathbf{L}}{\partial \boldsymbol{\beta}_p(t)}\right\} \tag{8.97}$$

Let us further define

$$\underset{p\times p}{Q[\beta(t)]} = \sum_{i=1}^{N}\left[S_i[\boldsymbol{\beta}(t)]S_i[\boldsymbol{\beta}(t)]'\right]$$

The value of the parameters at the next iteration is given by Eq. (8.98):

$$\underset{p\times 1}{\boldsymbol{\beta}(t+1)} = \boldsymbol{\beta}(t) + \left[Q[\boldsymbol{\beta}(t)]^{-1}S[\boldsymbol{\beta}(t)]\right] \tag{8.98}$$

The parameter estimates are obtained by convergence when the gradient vector approaches zero.

8.2.3.2 Conditional Logit: Case 2

This second case corresponds to the multinomial logit model (or binomial logit in the case of only two alternatives to choose from). In this case, the utility of an option varies because of different values of the marginal utilities $\boldsymbol{\beta}$s and because the factors predicting the utilities are the same across options:

$$P_{ij} = \frac{e^{\mathbf{x}'_i\boldsymbol{\beta}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}'_i\boldsymbol{\beta}_k}} \tag{8.99}$$

For identification, it is necessary to set $\boldsymbol{\beta}_1 = \mathbf{0}$.

The estimation of the model follows the same procedure as in the first case of the conditional model discussed in the prior section. We maximize the likelihood that is expressed as

$$\boldsymbol{\ell} = \prod_{i=1}^{N}\prod_{j=1}^{K}\left(\frac{e^{\mathbf{x}'_i\boldsymbol{\beta}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}'_i\boldsymbol{\beta}_k}}\right)^{y_{ij}} \tag{8.100}$$

Taking the logarithms,

$$\mathbf{L} = \sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\mathbf{x}'_i\boldsymbol{\beta}_j - \sum_{i=1}^{N}\mathrm{Ln}\sum_{j=1}^{K} e^{\mathbf{x}'_i\boldsymbol{\beta}_j} \tag{8.101}$$

An iterative procedure similar to case 1 above is used to obtain the maximum likelihood estimates. The only difference compared with case 1 comes from the larger size of the vector of parameters. The vector of all coefficients at iteration t is the vector with $(K-1)p$ elements $\boldsymbol{\beta}_{(K-1)p\times 1}(t)$.

The interpretation is, therefore, somewhat more complex in the case 2 model. The marginal utilities due to the increase of a unit of an explanatory variable are different across alternatives. Therefore, for example, variable x_1 may contribute marginally to the utility of alternative j but not significantly to the utility of alternative k.

8.2.4 Fit Measures

The fit measures follow for the most part those used in discriminant analysis, which are based on the classification table. We summarize them in Sect. 8.2.4.1. However, some additional measures are available because of the maximum likelihood estimation and its properties. These fit statistics are presented in Sect. 8.2.4.2.

8.2.4.1 Classification Table

These measures are the same as in discriminant analysis:

- Percentage of observations correctly classified
- Maximum chance criterion
- Proportional chance criterion
- Tau statistic

8.2.4.2 Statistics of Fit

Because of the properties of the likelihood function, two statistics can be used to test the model.

Log Likelihood Chi-Square Test

The null model is that the marginal utilities, apart from the constant term, are zero:

$$\mathrm{H_o} : \mathbf{b}_{slopes} = 0$$

If n is the number of successes ($y_i = 1$) observed in T observations, e.g., in the binary case

$$\text{under } \mathrm{H_o} : \boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_0\right) = \left(\frac{n}{T}\right)^n \left(\frac{T-n}{T}\right)^{T-n} \tag{8.102}$$

where $\hat{\boldsymbol{\beta}}_0$ represents the maximum likelihood estimates of the parameters of the reduced model with no slopes and $\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_0\right)$ is the value of the likelihood function obtained with these parameter estimates.

Taking the logarithm

$$\mathrm{Ln}\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_0\right) = \mathrm{nLn}\frac{n}{T} + (T-n)\mathrm{Ln}\left(\frac{T-n}{T}\right) \tag{8.103}$$

If $\hat{\boldsymbol{\beta}}_1$ is the value of the likelihood function estimated at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_1$, then

$$-2\left[\mathrm{Ln}\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_0\right) - \mathrm{Ln}\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_1\right)\right] \sim \chi^2_{(p-1)} \tag{8.104}$$

Therefore, an obvious advantage of the logit model vis-a-vis discriminant analysis is that it offers the possibility of testing the significance of the model.

Likelihood Ratio Index or Pseudo-R^2

Based on the same properties, the following index can be used:

$$\rho^2 = 1 - \frac{\mathrm{Ln}\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_1\right)}{\mathrm{Ln}\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_0\right)} \tag{8.105}$$

If the model is a perfect predictor in the sense that $\hat{P}_i = 1$ when $y_i = 1$ and $\hat{P}_i = 0$ when $y_i = 0$, then

$$\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_1\right) = 1 \Rightarrow \mathrm{Ln}\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_1\right) = 0 \Rightarrow \rho^2 = 1 \tag{8.106}$$

When there is no improvement in fit due to the predictor variables, then

$$\mathrm{Ln}\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_1\right) = \mathrm{Ln}\boldsymbol{\ell}\left(\hat{\boldsymbol{\beta}}_0\right) \Rightarrow \rho^2 = 0$$

8.3 Examples

8.3.1 Example of Discriminant Analysis

In Fig. 8.3, the SAS procedure "discrim" is used (highlighted in grey in the figure). The command "canonical" is also highlighted in grey because this gives the instructions to output all the relevant coefficients presented in Sect. 8.1. The variables used to discriminate are listed after the "var" term (highlighted in grey) and then the variable that contains the group numbering follows the term "class" (highlighted in grey) to indicate that it is a categorical variable.

Similarly, the same data file is read in STATA using the description of the format given in Fig. 8.4.

STATA commands indicate that there are three lines of data per record (i.e., observation) and the variables are then listed for each line. This dictionary file is called by the do-file that is shown in Fig. 8.5. The categorical variable is indicated by "group(*variable*)" where "*variable*" is the name of the variable being analyzed. The command "candisc" is used to perform canonical linear discriminant analysis. The command "discrim" can be used but only reports the classification table, as shown at the bottom of the figure.

The key sections of the SAS output are shown in Fig. 8.6. The output of discriminant analysis clearly shows the within-group SSCP matrices (separately for each group), the pooled-within SSCP matrix **W**, the between-group SSCP matrix **B**, and the total-sample SSCP matrix **T**. The raw (unstandardized) and standardized (correcting for the different units and variances of each of the variables) canonical coefficients, that is the discriminant coefficients, are then listed (highlighted in grey in Fig. 8.6). The raw coefficients indicate the weights to apply

```
OPTIONS LS=80;
DATA ALLIANCE;
INFILE "c:\SAMD\Chapter8\Examples\al8.dat";
INPUT   #1 choice dunc techu grow
        #2 firmsiz x1 7.4 x2 x3 asc
        #3 nccc;

proc discrim bsscp psscp wsscp tsscp canonical ;
        var dunc techu grow firmsiz asc nccc;
        class choice;
run;
```

Fig. 8.3 Example of SAS file for discriminant analysis (examp8-1.sas)

```
  infile dictionary using "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter8-MDA-
LOGIT/al8.dat" {
  _lines(3)
  _line(1)
                          choice dunc techu grow z1 z2 z3
  _line(2)
                               firmsiz x1 x2 x3 asc
  _line(3)
                                nccc z4 z5 z6 ads
  }
```

Fig. 8.4 STATA dictionary file for reading the example data (Al8dic_Mac.dct)

```
infile using "/users/gatignon/Documents/WORK_STATA/SAMD/Chapter8_MDA-
LOGIT/al8dic_Mac.dct", clear
candisc dunc techu grow firmsiz asc nccc, group(choice)
discrim lda dunc techu grow firmsiz asc nccc, group(choice)
```

Fig. 8.5 Example of STATA file for discriminant analysis (examp8-1.do)

```
                          The DISCRIM Procedure

          Observations     200          DF Total               199
          Variables          6          DF Within Classes      198
          Classes            2          DF Between Classes       1


                        Class Level Information

          Variable                                                  Prior
  choice  Name        Frequency       Weight    Proportion    Probability

       1  _1                155     155.0000      0.775000       0.500000
       2  _2                 45      45.0000      0.225000       0.500000
                            The SAS System                                2

                        The DISCRIM Procedure
                      Within-Class SSCP Matrices

                                choice = 1

       Variable              dunc              techu              grow

       dunc                 113.3               39.3               9.5
       techu                 39.3               79.4              48.0
       grow                   9.5               48.0              99.8
       firmsiz            -9339.1            -9615.8           -7354.2
       asc                  -27.4              1.7                6.4
       nccc                 -23.0              0.3                1.8

                                choice = 1

       Variable           firmsiz               asc              nccc

       dunc               -9339.1             -27.4             -23.0
       techu              -9615.8               1.7               0.3
       grow               -7354.2               6.4               1.8
       firmsiz        184070705.5           24104.1            9078.2
       asc                24104.1             132.8              21.5
       nccc                9078.2              21.5              83.9

--------------------------------------------------------------------------------


                                choice = 2
```

Fig. 8.6 SAS output for discriminant analysis (examp8-1.lst)

Variable	dunc	techu	grow
dunc	30.27	14.71	11.68
techu	14.71	26.14	14.97
grow	11.68	14.97	31.81
firmsiz	981.28	4710.89	12027.31
asc	-4.70	1.40	6.08
nccc	0.10	6.50	0.38

choice = 2

Variable	firmsiz	asc	nccc
dunc	981.28	-4.70	0.10
techu	4710.89	1.40	6.50
grow	12027.31	6.08	0.38
firmsiz	64024111.11	2718.02	-418.20
asc	2718.02	22.14	8.67
nccc	-418.20	8.67	22.90

Pooled Within-Class SSCP Matrix

Variable	dunc	techu	grow
dunc	143.6	54.0	21.2
techu	54.0	105.5	62.9
grow	21.2	62.9	131.6
firmsiz	-8357.8	-4904.9	4673.2
asc	-32.1	3.1	12.5
nccc	-22.9	6.8	2.2

Pooled Within-Class SSCP Matrix

Variable	firmsiz	asc	nccc
dunc	-8357.8	-32.1	-22.9
techu	-4904.9	3.1	6.8
grow	4673.2	12.5	2.2
firmsiz	248094816.6	26822.2	8660.0
asc	26822.2	154.9	30.2
nccc	8660.0	30.2	106.8

Between-Class SSCP Matrix

Variable	dunc	techu	grow
dunc	0.6129	-0.4180	0.7117
techu	-0.4180	0.2851	-0.4854
grow	0.7117	-0.4854	0.8264
firmsiz	-467.3848	318.7464	-542.7074
asc	-1.7287	1.1790	-2.0074
nccc	0.2759	-0.1881	0.3203

Between-Class SSCP Matrix

Variable	firmsiz	asc	nccc
dunc	-467.3848	-1.7287	0.2759
techu	318.7464	1.1790	-0.1881
grow	-542.7074	-2.0074	0.3203
firmsiz	356391.4050	1318.2102	-210.3682
asc	1318.2102	4.8758	-0.7781
nccc	-210.3682	-0.7781	0.1242

Total-Sample SSCP Matrix

Variable	dunc	techu	grow

Fig. 8.6 (continued)

```
dunc                144.2              53.6              21.9
techu                53.6             105.8              62.5
grow                 21.9              62.5             132.4
firmsiz           -8825.2           -4586.1            4130.4
asc                 -33.9               4.3              10.5
nccc                -22.6               6.7               2.5

                    Total-Sample SSCP Matrix

Variable           firmsiz               asc              nccc

dunc              -8825.2             -33.9             -22.6
techu             -4586.1               4.3               6.7
grow               4130.4              10.5               2.5
firmsiz       248451208.0           28140.4            8449.6
asc             28140.4               159.8              29.4
nccc             8449.6                29.4             106.9


              Pooled Covariance Matrix Information

                                 Natural Log of the
                  Covariance     Determinant of the
                 Matrix Rank      Covariance Matrix

                           6              11.06578

      Pairwise Generalized Squared Distances Between Groups

                2                          -1
               D (i|j) = (X - X )' COV    (X - X )
                           i   j           i   j


            Generalized Squared Distance to choice

          From choice             1             2

                    1             0       0.39588
                    2       0.39588             0


                 Canonical Discriminant Analysis

                                Adjusted    Approximate        Squared
                  Canonical     Canonical       Standard      Canonical
                 Correlation   Correlation          Error    Correlation

1                0.255312      0.209914       0.066267       0.065184

                              Eigenvalues of Inv(E)*H
                                = CanRsq/(1-CanRsq)

                Eigenvalue    Difference    Proportion    Cumulative

      1          0.0697                       1.0000        1.0000

           Test of H0: The canonical correlations in the
             current row and all that follow are zero

           Likelihood    Approximate
                Ratio        F Value    Num DF    Den DF    Pr > F

            1  0.93481561          2.24         6       193    0.0408

                 NOTE: The F statistic is exact.

                 Canonical Discriminant Analysis

                    Total Canonical Structure

                    Variable              Can1

                    dunc              0.255331
                    techu            -0.203296
```

Fig. 8.6 (continued)

grow 0.309396
firmsiz -0.148344
asc -0.684158
nccc 0.133471

Between Canonical Structure

Variable	Can1
dunc	1.000000
techu	-1.000000
grow	1.000000
firmsiz	-1.000000
asc	-1.000000
nccc	1.000000

Pooled Within Canonical Structure

Variable	Can1
dunc	0.247396
techu	-0.196824
grow	0.300080
firmsiz	-0.143531
asc	-0.671812
nccc	0.129123

Canonical Discriminant Analysis

Total-Sample Standardized Canonical Coefficients

Variable	Can1
dunc	0.3875344511
techu	-.7516524862
grow	0.7000312218
firmsiz	-.0910945522
asc	-.7239897268
nccc	0.4082828732

Pooled Within-Class Standardized Canonical Coefficients

Variable	Can1
dunc	0.3876854449
techu	-.7525324874
grow	0.6996037720
firmsiz	-.0912587756
asc	-.7146572217
nccc	0.4090748704

Raw Canonical Coefficients

Variable	Can1
dunc	0.455199542
techu	-1.030770927
grow	0.858082117
firmsiz	-0.000081526
asc	-0.807915970
nccc	0.556967570

Class Means on Canonical Variables

choice	Can1
1	0.1415685600
2	-.4876250399

Fig. 8.6 (continued)

Linear Discriminant Function

$$\text{Constant} = -.5\ \bar{X}_j'\ COV^{-1}\ \bar{X}_j \qquad \text{Coefficient Vector} = COV^{-1}\ \bar{X}_j$$

Linear Discriminant Function for choice

Variable	1	2
Constant	-0.49854	-1.04596
dunc	0.28366	-0.00275
techu	-0.21886	0.42970
grow	0.03893	-0.50097
firmsiz	0.0004568	0.0005081
asc	0.94089	1.44923
nccc	-0.24923	-0.59967

Classification Summary for Calibration Data: WORK.ALLIANCE
Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X-\bar{X}_j)'\ COV^{-1}\ (X-\bar{X}_j)$$

Posterior Probability of Membership in Each choice

$$Pr(j|X) = \exp(-.5\ D_j^2(X)) \ /\ SUM_k \exp(-.5\ D_k^2(X))$$

Number of Observations and Percent Classified into choice

From choice	1	2	Total
1	97	58	155
	62.58	37.42	100.00
2	12	33	45
	26.67	73.33	100.00
Total	109	91	200
	54.50	45.50	100.00
Priors	0.5	0.5	

Error Count Estimates for choice

	1	2	Total
Rate	0.3742	0.2667	0.3204
Priors	0.5000	0.5000	

Fig. 8.6 (continued)

to the p variates in order to form the most discriminating linear function. In the example, $y_i = 0.455*DUNC_i - 1.031*TECHU_i + 0.858*GROW_i - 0.00008*FIRMSIZ_i - 0.808*ASC_i + 0.557*NCCC_i$. In the particular case where only two groups are analyzed, a single discriminant function exists; there is only one eigenvector. The eigenvectors, or the discriminant functions, discussed earlier are interpretable in such a way that a positive (negative) sign of the discriminant function coefficients (weights) indicates that the corresponding variable contributes positively (negatively) to the discriminant function. A comparison with the group

means on the discriminant function indicates in what way the variates discriminate among the groups. For example, considering the values highlighted in grey in Fig. 8.6, choice 1 has a higher (positive) mean value (0.142) on the discriminant function y (the mean for choice 2 is negative, i.e., −0.488). Therefore, the positive coefficient of DUNC means that the higher the demand uncertainty (the higher the value of DUNC), the higher the discriminant function and, consequently, the more likely is choice 1 (corresponding to internal development mode). In contrast to DUNC that has a positive coefficient, the negative coefficient of TECHU means that higher technological uncertainty makes choice 2 (corresponding to using an alliance) more likely.

In addition, the absolute value of the standardized discriminant function coefficients (where the raw coefficients are multiplied by the standard deviation of the corresponding variables) reflects the contribution of the variables to that discriminant function so that a larger standardized weight indicates a bigger role of that variable in discriminating between the options. For example, the variable technology uncertainty ("techu") appears to be the most discriminant variable (−0.75), followed closely by the variables "asc" (−0.71) and "grow" (0.69), although observations with higher values of growth ("grow") are likely to belong to different groups from those with high ratings on "asc" and "techu" because of the opposite signs of these coefficients. Therefore, these standardized coefficients explain the contribution (extent and direction) of each variable for discriminating between the two groups.

For two-group discriminant analysis, the interpretation of the discriminant function weights is relatively clear, as presented above. When there are more than two groups, each discriminant function represents different dimensions on which the discrimination between groups would occur. For example, the first discriminant function could discriminate between groups 1 and 3 versus group 2, and the second discriminant function could discriminate between groups 1 and 2 on the one hand and group 3 on the other hand. The interpretation in such cases requires the comparison of the group means on the discriminant function values (y). Because it can sometimes be difficult to find an interpretation for the discriminant functions, it helps to plot the group means or centroids on the discriminant functions (as axes). It is also very useful to analyze the profiles of each group in terms of the means of the predictor variables for each group.

In Fig. 8.6, a vector of coefficients for each group is printed under the heading of "linear discriminant function." These are not, however, the discriminant functions discussed earlier; they are the classification functions. Indeed, in this particular example with only two choices, there could not be two discriminant functions. The SAS output shows the classification functions, which are the two components of Eq. (8.33), i.e., $\mathbf{W}^{-1}\overline{\mathbf{x}}_1$ and $\mathbf{W}^{-1}\overline{\mathbf{x}}_2$.

The classification table is also shown in Fig. 8.6 and the number of observations correctly classified is highlighted in grey. In this example, 62.58% of the observations in group 1 were classified in the correct group and 73.3% for group 2.

Equivalent output is given by STATA, as shown in Fig. 8.7.

```
. candisc dunc techu grow firmsiz asc nccc, group(choice)

Canonical linear discriminant analysis

      |                                  | Like-
      |  Canon.   Eigen-     Variance    | lihood
  Fcn |  Corr.    value   Prop.   Cumul. | Ratio     F      df1    df2  Prob>F
 -----+----------------------------------+-------------------------------------
    1 | 0.2553   .06973  1.0000  1.0000 | 0.9348   2.243     6    193  0.0408 e
 ---------------------------------------------------------------------------
  Ho: this and smaller canon. corr. are zero;                      e = exact F

Standardized canonical discriminant function coefficients

                | function1
    ------------+-----------
            dunc |   .3876854
          techu | -.7525325
            grow |   .6996038
         firmsiz | -.0912588
             asc | -.7146572
            nccc |   .4090749

Canonical structure

                | function1
    ------------+-----------
            dunc |   .2473955
          techu | -.1968236
            grow |   .3000803
         firmsiz | -.1435311
             asc | -.6718125
            nccc |   .1291226

Group means on canonical variables

          choice | function1
    -------------+-----------
               1 |   .1415686
               2 |   -.487625

Resubstitution classification summary

    +---------+
    | Key     |
    |---------|
    | Number  |
    | Percent |
    +---------+
                | Classified
    True choice |         1        2 |   Total
    ------------+-------------------+-------
               1 |        97       58 |     155
                 |     62.58    37.42 |  100.00
                 |                    |
               2 |        12       33 |      45
                 |     26.67    73.33 |  100.00
    ------------+-------------------+-------
          Total |       109       91 |     200
                 |     54.50    45.50 |  100.00
                 |                    |
         Priors |    0.5000   0.5000 |

. discrim lda dunc techu grow firmsiz asc nccc, group(choice)

Linear discriminant analysis
Resubstitution classification summary

    +---------+
    | Key     |
    |---------|
    | Number  |
    | Percent |
    +---------+
                | Classified
```

Fig. 8.7 STATA output for discriminant analysis (examp8-1.log)

```
True choice  |      1       2  |  Total
-------------+-----------------+-------
           1 |     97      58  |    155
             |  62.58   37.42  | 100.00
             |                 |
           2 |     12      33  |     45
             |  26.67   73.33  | 100.00
-------------+-----------------+-------
       Total |    109      91  |    200
             |  54.50   45.50  | 100.00
             |                 |
      Priors | 0.5000  0.5000  |
```

Fig. 8.7 (continued)

```
read; nrec = 4648; nvar=14; file = scanner.dat;
format = (f8.0,f4.0,2f2.0,f3.0,2f5.2,f2.0,f9.6,5f2.0);
names(x1  = panelid,
      x2  = week,
      x3  = purchase,
      x4  = count,
      x5  = brand,
      x6  = price,
      x7  = prcut,
      x8  = feature,
      x9  = loy,
      x10 = dum1,
      x11 = dum2,
      x12 = dum3,
      x13 = dum4,
      x14 = dum5);
$
open; output=c:\SAMD\Chapter8\Examples\Examp8-2.out$
discrete choice; lhs=purchase, count;
        rhs=price, prcut, feature, loy, dum1, dum2,dum3, dum4, dum5$
close$
```

Fig. 8.8 Example of LIMDEP file for logit model—case 1 (examp8-2.lim)

8.3.2 Example of Multinomial Logit: Case 1 Analysis Using LIMDEP

Figure 8.8 presents a typical input file using LIMDEP to estimate a conditional logit model of the case 1 type. The data set used for this example, scanner.dat, has the same structure as the data scan.dat described in Appendix C (Chap. 14). The first part of the file defines the data variables and reads them from the data file. The specification of the analysis follows in the second part with the command "discrete choice" (highlighted in grey in the figure). The variables on the left side of the

```
Normal exit from iterations. Exit status=0.
 : LIMDEP Estimation Results                    Run log line    3  Page   1 :
 : Current sample contains    4648 observa  tions.                           :


                +-----------------------------------------------  +
                | Discrete choice (multinomial logit) model    |
                | Maximum Likelihood Estimates                 |
                | Dependent  variable               Choice     |
                | Weighting variable                   ONE     |
                | Number of observations               949     |
                | Iterations completed                   6     |
                | Log likelihoo d function         -814.1519     |
                | Log-L for Choice    model =      -814.1519     |
                | R2=1 -LogL/LogL*  Log -L fncn  R-sqrd  RsqAdj |
                | No coefficients   -1700.3797  .52119  .52003 |
                | Constants only.   Must be computed directly. |
                |                  Use NLOGIT ;...; RHS=ONE $ |
                | Response data are given as ind. choice.      |
                | Number of obs.=   949, skipped   0 bad obs. |
                +---------------------- ---------------------- +

+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error | b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
 PRICE     -2.372695061       .33603584       -7.061    .0000
 PRCUT      1.973968500       .35129043        5.619    .0000
 FEATURE    .7023317528       .13901356        5.052    .0000
 LOY        3.791733215       .15780806       24.028    .0000
 DUM1       .97173189 76E-01  .24160340         .402    .6875
 DUM2       .9067318292       .25947016        3.495    .0005
 DUM3       .9511561911       .31347219        3.034    .0024
 DUM4       .4835120963       .25106381        1.926    .0541
 DUM5       .9019121730       .3899720 9       2.313    .0207
```

Fig. 8.9 LIMDEP output for logit model—case 1 (examp8-2.out)

equation are then specified (purchase) following the code "lhs=" (highlighted in grey). Finally, the explanatory variables are listed after the code "rhs=" for the right side of the equation (highlighted in grey). It is important to note that in LIMDEP the options must be coded from 0 to $K-1$. The predicted variables in the example of Fig. 8.8 consist of the price of each brand, any price cut applied to each transaction, and whether or not the brand was on display. Each brand is also specified as having a different intrinsic preference or utility that is modeled as a different constant term with dummy variables (the reference where all brand dummies are zero corresponds to private labels). Some heterogeneity in preferences across consumers is also captured by a loyalty measure representing past purchases of the brand.

The LIMDEP output is shown in Fig. 8.9.

The output shown in Fig. 8.9 should be self-explanatory. The gradient is printed at each iteration until convergence is achieved. Then, the estimated parameters are listed with the usual statistics that enable us to test the hypotheses and compute the fit statistics based on the likelihood function. The coefficients represent the marginal utility of each choice option (brand) of one additional unit of the corresponding variable. In the example in Fig. 8.9, price has a significant negative impact while price cuts and being on display add to the brand utility.

```
infile dictionary using "/users/gatignon/Documents/WORK_STATA/SAMD/Chapter8_MDA-
LOGIT/scan.dat" {
_lines(1)
_line(1)
panelid week purchase count brand price prcut feature loy dum1 dum2 dum3 dum4 dum5
}
```

Fig. 8.10 Dictionary file for reading the scanner data file in STATA (scandic.dct)

```
infile using "/users/gatignon/Documents/WORK_STATA/SAMD/Chapter8_MDA-
LOGIT/Scandic_Mac.dct", clear
gen pan1=string(panelid,"%03.0f")
gen week1=string(week,"%03.0f")
gen panelidnew=pan1+week1
encode panelidnew, generate(case2)
asclogit purchase price prcut feature loy dum1 dum2 dum3 dum4 dum5, case(case2)
alternative(brand) noconstant
```

Fig. 8.11 Example of STATA file for logit model—case 1 (examp8-2.do)

For STATA, we first define a dictionary to read the scanner data file. Fig. 8.10 shows that file.

The commands for estimating the conditional logit choice model in STATA (alternative-specific conditional logit) are shown in Fig. 8.11.

However, before explaining the logit command, we must use the panelid and the week variable to generate the variable that identifies each case. We first generate strings from these numeric variables, then we concatenate them with the "+" operator, and finally, we reconvert the new identification number by combining panelid and week. This means that each case corresponds to a panelist for a given week. The command for the conditional logit—case 1 model is "asclogit" (for alternative-specific conditional logit). Because multiple units of the brands are chosen at each purchase occasion (indicated by the variable "count"), this information was used in the example using LIMDEP. Similarly in STATA, the command line would simply become

`asclogit purchase price prcut feature loy dum1 dum2 dum3 dum4 dum5 [fweight = count], case(case2) alternative (brand) noconstant`

The output is shown in Fig. 8.12.

8.3.3 Example of Conditional Logit: Case 2 Analysis Using LIMDEP and STATA

Figure 8.13 shows the LIMDEP file that estimates the same choice as for the discriminant analysis example above. The "logit" command is highlighted in grey in the figure and the variables on the right and left sides of the equation follow the

```
Iteration 0:   log likelihood = -566.63846
Iteration 1:   log likelihood = -544.56834
Iteration 2:   log likelihood = -530.19524
Iteration 3:   log likelihood = -530.13254
Iteration 4:   log likelihood = -530.13251

Alternative-specific conditional logit          Number of obs      =       2967
Case variable: case2                            Number of cases    =        634

Alternative variable: brand                     Alts per case: min =          2
                                                               avg =        4.7
                                                               max =          6

                                                  Wald chi2(9)     =     472.21
Log likelihood = -530.13251                       Prob > chi2      =     0.0000

------------------------------------------------------------------------------
    purchase |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
brand        |
       price |  -2.323626   .4515947    -5.15   0.000    -3.208736   -1.438517
       prcut |   2.476117   .4170589     5.94   0.000     1.658697    3.293538
     feature |   .6756557   .1672598     4.04   0.000     .3478326    1.003479
         loy |   3.810995   .2002705    19.03   0.000     3.418472    4.203518
        dum1 |   .7056693   .3777715     1.87   0.062    -.0347492    1.446088
        dum2 |   1.490968   .4035483     3.69   0.000     .7000282    2.281908
        dum3 |   1.458424   .4700251     3.10   0.002     .5371918    2.379656
        dum4 |   1.029559   .4067982     2.53   0.011     .2322488    1.826868
        dum5 |   1.281715   .5588563     2.29   0.022     .1863768    2.377053
------------------------------------------------------------------------------
```

Fig. 8.12 STATA output for logit model—case 1 (examp8-2.do)

```
read; nrec = 200; nvar=8; file = al8.dat;
format = (f1.0,3f8.4/f17.4,24x,f8.4/f17.4,24x,f8.4);
names(x1 = rdmode,
       x2 = dunc,
       x3 = techu,
       x4 = grow,
       x5 = firmsiz,
       x6 = as2,
       x7 = nccc,
       x8 = ads,
$
create; rdmode= rdmode-1$
open; output=c:\SAMD\chapter8\examples\examp8-3.out$
       logit; lhs=rdmode;
              rhs=one, dunc, techu, grow, firmsiz, as2, nccc, ads$
close$
```

Fig. 8.13 Example of input for logit model using LIMDEP (examp8-3.lim)

same rules as those described for the "discrete choice" command in the prior section. These commands "lhs=" and "rhs=" are highlighted in grey in Fig. 8.13. Two aspects of the file require particular attention:

1. The choice variables should have a value of zero for the base case, up to the number of choice options minus one. In the example, the choice variable, which is the R&D mode, is re-coded to take the value 0 or 1 depending on whether the original variable read from the data file is 1 or 2.
2. LIMDEP does not automatically estimate a constant term. Therefore, if one expects different proportions to be chosen for the same values of the independent variables, then the variable called "one" in LIMDEP serves to add the constant term.

It can be seen from the LIMDEP output, shown in Fig. 8.14, that the results are displayed with the parameter estimates and the classification table, as was described previously (the key results are highlighted in grey in Fig. 8.14). The information necessary to compute the likelihood ratio test is also given with the log-likelihood functions for the full model and for the restricted version (no slopes). The chi-square statistic is also provided. The pseudo R-squared can be computed with this information as well.

The STATA input for this conditional logit—case 2 model is shown in Fig. 8.15. The STATA command corresponding to this model is "mlogit" for multinomial logistic regression.

The results shown in Fig. 8.16 reproduce those obtained with LIMDEP.

Because of the binary nature of the dependent variable "choice," the binomial model command "logit" could be used in place of "mlogit." This assumes, however, that the dependent variable takes the values 0 and 1. In the input file shown in Fig. 8.17, a new variable "rdchoice" is generated to satisfy this assumption.

The STATA output is shown in Fig. 8.18, including the classification table with the percentage of correctly classified observations.

In the case of the multinomial model, the classification table could be derived but the various thresholds need to be determined. Instead, the focus is on the interpretation of the coefficients, which, however, do not represent the effect of the x variables on the probabilities of each option but rather represent the changes in the utility of each option. The STATA post-estimation command ". margins, dydx (*) predict(outcome(2))" (highlighted in grey in Fig. 8.19) estimates the average marginal effects of each variable on choice option 2, as shown in Fig. 8.19.

8.4 Assignment

Use SURVEY.ASC data (described in Chap. 14, Appendix C) to run a model where the dependent variable is a categorical scale (choose preferably a variable with more than two categories). For example, you may want to address the following questions:

- Can purchase process variables be explained by psychographics?
- Are demographics and/or psychographics determinants of media habits?

```
: LIMDEP Estimation Results                        Run log line    4  Page   1 :
: Current sample contains     200 observations.                                :

            +-----------------------------------------------+
            | Multinomial logit model                       |
            | There are  2 outcomes for LH variable RDMODE  |
            | These are the OLS start values based on the   |
            | binary variables for each outcome Y(i) = j.   |
            | Coefficients for LHS=0 outcome are set to 0.0 |
            +-----------------------------------------------+

+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
         Characteristics in numerator of Prob[Y = 1]
 Constant  .2071964751      .38692314E-01    5.355    .0000
 DUNC     -.1371394030E-01  .36081756E-01    -.380    .7039 -.22794000E-01
 TECHU     .5819511367E-01  .47307026E-01    1.230    .2186 -.11773500E-01
 GROW     -.8250070805E-01  .37666224E-01   -2.190    .0285  .19359500E-01
 FIRMSIZ   .1275473374E-04  .23474470E-04     .543    .5869  706.10000
 AS2       .1665741370E-01  .32130499E-01     .518    .6042  .79726850
 NCCC     -.4558722443E-01  .37072263E-01   -1.230    .2188 -.24566500E-01
 ADS       .2261545326      .31404416E-01    7.201    .0000 -.16065145E-01

Normal exit from iterations. Exit status=0.

: LIMDEP Estimation Results                        Run log line    4  Page   2 :
: Current sample contains     200 observations.                                :

              +-----------------------------------------+
              | Multinomial Logit Model                 |
              | Maximum Likelihood Estimates            |
              | Dependent variable            RDMODE    |
              | Weighting variable               ONE    |
              | Number of observations           200    |
              | Iterations completed               7    |
              | Log likelihood function       -73.57682  |
              | Restricted log likelihood     -106.6328  |
              | Chi-squared                    66.11190  |
              | Degrees of freedom                 7    |
              | Significance level           .0000000   |
              +-----------------------------------------+

+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
         Characteristics in numerator of Prob[Y = 1]
 Constant -2.247599465      .40470933       -5.554    .0000
 DUNC     -.1341133808      .30167521        -.445    .6566 -.22794000E-01
 TECHU     .5217618767      .38423330        1.358    .1745 -.11773500E-01
 GROW     -.7767888885      .32769650       -2.370    .0178  .19359500E-01
 FIRMSIZ   .1237468921E-03  .17355371E-03     .713    .4758  706.10000
 AS2       .1825140247      .27622638         .661    .5088  .79726850
 NCCC     -.6736865330      .31454643       -2.142    .0322 -.24566500E-01
 ADS       2.038879995      .36284355        5.619    .0000 -.16065145E-01


Frequencies of actual & predicted outcomes
Predicted outcome has maximum probability.

             Predicted
------  ----------  +  -----
Actual       0    1  |  Total
------  ----------  +  -----
  0        143   12  |    155
  1         25   20  |     45
------  ----------  +  -----
Total      168   32  |    200
```

Fig. 8.14 Example of LIMDEP output for logit model (examp8-3.out)

```
infile using "/users/gatignon/Documents/WORK_STATA/SAMD/Chapter8_MDA-
LOGIT/al8dic_Mac.dct", clear
mlogit choice dunc techu grow firmsiz asc nccc ads
```

Fig. 8.15 Example of STATA input command for multinomial logit model (examp8-3_Mac.do)

```
. mlogit choice dunc techu grow firmsiz asc nccc ads

Iteration 0:   log likelihood = -106.63277
Iteration 1:   log likelihood = -78.511669
Iteration 2:   log likelihood = -73.675095
Iteration 3:   log likelihood = -73.576887
Iteration 4:   log likelihood =   -73.5768
Iteration 5:   log likelihood =   -73.5768

Multinomial logistic regression                 Number of obs   =        200
                                                LR chi2(7)      =      66.11
                                                Prob > chi2     =     0.0000
Log likelihood =   -73.5768                     Pseudo R2       =     0.3100

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
        dunc |  -.1341132   .3016753    -0.44   0.657     -.725386    .4571595
       techu |   .5217616   .3842333     1.36   0.174    -.2313218    1.274845
        grow |  -.7767891   .3276965    -2.37   0.018    -1.419062   -.1345157
     firmsiz |   .0001237   .0001736     0.71   0.476    -.0002164    .0004639
         asc |   .1825139   .2762264     0.66   0.509    -.3588799    .7239078
       nccc |  -.6736857   .3145464    -2.14   0.032    -1.290185   -.0571861
        ads |   2.038876   .3628429     5.62   0.000     1.327717    2.750035
      _cons |  -2.247596   .4047093    -5.55   0.000    -3.040812   -1.454381
------------------------------------------------------------------------------
```

Fig. 8.16 Example of STATA output of mlogit (examp8-3.log)

```
infile using "/users/gatignon/Documents/WORK_STATA/SAMD/Chapter8_MDA-
LOGIT/al8dic2_Mac.dct", clear
gen rdchoice=choice-1
logit rdchoice dunc techu grow firmsiz asc nccc ads
estat classification
```

Fig. 8.17 STATA command for binomial logit model

Note that for these analyses, you can use discriminant analysis with SAS or STATA, or the conditional logit—case 2 model estimated using STATA or LIMDEP. In both cases (discriminant analysis and multinomial logit model), provide fit statistics in addition to the explanation of the coefficients. Compare the results of both analyses. Pay particular attention to the format for reading the variables in LIMDEP, because the Windows version does not recognize format *i* for integers.

```
. gen rdchoice=choice-1

. logit rdchoice dunc techu grow firmsiz asc nccc ads

Iteration 0:   log likelihood = -106.63277
Iteration 1:   log likelihood = -78.511669
Iteration 2:   log likelihood = -73.675095
Iteration 3:   log likelihood = -73.576887
Iteration 4:   log likelihood =   -73.5768
Iteration 5:   log likelihood =   -73.5768

Logistic regression                               Number of obs   =        200
                                                  LR chi2(7)      =      66.11
                                                  Prob > chi2     =     0.0000
Log likelihood =   -73.5768                       Pseudo R2       =     0.3100

------------------------------------------------------------------------------
    rdchoice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dunc |  -.1341132   .3016753    -0.44   0.657     -.725386    .4571595
       techu |   .5217616   .3842333     1.36   0.174    -.2313218    1.274845
        grow |  -.7767891   .3276965    -2.37   0.018    -1.419062   -.1345157
     firmsiz |   .0001237   .0001736     0.71   0.476    -.0002164    .0004639
         asc |   .1825139   .2762264     0.66   0.509    -.3588799    .7239078
       nccc |  -.6736857   .3145464    -2.14   0.032    -1.290185   -.0571861
         ads |   2.038876   .3628429     5.62   0.000     1.327717    2.750035
       _cons |  -2.247596   .4047093    -5.55   0.000    -3.040812   -1.454381
------------------------------------------------------------------------------

. estat classification

Logistic model for rdchoice

             -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        20            12  |         32
     -     |        25           143  |        168
-----------+--------------------------+-----------
   Total   |        45           155  |        200

Classified + if predicted Pr(D) >= .5
True D defined as rdchoice != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   44.44%
Specificity                     Pr( -|~D)   92.26%
Positive predictive value       Pr( D| +)   62.50%
Negative predictive value       Pr(~D| -)   85.12%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    7.74%
False - rate for true D         Pr( -| D)   55.56%
False + rate for classified +   Pr(~D| +)   37.50%
False - rate for classified -   Pr( D| -)   14.88%
--------------------------------------------------
Correctly classified                        81.50%
--------------------------------------------------
```

Fig. 8.18 STATA output for binomial logit model

Using grocery scanner data in the file SCAN.DAT (the description of the file can be found in Chap. 14, Appendix C), model the brand choice of the frequently purchased grocery product for which the data has been collected. Use LIMDEP (discrete choice) or STATA (alternative-specific conditional logit) to estimate the conditional logit—case 1 models.

You may want to consider the following ideas for possible analysis:

- How does the inclusion of the "loyalty" variable (i.e., a measure of cross-sectional heterogeneity and nonstationarity) affect the brand choice model?

```
. margins, dydx(*) predict(outcome(2))

Average marginal effects                          Number of obs   =        200
Model VCE    : OIM

Expression   : Pr(choice==2), predict(outcome(2))
dy/dx w.r.t. : dunc techu grow firmsiz asc nccc ads

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dunc |  -.0157865   .0354588    -0.45   0.656    -.0852846    .0537116
       techu |   .0614167   .0445037     1.38   0.168     -.025809    .1486423
        grow |   -.091436   .0363809    -2.51   0.012    -.1627413   -.0201307
     firmsiz |   .0000146   .0000203     0.72   0.474    -.0000253    .0000544
         asc |   .0214838   .0323754     0.66   0.507    -.0419708    .0849383
        nccc |  -.0792997    .035683    -2.22   0.026     -.149237   -.0093623
         ads |   .2399966   .0278678     8.61   0.000     .1853767    .2946165
------------------------------------------------------------------------------
```

Fig. 8.19 STATA estimates of average marginal effects of the predictor variables on a given option

- What do we gain, if anything, by separating price paid into its two components?
- Are there brand-specific price effects?

Bibliography

Basic Technical Readings

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics* [Chap. 3 and Chap. 4]. Cambridge: Cambridge University Press.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics*. New York, NY: Academic.

McFadden, D. (1980). Econometric models of probabilistic choice among products. *Journal of Business, 53*(3), S13–S29.

Morrison, D. G. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research, 6*(2), 156–163.

Press, S. J., & Sandra, W. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*(364), 699–705.

Schmidt, P., & Strauss, R. P. (1975). The prediction of occupation using multiple logit models. *International Economic Review, 16*(2), 471–486.

Application Readings

Adapa, S. (2008). Discriminant analysis of adopters and non-adopters of global brands: Empirical evidence from India and Malaysia. *The ICFAI University Journal of Brand Management, 5*(4), 7–25.

Aviv, S., & Ruvio, A. (2008). Opinion leaders and followers: A replication and extension. *Psychology and Marketing, 25*(3), 280–297.

Bruderl, J., & Schussler, R. (1990). Organizational mortality: The liabilities of newness and adolescence. *Administrative Science Quarterly, 35*(3), 530–547.

Corstjens, M. L., & Gautschi, D. A. (1983). Formal choice models in marketing. *Marketing Science, 2*(1), 19–56.

Deshpandé, R., Farley, J. U., & Webster, F. E. J. (1993, January). Corporate culture, customer orientation, and innovativeness in Japanese firms: A quadrad analysis. *Journal of Marketing, 57*, 23–37.

Duffy, R. S. (2008). Towards a better understanding of partnership attributes: An exploratory analysis of relationship type classification. *Industrial Marketing Management, 37*(2), 228–244.

Fader, P. S., & Lattin, J. M. (1993). Accounting for heterogeneity and nonstationarity in a cross-sectional model of consumer purchase behavior. *Marketing Science, 12*(3), 304–317.

Fader, P. S., Lattin, J. M., & Little, J. D. C. (1992). Estimating nonlinear parameters in the multinomial logit model. *Marketing Science, 11*(4), 372–385.

Foekens, E. W., Leeflang, P. S. H., & Wittink, D. (1997). Hierarchical versus other market share models for markets with many items. *International Journal of Research in Marketing, 14*, 359–378.

Fotheringham, A. S. (1988). Consumer store choice and choice set definition. *Marketing Science, 7*(3), 299–310.

Gatignon, H., & Anderson, E. (1988). The multinational corporation's degree of control over foreign subsidiaries: An empirical test of a transaction cost explanation. *Journal of Law, Economics and Organization, 4*(2), 89–120.

Gatignon, H., & Reibstein, D. J. (1986). Pooling logit models. *Journal of Marketing Research, 23*(3), 281–285.

Guadagni, P. M., & Little, J. D. C. (1983). A logit model brand choice calibrated on scanner data. *Marketing Science, 2*(3), 203–238.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research, 25*(4), 342–355.

Gupta, S., Chintagunta, P. K., & Wittink, D. R. (1997). Household heterogeneity and state dependence in a model of purchase strings: Empirical results and managerial implications. *International Journal of Research in Marketing, 14*, 341–357.

Hall, E. H., Jr., & St. John, C. H. (1994). A methodological note on diversity measurement. *Strategic Management Journal, 15*(2), 153–168.

Hardie, B. G. S., Johnson, E. J., & Fader, P. S. (1992). Modeling loss aversion and reference dependence effects on brand choice. *Marketing Science, 12*(4), 378–394.

Pieper, T. M., Klein, S. B., & Jaskiewicz, P. (2008). The impact of goal alignment on board existence and top management team composition: Evidence from family-influenced businesses. *Journal of Small Business Management, 46*(3), 372–394.

Robertson, T. S., & Gatignon, H. (1998). Technology development mode: A transaction cost conceptualization. *Strategic Management Journal, 19*(6), 515–532.

Sinha, A. (2000). Understanding supermarket competition using choice maps. *Marketing Letters, 11*(1), 21–35.

Tallman, S. B. (1991, Summer). Strategic management models and resource-based strategies among MNEs in a host market. *Strategic Management Journal, 12*, 69–82.

Torben, H. (2005). Consumer adoption of online grocery buying: A discriminant analysis. *International Journal of Retail & Distribution Management, 33*(2), 101–121.

Vasudevan, R., Venkatraman, N., & Camillus, J. C. (1986). Multi-objective assessment of effectiveness of strategic planning: A discriminant analysis approach. *Academy of Management Journal, 29*(2), 347–372.

Wiggins, R. R., & Ruefli, T. W. (1995). Necessary conditions for the predictive validity of strategic groups: Analysis without reliance on clustering techniques. *Academy of Management Journal, 38*(6), 1635–1656.

Yapa, L. S., & Mayfield, R. C. (1978). Non-adoption of innovations: Evidence from discriminant analysis. *Economic Geography, 54*(2), 145–156.

Chapter 9
Rank-Ordered Data

When the criterion variable is defined on an ordinal scale, the typical analyses based on correlations or covariances are not appropriate. The methods described in Chap. 6 do not use the ordered nature of the data and, consequently, do not use all the information available. In this chapter, we present methodologies that take into account the ordinal property of the dependent variable.

A particular methodology that typically uses ordinal dependent variables is based on experimental designs to obtain preferences of respondents to different stimuli: conjoint analysis. We first discuss the methodology involved in conjoint analysis and methods used to estimate the parameters of the conjoint models, i.e., monotone analysis of variance (MONANOVA). We then discuss a choice probability model that takes into consideration the ordinal property of the dependent variable, the ordered probit model.

9.1 Conjoint Analysis: MONANOVA

In the conjoint problem, preference responses to stimuli are obtained. These stimuli are designed to represent a combination of characteristics or attributes. Therefore, we start by discussing the design itself that defines the independent or the predictor variables and the manners in which the combination of attributes can be coded for analysis.

9.1.1 Effect Coding Versus Dummy Variable Coding

In a typical experimental setting, the independent variables that characterize the conditions of a cell or a stimulus are discrete categories or levels of attributes. For example, the color of the packaging of a product is red or yellow. It can be ordered (for example, a "low," "medium," or "high" value) or not (for example, colors).

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_9, © Springer Science+Business Media New York 2014

Table 9.1 A 2 × 2 factorial design

		A		
		$\overline{a}$	a	
B	$\overline{b}$	40.9(1)	47.8(a)	44.4
	b	42.4(b)	50.2(ab)	46.3
		41.6	49.0	45.3

Each combination of levels of all the attributes can correspond in principle to a stimulus, although responses to all the combinations may not be necessary. Two methods can be used to code these combinations of levels of attributes. Effect coding is the traditional method in experimental research using analyses of variance models. Dummy variables are typically used in regression analysis. We present each coding scheme and discuss the differences.

The coding principle is best described by taking an example of a two-by-two factorial design. This means that there are two factors in the experiment, each with two levels. For example, the stimulus may or may not have property A and may or may not have property B. This is illustrated in Table 9.1.

This 2^2 factorial design can easily be generalized to the 2^n design or any design $m \times n \times \cdots \times k$.

In Table 9.1, the stimulus possesses the attribute A or not. If it does, the condition is noted as a, and if it does not, it is noted as $\overline{a}$. The same two cases for attribute B are noted as b and $\overline{b}$. The combinations of levels of the two attributes lead to four cases (conditions) that can be labeled and described as follows:

(1) = Treatment combination that consists of the first level of all factors
(a) = Treatment combination that consists of the second level of the first factor and the first level of the second factor
(b) = Treatment combination that consists of the first level of the first factor and the second level of the second factor
(ab) = Treatment combination that consists of the second level of the two factors

The labels for these cases are shown in the cells of Table 9.1. Assuming that the various stimuli are evaluated on an interval scale response measure, the values, which are also shown in the cells of Table 9.1, are the average ratings provided by respondents in each of these conditions. Assuming that the number of respondents in each cell is the same, one can derive the grand mean rating, the main effects of each attribute or factor, and the specific incremental effect of the combination of A and B.

The grand mean is the average value across the four cells:

$$M = \text{Grand Mean} = \frac{1}{4}(ab + a + b + (1)) \tag{9.1}$$

The main effect of A is the average of the effect of the presence of A (i.e., the difference in the ratings whether A is present or not) across the two conditions determined by whether B is present or not. If B is present, the effect of A is (ab)−(b); if B is not present, it is (a)−(1) or

$$A(\text{Main Effect of } A) = \frac{1}{2}[\{(ab) - (b)\} + \{(a) - (1)\}] \tag{9.2}$$

Similarly, the main effect of B is the average of the effect of the presence of B (i.e., the difference in the ratings whether B is present or not) across the two conditions determined by whether A is present or not. If A is present, the effect of B is $(ab)-(a)$; if B is not present, it is $(b)-(1)$ or

$$B(\text{Main Effect of } B) = \frac{1}{2}[\{(ab) - (a)\} + \{(b) - (1)\}] \tag{9.3}$$

The joint effect of A and B beyond the main effects of A and B is given by the difference between the value of the criterion variable when both effects are present and its value when none are present (i.e., $(ab)-(1)$), after removing the main effect of A (i.e., $(a)-(1)$) and the main effect of B (i.e., $(b)-(1)$):

$$\begin{aligned} AB &= [\{(ab) - (1)\} - \{(b) - (1)\} - \{(a) - (1)\}] \\ &= [(ab) - (b) - (a) + (1)] \end{aligned} \tag{9.4}$$

Using the data in Table 9.1

(1) = 40.9
(ab) = 50.2
(a) = 47.8
(b) = 42.4

Therefore, using Eqs. (9.2), (9.3), and (9.4)

$$A = \frac{1}{2}[50.2 - 42.4 + 47.8 - 40.9] = \frac{1}{2}(7.8 + 6.9) = 7.4$$

$$B = \frac{1}{2}[50.2 - 47.8 + 42.4 - 40.9] = \frac{1}{2}(2.4 + 1.5) = 1.9$$

$$AB = [50.2 - 42.4 - 47.8 + 40.9] = 0.9$$

These effects can easily be computed using a linear model where the independent variables are coded using a specific scheme. The coding scheme is different depending on whether the effects are coded directly (effect coding) or whether the levels are coded (dummy coding).

9.1.1.1 Effect Coding

A variable is created for each factor, for example x_1 for factor A and x_2 for factor B. We first present the coding scheme with two levels, and then with more than two levels.

Effect Coding with Two Levels

Let us assume a factor with two levels. The upper level is coded "+1" and the lower level "−1."

Therefore, a stimulus (a cell) is represented by the vector $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, which for the four cells in Table 9.1 gives the following combinations:

$$\begin{array}{cc} 1\begin{pmatrix} -1 \\ -1 \end{pmatrix} & a\begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ b\begin{pmatrix} -1 \\ 1 \end{pmatrix} & ab\begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{array}$$

A main effect model can be represented by the linear model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{9.5}$$

The ratings of the individual cells can then be obtained by combining the values of x_1 and x_2:

	x_1	x_2
(1)	-1	-1
(a)	1	-1
(b)	-1	1
(ab)	1	1

For each cell, this leads to the equations

$$\begin{array}{ll} (1) & y = \beta_0 - \beta_1 - \beta_2 \\ (a) & y = \beta_0 + \beta_1 - \beta_2 \\ (b) & y = \beta_0 - \beta_1 + \beta_2 \\ (ab) & y = \beta_0 + \beta_1 + \beta_2 \end{array}$$

The effects of each factor are therefore represented by the values of the βs:

$$\begin{aligned} A &= \frac{1}{2}(\beta_0 + \beta_1 - \beta_2) - (\beta_0 - \beta_1 - \beta_2) + (\beta_0 + \beta_1 + \beta_2) - (\beta_0 - \beta_1 + \beta_2) \\ &= \beta_1 - \beta_2 + \beta_1 + \beta_2 + \beta_1 + \beta_2 + \beta_1 - \beta_2 \\ &= 2\beta_1 \\ B &= \beta_1 + \beta_2 - (\beta_1 - \beta_2) = \beta_1 + \beta_2 - \beta_1 + \beta_2 - \beta_1 + \beta_2 - (-\beta_1 - \beta_2) \\ &= -\beta_1 + \beta_2 + \beta_1 + \beta_2 \\ &= 2\beta_2 \end{aligned}$$

Table 9.2 Linear effect coding for three-level variable

Level	1	2	3
Coding	−1	0	+1

Table 9.3 Quadratic effect coding for three-level variable

Level	1	2	3
Coding	+1	−2	+1

Effect Coding with More Than Two Levels

When more than two levels are involved, the coding scheme depends on the assumptions made about the functional form of the relationship between the factor (independent variable) and the dependent variable. This issue obviously does not arise in the case of only two levels.

We present below the case of three levels of a variable. The effects can be coded to reflect either a linear or a nonlinear relationship.

Linear Effect Coding

Let us consider first the coding scheme for a linear effect. Such a coding is represented in Table 9.2.

It can be seen that the difference between level one and level two is the same as the difference between level two and level three, i.e., one unit. The difference between level one and level three is twice the difference between level one and level two. Therefore, the effect is linear.

Nonlinear Effect Coding

The coding of nonlinear effects varies depending on the functional form that the researcher wants to represent and test. Table 9.3 shows the coding scheme for a quadratic form.

The shape of the function shows symmetry around level two, and the values depend on the coefficient that multiplies this variable. Furthermore, a positive value of the coefficient would imply a decreasing and then increasing function, and vice versa for a negative coefficient.

The coding scheme can become quite complex. Table 9.4 provides the appropriate schemes for more than three levels.

9.1.1.2 Dummy Variable

Dummy coding corresponds to creating a variable for each level of each factor minus one. Therefore, for a design where a factor has three levels, two variables are created: variable x_1 takes the value 0 for level one and level three, and 1 for level

Table 9.4 Coefficient of orthogonal polynomials

Number of levels	Polynomial	Coefficients (d_i)									$\sum d_i^2$
3	Linear	−1	0	1							2
	Quadratic	1	−2	1							6
	Linear	−3	−1	1	3						20
4	Quadratic	1	−1	−1	1						4
	Cubic	−1	3	−3	1						20
	Linear	−2	−1	0	1	2					10
5	Quadratic	2	−1	−2	−1	2					14
	Cubic	−1	2	0	−2	1					10
	Quartic	1	−4	6	−4	1					70
	Linear	−5	−3	−1	1	3	5				70
6	Quadratic	5	−1	−4	−4	−1	5				84
	Cubic	−5	7	4	−4	−7	5				180
	Quartic	1	−3	2	2	−3	1				28
	Linear	−3	−2	−1	0	1	2	3			28
7	Quadratic	5	0	−3	−4	−3	0	5			84
	Cubic	−1	1	1	0	−1	−1	1			6
	Quartic	3	−7	1	6	1	−7	3			154
	Linear	−7	−5	−3	−1	1	3	5	7		168
8	Quadratic	7	1	−3	−5	−5	−3	1	7		168
	Cubic	−7	5	7	3	−3	−7	−5	7		264
	Quartic	7	−13	−3	9	9	−3	−13	7		616
	Quintic	−7	23	−17	−15	15	17	−23	7		2,184
	Linear	−4	−3	−2	−1	0	1	2	3	4	60
9	Quadratic	28	7	−8	−17	−20	−17	−8	7	28	2,772
	Cubic	−14	7	13	9	0	−9	−13	−7	14	990
	Quartic	14	−21	−11	9	18	9	−11	−21	14	2,002
	Quintic	−4	11	−4	−9	0	9	4	−11	4	468
	Linear	−9	−7	−5	−3	−1	1	3	5	7	9 330
10	Quadratic	6	2	−1	−3	−4	−4	−3	−1	2	6 132
	Cubic	−42	14	35	31	12	−12	−31	−35	−14	42 8,580
	Quartic	18	−22	−17	3	18	18	3	−17	−22	18 2,860
	Quintic	−6	14	−1	−11	−6	6	11	1	−14	6 780

Adapted from Fisher and Yates, *Statistical Tables for Biological*, *Agricultural and Medical Research*, published by Oliver and Boyd Ltd., Edinburgh (Table 23).

two, and x_2 takes the value 0 for level one and level two, and 1 for level three. This implies that a separate coefficient will be estimated for each level, relative to the reference cell where all the dummy variables are 0.

9.1.1.3 Decomposing the Effects in a Regression Model

Let us assume the following model:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i \tag{9.6}$$

where the variables are coded (−1, +1) and each group is balanced because $N/2$ observations are coded −1 and $N/2$ observations are coded +1.

The dependent variable y_i is assumed to be mean centered or to have a mean of zero.

The three variables are orthogonal so that the effects can be analyzed independently. Indeed, it can be shown that the interaction term is independent of the other effects.

The covariance between the product term of two variables x_1 and x_2 with one of its components x_1 is

$$\mathrm{V}[x_1, x_1x_2] = \mathrm{V}[x_1x_2]\mathrm{E}[x_1] + \mathrm{E}\Big[(x_1 - \overline{x}_1)^2(x_2 - \overline{x}_2)\Big] + \mathrm{E}[x_2]\mathrm{V}[x_1] \tag{9.7}$$

In ANOVA, the mean of the two variables that are coding the effects is zero. Consequently, the expression reduces to

$$\mathrm{V}[x_1, x_1x_2] = \mathrm{V}[x_1x_2].0 + \mathrm{E}\Big[(x_1 - \overline{x}_1)^2(x_2 - \overline{x}_2)\Big] + 0.\mathrm{V}[x_1] \tag{9.8}$$

or

$$\mathrm{V}[x_1, x_1x_2] = \mathrm{E}\Big[(x_1 - \overline{x}_1)^2(x_2 - \overline{x}_2)\Big] \tag{9.9}$$

But in ANOVA, the covariance of the two variables coding the effects is also zero (they are independent). Therefore,

$$\mathrm{E}\Big[(x_1 - \overline{x}_1)^2(x_2 - \overline{x}_2)\Big] = \mathrm{E}\big[{x_1}^2x_2\big] = 0 \tag{9.10}$$

Therefore,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} = \frac{\sum_{i|x_i>0}^{N} y_i - \sum_{i|x_i<0}^{N} y_i}{N} = 2(\overline{y}_2 - \overline{y}_1) \tag{9.11}$$

where $\overline{y}_1$= the mean of the dependent variable over the observations when x_1 is coded −1, and $\overline{y}_2$= the mean of the dependent variable over the observations when x_1 is coded +1.

This means that the coefficient of x_1 can be interpreted as the difference in group means due to that variable.

9.1.1.4 Comparing Effect Coding and Dummy Coding

The two coding schemes do not give identical results because, as shown in the presentation above, effect coding places a restriction on the relationship that does not apply to dummy variable coding. Consequently, like any restricted form of a relationship compared to its unrestricted form, a test of the appropriateness of the restriction can be performed. The two approaches can consequently be combined to perform tests about the functional forms.

In summary, effect coding is appropriate when testing for the significance of the effect of a variable (conditionally, on assuming a specific form of the relationship). Dummy coding is used to estimate and test the effects of each level of a variable independently from the other levels.

9.1.2 Design Programs

A particularity of conjoint analysis concerns the generation of the experimental design itself. Recently, several companies have developed PC-based software for generating stimuli reflecting the combinations of the levels of attributes. Two such software packages are Conjoint Designer by Bretton-Clark and Consurv by IMS Inc. Each of these packages offers similar services that, once the attributes and their levels are determined, generate the combination of the attributes in the form of the description of the stimuli, enable the entry of the data by respondents, and analyze the data.

9.1.3 Estimation of Part-Worth Coefficients

In Sect. 9.1.1, we discussed one of the characteristics of conjoint analysis: the specific nature of the independent variables. The other characteristic of conjoint analysis concerns the rank-ordered nature of the dependent variable. Although the term "conjoint" has recently been used more broadly, these two characteristics were initially what distinguished conjoint analysis from other methodologies. MONANOVA was developed as an appropriate methodology for estimating the effects of variables using the rank-ordered nature of the dependent variable. More recently, as conjoint studies have been successfully developed in industry, the simpler ordinary least squares (OLS) estimation has replaced the use of MONANOVA. This is due to not only the simplicity of OLS but also two other factors: (1) the robustness of OLS that gives generally similar results to those obtained from MONANOVA and (2) the increased usage of ratings instead of rankings for the dependent variables.

Table 9.5 Example of input data for a 2 × 3 design

		Second factor		
	Levels	1	2	3
First factor	1	δ_{11}	δ_{12}	δ_{13}
	2	δ_{21}	δ_{22}	δ_{23}

First, we present MONANOVA and illustrate the estimation using PC-MDS running under the Windows operating system. PC-MDS is one of several proprietary software packages that offer the ability to estimate MONANOVA models (http://www.surveypro.com/info/pcmds.html). XLSTAT is another package that uses Microsoft Excel running under Windows or OS Mac (http://www.xlstat.com/en/download.html). Then, we show how to perform OLS estimations using the SAS GLM procedure and STATA commands.

9.1.3.1 MONANOVA

Monotone analysis of variance is an estimation procedure based on an algorithm that transforms the dependent variable using a monotonic transformation so that the data can best be explained by a linear model of main effects of the independent variables or factors. More formally, let the data be represented by the set of values $\{\delta_{ij}\}$, each corresponding to the evaluation of alternative j by individual i ($i = 1, \ldots, I; j = 1, \ldots, J$). For each individual, the data are presented in a separate table, as shown in Table 9.5.

The objective is, therefore, to estimate the main effects of each factor to best fit the relationship:

$$f(\delta_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \varepsilon_{ij} \tag{9.12}$$

where $f(.)$ is a monotonic transformation of the rank-ordered dependent variable and x_1 and x_2 are the variables representing the main effects of the two factors using effect coding.

The monotone transformations are performed using an algorithm to improve the fit.

9.1.3.2 OLS Estimation

The GLM procedure found in SAS automatically creates the dummy variables that correspond to the design. When a variable is defined as a discrete variable using the CLASS function, the levels of the variable are automatically generated with the proper dummy variables. The model is linear and the estimation follows the OLS estimation described in Chap. 5.

It remains that MONANOVA is technically more appropriate when rank data are obtained and used as a dependent variable. This is particularly important for

academic research where technically inappropriate methods should not be used, even if they provide generally robust results. Obviously, the use of ratings makes OLS a perfectly appropriate methodology.

9.2 Ordered Probit

Ordered probit modeling is a relatively recent approach to analyzing rank-ordered dependent variables (McKelvey and Zavoina 1975). Let us assume that there exists an unobserved variable y that can be expressed as a linear function of a single predictor variable x. Furthermore, the variable y is not observed; only discrete levels of that variable can be observed (e.g., levels one, two, and three).

Figure 9.1 illustrates the case of a trichotomous dependent variable (observed variable) with a single independent variable.

It is important to distinguish between the theoretical dependent variable y and the observed dependent variable z, which, in the example of Fig. 9.1, takes three possible values.

The variable y is an interval scale variable and, if we could observe it, it would fit a linear model $\mathbf{y} = \mathbf{x\beta} + \mathbf{u}$.

The variable z is ordinal and generally presents M observed response categories $R_1, \ldots, R_M$.

The model of the unobserved dependent variable y follows the usual linear model assumptions with multiple explanatory variables $\mathbf{X}$:

$$\mathbf{y}=\mathbf{X\beta}+\mathbf{u} \tag{9.13}$$

with

$$\mathbf{u} \sim N\left(0, \sigma^2 I\right) \tag{9.14}$$

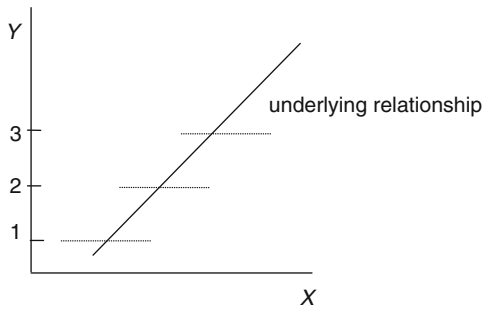


Fig. 9.1 The underlying linear relationship of the ordered probit model

We define $M + 1$ real numbers $\mu_0, \ldots, \mu_M$ with the following prespecified values:

$$\mu_0 = -\infty$$
$$\mu_M = +\infty$$

These values are rank ordered such that $\mu_0 \leq \mu_1 \leq \cdots \leq \mu_M$.

Let us consider an individual observation i. The value of the dependent variable z_{ij} will be 1 if the underlying unobserved variable falls within the values of y_i in the range of $[\mu_{j-1}, \mu_j]$. This can be expressed as

$$\mu_{j-1} < y_i \leq \mu_j \Leftrightarrow z_{ij} = 1; \quad \forall k \neq j: \ z_{ik} = 0 \tag{9.15}$$

Let us focus our attention on the interval in which the value of y_i falls:

$$\mu_{j-1} < y_i \leq \mu_j \tag{9.16}$$

We can replace the unobserved variable by the linear function of observed variables that determines it:

$$\mu_{j-1} < \mathbf{x}_i\boldsymbol{\beta} + u_i \leq \mu_j \tag{9.17}$$

Subtracting the deterministic component from the boundaries, we obtain

$$\mu_{j-1} - \mathbf{x}_i\boldsymbol{\beta} < u_i \leq \mu_j - \mathbf{x}_i\boldsymbol{\beta} \tag{9.18}$$

We can now standardize the values by dividing each element of the inequality by the standard deviation of the error term:

$$\frac{\mu_{j-1} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} < \frac{u_i}{\sigma} \leq \frac{\mu_j - \mathbf{x}_i\boldsymbol{\beta}}{\sigma} \tag{9.19}$$

The central element is a random variable with the normal distribution:

$$\frac{u_i}{\sigma} \sim N(0, 1) \tag{9.20}$$

We can, therefore, write the probability that this variable is within the range given by Eq. (9.19) by subtracting the cumulative density functions at the upper and lower levels:

$$P\left[z_{ij} = 1\right] = \phi\left[\frac{\mu_j - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right] - \phi\left[\frac{\mu_{j-1} - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right] \tag{9.21}$$

where ϕ is the cumulative density function:

$$\phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\Pi}} e^{-x^2/2} dx \tag{9.22}$$

In order to identify the model, we need to impose the restrictions

$$\mu_1 = 0$$
$$\sigma = 1$$

The first restriction has no consequence and the unit variance of the unobserved variable simply standardizes that variable. Consequently, Eq. (9.21) reduces to

$$P\left[z_{ij} = 1\right] = \phi\left[\mu_j - \mathbf{x}_i\boldsymbol{\beta}\right] - \phi\left[\mu_{j-1} - \mathbf{x}_i\boldsymbol{\beta}\right] \tag{9.23}$$

The parameters that need to be estimated are

$$\underset{K\times 1}{\boldsymbol{\beta}}\ ; \mu_2, \ldots, \mu_{M-1}$$

This means that there are $(K + M-2)$ parameters to be estimated.
The estimation is obtained by maximum likelihood.
Let

$$y_{ij} = \mu_j - \mathbf{x}_i\boldsymbol{\beta} \tag{9.24}$$

and, for simplification of the notation,

$$\phi_{i,j} = \phi\left(y_{ij}\right) \tag{9.25}$$

Then, the probability of z_{ij} being in the interval $[\mu_{j-1}, \mu_j]$ is

$$P\left[z_{ij} = 1\right] = \phi_{i,j} - \phi_{i,j-1} \tag{9.26}$$

Consequently, the likelihood of observing all the values of Z for all the observations in the data set is

$$\mathbf{L} = \mathbf{L}(\mathbf{z}|\boldsymbol{\beta}; \mu_2, \ldots \mu_{M-1}) \tag{9.27}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{M} \left(\phi_{i,j} - \phi_{i,j-1}\right)^{z_{ij}} \tag{9.28}$$

The logarithm of the likelihood is

$$\boldsymbol{\ell} = Ln\mathbf{L} = \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} Ln\left(\phi_{i,j} - \phi_{i,j-1}\right) \tag{9.29}$$

The estimation problem consists in finding the values of the parameters that maximize the logarithm of the likelihood function ℓ, subject to the inequality constraints about the values of μs, i.e.:

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{M-1}$$

One potential issue is that it is not always clear whether or not the dependent variable is ordered. The question, then, is whether one is better off using an ordered or an unordered model.

On the one hand, using an ordered model assumption when the true model is unordered creates a bias of the parameter estimates. On the other hand, using an unordered model when the true model is ordered does not create a bias but a loss of efficiency rather than consistency (Amemiya 1985, p. 293). Consequently, if the data are indeed ordered, the efficient and unbiased estimator will be provided by the ordered model. Using an unordered model may lead to parameters that are not significant but that would have been significant had the most efficient model been used. Of course, this is not an issue if all the parameters are significant. Using an ordered model if the data are not ordered is more risky because the parameter estimates are biased. Consequently, unless there is a strong theoretical reason for using an ordered model, it is best to use a non-ordered model when the order property of the dependent variable is not clearly proven.

9.3 Examples

9.3.1 Example of MONANOVA Using PC-MDS and XLSTAT

To illustrate the use of the PC-MDS software for MONANOVA analysis, we take the example of a $2 \times 2 \times 2$ design where the data are as given in Table 9.6.

The MONANOVA program used to analyze data that are structured as in Fig. 9.6 is run by clicking on the monanova.exe file from Windows Explorer. The data as well as the information about the run are contained in an input file. An example of a full input file is shown in Fig. 9.2. The first line of this input file contains the parameters that characterize the analysis. The meaning of the numbers on that first line is detailed in Table 9.7.

Table 9.6 Example of data for data entry using PC-MDS MONANOVA (a 2^3 design)

		Third factor			
		Second factor		Second factor	
	Level	1	2	1	2
First factor	1	x_{111}	x_{121}	x_{112}	x_{122}
	2	x_{211}	x_{221}	x_{212}	x_{222}

```
3  2  2  2  1
(8F10.2)
98.18     65.62     39.97     7.41      87.08     54.52     28.86     .0
```

Fig. 9.2 Example of input file for MONANOVA using PC-MDS (examp9-1.dat)

Table 9.7 Parameter line for reading data shown in Table 9.6

Parameter line	3	2	2	2	1
	# of factors	# of levels of first factor	# of levels of second factor	# of levels of third factor	# of replications

The second line corresponds to the format in which the data can be read using FORTRAN conventions.

The third line (and subsequent lines if there is more than one replication) corresponds to the data line(s). The data must be entered in a specific sequence. This sequence is best described through an example. In our $2 \times 2 \times 2$ example, the indices of the x variable are such that the first index represents the level on the first factor, the second represents the level on the second factor, and the third the level on the third factor. The sequence should then appear as

$$\begin{array}{cccc} 111 & 112 & 121 & 122 \\ 211 & 212 & 221 & 222 \end{array}$$

The results of the MONANOVA analysis are shown in Fig. 9.3.

The utilities for the levels within each factor are shown under the heading "UTILITIES OUTPUT FOR LEVELS WITHIN FACTORS" (highlighted in grey in Fig. 9.3).

For illustrating MONANOVA with XLSTAT, we use the responses of several individuals to a conjoint analysis regarding academic job preferences. Ten individuals provide their rankings of several profiles that are a full factorial combination of the various levels of three factors: compensation (higher vs. average), research reputation (excellent vs. good), and geographical location (North America, Europe, or Asia). The data are simply entered into an Excel spreadsheet. The beginning of the spreadsheet is shown in Fig. 9.4: the rows correspond to the responses of individuals to the profiles and the columns represent the profile descriptions that will be used as predictor variables and the corresponding profile ranking.

In the MONANOVA analysis module within the XLSTAT-Conjoint menu, a simple dialog box appears that is filled in using the cell location of the dependent and explanatory variables. This dialog box is shown in Fig. 9.5.

Drawn from the detailed statistical output that appears in a separate worksheet, the standardized utilities are plotted as shown in Fig. 9.6.

The analysis pooled the data across the ten respondents. MONANOVA can also be used in XLSTAT from the Conjoint Analysis menu to be performed by a respondent. When we know the utilities of each individual respondent, we can then

```
                              M O N A N O V A
                        MONOTONE  ANALYSIS  OF  VARIANCE
                          WRITTEN BY DR. J. B. KRUSKAL
                                PC-MDS VERSION




ANALYSIS TITLE:              Monanova
DATA IS READ FROM FILE:      examp9-1.dat
OUTPUT FILE IS:              examp9-1.out


INPUT DATA FILE PARAMETERS:  3  2  2  2  1
INPUT FORMAT:          (8F10.2)


SEQ. NO.   DATA      SUBSCRIPTS

    1      98.18000    1    1    1
    2      65.62000    1    1    2
    3      39.97000    1    2    1
    4       7.41000    1    2    2
    5      87.08000    2    1    1
    6      54.52000    2    1    2
    7      28.86000    2    2    1
    8        .00000    2    2    2

HISTORY OF COMPUTATION.

ITERAT STRESS  SRAT  SRTAVG CAGRGL  COSAV  ACSAV    GRMAG   GRMULT    STEP
   0   .000   .0000  1.2000   .000   .000   .200   .00000   .00000  .00000

ZERO STRESS WAS REACHED
MINIMUM WAS ACHIEVED
SATISFACTORY STRESS WAS REACHED
FINAL CONFIGURATION HAS STRESS OF     .0 PERCENT.

Monanova

UTILITIES OUTPUT FOR LEVELS WITHIN FACTORS


  2   .266  -.266
  2  1.498 -1.498
  2   .827  -.827


              .   5.8908.  27.4904.  49.0900.  70.6896.  92.2892.
           -4.9090   16.6906   38.2902   59.8898   81.4894  103.0890
            *.****.****.****.****.****.****.****.****.****.****.*
      2.85 ..                                                  ..   2.85
 T    2.64 ..                                                0 ..   2.64
 H    2.43 ..                                                   ..   2.43
 E    2.22 ..                                                   ..   2.22
      2.01 ..                                              0    ..   2.01
 X    1.80 ..                                                   ..   1.80
      1.58 ..                                                   ..   1.58
 A    1.37 ..                                                   ..   1.37
 R    1.16 ..                                                   ..   1.16
 E     .95 ..                                    0              ..    .95
       .74 ..                                                   ..    .74
```

Fig. 9.3 Output file for MONANOVA example (examp9-1.out)

```
  T       .53 ..                                                    ..    .53
  H       .32 ..                              0                     ..    .32
  E       .11 ..                                                    ..    .11
         -.11 ..                                                    ..   -.11
  L      -.32 ..                         0                          ..   -.32
  I      -.53 ..                                                    ..   -.53
  N      -.74 ..                                                    ..   -.74
  E      -.95 ..                    0                               ..   -.95
  A     -1.16 ..                                                    ..  -1.16
  R     -1.37 ..                                                    ..  -1.37
        -1.58 ..                                                    ..  -1.58
  M     -1.80 ..                                                    ..  -1.80
  O     -2.01 ..        0                                           ..  -2.01
  D     -2.22 ..                                                    ..  -2.22
  E     -2.43 ..                                                    ..  -2.43
  L     -2.64 ..  0                                                 ..  -2.64
        -2.85 ..                                                    ..  -2.85
              *.****.****.****.****.****.****.****.****.****.****.*
                .   5.8908.  27.4904.  49.0900.  70.6896.  92.2892.
             -4.9090   16.6906   38.2902   59.8898   81.4894  103.0890


 SEQ NO  DATA     LINEAR   MONOTONE MODELS
    1    98.180     2.592     2.592
    2    65.620      .937      .937
    3    39.970     -.405     -.405
    4     7.410    -2.059    -2.059
    5    87.080     2.059     2.059
    6    54.520      .405      .405
    7    28.860     -.937     -.937
    8      .000    -2.592    -2.592


                                ************

          SPEARMAN-S RANK DIFFERENCE CORRELATION COEFFICIENT (RD)
                             RD           =  1.000000

                             RD - SQUARED =  1.000000



 END MONANOVA RUN: Monanova
```

Fig. 9.3 (continued)

form clusters that correspond to market segments where each cluster (segment) has different preferences. Predictions of market share are also available as a function of the variations of offerings in the market.

9.3.2 Example of Conjoint Analysis with Interval Scale Rating Data

When the dependent variable is interval scale, the analysis is done with regression. The example in Fig. 9.7, which gives the sample input file used with SAS, provides a variant of the illustration in Fig. 9.6 where potential students rated different

	A	B	C	D	E	F
1	Individual	Observation	Compensation	Research Rep	Geographical Loc	Ranking
2	1	Profile1	Higher	Excellent	North America	1
3	1	Profile2	Higher	Excellent	Europe	2
4	1	Profile3	Higher	Excellent	Asia	3
5	1	Profile4	Higher	Good	North America	7
6	1	Profile5	Higher	Good	Europe	8
7	1	Profile6	Higher	Good	Asia	9
8	1	Profile7	Average	Excellent	North America	4
9	1	Profile8	Average	Excellent	Europe	5
10	1	Profile9	Average	Excellent	Asia	6
11	1	Profile10	Average	Good	North America	10
12	1	Profile11	Average	Good	Europe	11
13	1	Profile12	Average	Good	Asia	12
14	2	Profile1	Higher	Excellent	North America	2
15	2	Profile2	Higher	Excellent	Europe	1
16	2	Profile3	Higher	Excellent	Asia	3
17	2	Profile4	Higher	Good	North America	8
18	2	Profile5	Higher	Good	Europe	7
19	2	Profile6	Higher	Good	Asia	9
20	2	Profile7	Average	Excellent	North America	5
21	2	Profile8	Average	Excellent	Europe	4
22	2	Profile9	Average	Excellent	Asia	6
23	2	Profile10	Average	Good	North America	11
24	2	Profile11	Average	Good	Europe	10
25	2	Profile12	Average	Good	Asia	12
26	3	Profile1	Higher	Excellent	North America	3
27	3	Profile2	Higher	Excellent	Europe	2

Fig. 9.4 Excel spreadsheet for MONANOVA analysis in XLSTAT (examp9-1.xlsx)

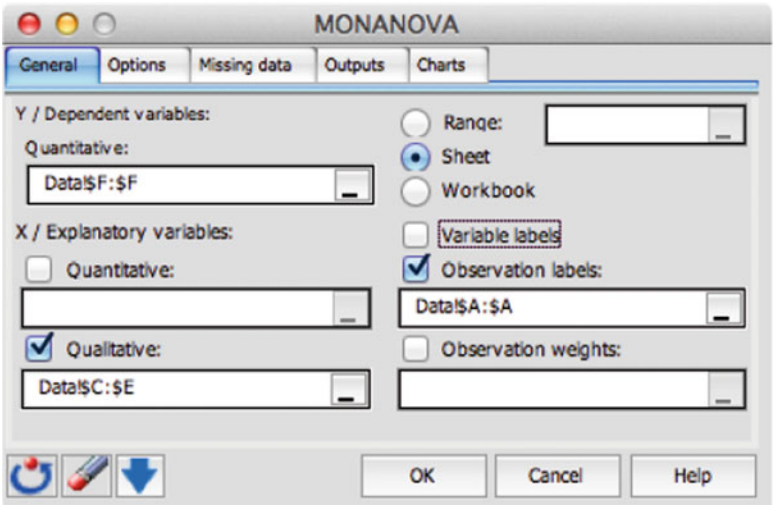


Fig. 9.5 Dialog box for MONANOVA analysis in XLSTAT (examp9-1.xlsx)

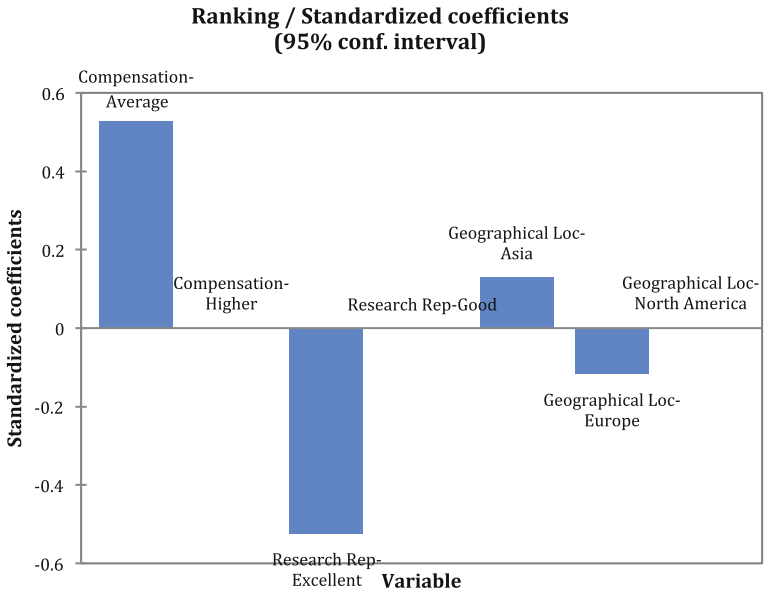


Fig. 9.6 XLSTAT output file for MONANOVA example (examp9-1.xlsx)

```
options ls=80;
DATA DATA1;
INFILE "C:\SAMD\Chapter9\Examples\Examp9-2.dat";
INPUT rating xidquant instruct resrep tearep prestige;
PROC glm;
CLASS xid quant instruct resrep tearep prestige;
MODEL rating = quant instruct resrep tearep prestige;
MEANS QUANT INSTRUCT RESREP TEAREP PRESTIGE;
estimate 'quant' quant 1-1;
estimate 'instr2 vs 1' instruct 1 -1 0;
estimate 'instr3 vs 1' instruct 1 0-1;
run;
```

Fig. 9.7 Example of input file for conjoint analysis using SAS (examp9-2.sas)

hypothetical schools. The hypothetical schools were described in terms of (1) being either (a) not very or (b) very quantitative; (2) using methods of instruction characterized by (a) the case method, (b) half case and half lectures, or (c) lectures only; (3) the research reputation of the faculty, which can be (a) low, (b) moderate, or (c) high; (4) the teaching reputation of the faculty, which can also be (a) low, (b) moderate, or (c) high; and (5) the overall prestige of the school as (a) one of the Ivy League colleges, (b) a private school but not part of the Ivy League, or (c) a state school. We then use the SAS procedure "glm" as highlighted in grey in Fig. 9.7. Within this highlighted area, discrete variables are identified by the "class" command, and the regression model specification is contained after the

"model" command. The command "means" requests that the means of the rating variable be displayed for each of the levels of the predictor variables. Finally, the last command highlighted in grey in Fig. 9.7 is the "estimate" command that is used to compare the means at different levels of a variable (i.e., to perform specific contrasts).

Fig. 9.8 gives the output of such analysis. The tests of significance of each factor are performed and then the marginal means of the dependent variable is shown for each level of each factor, one at a time. The example also illustrates the test for restrictions on the parameters such as for linear effects. The statistics at the bottom of the output in Fig. 9.8 show that the difference between levels 1 and 2 of instruction (i.e., "instr2 vs 1") is insignificant ($t = 0.42$) but the difference between levels 3 and 1 (i.e., "instr3 vs 1") is significant ($t = 2.04$).

In STATA, we use the "regress" command. Quantitative variables that have finite levels are converted and treated as categorical variables by specifying the prefix "i." attached to the variable. For example, the variable "quant" is treated as a discrete-level factor by specifying "i.quant." When the independent variables are all qualitative factors of that sort, the regression analysis is an ANOVA. The same

```
The SAS System

                        General Linear Models Procedure
                            Class Level Information

                      Class     Levels    Values

                      XID            9    1 2 3 4 5 6 7 8 9

                      QUANT          2    1 2

                      INSTRUCT       3    1 2 3

                      RESREP         3    1 2 3

                      TEAREP         3    1 2 3

                      PRESTIGE       3    1 2 3

                    Number of observations in data set = 162


                        General Linear Models Procedure

Dependent Variable: RATING
                                    Sum of            Mean
Source                    DF       Squares          Square   F Value     Pr > F

Model                      9     465.30941469     51.70104608     24.01     0.0001

Error                    152     327.33256062      2.15350369

Corrected Total          161     792.64197531

                  R-Square             C.V.        Root MSE          RATING Mean

                  0.587036         33.76876       1.4674821            4.3456790
```

Fig. 9.8 Output for GLM procedure using SAS example (examp9-2.lst)

```
Source                    DF        Type I SS     Mean Square   F Value     Pr > F

QUANT                      1       0.00308642      0.00308642      0.00     0.9699
INSTRUCT                   2      21.48504274     10.74252137      4.99     0.0080
RESREP                     2      75.94088319     37.97044160     17.63     0.0001
TEAREP                     2     332.45486111    166.22743056     77.19     0.0001
PRESTIGE                   2      35.42554123     17.71277062      8.23     0.0004

Source                    DF      Type III SS     Mean Square   F Value     Pr > F

QUANT                      1       0.02816755      0.02816755      0.01     0.9091
INSTRUCT                   2      14.50887457      7.25443728      3.37     0.0370
RESREP                     2      64.20418593     32.10209297     14.91     0.0001
TEAREP                     2     302.01431665    151.00715833     70.12     0.0001
PRESTIGE                   2      35.42554123     17.71277062      8.23     0.0004

                        General Linear Models Procedure

                  Level of          ------------RATING-----------
                  QUANT         N       Mean                SD

                  1           108     4.34259259       2.27609289
                  2            54     4.35185185       2.12049663

                  Level of          ------------RATING-----------
                  INSTRUCT      N       Mean                SD

                  1            63     4.30158730       2.35300029
                  2            45     3.86666667       1.64593162
                  3            54     4.79629630       2.41363759

                  Level of          ------------RATING-----------
                  RESREP        N       Mean                SD

                  1            54     3.37037037       1.85610971
                  2            54     4.72222222       1.99448926
                  3            54     4.94444444       2.46037784

                  Level of          ------------RATING-----------
                  TEAREP        N       Mean                SD

                  1            54     2.46296296       1.29895405
                  2            54     4.62962963       1.61708212
                  3            54     5.94444444       2.08694655

                  Level of          ------------RATING-----------
                  PRESTIGE      N       Mean                SD

                  1            54     4.94444444       2.34252457
                  2            45     4.95555556       2.22542698
                  3            63     3.39682540       1.75554130

                        General Linear Models Procedure

Dependent Variable: RATING

                                          T for H0:    Pr > |T|   Std Error of
Parameter                  Estimate    Parameter=0                  Estimate

quant                   0.02821743           0.11      0.9091     0.24672641
instr2 vs 1             0.13109512           0.42      0.6762     0.31324315
instr3 vs 1            -0.57290168          -2.04      0.0432     0.28097845
```

Fig. 9.8 (continued)

analysis as in Fig. 9.8 can then be easily performed using the "regress" command in STATA. Fig. 9.9 lists a dictionary file to describe the format in which the data in file examp9-2.dat should be read.

The commands to run the analysis in STATA are shown in Fig. 9.10.

```
infile dictionary using "/users/gatignon/Documents/WORK_STATA/SAMD/Chapter9_MONANOVA-
OL/Examp9-2.dat" {
_lines(2)
_line(1)
                        rating xid quant instruct resrep tearep
_line(2)
                                        prestige
}
```

Fig. 9.9 STATA dictionary file to read data in examp9-2.dat (examp9-2.dct)

```
infile using "/users/gatignon/Documents/WORK_STATA/SAMD/Chapter9_MONANOVA-OL/Examp9-
2_Mac.dct", clear
anova rating i.quant i.instruct i.resrep i.tearep i.prestige
regress rating i.quant i.instruct i.resrep i.tearep i.prestige
margins i.quant i.instruct i.resrep i.tearep i.prestige
lincom _b[1.quant] -_b[2.quant]
lincom _b[1.instruct] - b[2.instruct]
lincom _b[1.instruct]- _b[3.instruct]
glm rating i.quant i.instruct i.resrep i.tearep i.prestige
```

Fig. 9.10 STATA command file to analyze data in examp9-2.dat (examp9-2_Mac.do)

The "anova" command (highlighted in grey in Fig. 9.10) provides the overall test of significance of each factor. The "regress" command performs a regression analysis where the coefficients of each level of each factor (except for the first level of each factor that serves as the base) are estimated. The highlighted "margins" command provides the estimates of the cell means. Finally, the "lincom" command is used to test the significance of the difference of level effects (or any linear combination). For example, "lincom _b[1.instruct] - _b[3.instruct]" (highlighted in grey in Fig. 9.10) compares the effect of the third level of the variable "instruct" with the first level of that variable. The last line of commands uses a "glm" command in STATA that parallels the "glm" procedure in SAS. In this case, the results are the same as in regression. The results of running these commands are shown in Fig. 9.11.

The command lines are highlighted in grey in Fig. 9.11 to better identify each section of the analysis. The results are obviously identical to those obtained with the "glm" procedure in SAS.

9.3.3 Example of Ordered Probit Analysis Using LIMDEP

The use of ordered probit is illustrated with two examples, one with LIMDEP and the other with STATA. The input file for LIMDEP, which enables the estimation of an ordered probit model, is straightforward (Fig. 9.12). The only difference with the statements for a logit-type model specification is the use of the command "ORDERED" (the command is highlighted in grey in the figure, as well as the commands that are identical to those used for the logit model explained in Chap. 8, i.e., "lhs" and "rhs"). It should be noted that the right side list of variables must

```
. anova rating i.quant i.instruct i.resrep i.tearep i.prestige

                          Number of obs =     162     R-squared     =  0.5870
                          Root MSE      = 1.46748     Adj R-squared =  0.5626

                  Source | Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  465.309415     9  51.7010461       24.01     0.0000

                   quant |  .028167554     1  .028167554        0.01     0.9091
              instruct   |  14.5088746     2  7.25443728        3.37     0.0370
                resrep   |  64.2041859     2   32.102093       14.91     0.0000
                tearep   |  302.014317     2  151.007158       70.12     0.0000
              prestige   |  35.4255412     2  17.7127706        8.23     0.0004

                Residual |  327.332561   152  2.15350369
              -----------+----------------------------------------------------
                   Total |  792.641975   161  4.92324208

. regress rating i.quant i.instruct i.resrep i.tearep i.prestige

      Source |       SS       df       MS              Number of obs =     162
-------------+------------------------------           F(  9,   152) =   24.01
       Model |  465.309415     9  51.7010461           Prob > F      =  0.0000
    Residual |  327.332561   152  2.15350369           R-squared     =  0.5870
-------------+------------------------------           Adj R-squared =  0.5626
       Total |  792.641975   161  4.92324208           Root MSE      =  1.4675

------------------------------------------------------------------------------
      rating |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     2.quant |  -.0282174   .2467264    -0.11   0.909    -.5156733    .4592384

    instruct |
           2 |  -.1310951   .3132432    -0.42   0.676    -.7499677    .4877775
           3 |   .5729017   .2809785     2.04   0.043     .0177743    1.128029

      resrep |
           2 |     1.2332   .2937212     4.20   0.000      .652897    1.813503
           3 |   1.477272   .2870521     5.15   0.000     .9101444    2.044399

      tearep |
           2 |   2.048015   .2937212     6.97   0.000     1.467712    2.628318
           3 |   3.384679   .2870521    11.79   0.000     2.817552    3.951806

    prestige |
           2 |    -.53749   .3144772    -1.71   0.089    -1.158801    .0838207
           3 |  -1.118305   .2759947    -4.05   0.000    -1.663586   -.5730242

       _cons |   2.070344   .3646644     5.68   0.000     1.349878    2.790809
------------------------------------------------------------------------------

. margins i.quant i.instruct i.resrep i.tearep i.prestige

Predictive margins                              Number of obs   =        162
Model VCE    : OLS

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       quant |
           1 |   4.355085   .1416227    30.75   0.000     4.077509     4.63266
           2 |   4.326867   .2008689    21.54   0.000     3.933172    4.720563

    instruct |
           1 |   4.191127    .193717    21.64   0.000     3.811449    4.570805
           2 |   4.060032    .229635    17.68   0.000     3.609956    4.510108
           3 |   4.764029    .200432    23.77   0.000     4.371189    5.156868

      resrep |
           1 |   3.442188    .203987    16.87   0.000     3.042381    3.841996
           2 |   4.675389   .2040573    22.91   0.000     4.275444    5.075334
           3 |    4.91946   .2008689    24.49   0.000     4.525764    5.313156
```

Fig. 9.11 STATA output of analysis for data in examp9-2.dat (examp9-2_Mac.log)

```
      tearep |
           1 |   2.534781   .203987    12.43   0.000     2.134974    2.934588
           2 |   4.582796   .2040573   22.46   0.000     4.182851    4.982741
           3 |    5.91946   .2008689   29.47   0.000     5.525764    6.313156
             |
    prestige |
           1 |   4.929878   .2027093   24.32   0.000     4.532575    5.327181
           2 |   4.392388   .2321975   18.92   0.000      3.93729    4.847487
           3 |   3.811573   .1891649   20.15   0.000     3.440817    4.182329
------------------------------------------------------------------------------

. lincom _b[1.quant] - _b[2.quant]

 ( 1)  1b.quant - 2.quant = 0

------------------------------------------------------------------------------
      rating |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0282174   .2467264     0.11   0.909    -.4592384    .5156733
------------------------------------------------------------------------------

. lincom _b[1.instruct] - _b[2.instruct]

 ( 1)  1b.instruct - 2.instruct = 0

------------------------------------------------------------------------------
      rating |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1310951   .3132432     0.42   0.676    -.4877775    .7499677
------------------------------------------------------------------------------

. lincom _b[1.instruct] - _b[3.instruct]

 ( 1)  1b.instruct - 3.instruct = 0

------------------------------------------------------------------------------
      rating |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.5729017   .2809785    -2.04   0.043    -1.128029   -.0177743
------------------------------------------------------------------------------

. glm rating i.quant i.instruct i.resrep i.tearep i.prestige

Iteration 0:   log likelihood = -286.84185

Generalized linear models                         No. of obs      =        162
Optimization     : ML                             Residual df     =        152
                                                  Scale parameter =   2.153504
Deviance         =  327.3325606                   (1/df) Deviance =   2.153504
Pearson          =  327.3325606                   (1/df) Pearson  =   2.153504

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   3.664714
Log likelihood   = -286.8418486                   BIC             =  -445.9821

------------------------------------------------------------------------------
             |                 OIM
      rating |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    2.quant  |  -.0282174   .2467264    -0.11   0.909    -.5117923    .4553574
             |
    instruct |
           2 |  -.1310951   .3132432    -0.42   0.676    -.7450404    .4828502
           3 |   .5729017   .2809785     2.04   0.041      .022194    1.123609
             |
      resrep |
           2 |     1.2332   .2937212     4.20   0.000     .6575172    1.808883
           3 |   1.477272   .2870521     5.15   0.000     .9146597    2.039883
             |
      tearep |
           2 |   2.048015   .2937212     6.97   0.000     1.472332    2.623698
           3 |   3.384679   .2870521    11.79   0.000     2.822067    3.947291
             |
    prestige |
           2 |    -.53749   .3144772    -1.71   0.087    -1.153854     .078874
           3 |  -1.118305   .2759947    -4.05   0.000    -1.659245   -.5773656
             |
       _cons |   2.070344   .3646644     5.68   0.000     1.355615    2.785073
------------------------------------------------------------------------------
```

Fig. 9.11 (continued)

```
read; file = Examp9-3.wks;
format = WKS ;
names
$

open; output = c:\SAMD\Chapter9\Examples\Examp9-3.out$
ORDERED; lhs = Rnk;
         rhs = ONE, MBA_rate, Div_rate, R_Drate $
close$
```

Fig. 9.12 Example of ordered probit estimation using LIMDEP (examp9-3.lim)

include "ONE" to specify a constant term. This particular example concerns the ranking of business schools (Rnk) as a function of ratings on the MBA programs (MBA_rate), the diversity of populations represented in the schools (Div_rate), and the ratings of the research activities of the schools (R_Drate).

Fig. 9.13 shows the results of this analysis where the main components are highlighted in grey. These include the log-likelihood at its maximum value and the parameter estimates.

Diversity (DIV_RATE) appears insignificant but the rating of the MBA program (MBA_RATE) as well as the rating of the school on research activities (R_DRATE) appear to strongly predict the overall ranking of the school.

For STATA, we use the same data as used for MONANOVA. The data file is an Excel spreadsheet ("examp9-4.xlxs") with the same structure as described in Fig. 9.4. The STATA commands are represented in Fig. 9.14.

The profiles in this data file have been entered as alphabetical characters (e.g., "Higher," "Average"). These levels of each factor need first to be converted into numerical values. We therefore generate new variables for the three factors ("Comp" for compensation, "Resrep" for research reputation, and "Loc" for geographical location) and replace the values by numbers. The ordered probit command is highlighted in grey. The dependent variable is the variable "Ranking" and the predictor variables are the three newly recreated factor variables. A prefix "i." is added in front of each of these factor variables in order to indicate the need to use appropriate automatically created dummy variables for each level.

The output is shown in Fig. 9.15.

These results show that the second level of compensation ("Average") has a significant positive marginal utility, which indicates that it contributes to the increase in the ranking (i.e., less preferred) relative to a "Higher" compensation. Similarly, having a "Good" (versus "Excellent") research reputation lowers the ranking. Finally, concerning the location effect, the difference between North America and Europe is not statistically significant, but Asia is less preferred as a whole. Further analysis could also be performed to identify individual differences that could be used as a basis for cluster/segmentation analysis.

The ordered logit model results can be compared to the ordered probit model. The command line is simply changed to ". ologit Ranking i.Comp i.Resrep i.Loc."

The same conclusions are drawn from such an analysis.

```
: LIMDEP Estimation Results                          Run log line    3  Page   1 :
 : Current sample contains      50 observations.                                 :
 +=============================================================================+

 +-----------------------------------------------------------------------------+
 | Dependent variable is binary, y=0 or y not equal 0                          |
 | Ordinary    least squares regression    Weighting variable = none           |
 | Dep. var. = Y=0/Not0 Mean=    .9000000000    , S.D.=    .3030457634         |
 | Model size: Observations =      50, Parameters =    4, Deg.Fr.=      46     |
 | Residuals:  Sum of squares= 1023.254148    , Std.Dev.=          4.71642     |
 | Fit:        R-squared=*********, Adjusted R-squared =       -241.21958     |
 | Diagnostic: Log-L =   -146.4149, Restricted(b=0) Log-L =      -10.7483     |
 |             LogAmemiyaPrCrt.=    3.179, Akaike Info. Crt.=       6.017     |
 +-----------------------------------------------------------------------------+
 +---------+-------------+----------------+--------+---------+----------+
 |Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
 +---------+-------------+----------------+--------+---------+----------+
  Constant  1.299653062      2.8185737          .461    .6447
  MBA_RATE -.1151966230E-02  .51232691E-01     -.022    .9821  52.493633
  DIV_RATE -.4042208781E-02  .12218127        -.033    .9736  9.5247202
  R_DRATE  -.8542081297E-02  .34109580E-01     -.250    .8023  35.200000

 Normal exit from iterations. Exit status=0.
 +=============================================================================+
 : LIMDEP Estimation Results                          Run log line    3  Page   2 :
 : Current sample contains      50 observations.                                 :
 +=============================================================================+

              +---------------------------------------------+
              | Ordered Probit Model                        |
              | Maximum Likelihood Estimates                |
              | Dependent variable                   RNK    |
              | Weighting variable                   ONE    |
              | Number of observations                50    |
              | Iterations completed                  27    |
              | Log likelihood function          -76.87438  |
              | Restricted log likelihood        -115.1293  |
              | Chi-squared                       76.50975  |
              | Degrees of freedom                   3      |
              | Significance level               .0000000  |
              |    Cell frequencies for outcomes            |
              |  Y Count Freq  Y Count Freq  Y Count Freq   |
              |  0     5 .100  1     5 .100  2     5 .100   |
              |  3     5 .100  4     5 .100  5     5 .100   |
              |  6     5 .100  7     5 .100  8     5 .100   |
              |  9     5 .100                              |
              +---------------------------------------------+

 +---------+-------------+----------------+--------+---------+----------+
 |Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
 +---------+-------------+----------------+--------+---------+----------+
          Index function for probability
  Constant  11.27611554      1.9470225        5.791    .0000
  MBA_RATE -.9184333969E-01  .14200317E-01    -6.468    .0000  52.493633
  DIV_RATE  .6944545143E-02  .33367796E-01      .208    .8351  9.5247202
  R_DRATE  -.8690702844E-01  .17888303E-01    -4.858    .0000  35.200000
          Threshold parameters for index
  Mu( 1)    1.132867414      .41874468        2.705    .0068
  Mu( 2)    2.318480730      .70743429        3.277    .0010
  Mu( 3)    3.227069878      .78748983        4.098    .0000
  Mu( 4)    3.929271873      .82592852        4.757    .0000
  Mu( 5)    4.474735177      .84409675        5.301    .0000
  Mu( 6)    5.000183482      .88623007        5.642    .0000
  Mu( 7)    5.573052169      .95533576        5.834    .0000
  Mu( 8)    6.311310116     1.0282479         6.138    .0000
```

Fig. 9.13 Output of ordered probit model using LIMDEP (examp9-3.out)

```
import excel "/Users/gatignon/Documents/WORK_STATA/SAMD/Chapter9_MONANOVA-OL/Examp9-
4.xlsx", sheet("Sheet1") firstrow clear
gen Comp=1 if Compensation=="Higher"
replace Comp=2 if Compensation=="Average"
gen Resrep=1 if ResearchRep=="Excellent"
replace Resrep=2 if ResearchRep=="Good"
gen Loc=1 if GeographicalLoc=="North America"
replace Loc=2 if GeographicalLoc=="Europe"
replace Loc=3 if GeographicalLoc=="Asia"
oprobit Ranking i.Comp i.Resrep i.Loc
```

Fig. 9.14 STATA commands for ordered probit model example (examp9-4.do)

```
. oprobit Ranking i.Comp i.Resrep i.Loc

Iteration 0:   log likelihood = -298.1888
Iteration 1:   log likelihood =-252.15288
Iteration 2:   log likelihood =-251.59271
Iteration 3:   log likelihood =-251.59193
Iteration 4:   log likelihood = -251.59193

Ordered probit regression                       Number of obs   =        120
                                                LR chi2(4)      =      93.19
                                                Prob > chi2     =     0.0000
Log likelihood = -251.59193                     Pseudo R2       =     0.1563

------------------------------------------------------------------------------
     Ranking |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      2.Comp |   1.365533   .2063363     6.62   0.000     .9611212    1.769945
    2.Resrep |   1.570412   .2121635     7.40   0.000     1.154579    1.986245
             |
         Loc |
          2  |  -.3640599   .2304488    -1.58   0.114    -.8157312    .0876114
          3  |   .5582931   .2338327     2.39   0.017     .0999894    1.016597
-------------+----------------------------------------------------------------
       /cut1 |  -.5730945   .2588788                     -1.080488   -.0657014
       /cut2 |   .0925388   .2402791                     -.3783997    .5634772
       /cut3 |   .5700556   .2382819                      .1030317    1.037079
       /cut4 |   .9128469   .2388882                      .4446345    1.381059
       /cut5 |   1.206877   .2410636                      .7344007    1.679353
       /cut6 |   1.508688   .2463593                      1.025833    1.991544
       /cut7 |   1.813262    .254846                      1.313773    2.312751
       /cut8 |   2.141778   .2663604                      1.619721    2.663835
       /cut9 |   2.534498   .2850483                      1.975814    3.093183
      /cut10 |   3.001398   .3108749                      2.392095    3.610702
      /cut11 |   3.611815   .3491821                       2.92743    4.296199
------------------------------------------------------------------------------
```

Fig. 9.15 STATA commands for ordered probit model example (examp9-4.log)

9.4 Assignment

1. Decide on an issue to be analyzed with a conjoint study and gather data yourself on a few (10–20) individuals. Make sure that at least one of the factors has more than two levels.

   Investigate issues concerned with the level of analysis and estimation procedures:

   Types of analysis

   Aggregate analysis
   Individual-level analysis

Estimation

GLM
Regression with dummy variables
Regression with effect coding
MONANOVA

2. Using data from the SURVEY (Appendix C, Chap. 14), choose a rank-ordered variable and develop a model to explain and predict this variable. Compare the multinomial logit model with the ordered logit or probit model. In addition, choose a categorical variable and illustrate the problem of using an ordered logit or probit model when it is not appropriate.

Bibliography

Basic Technical Readings

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MS: Harvard University Press [Chap. 9].

Cattin, P., Gelfand, A. E., & Danes, J. (1983). A simple Bayesian procedure for estimation in a conjoint model. *Journal of Marketing Research, 20*(1), 29–35.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research, 8*(3), 355–363.

Louviere, J. J. (1988). Conjoint analysis modeling of stated preferences. *Journal of Transport Economics and Policy, 22*(1), 93–119.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*, 103–120.

Application Readings

Beggs, S., Cardell, S., & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics, 16*, 1–19.

Bowman, D., & Gatignon, H. (1995). Determinants of competitor response time to a New product introduction. *Journal of Marketing Research, 32*(1), 42–53.

Bunch, D. S., & Smiley, R. (1992). Who deters entry? evidence on the Use of strategic entry deterrents. *The Review of Economics and Statistics, 74*(3), 509–521.

Chu, W., & Anderson, E. (1992). Capturing ordinal properties of categorical dependent variables: a review with applications to modes of foreign entry and choice of industrial sales force. *International Journal of Research in Marketing, 9*, 149–160.

Green, P. E. (1984). Hybrid models for conjoint analysis: an expository review. *Journal of Marketing Research, 21*(2), 155–169.

Green, P. E., Krieger, A. M., & Agarwal, M. K. (1991). Adaptive conjoint analysis: some caveats and suggestions. *Journal of Marketing Research, 28*(2), 215–222.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: issues and outlook. *Journal of Consumer Research, 5*(2), 103–123.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with applications for research and practice. *Journal of Marketing, 54*(4), 3–19.

Green, P. E., & Wind, Y. (1975). *A New Way to measure consumers' judgements* (pp. 107–117). July-August: Harvard Business Review.

Jain, D. C., Muller, E., et al. (1999). Pricing patterns of cellular phones and phonecalls: a segment-level analysis. *Management Science, 45*(2), 131–141.

Mahajan, V., Green, P. E., & Goldberg, S. M. (1982). A conjoint model for measuring self- and cross-price/demand relationships. *Journal of Marketing Research, 19*(3), 334–342.

Page, A. L., & Rosenbaum, H. F. (1987). Redesigning product lines with conjoint analysis: How sunbeam does it. *Journal of Product Innovation Management, 4*, 120–137.

Priem, R. L. (1992). An application of metric conjoint analysis for the evaluation of top managers' individual strategic decision making processes: a research note. *Strategic Management Journal, 13*, 143–151.

Rangaswami, A., & Richard Shell, G. (1997). Using computers to realize joint gains in negociations: toward an "Electronic Bargaining Table". *Management Science, 43*(8), 1147–1163.

Srinivasan, V., & Chan Su Park. (1997). Surprising robustness of the self-explicated approach to customer preference structure measurement. *Journal of Marketing Research, 34*(2), 286–291.

Wind, J., Green, P. E., Shifflet, D., & Scarbrough, M. (1989). Courtyard by Marriott: designing a hotel facility with consumer-based marketing models. *Interfaces, 19*(1), 25–47.

Chapter 10
Error in Variables: Analysis of Covariance Structure – Structural Equation Models

In this chapter, we bring together the notions of measurement error discussed in Chaps. 3 and 4 with the structural modeling of simultaneous relationships presented in Chap. 6. We demonstrate that a bias is introduced when estimating the relationship between two variables measured with error if that measurement error is ignored. We then present a methodology for estimating the parameters of structural relationships between variables that are not observed directly: analysis of covariance structures. We focus on the role of the measurement model as discussed in Chap. 4 with the confirmatory factor analytic model.

10.1 Impact of Imperfect Measures

In this section, we discuss the bias introduced by estimating a regression model with variables that are measured with error.

10.1.1 Effect of Errors-in-Variables

Let us assume two mean-centered variables, a dependent variable and an independent variable, y_t and x_t respectively, that are observed. However, these variables are imperfect measures of the true unobserved variables y_t^* and x_t^*. The measurement models for both variables are expressed by the equations

$$x_t = x_t^* + u_t \tag{10.1}$$

$$y_t = y_t^* + v_t \tag{10.2}$$

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_10, © Springer Science+Business Media New York 2014

There exists a structural relationship between these two unobserved variables, as indicated by the equation below:

$$y_t^* = x_t^* \beta \tag{10.3}$$

This equation can be expressed in terms of the observed variables by replacing each unobserved variable with its expression as a function of the observed variable obtained from Eqs. (10.1) and (10.2):

$$y_t = (x_t - u_t)\beta + v_t \tag{10.4}$$

or by placing the random error term at the end:

$$y_t = x_t \beta + v_t - u_t \beta \tag{10.5}$$

It should be noted that the error on the dependent variable y is similar to the error on the structural relationship. Indeed, if we had added an error term to Eq. (10.3), it would have been confounded with the measurement error of the dependent variable ν_t.

Because these variables are not observed, only the relationship between the observed variables can be estimated. This can be done by using the ordinary least squares estimator of the regression of y_t on x_t:

$$\hat{\beta}_{OLS_{1\times 1}} = \left(\underset{1\times T}{\mathbf{x}'} \underset{T\times 1}{\mathbf{x}} \right)^{-1} \underset{1\times T}{\mathbf{x}'} \underset{T\times 1}{\mathbf{y}} \tag{10.6}$$

The bias can be evaluated by taking the expectation of the OLS estimator:

$$\begin{aligned} \mathrm{E}\left[\hat{\beta}_{OLS}\right] &= \mathrm{E}\left[\left(\mathbf{x}'\mathbf{x}\right)^{-1} \mathbf{x}'\mathbf{y} \right] \\ &= \mathbf{E}\left[\left(\mathbf{x}'\mathbf{x}\right)^{-1} \mathbf{x}'(\mathbf{x}\beta + \mathbf{v} - \mathbf{u}\beta) \right] \\ &= \beta + \left(\mathbf{x}'\mathbf{x}\right)^{-1} \mathrm{E}\left[\mathbf{x}'(\mathbf{v} - \mathbf{u}\beta)\right] \\ &= \beta + \left(\mathbf{x}'\mathbf{x}\right)^{-1} \mathrm{E}\left[(\mathbf{x}^* + \mathbf{u})'(\mathbf{v} - \mathbf{u}\beta) \right] \\ \mathrm{E}[\beta_{OLS}] &= \beta + \mathrm{E}\left[\left(\mathbf{x}'\mathbf{x}\right)^{-1} \left(-\beta \mathbf{u}'\mathbf{u}\right) \right] \end{aligned} \tag{10.7}$$

Let

$$\mathrm{E}\left[\mathbf{u}'\mathbf{u}\right] = \sigma_u^2$$

If x has a mean of 0, the bias is

$$\mathrm{E}\left[\hat{\beta}_{OLS}\right] - \beta = -\beta \frac{\sigma_u^2}{\sigma_x^2} \tag{10.8}$$

Since $\sigma_x^2 = \sigma_{x^*}^2 + \sigma_u^2$, the bias can be expressed as

$$-\beta \frac{\sigma_u^2}{\sigma_{x^*}^2 + \sigma_u^2} = -\beta \frac{1}{1+\rho} \tag{10.9}$$

where $\rho = \frac{\sigma_{x^*}^2}{\sigma_u^2}$ is the signal-to-noise ratio.

From Eq. (10.9), we cannot only assert that there is a bias but we can also indicate properties about this bias. Because the variances in the signal-to-noise ratio are positive $\left(\sigma_u^2, \sigma_{x^*}^2 > 0\right)$, this means that the bias is always negative, i.e., Eq. (10.9) is always negative, thus the OLS estimates are underestimated when using a predictor variable with error. This is known as the attenuation effect. It may lead to failing to reject the null hypothesis that the effect of the independent variable on the dependent variable is insignificant.

As the signal-to-noise ratio ρ increases, the bias decreases, i.e., $1/(1 + \rho)$ becomes smaller. Therefore, we can summarize the results as follows:

1. We have found a lower bound for β. Indeed, we have shown that the OLS estimator $\hat{\beta}_{OLS}$ is smaller than the true β.
2. Error in measurement of x attenuates the effect of x.
3. Error in measurement of y does not bias the effect of x (the measurement error is then confounded with the noise in the relationship between the independent and dependent variables).

10.1.2 Reverse Regression

Let us write the equation that expresses the independent variable x_t as a function of the dependent variable y_t:

$$x_t = \gamma y_t + \varepsilon_t \tag{10.10}$$

Or, for all the observations:

$$\mathbf{x} = \gamma \mathbf{y} + \boldsymbol{\varepsilon} \tag{10.11}$$

The OLS estimator of the parameter γ is

$$\hat{\gamma}_{OLS} = \left(\mathbf{y}'\mathbf{y}\right)^{-1} \mathbf{y}'\mathbf{x} \tag{10.12}$$

Let

$$\hat{\beta}^R = \frac{1}{\hat{\gamma}_{OLS}} = \frac{\mathbf{y}'\mathbf{y}}{\mathbf{y}'\mathbf{x}} \tag{10.13}$$

If the variables are centered to zero mean,

$$\hat{\beta}^R = \frac{\mathrm{V}[y]}{\mathrm{Cov}[x,y]} \tag{10.14}$$

However, from Eqs. (10.2) and (10.3),

$$\mathbf{y} = \mathbf{x}^*\beta + \mathbf{v} \tag{10.15}$$

Consequently,

$$\mathrm{V}[y] = \beta^2\sigma_{x^*}^2 + \sigma_v^2 \tag{10.16}$$

and

$$\mathrm{Cov}[x,y] = \beta\sigma_{x^*}^2 \tag{10.17}$$

Therefore, Eq. (10.14) can be expressed as

$$\hat{\beta}^R = \frac{\beta^2\sigma_{x^*}^2 + \sigma_v^2}{\beta\sigma_{x^*}^2} = \beta\left(\frac{\beta^2\sigma_{x^*}^2 + \sigma_v^2}{\beta\sigma_{x^*}^2}\right) = \beta\left(1 + \frac{\sigma_v^2}{\beta^2\sigma_{x^*}^2}\right) = \beta(1+\omega) \tag{10.18}$$

where $\omega = \frac{\sigma_v^2}{\beta^2\sigma_{x^*}^2}$, which is always positive.

Because ω is positive, it follows that $\hat{\beta}^R$ overestimates β.

If we recall that the coefficient obtained from a direct regression Eq.(10.6), which we may call $\hat{\beta}^D$, always underestimates the true value of β, we then have shown that $\hat{\beta}^D$ and $\hat{\beta}^R$ provide bounds in the range where the true β falls.

Consequently, the choice of the dependent variable in a simple regression has nothing to do with causality. It follows from the analysis presented above that if ${\sigma_v}^2$ is small, we should use reverse regression (ω in Eq. (10.18) is then close to 0 and the bias is small). If, however, ${\sigma_u}^2$ is small (i.e., little measurement error in the predictor variable), direct regression should be used because the bias in Eq. (10.6) is then small. From this discussion it follows that the variable with the largest measurement error should be selected as the dependent variable.

10.1.3 Case with Multiple Independent Variables

The case with several independent variables is more complex. Let us consider Eq. (10.19), where some variables $\mathbf{z}_t$ are estimated without measurement error and others ${\mathbf{x}_t}^*$ are estimated with measurement error:

$$y_t^* = \mathbf{z}_t\gamma + \mathbf{x}_t^*\beta \tag{10.19}$$

In such cases, the direction of the bias is not easy to determine. Some conclusions are possible, however, in the special case when only one of the independent variables is measured with error, i.e., $\mathbf{x}_t$ contains a single variable. Then, it can be shown that the bias can be expressed as follows:

$$-\beta \frac{\sigma_u^2}{\sigma_x^2\left(1 - R_{x\mathbf{z}}^2\right)} \tag{10.20}$$

where ${R_{x\mathbf{z}}}^2$ is the R^2 of the regression of the variable measured with error (x_t) on those measured without error ($\mathbf{z}_t$).

Because the ratio that multiplies $-\beta$ in Eq. (10.20) is always positive, the coefficient is, therefore, always underestimated.

It should be noted that having one of the independent variables measured with error not only affects the estimation of the impact of that variable, but also the coefficients of the variables measured without error. Furthermore, both the overall F statistics and the individual coefficient variances are affected. The Fstatistic is always understated. Therefore, we would expect to reject the models more often than we should. The impact on individual statistics is not as clear, however, as there is no unambiguous bias.

This case of a single variable measured with error is, however, unusual. Most of the research in the social sciences involves the formation of scales that cannot be considered to be without measurement error. In such cases, the analysis shown in this section does not provide any guidance. The next section presents a methodology called analysis of covariance structure that resolves the problems associated with measurement errors.

10.2 Analysis of Covariance Structures

In the analysis of covariance structures, both the measurement errors and the structural relationships between the variables of interest are modeled.

10.2.1 Description of Model

We start with a system of simultaneous equations identical to the ones analyzed in Chap. 6:

$$\underset{m\times n}{\mathbf{B}}\ \underset{m\times 1}{\boldsymbol{\eta}} = \underset{m\times n}{\boldsymbol{\Gamma}}\ \underset{n\times 1}{\boldsymbol{\xi}} + \underset{m\times 1}{\boldsymbol{\zeta}} \tag{10.21}$$

where

m = Number of endogenous constructs
n = Number of exogenous constructs
$\boldsymbol{\eta}$ = Column vector of m endogenous constructs
$\boldsymbol{\xi}$ = Column vector of n endogenous constructs
$\boldsymbol{\zeta}$ = Column vector of m distrubance terms
$\mathbf{B}$ = Matrix of structural parameters of endogenous variables
$\boldsymbol{\Gamma}$ = Matrix of structural parameters of exogenous variables

The endogenous constructs are represented by the vector $\boldsymbol{\eta}$ and the exogenous ones by $\boldsymbol{\xi}$. Eq. (10.21) represents the structural relationships that exist among the constructs $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ with a random disturbance $\boldsymbol{\zeta}$. The diagonal elements of the matrix $\mathbf{B}$ are specified as being equal to one without affecting the generality of the model. The endogenous and exogenous constructs $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are not observed but are, instead, measured with error using multiple items. Before defining the measurement models, we should note that these unobserved constructs are defined as centered with zero mean without any loss of generality:

$$\mathrm{E}[\boldsymbol{\eta}] = 0; \quad \mathrm{E}[\boldsymbol{\xi}] = 0 \tag{10.22}$$

Like for the regression model, the error term is assumed to have zero mean:

$$\mathrm{E}[\boldsymbol{\zeta}] = 0 \tag{10.23}$$

In addition, the matrix of parameters $\mathbf{B}$ should be non-singular.

Let us now define the factor analytic measurement models. These are represented by Eqs. (10.24) and (10.25). There are p items or observable variables reflecting the m endogenous constructs and there are q items or observable variables reflecting the n exogenous constructs:

$$\underset{p\times 1}{\mathbf{y}} = \underset{p\times m}{\boldsymbol{\Lambda}_{\mathrm{y}}}\ \underset{m\times 1}{\boldsymbol{\eta}} + \underset{p\times 1}{\boldsymbol{\varepsilon}} \tag{10.24}$$

where

p = Number of items measuring the m endogenous constructs
$\mathbf{y}$ = Column vector of the p items or observable variable reflecting the endogenous constructs
$\boldsymbol{\Lambda}_{\mathrm{y}}$ = Matrix of factor loadings
$\boldsymbol{\varepsilon}$ = Column vector of measurement errors

The elements of the matrix $\boldsymbol{\Lambda}_{\mathrm{y}}$ represent the factor loadings. Similarly, for the measurement model of the exogenous constructs

$$\underset{q\times 1}{\mathbf{x}} = \underset{q\times n}{\boldsymbol{\Lambda}_{\mathrm{x}}}\ \underset{n\times 1}{\boldsymbol{\xi}} + \underset{q\times 1}{\boldsymbol{\delta}} \tag{10.25}$$

where

q	=	Number of items measuring the n endogenous constructs
$\mathbf{x}$	=	Column vector of the q items or observable variable reflecting the endogenous constructs
$\boldsymbol{\Lambda}_x$	=	Matrix of factor loadings
$\boldsymbol{\delta}$	=	Column vector of measurement errors

Furthermore, we can express the covariances of the latent variables and of the error terms according to Eqs. (10.26)–(10.29):

$$\mathrm{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}'\right] = \underset{n\times m}{\boldsymbol{\Phi}} \tag{10.26}$$

$$\mathrm{E}\left[\boldsymbol{\zeta}\boldsymbol{\zeta}'\right] = \underset{m\times m}{\boldsymbol{\psi}} \tag{10.27}$$

$$\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] = \underset{p\times p}{\boldsymbol{\Theta}_\varepsilon} \tag{10.28}$$

$$\mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}'\right] = \underset{q\times q}{\boldsymbol{\Theta}_\varepsilon} \tag{10.29}$$

We can now write the expression of what would theoretically be the covariance matrix of all the observed variables ($\mathbf{x}$ and $\mathbf{y}$), assuming the model expressed in the equations above.

Let

$$\underset{(p+q)\times 1}{\mathbf{z}} = \begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix} \tag{10.30}$$

The theoretical covariance matrix of $\mathbf{z}$ is

$$\Sigma = \mathrm{E}\left[\mathbf{z}\mathbf{z}'\right] = \mathrm{E}\left[\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix}\left(\mathbf{y}'\mathbf{x}'\right)\right] = \mathrm{E}\begin{bmatrix}\mathbf{y}\mathbf{y}' & \mathbf{y}\mathbf{x}'\\ \mathbf{x}\mathbf{y}' & \mathbf{x}\mathbf{x}'\end{bmatrix} \tag{10.31}$$

We derive the expression of each of the four submatrices in Eq. (10.31) with the following three blocks (the off-diagonal blocks are symmetric):

$$\begin{aligned}\mathrm{E}\left[\mathbf{x}\mathbf{x}'\right] &= \mathrm{E}\left[\left(\boldsymbol{\Lambda}_x\boldsymbol{\xi}+\boldsymbol{\delta}\right)\left(\boldsymbol{\Lambda}_x\boldsymbol{\xi}+\boldsymbol{\delta}\right)'\right]\\ &= \mathrm{E}\left[\left(\boldsymbol{\Lambda}_x\boldsymbol{\xi}\boldsymbol{\xi}'\boldsymbol{\Lambda}'_x\right)\right] + \mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}'\right]\\ &= \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}'_x + \boldsymbol{\Theta}_\delta\end{aligned} \tag{10.32}$$

$$\begin{aligned}\mathrm{E}\left[\mathbf{y}\mathbf{y}'\right] &= \left[\left(\boldsymbol{\Lambda}_y\boldsymbol{\eta}+\boldsymbol{\varepsilon}\right)\left(\boldsymbol{\Lambda}_y\boldsymbol{\eta}+\boldsymbol{\varepsilon}\right)'\right]\\ &= \boldsymbol{\Lambda}_y\mathrm{E}\left[\boldsymbol{\eta}\boldsymbol{\eta}'\right]\boldsymbol{\Lambda}'_y + \boldsymbol{\Theta}_\varepsilon\\ &= \boldsymbol{\Lambda}_y\mathrm{E}\left[\mathbf{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi}\boldsymbol{\xi}\boldsymbol{\Gamma}'\mathbf{B}^{-1\prime} + \mathbf{B}^{-1}\boldsymbol{\xi}\boldsymbol{\xi}'\mathbf{B}^{-1\prime}\right]\boldsymbol{\Lambda}'_y + \boldsymbol{\Theta}_\varepsilon\\ &= \boldsymbol{\Lambda}_y\left(\mathbf{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\mathbf{B}^{-1\prime} + \mathbf{B}^{-1}\boldsymbol{\Psi}\mathbf{B}^{-1\prime}\right)\boldsymbol{\Lambda}'_y + \boldsymbol{\Theta}_\varepsilon\end{aligned} \tag{10.33}$$

$$\mathrm{E}\left[\mathbf{y}\mathbf{x}'\right] = \mathrm{E}\left[\left(\boldsymbol{\Lambda}_y\boldsymbol{\eta}+\boldsymbol{\varepsilon}\right)\left(\boldsymbol{\Lambda}_x\boldsymbol{\xi}+\boldsymbol{\delta}\right)'\right]$$
$$= \mathrm{E}\left[\left(\boldsymbol{\Lambda}_y\mathbf{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi}+\mathbf{B}^{-1}\boldsymbol{\xi}+\boldsymbol{\varepsilon}\right)\left(\boldsymbol{\Lambda}_x\boldsymbol{\xi}+\boldsymbol{\delta}\right)'\right] = \boldsymbol{\Lambda}_y\mathbf{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}'_x \qquad (10.34)$$

Equations (10.32)–(10.34) provide the information to complete the covariance matrix in Eq. (10.31).

In fact, the observed covariance matrix can be computed from the sample of observations:

$$\mathbf{S} = \begin{bmatrix}\mathbf{S}_{yy} & \mathbf{S}_{yx}\\ \mathbf{S}_{xy} & \mathbf{S}_{xx}\end{bmatrix} \qquad (10.35)$$

10.2.2 Estimation

The estimation consists in finding the parameters of the model that will replicate as closely as possible the observed covariance matrix in Eq. (10.35). The maximum likelihood estimation compares the matrices $\mathbf{S}$ and $\sum$ using the following expression, which is derived from the likelihood function as presented in Chap. 4 for the confirmatory factor analytic model:

$$F = \mathrm{Ln}\left|\sum\right| + \mathrm{tr}\left(\mathbf{S}\sum^{-1}\right) - \mathrm{Ln}|\mathbf{S}| - (p+q) \qquad (10.36)$$

The only difference with the derivations in Chap. 4 is inherent in the fact that the covariance matrices contain the variances and covariances among the $(p + q)$ $\mathbf{x}$ and $\mathbf{y}$ variables. Therefore, under the assumption that the observed variables $\binom{\mathbf{y}}{\mathbf{x}}$ are distributed as a multivariate normal distribution, the parameter estimates that minimize Eq. (10.36) are the maximum likelihood estimates.

There are $\frac{1}{2}(p+q)(p+q+1)$ distinct elements that constitute the data. This expression comes from half of the symmetric matrix to which one needs to add back half of the diagonal in order to include the variances of the variables, i.e., $[(p + q) \times (p + q)/2 + (p + q)/2]$. Consequently, the number of degrees of freedom corresponds to the number of distinct data points as defined above minus the number of parameters in the model to estimate.

An example will illustrate the model and the degrees of freedom. MacKenzie, Lutz, and Belch (1986) compare several models of the role of attitude toward the ad on brand attitude and purchase intentions. Focusing on their dual-mediation hypothesis model (DMH) represented in Fig. 10.1, which they found to be supported by the data, three types of cognitive responses to advertising (about the ad execution, about the source, and about repetition) are the three exogenous constructs explaining the attitude toward the ad. Attitude toward the ad, according to that DMH theory, affects the attitude toward the brand not only directly but also indirectly by affecting brand cognitions that, in turn, affect the attitude toward the brand. Purchase intentions are affected by the attitude toward the brand as well as

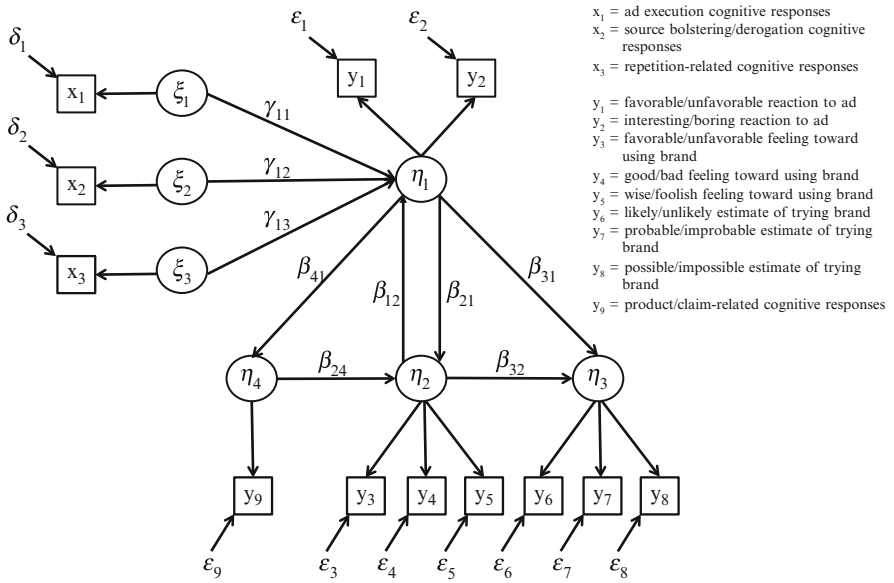


Fig. 10.1 A graphical representation of MacKenzie, Lutz, and Belch (1986)'s model of the role of attitude toward the ad. Adapted from MacKenzie et al. (1986)

directly by the attitude toward the ad. The relationships among the three exogenous constructs (the three types of cognitive responses) and the four endogenous constructs (attitude toward the ad, attitude toward the brand, brand cognitions, and purchase intentions) are drawn in Fig. 10.1. They can be expressed by the following system of four equations:

$$\begin{aligned}
\eta_1 &= \beta_{12}\eta_2 + \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \gamma_{13}\xi_3 + \zeta_1 \\
\eta_2 &= \beta_{21}\eta_1 + \beta_{24}\eta_4 + \zeta_2 \\
\eta_3 &= \beta_{31}\eta_1 + \beta_{32}\eta_2 + \zeta_3 \\
\eta_4 &= \beta_{41}\eta_1 + \zeta_4
\end{aligned} \tag{10.37}$$

or

$$\begin{bmatrix} 1 & -\beta_{12} & 0 & 0 \\ -\beta_{21} & 1 & 0 & -\beta_{24} \\ -\beta_{31} & -\beta_{32} & 1 & 0 \\ -\beta_{41} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \end{bmatrix} \tag{10.38}$$

In addition, Fig. 10.1 indicates that each of the exogenous constructs is measured by a single item, x_1 for ξ_1, x_2 for ξ_2, and x_3 for ξ_3. The attitude toward the ad (η_1) is measured by two items y_1 and y_2. The attitude toward the brand (η_2) and the purchase intentions (η_3) are both measured by three items:y_3, y_4, and y_5 for η_2,

and y_6, y_7, and y_8 for η_3. Finally, the brand cognitions (η_4) are measured by a single indicator y_9. The measurement model for the endogenous constructs can then be represented by Eq. (10.39), and the measurement model for the exogenous constructs can be expressed by Eq. (10.40):

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} \lambda_{y1} & 0 & 0 & 0 \\ \lambda_{y2} & 0 & 0 & 0 \\ 0 & \lambda_{y3} & 0 & 0 \\ 0 & \lambda_{y4} & 0 & 0 \\ 0 & \lambda_{y5} & 0 & 0 \\ 0 & 0 & \lambda_{y6} & 0 \\ 0 & 0 & \lambda_{y7} & 0 \\ 0 & 0 & \lambda_{y8} & 0 \\ 0 & 0 & 0 & \lambda_{y9} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{bmatrix} \tag{10.39}$$

and

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \lambda_{x1} & 0 & 0 \\ 0 & \lambda_{x2} & 0 \\ 0 & 0 & \lambda_{x3} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} \tag{10.40}$$

It should be noted that some restrictions on the measurement model parameters must be made for identification purposes. For each construct, the unit or scale of measurement must be defined. This is accomplished by setting one of the lambdas for a given construct to one; the corresponding variable will then serve as the unit of reference for that construct. For example, we can define $\lambda_{y1} = \lambda_{y3} = \lambda_{y6} = \lambda_{y9} = \lambda_{x1} = \lambda_{x2} = \lambda_{x3} = 1$. Alternatively, especially in the case of confirmatory factor analysis, the variance of the constructs could be set to unity.

We also need to impose some restrictions on some parameters in the cases where the constructs are measured by a single item. In such cases, the loading parameter is set to one, as discussed above and the error term is necessarily equal to zero. This means that the variance of the error term of that measurement equation must be constrained to be zero. This is the case for the example with $\theta_{\varepsilon 9}$, $\theta_{\delta 1}$, $\theta_{\delta 2}$, and $\theta_{\delta 3}$. Normally, the covariance matrices $\boldsymbol{\theta}_\delta$ and $\boldsymbol{\theta}_\varepsilon$ are assumed to be diagonal. In a few, exceptional cases, correlations between error terms of measurement equations can be estimated. This was the case in the example reported above from MacKenzie et al. (1986). However, estimating such correlations should only be done with great care, as the interpretation may be difficult.

The covariance matrix of the exogenous constructs can be symmetric or, with orthogonal factors, it can be defined as diagonal with zero covariances. In the example with orthogonal factors, three variances $\boldsymbol{\Psi}$ must be estimated.

Finally, the covariance matrix $\boldsymbol{\Phi}$ must be specified. It can be symmetric in the general case where the error terms of the structural equations are correlated. In such an example, there would be four variances and six covariances to estimate. The matrix is often assumed to be diagonal, in which case only four parameters (four variances) need to be estimated.

The equations described and the restrictions applied above indicate that 29 parameters must be estimated: five lambdas, six betas, three gammas, eight thetas, four phis, and three psis. Given that with 12 observed variables the covariance matrix consists of 78 different data points (i.e., (12 × 13)/2), this leaves 49 degrees of freedom.

10.2.3 Model Fit

We refer here to Sect. 4.2.1 in Chap. 4, since the measures of fit are identical to the description given when discussing the confirmatory factor analytic model. It should be noted that, for the adjusted goodness-of-fit index (AGFI), the adjustment for the degrees of freedom must take into account the $p + q$ variables, instead of just the q variables in confirmatory factor analysis:

$$\text{AGFI} = 1 - \left[\frac{(p+q)(p+q+1)}{(p+q)(p+q+1) - 2T}\right][1 - \text{GFI}] \tag{10.41}$$

where T is the number of estimated parameters.

The same change must be applied to the formula for the root mean square error of approximation (RMSEA) as the degrees of freedom d is given by

$$d = [(p+q)(p+q+1)/2] - T \tag{10.42}$$

10.2.4 Test of Significance of Model Parameters

The significance of each parameter can be tested using the standard t statistics formed by the ratio of the parameter estimate and its standard deviation. It should be recalled that this is possible because of the assumption about the normal distribution of the variables that enabled us to perform a maximum likelihood estimation.

10.2.5 Simultaneous Estimation of Measurement Model Parameters with Structural Relationship Parameters Versus Sequential Estimation

It can be noted that in the estimation method described above, the measurement model parameters are estimated at the same time as the structural model parameters. This means that the fit of the structural model and the structural model parameters are affected by the measurement model parameters. The rationale for the approach was to correct the bias produced by errors in measurement. However, the simultaneity of the estimation of all the parameters (measurement

model and structural model) implies that a trade-off is made between the values estimated for the measurement model and those for the structural model. To avoid this problem, it is best to first estimate the measurement model and then estimate the structural model parameters in a fully specified model (i.e., with the measurement model) where the parameters of the measurement model are fixed to the values estimated when the measurement model is estimated alone (Anderson and Gerbing 1988). This procedure does take into account the fact that the variables in the structural model are measured with error in order to estimate the structural model parameters, but it does not let the estimation of the measurement model interfere with the estimation of the structural model and vice versa.

10.2.6 Identification

As discussed earlier in Chap. 6, a model is identified if its parameters are identified, which means that there is only one set of values of the parameters that generates the covariance matrix. There are no general necessary and sufficient conditions to identify the general model discussed here; however, if the information matrix is not positive definite, the model is not identified. Furthermore, it appears logical that the structural model should be identified independently of the measurement model. Consequently, the order and rank conditions presented in Chap. 6 should be used to verify the identification of the structural relationships in an analysis of covariance structure model.

10.2.7 Special Cases of Analysis of Covariance Structure

The system of equations discussed in Chap. 6 and, a fortiori, the multiple regression analysis presented in Chap. 5 are obviously directly related to the general analysis of covariance models described above. This is because the fundamental relationships establishing the structural model follow the linear model. The distinguishing feature is the simultaneous modeling of measurement errors. If, however, each unobserved construct is defined by a single indicator (therefore fixing the factor loading to one and the error variance to zero), the models described in Chaps. 5 and 6 are reproduced.

Although less obvious, three of the analytical methods discussed in earlier chapters are also special cases of the general model we presented in this chapter: confirmatory factor analysis, second-order factor analysis and canonical correlation analysis. We show in this section how the general model reduces to each of these special cases.

10.2.7.1 Confirmatory Factor Analysis

In confirmatory factor analysis, there is no endogenous latent construct. The model simply reduces to the measurement model expressed in Eq. (10.25). Consequently, only the submatrix corresponding to the covariances among the exogenous items is considered in Eq. (10.31), i.e., the part given in Eq. (10.32).

10.2.7.2 Second-Order Factor Analysis

It is less obvious how the general model can reduce to the second-order factor analytic model. However, the relationships between the second order factors and the first order factors are established through Eq. (10.21) but with the peculiarity that $\mathbf{B} = \mathbf{I}$. This means that there is no endogeneity and that each η is a function of only the exogenous constructs. It may be confusing that, in this particular case, the structural relationships expressed by Eq. (10.21) represent a measurement model; however, this representation is in fact mathematically and statistically equivalent to a second-order factor analytic model. The other distinction with the general model is the lack of exogenous indicators. Indeed, the η's are considered as the reflective measures for the ξ's. Consequently, we are only interested in reproducing the submatrix in Eq. (10.31) that deals with the $\mathbf{y}$ variables, i.e., the covariances represented in Eq. (10.33).

10.2.7.3 Canonical Correlation Analysis

The equivalence of canonical correlation analysis with the general model described in this chapter is even more subtle. Again, the structural parameters are not truly considered as such. Let us consider a case where the exogenous constructs (the $\boldsymbol{\xi}$'s) are each measured by a single indicator. This means that the corresponding factor loadings will be set to one and the corresponding measurement error variances will be zero. If we now consider a single endogenous construct $\boldsymbol{\eta}$, the "structural relationship parameters" can be interpreted as the weights applied to the $\mathbf{x}$'s to form a linear combination of these variables. This can be seen more clearly by considering the graphical representation in Fig. 10.2.

The dotted box in Fig. 10.2 shows the part of the graphic that corresponds to the right side of the canonical correlation equation. Then, the relationships between the single η and the $\mathbf{y}$ variables are established through the "measurement" parameters in $\boldsymbol{\Lambda}_y$, combined with the specification of a full covariance matrix among the error terms $\boldsymbol{\varepsilon}$'s. Once again, the role of structural and measurement parameters as described earlier in this chapter does not hold; however, there is a statistical equivalence between these representations. This model is expressed as a multiple indicators/multiple causes (MIMIC) model of a single latent construct. We should be careful not to interpret the parameters $\boldsymbol{\Lambda}_y$ as being equivalent to the weights of

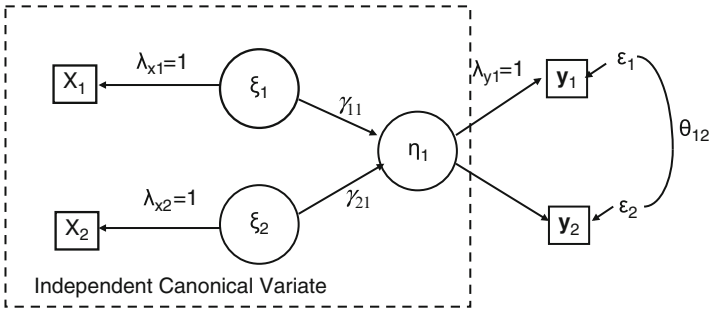


Fig. 10.2 Graphical representation of a canonical correlation model within the general analysis of covariance structure model

the dependent variables in the canonical correlation model specification. However, these parameters and weights are directly related to each other, and the canonical weights could then easily be inferred from the estimated parameters $\mathbf{\Lambda}_y$.

Indeed, λ_{y1} in Fig. 10.2 represents the correlation between y_1 and η_1 (assuming that η_1 has unit variance). However, η_1 is the canonical variate corresponding to the $\mathbf{x}$ variables in canonical correlation analysis (i.e., $\eta = z = \lambda_{x1}x_1 + \lambda_{x2}x_2$). But the squared correlation between y and z is precisely the definition of the redundancy measure in canonical correlation analysis (see Chap. 7). Therefore,

$$\underset{2\times 1}{\mathbf{\Lambda}_y} = \underset{1\times 1}{\mu}\ \underset{2\times 2}{\mathbf{R}_{yy}}\ \underset{2\times 1}{\mathbf{v}} \tag{10.43}$$

where $\mathbf{v}$ are the weights applied to the $\mathbf{y}$ variables to form the $\mathbf{y}$ canonical variate, μ is the correlation between the two canonical variates, and $\mathbf{R}_{yy}$ is the correlation matrix among the $\mathbf{y}$ variables (note in this case that $q = 1$).

Consequently, there is equivalence between the factor loadings in the analysis of covariance specification of the canonical correlation model and the weights of the linear combination of the y variables, as seen in Chap. 7 when canonical correlation analysis is performed.

10.3 Analysis of Covariance Structure with Means

Just as we introduced means and scalar constants in multi-group confirmatory factor analysis (Chap. 4), we now introduce them in the general model not only for the exogenous variables Eq. (10.45) but also for the endogenous variables Eq. (10.44):

$$\underset{p\times 1}{\mathbf{y}} = \underset{p\times 1}{\boldsymbol{\tau}_y} + \underset{p\times m}{\mathbf{\Lambda}_y}\ \underset{m\times 1}{\boldsymbol{\eta}} + \underset{p\times 1}{\boldsymbol{\varepsilon}} \tag{10.44}$$

$$\underset{q\times 1}{\mathbf{x}} = \underset{q\times 1}{\boldsymbol{\tau}_x} + \underset{q\times n}{\boldsymbol{\Lambda}_x}\ \underset{q\times 1}{\boldsymbol{\xi}} + \underset{q\times 1}{\boldsymbol{\delta}} \tag{10.45}$$

In addition, we introduce constant terms (intercepts) in the structural relationships to acknowledge the fact that the means of the unobserved constructs are not zero. These constant terms are the $\boldsymbol{\alpha}$'s in Eq. (10.46):

$$\underset{m\times 1}{\boldsymbol{\eta}} = \underset{m\times 1}{\boldsymbol{\alpha}} + \underset{m\times n}{\mathbf{B}}\ \underset{m\times 1}{\boldsymbol{\eta}} + \underset{m\times n}{\boldsymbol{\Gamma}}\ \underset{n\times 1}{\boldsymbol{\xi}} + \underset{m\times 1}{\boldsymbol{\zeta}} \tag{10.46}$$

with

$$\mathrm{E}[\boldsymbol{\xi}] = \underset{m\times 1}{\boldsymbol{\kappa}} \tag{10.47}$$

$$\mathrm{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}'\right] = \underset{m\times n}{\boldsymbol{\psi}} \tag{10.48}$$

$$\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] = \underset{p\times p}{\boldsymbol{\theta}_\varepsilon} \tag{10.49}$$

$$\mathrm{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}'\right] = \underset{q\times q}{\boldsymbol{\theta}_\delta} \tag{10.50}$$

It follows that the expected values of the observed exogenous variables are

$$\mathrm{E}\left[\underset{q\times 1}{\mathbf{x}}\right] = \underset{q\times 1}{\boldsymbol{\mu}_x} = \underset{q\times 1}{\boldsymbol{\tau}_x} + \underset{p\times n}{\boldsymbol{\Lambda}_x}\ \underset{n\times 1}{\boldsymbol{\kappa}} \tag{10.51}$$

The means of the endogenous constructs follow from Eqs. (10.46) and (10.51). From Eq. (10.46)

$$(\mathbf{I} - \mathbf{B})\mathrm{E}[\boldsymbol{\eta}] = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathrm{E}[\boldsymbol{\xi}] \tag{10.52}$$

$$\mathrm{E}[\boldsymbol{\eta}] = (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\kappa}) \tag{10.53}$$

However, the expected value of the endogenous observable items is

$$E[\mathbf{y}] = \boldsymbol{\mu}_y = \boldsymbol{\tau}_y + \boldsymbol{\Lambda}_y\mathrm{E}[\boldsymbol{\eta}] \tag{10.54}$$

Consequently, the expected value of the endogenous observable items expressed as a function of the theoretical parameters is

$$\boldsymbol{\mu}_y = \boldsymbol{\tau}_y + \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\kappa}). \tag{10.55}$$

Similar to the likelihood function discussed for confirmatory factor analysis with means and multiple groups in Chap. 4, the log likelihood function also contains the

parameters that model the means. Generalizing to the case of multiple groups, as was done in Chap. 4, leads to the log likelihood:

$$\mathbf{L} = -\frac{1}{2}\sum_{g=1}^{G} N^{(g)}\left[\left(p^{(g)}+q^{(g)}\right)Ln(2\pi)+Ln\left|\Sigma^{(g)}\right|+tr\left\{\left(\mathbf{Z}^{(g)}-\mu_z{}^{(g)}\right)\left(\mathbf{Z}^{(g)}-\mu_z{}^{(g)}\right)'\Sigma^{(g)^{-1}}\right\}\right] \tag{10.56}$$

If the theoretical model fits the data, Eqs. (10.51) and (10.54) provide the constraints that must be met to replicate the means of the observed variables. Therefore, in the full model with means, we model simultaneously the covariance matrix and the means of the observed variables in order to replicate as closely as possible the observed values. Then, the data will not only consist of the covariance matrix but also of the mean values of the observed variables. This is particularly useful in the presence of multiple groups where means across groups are likely to differ. Such multi-group analyses are common when testing the homogeneity of coefficients across groups. This is the type of analysis presented with multiple regression in Chap. 5 where we performed pooling tests. However, with this general model, the structural relationships tested take into consideration the measurement errors that would introduce a bias if ignored.

A particular type of test of homogeneity occurs when a moderator variable explains differences in structural relationships. In this case, one frequently used approach consists in splitting the observations into two (or more) groups according to the values of the moderator variable. Then a rejection of the homogeneity of coefficients hypothesis lends support to the moderating hypothesis.

10.4 Examples

We now present examples of analysis of covariance structure using LISREL for Windows, STATA or AMOS. These examples illustrate full structural models with error in measurement. The multi-group structural modeling described in Sect. 10.3 is illustrated with examples in Chap. 11 that compares alternative methods to test moderating effects.

10.4.1 Example of Structural Model with Measurement Models

Examples were given in prior chapters that were concerned exclusively with measurement models or confirmatory factor analysis. As shown earlier in this chapter, this is only one component of analysis of covariance structures. The full model also contains structural relationships among the unobserved constructs that

```
!Examp10-1.spl
!Raw Data From File: Examp10-1.txt
!Path Diagram

DA NI=19 MA = KM
RA FI=C:\SAMD\Chapter10\Examples\Examp10-1.txt
MO NX = 19 NK = 4 PH = SY TD = SY
FI    LX(1,1) LX(4,2) LX(9,3) LX(15,4)
VA 1 LX(1,1) LX(4,2) LX(9,3) LX(15,4)
LA
Q46 Q47 Q48 Q40 Q42 Q43 Q44 Q45 Q5 Q7 Q8 Q12 Q13 Q14 Q19r Q20 Q21 Q22 Q23
LK
Success Org2 CompEnh Radical
FR         LX(2,1)   LX(3,1)                              C
   LX(5,2)   LX(6,2)   LX(7,2)   LX(8,2)                   C
             LX(10,3) LX(11,3) LX(12,3) LX(13,3) LX(14,3)     C
             LX(16,4) LX(17,4) LX(18,4) LX(19,4) C
   TD(14,11)
Path Diagram
OU SE TV AD = 50 MI
```

Fig. 10.3 Step 1: Input of measurement model for exogenous and endogenous constructs—LISREL (examp10-1.spl)

need to be estimated. An example is provided below, where two characteristics of innovations (the extent to which an innovation is radical and the extent to which it is competence enhancing) are hypothesized to affect two constructs, one being changes in the management of the organization and the other being the success of that organization. We illustrate how to set up the problems in LISREL and then in STATA. Figure 10.3 presents the LISREL input file for step 1 of the analysis, i.e., the measurement model for all the constructs (including both the exogenous and endogenous constructs, although it would be feasible to estimate a separate measurement model for each).

The output results of the measurement model are shown in Fig. 10.4.

The output results shown in Fig. 10.4 are now represented graphically in Fig. 10.5.

The same model is now set up in STATA as shown in Fig. 10.6.

The STATA output results are shown in Fig. 10.7.

The values obtained in step 1 are then used as input for step 2, which consists in estimating the structural model parameters with the measurement parameters fixed to the values obtained in step 1. The LISREL input file for step 2 is shown in Fig. 10.8. The estimation of the model presented in that figure leads to maximum likelihood structural parameter estimates that take into consideration the fact that the constructs are measured with error.

The resulting parameter estimates of the structural relationships are shown graphically by LISREL in Fig. 10.9. Also included in that figure are the values of the measurement model parameters estimated in Fig. 10.4 and fixed to estimate the structural model parameters.

The full LISREL output is listed in Fig. 10.10. The example given in the figure is for illustrative purposes only because the results do not indicate that the fit between the model and the data is sufficiently close.

The corresponding input file for step 2 in STATA is shown in Fig. 10.11.

The output results of step 2 using STATA are given in Fig. 10.12.

```
                                L I S R E L  8.30

                                      BY

                           Karl G. Jöreskog & Dag Sörbom

                     This program is published exclusively by
                     Scientific Software International, Inc.
                        7383 N. Lincoln Avenue, Suite 100
                         Chicago, IL 60646-1704, U.S.A.
             Phone: (800)247-6113, (847)675-0720, Fax: (847)675-2140
       Copyright by Scientific Software International, Inc., 1981-99
         Use of this program is subject to the terms specified in the
                     Universal Copyright Convention.
                       Website: www.ssicentral.com

The following lines were read from file C:\SAMD\CHAPTER8\EXAMPLES\Examp8-6.SPL:

!Examp10-1.spl
!Raw Data From File: Examp10-1.txt
!Path Diagram

DA NI=19 MA = KM
RA FI=C:\SAMD\Chapter10\Examples\Examp10-1.txt
MO NX = 19 NK = 4 PH = SY TD = SY
FI LX(1,1) LX(4,2) LX(9,3) LX(15,4)
VA 1 LX(1,1) LX(4,2) LX(9,3) LX(15,4)
LA
Q46 Q47 Q48 Q40 Q42 Q43 Q44 Q45 Q5 Q7 Q8 Q12 Q13 Q14 Q19r Q20 Q21 Q22 Q23
LK
Success Org2 CompEnh Radical
FR        LX(2,1)  LX(3,1)            C
LX(5,2)   LX(6,2)  LX(7,2)  LX(8,2)         C
LX(10,3) LX(11,3) LX(12,3) LX(13,3) LX(14,3)     C
LX(16,4) LX(17,4) LX(18,4) LX(19,4)   C
TD(14,11)         !LX(1,1) LX(4,2) LX(9,3) LX(15,4)
Path Diagram
OU SE TV AD = 50 MI

!Examp10-1.spl

                          Number of Input Variables 19
                          Number of Y - Variables    0
                          Number of X - Variables   19
                          Number of ETA - Variables  0
                          Number of KSI - Variables  4
                          Number of Observations   146


        Covariance Matrix to be Analyzed

              Q46         Q47         Q48         Q40         Q42         Q43
         --------    --------    --------    --------    --------    --------
     Q46       1.00
     Q47       0.71        1.00
     Q48       0.60        0.83        1.00
     Q40       0.25        0.26        0.24        1.00
     Q42       0.31        0.30        0.25        0.59        1.00
     Q43       0.29        0.27        0.21        0.72        0.60        1.00
     Q44       0.32        0.36        0.27        0.60        0.59        0.67
     Q45       0.26        0.31        0.25        0.63        0.56        0.70
      Q5       0.28        0.30        0.25        0.22        0.05        0.20
      Q7       0.26        0.32        0.30        0.19        0.06        0.17
      Q8       0.18        0.15        0.11        0.26        0.37        0.36
     Q12       0.22        0.26        0.25        0.11       -0.02        0.08
     Q13       0.26        0.14        0.09        0.28        0.39        0.32
     Q14       0.17        0.11        0.11        0.28        0.41        0.31
    Q19r       0.26        0.21        0.15        0.29        0.38        0.32
     Q20       0.30        0.19        0.18        0.40        0.42        0.39
     Q21       0.18        0.11        0.09        0.32        0.39        0.33
     Q22       0.22        0.23        0.20        0.16        0.28        0.18
     Q23       0.33        0.23        0.21        0.31        0.35        0.28
```

Fig. 10.4 Step 1: The measurement model output results—LISREL (examp10-1.out)

Covariance Matrix to be Analyzed

	Q44	Q45	Q5	Q7	Q8	Q12
Q44	1.00					
Q45	0.63	1.00				
Q5	0.18	0.19	1.00			
Q7	0.13	0.15	0.62	1.00		
Q8	0.35	0.36	-0.03	0.01	1.00	
Q12	0.08	0.11	0.63	0.72	-0.06	1.00
Q13	0.21	0.26	-0.10	-0.08	0.56	-0.08
Q14	0.31	0.32	-0.12	-0.07	0.65	-0.13
Q19r	0.28	0.34	0.00	0.00	0.35	-0.09
Q20	0.30	0.32	0.00	0.01	0.47	-0.03
Q21	0.24	0.35	0.07	0.02	0.36	0.03
Q22	0.15	0.17	0.12	0.21	0.29	0.15
Q23	0.20	0.16	0.23	0.20	0.24	0.16

Covariance Matrix to be Analyzed

	Q13	Q14	Q19r	Q20	Q21	Q22
Q13	1.00					
Q14	0.60	1.00				
Q19r	0.48	0.38	1.00			
Q20	0.55	0.47	0.60	1.00		
Q21	0.43	0.35	0.63	0.77	1.00	
Q22	0.29	0.32	0.48	0.40	0.45	1.00
Q23	0.40	0.26	0.61	0.60	0.61	0.59

Covariance Matrix to be Analyzed

	Q23
Q23	1.00

Parameter Specifications

LAMBDA-X

	Success	Org2	CompEnh	Radical
Q46	0	0	0	0
Q47	1	0	0	0
Q48	2	0	0	0
Q40	0	0	0	0
Q42	0	3	0	0
Q43	0	4	0	0
Q44	0	5	0	0
Q45	0	6	0	0
Q5	0	0	0	0
Q7	0	0	7	0
Q8	0	0	8	0
Q12	0	0	9	0
Q13	0	0	10	0
Q14	0	0	11	0
Q19r	0	0	0	0
Q20	0	0	0	12
Q21	0	0	0	13
Q22	0	0	0	14
Q23	0	0	0	15

PHI

	Success	Org2	CompEnh	Radical
Success	16			
Org2	17	18		
CompEnh	19	20	21	
Radical	22	23	24	25

THETA-DELTA

	Q46	Q47	Q48	Q40	Q42	Q43

Fig. 10.4 (continued)

```
      Q46          26
      Q47           0          27
      Q48           0           0          28
      Q40           0           0           0          29
      Q42           0           0           0           0          30
      Q43           0           0           0           0           0          31
      Q44           0           0           0           0           0           0
      Q45           0           0           0           0           0           0
       Q5           0           0           0           0           0           0
       Q7           0           0           0           0           0           0
       Q8           0           0           0           0           0           0
      Q12           0           0           0           0           0           0
      Q13           0           0           0           0           0           0
      Q14           0           0           0           0           0           0
    Q19r            0           0           0           0           0           0
      Q20           0           0           0           0           0           0
      Q21           0           0           0           0           0           0
      Q22           0           0           0           0           0           0
      Q23           0           0           0           0           0           0

        THETA-DELTA

             Q44         Q45         Q5          Q7          Q8          Q12
        ---------   ---------   ---------   ---------   ---------   ---------
      Q44          32
      Q45           0          33
       Q5           0           0          34
       Q7           0           0           0          35
       Q8           0           0           0           0          36
      Q12           0           0           0           0           0          37
      Q13           0           0           0           0           0           0
      Q14           0           0           0           0          39           0
    Q19r            0           0           0           0           0           0
      Q20           0           0           0           0           0           0
      Q21           0           0           0           0           0           0
      Q22           0           0           0           0           0           0
      Q23           0           0           0           0           0           0

        THETA-DELTA

             Q13         Q14        Q19r         Q20         Q21         Q22
        ---------   ---------   ---------   ---------   ---------   ---------
      Q13          38
      Q14           0          40
    Q19r            0           0          41
      Q20           0           0           0          42
      Q21           0           0           0           0          43
      Q22           0           0           0           0           0          44
      Q23           0           0           0           0           0           0

        THETA-DELTA

             Q23
        ---------
      Q23          45


 Number of Iterations = 10

 LISREL Estimates (Maximum Likelihood)

        LAMBDA-X

            Success        Org2     CompEnh     Radical
        ---------   ---------   ---------   ---------
      Q46       1.00       - -         - -         - -

      Q47       1.34       - -         - -         - -
               (0.12)
                10.86

      Q48       1.16       - -         - -         - -
               (0.11)
                10.37
```

Fig. 10.4 (continued)

```
      Q40         - -        1.00         - -         - -

      Q42         - -        0.90         - -         - -
                           (0.10)
                            9.43

      Q43         - -        1.07         - -         - -
                           (0.09)
                           11.74

      Q44         - -        0.96         - -         - -
                           (0.09)
                           10.17

      Q45         - -        0.99         - -         - -
                           (0.09)
                           10.58

       Q5         - -         - -        1.00         - -

       Q7         - -         - -        1.13         - -
                                        (0.12)
                                         9.48

       Q8         - -         - -       -0.04         - -
                                        (0.12)
                                        -0.31

      Q12         - -         - -        1.13         - -
                                        (0.12)
                                         9.47

      Q13         - -         - -       -0.14         - -
                                        (0.12)
                                        -1.16

      Q14         - -         - -       -0.17         - -
                                        (0.12)
                                        -1.46

     Q19r         - -         - -         - -        1.00

      Q20         - -         - -         - -        1.13
                                                    (0.11)
                                                    10.18

      Q21         - -         - -         - -        1.15
                                                    (0.11)
                                                    10.36

      Q22         - -         - -         - -        0.75
                                                    (0.11)
                                                     6.63

      Q23         - -         - -         - -        1.00
                                                    (0.11)
                                                     8.93


        PHI

             Success       Org2     CompEnh     Radical
             --------    --------   --------    --------
Success         0.53
               (0.11)
                4.95

   Org2         0.22        0.65
               (0.06)      (0.11)
                3.71        5.70

CompEnh         0.20        0.11        0.56
               (0.06)      (0.06)      (0.11)
                3.47        1.86        4.96

Radical         0.13        0.30        0.03        0.56
```

Fig. 10.4 (continued)

```
                (0.05)      (0.07)      (0.05)      (0.11)
                  2.49        4.41        0.59        5.10


        THETA-DELTA

                   Q46         Q47         Q48         Q40         Q42         Q43
            --------    --------    --------    --------    --------    --------
Q46            0.47
              (0.06)
               7.71

Q47             - -        0.05
                          (0.04)
                           1.09

Q48             - -         - -        0.28
                                      (0.05)
                                       5.95

Q40             - -         - -         - -        0.35
                                                  (0.05)
                                                   6.80

Q42             - -         - -         - -         - -        0.47
                                                              (0.06)
                                                               7.49

Q43             - -         - -         - -         - -         - -        0.25
                                                                          (0.04)
                                                                           5.80

Q44             - -         - -         - -         - -         - -         - -

Q45             - -         - -         - -         - -         - -         - -

 Q5             - -         - -         - -         - -         - -         - -

 Q7             - -         - -         - -         - -         - -         - -

 Q8             - -         - -         - -         - -         - -         - -

Q12             - -         - -         - -         - -         - -         - -

Q13             - -         - -         - -         - -         - -         - -

Q14             - -         - -         - -         - -         - -         - -

Q19r            - -         - -         - -         - -         - -         - -

Q20             - -         - -         - -         - -         - -         - -

Q21             - -         - -         - -         - -         - -         - -

Q22             - -         - -         - -         - -         - -         - -

Q23             - -         - -         - -         - -         - -         - -


        THETA-DELTA

                   Q44         Q45          Q5          Q7          Q8         Q12
            --------    --------    --------    --------    --------    --------
Q44            0.40
              (0.06)
               7.17

Q45             - -        0.37
                          (0.05)
                           6.93

 Q5             - -         - -        0.44
                                      (0.07)
                                       6.74

 Q7             - -         - -         - -        0.28
```

Fig. 10.4 (continued)

```
                                                (0.06)
                                                 4.69

      Q8          - -         - -         - -         - -        1.00
                                                                (0.12)
                                                                 8.51

     Q12          - -         - -         - -         - -         - -        0.29
                                                                           (0.06)
                                                                            4.79

     Q13          - -         - -         - -         - -         - -         - -

     Q14          - -         - -         - -         - -        0.65         - -
                                                                (0.10)
                                                                 6.59

    Q19r          - -         - -         - -         - -         - -         - -

     Q20          - -         - -         - -         - -         - -         - -

     Q21          - -         - -         - -         - -         - -         - -

     Q22          - -         - -         - -         - -         - -         - -

     Q23          - -         - -         - -         - -         - -         - -


        THETA-DELTA

                Q13         Q14         Q19r        Q20         Q21         Q22
            --------    --------    --------    --------    --------    --------
     Q13        0.99
               (0.12)
                8.50

     Q14         - -        0.98
                           (0.12)
                            8.49

    Q19r         - -         - -        0.44
                                       (0.06)
                                        7.22

     Q20         - -         - -         - -        0.28
                                                   (0.05)
                                                    5.92

     Q21         - -         - -         - -         - -        0.26
                                                                (0.05)
                                                                 5.53

     Q22         - -         - -         - -         - -         - -        0.68
                                                                           (0.08)
                                                                            8.05

     Q23         - -         - -         - -         - -         - -         - -


        THETA-DELTA

                Q23
            --------
     Q23        0.44
               (0.06)
                7.24


        Squared Multiple Correlations for X - Variables

                Q46         Q47         Q48         Q40         Q42         Q43
            --------    --------    --------    --------    --------    --------
                0.53        0.95        0.72        0.65        0.53        0.75

        Squared Multiple Correlations for X - Variables
```

Fig. 10.4 (continued)

```
          Q44       Q45        Q5        Q7        Q8       Q12
        --------  --------  --------  --------  --------  --------
          0.60      0.63      0.56      0.72      0.00      0.71

  Squared Multiple Correlations for X - Variables

          Q13       Q14       Q19r      Q20       Q21       Q22
        --------  --------  --------  --------  --------  --------
          0.01      0.02      0.56      0.72      0.74      0.32

  Squared Multiple Correlations for X - Variables

          Q23
        --------
          0.56


                    Goodness of Fit Statistics

                      Degrees of Freedom = 145
          Minimum Fit Function Chi-Square = 332.35 (P = 0.00)
  Normal Theory Weighted Least Squares Chi-Square = 330.77 (P = 0.00)
          Estimated Non-centrality Parameter (NCP) = 185.77
      90 Percent Confidence Interval for NCP = (136.72 ; 242.54)

                Minimum Fit Function Value = 2.29
          Population Discrepancy Function Value (F0) = 1.28
        90 Percent Confidence Interval for F0 = (0.94 ; 1.67)
       Root Mean Square Error of Approximation (RMSEA) = 0.094
      90 Percent Confidence Interval for RMSEA = (0.081 ; 0.11)
        P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00

          Expected Cross-Validation Index (ECVI) = 2.90
      90 Percent Confidence Interval for ECVI = (2.56 ; 3.29)
                ECVI for Saturated Model = 2.62
            ECVI for Independence Model = 12.01

Chi-Square for Independence Model with 171 Degrees of Freedom = 1702.81
                      Independence AIC = 1740.81
                          Model AIC = 420.77
                        Saturated AIC = 380.00
                      Independence CAIC = 1816.49
                          Model CAIC = 600.03
                        Saturated CAIC = 1136.89

              Root Mean Square Residual (RMR) = 0.17
                    Standardized RMR = 0.17
                  Goodness of Fit Index (GFI) = 0.81
            Adjusted Goodness of Fit Index (AGFI) = 0.75
          Parsimony Goodness of Fit Index (PGFI) = 0.62

                  Normed Fit Index (NFI) = 0.80
                Non-Normed Fit Index (NNFI) = 0.86
            Parsimony Normed Fit Index (PNFI) = 0.68
              Comparative Fit Index (CFI) = 0.88
              Incremental Fit Index (IFI) = 0.88
               Relative Fit Index (RFI) = 0.77

                    Critical N (CN) = 82.82


Modification Indices and Expected Change

      Modification Indices for LAMBDA-X

          Success     Org2     CompEnh    Radical
        ----------  --------  --------  --------
  Q46       - -       2.36      0.26      6.54
  Q47       - -       0.16      0.40      1.64
  Q48       - -       0.40      0.14      0.09
  Q40      0.76        - -      0.87      0.08
  Q42      0.57        - -      3.71      6.56
  Q43      2.02        - -      0.00      0.26
  Q44      2.65        - -      0.02      1.94
  Q45      0.18        - -      0.22      0.27
   Q5      0.76       3.01       - -      0.94
```

Fig. 10.4 (continued)

```
        Q7        0.48        0.87         - -        0.47
        Q8        0.87        5.43         - -        5.36
       Q12        1.04        2.45         - -        0.22
       Q13        6.24       22.55         - -       44.94
       Q14        1.22        5.46         - -        7.33
     Q19r        0.73        0.53        2.85         - -
       Q20       0.01        1.56        2.90         - -
       Q21       6.79        0.59        0.36         - -
       Q22       2.45        0.52        5.18         - -
       Q23       2.10        0.71       10.40         - -

          Expected Change for LAMBDA-X

              Success        Org2     CompEnh     Radical
              --------    --------    --------    --------
       Q46         - -        0.13        0.05        0.22
       Q47         - -       -0.03       -0.05       -0.09
       Q48         - -       -0.04        0.03       -0.02
       Q40       -0.07         - -        0.08        0.03
       Q42        0.07         - -       -0.17        0.26
       Q43       -0.11         - -        0.00       -0.04
       Q44        0.14         - -        0.01       -0.13
       Q45        0.04         - -        0.04       -0.05
        Q5        0.08        0.14         - -        0.08
        Q7        0.06        0.07         - -        0.06
        Q8        0.09        0.19         - -        0.20
       Q12       -0.09       -0.12         - -       -0.04
       Q13        0.32        0.52         - -        0.78
       Q14        0.11        0.19         - -        0.24
     Q19r        0.07        0.07       -0.14         - -
       Q20       0.01        0.10       -0.13         - -
       Q21       -0.20       -0.06       -0.04         - -
       Q22       0.16       -0.08        0.23         - -
       Q23       0.13       -0.08        0.28         - -

No Non-Zero Modification Indices for PHI

          Modification Indices for THETA-DELTA

                Q46         Q47         Q48         Q40         Q42         Q43
               --------    --------    --------    --------    --------    --------
       Q46         - -
       Q47        0.05         - -
       Q48        0.97        3.24         - -
       Q40        0.14        1.68        1.11         - -
       Q42        0.64        0.04        0.01        0.02         - -
       Q43        0.79        0.47        0.58        2.47        2.32         - -
       Q44        0.00        2.28        0.53        1.34        1.17        0.03
       Q45        1.10        0.62        0.01        0.43        0.40        0.46
        Q5        1.34        0.01        0.49        0.77        1.31        0.30
        Q7        0.00        0.01        0.15        0.58        0.15        0.17
        Q8        0.19        0.15        0.60        1.69        0.01        0.82
       Q12        0.02        0.34        0.22        0.17        0.48        0.24
       Q13        7.38        0.24        0.65        0.02        5.51        0.21
       Q14        0.73        0.77        0.50        0.36        3.35        0.60
     Q19r        0.02        1.95        0.92        1.43        0.19        0.17
       Q20       2.17        1.17        0.45        2.24        0.00        0.72
       Q21       0.86        0.33        0.15        0.57        0.08        0.27
       Q22       0.18        0.56        0.03        0.91        1.23        0.44
       Q23       2.78        0.71        0.26        1.70        0.85        0.01

          Modification Indices for THETA-DELTA

                Q44         Q45          Q5          Q7          Q8         Q12
               --------    --------    --------    --------    --------    --------
       Q44         - -
       Q45        0.32         - -
        Q5        0.13        0.07         - -
        Q7        0.32        0.32        1.58         - -
        Q8        0.60        0.69        0.11        0.18         - -
       Q12        0.04        0.36        0.03        1.10        0.60         - -
       Q13        1.26        0.04        0.19        0.04        7.22        0.01
       Q14        0.02        0.00        0.38        0.27         - -        0.00
     Q19r        0.28        2.10        0.02        0.33        0.02        3.81
       Q20       0.00        1.58        1.51        0.02        2.41        0.07
       Q21       1.30        5.34        0.26        2.95        0.11        2.23
```

Fig. 10.4 (continued)

```
Q22        0.03        0.09        0.68        3.25        0.04        0.24
Q23        0.15       10.86        2.80        0.42        1.17        0.11

   Modification Indices for THETA-DELTA

            Q13         Q14        Q19r         Q20         Q21         Q22
        --------    --------    --------    --------    --------    --------
Q13         - -
Q14       13.55         - -
Q19r       2.45        0.33         - -
Q20        7.22        1.08        2.97         - -
Q21        0.20        0.55        0.70       17.27         - -
Q22        0.04        1.50        1.75        8.00        1.82         - -
Q23        0.31        0.02        2.32        2.20        3.37       18.40

   Modification Indices for THETA-DELTA

            Q23
        --------
Q23         - -

   Expected Change for THETA-DELTA

            Q46         Q47         Q48         Q40         Q42         Q43
        --------    --------    --------    --------    --------    --------
Q46         - -
Q47       -0.02         - -
Q48       -0.06        0.23         - -
Q40       -0.01       -0.03        0.03         - -
Q42        0.03        0.01        0.00        0.01         - -
Q43        0.03       -0.02       -0.02        0.06       -0.06         - -
Q44        0.00        0.04       -0.02       -0.05        0.05       -0.01
Q45       -0.04        0.02        0.00       -0.03       -0.03        0.03
  Q5        0.05        0.00       -0.02        0.03       -0.05        0.02
  Q7        0.00        0.00        0.01        0.03        0.02        0.01
  Q8        0.02        0.01       -0.03       -0.05        0.00        0.03
Q12         0.00       -0.02        0.01       -0.01       -0.03       -0.02
Q13         0.16       -0.02       -0.04        0.01        0.14        0.02
Q14         0.04       -0.03        0.02        0.02        0.08       -0.03
Q19r        0.01        0.04       -0.03       -0.05        0.02       -0.01
Q20         0.05       -0.03        0.02        0.05        0.00        0.03
Q21       -0.03       -0.01       -0.01       -0.02        0.01       -0.02
Q22       -0.02        0.03        0.01       -0.04        0.06       -0.03
Q23        0.07       -0.02        0.02        0.05        0.04        0.00

   Expected Change for THETA-DELTA

            Q44         Q45          Q5          Q7          Q8         Q12
        --------    --------    --------    --------    --------    --------
Q44         - -
Q45        0.02         - -
  Q5       0.01        0.01         - -
  Q7      -0.02       -0.02       -0.13         - -
  Q8       0.03        0.03        0.02        0.02         - -
Q12       -0.01        0.02        0.02        0.15       -0.03         - -
Q13       -0.06       -0.01       -0.03        0.01        0.17        0.00
Q14       -0.01        0.00       -0.03        0.02         - -        0.00
Q19r       0.02        0.06       -0.01        0.02        0.01       -0.08
Q20        0.00       -0.04       -0.05        0.00        0.06       -0.01
Q21       -0.04        0.08        0.02       -0.06        0.01        0.05
Q22       -0.01       -0.01       -0.04        0.08       -0.01        0.02
Q23       -0.02       -0.13        0.07        0.03       -0.05        0.01

   Expected Change for THETA-DELTA

            Q13         Q14        Q19r         Q20         Q21         Q22
        --------    --------    --------    --------    --------    --------
Q13         - -
Q14        0.23         - -
Q19r       0.09        0.03         - -
Q20        0.14        0.04       -0.08         - -
Q21       -0.02       -0.03       -0.04        0.20         - -
Q22        0.01        0.06        0.07       -0.13       -0.06         - -
Q23        0.03        0.01        0.07       -0.07       -0.08        0.22

   Expected Change for THETA-DELTA

            Q23
        --------
Q23         - -

Maximum Modification Index is   44.94 for Element (13, 4) of LAMBDA-X
```

Fig. 10.4 (continued)

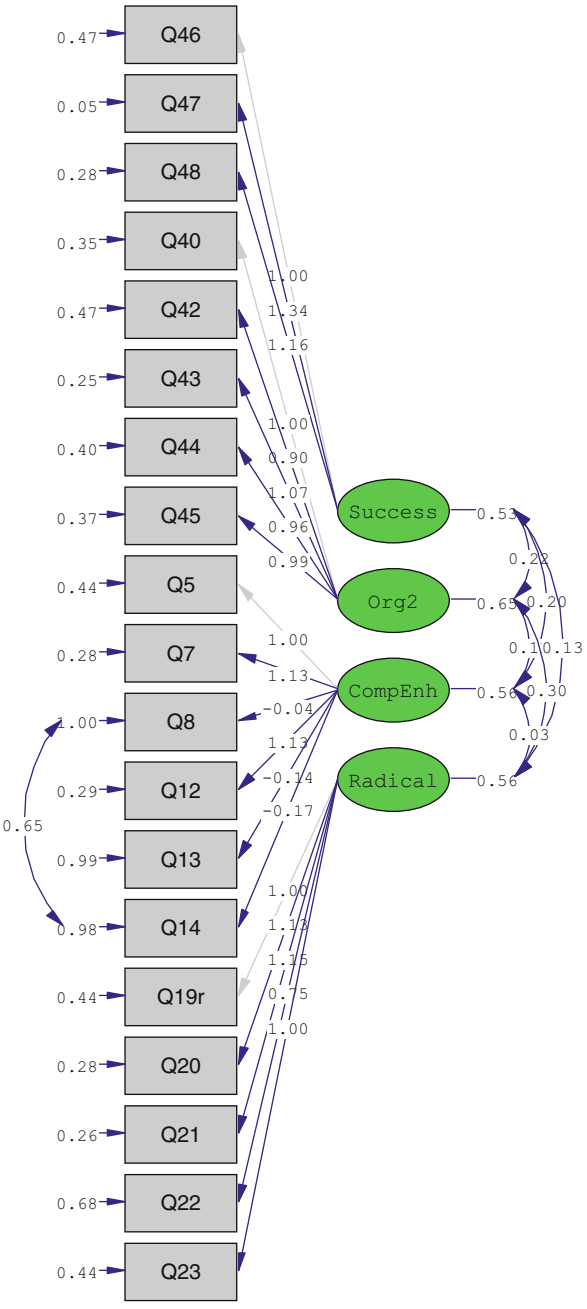


Fig. 10.5 Step 1: Graphical representation of measurement model for exogenous and endogenous constructs—LISREL (examp10-1.pth)

```
*Examp10-1.do
insheet q46 q47 q48 q40 q42 q43 q44 q45 q5 q7 q8 q12 q13 q14 q19r q20 q21 q22 q23
using "/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter10_ACS-SEM/Examp10-1.txt",
clear
replace q46 =. if q46 == 9
replace q47 =. if q47 == 9
replace q48 =. if q48 == 9
replace q40 =. if q40 == 9
replace q42 =. if q42 == 9
replace q43 =. if q43 == 9
replace q44 =. if q44 == 9
replace q45 =. if q45 == 9
replace q5 =. if q5 == 9
replace q7 =. if q7 == 9
replace q8 =. if q8 == 9
replace q12 =. if q12 == 9
replace q13 =. if q13 == 9
replace q14 =. if q14 == 9
replace q19r =. if q19r == 9
replace q20 =. if q20 == 9
replace q21 =. if q21 == 9
replace q22 =. if q22 == 9
replace q23 =. if q23 == 9
sem (Success -> q46 q47 q48) ///
(Org2 -> q40 q42 q43 q44 q45) ///
(CompEnh -> q5 q7 q8 q12 q13 q14) ///
(Radical -> q19r q20 q21 q22 q23) ///
, nomeans latent(Success Org2 CompEnh Radical) ///
cov(e.q8*e.q14)
estat gof, stats(all)
estat framework, fitted
estat mindices
```

Fig. 10.6 Step 1: The measurement model input—STATA (examp10-1.do)

```
. *Examp10-1.do
. insheet q46 q47 q48 q40 q42 q43 q44 q45 q5 q7 q8 q12 q13 q14 q19r q20 q21 q22 q23
using "/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter10_ACS-SEM/Examp10-1.txt",
clear
(19 vars, 146 obs)
. replace q46 =. if q46 == 9
(12 real changes made, 12 to missing)
. replace q47 =. if q47 == 9
(19 real changes made, 19 to missing)
. replace q48 =. if q48 == 9
(23 real changes made, 23 to missing)
. replace q40 =. if q40 == 9
(12 real changes made, 12 to missing)
. replace q42 =. if q42 == 9
(16 real changes made, 16 to missing)
. replace q43 =. if q43 == 9
(12 real changes made, 12 to missing)
. replace q44 =. if q44 == 9
(17 real changes made, 17 to missing)
. replace q45 =. if q45 == 9
(11 real changes made, 11 to missing)
. replace q5 =. if q5 == 9
(6 real changes made, 6 to missing)
. replace q7 =. if q7 == 9
(6 real changes made, 6 to missing)
. replace q8 =. if q8 == 9
(6 real changes made, 6 to missing)
. replace q12 =. if q12 == 9
(5 real changes made, 5 to missing)
. replace q13 =. if q13 == 9
(8 real changes made, 8 to missing)
. replace q14 =. if q14 == 9
```

Fig. 10.7 Step 1: The measurement model results—STATA (examp10-1.log)

```
(6 real changes made, 6 to missing)
. replace q19r =. if q19r == 9
(6 real changes made, 6 to missing)
. replace q20 =. if q20 == 9
(6 real changes made, 6 to missing)
. replace q21 =. if q21 == 9
(18 real changes made, 18 to missing)
. replace q22 =. if q22 == 9
(12 real changes made, 12 to missing)
. replace q23 =. if q23 == 9
(7 real changes made, 7 to missing)
. sem (Success -> q46 q47 q48) ///
> (Org2 -> q40 q42 q43 q44 q45) ///
> (CompEnh -> q5 q7 q8 q12 q13 q14) ///
> (Radical -> q19r q20 q21 q22 q23) ///
> , nomeans latent(Success Org2 CompEnh Radical) ///
> cov(e.q8*e.q14)
(50 observations with missing values excluded;
 specify option 'method(mlmv)' to use all observations)
Endogenous variables
Measurement:  q46 q47 q48 q40 q42 q43 q44 q45 q5 q7 q8 q12 q13 q14 q19r q20 q21 q22
q23
Exogenous variables
Latent:       Success Org2 CompEnh Radical
Fitting target model:
Iteration 0:   log likelihood = -3174.5611
Iteration 1:   log likelihood = -3158.2527
Iteration 2:   log likelihood = -3146.6956
Iteration 3:   log likelihood = -3145.1899
Iteration 4:   log likelihood = -3145.1747
Iteration 5:   log likelihood = -3145.1746

Structural equation model                       Number of obs      =         96
Estimation method  = ml
Log likelihood     = -3145.1746

 ( 1)  [q46]Success = 1
 ( 2)  [q40]Org2 = 1
 ( 3)  [q5]CompEnh = 1
 ( 4)  [q19r]Radical = 1
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Measurement  |
  q46 <-     |
     Success |          1  (constrained)
  -----------+----------------------------------------------------------------
  q47 <-     |
     Success |   1.165879   .1071507    10.88   0.000     .9558676    1.375891
  -----------+----------------------------------------------------------------
  q48 <-     |
     Success |   1.221725   .1158333    10.55   0.000     .9946955    1.448754
  -----------+----------------------------------------------------------------
  q40 <-     |
       Org2  |          1  (constrained)
  -----------+----------------------------------------------------------------
  q42 <-     |
       Org2  |   1.030687   .1826407     5.64   0.000     .6727175    1.388656
  -----------+----------------------------------------------------------------
  q43 <-     |
       Org2  |   1.161106   .1757181     6.61   0.000     .8167046    1.505507
  -----------+----------------------------------------------------------------
  q44 <-     |
       Org2  |   1.244799   .1899297     6.55   0.000     .8725435    1.617054
  -----------+----------------------------------------------------------------
  q45 <-     |
       Org2  |   1.107081   .1890572     5.86   0.000     .7365352    1.477626
  -----------+----------------------------------------------------------------
  q5 <-      |
     CompEnh |          1  (constrained)
  -----------+----------------------------------------------------------------
```

Fig. 10.7 (continued)

```
  q7 <-      |
     CompEnh |   .9699365   .1260964     7.69   0.000     .7227921    1.217081
  -----------+------------------------------------------------------------------
  q8 <-      |
     CompEnh |  -.4558465   .1092636    -4.17   0.000    -.6699992   -.2416939
  -----------+------------------------------------------------------------------
  q12 <-     |
     CompEnh |   1.146397   .1349254     8.50   0.000     .8819478    1.410845
  -----------+------------------------------------------------------------------
  q13 <-     |
     CompEnh |  -.4299586   .1283432    -3.35   0.001    -.6815065   -.1784106
  -----------+------------------------------------------------------------------
  q14 <-     |
     CompEnh |  -.5356665    .138298    -3.87   0.000    -.8067256   -.2646073
  -----------+------------------------------------------------------------------
  q19r <-    |
     Radical |           1  (constrained)
  -----------+------------------------------------------------------------------
  q20 <-     |
     Radical |   1.152578   .2463343     4.68   0.000     .6697716    1.635384
  -----------+------------------------------------------------------------------
  q21 <-     |
     Radical |   1.315215   .2619659     5.02   0.000     .8017709    1.828658
  -----------+------------------------------------------------------------------
  q22 <-     |
     Radical |   1.003016   .2360554     4.25   0.000     .5403556    1.465676
  -----------+------------------------------------------------------------------
  q23 <-     |
     Radical |   1.284173   .2427181     5.29   0.000     .8084539    1.759892
-------------+------------------------------------------------------------------
Variance     |
       e.q46 |   .5042322    .087704                      .3585718    .7090633
       e.q47 |   .2282654   .0722405                      .1227595    .4244485
       e.q48 |   .3338969   .0861639                      .2013516    .5536938
       e.q40 |   2.630155   .4187887                      1.925077    3.593475
       e.q42 |    2.11514    .349408                      1.530121    2.923831
       e.q43 |   1.064265   .2220603                      .7070423    1.601969
       e.q44 |   1.015812   .2316045                      .6497405    1.588133
       e.q45 |   2.159306   .3636282                      1.552284    3.003703
        e.q5 |   1.571541   .2550119                      1.143411    2.159977
        e.q7 |   .8801118   .1603103                      .6158763    1.257715
        e.q8 |    1.49632   .2210815                      1.120106    1.998893
       e.q12 |   .2463465   .1462531                       .076948    .7886698
       e.q13 |   2.194942   .3217872                      1.646773    2.925585
       e.q14 |   2.453943   .3624541                      1.837129    3.277852
     e.q19r  |   1.990048   .3289092                      1.439393    2.751361
       e.q20 |   1.660394   .3014176                      1.163293    2.369916
       e.q21 |   1.135532   .2721594                      .7098835    1.816401
       e.q22 |   2.539184   .4101386                      1.850141    3.484845
       e.q23 |   1.333542   .2825836                      .8803052    2.020134
     Success |   .9940317   .2094393                      .6577409    1.502262
        Org2 |   1.809841   .5457518                       1.00222    3.268267
     CompEnh |   1.720017   .4334061                       1.04966    2.818491
     Radical |   1.044673    .367468                      .5242839    2.081584
-------------+------------------------------------------------------------------
Covariance   |
  e.q8       |
       e.q14 |   .6273124   .2122972     2.95   0.003     .2112175    1.043407
  -----------+------------------------------------------------------------------
  Success    |
        Org2 |   .3338755   .1626949     2.05   0.040     .0149994    .6527516
     CompEnh |   .3328585   .1541534     2.16   0.031     .0307233    .6349937
     Radical |   .2285175     .13161     1.74   0.083    -.0294333    .4864684
  -----------+------------------------------------------------------------------
  Org2       |
     CompEnh |  -.1247411   .2004635    -0.62   0.534    -.5176423    .2681601
     Radical |   .4593763   .1944927     2.36   0.018     .0781777    .8405749
  -----------+------------------------------------------------------------------
  CompEnh    |
     Radical |   .1002444   .1598905     0.63   0.531    -.2131352    .4136241
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(145) =    216.49, Prob > chi2 = 0.0001
```

Fig. 10.7 (continued)

```
. estat gof, stats(all)

------------------------------------------------------------------------------
Fit statistic         |      Value   Description
----------------------+-------------------------------------------------------
Likelihood ratio      |
        chi2_ms(145)  |    216.489   model vs. saturated
           p > chi2   |      0.000
        chi2_bs(171)  |    999.476   baseline vs. saturated
           p > chi2   |      0.000
----------------------+-------------------------------------------------------
Population error      |
                RMSEA |      0.072   Root mean squared error of approximation
 90% CI, lower bound  |      0.051
         upper bound  |      0.091
              pclose  |      0.044   Probability RMSEA <= 0.05
----------------------+-------------------------------------------------------
Information criteria  |
                 AIC  |   6380.349   Akaike's information criterion
                 BIC  |   6495.745   Bayesian information criterion
----------------------+-------------------------------------------------------
Baseline comparison   |
                 CFI  |      0.914   Comparative fit index
                 TLI  |      0.898   Tucker-Lewis index
----------------------+-------------------------------------------------------
Size of residuals     |
                SRMR  |      0.093   Standardized root mean squared residual
                  CD  |      1.000   Coefficient of determination
------------------------------------------------------------------------------

. estat framework, fitted

Endogenous variables on endogenous variables
…

Exogenous variables on endogenous variables

                 | latent
           Gamma |   Success       Org2    CompEnh    Radical
    -------------+----------------------------------------------
    observed     |
             q46 |          1          0          0          0
             q47 |  1.165879          0          0          0
             q48 |  1.221725          0          0          0
             q40 |          0          1          0          0
             q42 |          0   1.030687          0          0
             q43 |          0   1.161106          0          0
             q44 |          0   1.244799          0          0
             q45 |          0   1.107081          0          0
              q5 |          0          0          1          0
              q7 |          0          0   .9699365          0
              q8 |          0          0  -.4558465          0
             q12 |          0          0   1.146397          0
             q13 |          0          0  -.4299586          0
             q14 |          0          0  -.5356665          0
            q19r |          0          0          0          1
             q20 |          0          0          0   1.152578
             q21 |          0          0          0   1.315215
             q22 |          0          0          0   1.003016
             q23 |          0          0          0   1.284173
    ------------------------------------------------------------

Covariances of error variables

 …

Covariances of exogenous variables

                 | latent
             Phi |   Success       Org2    CompEnh    Radical
    -------------+----------------------------------------------
    latent       |
         Success |  .9940317
```

Fig. 10.7 (continued)

```
           Org2 |  .3338755   1.809841
        CompEnh |  .3328585  -.1247411   1.720017
        Radical |  .2285175   .4593763   .1002444   1.044673
    ------------------------------------------------------------

Fitted covariances of observed and latent variables

…

                | latent
          Sigma |   Radical
    ------------+-----------
    latent      |
        Radical |  1.044673
    ------------------------

. estat mindices

Modification indices

------------------------------------------------------------------
             |                                          Standard
             |         MI     df   P>MI        EPC         EPC
-------------+----------------------------------------------------
Measurement  |
  q19r <-    |
     CompEnh |      6.762      1   0.01  -.3159788  -.2378839
  -----------+----------------------------------------------------
  q21 <-     |
     Success |      3.924      1   0.05  -.2883853  -.1676131
  -----------+----------------------------------------------------
  q22 <-     |
     Success |      8.105      1   0.00    .523498   .2754596
-------------+----------------------------------------------------
Covariance   |
  e.q46      |
       e.q45 |     12.756      1   0.00  -.4379273   -.419691
       e.q23 |     12.716      1   0.00   .3665776   .4470409
  -----------+----------------------------------------------------
  e.q48      |
       e.q45 |      7.918      1   0.00   .3169482   .3732717
        e.q8 |      4.925      1   0.03   .1856294   .2626204
  -----------+----------------------------------------------------
  e.q42      |
        e.q5 |      4.707      1   0.03     -.4444  -.2437483
  -----------+----------------------------------------------------
  e.q44      |
       e.q14 |      4.745      1   0.03   .4012734   .2541566
       e.q21 |      3.993      1   0.05  -.3116198  -.2901474
  -----------+----------------------------------------------------
  e.q45      |
       e.q21 |      8.008      1   0.00   .5655751   .3611882
       e.q23 |     13.449      1   0.00  -.7671903  -.4521085
  -----------+----------------------------------------------------
  e.q8       |
       e.q20 |      5.307      1   0.02    .380359   .2413103
  -----------+----------------------------------------------------
  e.q12      |
      e.q19r |      5.081      1   0.02  -.2952195  -.4216384
  -----------+----------------------------------------------------
  e.q20      |
       e.q21 |      9.678      1   0.00   .7361788   .5361398
       e.q22 |      5.489      1   0.02  -.5895306  -.2871136
------------------------------------------------------------------
EPC = expected parameter change
```

Fig. 10.7 (continued)

```
!Examp10-2.spl
!Raw Data From File: Examp10-6.txt
!Path Diagram

DA NI=19 MA = KM XM = 9
RA FI=C:\SAMD\Chapter10\Examples\Examp10-1.txt
MO NY = 8 NX = 11 NE = 2 NK = 2 PH = SY TD = SY

FI LY(1,1) LY( 2,1) LY(3,1)                                          C
   LY(4,2) LY( 5,2) LY(6,2) LY(7,2)  LY(8,2)                         C
   LX(1,1) LX( 2,1) LX(3,1) LX(4,1)  LX(5,1) LX(6,1)                 C
   LX(7,2) LX( 8,2) LX(9,2) LX(10,2) LX(11,2)                        C
   TE(1,1) TE( 2,2) TE(3,3) TE(4 ,4) TE( 5,5) TE(6,6) TE(7,7)        C
   TE(8,8)                                                           C
   TD(1,1) TD( 2,2) TD(3,3) TD(4 ,4) TD( 5,5) TD(6,6) TD(7,7)        C
   TD(8,8) TD( 9,9) TD(10,10) TD(11,11)                              C
   PH(1,1) PH( 2,1) PH( 2, 2)

VA  1       LY( 1,1) LX(1,1) LY(4,2)  LX(7,2)
VA  1.34    LY( 2,1)
VA  1.16    LY( 3,1)
VA  0.90    LY( 5,2)
VA  1.07    LY( 6,2)
VA  0.96    LY( 7,2)
VA  0.99    LY( 8,2)
VA  1.13    LX( 2,1)
VA -0.04    LX( 3,1)
VA  1.13    LX( 4,1)
VA -0.14    LX( 5,1)
VA -0.17    LX( 6,1)
VA  1.13    LX( 8,2)
VA  1.15    LX( 9,2)
VA  0.75    LX(10,2)
VA  1.00    LX(11,2)

VA  0.47    TE( 1,1)
VA  0.05    TE( 2,2)
VA  0.28    TE( 3,3)
VA  0.35    TE( 4,4)
VA  0.47    TE( 5,5)
VA  0.25    TE( 6,6)
VA  0.40    TE( 7,7)
VA  0.37    TE( 8,8)

VA  0.44    TD( 1,1)
VA  0.28    TD( 2,2)
VA  1.00    TD( 3,3)
VA  0.29    TD( 4,4)
VA  0.99    TD( 5,5)
VA  0.98    TD( 6,6)
VA  0.44    TD( 7,7)
VA  0.28    TD( 8,8)
VA  0.26    TD( 9,9)
VA  0.68    TD(10,10)
VA  0.44    TD(11,11)
VA  0.65    TD( 6,3)

VA  0.56    PH( 1,1)
VA  0.03    PH( 2,1)
VA  0.56    PH( 2,2)

LA
```

Fig. 10.8 Step 2: Input of full structural model—LISREL (examp10-2.spl)

```
Q46 Q47 Q48 Q40 Q42 Q43 Q44 Q45  Q5 Q7 Q8 Q12 Q13 Q14 Q19r Q20 Q21 Q22 Q23
LE
Success Org2
LK
CompEnh Radical

Path Diagram
OU SE TV AD = 50 MI
```

Fig. 10.8 (continued)

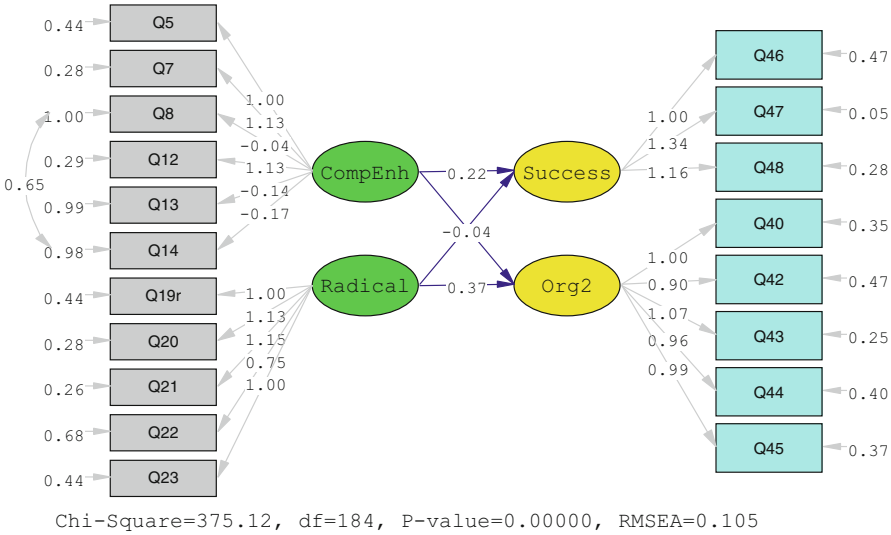


Fig. 10.9 Step 2: Graphical representation of full structural model—LISREL (examp10-2.pth)

```
                    L I S R E L  8.30

                          BY

                  Karl G. Jöreskog & Dag Sörbom

            This program is published exclusively by
            Scientific Software International, Inc.
               7383 N. Lincoln Avenue, Suite 100
                Chicago, IL 60646-1704, U.S.A.
        Phone: (800)247-6113, (847)675-0720, Fax: (847)675-2140
    Copyright by Scientific Software International, Inc., 1981-99
      Use of this program is subject to the terms specified in the
                  Universal Copyright Convention.
```

Fig. 10.10 Step 2: Output results of full structural model—LISREL (examp10-2.out)

```
                        Website: www.ssicentral.com

The following lines were read from file C:\SAMD\CHAPTER8\EXAMPLES\EXAMP8-7.SPL:

!Examp10-2.spl
!Raw Data From File: Examp10-1.txt
!Path Diagram

DA NI=19 MA = KM XM = 9
RA FI=C:\SAMD\Chapter10\Examples\Examp10-1.txt
MO NY = 8 NX = 11 NE = 2 NK = 2 PH = SY TD = SY

FI LY(1,1) LY( 2,1) LY(3,1)                  C
LY(4,2) LY( 5,2) LY(6,2) LY(7,2)  LY(8,2)      C
LX(1,1) LX( 2,1) LX(3,1) LX(4,1)  LX(5,1) LX(6,1)        C
LX(7,2) LX( 8,2) LX(9,2) LX(10,2) LX(11,2)         C
TE(1,1) TE( 2,2) TE(3,3) TE(4 ,4) TE( 5,5) TE(6,6) TE(7,7)  C
TE(8,8)        C
TD(1,1) TD( 2,2) TD(3,3) TD(4 ,4) TD( 5,5) TD(6,6) TD(7,7)  C
TD(8,8) TD( 9,9) TD(10,10) TD(11,11)     C
PH(1,1) PH( 2,1) PH( 2, 2)

VA  1     LY( 1,1) LX(1,1) LY(4,2)  LX(7,2)
VA  1.34   LY( 2,1)
VA  1.16   LY( 3,1)
VA  0.90   LY( 5,2)
VA  1.07   LY( 6,2)
VA  0.96   LY( 7,2)
VA  0.99   LY( 8,2)
VA  1.13   LX( 2,1)
VA -0.04   LX( 3,1)
VA  1.13   LX( 4,1)
VA -0.14   LX( 5,1)
VA -0.17   LX( 6,1)
VA  1.13   LX( 8,2)
VA  1.15   LX( 9,2)
VA  0.75   LX(10,2)
VA  1.00   LX(11,2)

VA  0.47   TE( 1,1)
VA  0.05   TE( 2,2)
VA  0.28   TE( 3,3)
VA  0.35   TE( 4,4)
VA  0.47   TE( 5,5)
VA  0.25   TE( 6,6)
VA  0.40   TE( 7,7)
VA  0.37   TE( 8,8)

VA  0.44   TD( 1,1)
VA  0.28   TD( 2,2)
VA  1.00   TD( 3,3)
VA  0.29   TD( 4,4)
VA  0.99   TD( 5,5)
VA  0.98   TD( 6,6)
VA  0.44   TD( 7,7)
VA  0.28   TD( 8,8)
VA  0.26   TD( 9,9)
VA  0.68   TD(10,10)
VA  0.44   TD(11,11)
VA  0.65   TD( 6,3)

VA  0.56   PH( 1,1)
VA  0.03   PH( 2,1)
VA  0.56   PH( 2,2)

LA
Q46 Q47 Q48 Q40 Q42 Q43 Q44 Q45  Q5 Q7 Q8 Q12 Q13 Q14 Q19r Q20 Q21 Q22 Q23
LE
Success Org2
LK
CompEnh Radical

Path Diagram
OU SE TV AD = 50 MI

!Examp10-2.spl
```

Fig. 10.10 (continued)

```
                    Number of Input Variables 19
                    Number of Y - Variables    8
                    Number of X - Variables   11
                    Number of ETA - Variables  2
                    Number of KSI - Variables  2
                    Number of Observations    96

        Covariance Matrix to be Analyzed

              Q46       Q47       Q48       Q40       Q42       Q43
        --------  --------  --------  --------  --------  --------
Q46         1.00
Q47         0.75      1.00
Q48         0.73      0.84      1.00
Q40         0.04      0.09      0.16      1.00
Q42         0.20      0.25      0.31      0.41      1.00
Q43         0.09      0.11      0.16      0.57      0.54      1.00
Q44         0.17      0.21      0.21      0.52      0.59      0.73
Q45         0.00      0.18      0.28      0.49      0.55      0.57
 Q5         0.21      0.13      0.13      0.11     -0.13      0.03
 Q7         0.26      0.20      0.25      0.10     -0.07      0.00
 Q8        -0.11     -0.04      0.05      0.21      0.19      0.26
Q12         0.23      0.20      0.24      0.05     -0.09     -0.03
Q13         0.06      0.06      0.04      0.07      0.12      0.14
Q14         0.01     -0.02      0.05      0.11      0.25      0.15
Q19r        0.19      0.17      0.12      0.05      0.26      0.14
Q20         0.13      0.08      0.08      0.20      0.15      0.21
Q21         0.08      0.07      0.04      0.17      0.22      0.24
Q22         0.28      0.34      0.31      0.07      0.24      0.14
Q23         0.33      0.17      0.13      0.12      0.14      0.16

        Covariance Matrix to be Analyzed

              Q44       Q45        Q5        Q7        Q8       Q12
        --------  --------  --------  --------  --------  --------
Q44         1.00
Q45         0.59      1.00
 Q5         0.04      0.00      1.00
 Q7        -0.04     -0.03      0.55      1.00
 Q8         0.19      0.22     -0.37     -0.31      1.00
Q12        -0.09     -0.02      0.69      0.77     -0.41      1.00
Q13         0.12      0.13     -0.25     -0.30      0.38     -0.32
Q14         0.29      0.20     -0.37     -0.32      0.45     -0.37
Q19r        0.19      0.27     -0.03     -0.06     -0.01     -0.19
Q20         0.19      0.17      0.00     -0.01      0.23     -0.05
Q21         0.18      0.30      0.17      0.14      0.02      0.13
Q22         0.23      0.10      0.09      0.19     -0.01      0.08
Q23         0.21      0.00      0.22      0.16     -0.08      0.08

        Covariance Matrix to be Analyzed

              Q13       Q14      Q19r       Q20       Q21       Q22
        --------  --------  --------  --------  --------  --------
Q13         1.00
Q14         0.38      1.00
Q19r        0.23      0.10      1.00
Q20         0.15      0.16      0.35      1.00
Q21        -0.03     -0.03      0.42      0.61      1.00
Q22         0.06      0.09      0.39      0.24      0.39      1.00
Q23         0.12     -0.05      0.48      0.49      0.56      0.47

        Covariance Matrix to be Analyzed

              Q23
        --------
Q23         1.00


Parameter Specifications

        GAMMA

            CompEnh    Radical
            --------   --------
 Success          1          2
```

Fig. 10.10 (continued)

```
    Org2          3          4

       PSI
       Note: This matrix is diagonal.

           Success       Org2
           --------   --------
                  5          6


Number of Iterations = 12

LISREL Estimates (Maximum Likelihood)

       LAMBDA-Y

           Success       Org2
           --------   --------
     Q46       1.00        - -

     Q47       1.34        - -

     Q48       1.16        - -

     Q40        - -       1.00

     Q42        - -       0.90

     Q43        - -       1.07

     Q44        - -       0.96

     Q45        - -       0.99


       LAMBDA-X

           CompEnh    Radical
           --------   --------
      Q5       1.00        - -

      Q7       1.13        - -

      Q8      -0.04        - -

     Q12       1.13        - -

     Q13      -0.14        - -

     Q14      -0.17        - -

    Q19r        - -       1.00

     Q20        - -       1.13

     Q21        - -       1.15

     Q22        - -       0.75

     Q23        - -       1.00


       GAMMA

           CompEnh    Radical
           --------   --------
 Success       0.22       0.18
             (0.11)     (0.10)
               2.06       1.73

    Org2      -0.04       0.37
             (0.11)     (0.11)
              -0.36       3.31
```

Fig. 10.10 (continued)

```
        Covariance Matrix of ETA and KSI

            Success       Org2    CompEnh    Radical
            --------   --------   --------   --------
Success       0.54
   Org2       0.04       0.60
CompEnh       0.13      -0.01       0.56
Radical       0.11       0.21       0.03       0.56

       PHI

            CompEnh    Radical
            --------   --------
CompEnh       0.56

Radical       0.03       0.56


       PSI
       Note: This matrix is diagonal.

            Success       Org2
            --------   --------
               0.49       0.52
             (0.08)     (0.09)
               6.50       5.98


       Squared Multiple Correlations for Structural Equations

            Success       Org2
            --------   --------
               0.09       0.13

       THETA-EPS

                Q46        Q47        Q48        Q40        Q42        Q43
            --------   --------   --------   --------   --------   --------
               0.47       0.05       0.28       0.35       0.47       0.25


       THETA-EPS

                Q44        Q45
            --------   --------
               0.40       0.37


       Squared Multiple Correlations for Y - Variables

                Q46        Q47        Q48        Q40        Q42        Q43
            --------   --------   --------   --------   --------   --------
               0.53       0.95       0.72       0.63       0.51       0.73

       Squared Multiple Correlations for Y - Variables

                Q44        Q45
            --------   --------
               0.58       0.61

       THETA-DELTA

                 Q5         Q7         Q8        Q12        Q13        Q14
            --------   --------   --------   --------   --------   --------
      Q5         0.44

      Q7          - -       0.28

      Q8          - -        - -       1.00

     Q12          - -        - -        - -       0.29

     Q13          - -        - -        - -        - -       0.99

     Q14          - -        - -       0.65        - -        - -       0.98
```

Fig. 10.10 (continued)

```
Q19r        - -         - -         - -         - -         - -         - -

Q20         - -         - -         - -         - -         - -         - -

Q21         - -         - -         - -         - -         - -         - -

Q22         - -         - -         - -         - -         - -         - -

Q23         - -         - -         - -         - -         - -         - -


   THETA-DELTA

          Q19r        Q20         Q21         Q22         Q23
       --------    --------    --------    --------    --------
Q19r       0.44

Q20         - -        0.28

Q21         - -         - -        0.26

Q22         - -         - -         - -        0.68

Q23         - -         - -         - -         - -        0.44


   Squared Multiple Correlations for X - Variables

           Q5          Q7          Q8         Q12         Q13         Q14
       --------    --------    --------    --------    --------    --------
          0.56        0.72        0.00        0.71        0.01        0.02

   Squared Multiple Correlations for X - Variables

          Q19r        Q20         Q21         Q22         Q23
       --------    --------    --------    --------    --------
          0.56        0.72        0.74        0.32        0.56


                    Goodness of Fit Statistics

                      Degrees of Freedom = 184
        Minimum Fit Function Chi-Square = 324.49 (P = 0.00)
  Normal Theory Weighted Least Squares Chi-Square = 375.12 (P = 0.00)
          Estimated Non-centrality Parameter (NCP) = 191.12
      90 Percent Confidence Interval for NCP = (139.61 ; 250.40)

               Minimum Fit Function Value = 3.42
          Population Discrepancy Function Value (F0) = 2.01
        90 Percent Confidence Interval for F0 = (1.47 ; 2.64)
        Root Mean Square Error of Approximation (RMSEA) = 0.10
      90 Percent Confidence Interval for RMSEA = (0.089 ; 0.12)
        P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00

          Expected Cross-Validation Index (ECVI) = 4.07
      90 Percent Confidence Interval for ECVI = (3.53 ; 4.70)
                 ECVI for Saturated Model = 4.00
              ECVI for Independence Model = 10.79

 Chi-Square for Independence Model with 171 Degrees of Freedom = 987.42
                      Independence AIC = 1025.42
                          Model AIC = 387.12
                        Saturated AIC = 380.00
                     Independence CAIC = 1093.14
                         Model CAIC = 408.50
                       Saturated CAIC = 1057.23

             Root Mean Square Residual (RMR) = 0.13
                   Standardized RMR = 0.13
              Goodness of Fit Index (GFI) = 0.75
          Adjusted Goodness of Fit Index (AGFI) = 0.74
        Parsimony Goodness of Fit Index (PGFI) = 0.72

                Normed Fit Index (NFI) = 0.67
             Non-Normed Fit Index (NNFI) = 0.84
          Parsimony Normed Fit Index (PNFI) = 0.72
```

Fig. 10.10 (continued)

```
                    Comparative Fit Index (CFI) = 0.83
                    Incremental Fit Index (IFI) = 0.83
                     Relative Fit Index (RFI) = 0.69

                         Critical N (CN) = 68.79


Modification Indices and Expected Change

        Modification Indices for LAMBDA-Y

            Success       Org2
            --------   --------
     Q46       0.46       0.17
     Q47       0.36       0.15
     Q48       0.05       3.16
     Q40       0.72       1.28
     Q42       3.70       0.01
     Q43       1.03       0.23
     Q44       1.31       1.27
     Q45       0.58       0.14

        Expected Change for LAMBDA-Y

            Success       Org2
            --------   --------
     Q46       0.07      -0.04
     Q47      -0.05      -0.03
     Q48       0.02       0.14
     Q40      -0.08      -0.10
     Q42       0.20      -0.01
     Q43      -0.09       0.04
     Q44       0.11       0.11
     Q45       0.07      -0.04

        Modification Indices for LAMBDA-X

            CompEnh    Radical
            --------   --------
      Q5       0.18       0.77
      Q7       0.06       0.34
      Q8       6.10       0.16
     Q12       0.20       0.86
     Q13       5.29       1.50
     Q14       0.20       0.10
    Q19r       7.40       1.82
     Q20       3.70       0.24
     Q21       4.57       0.17
     Q22       1.64       0.07
     Q23       2.92       0.07

        Expected Change for LAMBDA-X

            CompEnh    Radical
            --------   --------
      Q5      -0.05       0.09
      Q7      -0.02       0.06
      Q8      -0.27       0.04
     Q12       0.04      -0.09
     Q13      -0.34       0.18
     Q14      -0.05       0.03
    Q19r      -0.29      -0.14
     Q20      -0.18      -0.04
     Q21       0.19       0.04
     Q22       0.16      -0.03
     Q23       0.18       0.03

No Non-Zero Modification Indices for GAMMA

        Modification Indices for PHI

            CompEnh    Radical
            --------   --------
 CompEnh       0.03
 Radical       0.07       1.14
```

Fig. 10.10 (continued)

Expected Change for PHI

	CompEnh	Radical
CompEnh	0.02	
Radical	0.02	-0.10

Modification Indices for PSI

	Success	Org2
Success	- -	
Org2	2.67	- -

Expected Change for PSI

	Success	Org2
Success	- -	
Org2	0.09	- -

Modification Indices for THETA-EPS

	Q46	Q47	Q48	Q40	Q42	Q43
Q46	1.20					
Q47	0.16	0.09				
Q48	2.98	0.07	0.29			
Q40	0.05	0.55	0.12	12.82		
Q42	0.46	0.00	1.11	3.94	1.04	
Q43	0.85	0.50	0.18	1.48	1.43	2.58
Q44	0.67	0.99	0.98	2.66	0.98	5.32
Q45	8.88	0.21	4.00	1.21	1.09	4.51

Modification Indices for THETA-EPS

	Q44	Q45
Q44	0.62	
Q45	0.00	3.34

Expected Change for THETA-EPS

	Q46	Q47	Q48	Q40	Q42	Q43
Q46	-0.08					
Q47	-0.01	0.01				
Q48	0.07	-0.01	-0.02			
Q40	-0.01	-0.02	0.01	0.22		
Q42	0.04	0.00	0.04	-0.10	0.08	
Q43	0.04	-0.02	-0.01	-0.05	-0.05	0.08
Q44	0.04	0.03	-0.04	-0.07	0.05	0.09
Q45	-0.14	0.02	0.08	-0.05	0.05	-0.08

Expected Change for THETA-EPS

	Q44	Q45
Q44	-0.05	
Q45	0.00	0.12

Modification Indices for THETA-DELTA-EPS

	Q46	Q47	Q48	Q40	Q42	Q43
Q5	0.58	0.01	0.79	0.33	4.75	0.17
Q7	0.55	0.92	0.92	1.17	0.08	0.02
Q8	3.69	0.01	1.32	1.35	1.25	5.03
Q12	0.06	0.03	0.43	0.02	0.06	0.01
Q13	0.12	0.26	0.01	0.25	0.01	0.31
Q14	2.46	0.53	0.06	2.15	3.19	4.43
Q19r	0.07	1.10	0.31	5.93	3.86	2.20
Q20	0.00	0.70	0.28	4.10	2.55	0.28
Q21	2.26	0.01	0.45	0.27	0.14	0.02
Q22	0.05	1.12	0.54	1.02	2.52	0.09
Q23	10.55	1.01	0.41	0.54	0.00	0.50

Fig. 10.10 (continued)

```
   Modification Indices for THETA-DELTA-EPS

           Q44        Q45
      --------   --------
Q5       1.64       0.06
Q7       0.12       0.72
Q8       3.41       0.00
Q12      1.07       0.65
Q13      0.03       0.19
Q14      7.70       0.23
Q19r     0.08       6.86
Q20      0.26       1.21
Q21      3.48       8.04
Q22      2.52       1.24
Q23      2.76      16.59

   Expected Change for THETA-DELTA-EPS

           Q46        Q47        Q48        Q40        Q42        Q43
      --------   --------   --------   --------   --------   --------
Q5       0.04       0.00      -0.04       0.03      -0.12       0.02
Q7       0.04      -0.03       0.04       0.05       0.01      -0.01
Q8      -0.10       0.00       0.05       0.06      -0.06       0.10
Q12     -0.01      -0.01       0.02      -0.01       0.01       0.00
Q13      0.03       0.03      -0.01      -0.03       0.01       0.03
Q14      0.08      -0.03       0.01      -0.07       0.10      -0.10
Q19r     0.01       0.04      -0.02      -0.12       0.10      -0.06
Q20      0.00      -0.03       0.02       0.08      -0.07       0.02
Q21     -0.07       0.00      -0.02      -0.02      -0.02       0.00
Q22     -0.01       0.05       0.04      -0.06       0.10      -0.02
Q23      0.17      -0.04      -0.03       0.04       0.00       0.03

   Expected Change for THETA-DELTA-EPS

           Q44        Q45
      --------   --------
Q5       0.07      -0.01
Q7      -0.02      -0.04
Q8      -0.10       0.00
Q12     -0.05       0.04
Q13      0.01       0.03
Q14      0.15       0.02
Q19r     0.01       0.13
Q20     -0.02      -0.05
Q21     -0.08       0.12
Q22      0.09      -0.06
Q23      0.08      -0.20

   Modification Indices for THETA-DELTA

           Q5         Q7         Q8        Q12        Q13        Q14
      --------   --------   --------   --------   --------   --------
Q5       0.20
Q7       6.16       0.03
Q8       0.43       0.03      12.62
Q12      2.93       1.86       3.30       3.81
Q13      0.02       0.46       2.99       0.90       0.11
Q14      0.36       0.01      20.67       0.28       1.38      10.70
Q19r     0.03       0.13       3.70       5.56       4.21       2.83
Q20      1.00       0.27       9.50       0.04       1.08       0.29
Q21      0.25       0.37       0.08       3.18       7.15       2.06
Q22      0.64       2.64       2.28       0.15       0.01       2.62
Q23      3.22       0.34       2.28       0.80       0.64       0.00

   Modification Indices for THETA-DELTA

          Q19r        Q20        Q21        Q22        Q23
      --------   --------   --------   --------   --------
Q19r    12.84
Q20     14.36      21.95
Q21      7.22       0.20       6.99
Q22      2.87      13.48       0.25       0.58
Q23      1.44       3.37       0.55       4.76       0.89

   Expected Change for THETA-DELTA
```

Fig. 10.10 (continued)

```
             Q5          Q7          Q8         Q12         Q13         Q14
        --------    --------    --------    --------    --------    --------
  Q5       0.03
  Q7      -0.12        0.01
  Q8      -0.04        0.01        0.29
 Q12       0.08        0.05       -0.09       -0.12
 Q13      -0.01       -0.05        0.13       -0.06       -0.05
 Q14      -0.03        0.00       -0.22        0.03        0.09        0.27
Q19r       0.01        0.02       -0.11       -0.11        0.15        0.09
 Q20      -0.05       -0.02        0.15        0.01        0.07       -0.03
 Q21       0.02       -0.02        0.01        0.07       -0.17       -0.07
 Q22      -0.05        0.09       -0.10       -0.02        0.01        0.10
 Q23       0.10        0.03       -0.08       -0.04        0.06        0.00

    Expected Change for THETA-DELTA

            Q19r         Q20         Q21         Q22         Q23
        --------    --------    --------    --------    --------
Q19r       0.26
 Q20      -0.17        0.26
 Q21      -0.12       -0.02        0.14
 Q22       0.10       -0.19       -0.03        0.08
 Q23       0.06       -0.08       -0.03        0.13        0.07

Maximum Modification Index is   21.95 for Element ( 8, 8) of THETA-DELTA
```

Fig. 10.10 (continued)

```
*Examp10-1.do
insheet q46 q47 q48 q40 q42 q43 q44 q45 q5 q7 q8 q12 q13 q14 q19r q20 q21 q22 q23
using "/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter10_ACS-SEM/Examp10-1.txt",
clear
replace q46 =. if q46 == 9
replace q47 =. if q47 == 9
replace q48 =. if q48 == 9
replace q40 =. if q40 == 9
replace q42 =. if q42 == 9
replace q43 =. if q43 == 9
replace q44 =. if q44 == 9
replace q45 =. if q45 == 9
replace q5 =. if q5 == 9
replace q7 =. if q7 == 9
replace q8 =. if q8 == 9
replace q12 =. if q12 == 9
replace q13 =. if q13 == 9
replace q14 =. if q14 == 9
replace q19r =. if q19r == 9
replace q20 =. if q20 == 9
replace q21 =. if q21 == 9
replace q22 =. if q22 == 9
replace q23 =. if q23 == 9
egen stdq46 =std(q46)
egen stdq47 =std(q47)
egen stdq48 =std(q48)
egen stdq40 =std(q40)
egen stdq42 =std(q42)
egen stdq43 =std(q43)
egen stdq44 =std(q44)
egen stdq45 =std(q45)
egen stdq5 =std(q5)
egen stdq7 =std(q7)
egen stdq8 =std(q8)
egen stdq12 =std(q12)
egen stdq13 =std(q13)
egen stdq14 =std(q14)
```

Fig. 10.11 Step 2: Input of full structural model—STATA (examp10-2.do)

```
egen stdq19r =std(q19r)
egen stdq20 =std(q20)
egen stdq21 =std(q21)
egen stdq22 =std(q22)
egen stdq23 =std(q23)
sem (Success -> q46@1 q47@1.165879 q48@1.221725) ///
(Org2 -> q40@1 q42@1.030687 q43@1.161106 q44@1.244799 q45@1.107081) ///
(CompEnh -> q5@1 q7@.9699365 q8@-.4558465 q12@1.146397 q13@-.4299586 q14@-.535665) ///
(Radical -> q19r@1 q20@1.152578 q21@1.315215 q22@1.003016 q23@1.284173) ///
(Success <- CompEnh Radical) ///
(Org2 <- CompEnh Radical) ///
, nomeans latent(Success Org2 CompEnh Radical) ///
var(e.q46@.5042322) ///
var(e.q47@.2282654) ///
var(e.q48@.3338969) ///
var(e.q40@2.630155) ///
var(e.q42@2.11514) ///
var(e.q43@1.064265) ///
var(e.q44@1.015812) ///
var(e.q45@2.159306) ///
var(e.q5@1.571541) ///
var(e.q7@.8801118) ///
var(e.q8@1.49632) ///
var(e.q12@.2463465) ///
var(e.q13@2.194942) ///
var(e.q14@2.453943) ///
var(e.q19r@1.990048) ///
var(e.q20@1.660394) ///
var(e.q21@1.135532) ///
var(e.q22@2.539184) ///
var(e.q23@1.333542) ///
var(CompEnh@1.720017) ///
var(Radical@1.044673) ///
cov(CompEnh*Radical@.1002444) ///
cov(e.q8*e.q14@.6273124)
estat gof, stats(all)
estat framework, fitted
estat mindices
```

Fig. 10.11 (continued)

```
. *Examp10-2.do
. insheet q46 q47 q48 q40 q42 q43 q44 q45 q5 q7 q8 q12 q13 q14 q19r q20 q21 q22 q23
using "/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter10_ACS-SEM/Examp10-1.txt",
clear
(19 vars, 146 obs)

. replace q46 =. if q46 == 9
(12 real changes made, 12 to missing)
…

. replace q23 =. if q23 == 9
(7 real changes made, 7 to missing)

. egen stdq46 =std(q46)
(12 missing values generated)

…

. egen stdq23 =std(q23)
(7 missing values generated)
```

Fig. 10.12 Step 2: Output of full structural model—STATA (examp10-2.log)

```
. sem (Success -> q46@1 q47@1.165879 q48@1.221725) ///
> (Org2 -> q40@1 q42@1.030687 q43@1.161106 q44@1.244799 q45@1.107081) ///
> (CompEnh -> q5@1 q7@.9699365 q8@-.4558465 q12@1.146397 q13@-.4299586 q14@-.535665)
///
> (Radical -> q19r@1 q20@1.152578 q21@1.315215 q22@1.003016 q23@1.284173) ///
> (Success <- CompEnh Radical) ///
> (Org2 <- CompEnh Radical) ///
> , nomeans latent(Success Org2 CompEnh Radical) ///
> var(e.q46@.5042322) ///
> var(e.q47@.2282654) ///
> var(e.q48@.3338969) ///
> var(e.q40@2.630155) ///
> var(e.q42@2.11514) ///
> var(e.q43@1.064265) ///
> var(e.q44@1.015812) ///
> var(e.q45@2.159306) ///
> var(e.q5@1.571541) ///
> var(e.q7@.8801118) ///
> var(e.q8@1.49632) ///
> var(e.q12@.2463465) ///
> var(e.q13@2.194942) ///
> var(e.q14@2.453943) ///
> var(e.q19r@1.990048) ///
> var(e.q20@1.660394) ///
> var(e.q21@1.135532) ///
> var(e.q22@2.539184) ///
> var(e.q23@1.333542) ///
> var(CompEnh@1.720017) ///
> var(Radical@1.044673) ///
> cov(CompEnh*Radical@.1002444) ///
> cov(e.q8*e.q14@.6273124)
(50 observations with missing values excluded;
 specify option 'method(mlmv)' to use all observations)

Endogenous variables

Measurement:  q46 q47 q48 q40 q42 q43 q44 q45 q5 q7 q8 q12 q13 q14 q19r q20 q21 q22
q23
Latent:       Success Org2

Exogenous variables

Latent:       CompEnh Radical

Fitting target model:

Iteration 0:   log likelihood = -3147.2285
Iteration 1:   log likelihood = -3147.0378
Iteration 2:   log likelihood = -3147.0376

Structural equation model                       Number of obs      =        96
Estimation method  = ml
Log likelihood     = -3147.0376

 ( 1)  [q46]Success = 1
 ( 2)  [q47]Success = 1.165879
 ( 3)  [q48]Success = 1.221725
 ( 4)  [q40]Org2 = 1
 ( 5)  [q42]Org2 = 1.030687
 ( 6)  [q43]Org2 = 1.161106
 ( 7)  [q44]Org2 = 1.244799
 ( 8)  [q45]Org2 = 1.107081
 ( 9)  [q5]CompEnh = 1
 (10)  [q7]CompEnh = .9699365
 (11)  [q8]CompEnh = -.4558465
 (12)  [q12]CompEnh = 1.146397
 (13)  [q13]CompEnh = -.4299586
 (14)  [q14]CompEnh = -.535665
 (15)  [q19r]Radical = 1
 (16)  [q20]Radical = 1.152578
 (17)  [q21]Radical = 1.315215
 (18)  [q22]Radical = 1.003016
 (19)  [q23]Radical = 1.284173



 (20)  [var(e.q46)]_cons = .5042322



 (21)  [var(e.q47)]_cons = .2282654
```

Fig. 10.12 (continued)

```
 (22)  [var(e.q48)]_cons = .3338969
 (23)  [var(e.q40)]_cons = 2.630155
 (24)  [var(e.q42)]_cons = 2.11514
 (25)  [var(e.q43)]_cons = 1.064265
 (26)  [var(e.q44)]_cons = 1.015812
 (27)  [var(e.q45)]_cons = 2.159306
 (28)  [var(e.q5)]_cons = 1.571541
 (29)  [var(e.q7)]_cons = .8801118
 (30)  [var(e.q8)]_cons = 1.49632
 (31)  [cov(e.q8,e.q14)]_cons = .6273124
 (32)  [var(e.q12)]_cons = .2463465
 (33)  [var(e.q13)]_cons = 2.194942
 (34)  [var(e.q14)]_cons = 2.453943
 (35)  [var(e.q19r)]_cons = 1.990048
 (36)  [var(e.q20)]_cons = 1.660394
 (37)  [var(e.q21)]_cons = 1.135532
 (38)  [var(e.q22)]_cons = 2.539184
 (39)  [var(e.q23)]_cons = 1.333542
 (40)  [var(CompEnh)]_cons = 1.720017
 (41)  [cov(CompEnh,Radical)]_cons = .1002444
 (42)  [var(Radical)]_cons = 1.044673
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  Success <- |
    CompEnh  |   .1799156   .0801579     2.24   0.025      .022809    .3370223
    Radical  |   .2155963   .1090425     1.98   0.048     .0018769    .4293157
  ---------- +----------------------------------------------------------------
  Org2 <-    |
    CompEnh  |  -.0955978   .1105855    -0.86   0.387    -.3123414    .1211459
    Radical  |   .4615901   .1506379     3.06   0.002     .1663454    .7568349
-------------+----------------------------------------------------------------
Measurement  |
  q46 <-     |
    Success  |           1  (constrained)
  ---------- +----------------------------------------------------------------
  q47 <-     |
    Success  |   1.165879  (constrained)
  ---------- +----------------------------------------------------------------
  q48 <-     |
    Success  |   1.221725  (constrained)
  ---------- +----------------------------------------------------------------
  q40 <-     |
      Org2   |           1  (constrained)
  ---------- +----------------------------------------------------------------
  q42 <-     |
      Org2   |   1.030687  (constrained)
  ---------- +----------------------------------------------------------------
  q43 <-     |
      Org2   |   1.161106  (constrained)
  ---------- +----------------------------------------------------------------
  q44 <-     |
      Org2   |   1.244799  (constrained)
  ---------- +----------------------------------------------------------------
  q45 <-     |
      Org2   |   1.107081  (constrained)
  ---------- +----------------------------------------------------------------
  q5 <-      |
    CompEnh  |           1  (constrained)
  ---------- +----------------------------------------------------------------
  q7 <-      |
    CompEnh  |   .9699365  (constrained)
  ---------- +----------------------------------------------------------------
  q8 <-      |
    CompEnh  |  -.4558465  (constrained)
  ---------- +----------------------------------------------------------------
  q12 <-     |
    CompEnh  |   1.146397  (constrained)
  ---------- +----------------------------------------------------------------



  q13 <-     |



    CompEnh  |  -.4299586  (constrained)
  ---------- +----------------------------------------------------------------
  q14 <-     |
```

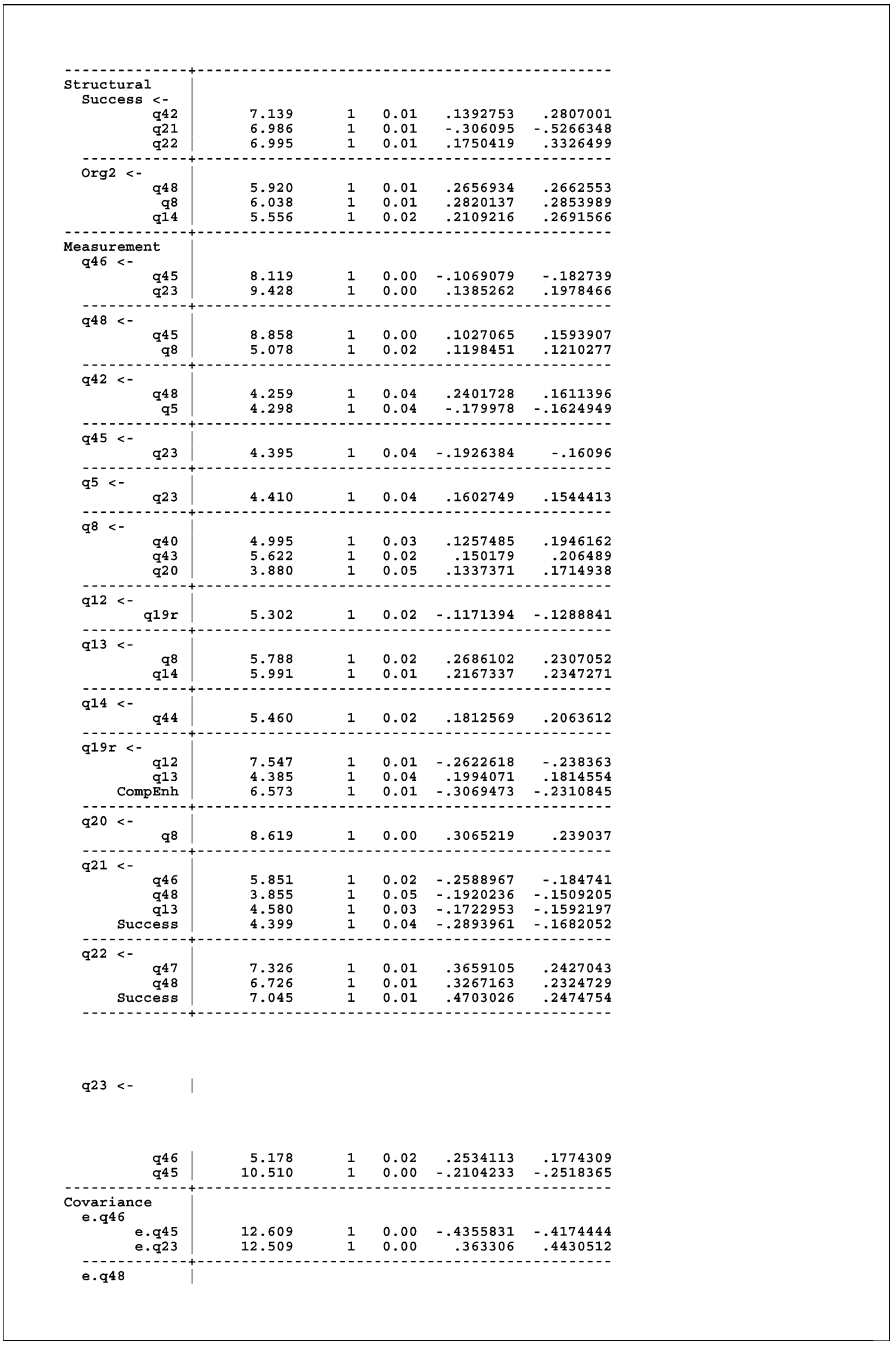
Fig. 10.12 (continued)

```
      CompEnh |   -.535665  (constrained)
  ------------+------------------------------------------------------------
    q19r <-   |
      Radical |          1  (constrained)
  ------------+------------------------------------------------------------
    q20 <-    |
      Radical |   1.152578  (constrained)
  ------------+------------------------------------------------------------
    q21 <-    |
      Radical |   1.315215  (constrained)
  ------------+------------------------------------------------------------
    q22 <-    |
      Radical |   1.003016  (constrained)
  ------------+------------------------------------------------------------
    q23 <-    |
      Radical |   1.284173  (constrained)
--------------+------------------------------------------------------------
Variance      |
         e.q46 |    .5042322  (constrained)
         e.q47 |    .2282654  (constrained)
         e.q48 |    .3338969  (constrained)
         e.q40 |    2.630155  (constrained)
         e.q42 |     2.11514  (constrained)
         e.q43 |    1.064265  (constrained)
         e.q44 |    1.015812  (constrained)
         e.q45 |    2.159306  (constrained)
          e.q5 |    1.571541  (constrained)
          e.q7 |    .8801118  (constrained)
          e.q8 |     1.49632  (constrained)
         e.q12 |    .2463465  (constrained)
         e.q13 |    2.194942  (constrained)
         e.q14 |    2.453943  (constrained)
        e.q19r |    1.990048  (constrained)
         e.q20 |    1.660394  (constrained)
         e.q21 |    1.135532  (constrained)
         e.q22 |    2.539184  (constrained)
         e.q23 |    1.333542  (constrained)
     e.Success |    .8820751   .1410789                      .6447102    1.206831
        e.Org2 |    1.580562    .269248                      1.131904    2.207056
       CompEnh |    1.720017  (constrained)
       Radical |    1.044673  (constrained)
--------------+------------------------------------------------------------
Covariance    |
    e.q8      |
         e.q14 |    .6273124  (constrained)
  ------------+------------------------------------------------------------
    CompEnh   |
       Radical |    .1002444  (constrained)
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(184) =    220.22, Prob > chi2 = 0.0351

. estat gof, stats(all)

------------------------------------------------------------------------------
Fit statistic         |      Value   Description
----------------------+-------------------------------------------------------
Likelihood ratio      |
         chi2_ms(184) |    220.215   model vs. saturated
             p > chi2 |      0.035
         chi2_bs(171) |    999.476   baseline vs. saturated
             p > chi2 |      0.000
----------------------+-------------------------------------------------------
Population error      |
                RMSEA |      0.045   Root mean squared error of approximation
 90% CI, lower bound  |      0.013
         upper bound  |      0.066
               pclose |      0.622   Probability RMSEA <= 0.05
----------------------+-------------------------------------------------------
Information criteria  |


                  AIC |   6306.075   Akaike's information criterion


                  BIC |   6321.461   Bayesian information criterion
----------------------+-------------------------------------------------------
Baseline comparison   |
                  CFI |      0.956   Comparative fit index
                  TLI |      0.959   Tucker-Lewis index
```

Fig. 10.12 (continued)

```
--------------------+-----------------------------------------------------
Size of residuals   |
               SRMR |       0.100   Standardized root mean squared residual
                 CD |       0.987   Coefficient of determination
-------------------------------------------------------------------------------

. estat framework, fitted

Endogenous variables on endogenous variables

…

Exogenous variables on endogenous variables

             |  latent
       Gamma |    CompEnh     Radical
-------------+----------------------
observed     |
          q46 |           0           0
          q47 |           0           0
          q48 |           0           0
          q40 |           0           0
          q42 |           0           0
          q43 |           0           0
          q44 |           0           0
          q45 |           0           0
           q5 |           1           0
           q7 |   .9699365           0
           q8 |  -.4558465           0
          q12 |   1.146397           0
          q13 |  -.4299586           0
          q14 |   -.535665           0
         q19r |           0           1
          q20 |           0   1.152578
          q21 |           0   1.315215
          q22 |           0   1.003016
          q23 |           0   1.284173
-------------+----------------------
latent       |
     Success |   .1799156   .2155963
        Org2 |  -.0955978   .4615901
-------------------------------------

Covariances of error variables

…

Covariances of exogenous variables

             |  latent
         Phi |    CompEnh     Radical
-------------+----------------------
latent       |
     CompEnh |   1.720017
     Radical |   .1002444   1.044673
-------------------------------------

Fitted covariances of observed and latent variables

…

             |  latent
       Sigma |    Radical
-------------+-----------
latent       |
     Radical |   1.044673
-------------------------



. estat mindices

Modification indices

-------------------------------------------------------------------
             |                                              Standard
             |         MI     df   P>MI        EPC            EPC
```

Fig. 10.12 (continued)



```
-------------+----------------------------------------------
Structural   |
  Success <- |
          q42|      7.139      1   0.01   .1392753   .2807001
          q21|      6.986      1   0.01   -.306095  -.5266348
          q22|      6.995      1   0.01   .1750419   .3326499
  -----------+----------------------------------------------
  Org2 <-    |
          q48|      5.920      1   0.01   .2656934   .2662553
           q8|      6.038      1   0.01   .2820137   .2853989
          q14|      5.556      1   0.02   .2109216   .2691566
-------------+----------------------------------------------
Measurement  |
  q46 <-     |
          q45|      8.119      1   0.00  -.1069079   -.182739
          q23|      9.428      1   0.00   .1385262   .1978466
  -----------+----------------------------------------------
  q48 <-     |
          q45|      8.858      1   0.00   .1027065   .1593907
           q8|      5.078      1   0.02   .1198451   .1210277
  -----------+----------------------------------------------
  q42 <-     |
          q48|      4.259      1   0.04   .2401728   .1611396
           q5|      4.298      1   0.04   -.179978  -.1624949
  -----------+----------------------------------------------
  q45 <-     |
          q23|      4.395      1   0.04  -.1926384    -.16096
  -----------+----------------------------------------------
  q5 <-      |
          q23|      4.410      1   0.04   .1602749   .1544413
  -----------+----------------------------------------------
  q8 <-      |
          q40|      4.995      1   0.03   .1257485   .1946162
          q43|      5.622      1   0.02    .150179    .206489
          q20|      3.880      1   0.05   .1337371   .1714938
  -----------+----------------------------------------------
  q12 <-     |
         q19r|      5.302      1   0.02  -.1171394  -.1288841
  -----------+----------------------------------------------
  q13 <-     |
           q8|      5.788      1   0.02   .2686102   .2307052
          q14|      5.991      1   0.01   .2167337   .2347271
  -----------+----------------------------------------------
  q14 <-     |
          q44|      5.460      1   0.02   .1812569   .2063612
  -----------+----------------------------------------------
  q19r <-    |
          q12|      7.547      1   0.01  -.2622618   -.238363
          q13|      4.385      1   0.04   .1994071   .1814554
      CompEnh|      6.573      1   0.01  -.3069473  -.2310845
  -----------+----------------------------------------------
  q20 <-     |
           q8|      8.619      1   0.00   .3065219    .239037
  -----------+----------------------------------------------
  q21 <-     |
          q46|      5.851      1   0.02  -.2588967   -.184741
          q48|      3.855      1   0.05  -.1920236  -.1509205
          q13|      4.580      1   0.03  -.1722953  -.1592197
      Success|      4.399      1   0.04  -.2893961  -.1682052
  -----------+----------------------------------------------
  q22 <-     |
          q47|      7.326      1   0.01   .3659105   .2427043
          q48|      6.726      1   0.01   .3267163   .2324729
      Success|      7.045      1   0.01   .4703026   .2474754
  -----------+----------------------------------------------



  q23 <-     |



          q46|      5.178      1   0.02   .2534113   .1774309
          q45|     10.510      1   0.00  -.2104233  -.2518365
-------------+----------------------------------------------
Covariance   |
  e.q46      |
        e.q45|     12.609      1   0.00  -.4355831  -.4174444
        e.q23|     12.509      1   0.00    .363306   .4430512
  -----------+----------------------------------------------
  e.q48      |
```

Fig. 10.12 (continued)

```
          e.q45 |       8.182      1    0.00    .3223735    .3796611
           e.q8 |       5.256      1    0.02      .19191    .2715059
         e.Org2 |       4.599      1    0.03    .2106348    .2899467
    ------------+----------------------------------------------------
    e.q42       |
          e.q5 |       4.794      1    0.03   -.4486444   -.2460763
     e.Success |       4.381      1    0.04    .3240388    .2372327
    ------------+----------------------------------------------------
    e.q44       |
         e.q14 |       4.889      1    0.03    .4078541    .2583247
         e.q21 |       4.685      1    0.03   -.3368807   -.3136677
    ------------+----------------------------------------------------
    e.q45       |
        e.q19r |       3.878      1    0.05    .4648556    .2242483
         e.q21 |       7.459      1    0.01    .5450434    .3480762
         e.q23 |      13.474      1    0.00    -.767121   -.4520677
    ------------+----------------------------------------------------
    e.q8        |
         e.q20 |       5.218      1    0.02    .3769523     .239149
    ------------+----------------------------------------------------
    e.q12       |
        e.q19r |       5.038      1    0.02   -.2936345   -.4193746
    ------------+----------------------------------------------------
    e.q20       |
         e.q21 |       5.372      1    0.02    .3994706     .290924
         e.q22 |       4.841      1    0.03   -.5136097   -.2501385
    ------------+----------------------------------------------------
    e.q21       |
     e.Success |       6.986      1    0.01   -.3475815   -.3472995
    ------------+----------------------------------------------------
    e.q22       |
     e.Success |       6.995      1    0.01    .4444633     .296986
    ------------------------------------------------------------------
    EPC = expected parameter change
```

Fig. 10.12 (continued)

10.5 Assignment

Using the SURVEY data described in Chap. 14 (Appendix C), develop a model that specifies structural relationships between unobservable constructs measured with multiple items. Develop a model with multiple equations and verify the identification of the structural model. Estimate the measurement model corresponding to a confirmatory factor analysis, and then estimate the structural model parameters.

Bibliography

Basic Technical Readings

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411–423.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*(Spring), 74–94.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*, 421–458.

Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of Marketing Research, 19*, 425–430.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31*, 419–456.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314.

Gerbin, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika, 52*(1), 99–111.

Joreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). NY: Seminar Press.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Dean Stilwell, C. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*(3), 430–445.

Application Readings

Ahearne, M., Gruen, T. W., & Jarvis, C. B. (1999). If looks could sell: Moderation and mediation of the attractiveness effect on salesperson performance. *International Journal of Research in Marketing, 16*(4), 269–284.

Anderson, J. C. (1987). An approach for confirmatory measurement and structural equation modeling of organizational properties. *Management Science, 33*(4), 525–541.

Anderson, J. C., & Narus, J. A. (1990). A model of distributor firm and manufacturer firm working partnerships. *Journal of Marketing, 54*, 42–58.

Bagozzi, R. P., & Dholakia, U. M. (2006). Open source software user communities: A study of participation in linux user groups. *Management Science, 52*(7), 1099–1115.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13*, 139–161.

Bentler, P. M., & Mooijaart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. *Psychological Bulletin, 106*(2), 315–317.

Capron, L. (1999). The long-term performance of horizontal acquisitions. *Strategic Management Journal, 20*(11), 987–1018.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*(2), 317–327.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–174.

Elsbach, K. D., & Bhattacharya, C. B. (2001). Defining who you are by what you're not : Organizational disidentification and the National Rifle Association. *Organization Science, 12* (4), 393–413.

Garbarino, E., & Johnson, M. S. (1999). The different roles of satisfaction, trust, and commitment in customer relationships. *Journal of Marketing, 63*(2), 70–87.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 25*(2), 186–192.

Gilbert, F. W., & Warren, W. E. (1995). Psychographic constructs and demographic segments. *Psychology and Marketing, 12*(3), 223–237.

Kuester, S., Homburg, C., & Robertson, T. S. (1999). Retaliatory behavior to new product entry. *Journal of Marketing, 63*(4), 90–106.

Kumar, A., & Dillon, W. R. (1990). On the use of confirmatory measurement models in the analysis of multiple-informant reports. *Journal of Marketing Research, 27*(1), 102–111.

MacKenzie, S. B., Lutz, R. J., & Belch, G. E. (1986). The role of attitude toward the ad as a mediator of advertising effectiveness: A test of competing explanations. *Journal of Marketing Research, 23*(2), 130–143.

Meyer, S. M., & Collier, D. A. (2001). An empirical test of the causal relationships in the Balridge health care pilot criteria. *Journal of Operations Management, 19*, 403–425.

Murtha, T. P., Lenway, S. A., & Bagozzi, R. P. (1998). Global mind-sets and cognitive shift in a complex multinational corporation. *Strategic Management Journal, 19*, 97–114.

Philips, L. W. (1981). Assessing measurement error in key informant reports: A methodological note on organizational analysis in marketing. *Journal of Marketing Research, 18*(4), 395–415.

Philips, L. W., Chang, D. R., & Buzzell, R. D. (1983). Product quality, cost position and business performance. *Journal of Marketing, 47*(2), 26–43.

Reddy, S. K., & LaBarbera, P. A. (1985). Hierarchical models of attitude. *Multivariate Behavioral Research, 20*, 451–471.

Stimpert, J. L., & Duhaime, I. M. (1997). In the eyes of the beholder: Conceptualizations of the relatedness held by the managers of large diversified firms. *Strategic Management Journal, 18* (2), 111–125.

Titman, S., & Wessels, R. (1988). The determinants of capital structure choice. *The Journal of Finance, 43*(1), 1–19.

Trieschmann, J. S., Dennis, A. R., Northcraft, G. B., & Niemi, A. W., Jr. (2000). Serving multiple constituencies in business schools: M.B.A. program versus research performance. *Academy of Management Journal, 43*(6), 1130–1141.

Vanden Abeele, P. (1989). Comment on: An investigation of the structure of expectancy-value attitude and its implications. *International Journal of Research in Marketing, 6*, 85–87.

Venkatraman, N., & Ramanujam, V. (1987). Planning system success: A conceptualization and an operational model. *Management Science, 33*(6), 687–705.

Walters, R. G., & MacKenzie, S. B. (1988). A structural equations analysis of the impact of price promotions on store performance. *Journal of Marketing Research, 25*, 51–63.

Yi, Y. (1989a). An investigation of the structure of expectancy-value attitude and its implications. *International Journal of Research in Marketing, 6*, 71–83.

Yi, Y. (1989b). Rejoinder to: An investigation of the structure of expectancy-value attitude and its implications. *International Journal of Research in Marketing, 6*, 89–94.

Chapter 11
Testing Mediation and Moderation Effects

In this chapter we present the methods for modeling and testing the relationships among variables that are characterized by mediation or moderation effects. As is illustrated throughout the chapter, the social sciences abound with theories that require the modeling and testing of such effects. We first describe the distinction between mediation and moderation and define the concept of moderated mediation as well as mediated moderation. We then discuss how such relationships can be estimated and the methods that are typically employed to test hypotheses corresponding to such effects.

11.1 Mediation vs. Moderation Effects

In this section we introduce the concepts of mediation and moderation and we illustrate them using examples selected from the literature.

11.1.1 Mediation Effects

In studying a mediation process, the focus is typically on the relationship between an independent variable X and a dependent variable Y. The independent variable can be either a continuous variable with at least interval-scale properties or a categorical or an ordinal variable such as with experimental data. The dependent variable is interval scale or ratio scale. In the simplest case that corresponds to most analyses encountered in the literature, the focus is on a single relationship between two variables, one independent variable represented by X and one dependent variable represented by Y. A theory usually predicts that there is a nonzero link (i.e., a direct effect) between these two variables with a causality that goes from X to Y. The mediation effect hypothesized typically corresponds to the explanatory mechanism of this causal relationship (i.e., the theory being tested).

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_11,

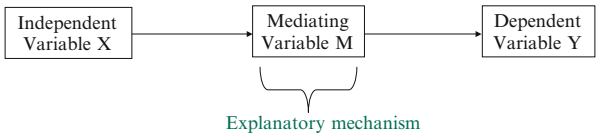


Fig. 11.1 Graphical representation of mediation effect

This explanatory mechanism is reflected by a measure of the process that causes the effect of X on Y. Such an intermediary explanatory mechanism corresponds to a mediating process that is represented by a variable M measured on an interval or a ratio scale. The relationships between these three variables X, Y, and M are illustrated in Fig. 11.1.

The effect of X on Y represented by the intervening or the mediating variable M is the indirect effect. Examples of such mediating effects can be found in all the disciplines of the social sciences. We will use two examples to illustrate the relevance of the concept to management science. First, we introduce an example in the field of organizational behavior, and then we present a more strategic application.

The role of leadership in explaining performance provides a first example. Hofmann and Jones (2005) explore the collective level of leadership in a store. They relate this level of leadership to store-level performance. More specifically, the focus is on transformational leadership, i.e., leadership designed to motivate subordinates to exceed expectations. A positive effect of transformational leadership on store performance is hypothesized; however, the thesis developed in this research is that this relationship is due, at least in part, to the personality traits collectively exhibited as a character of the organizational unit. These include collective conscientiousness, collective agreeableness, collective openness, and collective emotional stability. Therefore, these collective personality characteristics are the mediating explanations for the superior performance of the transformational leaders.

Innovation is at the core of a firm's business strategy. The role of R&D spending and of the innovativeness of firms in predicting their market performance or profit performance has long been investigated in the marketing and strategy literature. In recent years, the strategic orientation of firms and more particularly the extent to which they are market orientated have been examined as determinants of performance. According to Gatignon and Xuereb (1997) and Han, Kim, and Srivastava (1998), the reason for the impact of strategic orientation on performance is that market-oriented organizations are more innovative. The mediating variable in this case is the extent to which a firm brings innovations to the market.

11.1.2 Moderation Effects

Let us now consider the relationship between X and Y. In the case of mediation, the purpose is to help explain the reason for the existence of the relationship. However, the relationship itself is fundamentally stable and generalizable. If this relationship

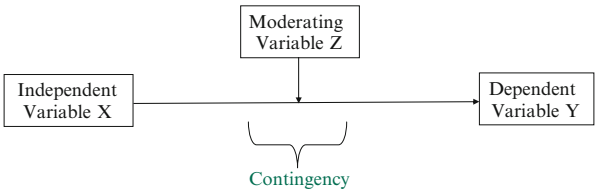


Fig. 11.2 Graphical representation of moderation effect

is contingent on particular conditions (also called contingencies), the conditions provide explanations for why the relationship between X and Y changes depending on these conditions. The conditions can be represented by variables Z, which can be either continuous or categorical factors. These conditioning variables are called moderator variables. A moderating variable is shown in Fig. 11.2 that impacts the relationship between X and Y; the relationship between X and Y is represented by the line and arrow going from X to Y. The moderating effect is represented by the arrow going from the moderator variable Z to the line between X and Y.

Much strategy research is based on contingency theories. Typically, environmental conditions impact the extent of the relationships between performance and its determinants. In more technical terminology, the impact of these determinants on performance is said to be ***conditional*** on environmental factors. Theories involving moderation effects are found throughout the literature. This can be explained in part by the complexity of the phenomena that are studied and that leave little room for unconditional effects. Such moderating effects are also often found in multi-level theories where both the organizational level and the individual level are interrelated.

For example, an interesting study by West and Broniarczyck (1998) analyzes how consumers respond when faced with critics whose collective opinions may either differ from each other or form a consensus. More specifically, the study hypothesizes that responses by consumers depend on their level of expectations of the decision outcome, i.e., their aspiration level. Consequently, the study hypothesizes that consumers will evaluate an alternative more favorably when there is more disagreement than agreement among critics, if the critics' opinions are on average below the consumer's aspiration level. However, consumers will evaluate an alternative more favorably when there is more agreement than disagreement among critics, if the critics' opinions are on average above the consumer's aspiration level. Thus, whether the critics' opinions on average are above or below the consumer's aspiration level moderates the effect of the critics' agreement or disagreement on the consumer's evaluation.

In the relationships between distribution channel partners, Kumar, Scheer, and Steenkamp (1998) analyze the asymmetries in interdependence between a dealer and a supplier. As the asymmetry in channel partners' punitive capabilities increases, (a) the firm with greater punitive capability is more likely to reciprocate punitive actions and (b) the firm with less punitive capability is less likely to reciprocate punitive actions. Thus, in this example, a firm's punitive capability is a moderator variable for the effect of the asymmetry in channel partners' punitive capabilities and reciprocation of punitive actions.

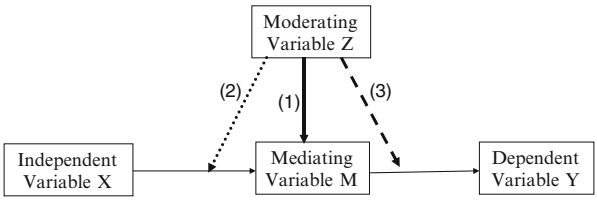


Fig. 11.3 Graphical representations of mediated moderation and moderated mediation effects

11.1.3 Mediated Moderation and Moderated Mediation Effects

Mediation and moderation are often both involved in a given theoretical explanation. In particular, as science advances, theories increase in complexity to better explain the diversity of phenomena. The concepts introduced in the previous sections can be combined to form multiple configurations. We can identify three basic configurations in which the fundamental theory states that the relationship between X and Y goes through an intermediary variable (a mediation) and that the indirect effect of X on Y is conditioned by a moderator variable. These three configurations are shown in Fig. 11.3; each corresponds to a different explanatory mechanism of the conditioning effect of variable Z.

In Fig. 11.3 the first mechanism is expressed by the arrow labeled (1) between the moderator variable Z and the mediating variable M. The reason that the indirect effect of X on Y differs depending on Z is that this variable Z changes the level of the mediating variable M. This process is called a mediated moderation. An example of this first mechanism is provided in Galunic and Anderson (2000) where the commitment of a sales agent to the firm he represents explains why relation-specific investments made by the firm and the agent, as well as generalized investments made by the firm, lead to enhanced performance: commitment of the sales agent (variable M) mediates the effect of investments in a relationship (variable X) on performance (variable Y). However, the level of commitment of the sales agent is not always the same depending on the characteristics of the relationship, such as the age of the relationship (variable Z). These variables condition the indirect effect by conditioning the explanatory (mediating) variable itself.

The other two mechanisms more clearly resemble the definition provided for a moderating effect. In both cases, the mediation is moderated by Z. Therefore, these cases are called moderated mediation. The difference between these two cases of moderated mediation stems from where the moderation occurs. In considering a mediated relationship (the indirect effect), there are really two relationships. The first relationship is the effect of X on M, and the second is the effect of M on Y. The first case of moderated mediation (expressed by the arrow labeled (2) in Fig. 11.3) is when the moderation conditions the effect of X on M. Edwards and Lambert (2007) refer to this as a first-stage moderated mediation. The second case (expressed by the arrow labeled (3) in Fig. 11.3) occurs when the moderation conditions the effect of

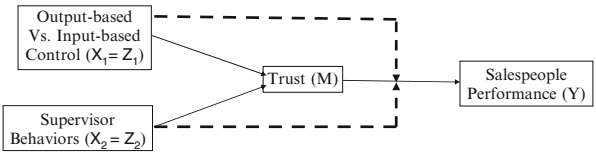


Fig. 11.4 Graphical representation of the hypotheses of Ganesan, Malter, and Rindfleisch (2005)

M on Y (second-stage moderated mediation). Each can be thought of as a simple case of moderation (as defined in Sect. 11.1.2) where the mediation variable is considered as a dependent variable (in the first case) or as an independent variable (in the second case). In other words, the relationships representing the mediation process form a recursive system (without feedback loop) where each relationship is treated separately; one is moderated and the other is not. Both can occur at the same time with the same or different moderator variables.

An example of moderated mediation corresponding to the first of these two cases is found in a study of the impact of new product development team compositions by Haon, Gotteland, and Fornerino (2009). The authors consider that the instrumental use of information (variable M) mediates the positive impact of group competence diversity (variable X) on new product performance (variable Y). However, the extent to which competence diversity (variable X) leads to instrumental information use (variable M) depends on the familiarity characterizing the group members (variable Z). Familiarity enhances the positive effect of diversity of competences on information use because it encourages group communication, in particular through enhanced interpersonal trust. Therefore, familiarity moderates the relationship between the independent variable (competence diversity) and the mediator variable (instrumental use of information).

An example of the second case of moderated mediation is provided in Ganesan, Malter, and Rindfleisch (2005), with the added particularity that the moderator variable Z is the same as the independent variable X. This case is illustrated in Fig. 11.4.

In analyzing the sales performance of salespeople (variable Y), trust (variable M) is specified as a mediating explanation for why the type of control used on salespeople—output-based vs. process-based—(variable X_1) and supervisor behaviors (variable X_2) affects performance. They also hypothesize that the independent variables (variables $Z_1 = X_1$ and $Z_2 = X_2$) moderate the extent to which better trust (variable M) translates into more or less performance (variable Y). Therefore, this is a case where the link between the mediator and the dependent variable is moderated by another factor, namely, the independent variable itself.

These relationships can be quite complex, and theory is critical since the methods may not always help to unequivocally discriminate between alternative effects. For example, even considering the relatively simple example provided in Fig. 11.4, it is probable that, except in experiments where X and Z are especially designed to be orthogonal, they are correlated with a causal relationship from X to Z or Z to X. The remaining sections of this chapter discuss the methods that can be

used to estimate these complex processes. Meanwhile, this first section provides an introduction to the way in which basic elements can be combined and points to the importance of specifying a model based on solid theoretical foundations.

11.2 Testing Mediation Effects

In this section we present the methods typically used in the social sciences to test mediation effects. We first describe the Baron and Kenny (1986) procedure, which is widely cited in the literature. Then we discuss the issues and the solutions that have been proposed to solve each particular problem associated with this procedure.

11.2.1 Baron and Kenny's Procedure

Baron and Kenny (1986) propose a method that is intended to test the explanation (or mediation role) represented by an intermediary variable (M) that intervenes in the process between the independent variable (X) and the dependent variable (Y). Assuming that there is an effect of X on Y, the question is to know whether that effect goes through M either completely or partially. Therefore, they suggest that the researcher run three regressions. All the variables are mean centered so that none of the three equations has an intercept and we can more easily focus on the other coefficients:

$$y_i = x_i c + u_i \tag{11.1}$$

$$m_i = x_i a + v_i \tag{11.2}$$

$$y_i = m_i b + x_i c^{'} + w_i \tag{11.3}$$

where i represents the unit of the observation (for example, an individual or an organization), and u_i, v_i, and w_i are the error terms of each of the three equations.

The first regression of y on x provides an estimate of the direct effect of the independent variable x on the dependent variable y through the coefficient c. The second regression of m on x indicates to what extent the proposed mediating variable m is related to the independent variable; this relationship is expressed by the coefficient a. Finally, the third regression of y on m and x estimates the coefficient b, which reflects the effect of m on y conditional on x, and the coefficient c', which represents the effect of x on y, controlling for m.

The conclusion depends on the comparison of the two coefficients c and c'. The rationale for performing this comparison comes from the relationship expressed in Eq. (11.4). Indeed, assuming that there is a single mediator (no missing data and correct model specification), then

$$c = \frac{Cov(Y,X)}{Var(X)} = \frac{c^{'}Var(X) + bCov(X,M)}{Var(X)} = c^{'} + b\left[\frac{Cov(X,M)}{Var(X)}\right] = c^{'} + ab \quad (11.4)$$

The first equality in Eq. (11.4) follows from Eq. (11.1) and the second equality follows from Eq. (11.3).

Therefore,

$$ab = c - c^{'} \quad (11.5)$$

It results in the global direct effect of x on y being fully explained through the indirect path reflected by the product term ab when c' is zero, i.e., once controlling for the effect of m, x has no more effect on y.

This also follows algebraically. Inserting the expression of m in Eq. (11.2) into Eq. (11.3) results in

$$y_i = b(ax_i + v_i) + c^{'}x_i + w_i = abx_i + c^{'}x_i + e_i = \left(ab + c^{'}\right)x_i + e_i \quad (11.6)$$

where $e_i = w_i + bv_i$.

Comparing with Eq. (11.1), which expresses the total effect of x on y just as expressed in Eq. (11.6), the coefficients of x_i in both equations must be equal. Therefore, $c = ab + c'$ and consequently $ab = c - c'$.

Yet another way to express the indirect effect of X on Y through M is by considering the sequence of the two functions. The marginal effect of X on Y through M is

$$\frac{\partial \mathrm{Y}}{\partial \mathrm{X}} = \frac{\partial \mathrm{Y}}{\partial \mathrm{M}} \frac{\partial \mathrm{M}}{\partial \mathrm{X}} = b \times a \quad (11.7)$$

Baron and Kenny indicate that (1) there must be a relationship between x and y that needs to be explained and (2) the comparison of c and c' in the estimated Eqs. (11.1) and (11.3) provides information about the extent to which the explanation is valid, i.e., the extent of the mediation. More specifically on this second point, if $c' = 0$, then the researcher concludes that m performs a full mediation, and if $c' < c$ then m performs a partial mediation.

This comparison does not provide a statistical test but, because of the relationship established in Eq. (11.5), the product ab provides the same information, since it corresponds to the same quantity $c - c'$. This product term ab can also be compared to c and it must be significantly different from zero. Therefore, Baron and Kenny suggest calculating the product ab and comparing this product term to $c - c'$ (or just to c if a full mediation is expected). They also suggest testing the significance of this indirect effect using the test proposed by Sobel (1982).

The test follows from the computation of the variance of the product of two random variables. When two variables X and Y are independent,

$$\mathrm{V[XY]} = (\mathrm{E[Y]})(\mathrm{E[Y]})\mathrm{V[X]} + (\mathrm{E[X]})(\mathrm{E[X]})\mathrm{V[Y]} + \mathrm{V[X]V[Y]} \quad (11.8)$$

Applying this equation to the parameters a and b in Eqs. (11.2) and (11.3),

$$s_{ab} = \sqrt{b^2 s_a^2 + a^2 s_b^2 + s_a^2 s_b^2} \tag{11.9}$$

where:

s_{ab} = standard deviation of the product term ab
s_a = standard deviation of the estimated parameter a
s_b = standard deviation of the estimated parameter b

Assuming a normal distribution of the product term, the Sobel test uses the standard deviation in Eq. (11.9) to compute the probability that the product term is greater than zero.

Because the last term in Eq. (11.9) is practically negligible (due to taking the product of squared terms that are small), it is often ignored and the standard deviation is approximated by Eq. (11.10):

$$s_{ab} = \sqrt{b^2 s_a^2 + a^2 s_b^2} \tag{11.10}$$

While much research in the behavioral sciences is based on these procedures and tests, some issues have been raised in the literature, and best practice has shifted over the last few years. In the next sections we discuss these issues and the best practice now recommended.

11.2.2 Best Practice

Zhao, Lynch, and Chen (2010) provide a clear presentation of the main issues that arise when using the Baron and Kenney procedure and the Sobel test and solutions. These issues have been recognized for years but there has been a recent convergence on what constitutes best practice.

11.2.2.1 Irrelevance of Direct Effect

The first issue with the Baron and Kenny procedure concerns the requirement, and even usefulness, of estimating the direct effect through Eq. (11.1) in order to compare c' with c or even ab with c'. Indeed, in most circumstances these comparisons are irrelevant. The reason is related to model misspecification. Thus far, the model we have presented has assumed that there is a single mediator variable M (simple mediation). If this assumption is correct, it is logical that the single mediator must explain a significant relationship between X and Y. The coefficient of the regression of Y on X provides an estimate of the direct effect of X on Y. If this correlation is zero, what is it that M could explain? Therefore, if

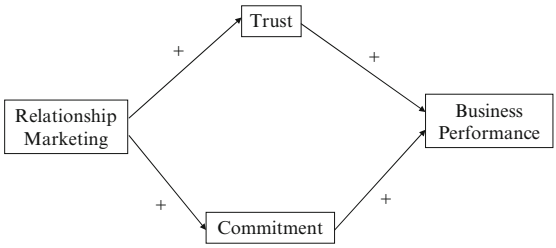


Fig. 11.5 Example of two complementary mediating variables

there were only a single plausible explanatory variable X, it is logical to assume that there must be a relationship between x and y. However, this situation rarely occurs because there are few theories where the only explanation of the effect of X on Y can be due to a single mediator variable. Once the relationship is not explained by a single mediator but by the coexistence of competing theories, the information on the direct effect from regressing Y on X can be misleading. We must distinguish, however, between two cases depending on whether the impact of the (multiple) mediators on the dependent variable Y is of the same sign or of the opposite sign (assuming that they are positively correlated to the independent variable X). Note that this assumption that the mediators are correlated positively to the independent variable does not reduce the generality of the discussion because a mediator can always be defined by its opposite and measured by its reversed scale to meet that condition.

Before analyzing more generally the case of multiple mediators, it is useful to present two examples that illustrate the distinction between these two cases (we use the same examples described by Zhao et al. (2010)). An example where both mediators impact the dependent variable positively is found in Morgan and Hunt (1994). The constructs involved and their relationships are represented in Fig. 11.5. Strong marketing relationships should relate positively to business performance through two mechanisms identified as trust and commitment (these constructs have been shown to have discriminant validity). Relationship marketing not only leads to greater trust between the parties but also to greater commitment toward each other. In turn, both trust and commitment enhance business performance. Therefore, these two mediators are complementary, and are referred to as complementary mediators.

Figure 11.5 shows the example where the direct relationship between the independent variable "relationship marketing" and the dependent variable "business performance" must be positive. As noted above, trust and commitment are two complementary explanations for why this positive relationship exists.

A different example is depicted in Fig. 11.6 where the impact of the mediators on the dependent variable goes in opposite directions. In studying the effect of advertising on consumer price sensitivity, Mitra and Lynch (1995) propose that two opposing mechanisms operate simultaneously. On the one hand, advertising increases the consideration set size, which leads to greater price sensitivity. On the other hand, it simultaneously exacerbates perceived differences in utilities among the brands, which leads to less price sensitivity. If both of these opposite effects

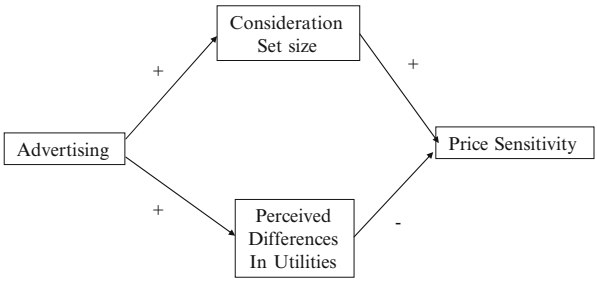


Fig. 11.6 Example of two competing mediating variables

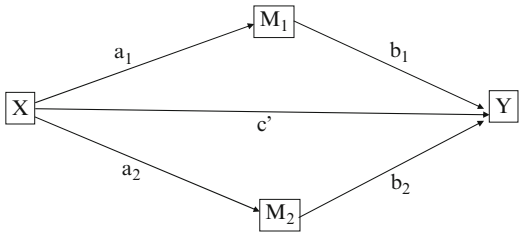


Fig. 11.7 General representation of two mediating variables

occur simultaneously, the resulting direct relationship between advertising exposure and price sensitivity depends entirely on the joint effects that are determined by the relative strength of each mediator. If the relative effects are weighted similarly, the resulting direct effect will be insignificant. Consequently, this demonstrates that the estimated coefficient c of the direct effect of X on Y is irrelevant.

This last example illustrates that when two mediating variables have competing effects, the only relevant information is provided by the estimation of the indirect effect.

It is, therefore, essential to specify the model correctly. We now examine the case of multiple mediators represented graphically in Fig. 11.7.

This model is represented algebraically by the following system of equations:

$$\begin{aligned} m_{1i} &= a_1 x_i + u_{11i} \\ m_{2i} &= a_2 x_i + u_{12i} \\ y_i &= b_1 m_{1i} + b_2 m_{2i} + c' x_i + u_{2i} \end{aligned} \tag{11.11}$$

It is important to note at this point that the third equation above must not omit any of the mediating variables, nor the independent variable for that matter. Any omission of the variables would lead to biased estimates because these three variables are structurally linked and therefore correlated. Consequently, any inference based on an analysis that would consider only, for example, M_1 through its coefficient b combined with a (the impact of X on that mediator) would be misleading. However, if the model is properly specified as in Eq. (11.11), then we can identify the indirect effects that are specific to each mediating variable (one for each) and the total indirect

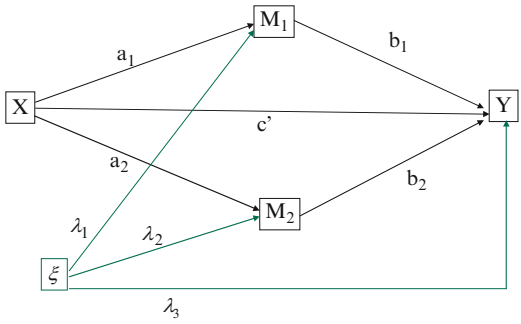


Fig. 11.8 Representation of two mediating variables with a covariate

effect corresponding to the sum of these effects. In the case depicted by Fig. 11.7, these effects are

Indirect effect specific to M_1: a_1b_1
Indirect effect specific to M_2: a_2b_2
Total indirect effect: $a_1b_1 + a_2b_2$

This notion of indirect effect is general and is very clear in this context. However, conceptually, the notion of a mediating process explaining a direct effect may be a bit confusing. These intermediary variables are mediating in the sense that they intervene in between the independent and the dependent variables. Nevertheless, the structure of such relationships does not distinguish between a model of the process mechanism that explains a direct relationship vs. a system of relationships among variables that are structurally related in that particular recursive sequence. Consequently, with more than one mediator, can these variables still be interpreted as "mediation," or are we simply considering a set of structural relationships among related constructs? Regardless of the label used for these "intermediary" variables, the notion of indirect effect remains valid and this is the critical factor for testing the theory.

In the discussion above, we have emphasized the necessity of including all the mediating variables in the model. However, as we also noted, it is critical to specify the direct link in the model in order to avoid a misspecification bias. The coefficient c' indicates whether any additional information is left in the relationship between X and Y, once the indirect effects have been considered.

The dependent variable Y may not be solely determined by the focal variables X and its mediators. Similar to the argument of model misspecification discussed above for the inclusion of X and of all the mediators in the third equation of Eq. (11.11), the relevant covariates must also be included in the model. Failing to incorporate these covariates has the same effect as that of the irrelevance of the direct effect estimate that would consider only the estimation of Eq. (11.1).

The model with a covariate is represented graphically in Fig. 11.8.

Such a model is represented by the following system of equations:

$$\begin{aligned} m_{1i} &= a_1 x_i + \lambda_1 \xi_i + u_{11i} \\ m_{2i} &= a_2 x_i + \lambda_2 \xi_i + u_{12i} \\ y_i &= b_1 m_{1i} + b_2 m_{2i} + c' x_i + \lambda_3 \xi_i + u_{2i} \end{aligned} \tag{11.12}$$

where ξ represents the covariate.

Although there is a single covariate expressed in Fig. 11.8 and in Eq. (11.12), it can easily be generalized to multiple covariates with the addition of the corresponding λ parameters.

11.2.2.2 Focus on Product of Indirect Effect Coefficients

The consequence of the prior discussion is that all the relevant information is contained in the product of the two coefficients a and b or the set of such coefficients when multiple mediators are involved. Therefore, the only requirement is that these coefficients be jointly significant and, more specifically, that each of the indirect effects reflected by the product terms $a_k b_k$ for each of the k indirect effects be significant. This is the rationale for the Sobel test. However, this test has been criticized for lacking power. We examine this test and alternatives in the next section. However, before proceeding with the ways in which the product term of the indirect effect can be tested for significance, it is useful to highlight the difference between the test that the product ab is different from zero and the joint test that the coefficients a and b are different from zero.

If coefficient a or b is equal to zero, then it follows that the product ab is zero. If we call the event A = {coefficient a is equal to zero} and event B = {coefficient b is equal to zero}, the null hypothesis in the test that the product ab is equal to zero corresponds to event A ∪ B. The null hypothesis of a joint test is a test of the event A ∩ B. Therefore, the probability corresponding to the joint test that a and b are each different from zero is much larger than the probability corresponding to the test of the product ab. We will, therefore, focus our attention on the test of the product ab rather than on joint tests.

11.2.2.3 Testing Indirect Effects with Bootstrap Confidence Intervals

The literature has clearly pointed out the deficiency of the Sobel test. It is low in power because (1) the product ab is not normally distributed and (2) a and b may not be independently distributed (when errors in each equation are correlated).

1. The product ab is not normally distributed

In Eq. (11.8) we have provided the formula for the calculation of the variance of the product term of two normally distributed random variables. However, the

variance is insufficient to perform a test of significance. The probability distribution of the product term needs to be defined. The Sobel test is based on the approximation of the distribution by a normal distribution. However, the non-normality of the distribution of the product of two normally distributed random variables is clearly established. More specifically, products of normal variables with positive means tend to have a positive skew and those with negative means tend to have a negative skew (Shrout & Bolger, 2002). The issue concerns the extent to which this approximation biases the results. As discussed below, an empirically based solution is recommended.

2. a and b may not be independently distributed

We should estimate the two key equations representing the mediation process (i.e., Eqs. (11.2) and (11.3)) simultaneously to recognize that the error terms of these two equations, rewritten in Eq. (11.13), can be correlated. Such correlated errors can be due to missing factors that affect the dependent variable and the mediating variable in a related manner:

$$\begin{aligned} m_i &= ax_i + u_{1i} \\ y_i &= bm_i + c^{'}x_i + u_{2i} \end{aligned} \tag{11.13}$$

We now rewrite these equations more generally to demonstrate the impact of the correlation between the error terms on the independence of the two coefficients composing the indirect effect.

We rewrite Eq. (11.13) using β to represent all the coefficients. This leads to Eq. (11.14) for each observation i:

$$\begin{cases} m_i = x_i\beta_{11} + u_{1i} \\ y_i = x_i\beta_{21} + m_i\beta_{22} + u_{2i} \end{cases} \tag{11.14}$$

For all observations and letting $\underset{N\times 2}{\mathbf{X}_2} = \begin{bmatrix} \underset{N\times 1}{\mathbf{x}} & \underset{N\times 1}{\mathbf{m}} \end{bmatrix}$

$$\begin{cases} \underset{N\times 1}{\mathbf{m}} = \underset{N\times 1}{\mathbf{x}}\,\beta_{11} + \underset{N\times 1}{\mathbf{u}_1} \\ \underset{N\times 1}{\mathbf{y}} = \underset{N\times 1}{\mathbf{x}}\,\beta_{21} + \underset{N\times 1}{\mathbf{m}}\,\beta_{22} + \underset{N\times 1}{\mathbf{u}_2} = \begin{bmatrix} \underset{N\times 1}{\mathbf{x}} & \underset{N\times 1}{\mathbf{m}} \end{bmatrix} \underset{2\times 1}{\begin{bmatrix} \beta_{21} \\ \beta_{22} \end{bmatrix}} + \underset{N\times 1}{\mathbf{u}_2} = \underset{N\times 2}{\mathbf{X}_2}\,\underset{2\times 1}{\beta_2} + \underset{N\times 1}{\mathbf{u}_2} \end{cases} \tag{11.15}$$

Let $\mathbf{x}_1 = \mathbf{x}$. Bringing the two equations in the system into a single equation

$$\underset{2N\times 1}{\begin{bmatrix} \mathbf{m} \\ \mathbf{y} \end{bmatrix}} = \underset{2N\times 3}{\begin{bmatrix} \underset{N\times 1}{\mathbf{x}_1} & 0 \\ 0 & \underset{N\times 2}{\mathbf{X}_2} \end{bmatrix}} \underset{3\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{2N\times 1}{\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}} \tag{11.16}$$

Let $\underset{2N\times 1}{\mathbf{w}} = \begin{bmatrix} \mathbf{m} \\ \mathbf{y} \end{bmatrix}$; $\underset{2N\times 3}{\mathbf{Z}} = \begin{bmatrix} \underset{N\times 1}{\mathbf{x}_1} & 0 \\ 0 & \underset{N\times 2}{\mathbf{X}_2} \end{bmatrix}$.

Then, the generalized least square estimator is the efficient estimator of the parameters:

$$\hat{\beta}_{GLS} = \left(\mathbf{Z}'\left(\underset{2\times 2}{\Sigma^{-1}} \otimes \mathbf{I}_N\right)\mathbf{Z}\right)^{-1}\mathbf{Z}'\left(\Sigma_{2\times 2}^{-1} \otimes \mathbf{I}_N\right)\mathbf{w} \tag{11.17}$$

And the variance of the estimator is

$$\underset{3\times 3}{V\left[\hat{\beta}_{GLS}\right]} = \left(\underset{3\times 2N}{\mathbf{Z}'}\left(\Sigma_{2\times 2}^{-1} \otimes \mathbf{I}_N\right)\underset{2N\times 3}{\mathbf{Z}}\right)^{-1} \tag{11.18}$$

We first consider the case where the correlation for each observation of the two equation errors (called contemporaneous correlation) is 0 ($\Sigma = diag$):

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}; \quad \Sigma^{-1} = \begin{bmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{bmatrix}$$

Let

$$\underset{2N\times 3}{\mathbf{Z}} = \begin{bmatrix} \underset{N\times 1}{\mathbf{x}_1} & 0 \\ 0 & \underset{N\times 2}{\mathbf{X}_2} \end{bmatrix}; \quad \underset{3\times 2N}{\mathbf{Z}'} = \begin{bmatrix} \underset{1\times N}{\mathbf{x}'_1} & 0 \\ 0 & \underset{2\times N}{\mathbf{X}'_2} \end{bmatrix}$$

Then,

$$\Sigma^{-1} \otimes \mathbf{I}_N = \begin{bmatrix} \sigma_1^{-2} & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_1^{-2} & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_1^{-2} & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \sigma_1^{-2} & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \sigma_2^{-2} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \sigma_2^{-2} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & \sigma_2^{-2} \end{bmatrix} \tag{11.19}$$

And

$$\underset{3\times3}{\mathrm{V}\left[\hat{\beta}_{GLS}\right]} = \left(\underset{3\times2\mathrm{N}}{\mathbf{Z}'}\left(\Sigma_{2\times2}^{-1}\otimes\mathbf{I}_{\mathrm{N}}\right)\underset{2\mathrm{N}\times3}{\mathbf{Z}}\right)^{-1}$$

$$= \left[\begin{pmatrix} x_1\sigma_1^{-2} & x_2\sigma_1^{-2} & \cdots & x_{\mathrm{N}}\sigma_1^{-2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_1\sigma_2^{-2} & x_2\sigma_2^{-2} & \cdots & x_{\mathrm{N}}\sigma_2^{-2} \\ 0 & 0 & 0 & 0 & m_1\sigma_2^{-2} & m_2\sigma_2^{-2} & \cdots & m_{\mathrm{N}}\sigma_2^{-2} \end{pmatrix}_{3\times 2N} \underset{2\mathrm{N}\times3}{\begin{pmatrix} \mathbf{x} & 0 & 0 \\ 0 & \mathbf{x} & \mathbf{m} \end{pmatrix}}\right]^{-1}$$

Consequently,

$$\underset{3\times3}{\mathrm{V}\left[\hat{\beta}_{GLS}\right]} = \begin{bmatrix} \sum_{i=1}^{\mathrm{N}} x_i^2\sigma_1^{-2} & 0 & 0 \\ 0 & \sum_{i=1}^{\mathrm{N}} x_i^2\sigma_2^{-2} & \sum_{i=1}^{\mathrm{N}} x_i m_i\sigma_2^{-2} \\ 0 & \sum_{i=1}^{\mathrm{N}} x_i m_i\sigma_2^{-2} & \sum_{i=1}^{\mathrm{N}} m_i^2\sigma_2^{-2} \end{bmatrix}^{-1} \tag{11.20}$$

The variances of the estimated coefficients a and b are the first two diagonal elements in Eq. (11.20). The off-diagonal elements of that submatrix represent the covariances that are zero. Therefore, the estimated coefficients a and b are uncorrelated.

This is different, however, in the case where the contemporaneous correlation is not 0, i.e., $\Sigma \neq diag$. In that case, if we note the inverse of the contemporaneous covariance matrix as in Eq. (11.21):

$$\Sigma^{-1} = \begin{bmatrix} \sigma_1^{-2} & \sigma_{12}^{-1} \\ \sigma_{12}^{-1} & \sigma_2^{-2} \end{bmatrix} \tag{11.21}$$

Then, it follows that

$$\Sigma^{-1}\otimes\mathbf{I}_{\mathrm{N}} = \begin{bmatrix} \sigma_1^{-2} & 0 & \cdots & 0 & \cdots & 0 & \sigma_{12}^{-1} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_1^{-2} & \cdots & 0 & \cdots & 0 & 0 & \sigma_{12}^{-1} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_1^{-2} & \cdots & 0 & 0 & 0 & \cdots & \sigma_{12}^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \sigma_1^{-2} & 0 & 0 & \cdots & 0 & \cdots & \sigma_{12}^{-1} \\ \sigma_{12}^{-1} & 0 & \cdots & 0 & \cdots & 0 & \sigma_2^{-2} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_{12}^{-1} & \cdots & 0 & \cdots & 0 & 0 & \sigma_2^{-2} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_{12}^{-1} & \cdots & 0 & 0 & 0 & \cdots & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \sigma_{12}^{-1} & 0 & 0 & \cdots & 0 & \cdots & \sigma_2^{-2} \end{bmatrix} \tag{11.22}$$

Consequently, the covariance matrix of the generalized least square estimator is

$$\underset{3\times 3}{\mathrm{V}\left[\hat{\beta}_{GLS}\right]} = \left(\underset{3\times 2\mathrm{N}}{\mathbf{Z}'}\left(\Sigma^{-1}_{2\times 2}\otimes \mathbf{I}_{\mathrm{N}}\right)\underset{2\mathrm{N}\times 3}{\mathbf{Z}}\right)^{-1}$$

$$= \left[\begin{pmatrix} x_1\sigma_1^{-2} & x_2\sigma_1^{-2} & \cdots & x_{\mathrm{N}}\sigma_1^{-2} & x_1\sigma_{12}^{-1} & x_1\sigma_{12}^{-1} & \cdots & x_{\mathrm{N}}\sigma_{12}^{-1} \\ x_1\sigma_{12}^{-1} & x_2\sigma_{12}^{-1} & \cdots & x_{\mathrm{N}}\sigma_{12}^{-1} & x_1\sigma_2^{-2} & x_2\sigma_2^{-2} & \cdots & x_{\mathrm{N}}\sigma_2^{-2} \\ m_1\sigma_{12}^{-1} & m_2\sigma_{12}^{-1} & \cdots & m_{\mathrm{N}}\sigma_{12}^{-1} & m_1\sigma_2^{-2} & m_2\sigma_2^{-2} & \cdots & m_{\mathrm{N}}\sigma_2^{-2} \end{pmatrix}_{3\times 2\mathrm{N}} \begin{pmatrix} \mathbf{x} & 0 & 0 \\ 0 & \mathbf{x} & \mathbf{m} \end{pmatrix}_{2\mathrm{N}\times 3}\right]^{-1}$$

which leads to

$$\underset{3\times 3}{\mathrm{V}\left[\hat{\beta}_{GLS}\right]} = \begin{bmatrix} \sum_{i=1}^{\mathrm{N}} x_i^2\sigma_1^{-2} & \sum_{i=1}^{\mathrm{N}} x_i^2\sigma_{12}^{-1} & \sum_{i=1}^{\mathrm{N}} x_i m_i\sigma_{12}^{-1} \\ \sum_{i=1}^{\mathrm{N}} x_i^2\sigma_{12}^{-1} & \sum_{i=1}^{\mathrm{N}} x_i^2\sigma_2^{-2} & \sum_{i=1}^{\mathrm{N}} x_i m_i\sigma_2^{-2} \\ \sum_{i=1}^{\mathrm{N}} x_i m_i\sigma_{12}^{-1} & \sum_{i=1}^{\mathrm{N}} x_i m_i\sigma_2^{-2} & \sum_{i=1}^{\mathrm{N}} m_i^2\sigma_2^{-2} \end{bmatrix}^{-1} \tag{11.23}$$

In Eq. (11.23), the off-diagonal elements of the submatrix corresponding to the variances and covariances of a and b are no longer zero. Therefore, the estimated coefficients a and b are correlated. The consequence of this correlation is that Eq. (11.8) no longer applies and the Sobel test based on this assumption of independence is inappropriate.

3. The bootstrapping approach and Preacher and Hayes' algorithm (2004, 2008)

The use of a bootstrapping method for estimating the distribution of the indirect effect (the product ab) was proposed by Bollen and Stine (1990) in order to compute the confidence interval of the non-symmetric distribution. MacKinnon, Lockwood, and Williams (2004) demonstrate the superiority of empirically estimating the distribution of the product term representing the indirect effect using such a bootstrapping approach. Multiple samples are drawn randomly with replacement from the original data, each with the same number of observations as in the original data set. Each sample provides an estimate of the effect for which the distribution is built over the repetitions. This empirically based distribution does not make any assumption about the form (normality) or about the independence of the elements of the product term. Preacher and Hayes (2004, 2008) made available a subroutine in SAS as part of several statistical software packages that popularized the use of the method.

Although Preacher and Hayes have proposed different SAS subroutines adapted to different models[1] that vary in complexity, the "*%INDIRECT.SAS*" subroutine is

[1] A more complete subroutine (PROCESS) for SPSS and SAS is available for download from Andrew F. Hayes at http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html.

```
capture drop program bootcm2med
program bootcm2med, rclass
sureg (med1 iv) (med2 iv) (dv med1 med2 iv)
return scalar indirect1 = [med1]_b[iv]*[dv]_b[med1]
return scalar direct1 = [dv]_b[iv]
return scalar indirect2 = [med2]_b[iv]*[dv]_b[med2]
return scalar direct2 = [dv]_b[iv]
end
```

Fig. 11.9 STATA subroutine with two mediators and no covariates (bootcm2med.do)

flexible enough to include multiple mediators and covariates (the "%" is used in SAS to indicate that it is a subroutine). By running the subroutine first, the researcher can then call this subroutine from within the SAS file. An example of how to use the method is described in the next section.

A similar subroutine can easily be written in STATA. Such a subroutine (a do-file in STATA to be run once in a STATA session) is shown in Fig. 11.9.

In this program, the subroutine consists in executing an estimation with seemingly unrelated regression (SUR) of three equations: (1) the first mediator variable (med1) on the independent variable, (2) the second mediator variable (med2) on the independent/exogenous variable (iv), and (3) the ultimate dependent variable in the system of equations on the two mediator variables and the independent/exogenous variable (iv). This program can easily be adapted to reflect the specifications of the equations, using the derivations to compute the indirect effects explained in Sect. 11.2.1, and in particular Eq. (11.7). The coefficients from the SUR estimation that come in the product term for the indirect effects are noted by **[var1]_b[var2]** where var1 is the name of the dependent variable of the relevant equation and var2 is the name of the variables of the coefficient that is of interest. For example, **indirect 1** is the product of **[med1]_b[iv]**, which is the coefficient of the iv variable in the med1 equation, and **[dv]_b[med1]**, which is the coefficient of the med1 variable in the equation where dv is the dependent variable. This do-file should be executed once so that it can be called from another do-file with the relevant analysis, as demonstrated below.

11.2.2.4 Example of Mediation in SAS and STATA

Multiple Mediators: No Covariates

Figure 11.10 shows an SAS file used to analyze a data set of 400 observations with x_1 as the independent variable, x_4 as the dependent variable, and x_2 and x_3 as two mediating variables. Standard OLS regressions are performed. Three equations are highlighted in grey, one for each mediator as a function of the independent variable, and a third one that models the dependent variable as a function of the two mediators, controlling for the possible direct effect of the independent variable remaining after the indirect effects are considered. After running these ordinary

```
/*Simdata.SAS*/
filename simdata 'C:\DATA\WORK_SAS\SAS_Mediation\simdata5.csv';
data simdata2;
infile  simdata firstobs=2 dlm=',';
input x1-x4;
iv=x1;
med1=x2;
med2=x3;
dv=x4;
Proc reg data=simdata2 ;
model dv = iv;
model med1 = iv;
model med2 = iv;
model dv = med1 Med2 iv;
run;

data simdata3;
set simdata2;
%indirect(data=simdata3, y=dv, x=iv, m=med1 med2, boot=5000);
run;
```

Fig. 11.10 SAS example with two mediators and no covariates (Examp11-1.sas)

```
insheet x1-x4 using "/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/SimData5_CorErr.csv", clear
generate iv = x1
generate med1 = x2
generate med2 = x3
generate dv = x4
sureg (med1 iv) (med2 iv) (dv med1 med2 iv)
bootstrap r(indirect1) r(indirect2), reps(5000) nodots: bootcm2med
estat boot, bc percentile
```

Fig. 11.11 STATA example with two mediators and no covariates (Examp11-1_Mac.do)

regressions, the "indirect" subroutine is called (also in grey in Fig. 11.10) with the following parameters in parentheses:

data = simdata3 to indicate the name of the SAS data set containing the data to be analyzed
y = dv to indicate which variable is the dependent variable (here it is dv)
x = iv to indicate which variable is the independent variable (here it is iv)
m = med1 med2 where the list on the right side of the equal sign indicates all the mediating variables (here the two mediators are med1 and med2)
boot = 5,000 indicates the number of repeated samples used for the bootstrapping method (5,000 is usually advised, e.g., Hayes, 2009)

The equivalent do-file in STATA is shown in Fig. 11.11.

The only difference between the two programs shown in Figs. 11.10 and 11.11 comes from the fact that the estimators in STATA are generalized least square estimators while those in the SAS program are ordinary least squares. However, both subroutines perform the bootstrapping for the estimation of the confidence intervals.

The output of running the SAS program appears in Fig. 11.12. The first regression estimating the direct effect with a simple regression of the dependent variable on the independent variable shows an insignificant effect (parameter = 0.00974 with a *t* value of 0.27). However, the next two regressions show that each mediating

variable is significantly related to the independent variable (0.37503 with $t = 7.56$ for med1 and 0.238 with $t = 6.48$ for med2). Also the mediating variables have a significant effect on the dependent variable with a coefficient of 0.47992 ($t = 22.22$) and -0.53416 ($t = 18.31$) for med1 and med2, respectively. The independent variable is still insignificant in that regression. This leads us to believe that the two mediators together mediate fully, but competitively, the impact of the independent variable on the dependent variable. These effects could not be estimated directly because of the counterbalancing effects of each mediator. But before concluding that this process correctly reflects the complex (invisible)

```
The REG Procedure
                              Model: MODEL1
                          Dependent Variable: dv

                    Number of Observations Read         400
                    Number of Observations Used         400

                              Analysis of Variance
                                     Sum of           Mean
 Source                    DF        Squares         Square    F Value    Pr > F

 Model                      1        0.03588        0.03588       0.07    0.7910
 Error                    398      203.09815        0.51030
 Corrected Total          399      203.13403


              Root MSE              0.71435    R-Square     0.0002
              Dependent Mean        0.98254    Adj R-Sq    -0.0023
              Coeff Var            72.70427

                              Parameter Estimates
                           Parameter       Standard
      Variable     DF       Estimate          Error    t Value    Pr > |t|

      Intercept     1        0.98277        0.03573      27.51      <.0001
      iv            1        0.00974        0.03672       0.27      0.7910

                              The REG Procedure
                                Model: MODEL2
                          Dependent Variable: med1

                    Number of Observations Read         400
                    Number of Observations Used         400

                              Analysis of Variance
                                     Sum of           Mean
 Source                    DF        Squares         Square    F Value    Pr > F

 Model                      1       53.23871       53.23871      57.11    <.0001
 Error                    398      371.03482        0.93225
 Corrected Total          399      424.27353


              Root MSE              0.96553    R-Square     0.1255
              Dependent Mean        1.99006    Adj R-Sq     0.1233
              Coeff Var            48.51755

                              Parameter Estimates
                           Parameter       Standard
      Variable     DF       Estimate          Error    t Value    Pr > |t|

      Intercept     1        1.99875        0.04829      41.39      <.0001
      iv            1        0.37503        0.04963       7.56      <.0001
```

Fig. 11.12 SAS output of bootstrapping estimates (Examp11-1.lst)

```
                          The REG Procedure


                              Model: MODEL3
                         Dependent Variable: med2

                    Number of Observations Read         400
                    Number of Observations Used         400

                              Analysis of Variance
                                     Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F

Model                       1       21.44091       21.44091      41.94    <.0001
Error                     398      203.47400        0.51124
Corrected Total           399      224.91491


              Root MSE              0.71501    R-Square     0.0953
              Dependent Mean        2.92571    Adj R-Sq     0.0931
              Coeff Var            24.43890

                           Parameter Estimates
                         Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|

     Intercept     1        2.93122        0.03576      81.97      <.0001
     iv            1        0.23800        0.03675       6.48      <.0001

                          The REG Procedure
                            Model: MODEL4
                       Dependent Variable: dv

                    Number of Observations Read         400
                    Number of Observations Used         400


                              Analysis of Variance
                                     Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F

Model                       3      134.85133       44.95044     260.69    <.0001
Error                     396       68.28270        0.17243
Corrected Total           399      203.13403


              Root MSE              0.41525    R-Square     0.6639
              Dependent Mean        0.98254    Adj R-Sq     0.6613
              Coeff Var            42.26260

                           Parameter Estimates
                         Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|

     Intercept     1        1.58928        0.09565      16.62      <.0001
     med1          1        0.47992        0.02160      22.22      <.0001
     med2          1       -0.53416        0.02917     -18.31      <.0001
     iv            1       -0.04312        0.02371      -1.82      0.0698

          Dependent, Independent, and Proposed Mediator Variables

                                    VS

                               DV =    DV
                               IV =    IV
                               MEDS =  MED1
                               ******* MED2


                               Sample size

                                    N
                                  400
```

Fig. 11.12 (continued)

```
                IV to Mediators (a paths)

                          BZXMAT
                Coeff        se         t          p

MED1           0.3750     0.0496     7.5570     0.0000
MED2           0.2380     0.0368     6.4760     0.0000


Direct Effects of Mediators on DV (b paths)

                          BYZX2MAT
                Coeff        se         t          p

MED1           0.4799     0.0216    22.2196     0.0000
MED2          -0.5342     0.0292   -18.3143     0.0000

     Total effect of IV on DV (c path)

                          BYXMAT
                Coeff        se         t          p

IV             0.0097     0.0367     0.2652     0.7910

   Direct Effect of IV on DV (c' path)

                          CPRIMMAT
                Coeff        se         t          p

IV            -0.0431     0.0237    -1.8182     0.0698

          Fit Statistics for DV Model

                           DVMS
  R-sq  adj R-sq            F        df1        df2          p

0.6639     0.6613   260.6865     3.0000   396.0000     0.0000

*****************************************************

       BOOTSTRAP RESULTS FOR INDIRECT EFFECTS

Indirect Effects of IV on DV through Mediators (ab paths)
                           RES
                Data       Boot       Bias         SE

TOTAL          0.0529     0.0534     0.0006     0.0304
MED1           0.1800     0.1806     0.0006     0.0256
MED2          -0.1271    -0.1272    -0.0000     0.0198


 Bias Corrected and Accelerated Confidence Intervals
                           CI
                         Lower      Upper

                TOTAL    -0.0049     0.1141
                MED1      0.1323     0.2330
                MED2     -0.1689    -0.0909


*****************************************************

     Level of Confidence for Confidence Intervals
                           CONF
                             95

          Number of Bootstrap Resamples
                           BTN
                           5000
```

Fig. 11.12 (continued)

relationship between the iv and the dv, we want to test that the indirect effects are significant. The bootstrapping tests are performed in the last part of the output. The subroutine performs the OLS estimations on the original sample and the same results as previously described are reported. These results are highlighted in grey. First, the indirect effect through med1 is shown to be 0.18 (the product of 0.37503×0.47992), and then the indirect effect through med2 is -0.1271 (the product of 0.238×-0.53416). Finally, the confidence intervals for both indirect effects at the 0.05 level exclude the value of zero, from which we infer that the indirect mediating effects are significant.

The STATA output follows in Fig. 11.13 with almost identical results.

The confidence intervals that are bias corrected take into account the skewness of the distribution of a product term.

Multiple Mediators: With Covariates

The syntax for introducing covariates in the subroutine "indirect" is

%indirect (data=filename, y=*Dv*, x=*Iv*, m=*MlistCovlist*, c=*Cov*, boot= *Z*)

where:

Dv is the unique dependent variable name.
Iv is the unique independent variable name.
Mlist is the list of the mediating variables, separated by a space.
Covlist is the list of the covariates separated by a space.
Cov is the number of covariates (this number is used to identify the last items in the names that follow the "m=" list to distinguish them from the mediating variables).
Z is the resampling number; Hayes (2009) recommends at least 5,000 sampling iterations.

Some additional optional parameters can be added but are not necessary. Except for the additional parameters related to the covariates, the interpretation of the output remains identical to the description of the example in Fig. 11.12.

In STATA, the solution to the problem of covariates is easily solved. It simply requires adding these covariates into the specification of the "sureg" equations. These covariates will impact the estimates of the components of the indirect effects but not their calculation in the subroutine.

11.2.3 Sequential Multiple Mediation Effects

The cases presented in Sect. 11.2.2.4 may include multiple mediators but there are no links among these mediators. Sometimes, however, the explanation for a phenomenon (or a relationship from x to y) may involve a sequence of factors influencing each other. For example, when consumers are exposed repeatedly to given advertising, they process more information about that ad, which in turn

```
Seemingly unrelated regression
------------------------------------------------------------------------------
Equation           Obs  Parms        RMSE    "R-sq"       chi2        P
------------------------------------------------------------------------------
med1               400      1    .9631132    0.1255      57.39   0.0000
med2               400      1    .7132216    0.0953      42.15   0.0000
dv                 400      3    .4131667    0.6639     789.96   0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
med1         |
          iv |   .3750253   .0495022     7.58   0.000     .2780028    .4720477
       _cons |   1.998754   .0481693    41.49   0.000     1.904344    2.093164
-------------+----------------------------------------------------------------
med2         |
          iv |   .2379953   .0366582     6.49   0.000     .1661465    .3098441
       _cons |   2.931225   .0356712    82.17   0.000      2.86131    3.001139
-------------+----------------------------------------------------------------
dv           |
        med1 |   .4799168   .0214905    22.33   0.000     .4377961    .5220375
        med2 |  -.5341623   .0290202   -18.41   0.000    -.5910409   -.4772838
          iv |  -.043117    .0235955    -1.83   0.068    -.0893635    .0031294
       _cons |   1.589283   .0951675    16.70   0.000     1.402758    1.775807
------------------------------------------------------------------------------

. bootstrap r(indirect1) r(indirect2), reps(5000) nodots: bootcm2med

Bootstrap results                                 Number of obs      =       400
                                                  Replications       =      5000

      command:  bootcm2med
        _bs_1:  r(indirect1)
        _bs_2:  r(indirect2)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   .1799809   .0257484     6.99   0.000     .1295151    .2304468
       _bs_2 |  -.1271281   .0195481    -6.50   0.000    -.1654417   -.0888146
------------------------------------------------------------------------------

. estat boot, bc percentile

Bootstrap results                                 Number of obs      =       400
                                                  Replications       =      5000

      command:  bootcm2med
        _bs_1:  r(indirect1)
        _bs_2:  r(indirect2)

------------------------------------------------------------------------------
             |   Observed               Bootstrap
             |      Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   .17998092   .0001965   .02574836    .1307582   .2315905   (P)
             |                                       .131224   .2318227   (BC)
       _bs_2 |  -.12712812   .0000494   .01954808   -.1665718  -.0890378   (P)
             |                                      -.1687979    -.09114   (BC)
------------------------------------------------------------------------------
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
```

Fig. 11.13 STATA output of bootstrapping estimates (Examp11-1_Mac.log)

affects their attitude toward the ad and subsequently toward the brand itself. This is reflected in the sequence of mediations represented graphically in Fig. 11.14 where X represents advertising exposures, M_1 the amount of information processing, M_2 the attitude toward the ad, and Y the attitude toward the brand advertised.

The indirect effect going through the path $X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ is given by Eq. (11.24):

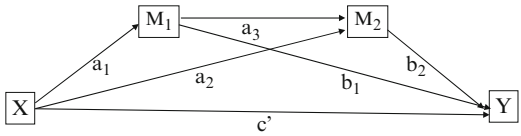


Fig. 11.14 Representation of a sequence of mediations

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial m_2}\frac{\partial m_2}{\partial m_1}\frac{\partial m_1}{\partial x} = b_2 a_3 a_1 \tag{11.24}$$

where b_2, a_3, and a_1 correspond to the coefficients of the respective linear functions.

The total effect of x on y corresponds to the sum of all the linear combinations that influence the dependent variable y. Therefore, the principles are exactly the same as for the cases discussed in the prior sections of this chapter.

It is then straightforward to modify the commands in STATA to obtain the parameter estimates from the "sureg" model specification and the appropriate bootstrapping estimation of the confidence intervals. In the example of Fig. 11.14, the indirect effect through M_1 and M_2 is given by the product $b_2 a_3 a_1$.

Another approach (Chandukala, Dotson, Brazell, and Allenby, 2011) has been proposed that allows for the introduction of more complex relationships. The joint distribution of all the variables involved in the sequential system is expressed as a set of marginal and conditional distribution. The joint distribution is decomposed according to the hypothesized process as the product of conditional distributions that are independent from each other. The total effect of a variable on the last dependent variable in the sequence can then be computed using the chain rule to compute derivatives. Thus, in the example illustrated in Fig. 11.14

$$f(x, m_1, m_2, y) = f\left(y\middle|m_2\right) f\left(m_2\middle|m_1\right) f\left(m_1\middle|x\right) f(x) \tag{11.25}$$

The functions in Eq. (11.25) do not necessarily correspond to simple regression equations, as we have discussed. In this more general modeling approach, we can assume a finite mixture of likelihoods to represent the heterogeneity of responses (the heterogeneity can be expressed only in selected parts of the chain of variables). The major difference with this approach, beyond its greater flexibility to incorporate more complex response functions, is in its estimation that uses a Bayesian approach. As noted by Chandukala et al. (2011), "computing the joint marginal density requires the specification of the marginal density of model factors that are not conditionally related to other variables" (p. 126). For example, in the model in Eq. (11.25), the specification of the marginal density of x is required. Assuming the normal distribution of the marginal distributions, the Bayesian approach provides estimates of the model parameters. This approach is mentioned here because it is specific to the modeling of the sequential mediation structure. However, its approach based on the joint distribution of all the variables in the sequence builds on the analysis of covariance structure approach to systems of equations, even if the variables are directly observed rather than relying on latent factor structures underlying the observed data.

11.2.4 Testing Mediation When Constituent Paths Are Nonlinear

Two cases can be found where the relationships, either between the independent variable and the mediator or between the mediator and the dependent variable, are not linear. In the first case, a particular functional form is specified for which the parameters can be estimated. The estimation can be ordinary least squares if the variables can be transformed to make the model linear in the parameters, or if this is not the case, maximum likelihood estimation can be used. The second case occurs when the dependent variable is not measured on an interval or a ratio scale. In that case, a logit type of model can be used.

In each of these cases, the conceptual analysis of mediating effects remains the same as what we have discussed thus far. The paths, however, may not correspond to a single parameter, so it is important to first address the question of the identification of the paths in terms of the nonlinear parameters. This leads to a complication when calculating and testing the indirect effect, which we discuss next. Finally, we present a subroutine proposed by Preacher and Hayes (2004, 2008) to find confidence intervals of the indirect effects based on the bootstrapping method.

11.2.4.1 Nonlinear Functional Form

Let us consider an example where the independent variable is the number of exposures to an ad for a brand. The dependent variable is the attitude toward the brand. Examining this relationship can be complex because a wear-out effect has been observed that leads to a decrease in advertising effectiveness. This wear-out could be explained by the extent of information processing, which then is a mediating factor. Therefore, as a first step in the mediation process, ad exposure leads to more processing as pieces of information in the ad get through to the audience. However, at a certain level of repetition, distraction starts to occur, so not only do increases in the processing level occur at a decreasing rate, but, overall, the audience starts to decrease its level of processing about both the ad and the brand. This would suggest a relationship between the number of exposures and the level of processing that could be represented by a second-order polynomial function. Considering now the second step in the mediation process, attitude toward the brand may improve with more processing about the brand (assuming no counterarguing and that positive information is conveyed in the ad). There must be, however, a saturation effect in such a relationship so that attitude does not improve to infinity as more processing occurs. Such a saturation effect could be reflected by a functional form such as a power function or a logarithmic transformation of the level of processing variable. These equations would, therefore, be

$$\begin{aligned} m_i &= a_0 + a_1 x_i + a_2 {x_i}^2 + v_i \\ y_i &= b_0 + b_1 Ln(m_i) + x_i c' + w_i \end{aligned} \tag{11.26}$$

This example illustrates the need to consider the theoretical functional form of each of the links involved in the mediating process. Apart from leading us to specify two equations that may not be linear, the conceptual approach to investigating mediation remains identical. We can estimate the parameters of the two equations, in this case using ordinary least squares, as the models are linear in the parameters. We can then easily test the significance of the various paths. However, the joint test or the test of the indirect effect due to the particular path that goes through the mediating variable is not straightforward. The issue stems from the fact that the indirect effect cannot be easily represented by the simple product of two of the estimated coefficients a and b. For such nonlinear relationships, the indirect effect is best described by referring back to Eq. (11.7). The indirect effect is still described by the derivatives

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial m} \frac{\partial m}{\partial x} \tag{11.27}$$

However, the result is a bit more complex. Considering the example expressed in Eq. (11.26), the two partial derivatives are

$$\frac{\partial y}{\partial m} = \frac{b_1}{m} \tag{11.28}$$

and

$$\frac{\partial m}{\partial x} = a_1 + 2a_2 x \tag{11.29}$$

The marginal effect is no longer a constant effect but varies depending on the level of the independent variable x. This function of x is clearly visible in Eq. (11.29). Moreover, in Eq. (11.28), the presence of m indicates that this derivative is also dependent on the level of the independent variable x, as m can be expressed in terms of x using the first equation in Eq. (11.26).

These indirect effects have been labeled instantaneous indirect effects (Hayes & Preacher, 2010), even though no time dimension is involved. The value theta represented by the partial derivatives is best thought of as the marginal indirect effect when moving slightly away from a reference value x_0:

$$\theta_{x0} = \frac{\partial y}{\partial m} \frac{\partial m}{\partial x} \tag{11.30}$$

Now we are confronted with the same issues as discussed earlier regarding the test of significance of the indirect linear effect, except that the test is complicated by the fact that the indirect effect could be significant at a particular level of x and not

```
/*SimdataMediationReg7.SAS*/
libname db 'C:\DATA\WORK_SAS\SAS_Mediation';
data simdata2;
 set db.simdata7;
 iv=x1;
 med1=x2;
 dv=x4;
 ivsq=iv*iv;
 logmed1=log(med1);
Proc reg data=simdata2 ;
model dv = iv;
model med1 = iv ivsq;
model dv = logmed1 iv;
run;

data simdata3;
set simdata2;
%medcurve (data=simdata3,y=dv,x=iv,m=med1,aform=4,bform=2,cpform=1,xval=1,boot=5000);
run;
```

Fig. 11.15 SAS example with nonlinear paths (Examp11-2.sas)

at another. At least, the empirically based method of bootstrapping provides a solution to the confidence interval estimation. The difference with the procedure we described earlier for the linear models is that now we need to test the significance for several values of x. Taking a large range of values, significance over the range provides a test that the mediation explains the relationship between x and y without ambiguity. However, significance that occurs only over a specific range indicates that there are values of x for which the mediation mechanism is not valid.

Hayes and Preacher (2010) provide a subroutine available in SAS (i.e., *% MEDCURVE.SAS*) to compute these confidence intervals based on the bootstrapping distribution of the indirect effects θ for nonlinear paths. A number of frequently used functional forms can be specified. We now present an example using simulated data.

We use the model specification presented in Eq. (11.26) where we rename the variables as iv, med1, dv, ivsq, and logmed1 for the independent variable, the mediator, the dependent variable, the squared independent variable, and the logarithm of the mediator variable, respectively. Figure 11.15 corresponds to the SAS input file.

As in the earlier example, the paths are estimated by ordinary least squares after having transformed the variables. Although not strictly necessary because these estimations are also performed within the bootstrapping procedure, it is useful to compare these OLS estimates with the estimates from the bootstrapping procedure in order to help follow the output of the procedure. The subroutine requires the following specifications:

%medcurve (data=filename, y=*Dv*, x=*Iv*, m=*Mv*, aform=*a*, bform=*b*, cpform=*c*, boot=*Z*)

where:

Dv is the unique dependent variable name;
Iv is the unique independent variable name;
Mv is the unique mediating variable (med1 in the illustration represented in Fig. 11.15).

a, *b*, and *c* each represents a number used as a code to indicate the specific form of the path, respectively, from the independent variable to the mediator (*a*), from

the mediator to the dependent variable (b), and, for the direct path, from the independent variable to the dependent variable (c), any of which can take any of the following possible values:

1 = linear relationship
2 = logarithmic relationship
3 = exponential relationship
4 = quadratic relationship
5 = inverse relationship

Z is the number of resamplings (i.e., sampling iterations); Hayes (2009) recommends at least 5,000 sampling iterations.

The results of running the file represented in Fig. 11.15 are shown in Fig. 11.16 (note that the subroutine "medcurve.sas" should first be submitted in SAS by opening and running the corresponding file).

Figure 11.16 first shows the SAS output for the OLS runs for each of the three models requested within the "regress" procedure, and then it shows the output of the "medcurv" subroutine. Although the results of this "medcurv" subroutine are identical to the key parameter estimates of the OLS "regress" procedure, we have chosen to highlight the "regress" output section because the variable names here are chosen by the researcher and thus the results may be easier to follow. However, the structure of the report of each model from the "medcurv" subroutine is also straightforward, even if the variable names are automatically generated by the subroutine.

The results in Fig. 11.16 show that the coefficients of all the paths are significant. Considering the mediator equation (model 2 in the output corresponding to the model with dependent variable "med1"), the quadratic terms contain a linear term of 0.39592 (the coefficient of "iv") and a square term of −0.25919 (the coefficient of ivsq), indicating an increasing function at a decreasing rate (they have t values of 8.51 and 7.82, respectively, indicating strong significance). The coefficient of the logarithm of the mediator (the coefficient of logmed1) in the dependent variable equation (model 3 corresponding to model with dependent variable "dv") is also significant with a value of 1.00073 with a t statistic of 23.82. The independent variable does not have any residual information that is not contained in the mediator, as the coefficient of this "iv" variable is not significant. Consequently, the mediation appears to provide a very good explanation for the dependent variable.

The estimates of the marginal indirect effects are highlighted in the "medcurve" section of the output. In that section, the dependent variable ("dv") is represented by Y, the independent variable ("iv") is represented by X, and the mediator ("med1") is represented by M. The estimates for the "instantaneous indirect effect (THETA)" are given in two tables, the first one presenting the mean value of the estimates and the second providing the confidence intervals. In both tables these estimates are listed under the "THETA" column at three values of the independent variable (the three rows in the "XVAL" column). The second row is the estimate of the indirect effect at the value of the mean of the independent variable ($x = -0.0156$). The first and third rows correspond to estimates evaluated at

minus and plus one standard deviation from the mean of the independent variable (i.e., $x = -1.0047$ and $x = 0.9734$). These marginal indirect effect values are 0.3942 when $x = -1.0047$, 0.1357 when $x = -0.0156$, and -0.0348 when $x = 0.9734$. More importantly, while the confidence intervals do not include the zero value when $x = -1.0047$ or when $x = -0.0156$ (the values of the lower bound, i.e., LowerCI, and higher bound, i.e., UpperCI, confidence intervals are

```
                          The REG Procedure
                            Model: MODEL1
                       Dependent Variable: dv
                 Number of Observations Read         400
                 Number of Observations Used         400

                          Analysis of Variance
                                    Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F

Model                      1       18.03829       18.03829      40.09    <.0001
Error                    398      179.06454        0.44991
Corrected Total          399      197.10283

          Root MSE              0.67075    R-Square     0.0915
          Dependent Mean        1.72458    Adj R-Sq     0.0892
          Coeff Var            38.89370

                       Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1        1.72794        0.03354      51.52      <.0001
iv            1        0.21498        0.03395       6.33      <.0001

                          The REG Procedure
                            Model: MODEL2
                       Dependent Variable: med1

                 Number of Observations Read         400
                 Number of Observations Used         400

                          Analysis of Variance
                                    Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F

Model                      2      105.14816       52.57408      62.52    <.0001
Error                    397      333.85463        0.84094
Corrected Total          399      439.00279

          Root MSE              0.91703    R-Square     0.2395
          Dependent Mean        2.72752    Adj R-Sq     0.2357
          Coeff Var            33.62132

                       Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1        2.98668        0.05615      53.20      <.0001
iv            1        0.39592        0.04653       8.51      <.0001
ivsq          1       -0.25919        0.03314      -7.82      <.0001

                          The REG Procedure
                            Model: MODEL3
                       Dependent Variable: dv
                 Number of Observations Read         400
                 Number of Observations Used         400

                          Analysis of Variance

                                    Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F

Model                      2      123.38614       61.69307     332.25    <.0001
Error                    397       73.71669        0.18568
Corrected Total          399      197.10283

          Root MSE              0.43091    R-Square     0.6260
          Dependent Mean        1.72458    Adj R-Sq     0.6241
          Coeff Var            24.98640
```

Fig. 11.16 SAS output for nonlinear path example (Examp11-2a.lst)

```
                    Parameter Estimates
                  Parameter       Standard
Variable     DF    Estimate          Error    t Value    Pr > |t|

Intercept     1     0.82686        0.04354      18.99      <.0001
logmed1       1     1.00073        0.04201      23.82      <.0001
iv            1   -0.00072168      0.02362      -0.03      0.9756

                        VARIABLES
                           IN
                        MEDIATION
                          MODEL

                        Y    DV
                        X    IV
                        M    MED1

                      SAMPLE SIZE
                           400
------------------------------------------------------------------------
                        MODEL OF M

                     Coeff          SE          t          p

          constant    2.9867      0.0561    53.1955     0.0000
          X           0.3959      0.0465     8.5094     0.0000
          X*X        -0.2592      0.0331    -7.8213     0.0000

            R       R-sq          F          p        df1        df2

       0.4894     0.2395    62.5180     0.0000     2.0000   397.0000
------------------------------------------------------------------------
                        MODEL OF Y

                     Coeff          SE          t          p

          constant    0.8269      0.0435    18.9921     0.0000
          X          -0.0007      0.0236    -0.0306     0.9756
          ln(M)       1.0007      0.0420    23.8191     0.0000

            R       R-sq          F          p        df1        df2

       0.7912     0.6260   332.2470     0.0000     2.0000   397.0000
========================================================================
  Instantaneous Indirect Effect (THETA) of X on Y through M at X = XVAL

                        XVAL       THETA          SE

                     -1.0047      0.3942      0.0438
                     -0.0156      0.1357      0.0162
                      0.9734     -0.0348      0.0249

Bias Corrected Bootstrap Confidence Interval for Instantaneous Indirect Effect

                    XVAL     LowerCI       THETA     UpperCI

                 -1.0047      0.3160      0.3942      0.4866
                 -0.0156      0.1041      0.1357      0.1669
                  0.9734     -0.0819     -0.0348      0.0170

                      BOOTSTRAP SAMPLES:
                           5000
------------------------------------------------------------------------
                          NOTES:

              LEVEL OF CONFIDENCE FOR CONFIDENCE INTERVALS:

   XVAL values above are the sample mean and plus/minus one SD from mean.

   SE for THETA is the standard deviation of the bootstrap estimates.
```

Fig. 11.16 (continued)

both positive), the value of theta when $x = 0.9734$ is not significant (the LowerCI bound is negative, i.e., -0.0819, and the HigherCI bound is positive, i.e., 0.0170). Therefore, one could conclude that there is a level of the independent variable at which the indirect effect is not significantly different from zero, i.e., the mediating variable does not explain the relationship from x to y at these levels.

```
Instantaneous Indirect Effect (THETA) of X on Y through M at X = XVAL

                    XVAL       THETA          SE

                  1.0000     -0.0392      0.0251

Bias Corrected Bootstrap Confidence Interval for Instantaneous Indirect Effect

                XVAL     LowerCI       THETA     UpperCI

              1.0000     -0.0873     -0.0392      0.0118
```

Fig. 11.17 SAS output for nonlinear path example at a specific value of x (Examp11-2b.lst)

It is also possible to request the computation of the confidence interval of the indirect effect at a specific value of the independent variable. This is indicated by adding another parameter "xval=" followed by the value at which the estimation is desired. In the line below, the request is made for a value of the independent variable of 1.0:

%medcurve(data=simdata3,y=dv,x=iv,m=med1,aform=4,bform=2, cpform=1,xval=1,boot=5000)

Figure 11.17 shows the output that is obtained.

The marginal indirect effect theta at $x = 1$ is -0.0392, and it is not significantly different from zero since the confidence interval contains the zero.

This raises an interesting question: Can the intermediary variable be thought of as a mediating explanation if the indirect effect going through that variable has no explanatory power at a range of values of the independent variable? Indeed, the paths from the independent variable to the mediator and from the mediator to the dependent variable are both clearly significant throughout the full range of that independent variable. More generally, could a mediator variable explain the lack of effect of x on y? This could generally occur when the paths from the independent variable to the mediator and from the mediator to the dependent variable are in opposite directions so that the direct effect cannot be observed without understanding the role played by the "mediator" in the lack of direct effect. Perhaps it is best, then, not to speak of mediation but simply of explanation, assuming that a mediation is more clearly explaining an effect rather than the lack of an effect. In the example above, this lack of indirect effect in spite of significant paths occurs only on a limited range of the independent variable, but it could happen on a larger range. This points out, however, that, if the significance of the paths alone is insufficient to evaluate a mediation, the insignificance of the indirect effect cannot ignore the explanation provided by the significance of the paths. This is especially important when nonlinear relationships are involved because effects can be significant only at some levels of the independent variable.

We have illustrated the problem using SAS but the STATA procedure presented earlier in Figs. 11.11 and 11.13 can also be adapted. The indirect effect is computed at different levels of the variable and the bootstrapping procedure is applied to these nonlinear combinations of parameters. The same model presented above is specified in the subroutine "bootcmNonLinear.do" shown in Fig. 11.18.

This estimation takes into account the correlations between the two mediating equations using a seemingly unrelated regression, as described earlier (see Fig. 11.9).

```
capture drop program bootcmNonLinear
program bootcmNonLinear, rclass
sureg (med iv ivSq) (dv Lnmed iv)
return scalar indirect =
([med]_b[iv]+2*[med]_b[ivSq]*1.9)*([dv]_b[Lnmed]/([med]_b[_cons]+[med]_b[iv]*1.9+[med]
_b[ivSq]*1.9*1.9))
return scalar direct = [dv]_b[iv]
end
```

Fig. 11.18 STATA subroutine for nonlinear indirect effects evaluated at a given value of the independent variable

```
insheet x1-x4 using "/Users/gatignon/Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/SimData5_CorErr.csv", clear
generate iv = x1
generate ivSq= x1*x1
generate med = x2
generate Lnmed = log(med)
generate dv = x4
sureg (med iv ivSq)  (dv Lnmed iv)
bootstrap r(indirect) r(direct), reps(5000) nodots: bootcmNonLinear
estat boot, bc percentile
```

Fig. 11.19 STATA input file for bootstrapping estimation of nonlinear indirect effects

```
. estat boot, bc percentile

Bootstrap results                                     Number of obs      =       391
                                                      Replications       =      5000

      command:  bootcmNonLinear
         _bs_1:  r(indirect)
         _bs_2:  r(direct)

------------------------------------------------------------------------------
             |    Observed               Bootstrap
             |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |    .0491899  -.0019151    .0245088   -.0009359   .0952567   (P)
             |                                       .0039051   .0996923  (BC)
       _bs_2 |  -.11448598  -.0015493   .03587136   -.1862767  -.0462359   (P)
             |                                      -.1842554  -.0434833  (BC)
------------------------------------------------------------------------------
 (P)    percentile confidence interval
 (BC)   bias-corrected confidence interval
```

Fig. 11.20 STATA output example for bootstrapping estimation of nonlinear indirect effects

The marginal indirect effect is estimated at a value of the independent variable of 1.9. This value can be changed to any value within the range of feasible values of that variable. This is also the model estimated from the data using the STATA input file shown in Fig. 11.19.

The results using STATA are shown in Fig. 11.20.

After considering the correction for bias, the 95% confidence interval [0.0039, 0.0997] does not contain 0 and, therefore, we can conclude that the indirect effect of iv is significant at the 0.05 level. The subroutine can be modified for estimation of the marginal indirect effect at different values of the independent variable. As mentioned earlier, the value of 1.9 was used as an illustration, but the estimation is often done at the mean value of the independent variable, at plus and minus one

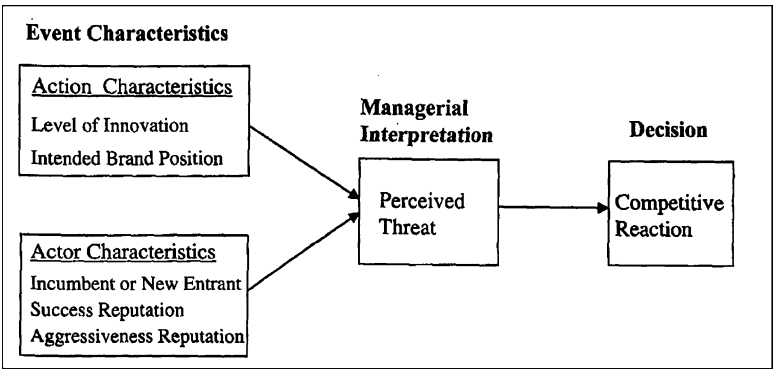


Fig. 11.21 The mediation model tested in Waarts and Wierenga (2000)

and two standard deviations from the mean and at the extreme range of the feasible values. The subroutine can also be easily adapted to reflect the derivatives of the specific mediating equations.

11.2.4.2 Dependent Variable Is Less Than Interval Scale

In the cases analyzed above, the mediator and the independent variables are continuous variables, at least with interval scales. This is not always the case. Let us consider, for example, the study by Waarts and Wierenga (2000) of how a firm reacts when a competing firm introduces a new product. In this case, the dependent variable is whether or not the firm responds, i.e., the dependent variable is binary. The model tested proposes that it is the perceived threat of the introduction that determines whether a firm reacts or not to a competitor, and that this perceived threat is the result of the characteristics of the new product and of the firm introducing it (see Fig. 11.21).

The first path from the independent variables to the mediator (i.e., perceived threat) is similar to what has been discussed thus far. The difference comes from the use of a binomial logit model to estimate the relationship from the mediator to the dichotomous competitive reaction dependent variable. The coefficient from the second path (mediator to dependent variable) determines the probability that the event (here a competitive reaction) will occur. Taking the sequence of paths, the product of the probability of the event multiplied by the marginal effect due to the first path gives the expected indirect effect. Therefore, the relevant calculation in this case is not the product ab but the probability associated with a value of x, which must first be computed. This probability is conditioned by a value of x, just as in the case of nonlinear component paths. In principle, researchers could use the same methods to estimate the indirect effect; however, using these methods with the bootstrapping calculations would be more complicated, although nothing prevents researchers from doing so.

Another example is provided by Chandon, Wesley Hutchinson, Bradlow, and Young (2009), who consider the indirect impact of in-store marketing on brand evaluation through the greater attention that it generates. Brand evaluation is assessed by an ordered scale on three levels: whether the brand was (1) neither chosen nor considered, (2) considered but not chosen, or (3) considered and chosen. They perform a path analysis as discussed earlier in this chapter, except that they estimate the model as a structural equation model as discussed in Sect. 11.2.6.4.

11.2.5 Experimental vs. Non-experimental Data

Thus far in this chapter, we have assumed that the mediating variables were at least interval scales. Except for the example where the dependent variable was a binary variable, the same assumption was made for the dependent variable. When analyzing experimental data, the dependent variables are typically also measured on interval scale items. The mediating variable is typically not part of the manipulated factors in the experimental design. It is usually measured on an interval scale. However, in experimental data (and sometimes with non-experimental data), the independent variable(s) is discrete. Statistically, this does not change the analysis since the use of dummy variables enables us to use the regression method to estimate the differences in effects across levels of the dependent variable. In Chap. 9 (Sect. 9.1.1), we showed that the regression estimation provides estimates of the differences in group means due to different levels of the independent variable. Therefore, a is no longer the marginal effect on the mediator of incrementing the independent variable by one unit, but is now interpretable as the effect of one of the levels vs. the other. Assuming only two levels, with one coded 0 and the other coded 1, the intercept of the moderator equation corresponds to the mean of the first level and the "slope" is the difference in effect for that group. The impact of the second sequence of the indirect path (b) is still the marginal effect of m on y. Therefore, the total indirect effect of switching from the first group (coded 0) to the second group (coded 1) is still the product ab. As there are $K - 1$ dummy variables (where K is the number of groups), there is an estimate of the indirect effect for each of the $K - 1$ groups, i.e., $a_k b_k$. The testing approach is identical whether analyzing experimental or non-experimental data. Note that if the independent variable is effect coded rather than dummy coded (see Chap. 9 on rank-ordered data for the explanation of effect coding vs. dummy variable coding of discrete variables), the effect estimate ab reflects the estimated differences from the grand mean due to the indirect path.

The mediator variable can be categorical. In such a case, the relationship from the independent variable (dummy variables with experimental data) to the mediator variable can be expressed by a logit type of probabilistic model. If the dependent variable is at least interval scale, the second path is identical to what was discussed above.

11.2.6 Regression vs. Structural Equation Modeling

The discussion in this chapter has presented regression approaches outlined in the Baron and Kenny or Preach and Hayes procedures. However, arguments have been presented for preferring a structural equation modeling (SEM) approach instead of testing the validity of paths that correspond to mediation hypotheses. In fact, Iacobucci, Saldanha, and Deng (2007) demonstrate with simulated data that over a significant number of cases, the SEM approach dominates the regression approach.

The test of the joint significance of an indirect path when using structural equation models comes from the likelihood ratio test that compares the significance of the worsening of the fit when constraining the indirect path(s) to be zero. However, in comparing the use of SEM vs. regression à la Baron and Kenney, it is useful to distinguish between three separate issues: correlated errors, measurement errors, and complex nomological networks of relationships.

11.2.6.1 Correlated Errors

As mentioned above, if the errors in the two equations corresponding to the mediator variable model and to the mediator variable effect model are independent, the ordinary least square estimator of individual equations will be identical to the generalized least squares or maximum likelihood estimator that would be derived from a simultaneous estimation of the two equations, i.e., using seemingly unrelated regression or a structural equation model. But in most cases, there is no reason to believe that the correlation would be zero in a simple model that is likely to omit variables, even if these variables are independent from those included in the model (i.e., X). This clearly would favor the use of SUR or SEM. However, a test of the correlation of the errors between the equations is straightforward (see Chap. 6). If the correlation is significant, a simultaneous estimation of the seemingly unrelated regressions would lead to asymptotically more efficient estimators. As indicated in Chap. 6, however, in the case of samples of limited size, the superiority of the simultaneous estimation is not clear and performing both analyses could be useful, as the parameter estimates differ but we do not know which are better.

11.2.6.2 Measurement Errors

The second aspect of the question concerns the incorporation of measurement errors. From that point of view, it is clear that ignoring errors in measurement leads to biased structural parameters, as discussed in Chap. 10. SEM provides a method for estimating structural relationships while taking into account errors in measurement. Therefore, the superiority of this approach is theoretically undeniable. The approach is particularly important as it corrects for a bias that would exist

in the simple models that do not take measurement errors into account. However, the empirical superiority from a practical point of view may not be quite as clear. This is due to the need for excellent measurement model fits in order to reliably estimate structural models of a certain complexity. However, in these more complex cases, the researcher knows that the corrections for measurement errors are small and, consequently, the bias will be minimal. Researchers then face a trade-off between a small theoretical gain and the stability of parameter estimates (especially over model specifications with different paths constrained to zero). Again, it is advisable to perform both the regression and SEM analyses.

11.2.6.3 Complex Nomological Network of Relationships

The third consideration is the number and the complexity of the relationships modeled and estimated. Here it is not an issue of using regressions (in a system of equations simultaneously estimated) vs. analysis of covariance structure modeling. The important point is that the model should not be misspecified by omitting relevant and possibly endogenous variables. The model should accurately reflect the complexity of the phenomenon in a set of nomological networks of relationships. This calls for avoiding simple mediation analyses that involve one independent variable and one mediating factor to explain the dependent variable. Even in experiments that consider a single manipulated variable, there may still be control variables that explain the mediator and the independent variable and there may not be a single mediating variable. The existence of feedback loops is also important to take into consideration, especially because of the ordering of the questions asked in the experiment and of the temporal proximity with which the observations are made. Any endogeneity that would not be considered explicitly in the estimation method leads to biased parameter estimates.

11.2.6.4 Requesting Estimates of Indirect Effects in Structural Equation Models

There are several possible solutions for estimating indirect effects in systems of equations that present some of the characteristics mentioned above. We start with the commands in LISREL and in STATA to obtain these estimates. We illustrate a full case where the constructs that follow a structural path are latent and observed with error. We then illustrate an application where the variables are observed without measurement error. We also suggest the possibility of using a combination where the factor scores reflecting the estimated measurement model are submitted to one of the subroutines proposed by Preacher and Hayes (2004, 2008) or Hayes and Preacher (2010).

```
!Simdata6.lpj
!raw data From File: Simdata6
DA NI=13
RA FI=D:\WORK_SAS\SAS_Mediation\SimData6_4.txt
LA
X5 X6 X7 X8 X9 X10 X11 X12 X13 X1 X2 X3 X4
MO NY=9 NX=4 NE=3 NK=1 BE=FU,FI GA=FU,FI PH=SY TD=DI TE=DI
FI LY(1,1)  LY(5,2) LY(9,3) LX(1,1) TE(9,9)
VA 1 LY(1,1)  LY(5,2) LY(9,3) LX(1,1)
VA 0 TE(9,9)
LK
IV
LE
M1 M2 DV
FR BE(3,1) BE(3,2) GA(1,1) GA(2,1) GA(3,1) C
LY(2,1) LY(3,1) LY(4,1) LY(6,2) LY(7,2) LY(8,2) C
   LX(2,1) LX(3,1) LX(4,1)
Path Diagram
OU SE TV AD=50 MI EF
```

Fig. 11.22 LISREL input file (Examp11-3.ls8)

Estimates of Indirect Effects in LISREL

The model represented in Fig. 11.14 (with two mediating variables following a sequence) is now estimated using data where the variables reflect measures from latent constructs. Therefore, the independent construct X (labeled here IV) influences the dependent variable Y (labeled here DV) through two mediating latent constructs. In the example, the dependent variable is an observed rather than a latent construct, but it could also be an unobserved construct. The LISREL file to estimate such a model is shown in Fig. 11.22.

In this example, there are 13 observed variables. Nine of these (X5 through X13) correspond to the three endogenous constructs (η_1, η_2, η_3, relabeled as M1, M2, and DV) and four (X1 through X4) correspond to the exogenous construct ξ_1 (relabeled IV). Each construct is measured by four variables, except for DV which is not a latent construct and corresponds exactly to the variable X13. Parameter matrices Γ and B are defined in the model specification (the line starting by "MO") as full and fixed (in LISREL a "full" matrix means each of the parameters is defined differently and a "fixed" matrix means the parameters are constant rather than estimated); the specific parameters to be estimated are defined subsequently in the line starting with "FR" (which stands for "FREE"). Note that, in this example, the measurement model parameters are estimated simultaneously with the structural parameters. Following the recommendations in Chap. 10, it is preferable to first estimate the measurement model parameters and then estimate the structural model parameters with the measurement model parameters constrained to the values estimated in the first stage.

The request for estimating the indirect (and the total) effects is included in the last line with all the output options. The option "EF" (highlighted in grey in the figure) instructs the program to provide that information. Figure 11.23 gives the graphical representation of the results (note that the measurement model fits particularly well, with factor loadings that are all rounded to one and variances of the measurement errors that are all rounded to zero). These simulated data serve to illustrate the possibility of using several different methods.

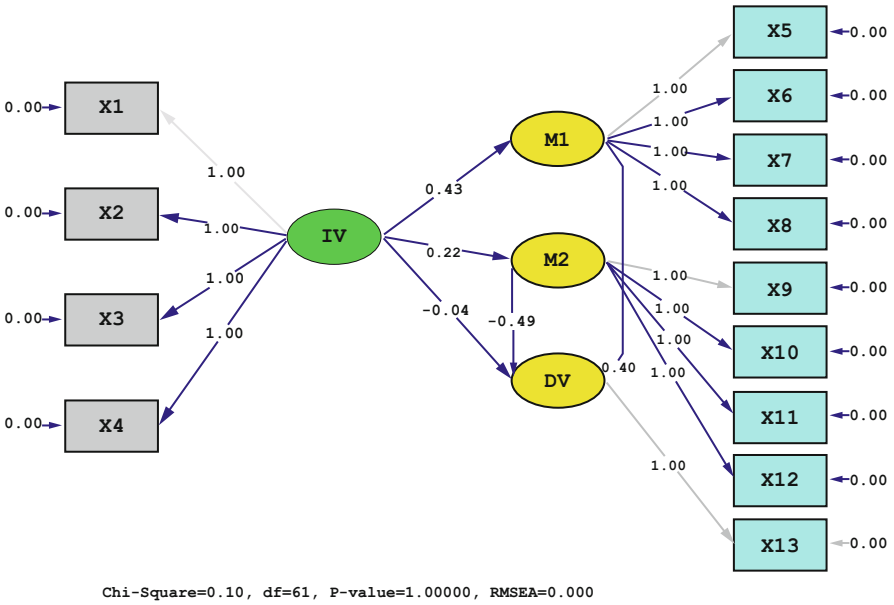


Fig. 11.23 LISREL output graphical representation (Examp11-3.pth)

```
Total and Indirect Effects
        Total Effects of KSI on ETA
                  IV
             --------
       M1         0.43
                 (0.04)
                  9.87

       M2         0.22
                 (0.03)
                  6.35

       DV         0.02
                 (0.03)
                  0.80

        Indirect Effects of KSI on ETA
                  IV
             --------
       M1          - -

       M2          - -

       DV         0.06
                 (0.03)
                  2.43
```

Fig. 11.24 LISREL output on indirect effects (Examp11-8.out)

The parameter estimates indicate that, while the effect of IV on DV, controlling for the indirect effects, is not significant (-0.04 which is not significant as indicated by the t value in the full output), the indirect effect through M1 is $0.43 \times 0.40 = 0.172$, and the indirect effect through M2 is $0.22 \times (-0.49) = -0.108$. The information concerning the indirect and total effects with their significance levels is provided toward the end of the output. The relevant section is reproduced in Fig. 11.24.

The indirect effect of IV on DV is shown to be 0.06, i.e., the sum of the effects through the two indirect paths through M1 and M2 calculated above (0.172 + (−0.108)). This estimate of the indirect effect (highlighted in grey in Fig. 11.24) is statistically significant as indicated by the t value of 2.43 (or the standard error of 0.03 in parentheses). The total effect is the sum of this indirect effect plus the direct path (−0.04 for the link from IV to DV in Fig. 11.23) and, therefore, 0.06 + (−0.04) = 0.02. Figure 11.24 gives the standard error and t statistic for that total effect, which in this case is not significant ($t = 0.80$).

We have illustrated the estimation of structural models using LISREL. As indicated in Chap. 4, AMOS is an alternative software for estimating such models. The structural model parameters can be estimated using different estimation methods, and in particular Bayesian estimations are effective for complex relationships such as those mentioned earlier with the example in Chandon et al. (2009) where the dependent variable is ordinal. The indirect effects are estimated using either the bootstrapping method or the Bayesian method. The interactive commands are straightforward to use and produce results similar to those described in this chapter.

To illustrate the use of STATA for estimating indirect effects in models with multiple mediators, let us consider data from a new product survey. Multiple items measure four constructs: the relative advantage of the new product, the new product complexity, its ease of comprehension by the user, and the respondent's attitude toward the new product. We propose a model where new product complexity affects attitudes toward the new product both directly and indirectly. The indirect path suggests that complexity has an effect through the new product's relative advantage and through the new product's ease of comprehension. The request to estimate such a model with STATA is shown in Fig. 11.25.

```
insheet using "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/NewProdSurvey.csv", clear
egen stx1 = std(x1)
egen stx2 = std(x2)
egen stx3 = std(x3)
egen stx4 = std(x4)
egen stx5 = std(x5)
egen stx6 = std(x6)
egen stx7 = std(x7)
egen stx8 = std(x8)
egen stx9 = std(x9)
egen stx10 = std(x10)
egen stx11 = std(x11)
egen stx12 = std(x12)
egen stx13 = std(x13)
egen stx14 = std(x14)
egen stx15 = std(x15)
egen stx16 = std(x16)
sem (Complex -> stx1 stx2 stx3) ///
(RelAdvan -> stx7 stx8 stx9) ///
(Compreh -> stx10 stx11 stx12) ///
(Attitude -> stx13 stx14 stx15) ///
(Complex -> RelAdvan) ///
(Complex -> Compreh) ///
(Complex RelAdvan Compreh -> Attitude) ///
, nomeans latent(Complex RelAdvan Compreh Attitude)
estat gof, stats(all)
estat framework, fitted
estat teffects
```

Fig. 11.25 STATA input for estimation of indirect effects (Examp11-4a_Mac.do)

The STATA output is shown in Fig. 11.26.

Alternatively, bootstrapping estimation can be requested as in Fig. 11.27.

The output for such bootstrapping estimation is shown in Fig. 11.28.

Instead of investigating the mediation by using a model that contains both the measurement model and the structural model, there are two alternatives using LISREL or SAS. Both of these methods require that the factor analytic model be estimated first, from which the factor scores are derived.

Exporting Factor Scores

The first step in the estimation is to perform a factor analysis (as set out in Chap. 4). We then compute the latent variable scores (Jöreskog, 2000), which are exported and appended to the data set for further utilization with any other statistical software package. In Chap. 4, we presented the "predict" command in STATA; this command computes the factor scores and saves them with a new variable name to the working data set that can then be saved as a ".dta" data file. With LISREL, these tasks are best performed through the interactive facilities. The steps to follow are described below.

Step 1. Create LISREL system file ".PSF"

(a) Import raw data file.

```
Structural equation model                       Number of obs      =       400
Estimation method  = ml
Log likelihood     = -4254.8456

 ( 1)  [stx7]RelAdvan = 1
 ( 2)  [stx10]Compreh = 1
 ( 3)  [stx13]Attitude = 1
 ( 4)  [stx1]Complex = 1
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  RelAdvan <-|
     Complex |  -.1006838   .0523984    -1.92   0.055    -.2033828    .0020152
  -----------+----------------------------------------------------------------
  Compreh <- |
     Complex |  -.4613167   .0491216    -9.39   0.000    -.5575933     -.36504
  -----------+----------------------------------------------------------------
  Attitude <-|
    RelAdvan |    .189158   .0513326     3.68   0.000     .0885479    .2897681
     Compreh |   .1291384   .0545001     2.37   0.018     .0223201    .2359567
     Complex |   -.289725   .0541575    -5.35   0.000    -.3958718   -.1835782
-------------+----------------------------------------------------------------
Measurement  |
  stx1 <-    |
     Complex |          1  (constrained)
  -----------+----------------------------------------------------------------
  stx2 <-    |
     Complex |   .9001634   .0322856    27.88   0.000     .8368847     .963442
  -----------+----------------------------------------------------------------
```

Fig. 11.26 STATA output for estimation of indirect effects—Maximum likelihood estimates (Examp11-4a.log)

```
  stx3 <-      |
       Complex |   .9326928   .0305902    30.49   0.000     .8727372    .9926484
  -------------+----------------------------------------------------------------
  stx7 <-      |
      RelAdvan |           1  (constrained)
  -------------+----------------------------------------------------------------
  stx8 <-      |
      RelAdvan |   .8939928   .0347232    25.75   0.000     .8259367    .9620489
  -------------+----------------------------------------------------------------
  stx9 <-      |
      RelAdvan |   .9217485    .033384    27.61   0.000     .8563171      .98718
  -------------+----------------------------------------------------------------
  stx10 <-     |
       Compreh |           1  (constrained)
  -------------+----------------------------------------------------------------
  stx11 <-     |
       Compreh |   .9638235   .0185384    51.99   0.000     .9274889    1.000158
  -------------+----------------------------------------------------------------
  stx12 <-     |
       Compreh |     .93556   .0216618    43.19   0.000     .8931037    .9780163
  -------------+----------------------------------------------------------------
  stx13 <-     |
      Attitude |           1  (constrained)
  -------------+----------------------------------------------------------------
  stx14 <-     |
      Attitude |   .9815344   .0218838    44.85   0.000      .938643    1.024426
  -------------+----------------------------------------------------------------
  stx15 <-     |
      Attitude |   .9274065   .0264099    35.12   0.000      .875644     .979169
-------------+----------------------------------------------------------------
Variance     |
       e.stx1 |   .0789193   .0166139                      .0522384    .1192276
       e.stx2 |   .2531796   .0220753                       .213408    .3003632
       e.stx3 |   .1984122   .0196626                      .1633859    .2409474
       e.stx7 |   .0880472   .0195938                      .0569235    .1361882
       e.stx8 |   .2706443    .024367                      .2268623    .3228758
       e.stx9 |   .2248104   .0223241                      .1850507    .2731129
      e.stx10 |   .0240058   .0075631                      .0129462    .0445135
      e.stx11 |   .0931669    .009488                      .0763092    .1137488
      e.stx12 |   .1454274   .0121895                      .1233956    .1713928
      e.stx13 |   .0578489   .0104297                      .0406284    .0823682
      e.stx14 |   .0922309   .0113158                      .0725175    .1173033
      e.stx15 |   .1893222   .0158327                      .1607003    .2230418
   e.RelAdvan |   .9001409   .0719922                      .7695422    1.052903
    e.Compreh |   .7780082   .0578284                      .6725358    .9000216
   e.Attitude |   .7572913   .0582738                      .6512725    .8805685
      Complex |   .9185806   .0720331                      .7877128     1.07119
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(49)  =    100.58, Prob > chi2 = 0.0000

…

. estat teffects


Direct effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Measurement  |
  stx1 <-    |
       Complex |           1  (constrained)
  -------------+----------------------------------------------------------------
  stx2 <-    |
       Complex |   .9001634   .0322856    27.88   0.000     .8368847     .963442
  -------------+----------------------------------------------------------------
  stx3 <-    |
       Complex |   .9326928   .0305902    30.49   0.000     .8727372    .9926484
  -------------+----------------------------------------------------------------
```

Fig. 11.26 (continued)

```
  stx7 <-     |
     RelAdvan |          1  (constrained)
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx8 <-     |
     RelAdvan |   .8939928   .0347232    25.75   0.000     .8259367    .9620489
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx9 <-     |
     RelAdvan |   .9217485    .033384    27.61   0.000     .8563171      .98718
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx10 <-    |
      Compreh |          1  (constrained)
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx11 <-    |
      Compreh |   .9638235   .0185384    51.99   0.000     .9274889    1.000158
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx12 <-    |
      Compreh |     .93556   .0216618    43.19   0.000     .8931037    .9780163
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx13 <-    |
     RelAdvan |          0  (no path)
      Compreh |          0  (no path)
     Attitude |          1  (constrained)
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx14 <-    |
     RelAdvan |          0  (no path)
      Compreh |          0  (no path)
     Attitude |   .9815344   .0218838    44.85   0.000      .938643    1.024426
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx15 <-    |
     RelAdvan |          0  (no path)
      Compreh |          0  (no path)
     Attitude |   .9274065   .0264099    35.12   0.000      .875644     .979169
      Complex |          0  (no path)
--------------+------------------------------------------------------------
Structural    |
  RelAdvan <- |
      Complex |  -.1006838   .0523984    -1.92   0.055    -.2033828    .0020152
  ------------+------------------------------------------------------------
  Compreh <-  |
      Complex |  -.4613167   .0491216    -9.39   0.000    -.5575933     -.36504
  ------------+------------------------------------------------------------
  Attitude <- |
     RelAdvan |    .189158   .0513326     3.68   0.000     .0885479    .2897681
      Compreh |   .1291384   .0545001     2.37   0.018     .0223201    .2359567
      Complex |   -.289725   .0541575    -5.35   0.000    -.3958718   -.1835782
------------------------------------------------------------------------------


Indirect effects
------------------------------------------------------------------------------
              |                 OIM
              |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Measurement   |
  stx1 <-     |
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx2 <-     |
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx3 <-     |
      Complex |          0  (no path)
  ------------+------------------------------------------------------------
  stx7 <-     |
     RelAdvan |          0  (no path)
```

Fig. 11.26 (continued)

```
      Complex |  -.1006838   .0523984    -1.92   0.055    -.2033828    .0020152
  ------------+------------------------------------------------------------------
    stx8 <-   |
     RelAdvan |          0  (no path)
      Complex |  -.0900106    .046911    -1.92   0.055    -.1819545    .0019333
  ------------+------------------------------------------------------------------
    stx9 <-   |
     RelAdvan |          0  (no path)
      Complex |  -.0928051   .0483226    -1.92   0.055    -.1875156    .0019054
  ------------+------------------------------------------------------------------
    stx10 <-  |
      Compreh |          0  (no path)
      Complex |  -.4613167   .0491216    -9.39   0.000    -.5575933     -.36504
  ------------+------------------------------------------------------------------
    stx11 <-  |
      Compreh |          0  (no path)
      Complex |  -.4446279   .0477683    -9.31   0.000     -.538252   -.3510038
  ------------+------------------------------------------------------------------
    stx12 <-  |
      Compreh |          0  (no path)
      Complex |  -.4315894   .0467166    -9.24   0.000    -.5231523   -.3400266
  ------------+------------------------------------------------------------------
    stx13 <-  |
     RelAdvan |    .189158   .0513326     3.68   0.000     .0885479    .2897681
      Compreh |   .1291384   .0545001     2.37   0.018     .0223201    .2359567
     Attitude |          0  (no path)
      Complex |  -.3683439   .0494342    -7.45   0.000    -.4652331   -.2714546
  ------------+------------------------------------------------------------------
    stx14 <-  |
     RelAdvan |   .1856651   .0503847     3.68   0.000     .0869128    .2844173
      Compreh |   .1267538   .0534938     2.37   0.018      .021908    .2315997
     Attitude |          0  (no path)
      Complex |  -.3615422   .0486824    -7.43   0.000    -.4569579   -.2661265
  ------------+------------------------------------------------------------------
    stx15 <-  |
     RelAdvan |   .1754264   .0476062     3.68   0.000     .0821199    .2687328
      Compreh |   .1197638   .0505438     2.37   0.018     .0206998    .2188278
     Attitude |          0  (no path)
      Complex |  -.3416045   .0464315    -7.36   0.000    -.4326085   -.2506005
--------------+------------------------------------------------------------------
Structural    |
  RelAdvan <- |
      Complex |          0  (no path)
  ------------+------------------------------------------------------------------
    Compreh <-|
      Complex |          0  (no path)
  ------------+------------------------------------------------------------------
    Attitude <-|
     RelAdvan |          0  (no path)
      Compreh |          0  (no path)
      Complex |  -.0786189   .0266094    -2.95   0.003    -.1307723   -.0264654
------------------------------------------------------------------------------

Total effects
------------------------------------------------------------------------------
              |                 OIM
              |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Measurement   |
  stx1 <-     |
      Complex |          1  (constrained)
  ------------+---------------------------------------------------------------
    stx2 <-   |
      Complex |   .9001634   .0322856    27.88   0.000     .8368847     .963442
  ------------+---------------------------------------------------------------
```

Fig. 11.26 (continued)

```
stx3 <-     |
    Complex |   .9326928   .0305902    30.49   0.000     .8727372    .9926484
------------+----------------------------------------------------------------
stx7 <-     |
   RelAdvan |          1  (constrained)
    Complex |  -.1006838   .0523984    -1.92   0.055    -.2033828    .0020152
------------+----------------------------------------------------------------



stx8 <-     |



   RelAdvan |   .8939928   .0347232    25.75   0.000     .8259367    .9620489
    Complex |  -.0900106    .046911    -1.92   0.055    -.1819545    .0019333
------------+----------------------------------------------------------------
stx9 <-     |
   RelAdvan |   .9217485    .033384    27.61   0.000     .8563171     .98718
    Complex |  -.0928051   .0483226    -1.92   0.055    -.1875156    .0019054
------------+----------------------------------------------------------------
stx10 <-    |
    Compreh |          1  (constrained)
    Complex |  -.4613167   .0491216    -9.39   0.000    -.5575933    -.36504
------------+----------------------------------------------------------------
stx11 <-    |
    Compreh |   .9638235   .0185384    51.99   0.000     .9274889    1.000158
    Complex |  -.4446279   .0477683    -9.31   0.000     -.538252   -.3510038
------------+----------------------------------------------------------------
stx12 <-    |
    Compreh |     .93556   .0216618    43.19   0.000     .8931037    .9780163
    Complex |  -.4315894   .0467166    -9.24   0.000    -.5231523   -.3400266
------------+----------------------------------------------------------------
stx13 <-    |
   RelAdvan |    .189158   .0513326     3.68   0.000     .0885479    .2897681
    Compreh |   .1291384   .0545001     2.37   0.018     .0223201    .2359567
   Attitude |          1  (constrained)
    Complex |  -.3683439   .0494342    -7.45   0.000    -.4652331   -.2714546
------------+----------------------------------------------------------------
stx14 <-    |
   RelAdvan |   .1856651   .0503847     3.68   0.000     .0869128    .2844173
    Compreh |   .1267538   .0534938     2.37   0.018      .021908    .2315997
   Attitude |   .9815344   .0218838    44.85   0.000      .938643    1.024426
    Complex |  -.3615422   .0486824    -7.43   0.000    -.4569579   -.2661265
------------+----------------------------------------------------------------
stx15 <-    |
   RelAdvan |   .1754264   .0476062     3.68   0.000     .0821199    .2687328
    Compreh |   .1197638   .0505438     2.37   0.018     .0206998    .2188278
   Attitude |   .9274065   .0264099    35.12   0.000      .875644     .979169
    Complex |  -.3416045   .0464315    -7.36   0.000    -.4326085   -.2506005
------------+----------------------------------------------------------------
Structural  |
  RelAdvan <-
    Complex |  -.1006838   .0523984    -1.92   0.055    -.2033828    .0020152
------------+----------------------------------------------------------------
  Compreh <-
    Complex |  -.4613167   .0491216    -9.39   0.000    -.5575933    -.36504
------------+----------------------------------------------------------------
  Attitude <-
   RelAdvan |    .189158   .0513326     3.68   0.000     .0885479    .2897681
    Compreh |   .1291384   .0545001     2.37   0.018     .0223201    .2359567
    Complex |  -.3683439   .0494342    -7.45   0.000    -.4652331   -.2714546
------------------------------------------------------------------------------
```

Fig. 11.26 (continued)

An easy way to create such a system data file is to import the data contained in a text file (or Excel or other format) by clicking on FilelImport Data in Free Format. This opens a dialog box where you can search through the computer directories for the proper file with the "txt" extension. Click on the file name

```
insheet using "/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/NewProdSurvey.csv", clear
egen stx1 = std(x1)
egen stx2 = std(x2)
egen stx3 = std(x3)
egen stx4 = std(x4)
egen stx5 = std(x5)
egen stx6 = std(x6)
egen stx7 = std(x7)
egen stx8 = std(x8)
egen stx9 = std(x9)
egen stx10 = std(x10)
egen stx11 = std(x11)
egen stx12 = std(x12)
egen stx13 = std(x13)
egen stx14 = std(x14)
egen stx15 = std(x15)
egen stx16 = std(x16)
sem (Complex -> stx1 stx2 stx3) ///
(RelAdvan -> stx7 stx8 stx9) ///
(Compreh -> stx10 stx11 stx12) ///
(Attitude -> stx13 stx14 stx15) ///
(Complex -> RelAdvan) ///
(Complex -> Compreh) ///
(Complex RelAdvan Compreh -> Attitude) ///
, nomeans latent(Complex RelAdvan Compreh Attitude) ///
vce(bootstrap)
estat gof, stats(all)
estat framework, fitted
estat teffects
```

Fig. 11.27 STATA input for estimation of indirect effects—Bootstrapping (Examp11-4b_Mac.do)

```
(running sem on estimation sample)

Bootstrap replications (50)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50

Structural equation model                       Number of obs      =       400
Log likelihood = -4254.8456                     Replications       =        50

 ( 1)  [stx7]RelAdvan = 1
 ( 2)  [stx10]Compreh = 1
 ( 3)  [stx13]Attitude = 1
 ( 4)  [stx1]Complex = 1
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  RelAdvan <-|
     Complex |  -.1006838   .0558701    -1.80   0.072    -.2101871    .0088195
  -----------+----------------------------------------------------------------
  Compreh <- |
     Complex |  -.4613167   .0379229   -12.16   0.000    -.5356441   -.3869892
  -----------+----------------------------------------------------------------
  Attitude <-|
    RelAdvan |    .189158   .0535395     3.53   0.000     .0842226    .2940934
     Compreh |   .1291384    .045151     2.86   0.004     .0406441    .2176328
     Complex |   -.289725   .0626669    -4.62   0.000    -.4125499   -.1669001
-------------+----------------------------------------------------------------
Measurement  |
```

Fig. 11.28 STATA output for estimation of indirect effects—Bootstrapping (Examp11-4b.log)

```
  stx1 <-     |
      Complex |          1  (constrained)
  ------------+------------------------------------------------------------
  stx2 <-     |
      Complex |  .9001634   .0315387    28.54   0.000     .8383486    .9619781
  ------------+------------------------------------------------------------
  stx3 <-     |
      Complex |  .9326928   .0329998    28.26   0.000     .8680144    .9973712
  ------------+------------------------------------------------------------
  stx7 <-     |
     RelAdvan |          1  (constrained)
  ------------+------------------------------------------------------------
  stx8 <-     |
     RelAdvan |  .8939928   .0301487    29.65   0.000     .8349025    .9530832
  ------------+------------------------------------------------------------
  stx9 <-     |
     RelAdvan |  .9217485   .0330453    27.89   0.000     .8569809    .9865162
  ------------+------------------------------------------------------------
  stx10 <-    |
      Compreh |          1  (constrained)
  ------------+------------------------------------------------------------
  stx11 <-    |
      Compreh |  .9638235   .0144151    66.86   0.000     .9355704    .9920766
  ------------+------------------------------------------------------------
  stx12 <-    |
      Compreh |    .93556   .0268384    34.86   0.000     .8829576    .9881623
  ------------+------------------------------------------------------------
  stx13 <-    |
     Attitude |          1  (constrained)
  ------------+------------------------------------------------------------
  stx14 <-    |
     Attitude |  .9815344   .0244403    40.16   0.000     .9336323    1.029436
  ------------+------------------------------------------------------------
  stx15 <-    |
     Attitude |  .9274065   .0303225    30.58   0.000     .8679756    .9868374
-------------+------------------------------------------------------------
Variance     |
       e.stx1 |  .0789193   .0140746                      .055639    .1119406
       e.stx2 |  .2531796   .0221214                     .2133317    .3004705
       e.stx3 |  .1984122   .0138016                     .1731244    .2273936
       e.stx7 |  .0880472   .0177275                     .0593382    .1306463
       e.stx8 |  .2706443    .025675                     .2247236    .3259486
       e.stx9 |  .2248104   .0219005                     .1857353    .2721062
      e.stx10 |  .0240058   .0083805                     .0121104    .0475854
      e.stx11 |  .0931669   .0145302                     .0686292    .1264778
      e.stx12 |  .1454274   .0173363                     .1151264    .1837036
      e.stx13 |  .0578489   .0130022                     .0372372    .0898695
      e.stx14 |  .0922309   .0168128                     .0645224    .1318386
      e.stx15 |  .1893222   .0201106                     .1537387    .2331416
   e.RelAdvan |  .9001409   .0797827                     .7565984    1.070916
    e.Compreh |  .7780082   .0522722                     .6820156    .8875116
   e.Attitude |  .7572913   .0585038                      .650885    .8810928
      Complex |  .9185806   .0780671                     .7776362    1.085071
------------------------------------------------------------------------------
…
. estat teffects

Direct effects
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Measurement  |
  stx1 <-    |
      Complex |          1  (constrained)
  ------------+------------------------------------------------------------
  stx2 <-    |
      Complex |  .9001634   .0315387    28.54   0.000     .8383486    .9619781
  ------------+------------------------------------------------------------
  stx3 <-    |
```

Fig. 11.28 (continued)

```
      Complex |   .9326928   .0329998    28.26   0.000     .8680144    .9973712
--------------+------------------------------------------------------------------
    stx7 <-   |
     RelAdvan |           1  (constrained)
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
    stx8 <-   |
     RelAdvan |   .8939928   .0301487    29.65   0.000     .8349025    .9530832
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
    stx9 <-   |
     RelAdvan |   .9217485   .0330453    27.89   0.000     .8569809    .9865162
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
    stx10 <-  |
      Compreh |           1  (constrained)
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
    stx11 <-  |
      Compreh |   .9638235   .0144151    66.86   0.000     .9355704    .9920766
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
    stx12 <-  |
      Compreh |     .93556   .0268384    34.86   0.000     .8829576    .9881623
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
    stx13 <-  |
     RelAdvan |           0  (no path)
      Compreh |           0  (no path)
     Attitude |           1  (constrained)
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
    stx14 <-  |
     RelAdvan |           0  (no path)
      Compreh |           0  (no path)
     Attitude |   .9815344   .0244403    40.16   0.000     .9336323    1.029436
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
    stx15 <-  |
     RelAdvan |           0  (no path)
      Compreh |           0  (no path)
     Attitude |   .9274065   .0303225    30.58   0.000     .8679756    .9868374
      Complex |           0  (no path)
--------------+------------------------------------------------------------------
Structural    |
  RelAdvan <- |
      Complex |  -.1006838   .0558701    -1.80   0.072    -.2101871    .0088195
--------------+------------------------------------------------------------------
  Compreh <-  |
      Complex |  -.4613167   .0379229   -12.16   0.000    -.5356441   -.3869892
--------------+------------------------------------------------------------------
  Attitude <- |
     RelAdvan |    .189158   .0535395     3.53   0.000     .0842226    .2940934
      Compreh |   .1291384    .045151     2.86   0.004     .0406441    .2176328
      Complex |   -.289725   .0626669    -4.62   0.000    -.4125499   -.1669001
------------------------------------------------------------------------------

Indirect effects
------------------------------------------------------------------------------
              |   Observed   Bootstrap                         Normal-based
              |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Measurement   |
  stx1 <-     |
      Complex |           0  (no path)
--------------+---------------------------------------------------------------
    stx2 <-   |
      Complex |           0  (no path)
--------------+---------------------------------------------------------------
    stx3 <-   |
      Complex |           0  (no path)
--------------+---------------------------------------------------------------
```

Fig. 11.28 (continued)

```
stx7 <-     |
   RelAdvan |          0  (no path)
    Complex |  -.1006838   .0558701    -1.80   0.072    -.2101871    .0088195
------------+------------------------------------------------------------------
stx8 <-     |
   RelAdvan |          0  (no path)
    Complex |  -.0900106   .0509192    -1.77   0.077    -.1898104    .0097892
------------+------------------------------------------------------------------
stx9 <-     |
   RelAdvan |          0  (no path)
    Complex |  -.0928051   .0510278    -1.82   0.069    -.1928178    .0072076
------------+------------------------------------------------------------------
stx10 <-    |
    Compreh |          0  (no path)
    Complex |  -.4613167   .0379229   -12.16   0.000    -.5356441   -.3869892
------------+------------------------------------------------------------------
stx11 <-    |
    Compreh |          0  (no path)
    Complex |  -.4446279   .0367042   -12.11   0.000    -.5165669   -.3726889
------------+------------------------------------------------------------------
stx12 <-    |
    Compreh |          0  (no path)
    Complex |  -.4315894   .0352511   -12.24   0.000    -.5006803   -.3624985
------------+------------------------------------------------------------------
stx13 <-    |
   RelAdvan |    .189158   .0535395     3.53   0.000     .0842226    .2940934
    Compreh |   .1291384    .045151     2.86   0.004     .0406441    .2176328
   Attitude |          0  (no path)
    Complex |  -.3683439   .0600795    -6.13   0.000    -.4860975   -.2505902
------------+------------------------------------------------------------------
stx14 <-    |
   RelAdvan |   .1856651   .0525508     3.53   0.000     .0826673    .2886628
    Compreh |   .1267538   .0443173     2.86   0.004     .0398936     .213614
   Attitude |          0  (no path)
    Complex |  -.3615422   .0608932    -5.94   0.000    -.4808907   -.2421937
------------+------------------------------------------------------------------
stx15 <-    |
   RelAdvan |   .1754264   .0496528     3.53   0.000     .0781086    .2727441
    Compreh |   .1197638   .0418733     2.86   0.004     .0376936     .201834
   Attitude |          0  (no path)
    Complex |  -.3416045   .0590251    -5.79   0.000    -.4572916   -.2259174
------------+------------------------------------------------------------------
Structural  |
  RelAdvan <- |
    Complex |          0  (no path)
------------+------------------------------------------------------------------
  Compreh <- |
    Complex |          0  (no path)
------------+------------------------------------------------------------------
  Attitude <- |
   RelAdvan |          0  (no path)

    Compreh |          0  (no path)
    Complex |  -.0786189   .0248345    -3.17   0.002    -.1272936   -.0299441
------------------------------------------------------------------------------

Total effects
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Measurement  |
  stx1 <-    |
    Complex |          1  (constrained)
------------+------------------------------------------------------------------
  stx2 <-    |
    Complex |   .9001634   .0315387    28.54   0.000     .8383486    .9619781
------------+------------------------------------------------------------------
  stx3 <-    |
    Complex |   .9326928   .0329998    28.26   0.000     .8680144    .9973712
------------+------------------------------------------------------------------
  stx7 <-    |
   RelAdvan |          1  (constrained)
```

Fig. 11.28 (continued)

```
          Complex |  -.1006838   .0558701    -1.80   0.072    -.2101871    .0088195
------------------+------------------------------------------------------------------
stx8 <-           |
         RelAdvan |   .8939928   .0301487    29.65   0.000     .8349025    .9530832
          Complex |  -.0900106   .0509192    -1.77   0.077    -.1898104    .0097892
------------------+------------------------------------------------------------------
stx9 <-           |
         RelAdvan |   .9217485   .0330453    27.89   0.000     .8569809    .9865162
          Complex |  -.0928051   .0510278    -1.82   0.069    -.1928178    .0072076
------------------+------------------------------------------------------------------
stx10 <-          |
          Compreh |          1  (constrained)
          Complex |  -.4613167   .0379229   -12.16   0.000    -.5356441   -.3869892
------------------+------------------------------------------------------------------
stx11 <-          |
          Compreh |   .9638235   .0144151    66.86   0.000     .9355704    .9920766
          Complex |  -.4446279   .0367042   -12.11   0.000    -.5165669   -.3726889
------------------+------------------------------------------------------------------
stx12 <-          |
          Compreh |    .93556   .0268384    34.86   0.000     .8829576    .9881623
          Complex |  -.4315894   .0352511   -12.24   0.000    -.5006803   -.3624985
------------------+------------------------------------------------------------------
stx13 <-          |
         RelAdvan |    .189158   .0535395     3.53   0.000     .0842226    .2940934
          Compreh |   .1291384    .045151     2.86   0.004     .0406441    .2176328
         Attitude |          1  (constrained)
          Complex |  -.3683439   .0600795    -6.13   0.000    -.4860975   -.2505902
------------------+------------------------------------------------------------------
stx14 <-          |
         RelAdvan |   .1856651   .0525508     3.53   0.000     .0826673    .2886628
          Compreh |   .1267538   .0443173     2.86   0.004     .0398936     .213614
         Attitude |   .9815344   .0244403    40.16   0.000     .9336323    1.029436
          Complex |  -.3615422   .0608932    -5.94   0.000    -.4808907   -.2421937
------------------+------------------------------------------------------------------
stx15 <-          |
         RelAdvan |   .1754264   .0496528     3.53   0.000     .0781086    .2727441
          Compreh |   .1197638   .0418733     2.86   0.004     .0376936     .201834
         Attitude |   .9274065   .0303225    30.58   0.000     .8679756    .9868374
          Complex |  -.3416045   .0590251    -5.79   0.000    -.4572916   -.2259174
------------------+------------------------------------------------------------------
Structural        |
  RelAdvan <-     |
          Complex |  -.1006838   .0558701    -1.80   0.072    -.2101871    .0088195
------------------+------------------------------------------------------------------
  Compreh <-      |
          Complex |  -.4613167   .0379229   -12.16   0.000    -.5356441   -.3869892
------------------+------------------------------------------------------------------
  Attitude <-     |
         RelAdvan |    .189158   .0535395     3.53   0.000     .0842226    .2940934
          Compreh |   .1291384    .045151     2.86   0.004     .0406441    .2176328
          Complex |  -.3683439   .0600795    -6.13   0.000    -.4860975   -.2505902
------------------------------------------------------------------------------
```

Fig. 11.28 (continued)

when found. This opens a new dialog box asking for the number of variables contained in the file (i.e., the number of columns). If the variable names (i.e., labels) are on the first row, check the box for "Variable names at top of file." Clicking "ok" will then create the PSF file.

(b) Rename variables if the names were not on the first row of raw data file. Renaming is done by clicking on Data|Define Variables, which opens a dialog box where you can select a variable by clicking on its name, then clicking on "Rename," and modifying the name. This should be done for all the variables in order to have a consistent data set with the labels used in the LISREL program or other statistical analysis software.

```
SYSTEM FILE from file 'D:\WORK_LISREL\LVSCORES\EXAMP10-1.DSF'
Latent Variables Success CompEnh Radical
Relationships
Q46 = 1.00*Success
Q47 = Success
Q48 = Success
Q5 = 1.00*CompEnh
Q7 = CompEnh
Q8 = CompEnh
Q12 = CompEnh
Q13 = CompEnh
Q14 = CompEnh
Q19r = 1.00*Radical
Q20 = Radical
Q21 = Radical
Q22 = Radical
Q23 = Radical
PSFfile Examp10-1.PSF
End of Problem
```

Fig. 11.29 PRELIS input example for factor analytic model

Step 2. Compute latent variable scores

(c) Create a LISREL file using the SIMPLIS commands.
A new file is created by clicking on File|New, which opens a dialog box where you can select "SIMPLIS Project." Enter a name for this file (which will have an "spj" extension). This opens a new dialog box that has a name corresponding to that project file, including the "spj" extension.

(d) Define the LISREL system file containing the data to be used.
This is done by clicking on Setup|Data, which opens a new dialog box. In that box, choose "LISREL System Data" and click on the "Browse" button to select the data file with the extension ".DSF" corresponding to the proper .PSF data file as created in step 1 above (note that when the PSF file is created, it also creates a file with the ".DSF" extension). Click on "ok" when done.

(e) Start setting up the SIMPLIS commands.
The SIMPLIS basic instructions are inserted at the beginning of the file by clicking on Setup|Build SIMPLIS Syntax.

(f) Define observed and latent variables.
Choose Setup|Variables and click on Add/Read Variables button on the left side of the label dialog box. Make sure that the "Read from file:" option is checked and "LISREL System File" is selected, and then browse through the directory to select the proper data file (with the "DSF" extension). The variables in the database are listed in the table. When the observed variables are read from the data file, add the names of the latent variables, one by one, by clicking on Add Latent Variables.

(g) Complete instructions by inserting commands between the "Relationships" and "End of Problem" lines.
The commands to include correspond to the measurement model. The example in Fig. 11.29 shows the measurement model for three of the constructs used in Chap. 10 (Examp10-1.spl).

(h) Note the command "PSFfile name.PSF" (highlighted in grey in Fig. 11.29).
This command appends the latent variable scores to the data file.

(i) Verify that the latent variable scores have been added to the data file.

```
!Simdata6WithFacScores.lpj
!raw data From File: Simdata6WithFacScores.psf
DA NI=4  MA=KM
RA FI=D:\WORK_SAS\SAS_Mediation\SimData6_3.txt
LA
M1 M2 X13 IV
MO NY=3 NX=1 NE=3 NK=1 BE=FU,FI GA=FU,FI PH=SY TD=DI TE=DI
FI LY(1,1)  LY(2,2) LY(3,3) LX(1,1) TE(1,1) TE(2,2) TE(3,3) TD(1,1)
VA 1 LY(1,1)  LY(2,2) LY(3,3) LX(1,1)
VA 0 TE(1,1) TE(2,2) TE(3,3) TD(1,1)
LK
Xsi1
LE
Eta1 Eta2 Eta3
FR BE(3,1) BE(3,2) GA(1,1) GA(2,1) GA(3,1)
Path Diagram
OU SE TV AD=50 MI EF
```

Fig. 11.30 LISREL input example for structural model using factor scores (Examp11-5.lpj)

Click on File|Open and when the dialog box opens, select "PRELIS Data (*.psf)" in the "Files of Type:" option. Then click on the appropriate data file ("Examp10-1.psf" in Fig. 11.29). Additional columns have been entered with the scores corresponding to the latent variables.

Estimation Using Factor Scores

Once the factor scores have been added to the data file, a structural model can be estimated using these new factor scores. Coming back to the previous example with IV, M1, M2, and DV, the structural model can be specified in LISREL as shown in Fig. 11.30.

The variables that are read from the file that contains the computed factor scores are labeled M1, M2, X13, and IV. The model is identical to the model represented in Fig. 11.23. Here the estimation is done through LISREL, but no measurement model is included with the variables specified in the model. Therefore, for all the constructs, i.e., those labeled Xsi1, Eta1, Eta2, and Eta3, a single measure is read from the data set (labeled M1, M2, X13, and IV). The measurement error variances (θ_δ and θ_ε) are fixed to zero and the factor loadings (Λ_y and Λ_x) are fixed to unity.

This results in the structural parameter estimates represented in Fig. 11.31.

From the results shown in Fig. 11.31, it can be seen that the indirect effect of Xsi1 on Eta3 through Eta1 is $0.44 \times 0.63 = 0.2772$. The indirect effect through Eta2 is $0.30 \times (-0.57) = -0.171$. The direct effect of Xsi1 on Eta3, controlling for the mediating factors, is -0.07.

The request for estimating indirect and total effects on the output parameter line leads to the section of the output shown in Fig. 11.32.

The indirect effect through all mediating factors, i.e., Eta1 and Eta2, is the sum of all the specific indirect effects, i.e., $0.2772 + (-0.171) = 0.1062$ (rounded to 0.11 in Fig. 11.32). This indirect effect is significantly different from zero ($t = 2.43$). The total indirect effect is the sum of the indirect effect (0.1062) and of the direct effect (-0.07), i.e., 0.0362 (rounded to 0.04 in Fig. 11.32).

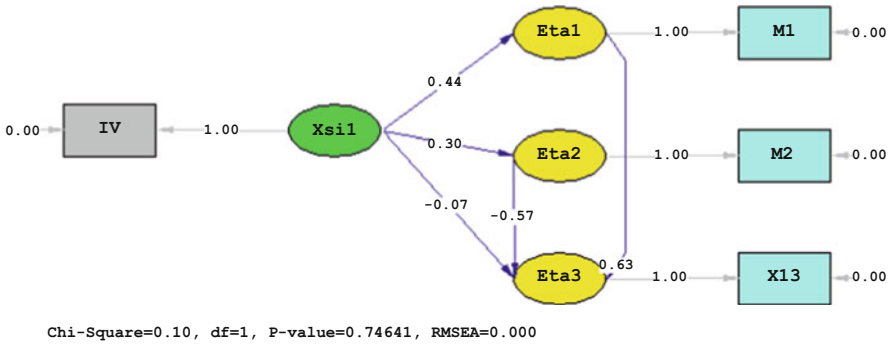


Fig. 11.31 LISREL output example for structural model using factor scores (Examp11-5.pth)

```
Total and Indirect Effects
        Total Effects of KSI on ETA
                Xsi1
             --------
     Eta1        0.44
                (0.04)
                 9.87

     Eta2        0.30
                (0.05)
                 6.35

     Eta3        0.04
                (0.05)
                 0.80

        Indirect Effects of KSI on ETA
                Xsi1
             --------
     Eta1         - -

     Eta2         - -

     Eta3        0.11
                (0.04)
                 2.43
```

Fig. 11.32 LISREL output example of indirect effects using factor scores (Examp11-5.out)

Estimation Using Factor Scores with STATA or SAS

Given the availability of the factor scores in a regular data file, nothing prevents us from using Preacher and Hayes (2004, 2008, 2010) subroutines in SAS, as illustrated earlier in this chapter (Sect. 11.2.2.3). The STATA subroutine presented earlier with Figs. 11.10, 11.11, and 11.12 can also be used to implement Preacher and Hayes' bootstrap procedure. An illustration for an SAS input file is shown in Fig. 11.33.

Even though the results are consistent, the parameter estimates are not quite the same due to the use of different estimation methods. The results from the bootstrapping method are shown in Fig. 11.34.

```
/*SimdataMediationReg6.SAS*/
libname db 'd:\WORK_SAS\SAS_Mediation';

data simdata2;
filename simdata 'D:\WORK_SAS\SAS_Mediation\SimData6_3.csv';
data simdata2;
infile  simdata dlm=';';
input med1 med2 dv iv;
Proc reg data=simdata2 ;
model dv = iv;
model med1 = iv;
model med2 = iv;
model dv = med1 Med2 iv;
run;

data simdata3;
set simdata2;
%indirect(data=simdata3, y=dv, x=iv, m=med1 med2, boot=5000);
run;
```

Fig. 11.33 SAS input example of indirect effects using factor scores (Examp11-6.sas)

Yet another alternative, especially when the error in measurement is negligible, is to perform the analysis using unweighted scores rather than factor scores to represent the constructs. This is particularly useful because the measures are not idiosyncratic to the sample, as they are when using factor scores that are derived from fitting the factor loading to the data sample.

Analysis With Unweighted Scores in SAS

Such an analysis does not require new commands. The scores are simply calculated as the averages of the variables measuring each construct and the analysis can be done easily using these constructed unweighted scores.

11.2.7 Other Issues

11.2.7.1 Standardized vs. Mean-Centered vs. Raw Variables

These transformations, whether we are mean centering all variables or standardizing them, present no particular issues in mediation analysis. Just as in regression analysis, mean centering necessitates estimating the regressions without intercept terms since the intercepts are zero. It does not affect the regression coefficients or their standard errors. The coefficients are interpretable in terms of marginal effects in the original unit of the variables, whether or not the variables are mean centered. Consequently, the size of the effects is readily interpretable in the units of the variables.

Standardization transforms the regression coefficients into partial correlations. In some cases, it may be easier to reason in terms of these relative effects. However,

```
Dependent, Independent, and Proposed Mediator Variables

VS
DV =    DV
IV =    IV
MEDS =  MED1
******* MED2

Sample size
N
400
IV to Mediators (a paths)
BZXMAT
Coeff        se          t          p

MED1    0.4344    0.0441    9.8603    0.0000
MED2    0.2191    0.0345    6.3438    0.0000


Direct Effects of Mediators on DV (b paths)

BYZX2MAT
Coeff        se          t          p

MED1    0.3956    0.0219   18.0740    0.0000
MED2   -0.4885    0.0279  -17.4943    0.0000

Total effect of IV on DV (c path)

BYXMAT
Coeff        se          t          p

IV    0.0249    0.0308    0.8080    0.4196

Direct Effect of IV on DV (c' path)

CPRIMMAT
Coeff        se          t          p

IV   -0.0400    0.0223   -1.7949    0.0734
Fit Statistics for DV Model

DVMS
R-sq  adj R-sq         F       df1       df2          p

0.6119    0.6089  208.1046    3.0000  396.0000    0.0000

*****************************************************
BOOTSTRAP RESULTS FOR INDIRECT EFFECTS

Indirect Effects of IV on DV through Mediators (ab paths)

RES
Data      Boot      Bias        SE

TOTAL    0.0648    0.0647   -0.0002    0.0297
MED1     0.1719    0.1716   -0.0003    0.0210
MED2    -0.1070   -0.1069    0.0001    0.0185


Bias Corrected and Accelerated Confidence Intervals

CI
Lower     Upper

TOTAL    0.0042    0.1203
MED1     0.1314    0.2135
MED2    -0.1465   -0.0726
```

Fig. 11.34 SAS output example of indirect effects using factor scores (Examp11-6.lst)

the indirect effect estimated by the *ab* product term must also be interpreted in terms of these transformed units (i.e., deviations from the mean or the standardized scale), which is not as intuitive as when thinking in the original units of the variables.

11.2.7.2 Multicollinearity Due to Correlation Between Mediator and Independent Variable

The regression of the dependent variable on the mediator and the independent variable may be affected by collinearity. Indeed, if x predicts m, then these two variables, which are included as regressors in the estimated equation, are correlated. At the extreme, a perfect mediator would predict both the outcome accurately without much noise as well as the original "cause" of the effect. This correlation could cause multicollinearity whereby the coefficients of the dependent variable regression would be unstable and unreliable with large standard errors. The regression coefficients would be insignificant even though the model could predict well. Fortunately, the theories expressed in a mediating process are not so strongly associated with the effect that this happens in practice. This correlation problem is related, however, to the next issue, which can be more critical in research.

11.2.7.3 Discriminant Validity of X and M and M with Y

The mediating construct and its measure cannot be too closely related to the independent variable on the one hand and to the dependent variable on the other hand. The constructs should conceptually correspond to differentiated concepts where the mediator is not merely reflecting the same construct as the independent variable. Constructs that were insufficiently differentiated would not only cause the multicollinearity issues alluded to in the previous section, but more critically, they would also simply state tautological relationships of no interest. While multicollinearity is not an issue if the similarity is too great between the mediator and the dependent variable, the tautological issue still applies. Therefore, the mediating variable must provide a true theoretical link between the independent variable (or experimental condition) and the dependent variable, where there is a clear conceptual and empirical discrimination among these three variables (i.e., independent variable, mediator, and dependent variable). This is an important aspect of the evaluation of a mediation model.

11.2.7.4 Temporal Proximity

Another aspect of the evaluation of a mediation model concerns the fact that measures of the effect and its consequences (mediation and dependent variable) are taken in a temporal sequence. The time during which the causal process occurs can be very short, especially in experimental conditions. If the causal mechanism occurs rapidly (temporally proximal), the effects will tend to be larger than if the process is temporally distant (or distal). Shrout and Bolger (2002) indicate that distal processes are smaller because (1) there are multiple causal mechanisms involved sequentially, (2) competing causal mechanisms appear, and (3) random

factors are more likely to intervene. Failing to incorporate all these factors (as discussed in Sect. 11.2.2.1) can make it difficult to detect a relationship between x and y, although the mediating relationships may be more easily observed. Nevertheless, the temporal proximity may not always be desirable, as demand characteristics can undermine the causal process.

11.2.7.5 Recursivity vs. Feedback Loops (Endogeneity)

In mediation models, the focus is clearly on the process where one variable (x) leads to a temporal sequence of events in a chain of causal links. This is indeed the case in experiments where the processes occur within short periods of time (proximal processes). Therefore, the recursive assumption of the mediation model appears to be well justified. When the processes and the measures are more distant, especially in non-experimental data analysis, the pure recursive nature of the system cannot be taken for granted. Mediation analysis as discussed above may become inappropriate because there may be feedback loops in the processes that would require modeling the endogeneity of some of the variables. For example, if there is too much temporal distance between the occurrences of the independent variable, on the one hand, and the measures of the mediating and the dependent variables, on the other hand, the mediating variable can be affected by the outcome. We would then estimate a system of equations with the appropriate structural form and identification, as discussed in Chap. 6. The methods discussed above to test mediation, especially the use of the bootstrapping method for estimating the indirect effects, can still be used. These methods simply need to be adapted to the corresponding parameters estimated with the proper statistical methods (generalized least squares or maximum likelihood). As discussed in Sect. 11.2.6, the joint test of the significance of the indirect parameters comparing the full model with a constrained model without the indirect path is a straightforward alternative.

11.3 Testing Moderation Effects

Moderated effects are represented in Fig. 11.2 above. Although the basic approach is identical whether or not errors in measurement are taken into account, two methodologies for estimating moderation effects are discussed in this section. The first methodology does not take into account errors in measurement and fundamentally applies a regression model that has been labeled moderated regression. The second methodology is based on the analysis of covariance structure, and therefore takes into account the errors in measurement in estimating the moderating effects. We consider several approaches, one which consists in dividing the sample into subgroups and in testing for the differences in structural parameters across groups. The other approaches are more like moderated regression but take into account the errors in measurement.

11.3.1 Moderated Regression

In the case of a single independent variable x and a moderator z, the model is represented algebraically by Eqs. (11.31) and (11.32):

$$y_i = \beta_0 + \beta_{1i} x_i + u_i \tag{11.31}$$

$$\beta_{1i} = \alpha_0 + \alpha_1 z_i \tag{11.32}$$

Equation (11.31) indicates how the dependent variable y responds to the focal independent variable x. Additional independent variables can be added in a linear fashion to control for the effects they may have on the dependent variable y. The only distinction between Eq. (11.31) and any ordinary regression equation is that the response coefficient of the focal variable x (i.e., β_{1i}) contains a subscript for the observation i. This is because this parameter is not fixed across all observations. Instead its value changes depending on the value of the moderating variable z. This variability is expressed by the linear function shown in Eq. (11.32). This second equation expresses the process driving the coefficient and has been called a process equation. It also corresponds to a level-two equation in multi-level modeling or hierarchical linear modeling.

When Eq. (11.32) is inserted into Eq. (11.31), the single moderated equation is

$$y_i = \beta_0 + \alpha_0 x_i + \alpha_1 x_i z_i + u_i \tag{11.33}$$

This last equation contains the focal independent variable x and the product term of the focal variable with the moderator variable z. Therefore, a test of a moderation effect can be performed by augmenting the equation of the focal variable (the response function) with a product term.

Such model specification being linear in the parameters, its estimation should a priori not raise any particular difficulty. Furthermore, the versatility of the regression model to analyze both experimental and non-experimental data allows the application of such a moderated regression model to both kinds of data. In the case of experimental data, the focal variable is represented by one or several dummy variables.

The moderator variable can be discrete as well. If the focal independent variable is continuous and the moderator variable is discrete, a test of moderation is simply a pooling test that the coefficients of the response function are equal across groups. Similar multi-group analysis of the constraint that the coefficients across groups are equal can be performed in analysis of covariance structure models. When both the focal variable and the moderator variable are discrete, i.e., x and z are dummy variables, the model becomes the usual ANOVA model. An example can be found in Franke, Schreier, and Kaiser (2009) where they look at the extent to which individuals who design their own products (for example, a pair of athletic shoes) are willing to pay for them and whether this willingness to pay is moderated by the

extent of the fit with their preferences. Franke et al. manipulate preference fit on two levels and end up with a traditional 2 × 2 ANOVA design.

11.3.1.1 Experimental Data

In experiments, the variables are treatments that are manipulated so as to be independent. In addition, the variables are typically coded so that the means are zero.

Independence of Interaction Term with Its Components in ANOVA

The covariance between the product term of two variables x and z with one of its components x is

$$\mathrm{V}[x, xz] = \mathrm{V}[xz]\mathrm{E}[x] + \mathrm{E}\left[(x - \overline{x})^2(z - \overline{z})\right] + \mathrm{E}[z]\mathrm{V}[x] \tag{11.34}$$

In ANOVA, as mentioned above, the mean of the two variables coding the effects is zero. Consequently, the expression reduces to

$$\mathrm{V}[x, xz] = \mathrm{V}[xz].0 + \mathrm{E}\left[(x - \overline{x})^2(z - \overline{z})\right] + 0.\mathrm{V}[x] \tag{11.35}$$

or

$$\mathrm{V}[x, xz] = \mathrm{E}\left[(x - \overline{x})^2(z - \overline{z})\right] \tag{11.36}$$

And, since the means are zero

$$\mathrm{E}\left[(x - \overline{x})^2(z - \overline{z})\right] = \mathrm{E}\left[x^2 z\right] \tag{11.37}$$

But in ANOVA, the covariance of the two variables coding the effects is zero (they are independent). Therefore,

$$\mathrm{E}\left[x^2 z\right] = 0 \tag{11.38}$$

Coding and Interpretation of Effects

Because it is important to understand the implications of using coding of effects with experimental data, the reader should review the discussion of this issue in Chap. 9, Sect. 9.1.1.

11.3.1.2 Mean Centering

Even if the moderated regression model seems to be flexible and essentially a traditional regression technique, much has been written about it in the literature. Much of this literature is concerned with not being able to detect moderation effects when using the model, especially considering the danger of multicollinearity introduced by having the product term and its components in the estimated regression. These concerns can be addressed by focusing on two issues: mean centering and whether to include the moderator as a single factor in the regression equation.

While much has been written about mean centering, Echambadi and Hess (2007) used a mathematical proof to clarify that mean centering the focal variable and the moderator variable before taking the product term has no effect on the multicollinearity inherent in the moderator regression. Thus, while many authors claim to perform mean centering of the variables to reduce multicollinearity, such mean centering is unnecessary. Mean centering has no impact on the parameter estimate of the product term, neither on the mean estimate nor on its standard deviation.

Mean centering changes the interpretation of the "main" effects simply because the coefficients obtained are estimates that correspond to different values of the moderator variable. More specifically, when using raw data, the coefficient of the focal variable is estimated when the moderator takes the value 0; with mean-centered variables, the coefficient of the focal variable is estimated when the moderator is at the mean value of the moderator. This can facilitate the interpretation in some cases but it depends on the cases. Regardless of mean centering or not, the parameter estimates at the same value of the moderator variable are identical.

What may appear troubling is that mean centering "orthogonalizes" the data matrix. Cronbach (1987) shows that the variance of the product term with the focal variable can be expressed according to Eq. (11.39):

$$\mathrm{V}[x, xz] = \mathrm{V}[xz](\mathrm{E}[x]) + \mathrm{E}\left[(x - \overline{x})^2(z - \overline{z})\right] + (\mathrm{E}[z])\mathrm{V}[x] \tag{11.39}$$

When mean-centered variables $x^d = x - \overline{x}$ and $z^d = z - \overline{z}$ are used, the expression above is reduced to

$$\mathrm{V}\left[x^d, x^d z^d\right] = \mathrm{E}\left[(x - \overline{x})^2(z - \overline{z})\right] \tag{11.40}$$

Because the variances of the raw and the mean-centered variables are equal $\left(\mathrm{V}[x] = \mathrm{E}\left[(x - \overline{x})^2\right] = \mathrm{E}\left[\left(x - \overline{x} - \overline{x}^d\right)^2\right] = \mathrm{V}\left[x^d\right]\right)$, we can compare correlations by comparing the covariances. Inserting Eq. (11.40) into Eq. (11.39) yields

$$\mathrm{V}[x, xz] = \mathrm{V}\left[x^d, x^d z^d\right] + \mathrm{V}[xz]\mathrm{E}[x] + \mathrm{E}[z]\mathrm{V}[x] \tag{11.41}$$

If the expected values of the raw variables are positive (the variances are positive by definition), the covariance of the raw variables is at least as large as the

covariance of the mean-centered variables. And as the expected values of the raw variables increase, so does the difference between the two covariances. So mean centering reduces the bivariate correlation between x and xz, and the greater the means of the variables, the greater this reduction.

How can the relationship described above be reconciled with the proof that multicollinearity is not affected and that none of the parameter estimates changes (given a constant value of the moderator)? The proof has thus far focused on the bivariate correlations. However, the correlations with the intercept term are also changed when mean centering is performed. In fact, the transformation does not change the multivariate relationships since the determinant of the transformed and of the untransformed independent variable covariance matrix remains identical.

This is consistent with Belsley (1984), who shows that coefficients from mean-centered data are as sensitive to the addition or the deletion of a few observations as are coefficients from raw data, and who concludes that typical multicollinearity indicators (such as the variance inflation indicator) are misleading.

Independence of X and Z

It is clear from Eq. (11.40), however, that if the focal variable and the moderator variable are independent, then all the terms in the regression are independent, including the product term. However, the correlation between x and z is rarely the source of a problem, as this correlation remains moderate.

11.3.1.3 Including All Components of the Product Term

The moderating hypothesis is specifically expressed in Eq. (11.33). This equation contains two terms on the right side: the focal variable and the product term of this focal variable with the moderator. The moderating variable does not appear on its own. Theories often do not exclude the possibility that the moderator also has a direct effect. Consequently, we include the moderator variable by itself, and the model to be estimated, then, is shown in Eq. (11.42):

$$y_i = \beta_0 + \alpha_0 x_i + \gamma_0 z_i + \alpha_1 x_i z_i + u_i \tag{11.42}$$

In addition, the inclusion of the moderator as a third term as in Eq. (11.42) is necessary if the focal variable is measured on an interval scale rather than a ratio scale. Because the measure is then only defined up to a scalar κ, a term in the moderating variable is introduced through the product term. This is demonstrated below where x_i is replaced with $(\kappa + x_i)$:

$$\begin{aligned} y_i &= \beta_0 + (\alpha_0 + \alpha_1 z_i)(\kappa + x_i) + u_i \\ &= (\beta_0 + \alpha_0\kappa) + \alpha_0 x_i + \alpha_1 \kappa z_i + \alpha_1 x_i z_i + u_i \\ &= \beta'_0 + \alpha_0 x_i + \gamma_0 z_i + \alpha_1 x_i z + u_i \end{aligned} \tag{11.43}$$

This contributes to a significant worsening of the collinearity problem in moderated regression. However, this problem is inherent in the model specification and mean centering does not solve it, especially for the estimation of the interaction coefficient (on which mean centering has no impact at all). In the case of ratio scales, if the theory postulates only a moderation effect and excludes the possibility of a direct effect of the moderating variable, it may be possible to minimize the problem by estimating a model without a direct effect, i.e., as specified in Eq. (11.33). However, in the case of interval scale variables, which is the most frequent, the complete model specification of Eq. (11.42) cannot be avoided. The major issue, however, concerns the estimation of the coefficients of the component terms (i.e., x and z). In such cases, the estimates, as shown in Eq. (11.43), contain an indeterminate constant, which makes the estimated coefficients impossible to interpret. Fortunately, the theories under investigation in the research typically do not concern these parameters but focus instead on the moderation effect.

11.3.1.4 Estimating the Effects of a Focal Variable Over the Range of the Moderator

In some cases, the test of the existence of a moderating effect is all that the test of the theory requires. This is particularly true in experimental designs where the range of the value of the variables (focal and mediator) is arbitrary, so the magnitude of the differences across groups may be difficult to interpret since the effect sizes are relative to the size of the manipulations. Yet, very often, the presence of a moderation is only one component of the hypothesis, which also requires that the effects be significant at some levels of the moderators and perhaps not significant at other levels or be of the opposite sign. Consequently, it is critical to provide tests of the significance of the effects over a range of values of the moderator variable. Fortunately, the moderation is linear and it is easy to calculate the variance and standard deviation of a linear function of a parameter. When the theory does not focus attention on particular values, it is common practice to evaluate the effect at the mean value of the moderator and at $\pm\ 1\sigma_z$ and $\pm\ 2\sigma_z$. The partial derivative $\frac{\partial y}{\partial x}$ can then be plotted as a function of z, as shown in Fig. 11.35 (Schoonhoven, 1981).

The conditional effects are linear combinations of normally distributed parameters, and thus are normally distributed as well. However, the computation involves the covariance of the estimated parameters. In STATA, the command "lincom" introduced in Chap. 5 for testing linear restrictions of parameters can be used to estimate such conditional effects. The STATA commands for estimating the effects of the focal variable x at two values (−2 and +2) of the moderator variable z are shown in Fig. 11.36.

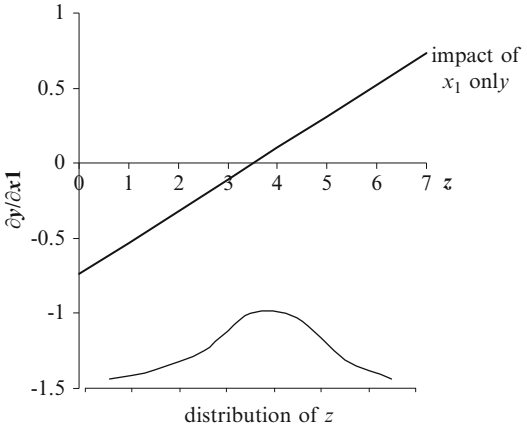


Fig. 11.35 Representation of marginal effect over range of mediator

```
…
generate xz = x*z
regress y x z xz control
*commands for estimating effect at a moderator value
scalar modlevel1=-2
scalar modlevel2=2
lincom _b[x]+_b[xz]*modlevel1
lincom _b[x]+_b[xz]*modlevel2
```

Fig. 11.36 STATA commands to estimate the conditional effect of a focal variable x

```
…
. lincom _b[x]+_b[xz]*modlevel1

 ( 1)  x - 2*xz = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.5693079   .1079681    -5.27   0.000    -.7815719   -.3570439
------------------------------------------------------------------------------

. lincom _b[x]+_b[xz]*modlevel2

 ( 1)  x + 2*xz = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.235332   .1177632    10.49   0.000     1.003811    1.466853
------------------------------------------------------------------------------
```

Fig. 11.37 STATA output of the estimates of conditional effects of a focal variable x

The estimated effects are shown in Fig. 11.37.

The effect of x when the moderator value is -2 is negative and significant (coef $= -0.5693079$, $t = -5.27$), while it is positive and significant at the moderator value of +2 (coef $= 1.235332$, $t = 10.49$). The estimated values of the effect of x at different levels of the moderator variable z can then be plotted in a graph similar to the one shown in Fig. 11.35.

It is recommended that the moderating variables be mean centered. While this has no effect on statistical results, as indicated above, it facilitates the interpretation of the parameter estimates. The coefficient of the focal variable can then be interpreted as the conditional effect of that variable at the mean value of the moderator. In the case of multiple moderator variables, it is the conditional effect of that focal variable when all the moderators are at their mean values and not just when a particular moderator is at its mean value. Mean centering becomes especially convenient when calculating the conditional effect of the focal variable at another value of a moderator variable. If considering the conditional effect when all other moderators are at their mean values, then the coefficients of the other moderators can be ignored since they are multiplied by zero.

Another question may be raised when computing these conditional effects. If the coefficient of the focal variable is insignificant, should it be included in the calculation of the conditional effect at some value of the moderator? Assuming that the moderator variable is mean centered, insignificance of the focal variable coefficient means that the effect of that variable at the mean value of the moderator variable is not significantly different from zero. This does not mean, however, that this coefficient should be ignored in the computation of the effect of the focal variable at another value of the moderator. The linear combination that forms the condition effect needs to take into consideration the parameter of the focal variable as well as the parameter of the interaction term and their variances and covariances. Similarly, when estimating the effect via the bootstrap method, the variance of that linear combination must be estimated.

11.3.1.5 Nonlinear Moderating Effects

As long as the model is linear in the parameters, regression estimations can be used. This requires that the nonlinear relationships be expressed in ways that can be transformed into a linear-in-parameter model. Just as was discussed in the case of nonlinear mediation analysis, a consequence of nonlinearity of moderating relationships is that the effects must be evaluated at a specific value of the moderator. The partial derivatives mentioned in the previous section will need to be estimated at a specific value of the moderator. A range of values of the moderator variable can then be used to assess the significance of such effects over this range. STATA is particularly well suited to perform these analyses by easily adapting the subroutine “bootcm2med.do” described earlier in this chapter.

11.3.1.6 Stochastic/Hierarchical Linear Models and Multi-Level Models

Thus far, moderation effects have been considered as strict constraints on model parameters. However, such effects can also be subject to stochastic error. Such a model specification is used particularly in the context of hierarchical linear models

(HLM). HLM specification leads to random coefficient models that impose a particular structure on the covariance matrix. When the coefficients are conditional on variables at a particular level, these coefficients also reflect moderating effects. While maximum likelihood estimation is possible in many cases, the complexity of the models leads to using Bayesian methods for estimation. Most standard statistical packages that include SEM provide estimations of these hierarchical models.

11.3.1.7 Double Moderation: Three-Way Interactions

When the effect of a moderator variable is itself conditional on a third factor, it leads to a three-way interaction. This creates additional product terms to represent such effects. However, the basic estimation approach is identical to what has been discussed so far in this chapter. Nevertheless, three-way interactions are not always easy to interpret. It is, therefore, highly recommended that there be a strong theory and hypotheses before embarking on higher level interactions.

11.3.1.8 Binomial Dependent Variable: Moderation Effect with Logit Model

The bootstrapping method can also be used with the logit model specification, which is appropriate when the dependent variable takes the values of 0 or 1. An example is shown in Fig. 11.38, where the subroutine calculates the values of the conditional effects of x1 at several values of the moderator variable ("Inter" represents the product term for x1 with the moderator variable). In this example, "effect1" and "effect2" are, respectively, the effect at minus and plus one standard deviation of the moderator (which is centered); "effect3" and "effect4" are, respectively, the effect at minus and plus two standard deviations of the moderator; and "effect5" and "effect6" are, respectively, the effect at the minimum and the maximum values of the moderator variable.

The relevant section of the do-file to estimate the parameters using the actual data is shown in Fig. 11.39.

11.3.2 Incorporating Moderating Effects in Analysis of Covariance Structure

When taking into account the fact that the relationships that are being estimated are among constructs that are measured with error, moderated regression is complicated by the product of the measurement errors. Indeed, the moderated regression is expressed in terms of the unobserved constructs; when computing the product term of these constructs, the measurement errors are multiplied. Therefore, the

```
capture drop program bootcmmodlogit
program bootcmmodlogit, rclass
logit (dv x1 x2 x3 Inter)
return scalar effect1 = ([dv]_b[x1]-([dv]_b[Inter]*0.91))
return scalar effect2 = ([dv]_b[x1]+([dv]_b[Inter]*0.91))
return scalar effect3 = ([dv]_b[x1]-([dv]_b[Inter]*1.82))
return scalar effect4 = ([dv]_b[x1]+([dv]_b[Inter]*1.82))
return scalar effect5 = ([dv]_b[x1]-([dv]_b[Inter]*1.331306))
return scalar effect6 = ([dv]_b[x1]+([dv]_b[Inter]*2.596558))
end
```

Fig. 11.38 STATA subroutine to estimate conditional effects in logit model with the bootstrap method

```
…
logit (dv x1 x2 x3 Inter)
bootstrap r(effect1) r(effect2) r(effect3) r(effect4) ///
       r(effect5) r(effect6), reps(5000) nodots: bootcmmodlogit
estat boot, bc percentile
```

Fig. 11.39 STATA example to estimate conditional effects with the bootstrap method

researcher cannot simply add another variable representing the product term to the analysis of covariance structure models proposed in Chap. 10. One straightforward approach to this issue—the subsample approach—has been widely applied in the social sciences literature and consists in dividing the sample into subsamples in order to show that the structural parameters (the betas and the gammas) are different across subsamples. A second type of approach specifies a moderated causal structural equation model that implies a particular measurement error structure of the product terms to be recognized for the estimation.

11.3.2.1 Multi-Group Analysis or Subgroup Approach

In order to test a moderating hypothesis using multi-group analysis of covariance structure, if the moderator variable is a continuous variable, it is necessary to determine cutoff points along the moderator variable that will be used to form a number of subsamples. The first step raises some important questions concerning how and how many subgroups should be formed. The number of subgroups depends in large part on the hypothesized functional form of the moderating effect. If the moderating effect is linear, two points are sufficient to test this effect. If the hypothesized moderation effect is nonlinear, usually U-shaped or inverted U-shaped, at least three points would be required. The sizes of the subsamples are also critical because they must be large enough to ensure the robustness of the analysis. Given that the unrestricted estimates will require a separate analysis for each group, the minimum size of a group should be similar to the minimum size of any analysis of covariance structure.

The appropriate cutoff values on the moderating factor are more uncertain. If the moderating effect is linear, for example, the estimated value of a subsample will reflect the average effect along the linear relationship. Assuming a normal distribution of the moderator variable and a median split, the effects for each group will

reflect average subgroup effects that will be relatively close to each other. This may lead to a lack of significance of the differences in the structural parameters across subgroups. An alternative would be to consider a larger number of quartiles and possibly delete observations that fall in the middle subgroup(s). However, this approach loses information in truncating the moderating variable, and the choice of cutoff points is somewhat arbitrary. One positive aspect is that, if the median split leads to significant differences in the structural parameters, it lends support to the hypothesis of a moderating effect. An issue arises, however, when this approach leads to insignificant differences. One could then test the sensitivity of the results to different cutoff values, or proceed with more complex approaches as discussed in the next section.

Given a number of groups defined along the moderating variable, the analysis consists in estimating a multi-group structural model as presented in Chap. 10. This provides an unrestricted model where the structural parameters may differ across groups. Just as in single-group analysis of covariance structure, it is appropriate to fix the measurement parameters to those obtained in an estimation of the measurement model without structural relationships. The problem when simultaneously estimating the measurement parameters and the structural parameters is illustrated in the example in Figs. 11.38 and 11.39. Similarly, the issues of measurement invariance discussed in the section on multiple group confirmatory factor analysis (Chap. 4, Sect. 4.5) apply equally in this context. This is due to the fact that the covariances that are analyzed according to the underlying structure among the unobserved constructs and that would be consistent with their empirical values are affected by the measurement parameters. Different factor loadings (lambdas) and different scalars (taus) affect these covariances, even if the means of the unobserved constructs were to be different (because the means do not affect the covariances). Again, the example in Figs. 11.38 and 11.39 illustrates the problem that arises if this precaution is not taken.

To illustrate with a simple example, a two-group sample is now analyzed according to age ("young" respondents and "old" respondents). The LISREL input is shown in Fig. 11.40.

The graphical representation of the output for each group is shown in Fig. 11.41.

The restricted LISREL input is shown in Fig. 11.42.

The results graphically represented in Fig. 11.43 clearly show that the structural parameters are now identical for the two groups.

However, in comparing the unrestricted and restricted estimates of the measurement model, we can see the problem that results from simultaneously estimating the structural parameters and the measurement model parameters. First, we can see that the measurement model of the old group for the second exogenous factor is much poorer than for the young group. The factor loadings are all of similar magnitude in the young group but three out of four are close to zero for the older group. One factor loading for that factor is fixed to one and only the loading for item 11 is estimated and significant (0.59). When imposing the restriction on the structural parameters, this factor loading goes from 0.59 to 0.15. This demonstrates the

```
YOUNG
DA NI=14 NG=2 NO=155 MA=CM
RA FI=d:\WORK_LISREL\Multigroup_SEM\data_young.csv
MO NY=5 NX=9 NE=1 NK=2 TX=FR TY=FR KA=FI PH=SY TD=DI,FR TE=DI,FR
LA
item1 item2 item3 item4 item5       !EndogenousFactorItems
item6 item7 item8 item9             !FirstExogenousFactorItems
item10 item11 item12 item13 item14  !SecondExogenousFactorItems
LE
Endo
LK
Exo1
Exo2
PA LX
4(1 0) 5(0 1)
PA LY
5(1)
FI LX(1,1) LX(5,2) LY(1,1)
VA 1 LX(1,1) LX(5,2) LY(1,1)
OU MI AD=OFF

OLD
DA NO=145
RA FI=d:\WORK_LISREL\Multigroup_SEM\data_old.csv
MO NY=5 NX=9 NE=1 NK=2 TX=FR TY=FR KA=FI PH=SY TD=DI,FR TE=DI,FR
LE
Endo
LK
Exo1
Exo2
PA LX
4(1 0) 5(0 1)
PA LY
5(1)
FI LX(1,1) LX(5,2) LY(1,1)
VA 1 LX(1,1) LX(5,2) LY(1,1)
PD !To get path diagram
OU MI AD=OFF
!End of the program
```

Fig. 11.40 LISREL input for multiple group structural equation model—Unrestricted (Examp11-7.spl)

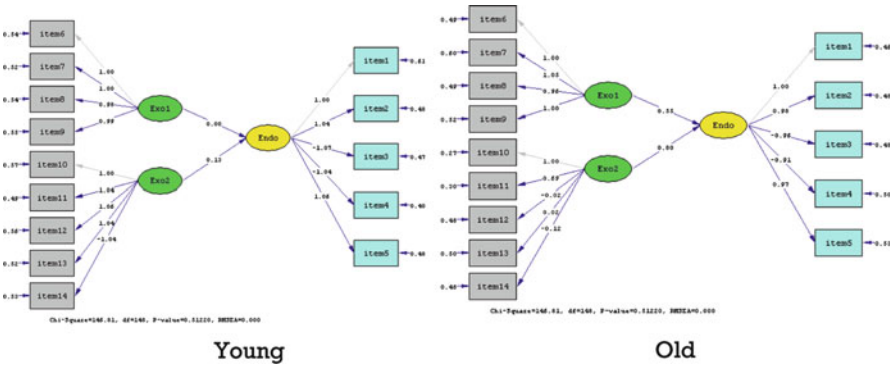


Fig. 11.41 LISREL output of unrestricted analysis (Examp11-7_Young.pth and Examp11-7_Old.pth)

inappropriate trade-off between structural parameter estimates and measurement model parameter estimates when these are estimated simultaneously.

In this example, the imposition of the restriction of equal structural parameters leads to a significantly worse fit. The difference in chi-square is obtained from the

```
YOUNG
DA NI=14 NG=2 NO=155 MA=CM
RA FI=d:\WORK_LISREL\Multigroup_SEM\data_young.csv
MO NY=5 NX=9 NE=1 NK=2 TX=FR TY=FR KA=FI PH=SY TD=DI,FR TE=DI,FR
LA
item1 item2 item3 item4 item5       !EndogenousFactorItems
item6 item7 item8 item9             !FirstExogenousFactorItems
item10 item11 item12 item13 item14  !SecondExogenousFactorItems
LE
Endo
LK
Exo1
Exo2
PA LX
4(1 0) 5(0 1)
PA LY
5(1)
FI LX(1,1) LX(5,2) LY(1,1)
VA 1 LX(1,1) LX(5,2) LY(1,1)
OU MI AD=OFF

OLD
DA NO=145
RA FI=d:\WORK_LISREL\Multigroup_SEM\data_old.csv
MO NY=5 NX=9 NE=1 NK=2 TX=FR TY=FR KA=FI BE=INV GA=INV PH=SY TD=DI,FR TE=DI,FR
LE
Endo
LK
Exo1
Exo2
PA LX
4(1 0) 5(0 1)
PA LY
5(1)
FI LX(1,1) LX(5,2) LY(1,1)
VA 1 LX(1,1) LX(5,2) LY(1,1)
PD !To get path diagram
OU MI AD=OFF
!End of the program
```

Fig. 11.42 LISREL input for multiple group structural equation model—Restricted (Examp11-8.spl)

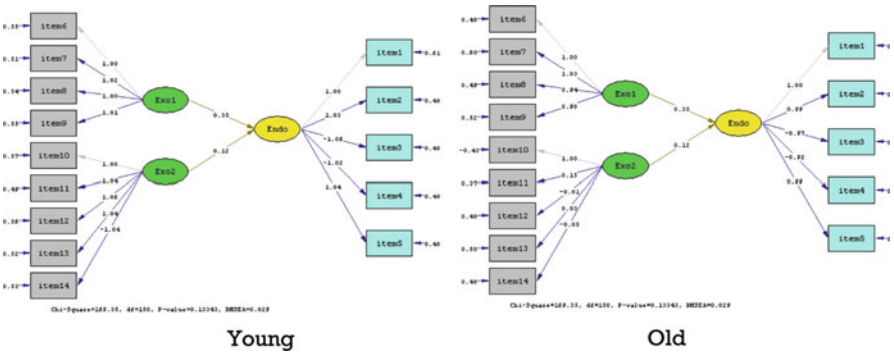


Fig. 11.43 LISREL output of restricted analysis (Examp11-8.pth)

output of the analyses shown in Figs. 11.41 and 11.43. The global (i.e., for all the groups) chi-square of the unrestricted model is 151.24 with 148 degrees of freedom while the global chi-square for the restricted model is 171.84 with 150 degrees of freedom. The difference of 20.60 with 2 degrees of freedom (the number of restricted parameters, i.e., the two structural parameters for group 2 in the example)

```
use "/Volumes/FREEAGENT D/My Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/MULTIGROUP SEM/data_young.dta", clear
append using "/Volumes/FREEAGENT D/My Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/MULTIGROUP SEM/data_old.dta", generate(AgeGroup)
label define AgeGroupLab 0 "Young" 1 "Old"
label values AgeGroup AgeGroupLab
sem (Endo-> item1-item5) ///
(Exo1 -> item6-item9) ///
(Exo2 -> item10-item14) ///
(Exo1 Exo2 -> Endo) ///
, group (AgeGroup)
estat ginvariant
```

Fig. 11.44 STATA input for testing invariance of structural parameters across groups (Examp11-9_Mac.do)

is statistically significant. Consequently, this test provides support for a moderated effect of age. It is important to recall that this example does not constrain the measurement model parameters. This allows us to illustrate the effect of simultaneous estimation (i.e., estimating both measurement and structural parameters simultaneously), as the consequences are more easily identifiable with multi-group analysis. However, actual tests of moderation using multi-group analysis of covariance structure should follow the procedure of stepwise estimation recommended in single-group analysis (Chap. 10).

The same analysis can be performed in STATA. Figure 11.44 lists the input for testing the equality of coefficients across groups using the "ginvariant" procedure. In that run, the two data sets (for young and old subjects) are combined (using the "append" STATA command) into a single data set with the creation of a new variable to indicate the group (young or old).

The model specification indicates that the model is estimated by groups determined by the "AgeGroup" variable, and the request "estat ginvariant" returns the statistics needed to test the hypothesis of equality of coefficients across groups. The output is shown in Fig. 11.45.

As shown in the grey-highlighted part of Fig. 11.45, both coefficients corresponding to the path from Exo1 to Endo and from Exo2 to Endo are significantly different with a chi-square of almost 20 and 1 degree of freedom. The coefficients are indeed different with a value of 0.0016 and 0.538, respectively, for the effect of Exo1 on the young group and the old group, and 0.137 and 0.913, respectively, for the effect of Exo2.

The simple model specification shown in the STATA example above illustrates how easy it is to perform global tests of equality of parameters. The commands simply required an analysis by group and some basic assumptions were automatically made to test the equality of structural parameters. It is also possible to specify which parameters should be equal across groups and which should be freely estimated as different parameters across groups. Table 11.1 shows the various possibilities.

The option ginvariant(class name) defines the parameters that should not vary across groups (to be invariant across groups). For example, ginvariant(scoef) constrains all the structural parameters to be the same in all the groups. The fit of such a constrained model can then be compared with the fit of a model where separate structural coefficients are estimated for each group. The difference in

```
Structural equation model                       Number of obs      =        300
Grouping variable  = AgeGroup                   Number of groups   =          2
Estimation method  = ml
Log likelihood     = -5772.2408

 ( 1)  [item1]0bn.AgeGroup#c.Endo = 1
 ( 2)  [item2]0bn.AgeGroup#c.Endo - [item2]1.AgeGroup#c.Endo = 0
 ( 3)  [item3]0bn.AgeGroup#c.Endo - [item3]1.AgeGroup#c.Endo = 0
 ( 4)  [item4]0bn.AgeGroup#c.Endo - [item4]1.AgeGroup#c.Endo = 0
 ( 5)  [item5]0bn.AgeGroup#c.Endo - [item5]1.AgeGroup#c.Endo = 0
 ( 6)  [item6]0bn.AgeGroup#c.Exo1 = 1
 ( 7)  [item7]0bn.AgeGroup#c.Exo1 - [item7]1.AgeGroup#c.Exo1 = 0
 ( 8)  [item8]0bn.AgeGroup#c.Exo1 - [item8]1.AgeGroup#c.Exo1 = 0
 ( 9)  [item9]0bn.AgeGroup#c.Exo1 - [item9]1.AgeGroup#c.Exo1 = 0
 (10)  [item10]0bn.AgeGroup#c.Exo2 = 1
 (11)  [item11]0bn.AgeGroup#c.Exo2 - [item11]1.AgeGroup#c.Exo2 = 0
 (12)  [item12]0bn.AgeGroup#c.Exo2 - [item12]1.AgeGroup#c.Exo2 = 0
 (13)  [item13]0bn.AgeGroup#c.Exo2 - [item13]1.AgeGroup#c.Exo2 = 0
 (14)  [item14]0bn.AgeGroup#c.Exo2 - [item14]1.AgeGroup#c.Exo2 = 0
 (15)  [item1]0bn.AgeGroup - [item1]1.AgeGroup = 0
 (16)  [item2]0bn.AgeGroup - [item2]1.AgeGroup = 0
 (17)  [item3]0bn.AgeGroup - [item3]1.AgeGroup = 0
 (18)  [item4]0bn.AgeGroup - [item4]1.AgeGroup = 0
 (19)  [item5]0bn.AgeGroup - [item5]1.AgeGroup = 0
 (20)  [item6]0bn.AgeGroup - [item6]1.AgeGroup = 0
 (21)  [item7]0bn.AgeGroup - [item7]1.AgeGroup = 0
 (22)  [item8]0bn.AgeGroup - [item8]1.AgeGroup = 0
 (23)  [item9]0bn.AgeGroup - [item9]1.AgeGroup = 0
 (24)  [item10]0bn.AgeGroup - [item10]1.AgeGroup = 0
 (25)  [item11]0bn.AgeGroup - [item11]1.AgeGroup = 0
 (26)  [item12]0bn.AgeGroup - [item12]1.AgeGroup = 0
 (27)  [item13]0bn.AgeGroup - [item13]1.AgeGroup = 0
 (28)  [item14]0bn.AgeGroup - [item14]1.AgeGroup = 0
 (29)  [item1]1.AgeGroup#c.Endo = 1
 (30)  [item6]1.AgeGroup#c.Exo1 = 1
 (31)  [item10]1.AgeGroup#c.Exo2 = 1
 (32)  [mean(Exo1)]0bn.AgeGroup = 0
 (33)  [mean(Exo2)]0bn.AgeGroup = 0
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  Endo <-    |
    Exo1     |
       Young |    .001572   .0981042     0.02   0.987    -.1907086    .1938527
         Old |    .538155   .0744472     7.23   0.000     .3922413    .6840688
    Exo2     |
       Young |   .1372153   .0976388     1.41   0.160    -.0541533    .3285839
         Old |   .9129961   .1485261     6.15   0.000     .6218902    1.204102
-------------+----------------------------------------------------------------
Measurement  |
  item1 <-   |
    Endo     |
         [*] |           1  (constrained)
    _cons    |
         [*] |   3.654508   .1687282    21.66   0.000     3.323807     3.98521
  -----------+----------------------------------------------------------------
  item2 <-   |
    Endo     |
         [*] |    1.01551   .0308701    32.90   0.000     .9550057    1.076014
    _cons    |
         [*] |   3.782679   .1701907    22.23   0.000     3.449111    4.116246
  -----------+----------------------------------------------------------------
  item3 <-   |
    Endo     |
         [*] |  -1.011684   .0309146   -32.73   0.000    -1.072276   -.9510927
    _cons    |
         [*] |   4.362747   .1697339    25.70   0.000     4.030075     4.69542
  -----------+----------------------------------------------------------------
  item4 <-   |
    Endo     |
```

Fig. 11.45 STATA output for testing invariance of structural parameters across groups (Examp11-9.log)

```
             [*] |  -.9864123   .0298751   -33.02   0.000    -1.044966   -.9278581
           _cons |
             [*] |   4.310584   .1650632    26.11   0.000     3.987066    4.634102
-----------------+------------------------------------------------------------------
item5 <-         |
  Endo           |
             [*] |   1.021744   .0312347    32.71   0.000     .9605248    1.082963
           _cons |
             [*] |   3.649016   .1712602    21.31   0.000     3.313352     3.98468
-----------------+------------------------------------------------------------------
item6 <-         |
  Exo1           |
             [*] |           1  (constrained)
           _cons |
             [*] |    2.90304   .1455207    19.95   0.000     2.617825    3.188255
-----------------+------------------------------------------------------------------
item7 <-         |
  Exo1           |
             [*] |   1.007961   .0330473    30.50   0.000     .9431893    1.072732
           _cons |
             [*] |   2.818037    .146074    19.29   0.000     2.531738    3.104337
-----------------+------------------------------------------------------------------
item8 <-         |
  Exo1           |
             [*] |   .9518236   .0315293    30.19   0.000     .8900273     1.01362
           _cons |
             [*] |    2.8699   .1387963    20.68   0.000     2.597864    3.141936
-----------------+------------------------------------------------------------------
item9 <-         |
  Exo1           |
             [*] |   .9878885   .0324063    30.48   0.000     .9243734    1.051404
           _cons |
             [*] |   2.968809    .143769    20.65   0.000     2.687027    3.250591
-----------------+------------------------------------------------------------------
item10 <-        |
  Exo2           |
             [*] |           1  (constrained)
           _cons |
             [*] |   3.595428    .146593    24.53   0.000     3.308111    3.882745
-----------------+------------------------------------------------------------------
item11 <-        |
  Exo2           |
             [*] |      1.063   .0395041    26.91   0.000     .9855733    1.140427
           _cons |
             [*] |   3.603053   .1538031    23.43   0.000     3.301604    3.904501
-----------------+------------------------------------------------------------------
item12 <-        |
  Exo2           |
             [*] |   1.052546   .0406796    25.87   0.000     .9728155    1.132277
           _cons |
             [*] |   3.595669   .1533595    23.45   0.000      3.29509    3.896248
-----------------+------------------------------------------------------------------
item13 <-        |
  Exo2           |
             [*] |   1.024192   .0397986    25.73   0.000     .9461882    1.102196
           _cons |
             [*] |   3.534309   .1492131    23.69   0.000     3.241857    3.826761
-----------------+------------------------------------------------------------------
item14 <-        |
  Exo2           |
             [*] |  -1.017279   .0396325   -25.67   0.000    -1.094957   -.9396009
           _cons |
             [*] |   4.464521   .1484101    30.08   0.000     4.173643      4.7554
-----------------+------------------------------------------------------------------
Mean             |
  Exo1           |
       Young     |           0  (constrained)
         Old     |   1.332677   .2096696     6.36   0.000     .9217319    1.743622
  Exo2           |
       Young     |           0  (constrained)
         Old     |  -1.850784   .1494709   -12.38   0.000    -2.143742   -1.557827
-----------------+------------------------------------------------------------------
Variance         |
  e.item1        |
```

Fig. 11.45 (continued)

```
  e.item1 |
       Young |   .6056648   .0817177                      .4649286    .7890026
         Old |   .4826563   .0720213                      .3602657     .646626
  e.item2 |
       Young |   .4714206   .0678175                      .3555958    .6249718
         Old |   .4840934   .0719491                      .3617579     .647799
  e.item3 |
       Young |   .4854182   .0699564                      .3659692    .6438542
         Old |   .4832038   .0715528                      .3614802    .6459163
  e.item4 |
       Young |   .4002522    .059983                      .2983803    .5369047
         Old |    .496537   .0720553                       .373619    .6598943
  e.item5 |
       Young |   .4752953   .0684789                      .3583648     .630379
         Old |   .5204365   .0758545                      .3911143    .6925191
  e.item6 |
       Young |   .5402566   .0829017                      .3999303    .7298201
         Old |   .4827213    .079043                      .3502005    .6653897
  e.item7 |
       Young |   .5068803   .0797299                      .3724049    .6899147
         Old |   .6088781   .0927966                      .4516502     .82084
  e.item8 |
       Young |   .5486635   .0811773                      .4105509    .7332385
         Old |    .479729   .0755335                      .3523498    .6531575
  e.item9 |
       Young |   .5413576   .0822101                      .4019959    .7290326
         Old |   .5166863   .0816531                      .3790615    .7042783
  e.item10 |
       Young |   .5641564   .0781854                       .429965    .7402287
         Old |   .4393857    .056197                      .3419621    .5645648
  e.item11 |
       Young |   .4828313   .0718367                      .3607047    .6463072
         Old |   .3706641   .0479804                      .2876055    .4777094
  e.item12 |
       Young |   .5519644   .0786768                      .4174276    .7298625
         Old |   .4651306   .0567188                      .3662504    .5907066
  e.item13 |
       Young |   .5191841   .0743148                      .3921772    .6873222
         Old |   .4905321   .0596569                      .3864979    .6225694
  e.item14 |
       Young |   .5293543   .0753676                      .4004565    .6997413
         Old |   .4731783   .0574044                      .3730439    .6001913
  e.Endo |
       Young |   4.000627    .494189                       3.14037    5.096537
         Old |   2.213287   .2919255                      1.709098    2.866213
  Exo1 |
       Young |    2.90013   .3637434                      2.268073    3.708325
         Old |   3.271281   .4189313                      2.545131    4.204609
  Exo2 |
       Young |   2.900906   .3715719                       2.25686    3.728744
         Old |   .0157472   .0127085                      .0032379    .0765859
-------------+----------------------------------------------------------------
Covariance |
  Exo1 |
    Exo2 |
       Young |   .2373585   .2435056     0.97   0.330    -.2399037    .7146206
         Old |  -.0105437    .048461    -0.22   0.828    -.1055255    .0844381
------------------------------------------------------------------------------
Note: [*] identifies parameter estimates constrained to be equal across
      groups.
LR test of model vs. saturated: chi2(171) =    187.70, Prob > chi2 = 0.1812

. estat ginvariant

Tests for group invariance of parameters

------------------------------------------------------------------------------
             |              Wald Test                     Score Test
             |      chi2         df    p>chi2       chi2          df    p>chi2
-------------+----------------------------------------------------------------
Structural |
  Endo <- |
        Exo1 |     19.021          1    0.0000          .           .         .
        Exo2 |     19.109          1    0.0000          .           .         .
```

Fig. 11.45 (continued)

Measurement						
item1 <-						
Endo	.	.	.	2.836	1	0.0922
_cons	.	.	.	0.477	1	0.4897
item2 <-						
Endo	.	.	.	0.406	1	0.5240
_cons	.	.	.	0.912	1	0.3395
item3 <-						
Endo	.	.	.	1.834	1	0.1756
_cons	.	.	.	3.496	1	0.0615
item4 <-						
Endo	.	.	.	0.846	1	0.3576
_cons	.	.	.	1.410	1	0.2351
item5 <-						
Endo	.	.	.	0.001	1	0.9728
_cons	.	.	.	0.098	1	0.7546
item6 <-						
Exo1	.	.	.	0.060	1	0.8058
_cons	.	.	.	1.555	1	0.2124
item7 <-						
Exo1	.	.	.	0.230	1	0.6312
_cons	.	.	.	0.601	1	0.4381
item8 <-						
Exo1	.	.	.	0.796	1	0.3722
_cons	.	.	.	0.755	1	0.3851
item9 <-						
Exo1	.	.	.	0.021	1	0.8853
_cons	.	.	.	0.107	1	0.7431
item10 <-						
Exo2	.	.	.	0.011	1	0.9161
_cons	.	.	.	0.000	1	0.9943
item11 <-						
Exo2	.	.	.	2.606	1	0.1064
_cons	.	.	.	2.441	1	0.1182
item12 <-						
Exo2	.	.	.	0.007	1	0.9336
_cons	.	.	.	0.001	1	0.9811
item13 <-						
Exo2	.	.	.	0.753	1	0.3856
_cons	.	.	.	0.649	1	0.4206
item14 <-						
Exo2	.	.	.	0.813	1	0.3673
_cons	.	.	.	0.686	1	0.4076
Mean						
Exo1	.	.	.	.	.	.
Exo2	.	.	.	.	.	.
Variance						
e.item1	1.265	1	0.2607	.	.	.
e.item2	0.016	1	0.8978	.	.	.
e.item3	0.000	1	0.9824	.	.	.
e.item4	1.052	1	0.3052	.	.	.
e.item5	0.196	1	0.6577	.	.	.
e.item6	0.255	1	0.6134	.	.	.
e.item7	0.704	1	0.4015	.	.	.
e.item8	0.391	1	0.5319	.	.	.
e.item9	0.046	1	0.8302	.	.	.

Fig. 11.45 (continued)

```
      e.item10 |      1.683          1    0.1945          .          .          .
      e.item11 |      1.686          1    0.1941          .          .          .
      e.item12 |      0.802          1    0.3704          .          .          .
      e.item13 |      0.090          1    0.7639          .          .          .
      e.item14 |      0.351          1    0.5537          .          .          .
        e.Endo |     10.645          1    0.0011          .          .          .
          Exo1 |      0.493          1    0.4825          .          .          .
          Exo2 |     60.394          1    0.0000          .          .          .
-------------+------------------------------------------------------------------
Covariance   |
  Exo1       |
          Exo2 |      0.997          1    0.3181          .          .          .
------------------------------------------------------------------------------
```

Fig. 11.45 (continued)

Table 11.1 STATA commands for tests of equality of parameters across groups

Class description	Class name
Structural coefficients	scoef
Structural intercepts	scons
Measurement coefficients	mcoef
Measurement intercepts	mcons

chi-square is also chi-square distributed and provides a joint test that all the structural coefficients are different across groups. Note that STATA automatically sets the measurement model parameters to be equal across groups when estimating multi-group structural relationships. These equal coefficients are indicated by a "*" in the output.

11.3.2.2 Moderated Causal Approach: Product of Constructs Measured with Error

When the moderator variable is discrete, the multi-group analysis presented in the previous section is appropriate but, as discussed above, if the moderator variable is continuous, discretization leads to a loss of information. Moderated regression allows us to keep the continuous moderator variable as such. However, when the variables involved are latent variables, measured with error that can be estimated with multiple-item indicators, a similar approach can be used in an analysis of covariance structure framework.

Jöreskog (2000) Procedure

The procedure presented earlier in the context of mediation effects with latent variables (Sect. 11.2.6) can also be used for interactions among latent variables. The procedure proposed by Jöreskog consists in estimating the factor scores for the unobserved, latent constructs and creating a product term of these scores for the

latent variables concerned. Then a moderated regression including the interaction term can be estimated.

With this approach, the errors in measurement are taken into consideration in the formation of the latent variable scores and, consequently, the regression results do take into account these measurement errors into account. However, the errors in measurement are then reflected only on the factor loadings and consequently on the weights applied to the items reflecting the latent construct. The same problem as indicated earlier applies here as well, i.e., the fact that the factor scores are specific to the sample. But more critical here is the fact that the reliability of the measures (the variance of the measurement error terms) is not used. This means that the regression weights are still biased due to the measurement uncertainty. The method proposed in the next section, the extended LISREL model, solves this problem. As such, it should now be recommended as best practice.

Jöreskog's method has an advantage, however, because it uses factor scores and is thus very robust, while the extended LISREL model may not easily converge. The fact that the least squares estimator also uses factor scores makes it possible to apply the bootstrapping estimation of a moderated mediation effect using the subroutine described earlier in Sect. 11.2.2.3. In fact, if the variances of the measurement errors are small, as they should be if the measurement model provides a good fit, the results of both methods should be very similar. In general, from a practical point of view, it is advisable to use both approaches and to compare their results.

The Extended LISREL Model

In this method, the approach used in moderated regression is applied to the case of latent variables. Let us consider the model represented graphically in Fig. 11.46.

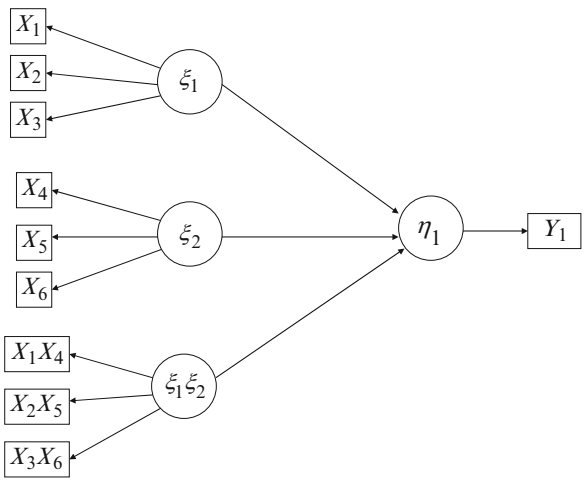


Fig. 11.46 Interactions among latent constructs

In Fig. 11.46, two latent constructs ξ_1 and ξ_2 interact (as a result of, for example, ξ_2 moderating the effect of ξ_1 on η_1). Each construct is measured by three manifest variables (items). The product term of the two latent constructs expresses the interaction just as in moderated regression analysis. The difference is that the components of this "product construct" are not measured without error. Nevertheless, each of the combinations of the items measuring its components is a measure of the product/interaction construct. If one would consider all the combinations of the items involved in the interaction term, it would lead to the full model as originally proposed by Jöreskog and Yang (1996), following the Kenny and Judd (1984) procedure. This, however, implies a repetition of each item in creating multiple product terms among the observed indicators. Therefore, the extended model typically considers the matched-pair strategy where each item is used only in one combination. This strategy is reflected in Fig. 11.46 where the product of latent constructs is reflected by the three product terms of the measures of the latent constructs ξ_1 and ξ_2.

This does not solve all the issues because the model described in Fig. 11.46 has implicit constraints on the parameters due to the repetition of the measures and of the latent variables. These restrictions should be imposed while estimating the model parameters. This will lead to the constrained extended interaction model. We now examine the nature of the restrictions that need to be imposed and then we illustrate the method with an example.

Let us consider two items i and j, each measuring a different latent variable ξ_1 and ξ_2. The measurement equations are given by Eqs. (11.44) and (11.45):

$$X_i = \lambda_{i1}\xi_1 + \delta_i \tag{11.44}$$

$$X_j = \lambda_{j2}\xi_2 + \delta_j \tag{11.45}$$

The product term item is $X_{ij} = X_i X_j$. Replacing these variables with their theoretical expression leads to

$$X_{ij} = (\lambda_{i1}\xi_1 + \delta_i)(\lambda_{j2}\xi_2 + \delta_j) \tag{11.46}$$

$$= \lambda_{i1}\lambda_{j2}\xi_1\xi_2 + \lambda_{i1}\xi_1\delta_j + \lambda_{j2}\xi_2\delta_i + \delta_i\delta_j \tag{11.47}$$

Grouping the last three terms in Eq. (11.47) as the measurement error term δ_{q+1},

$$X_{ij} = \lambda_{q+1}\xi_1\xi_2 + \delta_{q+1} \tag{11.48}$$

The subscript $(q + 1)$ corresponds to the fact that there are q exogenous items and that each product term of the components of the interaction term adds to the number of exogenous variables. The expression in Eq. (11.48) implies a particular structure of the variance:

$$\mathrm{V}\left(X_{ij}\right) = \lambda_{q+1}^2 \mathrm{V}(\xi_1 \xi_2) + \theta_{\delta_{q+1}} \tag{11.49}$$

Furthermore, the augmented factor loading matrix, the Φ matrix of the covariance of the latent constructs, and the θ_δ matrix of the covariance of measurement errors must be constrained to reflect the specific structure imposed by the interaction terms:

$$\lambda_{q+1}^2 = \lambda_{i1}^2 \lambda_{j2}^2 \tag{11.50}$$

$$\theta_{\delta_{q+1}} = \lambda_{i1}^2 \mathrm{V}[\xi_1]\theta_{\delta_j} + \lambda_{j2}^2 \mathrm{V}[\xi_2]\theta_{\delta_i} + \theta_{\delta_i}\theta_{\delta_j} = \lambda_{i1}^2 \Phi_1 \theta_{\delta_j} + \lambda_{j2}^2 \Phi_2 \theta_{\delta_i} + \theta_{\delta_i}\theta_{\delta j} \tag{11.51}$$

$$\mathrm{V}[\xi_1 \xi_2] = \mathrm{V}[\xi_1]\mathrm{V}[\xi_2] + \left(\mathrm{Cov}[\xi_1, \xi_2]\right)^2 = \Phi_1 \Phi_2 + \Phi_{12}^2 \tag{11.52}$$

Equations (11.51) and (11.52) follow from the formula for the variance of the product of two normally distributed random variables. The general formula for two random variables X and Y is

$$\begin{aligned}\mathrm{V}[XY] = {} & (\mathrm{E}[X])^2 \mathrm{V}[Y] + (\mathrm{E}[Y])^2 \mathrm{V}[X] + \mathrm{V}[X]\mathrm{V}[Y] + 2\mathrm{E}[X]\mathrm{E}[Y]\mathrm{Cov}[X, Y] \\ & + (\mathrm{Cov}[X, Y])^2\end{aligned} \tag{11.53}$$

In single-group analysis of covariance, the variables are mean centered (i.e., have a mean of 0) because the mean of the unobserved constructs cannot be estimated. Then, when the variables in Eq. (11.53) are mean centered, this expression reduces to

$$\mathrm{V}[XY] = +\mathrm{V}[X]\mathrm{V}[Y] + (\mathrm{Cov}[X, Y])^2 \tag{11.54}$$

This expression reduces further when the two random variables are independently distributed to

$$\mathrm{V}[XY] = +\mathrm{V}[X]\mathrm{V}[Y] \tag{11.55}$$

Equation (11.51) follows directly from this last formula applied to Eq. (11.47) because of the independence of the unobserved construct with the measurement errors. Equation (11.52) is purely derived from the formula in Eq. (11.54). In addition, the means of the product term of the latent constructs is simply the covariance between the two constructs:

$$\mathrm{E}[\xi_1 \xi_2] = \Phi_{12} \tag{11.56}$$

Consequently, the mean vector κ contains these constraints for the product term means. This implies that a mean structure must always be specified when estimating a model with interaction terms among the latent constructs.

One additional covariance must be examined; it is the covariance between the product term of the unobserved constructs with each of its components, i.e., Cov $[\xi_1, \xi_1 \xi_2]$.

The covariance between a product term XY and one of its components X is given by

$$\mathrm{Cov}[X, XY] = \mathrm{V}[XY]\mathrm{E}[X] + \mathrm{E}\left[\left(X - \overline{X}\right)^2\left(Y - \overline{Y}\right)\right] + \mathrm{E}[Y]\mathrm{V}[X] \tag{11.57}$$

When the variables are mean centered, Eq. (11.57) reduces to

$$\mathrm{Cov}[X, XY] = \mathrm{E}\left[\left(X - \overline{X}\right)^2\left(Y - \overline{Y}\right)\right] = \mathrm{E}\left[X^2 Y\right] \tag{11.58}$$

Applying Eq. (11.58) to the latent variables ξ_1 and ξ_2,

$$\mathrm{Cov}[\xi_1, \xi_1\xi_2] = \mathrm{E}\left[\xi_1^2\xi_2\right] \tag{11.59}$$

It should be noted that, as shown in the literature on mean centering in moderated regression, the expression in Eq. (11.59) is not zero. The fact that the variables have zero mean, however, reduces the correlation. McClelland and Judd (1993) note that when the two components "are centered and are either jointly symmetric or stochastically independent," the expression in Eq. (11.58) is equal to zero (p. 378). This may explain why this expression has been constrained to zero, even though the latent constructs are correlated ($\Phi_{12} \neq 0$). Furthermore, as noted by Kenny and Judd (1984), this model and estimation through maximum likelihood implies that the product terms are normally distributed, which is not the case if the individual components are themselves normally distributed.

We now illustrate the method with an example involving three latent constructs, each measured with four items. The second latent factor is expected to moderate the effect of the first latent factor; consequently, an interaction between these two latent variables is specified. The endogenous variable is observed, i.e., has only one measure for which the loading is constrained to unity and the corresponding error variance to zero.

The LISREL file corresponding to the problem is shown in Fig. 11.47.

The code highlighted in grey indicates the constraints that are imposed on the parameters. The first set of constraints applies to the factor loadings of the product term latent variable. These correspond to Eq. (11.50):

$$\begin{aligned}
&\mathrm{CO\ LX}(14,4) = \mathrm{LX}(2,1)^*\mathrm{LX}(6,2)\\
&\mathrm{CO\ LX}(15,4) = \mathrm{LX}(3,1)^*\mathrm{LX}(7,2)\\
&\mathrm{CO\ LX}(16,4) = \mathrm{LX}(4,1)^*\mathrm{LX}(8,2)
\end{aligned}$$

The constraint on the variance of the interaction latent variable corresponds to Eq. (11.52) as follows:

$$\mathrm{CO\ \ PH}(4,4) = \mathrm{PH}(1,1)^*\mathrm{PH}(2,2) + \mathrm{PH}(2,1)^{**}2$$

```
DA NI=17 NO=400
RA FI='c:\DATA\WORK_SAS\Interactions\SimData8_Iteractions-Augmented_NOVarName.txt'
LA
Y X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X1X5 X2X6 X3X7 X4X8
MO NY=1 NX=16 NE=1 NK=4 LY=FU,FI LX=FU,FR GA=FU,FR PH=SY TE=DI,FI TD=SY PS=DI,FR AL=FR
KA=FI
LK
XI1 XI2 XI3 XI1xXI2
LE
ETA1
PA LY
1
VA 1 LY(1,1)
PA LX
0 0 0 0
1 0 0 0
1 0 0 0
1 0 0 0
0 0 0 0
0 1 0 0
0 1 0 0
0 1 0 0
0 0 0 0
0 0 1 0
0 0 1 0
0 0 1 0
0 0 0 0
0 0 0 1
0 0 0 1
0 0 0 1
VA 1 LX(1,1) LX(5,2) LX(9,3) LX(13, 4)

FI PH(4,1) PH(4,2) PH(4,3)
VA 0 PH(4,1) PH(4,2) PH(4,3)

CO LX(14,4)=LX(2,1)*LX(6,2)
CO LX(15,4)=LX(3,1)*LX(7,2)
CO LX(16,4)=LX(4,1)*LX(8,2)

CO PH(4,4)=PH(1,1)*PH(2,2)+PH(2,1)**2
PA KA
0 0 0 1
CO KA(4)=PH(2,1)

CO TD(13,13)=PH(1,1)*TD(5,5)+PH(2,2)*TD(1,1)+TD(1,1)*TD(5,5)
CO TD(14,14)=LX(2,1)**2*PH(1,1)*TD(6,6)+LX(6,2)**2*PH(2,2)*TD(2,2)+TD(2,2)*TD(6,6)
CO TD(15,15)=LX(3,1)**2*PH(1,1)*TD(7,7)+LX(7,2)**2*PH(2,2)*TD(3,3)+TD(3,3)*TD(7,7)
CO TD(16,16)=LX(4,1)**2*PH(1,1)*TD(8,8)+LX(8,2)**2*PH(2,2)*TD(4,4)+TD(4,4)*TD(8,8)
Path Diagram
OU ME=ML IT=3500 AD=OFF ND=3 RES
```

Fig. 11.47 LISREL input for model of interactions between latent variables—Constrained (Examp11-10.ls8)

The mean of the latent interaction variable corresponds to Eq. (11.56) as follows:

$$\text{CO KA}(4) = \text{PH}(2,1)$$

Finally, according to Eq. (11.51), the variances of the measurement errors corresponding to the four products used as indicators of the interaction latent construct are specified as follows:

$$\text{CO TD}(13,13) = \text{PH}(1,1)^{*}\text{TD}(5,5) + \text{PH}(2,2)^{*}\text{TD}(1,1) + \text{TD}(1,1)^{*}\text{TD}(5,5)$$
$$\text{CO TD}(14,14) = \text{LX}(2,1)^{**}2^{*}\text{PH}(1,1)^{*}\text{TD}(6,6) + \text{LX}(6,2)^{**}2^{*}\text{PH}(2,2)^{*}\text{TD}(2,2) + \text{TD}(2,2)^{*}\text{TD}(6,6)$$
$$\text{CO TD}(15,15) = \text{LX}(3,1)^{**}2^{*}\text{PH}(1,1)^{*}\text{TD}(7,7) + \text{LX}(7,2)^{**}2^{*}\text{PH}(2,2)^{*}\text{TD}(3,3) + \text{TD}(3,3)^{*}\text{TD}(7,7)$$
$$\text{CO TD}(16,16) = \text{LX}(4,1)^{**}2^{*}\text{PH}(1,1)^{*}\text{TD}(8,8) + \text{LX}(8,2)^{**}2^{*}\text{PH}(2,2)^{*}\text{TD}(4,4) + \text{TD}(4,4)^{*}\text{TD}(8,8)$$

The results are shown in the LISREL output reproduced in Fig. 11.48.

The model fits the data matrix very well, providing a nonsignificant chi-square of 145.245. The results are displayed graphically in Fig. 11.49.

As shown in the output tables (Fig. 11.48) and graph (Fig. 11.49), the interaction latent variable is positive (i.e., 0.37), which is consistent with the moderating effect of XI2 on the effect of XI1 on the dependent variable Y.

It has also been suggested that the model could be estimated without specifying these constraints (Kelava, Moosbrugger, Dimitruk, & Schermelleh-Engel, 2008). The unconstrained model specification is shown in Fig. 11.50.

```
 The following lines were read from file
C:\DATA\WORK_SAS\Interactions\ProductTermsConstraints20110915.LS8:

 DA NI=17 NO=400
 RA FI='c:\DATA\WORK_SAS\Interactions\SimData8_Iteractions-Augmented_NOVarName.txt'
 LA
 Y X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X1X5 X2X6 X3X7 X4X8
 MO NY=1 NX=16 NE=1 NK=4 LY=FU,FI LX=FU,FR GA=FU,FR PH=SY TE=DI,FI TD=SY PS=DI,FR
AL=FR KA=FI
 LK
 XI1 XI2 XI3 XI1xXI2
 LE
 ETA1
 PA LY
 1
 VA 1 LY(1,1)
 PA LX
 0 0 0 0
 1 0 0 0
 1 0 0 0
 1 0 0 0
 0 0 0 0
 0 1 0 0
 0 1 0 0
 0 1 0 0
 0 0 0 0
 0 0 1 0
 0 0 1 0
 0 0 1 0
 0 0 0 0
 0 0 0 1
 0 0 0 1
 0 0 0 1
 VA 1 LX(1,1) LX(5,2) LX(9,3) LX(13, 4)

 FI PH(4,1) PH(4,2) PH(4,3)
 VA 0 PH(4,1) PH(4,2) PH(4,3)

 CO LX(14,4)=LX(2,1)*LX(6,2)
 CO LX(15,4)=LX(3,1)*LX(7,2)
 CO LX(16,4)=LX(4,1)*LX(8,2)
 CO PH(4,4)=PH(1,1)*PH(2,2)+PH(2,1)**2
 PA KA
 0 0 0 1
 CO KA(4)=PH(2,1)
 CO TD(13,13)=PH(1,1)*TD(5,5)+PH(2,2)*TD(1,1)+TD(1,1)*TD(5,5)
 CO TD(14,14)=LX(2,1)**2*PH(1,1)*TD(6,6)+LX(6,2)**2*PH(2,2)*TD(2,2)+TD(2,2)*TD(6,6)
 CO TD(15,15)=LX(3,1)**2*PH(1,1)*TD(7,7)+LX(7,2)**2*PH(2,2)*TD(3,3)+TD(3,3)*TD(7,7)
 CO TD(16,16)=LX(4,1)**2*PH(1,1)*TD(8,8)+LX(8,2)**2*PH(2,2)*TD(4,4)+TD(4,4)*TD(8,8)
 Path Diagram
 OU ME=ML IT=3500 AD=OFF ND=3 RES

 DA NI=17 NO=400

                          Number of Input Variables 17
                          Number of Y - Variables    1
                          Number of X - Variables   16
                          Number of ETA - Variables  1
                          Number of KSI - Variables  4
                          Number of Observations   400
…
Parameter Specifications

         LAMBDA-X

                XI1         XI2         XI3    XI1xXI2
           --------    --------    --------    --------
      X1           0           0           0           0
      X2           1           0           0           0
      X3           2           0           0           0
      X4           3           0           0           0
      X5           0           0           0           0
      X6           0           4           0           0
```

Fig. 11.48 LISREL output for model of interactions between latent variables—Constrained (Examp11-10.out)

```
          X7           0           5           0           0
          X8           0           6           0           0
          X9           0           0           0           0
         X10           0           0           7           0
         X11           0           0           8           0
         X12           0           0           9           0
        X1X5           0           0           0           0
        X2X6           0           0           0   Constr'd
        X3X7           0           0           0   Constr'd
        X4X8           0           0           0   Constr'd

        GAMMA

                    XI1         XI2         XI3      XI1xXI2
               --------    --------    --------    --------
        ETA1          10          11          12          13

        PHI

                    XI1         XI2         XI3      XI1xXI2
               --------    --------    --------    --------
         XI1          14
         XI2          15          16
         XI3          17          18          19
      XI1xXI2          0           0           0   Constr'd

        PSI

                   ETA1
               --------
                    20

        THETA-DELTA

                     X1          X2          X3          X4          X5          X6
               --------    --------    --------    --------    --------    --------
                    21          22          23          24          25          26

        THETA-DELTA

                     X7          X8          X9         X10         X11         X12
               --------    --------    --------    --------    --------    --------
                    27          28          29          30          31          32

        THETA-DELTA

                   X1X5        X2X6        X3X7        X4X8
               --------    --------    --------    --------
               Constr'd    Constr'd    Constr'd    Constr'd

        ALPHA

                   ETA1
               --------
                    33

        KAPPA

                    XI1         XI2         XI3      XI1xXI2
               --------    --------    --------    --------
                      0           0           0   Constr'd


 DA NI=17 NO=400

 Number of Iterations =  6

 LISREL Estimates (Maximum Likelihood)

        LAMBDA-Y

                   ETA1
```

Fig. 11.48 (continued)

It should be noted that, while there is no constraint imposed on the covariances, it is important to maintain the constraints on the factor loadings and on the mean of the latent product term. The output of the estimation of such a model is shown in Fig. 11.51.

```
          --------
Y           1.225

 LAMBDA-X

          XI1         XI2         XI3     XI1xXI2
        --------    --------    --------  --------
X1        1.000       - -         - -       - -

X2        0.947       - -         - -       - -
         (0.020)
          48.146

X3        0.888       - -         - -       - -
         (0.017)
          53.178

X4        0.968       - -         - -       - -
         (0.015)
          63.609

X5          - -       1.000       - -       - -

X6          - -       0.914       - -       - -
                     (0.021)
                      44.208

X7          - -       0.919       - -       - -
                     (0.017)
                      53.671

X8          - -       0.969       - -       - -
                     (0.016)
                      61.109

X9          - -         - -       1.000     - -

X10         - -         - -       0.910     - -
                               (0.022)
                                42.072

X11         - -         - -       0.926     - -
                               (0.020)
                                47.467

X12         - -         - -       0.950     - -
                               (0.016)
                                59.768

X1X5        - -         - -         - -     1.000

X2X6        - -         - -         - -     0.866
                                          (0.022)
                                           39.637

X3X7        - -         - -         - -     0.816
                                          (0.018)
                                           46.035

X4X8        - -         - -         - -     0.938
                                          (0.018)
                                           53.558


 GAMMA

          XI1         XI2         XI3     XI1xXI2
        --------    --------    --------  --------
ETA1      0.273      -0.387       0.104     0.370
         (0.040)     (0.040)     (0.038)   (0.040)
           6.821      -9.570       2.727     9.148

 Covariance Matrix of ETA and KSI

          ETA1        XI1         XI2         XI3     XI1xXI2
        --------    --------    --------    --------  --------
ETA1      1.000
```

Fig. 11.48 (continued)

```
            XI1       0.286       0.997
            XI2      -0.388      -0.049       0.974
            XI3       0.095      -0.046       0.016       1.096
         XI1xXI2      0.360       - -         - -         - -         0.974

        Mean Vector of Eta-Variables

               ETA1
            --------
               0.001

        PHI

                  XI1         XI2         XI3     XI1xXI2
            --------    --------    --------    --------
            XI1       0.997
                     (0.056)
                      17.822

            XI2      -0.049       0.974
                     (0.035)     (0.056)
                      -1.399      17.539

            XI3      -0.046       0.016       1.096
                     (0.052)     (0.052)     (0.078)
                      -0.878      0.315       14.091

         XI1xXI2      - -         - -         - -         0.974
                                                         (0.051)
                                                          19.158


        PSI

               ETA1
            --------
               0.629
              (0.045)
               14.087


        Squared Multiple Correlations for Structural Equations

               ETA1
            --------
               0.371

        Squared Multiple Correlations for Y - Variables

                 Y
            --------
               1.000

        THETA-DELTA

                 X1          X2          X3          X4          X5          X6
            --------    --------    --------    --------    --------    --------
               0.003       0.188       0.139       0.109       0.003       0.217
              (0.004)     (0.013)     (0.010)     (0.008)     (0.005)     (0.014)
               0.637      14.408      14.381      13.181       0.660      15.123


        THETA-DELTA

                 X7          X8          X9         X10         X11         X12
            --------    --------    --------    --------    --------    --------
               0.139       0.119       0.000       0.197       0.159       0.102
              (0.010)     (0.009)     (0.005)     (0.015)     (0.012)     (0.008)
               14.079      13.492      0.092       13.563      13.237      12.084

        THETA-DELTA

               X1X5        X2X6        X3X7        X4X8
            --------    --------    --------    --------
               0.006       0.388       0.243       0.224
              (0.006)     (0.019)     (0.012)     (0.012)
               1.015      20.336      20.177      19.403
```

Fig. 11.48 (continued)

```
…
     ALPHA

           ETA1
        --------
           0.019
         (0.040)
           0.481


     KAPPA

            XI1         XI2         XI3     XI1xXI2
        --------    --------    --------    --------
            - -         - -         - -      -0.049
                                            (0.035)
                                             -1.399


                         Goodness of Fit Statistics

                           Degrees of Freedom = 137
          Minimum Fit Function Chi-Square = 145.245 (P = 0.299)
  Normal Theory Weighted Least Squares Chi-Square = 141.689 (P = 0.374)
               Estimated Non-centrality Parameter (NCP) = 4.689
          90 Percent Confidence Interval for NCP = (0.0 ; 37.107)

                    Minimum Fit Function Value = 0.364
            Population Discrepancy Function Value (F0) = 0.0118
           90 Percent Confidence Interval for F0 = (0.0 ; 0.0930)
         Root Mean Square Error of Approximation (RMSEA) = 0.00926
         90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0261)
            P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

                Expected Cross-Validation Index (ECVI) = 0.566
         90 Percent Confidence Interval for ECVI = (0.489 ; 0.582)
                   ECVI for Saturated Model = 0.767
                ECVI for Independence Model = 19.579

  Chi-Square for Independence Model with 136 Degrees of Freedom = 7778.189
                          Independence AIC = 7812.189
                              Model AIC = 225.689
                            Saturated AIC = 306.000
                          Independence CAIC = 7897.044
                              Model CAIC = 435.330
                            Saturated CAIC = 1069.694

                        Normed Fit Index (NFI) = 0.981
                     Non-Normed Fit Index (NNFI) = 0.999
                  Parsimony Normed Fit Index (PNFI) = 0.989
                     Comparative Fit Index (CFI) = 0.999
                     Incremental Fit Index (IFI) = 0.999
                       Relative Fit Index (RFI) = 0.981

                          Critical N (CN) = 491.137


                    Root Mean Square Residual (RMR) = 0.0458
                          Standardized RMR = 0.0444



                       Goodness of Fit Index (GFI) = 0.964



                  Adjusted Goodness of Fit Index (AGFI) = 0.960
                  Parsimony Goodness of Fit Index (PGFI) = 0.864
```

Fig. 11.48 (continued)

This results in very similar parameter estimates with a chi-square that is slightly smaller (with five more parameters estimated leading to 132 degrees of freedom vs. 137 for the constrained estimation). The structural relationships are shown in Fig. 11.52.

These results are consistent and almost identical to those obtained with the constrained estimation.

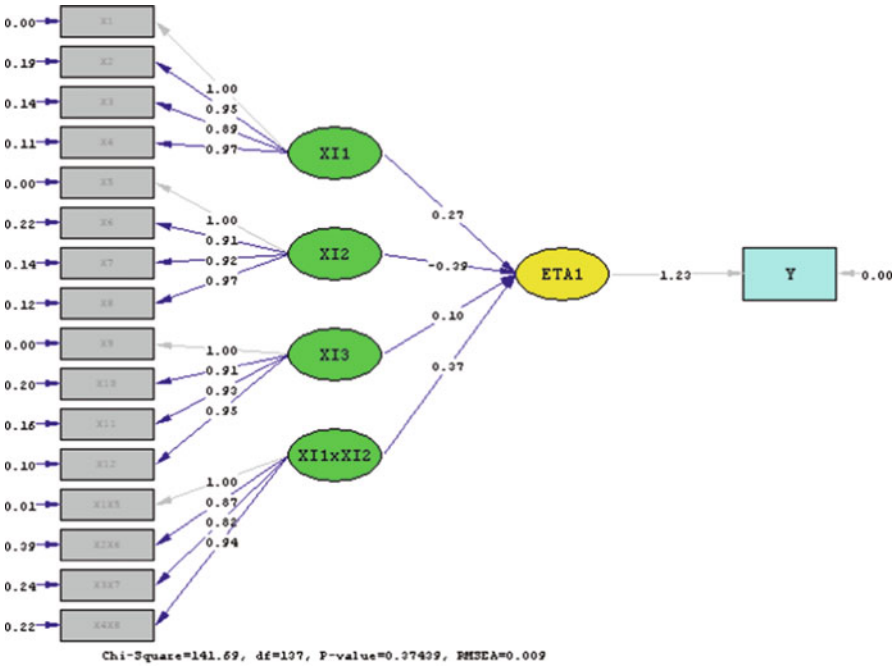


Fig. 11.49 LISREL output graph for model of interactions between latent variables—Constrained (Examp11-10.pth)

To illustrate the use of STATA with such models, we will take two different approaches. It should be noted that the LISREL example used above estimates simultaneously the measurement and the structural parameters. However, as discussed in Chap. 10, it is more appropriate not to trade off measurement fit with structural tests. Consequently, we concluded that an appropriate procedure consists in estimating the measurement model parameters first. In a second stage, the structural parameters can be estimated fixing the measurement model parameters estimated via a confirmatory factor analysis. This means that, in the presence of interactions, the constraints presented above for the interaction components are known. The factor loading of the product of two items is the product of the estimated parameters of these items taken individually. Also, the variance of the product term construct can be set to the estimated value of the covariance of the two relevant constructs. These being the most critical constraints and the variances being fully specified by these values, a simple STATA model can be defined as provided in Fig. 11.53.

The values defined for the factor loadings are those obtained from a prior confirmatory factor analysis. The loadings of the interaction items are the products of the respective loadings of each component item; for example, the loading of x2x6 is the product of the loading of x2 on XI1 (i.e., 0.944) and the loading of x6 on

```
DA NI=17 NO=400
RA FI='c:\DATA\WORK_SAS\Interactions\SimData8_Iteractions-Augmented_NOVarName.txt'
LA
Y X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X1X5 X2X6 X3X7 X4X8
MO NY=1 NX=16 NE=1 NK=4 LY=FU,FI LX=FU,FR GA=FU,FR PH=SY TE=DI,FI TD=SY PS=DI,FR AL=FR
KA=FI
LK
XI1 XI2 XI3 XI1xXI2
LE
ETA1
PA LY
1
VA 1 LY(1,1)
PA LX
0 0 0 0
1 0 0 0
1 0 0 0
1 0 0 0
0 0 0 0
0 1 0 0
0 1 0 0
0 1 0 0
0 0 0 0
0 0 1 0
0 0 1 0
0 0 1 0
0 0 0 0
0 0 0 1
0 0 0 1
0 0 0 1
VA 1 LX(1,1) LX(5,2) LX(9,3) LX(13, 4)

FI PH(4,1) PH(4,2) PH(4,3)
VA 0 PH(4,1) PH(4,2) PH(4,3)

CO LX(14,4)=LX(2,1)*LX(6,2)
CO LX(15,4)=LX(3,1)*LX(7,2)
CO LX(16,4)=LX(4,1)*LX(8,2)

PA KA
0 0 0 1
CO KA(4)=PH(2,1)

Path Diagram
OU ME=ML IT=3500 AD=OFF ND=3 RES
```

Fig. 11.50 LISREL input for model of interactions between latent variables—Partially constrained (Examp11-11.ls8)

XI2 (i.e., 0.913), which gives $0.944 \times 0.913 = 0.8619$. The means of the exogenous latent variables are set to a value of 0, except for the interaction latent construct that is constrained to equal the covariance of the latent variables XI1 and XI2 (i.e., −0.0504738, as per the output of the confirmatory factor analysis—not shown here). In complex models, it is often necessary to provide initial values in order to prevent non-concavity and endless iterations of the optimization routine. This is done by indicating starting values for the structural parameters to be estimated in the structural equation specifications. The commands highlighted in grey, "init(*value*)" in Fig. 11.53, are for that purpose. The specific values can be taken from a different model estimation, for example as suggested below using least square estimation methods with factor scores or unweighted composite scales.

The results, as seen in Fig. 11.54, show values of the structural parameters similar to those obtained via LISREL above. The constraint on the variance of the latent product construct can also be added at the end of the option line:

", var(e.y@0) means(XI1@0 XI2@0 XI3@0 XI1xXI2@-0.0504738) var (XI1xXI2@1.3032)"

```
 The following lines were read from file
C:\DATA\WORK_SAS\Interactions\ProductTermsNOConstraints20110915.LS8:

 DA NI=17 NO=400
 RA FI='c:\DATA\WORK_SAS\Interactions\SimData8_Iteractions-Augmented_NOVarName.txt'
 LA
 Y X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X1X5 X2X6 X3X7 X4X8
 MO NY=1 NX=16 NE=1 NK=4 LY=FU,FI LX=FU,FR GA=FU,FR PH=SY TE=DI,FI TD=SY PS=DI,FR
AL=FR KA=FI
 LK
 XI1 XI2 XI3 XI1xXI2
 LE
 ETA1
 PA LY
 1
 VA 1 LY(1,1)
 PA LX
 0 0 0 0
 1 0 0 0
 1 0 0 0
 1 0 0 0
 0 0 0 0
 0 1 0 0
 0 1 0 0
 0 1 0 0
 0 0 0 0
 0 0 1 0
 0 0 1 0
 0 0 1 0
 0 0 0 0
 0 0 0 1
 0 0 0 1
 0 0 0 1
 VA 1 LX(1,1) LX(5,2) LX(9,3) LX(13, 4)

 FI PH(4,1) PH(4,2) PH(4,3)
 VA 0 PH(4,1) PH(4,2) PH(4,3)

 CO LX(14,4)=LX(2,1)*LX(6,2)
 CO LX(15,4)=LX(3,1)*LX(7,2)
 CO LX(16,4)=LX(4,1)*LX(8,2)

 PA KA
 0 0 0 1
 CO KA(4)=PH(2,1)

 Path Diagram
 OU ME=ML IT=3500 AD=OFF ND=3 RES

 DA NI=17 NO=400

                           Number of Input Variables 17
                           Number of Y - Variables    1
                           Number of X - Variables   16
                           Number of ETA - Variables  1
                           Number of KSI - Variables  4
                           Number of Observations   400
…
 Number of Iterations =  5

 LISREL Estimates (Maximum Likelihood)

         LAMBDA-Y

               ETA1
            --------
        Y      1.228


         LAMBDA-X

                 XI1        XI2        XI3    XI1xXI2
```

Fig. 11.51 LISREL output for model of interactions between latent variables—Partially constrained (Examp11-11.out)

```
          --------    --------    --------    --------
     X1      1.000       - -         - -         - -
     X2      0.944       - -         - -         - -
            (0.019)
            48.526

     X3      0.891       - -         - -         - -
            (0.017)
            53.139

     X4      0.968       - -         - -         - -
            (0.015)
            64.006

     X5        - -       1.000       - -         - -

     X6        - -       0.911       - -         - -
                        (0.020)
                        44.847

     X7        - -       0.922       - -         - -
                        (0.017)
                        53.825

     X8        - -       0.969       - -         - -
                        (0.016)
                        61.703

     X9        - -         - -       1.000       - -

    X10        - -         - -       0.910       - -
                                    (0.022)
                                    42.073

    X11        - -         - -       0.926       - -
                                    (0.020)
                                    47.469

    X12        - -         - -       0.950       - -
                                    (0.016)
                                    59.771

   X1X5        - -         - -         - -       1.000

   X2X6        - -         - -         - -       0.860
                                              (0.022)
                                              38.370

   X3X7        - -         - -         - -       0.822
                                              (0.018)
                                              45.501

   X4X8        - -         - -         - -       0.938
                                              (0.018)
                                              52.652


    GAMMA

              XI1         XI2         XI3      XI1xXI2
          --------    --------    --------    --------
   ETA1      0.272      -0.385       0.103       0.370
            (0.039)     (0.039)     (0.038)     (0.041)
             6.943      -9.776       2.728       8.962


    Covariance Matrix of ETA and KSI

              ETA1        XI1         XI2         XI3      XI1xXI2
          --------    --------    --------    --------    --------
   ETA1      1.000
    XI1      0.296       1.035
```

Fig. 11.51 (continued)

where "1.3032" corresponds to Eq. (11.52), in our case $1.3032 = 1.03 \times 1.0179 + (-0.504738)^2$. The results of these alternative commands are not shown since they are essentially identical to those reported in Fig. 11.54.

```
          XI2      -0.405      -0.052       1.019
          XI3       0.094      -0.048       0.017       1.096
       XI1xXI2      0.346       - -         - -         - -         0.936

       Mean Vector of Eta-Variables

              ETA1
          --------
             0.000

       PHI

                XI1         XI2         XI3     XI1xXI2
          --------    --------    --------    --------
          XI1       1.035
                   (0.073)
                    14.084

          XI2      -0.052       1.019
                   (0.035)     (0.072)
                    -1.464      14.079

          XI3      -0.048       0.017       1.096
                   (0.053)     (0.053)     (0.078)
                    -0.894       0.327      14.091

       XI1xXI2       - -         - -         - -         0.936
                                                        (0.067)
                                                         13.893


       PSI

              ETA1
          --------
             0.626
            (0.044)
             14.084


       Squared Multiple Correlations for Structural Equations

              ETA1
          --------
             0.374

       Squared Multiple Correlations for Y - Variables

                 Y
          --------
             1.000

       THETA-DELTA

                X1          X2          X3          X4          X5          X6
          --------    --------    --------    --------    --------    --------
             0.002       0.182       0.144       0.110       0.002       0.209
            (0.005)     (0.014)     (0.011)     (0.009)     (0.005)     (0.015)
             0.490      13.386      13.220      12.257       0.482      13.551


       THETA-DELTA

                X7          X8          X9         X10         X11         X12
          --------    --------    --------    --------    --------    --------
             0.145       0.120       0.000       0.197       0.159       0.102
            (0.011)     (0.010)     (0.005)     (0.015)     (0.012)     (0.008)
             13.018      12.342       0.092      13.563      13.237      12.084


       THETA-DELTA

              X1X5        X2X6        X3X7        X4X8
```

Fig. 11.51 (continued)

```
        --------    --------    --------    --------
           0.008       0.411       0.225       0.222
         (0.009)     (0.030)     (0.017)     (0.018)
           0.804      13.644      13.028      12.407

…
    ALPHA

           ETA1
        --------
           0.019
         (0.040)
           0.485


    KAPPA

            XI1         XI2         XI3      XI1xXI2
        --------    --------    --------    --------
            - -         - -         - -      -0.052
                                            (0.035)
                                            -1.464

                      Goodness of Fit Statistics

                        Degrees of Freedom = 132
          Minimum Fit Function Chi-Square = 141.303 (P = 0.274)
Normal Theory Weighted Least Squares Chi-Square = 137.203 (P = 0.361)
             Estimated Non-centrality Parameter (NCP) = 5.203
         90 Percent Confidence Interval for NCP = (0.0 ; 37.233)

                    Minimum Fit Function Value = 0.354
          Population Discrepancy Function Value (F0) = 0.0130
         90 Percent Confidence Interval for F0 = (0.0 ; 0.0933)
        Root Mean Square Error of Approximation (RMSEA) = 0.00994
        90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0266)
           P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00

              Expected Cross-Validation Index (ECVI) = 0.554
        90 Percent Confidence Interval for ECVI = (0.489 ; 0.582)
                  ECVI for Saturated Model = 0.767
                ECVI for Independence Model = 19.579

Chi-Square for Independence Model with 136 Degrees of Freedom = 7778.189
                      Independence AIC = 7812.189
                         Model AIC = 221.203
                       Saturated AIC = 306.000
                      Independence CAIC = 7897.044
                          Model CAIC = 430.845
                        Saturated CAIC = 1069.694

                     Normed Fit Index (NFI) = 0.982
                  Non-Normed Fit Index (NNFI) = 0.999
               Parsimony Normed Fit Index (PNFI) = 0.953
                  Comparative Fit Index (CFI) = 0.999
                  Incremental Fit Index (IFI) = 0.999
                    Relative Fit Index (RFI) = 0.981

                       Critical N (CN) = 488.686


                Root Mean Square Residual (RMR) = 0.0443
                     Standardized RMR = 0.0429
                   Goodness of Fit Index (GFI) = 0.965
              Adjusted Goodness of Fit Index (AGFI) = 0.960
              Parsimony Goodness of Fit Index (PGFI) = 0.833
```

Fig. 11.51 (continued)

The structural coefficients are highlighted in grey in Fig. 11.54 with values of 0.33, −0.47, 0.12, and 0.45, respectively, for the impact of XI1, XI2, XI3, and the interaction XI1xXI2.

Yet another solution would be to compute the factor scores corresponding to the confirmatory factor analysis. As mentioned earlier, when the factor loadings are fixed, the problem is equivalent to estimating the structural parameters where the endogenous and exogenous variables are derived from the factor loadings. As shown in Chap. 3, it is possible to go from factor loadings to the linear combination (weighted composite scale) of the observed variables. Therefore, by using the factor

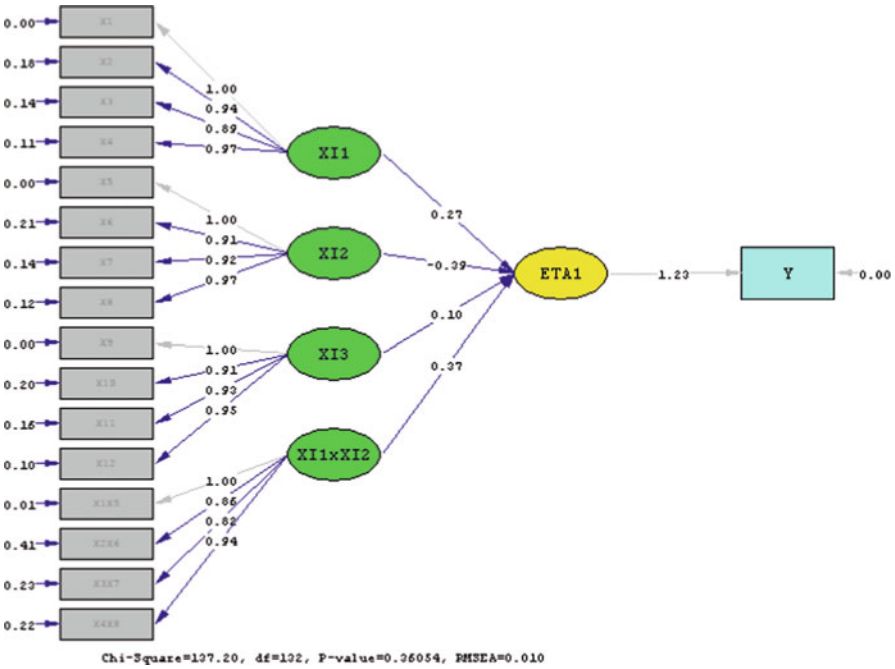


Fig. 11.52 LISREL output graph for model of interactions between latent variables—Partially constrained

```
*SEM model
*SEM with interactions-Restricted Lambdas and Mean of Latent Interaction.do
infile y x1-x12 x1x5 x2x6 x3x7 x4x8 using
"/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/MODERATION/SimData8_Iteractions-Augmented_NOVarName.txt", clear
sem (ETA1@1 _cons@0-> y) ///
(XI1@1 _cons@0-> x1) ///
(XI1@.944 _cons@0-> x2) ///
(XI1@.884 _cons@0-> x3) ///
(XI1@.956 _cons@0-> x4) ///
(XI2@1 _cons@0 -> x5) ///
(XI2@.913 _cons@0 -> x6) ///
(XI2@.914 _cons@0 -> x7) ///
(XI2@.954 _cons@0 -> x8) ///
(XI3@1 _cons@0 -> x9) ///
(XI3@.91 _cons@0 -> x10) ///
(XI3@.926 _cons@0 -> x11) ///
(XI3@.95 _cons@0 -> x12) ///
(XI1xXI2@1 _cons@0 -> x1x5) ///
(XI1xXI2@.8619 _cons@0 -> x2x6) ///
(XI1xXI2@.808 _cons@0 -> x3x7) ///
(XI1xXI2@.912 _cons@0 -> x4x8) ///
(ETA1 <- _cons (XI1, init(.3)) (XI2, init(-.5)) (XI3, init(.1)) (XI1xXI2, init(.4)))
///
, var(e.y@0) means(XI1@0 XI2@0 XI3@0 XI1xXI2@-0.0504738)
estat gof
estat framework
```

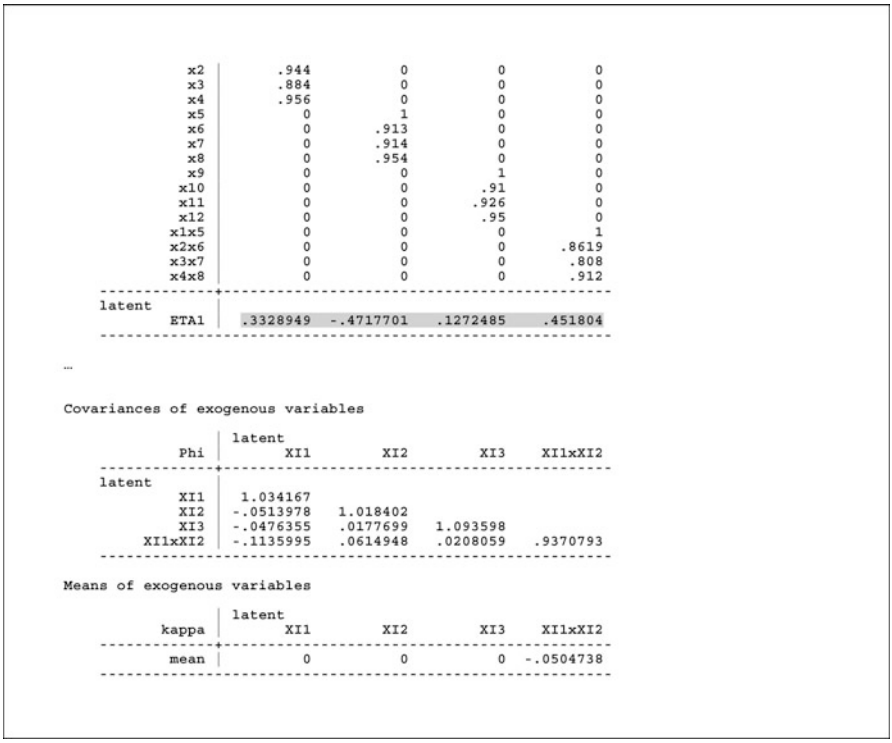
Fig. 11.53 STATA input for model of interactions between latent variables—Constrained (Examp11-11.do)

```
. *SEM model
. *SEM with interactions-Restricted Lambdas and Mean of Latent Interaction.do
…
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  ETA1 <-    |
         XI1 |   .3328949   .0484233     6.87   0.000      .237987    .4278028
         XI2 |  -.4717701   .0485338    -9.72   0.000     -.5668947   -.3766455
         XI3 |   .1272485   .0465688     2.73   0.006      .0359753    .2185217
      XI1xXI2 |    .451804   .0509705     8.86   0.000      .3519036    .5517045
        _cons |   .0189885    .048768     0.39   0.697     -.0765952    .1145721
-------------+----------------------------------------------------------------
Measurement  |
  y <-       |
         ETA1 |           1  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x1 <-      |
         XI1 |           1  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x2 <-      |
         XI1 |        .944  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x3 <-      |
         XI1 |        .884  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x4 <-      |
         XI1 |        .956  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x5 <-      |
         XI2 |           1  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x6 <-      |
         XI2 |        .913  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x7 <-      |
         XI2 |        .914  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x8 <-      |
         XI2 |        .954  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x9 <-      |
         XI3 |           1  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x10 <-     |
         XI3 |         .91  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x11 <-     |
         XI3 |        .926  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x12 <-     |
         XI3 |         .95  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x1x5 <-    |
      XI1xXI2 |           1  (constrained)
        _cons |           0  (constrained)
  -----------+----------------------------------------------------------------
  x2x6 <-    |
```

Fig. 11.54 STATA output for model of interactions between latent variables—Constrained (Examp11-11.log)

```
      XI1xXI2 |      .8619  (constrained)
        _cons |         0  (constrained)
  ------------+------------------------------------------------------------
  x3x7 <-     |
      XI1xXI2 |        .808  (constrained)
        _cons |         0  (constrained)
  ------------+------------------------------------------------------------
  x4x8 <-     |
      XI1xXI2 |        .912  (constrained)
        _cons |         0  (constrained)
--------------+------------------------------------------------------------
Mean          |
      XI1xXI2 |  -.0504738  (constrained)
--------------+------------------------------------------------------------
Variance      |
           e.y |         0  (constrained)
          e.x1 |  .0009124   .0044234                      6.82e-08    12.21438
          e.x2 |  .1830927   .0134882                      .1584762     .211533
          e.x3 |  .1449931   .0107996                      .1252987     .167783
          e.x4 |  .1107293   .0088866                      .0946127    .1295913
          e.x5 |  .0009481   .0046693                      6.09e-08    14.76247
          e.x6 |  .2089392   .0151868                      .1811965    .2409294
          e.x7 |  .1456053    .010988                      .1255862    .1688155
          e.x8 |  .1206175   .0096019                      .1031927    .1409846
          e.x9 |   .000483   .0045014                      5.64e-12    41362.64
         e.x10 |  .1965459   .0143044                      .1704175    .2266802
         e.x11 |  .1583804   .0120239                      .1364836    .1837903
         e.x12 |  .1020704   .0082362                      .0871394    .1195597
       e.x1x5  |  .0032845   .0089323                      .0000159    .6781397
       e.x2x6  |  .4134866   .0303629                      .3580605    .4774923
       e.x3x7  |  .2278608    .017267                      .1964113     .264346
       e.x4x8  |   .226971    .017637                      .1949067    .2643102
       e.ETA1  |  .9441545   .0670024                      .8215559    1.085048
          XI1  | 1.034167   .0733194                      .9000008    1.188333
          XI2  | 1.018402   .0722193                      .8862508    1.170258
          XI3  | 1.093598   .0774902                       .951795    1.256529
      XI1xXI2  |  .9370793   .0666302                       .815178     1.07721
---------------+------------------------------------------------------------
Covariance    |
  XI1         |
          XI2 |  -.0513978   .0514248    -1.00   0.318    -.1521885     .049393
          XI3 |  -.0476355   .0532773    -0.89   0.371    -.1520571     .056786
      XI1xXI2 |  -.1135995   .0496645    -2.29   0.022    -.2109401   -.0162588
  ------------+------------------------------------------------------------
  XI2         |
          XI3 |   .0177699   .0528352     0.34   0.737    -.0857853    .1213251
      XI1xXI2 |   .0614948   .0490492     1.25   0.210      -.03464    .1576295
  ------------+------------------------------------------------------------
  XI3         |
      XI1xXI2 |   .0208059   .0507309     0.41   0.682    -.0786249    .1202367
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(138) =    137.59, Prob > chi2 = 0.4938

. estat gof

------------------------------------------------------------------------------
Fit statistic         |      Value   Description
----------------------+-------------------------------------------------------
Likelihood ratio      |
        chi2_ms(138)  |    137.592   model vs. saturated
            p > chi2  |      0.494

        chi2_bs(136)  |   8898.464   baseline vs. saturated
            p > chi2  |      0.000
------------------------------------------------------------------------------

. estat framework

Exogenous variables on endogenous variables

                 | latent
           Gamma |       XI1         XI2         XI3     XI1xXI2
    -------------+--------------------------------------------
    observed     |
               y |         0           0           0           0
              x1 |         1           0           0           0
```

Fig. 11.54 (continued)



```
              x2 |       .944          0          0          0
              x3 |       .884          0          0          0
              x4 |       .956          0          0          0
              x5 |         0           1          0          0
              x6 |         0         .913         0          0
              x7 |         0         .914         0          0
              x8 |         0         .954         0          0
              x9 |         0           0          1          0
             x10 |         0           0        .91          0
             x11 |         0           0       .926          0
             x12 |         0           0        .95          0
            x1x5 |         0           0          0          1
            x2x6 |         0           0          0       .8619
            x3x7 |         0           0          0        .808
            x4x8 |         0           0          0        .912
-------------+------------------------------------------------
  latent     |
        ETA1 |  .3328949  -.4717701   .1272485    .451804
-------------------------------------------------------------

…

Covariances of exogenous variables

             |  latent
         Phi |       XI1        XI2        XI3     XI1xXI2
-------------+------------------------------------------------
  latent     |
         XI1 |   1.034167
         XI2 |  -.0513978   1.018402
         XI3 |  -.0476355   .0177699   1.093598
      XI1xXI2 |  -.1135995   .0614948   .0208059   .9370793
-------------------------------------------------------------

Means of exogenous variables

             |  latent
       kappa |       XI1        XI2        XI3     XI1xXI2
-------------+------------------------------------------------
        mean |         0          0          0  -.0504738
-------------------------------------------------------------
```

Fig. 11.54 (continued)

```
use "/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/MODERATION/SimData8_Iteractions-Augmented_NOVarName_WithFactorScores.dta",
clear
generate inter = xi1*xi2
regress eta1 xi1 xi2 xi3 inter
```

Fig. 11.55 STATA input for model of interactions using factor scores (Examp11-12.do)

scores from a confirmatory factor analysis, least square methods can be used to estimate the structural parameters. Depending on the nature of the model, ordinary least squares, seemingly related regression, and two- or three-stage least squares can be used. The conditions for which estimation is appropriate are presented in Chap. 6. A simple example is shown with the STATA input file in Fig. 11.55.

We first generate the interaction term of the factor scores corresponding to the moderation effect. Then, in this example, we use simple multiple regression (regress) because there is a single endogenous variable. When there are multiple endogenous variables, seemingly unrelated regression ("sureg" in STATA) or three-stage least squares ("reg3" in STATA) is used instead to take into account the contemporaneous correlations of the error terms of each equation. The results are shown in Fig. 11.56.

```
. use "/Users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter11_Mediation-
Moderation/MODERATION/SimData8_Iteractions-Augmented_NOVarName_WithFactorScores.dta",
clear

. generate inter = xi1*xi2

. regress eta1 xi1 xi2 xi3 inter

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  4,   395) =   52.42
       Model |  200.544029     4  50.1360071            Prob > F      =  0.0000
    Residual |  377.768637   395  .956376297           R-squared     =  0.3468
-------------+------------------------------           Adj R-squared =  0.3402
       Total |  578.312666   399  1.44940518           Root MSE      =  .97794

------------------------------------------------------------------------------
        eta1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         xi1 |   .3330121    .048613     6.85   0.000     .2374395    .4285847
         xi2 |  -.4722278   .0486366    -9.71   0.000    -.5678468   -.3766087
         xi3 |   .1273392   .0468315     2.72   0.007     .0352691    .2194094
       inter |     .45116   .0509888     8.85   0.000     .3509167    .5514033
       _cons |   .0188738    .049053     0.38   0.701    -.0775637    .1153113
------------------------------------------------------------------------------
```

Fig. 11.56 STATA output for model of interactions using factor scores (Examp11-12.log)

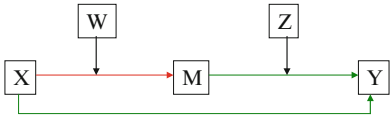


Fig. 11.57 Formalization of model with one mediator and two moderators

As can be seen, the results are almost identical to those obtained with the structural equation model estimations. The estimates are exactly the same with two digits after the decimal point for the effects of xi1, xi2, xi3, and the interaction of, respectively, 0.33, −0.47, 0.12, and 0.45. If mediating variables are involved, it is then possible to use the bootstrapping procedure described earlier combined with the seemingly unrelated regression estimation.

11.4 Testing Moderated Mediation Effects

We do not discuss here the case of mediated moderation because, as discussed in Sect. 11.1.3, this leads to a system of recursive multiple equations and we refer the researcher to Chap. 6 to simultaneously estimate all the relevant paths. Here we consider models of moderated mediation effects. The two cases schematized in Fig. 11.3 are more formally represented in Fig. 11.57.

In Fig. 11.57 the independent variable is noted by X, the dependent variable by Y, and the mediating variable by M. The two moderating variables are noted by W and Z, respectively, for the moderation of the mediating equation (from X to M) and for the dependent variable equation (from M to Y). Note that the variables W and Z could represent the same variable, but here we use different symbols to provide the more general case where they correspond to different conditions. As indicated earlier in this chapter, W illustrates a case of stage-one moderated

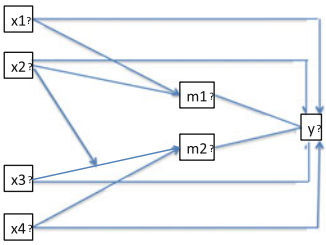


Fig. 11.58 Example of stage-one moderated mediation

mediation and Z illustrates a case of stage-two moderated mediation. These two effects are represented algebraically in Eqs. (11.60) and (11.61):

$$m_i = a_0 + (a_1 + a_2 w_i)x_i + u_i^m \tag{11.60}$$

$$y_i = b_0 + (b_1 + b_2 z_i)m_i + u_i^y \tag{11.61}$$

Similar to the derivations in Sect. 11.2.4.1 with nonlinear effects, the marginal indirect effect theta is the partial derivative that makes use of the chain rule so that

$$f(\theta|w,z) = (a_1 + a_2 w)(b_1 + b_2 z) \tag{11.62}$$

θ is a function of the two moderators. Therefore, the indirect effect is not fixed but depends on the values taken by these moderator variables. Consequently, just as in the case of nonlinear effects, the researcher must evaluate the indirect effects over a range of values of the moderators.

Therefore, conceptually, these models of moderated mediation do not require any new analytical treatment. The same problems as identified above for testing mediation arise here as well. The bootstrapping method provides once more a solution to the estimation of the product terms shown in Eq. (11.62). Preacher, Rucker, and Hayes (2007) provide an algorithm in SPSS to provide a distribution of this product term for different levels of the moderators. In addition to providing the confidence intervals for the standard $\pm\ 1\sigma_z$, it is possible to request the estimation at a particular value of the moderators from the minimum to the maximum values in the data. The procedure proposed earlier in this chapter in STATA (Fig. 11.10) can easily be adapted to the system of equations illustrated in Fig. 11.57: the use of the procedure "sureg" (seemingly unrelated regression) takes into account the correlations of error terms across equations and can include all the relevant product terms implied by the moderated effects, as discussed in the previous section of this chapter. Such a procedure is illustrated by the example in Fig. 11.58, which represents a more complex model with two mediator variables where variable x2 acts as a moderator of the effect of x3 on the second mediator (m2).

The STATA subroutine for performing a bootstrap estimation is shown in Fig. 11.59.

In this example, the variables are all mean centered, so constant terms are not estimated. The interaction term x3x2 has been generated as a variable introduced in

```
capture drop program bootcmmod1stdv
program bootcmmod1stdv, rclass
sureg (m1 x1 x2, noconstant) ///
      (m2 x3 x4 x3x2, noconstant) ///
      (y m1 m2 x1 x2 x3 x4, noconstant)
return scalar indirect1 = ([m2]_b[x3]-([m2]_b[x3x2]*1))*[y]_b[m2]
return scalar direct1 = [y]_b[x3]
return scalar indirect2 = ([m2]_b[x3]+([m2]_b[x3x2]*1))*[y]_b[m2]
return scalar direct2 = [y]_b[x2]
end
```

Fig. 11.59 STATA example of bootstrap subroutine

```
…
sureg (m1 x1 x2, noconstant) ///
      (m2 x3 x4 x3x2, noconstant) ///
      (y m1 m2 x1 x2 x3 x4, noconstant)
bootstrap r(indirect1) r(indirect2), reps(5000) nodots: bootcmmod1stdv
estat boot, bc percentile
```

Fig. 11.60 STATA example of bootstrap estimation of indirect effects in moderated mediation

the data set. Here, indirect1 and indirect2 are the estimates of the indirect effect at minus one standard deviation (1 in this case because the moderator variable has been standardized) and plus one standard deviation from the mean of the moderator variable x2. The value 1 can easily be replaced by any other value, for example at the extreme values of the range of the moderator variable.

Figure 11.60 gives the STATA code for doing the estimation on a specific data set.

The STATA output is identical to the one shown earlier in this chapter with the mean and confidence intervals of these indirect effects.

Just as we discussed the possibility of estimating moderated effects with structural equation models that take into account measurement errors, the same methods can be applied in the case of moderated mediation models. If the moderator variable is a nominal variable, multiple-group analysis can be performed to test the equality of the structural coefficients across groups (i.e., levels of the moderator variable). Note that the practice of dichotomizing a continuous moderator variable for group analysis is indeed one way to deal with moderated mediation models but note also that this reflects a loss of information and is not generally recommended. As discussed in Chap. 5, tests of equality of coefficients or pooling tests allow us to first perform a joint test of the overall equality of structural parameters. If the joint test fails, the null hypothesis of no moderated effect cannot be rejected. If significant, further tests can be performed about the source of the moderation to discriminate between first-stage, second-stage, and direct-moderated mediation if these possibilities all have theoretical meaning.

When the moderator variable is continuous, the models of moderated mediation are no different from any structural equation models with interactions. Therefore, the estimation methods presented in the section on testing moderation effects are perfectly appropriate and do not require any particular adaptation.

11.5 Stating Mediation and Moderation Effect Hypotheses

In this section, we raise several issues that appear when developing and presenting hypotheses involving mediation and moderation.

11.5.1 Stating Hypotheses About Mediation

We have discussed the issue raised by the fact that a mediation process could be significant, even if the direct relationship between an independent variable and a dependent variable may not be significant. This was due to the possibility of missing factors, especially missing mediation explanations that work in parallel with the central explanation being tested. In spite of this, the theory development process rarely starts with an explanatory mechanism in search of relationships between variables to be explained. More commonly, it is the observation of a phenomenon that occurs regularly that raises the question of why this phenomenon may be occurring. At times, the relationship between x and y has been established and the research is only about testing the explanation. It does occur, however, that both the effect of x on y and the explanatory mechanism are to be demonstrated together. In such cases, it is more effective to hypothesize an effect of x on y before an explanatory mechanism is advanced.

Another critical issue when considering mediation hypotheses concerns the discriminant validity of the variables under investigation. M should not be just an operational measurement of X or the manipulation levels of X.

11.5.2 Stating Hypotheses About Moderation

It may appear logical to first state a general hypothesis before introducing moderating effects for that effect. However, this may not always make sense.

If an effect (in a particular direction) dominates over the range of a moderator variable, then there is no issue in stating such a "main effect" hypothesis. However, if the sign of the effect depends on the level of the moderator, stating a general positive (or negative) effect first does not make sense, since it is contradicted by the moderating hypothesis.

Consequently, if a moderating process is hypothesized, it appears more logical to start with the development of the complete theory, including the moderation effects, and state the moderating hypothesis first. The effects for particular levels of the moderator or throughout ranges of values may then be described as additional hypotheses.

Regardless, the test for the existence of a moderation effect should always come first. The reason for this precedence is that estimates obtained from a model that would ignore moderator effects when they would be significant are biased (since the interactions are generally correlated with the model's components).

```
/*     AssignMediation.sas     */
filename survey 'd:\WORK_SAS\SASMVS\survey.asc';
data new;
infile survey firstobs=19;
input   (Age Marital Income Educatn HHSize Occuptn Location
         TryHair LatStyle DrssSmrt BlndsFun LookDif
         LookAttr GrocShp LikeBkng ClthFrsh WashHnds Sportng LikeClrs
         FeelAttr TooMchSx Social LikeMaid ServDnrs SaveRcps LikeKtch) (3.)
        #2 (LoveEat SpirtVal Mother ClascMsc Children Applianc ClsFamly
         LovFamly TalkChld Exercise LikeSelf CareSkin MedChckp
         EvngHome TripWrld HomeBody LondnPrs Comfort Ballet Parties
         WmnNtSmk BrghtFun Seasonng ColorTV SlppyPpl Smoke) (3.)
        #3 (Gasoline Headache Whiskey Bourbon FastFood Restrnts OutFrDnr
         OutFrLnc RentVide Catsup KnowSont PercvDif BrndLylt
         CatgMotv BrndMotv OwnSonit NecssSon OthrInfl DecsnTim
         RdWomen RdHomSrv RdFashn RdMenMag RdBusMag RdNewsMg
         RdGlMag) (3.)
        #4 (RdYouthM RdNwsppr WtchDay WtchEve WtchPrm
         WTchLate WtchWknd WtchCsby WtchFmTs WtchChrs WtchMoon
         WtchBoss WtchGrwP WtchMiaV WtchDns WtchGold WtchBowl) (3.);
idobs=_n_;

Proc reg data=new ;
model LikeSelf = HHSize Age;
model ClsFamly = HHSize Age;
model LovFamly = HHSize Age;
model TalkChld = HHSize Age;
model Exercise = HHSize Age;
model Restrnts = LikeSelf ClsFamly LovFamly TalkChld Exercise HHSize Age;
run;
data new2;
set new;
%indirect(data=new2, y=Restrnts, x=HHSize, m=LikeSelf ClsFamly LovFamly TalkChld
Exercise Age, c=1, boot=5000);
run;
```

Fig. 11.61 Assignment example for investigating mediation and moderation

11.6 Assignment

Using either the data from the SURVEY or from the INDUP and PANEL files (described in Chap. 14), develop and test hypotheses where a mediation process is involved as well as conditional hypotheses that you can investigate with tests of moderation. The variables corresponding to the various files are described in Appendix C (Chap. 14). Figure 5.4 in Chap. 5 shows how to read and merge the data that are in the INDUP and PANEL files. Figure 11.61 provides an example of an input file in SAS for a mediation test using the data in the SURVEY file.

Bibliography

Basic Technical Readings

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology, 20*, 115–140.

Belsley, David A. (1984). Demeaning Conditioning Diagnostics Through Centering. *The American Statistician, 38*(2), 73–77.

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin, 102*, 414–417.

Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science, 26*(3), 438–445.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods, 12*(1), 1–22.

Hayes, A. F. (2008). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420.

Hayes, A. F., & Preacher, K. J. (2010). Quantifying and testing indirect effects in simple mediation models when constituent paths are nonlinear. *Multivariate Behavioral Research, 45*(4), 627–660.

Iacobucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regressions. *Journal of Consumer Psychology, 17*(2), 139–153.

Jöreskog K. G. (2000). Latent variable scores and their uses. Working Paper, July 10.

Jöreskog, K. G., & Yang, F. (1996). *Nonlinear structural equation models: The Kenny-Judd model with interaction effects* (Advanced structural equation modeling, pp. 57–87). Mahwah, NJ: Lawrence Erlbaum Associates.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*, 602–619.

Kelava, A., Moosbrugger, H., Dimitruk, P., & Schermelleh-Engel, K. (2008). Multicollinearity and missing constraints: A comparison of three approaches for the analysis of latent nonlinear effects. *Methodology, 4*, 51–66.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1). New York: McGraw-Hill.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96*(1), 201–210.

MacKinnon, D. P., Lockwood, C. M., & Williams J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*(1), 99–128.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614.

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*(3), 275–300.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*(2), 376–390.

Muller, D., Juddand, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*(6), 852–863.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717–731.

Preacher, K. J., & Hayes A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891.

Preacher, Kristopher J., Derek D. Rucker, & Andrew F. Hayes (2007). Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions. *Multivariate Behavioral Research, 42*(1), 185–227.

Schoonhoven, C. B. (1981). Problems with contingency theory: Testing assumptions hidden within the language of contingency "theory". *Administrative Science Quarterly, 26*, 349–377.
Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*(4), 422–445.
Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–312). Washington, DC: American Sociological Association.
Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. In N. Tuma (Ed.), *Sociological methodology*. Washington, DC: American Sociological Association.
Taylor, A. B., MacKinnon, D. P., & Tein, J.-Y. (2008). Tests of the three-path mediated effect. *Organizational Research Methods, 11*(2), 241–269.
Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37*(3), 197–206.

Application Readings

Atuahene-Gima, K., Slater, S. F., & Olson, E. M. (2005). The contingent value of responsive and proactive market orientations for new product program performance. *Journal of Product Innovation Management, 22*(6), 464–482.
Avnet, T., & Higgins, E. T. (2006). How regulatory fit affects value in consumer choices and opinions. *Journal of Marketing Research, 43*(1), 1–10.
Birnbaum, M. H., & Mellers, B. A. (1979). Stimulus recognition may mediate exposure effects. *Journal of Personality and Social Psychology, 37*(3), 391–394.
Bolton, L. E., Cohen, J. B., & Bloom, P. N. (2006 June). Does marketing products as remedies create 'get out of jail free cards'? *Journal of Consumer Research, 33*, 71–81.
Brown, S. P., Jones, E., & Leigh, T. W. (2005). The attenuating effect of role overload on relationships linking self-efficacy and goal level to work performance. *Journal of Applied Psychology, 90*(5), 972–979.
Chandon, P., Wesley Hutchinson, J., Bradlow, E. T., & Young, S. H. (2009). Does in-store marketing work? Effects of the number and position of shelf facings on brand attention and evaluation at the point of purchase. *Journal of Marketing, 73*(4), 1–17.
Chandukala, S. R., Dotson, J. P., Brazell, J. D., & Allenby, G. M. (2011). Bayesian analysis of hierarchical effects. *Marketing Science, 30*(1), 123–133.
Deng, X., & Kahn, B. E. (2009). Is your product on the right side? The 'location effect' on perceived product heaviness and package evaluation. *Journal of Marketing Research, 46*(6), 725–738.
Dholakia, U. M. (2006). How customer self-determination influences relational marketing outcomes: Evidence from longitudinal field studies. *Journal of Marketing Research, 43*(1), 109–120.
Elliott, M., & Armitage, C. J. (2006). Effects of implementation intentions on the self-reported frequency of drivers' compliance with speed limits. *Journal of Experimental Psychology. Applied, 12*(2), 108–117.
Franke, N., Schreier, M., & Kaiser, U. (2009). The "I designed it myself" effect in mass customization. *Management Science, 56*(1), 125–140.
Galunic, D. C., & Anderson, E. (2000). From security to mobility: Generalized investments in human capital and agent commitment. *Organization Science, 11*(1), 1–20.
Ganesan, S., Malter, A. J., & Rindfleisch, A. (2005). Does distance still matter? Geographic proximity and new product development. *Journal of Marketing, 69*(4), 44–60.

Gardner, H. K., Gino, F., & Staats, B. R. (2012). Dynamically integrating knowledge in teams: Transforming resources into performance. *Academy of Management Journal, 55*(4), 998–1023.

Gatignon, H., & Xuereb, J.-M. (1997). Strategic orientation of the firm and new product performance. *Journal of Marketing Research, 34*(1), 77–90.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters, 12*, 211–223.

Gosserand, R. H., & Diefendorff, L. M. (2005). Emotional display rules and emotional labor: The moderating role of commitment. *The Journal of Applied Psychology, 90*(6), 1256–1264.

Gotteland, D., & Boule, J. (2006). The market orientation—new product performance relationship: Redefining the moderating role of environmental conditions. *International Journal of Research in Marketing, 23*(2), 171–185.

Grandey, A. A., Fisk, G. M., & Steiner, D. D. (2005). Must "service with a smile" be stressful? The moderating role of personal control for American and French employees. *The Journal of Applied Psychology, 90*(5), 893–904.

Han, J. K., Kim, N., & Srivastava, R. K. (1998). Market orientation and organizational performance: Is innovation a missing link? *Journal of Marketing, 62*(4), 30–45.

Haon, C., Gotteland, D., & Fornerino, M. (2009). Familiarity and competence diversity in new product development teams: Effects on new product performance. *Marketing Letters, 20*(1), 75–89.

Harmancioglu, N., Grinstein, A., & Goldman, A. (2010). Innovation and performance outcomes of market information collection efforts: The role of top management team involvement. *International Journal of Research in Marketing, 27*(1), 33–43.

Hennig-Thurau, T., Henning, V., & Sattler, H. (2007). Consumer file sharing of motion pictures. *Journal of Marketing, 71*(4), 1–18.

Ho, J. Y. C., Dhar, T., & Weinberg, C. B. (2009). Playoff payoff: Super bowl advertising for movies. *International Journal of Research in Marketing, 26*(3), 168–179.

Hofmann, D. A., & Jones, L. M. (2005). Leadership, collective personality, and performance. *The Journal of Applied Psychology, 90*(3), 509–522.

Hull, C. E., & Rothenberg, S. (2008). Research notes and commentaries firm performance: The interactions of corporate social performance with innovation and industry differentiation. *Strategic Management Journal, 29*, 781–789.

Hurley, R. F., & Hult, G. T. M. (1998 July). Innovation, market orientation, and organizational learning: An integration and empirical examination. *Journal of Marketing, 62*, 42–54.

Karim, S. (2009). Business unit reorganization and innovation in new product markets. *Management Science, 55*(7), 1237–1254.

Keller, R. T. (2001). Cross-functional project groups in research and new product development: Diversity, communications, job stress, and outcomes. *The Academy of Management Journal, 44*(3), 547–555.

Kirca, A. H., Jayachandran, S., & Bearden, W. O. (2005). Market orientation: A meta-analytic review and assessment of its antecedents and impact on performance. *Journal of Marketing, 69* (2), 24–41.

Kirkman, B. L., Chen, G., Farh, J.-L., Chen, Z. X., & Lowe, K. B. (2009). Individual power distance orientation and follower reactions to transformational leaders: A cross-level, cross-cultural examination. *Academy of Management Journal, 52*(4), 744–764.

Kramer, T., Spolter-Weisfeld, S., & Thakkar, M. (2007). The effect of cultural orientation on consumer responses to personalization. *Marketing Science, 26*(2), 246–258.

Kumar, N., Scheer, L. K., & Steenkamp, J.-B. E. M. (1998). Interdependence, punitive capability, and the reciprocation of punitive actions in channel relationships. *Journal of Marketing Research, 35*(2), 225–235.

Kyriakopoulos, K., & Moorman, C. (2004). Tradeoffs in marketing exploitation and exploration strategies: The overlooked role of market orientation. *International Journal of Research in Marketing, 21*(3), 219–240.

Langerak, F., Rijsdijk, S., & Dittrich, K. (2009). Development time and new product sales: A contingency analysis of product innovativeness and price. *Marketing Letters, 20*(4), 399–413.

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology, 101*(6), 1174–1188.

Luo, X., Slotegraaf, R. J., & Pan, X. (2006). Cross-functional "coopetition": The simultaneous role of cooperation and competition within firms. *Journal of Marketing, 70*(2), 67–80.

Martin-Tapia, I., Aragon-Correa, J. A., & Senise-Barrio, M. E. (2008). Being green and export intensity of SMEs: The moderating influence of perceived uncertainty. *Ecological Economics, 68*(1–2), 56–67.

Menguc, B., & Auh, S. (2008). Conflict, leadership, and market orientation. *International Journal of Research in Marketing, 25*(1), 34–45.

Mitra, A., & Lynch, J. G., Jr. (1995 March). Toward a reconciliation of market power and information theories of advertising effects on price elasticity. *Journal of Consumer Research, 21*, 644–659.

Moorman, C. (1995). Organizational market information processes: Cultural antecedents and new product outcomes. *Journal of Marketing Research, 32*(3), 318–335.

Morgan, R. M., & Hunt, S. D. (1994 July). The commitment-trust theory of relationship marketing. *Journal of Marketing, 58*, 20–38.

Noble, C. H., Sinha, R. K., & Kumar, A. (2002). Market orientation and alternative strategic orientations: A longitudinal assessment of performance implications. *Journal of Marketing, 66* (4), 25–39.

Palmatier, R. W., Dant, R. P., Grewal, D., & Evans, K. R. (2006 October). Factors influencing the effectiveness of relationship marketing: A meta-analysis. *Journal of Marketing, 70*, 136–153.

Palmatier, R. W., Jarvis, C. B., Bechkoff, J. R., & Kardes, F. R. (2009 September). The role of customer gratitude in relationship marketing. *Journal of Marketing, 73*, 1–18.

Porter, C. O. L. H. (2005). Goal orientation: Effects on backing up behavior, performance, efficacy, and commitment in teams. *The Journal of Applied Psychology, 90*(4), 811–818.

Prabhu, J. C., Chandy, R. K., & Ellis, M. E. (2005). The impact of acquisitions on innovation: Poison pill, placebo, or tonic? *Journal of Marketing, 69*(1), 114–130.

Redman, T., & Snape, E. (2005). Exchange ideology and member-union relationships: An evaluation of moderation effects. *The Journal of Applied Psychology, 90*(4), 765–773.

Rueda-Manzanares, A., Aragón-Correa, J. A., & Sharma, S. (2008). The influence of stakeholders on the environmental strategy of service firms: The moderating effects of complexity, uncertainty and munificence. *British Journal of Management, 19*(2), 185–203.

Ryu, E., West, S. G., & Sousa, K. H. (2009). Mediation and moderation: Testing relationships between symptom status, functional health, and quality of life in HIV patients. *Multivariate Behavioral Research, 44*(2), 213–232.

Schilling, M. A., & Phelps, C. C. (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science, 53*(7), 1113–1126.

Shiv, B., Carmon, Z., & Ariely, D. (2005). Placebo effects of marketing actions: Consumers may get what they pay for. *Journal of Marketing Research, 42*(4), 383–393.

Sparrowe, R. T., & Liden, R. C. (2005). Two routes to influence: Integrating leader-member exchange and social network perspectives. *Administrative Science Quarterly, 50*, 505–535.

Stremersch, S., & Van Dyck, W. (2009). Marketing of the life sciences: A new framework and research agenda for a nascent field. *Journal of Marketing, 73*(4), 4–30.

Thompson, D. V., Hamilton, R. W., & Rust, R. T. (2005). Feature fatigue: When product capabilities become too much of a good thing. *Journal of Marketing Research, 42*(4), 431–442.

Zhang, J., Wedel, M., & Pieters, R. (2009). Sales effects of attention to feature advertisements: A Bayesian mediation analysis. *Journal of Marketing Research, 46*(5), 669–681.

Zhao, X. (1997). Clutter and serial order redefined and retested. *Journal of Advertising Research, 37*(5), 57–74.

Waarts, E., & Wierenga, B. (2000). Explaining competitors' reactions to new product introductions: The roles of event characteristics, managerial interpretation, and competitive context. *Marketing Letters, 11*(1), 67–79.

Wang, J., & Lee, A. Y. (2006). The role of regulatory focus in preference construction. *Journal of Marketing Research, 43*(1), 28–38.

West, P. M., & Broniarczyk, S. M. (1998 June). Integrating multiple opinnions: The role of aspiration level on consumer response to critic consensus. *Journal of Consumer Research, 25*, 38–51.

Chapter 12
Cluster Analysis

The objective of cluster analysis is to group observations (e.g., individuals) in such a way that the groups formed are as homogeneous as possible within each group and as different as possible across groups.

These criteria remind us of those for discriminant analysis where the objective is to derive a linear combination of the variables such that this transformed linear combination would exhibit the largest difference between centroids but the smallest variance within groups. However, in discriminant analysis, the groups are known a priori. The purpose of cluster analysis is to form such groups. These groups are called "clusters."

This type of analysis is particularly relevant in market research where market segments are sought out in order to address, from a practical point of view, the heterogeneity of consumers. This technique is generally most useful in any situation where the analyst needs to reduce the heterogeneity of observations by forming a manageable number of groups that should reflect the diversity of these observations.

12.1 The Clustering Methods

The clustering solutions are found by applying an algorithm that determines the rules by which observations are aggregated. A number of algorithms can be found in the literature. They are more or less complicated procedures based on "rules" that lead to reasonable solutions, although these procedures are clearly not grounded in statistical theory, and different algorithms often lead to different solutions. For this reason, it is particularly critical to understand the specific "rules" used in each method and to identify the specific method used in reporting the clustering solution found.

Algorithms can be classified into two groups: hierarchical methods and nonhierarchical methods.

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_12, © Springer Science+Business Media New York 2014

Hierarchical algorithms are the most common methods of cluster analysis. In such algorithms, observations are added to each other one by one in a treelike fashion. Such a tree can be graphed to form a dendrogram showing the aggregation process from the N groups made up of the N individuals to any level of K groups.

In nonhierarchical algorithms, the number of groups K is known (or assumed) a priori, and each observation is assigned one of the K groups according to its distance to the group centroid and keeps being relocated until a stopping rule criterion is verified.

12.1.1 Similarity Measures

Any of these methods requires the proximity of the observations to be measured. These proximity measures can take multiple forms although, given the multivariate description of the observations through P variates, the Euclidean distance or related measures come immediately to mind. The squared distance between objects i and j is therefore

$$d_{ij}^2 = \sum_{p=1}^{P} \left(x_{pi} - x_{pj}\right)^2 = \underset{P\times 1}{\left(\mathbf{x}_i - \mathbf{x}_j\right)'} \left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{12.1}$$

where x_{pi} = value of observation i on variate p.

It is clear from this expression in Eq. (12.1) that the scale of each variable can have a large impact on the distance measure. Therefore, the question of standardization of the variable is a pertinent one. Unfortunately, there is no obvious response to that question. It is therefore important to compare results using Euclidean distances with those using standardized measures of similarity. The standardized Euclidean distance is given by Eq. (12.2):

$$d_{ij}^2 = \left(\mathbf{x}_i - \mathbf{x}_j\right)' \mathbf{D}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{12.2}$$

where the diagonal matrix $\mathbf{D}$ contains the variances of the variables across observations σ_p^2. The more general Mahalanobis' ellipse measure considers the correlations between the variables, represented in the off-diagonal elements of the covariance matrix Σ in Eq. (12.3):

$$d_{ij}^2 = \left(\mathbf{x}_i - \mathbf{x}_j\right)' \Sigma^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{12.3}$$

12.1.2 The Centroid Method

The centroid is the average value of each variate across the observations in a group. The algorithm of the centroid method starts by bringing together into the first group

Table 12.1 Sample data

Individual	Variable 1	Variable 2
1	15	12
2	10	20
3	14	18
4	10	14
5	16	15
6	8	20

Table 12.2 Dissimilarity measures based on Euclidean distances (only the upper half of the symmetric matrix is shown)

Individuals	1	2	3	4	5	6
1	0	89	37	29	10	113
2		0	20	36	61	4
3			0	32	13	40
4				0	37	40
5					0	89
6						0

observations that exhibit the smallest distance from each other. These two observations that are the closest form the first group. In a second step, the centroid formed by this group is computed. The observations that have not yet been assigned to the group are then assessed based on their distances from each other as well as their distance to the centroid of the first group formed. The observations or group corresponding to the smallest distance of all combinations are grouped together. That is, if the smallest distance formed by a pair of observations is not yet part of a group, then a new group is formed. Otherwise, the observation with the smallest distance to the centroid of an existing group joins this group. We then continue the process of step 2 until all observations fall within a single group.

We now illustrate the centroid method using a small sample data set.

Let us consider the data in Table 12.1 where six individuals are characterized by two variables, variable 1 and variable 2.

Step 1 ($s = 1$): Calculate the Euclidean distances between all pairs of observations according to Eq. (12.1). These calculations lead to the matrix of similarities between each of the six individuals as shown in Table 12.2.

Taking the smallest distance indicates that we should group individuals 2 and 6, as they are the closest together with a distance of only 4.

Step 2 ($s = 2$): In this step, we need to first compute the means (centroids) of the variates for the first cluster (2,6), with the value of the variates for the other individuals remaining and the value of the variable of each individual since, at this stage, each individual constitutes its own cluster. This forms the $N - 1$ cluster solution (i.e., five clusters in our example). Then, the new distance matrix can be computed between this first cluster and each of the other individuals.

(i) Compute centroids of $N - 1 = 5$ clusters.
The averages lead to the new table of data shown in Table 12.3. The average value on variable 1 for cluster (2,6) is the average of the values of that variable

Table 12.3 Centroids for 5-cluster solution

Individuals (clusters)	Variable 1	Variable 2
(2,6)	9.0	20.0
1	15.0	12.0
3	14.0	18.0
4	10.0	14.0
5	16.0	15.0

Table 12.4 Five-cluster dissimilarity matrix

Individuals (clusters)	(2,6)	1	3	4	5
(2,6)	0	100.00	29.00	37.00	74.00
1		0	37.00	29.00	10.00
3			0	32.00	13.00
4				0	40.00
5					0

Table 12.5 Centroids for 4-cluster solution

Individuals (clusters)	Variable 1	Variable 2
(2,6)	9.0	20.0
(1,5)	15.5	13.5
3	14.0	18.0
4	10.0	14.0

for individuals 2 and 6, i.e., $(10 + 8)/2 = 9$. The same calculation is made for variable 2, which results in a value of 20.

(ii) Compute Euclidean distances between each group centroid.
The Euclidean distance between each of these five groups is computed using Eq. (12.1) applied to the data in Table 12.3. This results in the new dissimilarity matrix shown in Table 12.4 between the first cluster (2,6) and each of the other individuals.

The smallest distance is between individuals 1 and 5 with a distance of 10.00, leading to grouping individuals 1 and 5 into a new cluster for an $N - 2$ or a 4-cluster solution.

Step 3 ($s = 3$): Step 2 is now repeated with $N - 2$ data points.

(i) Compute centroids of $N - 2 = 4$ clusters.
First we compute the average values of each variate for the two clusters found, with the values of the other individuals remaining unchanged. This gives the new data matrix as shown in Table 12.5. For example, the average value of variate 1 for cluster (1,5) is $(15 + 16)/2 = 15.5$.

(ii) Compute Euclidean distances between each group centroid.
We can then compute the dissimilarity matrix, which results in Table 12.6.

The smallest distance is now between individual 3 and cluster (1,5), leading to a change in one cluster from two individuals (1,5) to three individuals (1,3,5).

Table 12.6 Four-cluster dissimilarity matrix

Individuals (clusters)	(2,6)	(1,5)	3	4
(2,6)	0	84.50	29.00	37.00
(1,5)		0	22.50	30.50
3			0	32.00
4				0

Table 12.7 Centroids for 3-cluster solution

Individuals (clusters)	Variable 1	Variable 2
(2,6)	9.0	20.0
(1,3,5)	15.0	15.0
4	10.0	14.0

Table 12.8 Three-cluster dissimilarity matrix

Individuals (clusters)	(2,6)	(1,3,5)	4
(2,6)	0	61.00	37.00
(1,3,5)		0	26.00
4			0

Therefore, we now have three clusters ($N - 3$) composed of cluster 1 = (2,6), cluster 2 = (1,3,5), and cluster 3 = (4). This is the 3-cluster solution.

Step 4 ($s = 4$): We now perform the same procedure for the $N - 3 = 3$ clusters.

(i) Compute centroids of $N - 3$ clusters. This is done in Table 12.7.
(ii) Compute Euclidean distances between each group. The results of these computations are shown in Table 12.8.

The clusters (1,3,5) and individual 4 are the least dissimilar with a distance of 26.00, which leads to forming a single cluster composed of these four individuals: 1, 3, 4, and 5. This gives us the 2-cluster solution: (2,6) and (1,3,4,5).

This is the last step that occurs when only two clusters remain (value of step s when $N - s + 1 = 2$), since only one way is left for them to be grouped together.

The dendrogram illustrated in Fig. 12.1 summarizes the results of the full process. The individuals are represented on the x-axis without reflecting any scale but simply in the order in which they enter the clustering hierarchy. The y-axis represents the Euclidean distance (on standardized variables in the figure) between each cluster for any solution of ($N - s + 1$) clusters (where s is the step of group formation).

12.1.3 Ward's Method

The criterion used in Ward's algorithm to add observations to a group is the within-clusters sum of squares (Eq. (12.5) gives the formal formula of the sum of squares measure). Therefore, at each step, the within-clusters sum of squares is computed for all possible combinations remaining.

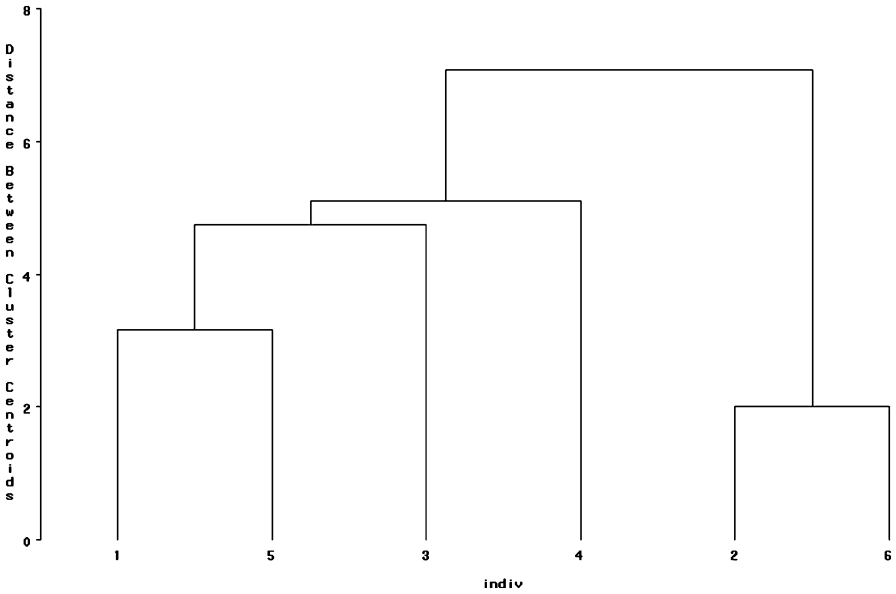


Fig. 12.1 Dendrogram for centroid method on illustrative sample

For the first step, all the possible combinations of pairs of observations are considered as potentially forming the first cluster, with each of the remaining observations forming each one of the $N - 2$ remaining clusters. The sum of squares of each of these pairs of observations is obtained by taking the deviations from the cluster mean, squaring it, and summing it over the P variates. In that first step, only the pairs of observations in the first potential cluster count since the other clusters have a single observation, showing zero deviation from that centroid.

Therefore the pair means or centroids are first computed according to Eq. (12.4):

$$\overline{x}_p(i,j) = \frac{1}{2}\left(x_{pi} + x_{pj}\right) \tag{12.4}$$

where i and j are the indices for the two individual observations.

p is the index for the variable or the variate.
x_{pi} is the coordinate or the value of observation i on variate p.
$\overline{x}_p(i,j)$ is the mean of variate p for observations i and j.

The squared deviations from the centroid can then be computed as

$$d(i,j) = \sum_{p=1}^{\mathrm{P}} \left\{ \left[x_{pi} - \overline{x}_p(i,j)\right]^2 + \left[x_{pj} - \overline{x}_p(i,j)\right]^2 \right\} \tag{12.5}$$

The combination that provides the smallest value $d(i,j)$ is chosen for the first cluster. As indicated above, the other $N - 2$ clusters are composed of the single remaining observations.

For the subsequent steps, all combinations for grouping two of the first-step clusters together are considered. These steps consist of adding to cluster 1 any observations not already in it, as well as considering grouping this observation with any of the other $N - 2$ clusters made up of single observations. The sum of squares is then computed for any such combination. More generally, at any step s, we will be considering $(N - s)$ clusters. A number of alternative combinations are then possible; let us index any of these alternatives by a. We designate the combination of a particular cluster formed within that alternative a as $C_k(a)$, where $k = 1,\ldots(N - s)$ and which contains a number of observations in the cluster, i.e., $C_k(a) = \{i,j,\ldots\}$. We first compute the centroid of the cluster made of the subset of observations $C_k(a)$:

$$\overline{x}_p(C_k(a)) = \frac{1}{n_{C_k(a)}} \sum_{j \in C_k(a)} x_{pj} \tag{12.6}$$

where p is the variate index and $n_{C_k(a)}$ is the number of observations in subset $C_k(a)$.

The squared deviations from the centroid are then

$$d(C_k(a)) = \sum_{p=1}^{P} \sum_{j \in C_k(a)} \left(x_{pj} - \overline{x}_p(C_k(a))\right)^2 \tag{12.7}$$

The sum of squares of a particular alternative a is the sum over the number of clusters at step s (i.e., $N - s$) of the deviations within each cluster. Therefore,

$$d(a) = \sum_{k=1}^{N-s} d(C_k(a)) \tag{12.8}$$

The alternative that provides the smallest value d(a) is then chosen for the next step.

In step 2, this choice could result in an observation being added to the two observations constituting cluster 1 or to any other observation, thus forming another cluster with more than one observation. The process continues until all observations are allocated to a cluster. Therefore, this procedure takes $N - 1$ steps.

We illustrate the process of Ward's method with the same data as used previously for the centroid method (Table 12.1).

Step 1: In this step, we consider all the alternatives for classifying the six individuals (observations) into five groups or clusters. We then select the alternative that provides the smallest sum of squares.

Assign to cluster 1 all possible combinations of pairs among the six individuals ($C_6^2 = 15$ combinations) and the remaining observations to each of the

Table 12.9 Possible alternatives of 2-individual clusters in step 1

Individuals	1	2	3	4	5	6
1	–	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2		–	(2,3)	(2,4)	(2,5)	(2,6)
3			–	(3,4)	(3,5)	(3,6)
4				–	(4,5)	(4,6)
5					–	(5,6)
6						–

Table 12.10 Composition of all possible groups of five clusters and corresponding sum of squares

	Cluster composition					Sum of
Alternative	CL1	CL2	CL3	CL4	CL5	squares
1	(1,2)	3	4	5	6	44.50
2	(1,3)	2	4	5	6	18.50
3	(1,4)	2	3	5	6	15.00
4	(1,5)	2	3	4	6	5.00
5	(1,6)	2	3	4	5	56.50
6	(2,3)	1	4	5	6	10.00
7	(2,4)	1	3	5	6	18.00
8	(2,5)	1	3	4	6	30.50
9	(2,6)	1	3	4	5	2.00
10	(3,4)	1	2	5	6	16.00
11	(3,5)	1	2	4	6	6.50
12	(3,6)	1	2	4	5	20.00
13	(4,5)	1	2	3	6	18.50
14	(4,6)	1	2	3	5	20.00
15	(5,6)	1	2	3	4	44.50

remaining clusters. These various alternatives can be considered by developing the 6-by-6 matrix (displayed in Table 12.9), where only the upper half needs to be considered because the bottom half represents the identical combinations.

To complete the example, we can represent the full set of alternatives with all the elements composing each of the five clusters as shown in Table 12.10.

Compute within-cluster sums of squares for each combination and pick the combination with the smallest sum of squares (ties are broken by picking one randomly).

These sums of squares for each combination are calculated and displayed in the last column of Table 12.10. The smallest value is for the combination where individuals 2 and 6 are grouped together to form cluster 1. Therefore, this cluster (2,6) becomes the first step in the hierarchy. We are now ready for step 2.

Step 2: We now consider all the alternative 4-cluster solutions given that cluster 1 contains individuals 2 and 6. These are represented in Table 12.11.

The full description of each alternative at this stage is shown in Table 12.12 with the computed values of the within-clusters sums of squares for each alternative.

The smallest within sum of squares indicates that a second cluster should be formed with individuals 1 and 5. At this stage, this gives us two clusters of two individuals (2,6) and (1,5) and two clusters with a single individual, i.e., individuals 3 and 4.

Table 12.11 Possible alternatives in step 2

Individuals	(2,6)	1	3	4	5
(2,6)	–	(2,6,1)	(2,6,3)	(2,6,4)	(2,6,5)
1		–	(1,3)	(1,4)	(1,5)
3			–	(3,4)	(3,5)
4				–	(4,5)
5					–

Table 12.12 Composition of all possible groups of four clusters and corresponding sum of squares

	Cluster composition				
Alternative	CL1	CL2	CL3	CL4	Sum of squares
1	(2,6,1)	3	4	5	68.67
2	(2,6,3)	1	4	5	21.34
3	(2,6,4)	1	3	5	34.67
4	(2,6,5)	1	3	4	227.12
5	(2,6)	(1,3)	4	5	20.50
6	(2,6)	(1,4)	3	5	17.00
7	(2,6)	(1,5)	3	4	7.00
8	(2,6)	(3,4)	5	6	18.00
9	(2,6)	(3,5)	1	4	8.50
10	(2,6)	(4,5)	1	3	20.50

Table 12.13 Possible alternatives in step 3

Individuals	(2,6)	(1,5)	3	4
(2,6)	–	(2,6,1,5)	(2,6,3)	(2,6,4)
(1,5)		–	(1,5,3)	(1,5,4)
3			–	(3,4)
4				–

Table 12.14 Composition of all possible groups of three clusters and corresponding sum of squares

	Cluster composition			
Alternative	CL1	CL2	CL3	Sum of squares
1	(2,6,1,5)	3	4	116.48
2	(2,6,3)	(1,5)	4	26.35
3	(2,6,4)	(1,5)	3	39.67
4	(2,6)	(1,5,3)	4	22.00
5	(2,6)	(1,5,4)	3	27.34
6	(2,6)	(1,5)	(3,4)	23.00

Step 3: We now consider all alternatives that would make three clusters. These combinations can be found in Table 12.13.

The complete description of each alternative is shown in Table 12.14 as well as the sum of squares for each alternative.

Based on the sum of squares, we now add individual 3 to cluster 2, so it is now composed of individuals 1, 5, and 3, while cluster 1 remains unchanged and cluster 3 contains a single observation, individual 4.

Table 12.15 Possible alternatives in step 4

Individuals	(2,6)	(1,5,3)	4
(2,6)	–	(2,6,1,5,3)	(2,6,4)
(1,5,3)		–	(1,5,3,4)
4			–

Table 12.16 Composition of all possible groups of two clusters and corresponding sum of squares

	Cluster composition		
Alternative	CL1	CL2	Sum of squares
1	(2,6,1,5,3)	4	95.20
2	(2,6,4)	(1,5,3)	54.67
3	(2,6)	(1,5,3,4)	41.48

Step 4: In the final step where only two clusters are considered, we identify the alternative combinations of two clusters as shown in Table 12.15.

The complete description of each alternative in this step is shown in Table 12.16 as well as the sum of squares for each alternative.

This step finalizes the process, and the best alternative combination results in two clusters, one composed of individuals 2 and 6 and one of individuals 1, 3, 4, and 5.

We can follow each step of this process using the dendrogram as shown in Fig. 12.2.

12.1.4 Nonhierarchical Clustering: K-Means Method

In a nonhierarchical clustering algorithm, the solution is conditional on a predetermined number of clusters selected a priori. If K is the number of groups or clusters, the algorithm follows the four basic steps described below:

Step 1: Assign each of the first K observations to the K clusters as the initial centroids (other assignment rules, such as random selection, offer variants of this method).

Step 2: Compute the distance from each of the other $N - K$ observations to the initial K cluster centroids and assign each observation to the cluster that has the shortest distance (a variant may consist in using a different distance measure). The commands "Kmeans" in STATA and "FASTCLUS" in SAS use the shortest distance between an observation and each of the elements contained in a cluster, instead of the distance to the centroid.

Step 3: Compute the centroids of the K clusters and recompute the distance of each observation not yet assigned to a cluster. Assign that observation to the cluster that has the shortest distance (a variant consists in recomputing the centroid after each observation is assigned).

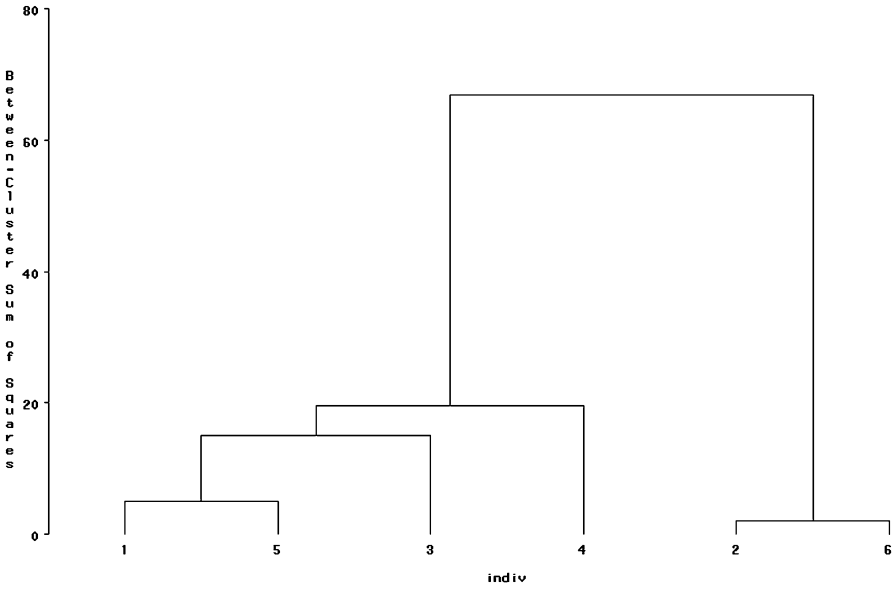


Fig. 12.2 Dendrogram for Ward's method on illustrative sample

Step 4: Repeat step three until the changes in centroids are smaller than a minimum amount (used as a stopping rule) or the maximum number of iterations has been reached.

12.2 Examples

We now illustrate the methods described above. We show how to perform such analyses with both SAS and STATA, using the same data for each. These data concern the assessment of innovations according to a number of variables reflecting different types of innovation characteristics.

12.2.1 Example of Clustering with the Centroid Method

The commands for performing cluster analysis are similar across methods whether in SAS or STATA. Figure 12.3 shows the SAS commands for the centroid method.

The SAS procedure "cluster" is used and the "method=centroid" command simply determines the method used. The dendrogram is requested as an output with the "out=tree" command. The observations classified are identified with the id for variable prod (individual product number contained in the variable named "prod").

```
/*   Examp11-1-Cluster-centroid.sas
*/
option ls=120;
data data1;
infile 'c:\SAMD\Chapter12\Examples\product.dat';
input prod rad it1 it2 it3 it4 it5 it6 it7 it8 it9;
if it1=9 then it1=.;
if it2=9 then it2=.;
if it3=9 then it3=.;
if it4=9 then it4=.;
if it5=9 then it5=.;
if it6=9 then it6=.;
if it7=9 then it7=.;
if it8=9 then it8=.;
if it9=9 then it9=.;
Proc cluster simple noeigen method=centroid rmsstd rsquare nonorm out=tree;
id prod;
var it1-it9;
proc tree data=tree out=clus2 nclusters=2;
id prod;
copy it1-it9;
proc sort; by cluster;
proc print; by cluster;
var prod it1-it9;
Title '2-cluster solution';
Run;
```

Fig. 12.3 SAS commands for centroid method (examp12-1.sas)

```
infile prod rad it1-it9 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter12_ClusterA/PRODUCT.DAT", clear
replace it1=. if it1==9
replace it2=. if it2==9
replace it3=. if it3==9
replace it4=. if it4==9
replace it5=. if it5==9
replace it6=. if it6==9
replace it7=. if it7==9
replace it8=. if it8==9
replace it9=. if it9==9
cluster centroidlinkage it1-it9, name(ClusterCentroid)
cluster dendrogram ClusterCentroid
cluster generate ClusterGroupC = groups(2), name(ClusterCentroid) ties(error)
mean it1-it9, over(ClusterGroupC)
```

Fig. 12.4 STATA commands for centroid method (examp12-1.do)

The variables used for the clustering are listed after the key word "var" and include all the product characteristics it1 through it9.

The observations (products) are sorted by cluster as determined from the results of the cluster analysis and the dendrogram, which is shown in Fig. 12.5.

However, before describing those results, the equivalent input file in STATA is shown in Fig. 12.4.

The centroid method of clustering is specified by the "centroidlinkage" command followed by the list of variables to use to measure distances. The results are saved in a variable using the "ClusterCentroid" name; this chosen name is inserted in the parentheses following the "name" option of the cluster command. Groups are then generated; here two groups are requested to track with the results with two groups. Finally, the means of the nine product characteristics by cluster are computed.

The CLUSTER Procedure
Centroid Hierarchical Cluster Analysis

Variable	Mean	Std Dev	Skewness	Kurtosis	Bimodality
it1	3.0000	1.7321	0.1856	-0.6429	0.2633
it2	4.7778	1.0929	-0.1885	-1.2322	0.3101
it3	3.0000	1.7321	0.1856	-0.6429	0.2633
it4	3.2222	2.2236	0.0808	-2.0806	0.4041
it5	3.1111	1.6915	0.3772	-0.5056	0.2809
it6	3.6667	1.7321	0.1031	-0.9603	0.2799
it7	3.4444	1.8782	0.3294	-1.5570	0.3677
it8	4.2222	1.4814	-1.3745	2.1862	0.4276
it9	4.6667	1.6583	-1.4332	2.2999	0.4444

Root-Mean-Square Total-Sample Standard Deviation = 1.715039

Cluster History

NCL	Clusters Joined		FREQ	RMS STD	SPRSQ	RSQ	Cent Dist	Tie
8	7	12	2	0.5774	0.0142	.986	2.4495	
7	3	CL8	3	0.7454	0.0331	.953	3.2404	
6	9	11	2	0.8165	0.0283	.924	3.4641	
5	2	CL6	3	0.9623	0.0504	.874	4	
4	CL5	13	4	1.0585	0.0641	.810	4.2557	
3	CL7	10	4	1.0000	0.0803	.730	4.761	
2	CL4	8	5	1.2156	0.1084	.621	5.3561	
1	CL2	CL3	9	1.7150	0.6213	.000	7.6948	

CLUSTER=.

Obs	prod	it1	it2	it3	it4	it5	it6	it7	it8	it9
1	1	1	6	2	5	.	5	5	5	5
2	4	2	5	2	.	5	.	5	5	2
3	5	6	.	6	2	2	2	2	2	2
4	6	2	5	2	4	.	5	5	5	6

CLUSTER=1

Fig. 12.5 SAS example output of cluster analysis using the centroid method (examp12-1.lst)

Obs	prod	it1	it2	it3	it4	it5	it6	it7	it8	it9
5	7	1	6	1	6	4	6	6	6	6
6	12	1	6	1	6	3	6	6	5	4
7	3	1	6	1	5	6	5	5	5	6
8	10	3	5	3	4	5	3	4	5	6

---------- CLUSTER=2 ----------

Obs	prod	it1	it2	it3	it4	it5	it6	it7	it8	it9
9	9	4	5	4	1	3	1	2	5	6
10	11	3	4	3	1	3	3	2	4	4
11	2	6	4	6	1	2	2	1	4	4
12	13	4	4	4	4	1	4	2	3	5
13	8	4	3	4	1	1	3	3	1	1

Fig. 12.5 (continued)

```
. infile prod rad it1-it9 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter12_ClusterA/PRODUCT.DAT", clear
(13 observations read)

…

. cluster centroidlinkage it1-it9, name(ClusterCentroid)

. cluster dendrogram ClusterCentroid

. cluster generate ClusterGroupC = groups(2), name(ClusterCentroid) ties(error)

. mean it1-it9, over(ClusterGroupC)

Mean estimation                     Number of obs    =         9

            1: ClusterGroupC = 1
            2: ClusterGroupC = 2

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
it1          |
           1 |        4.2   .4898979      3.070293    5.329707
           2 |        1.5         .5      .3469979    2.653002
-------------+------------------------------------------------
it2          |
           1 |          4   .3162278      3.270777    4.729223
           2 |       5.75        .25      5.173499    6.326501
-------------+------------------------------------------------
it3          |
           1 |        4.2   .4898979      3.070293    5.329707
           2 |        1.5         .5      .3469979    2.653002
-------------+------------------------------------------------
it4          |
           1 |        1.6         .6      .2163975    2.983602
           2 |       5.25   .4787136      4.146085    6.353915
-------------+------------------------------------------------
it5          |
           1 |          2   .4472136      .9687236    3.031276
           2 |        4.5   .6454972      3.011481    5.988519
-------------+------------------------------------------------
it6          |
           1 |        2.6    .509902      1.424164    3.775836
           2 |          5   .7071068      3.369409    6.630591
-------------+------------------------------------------------
it7          |
           1 |          2   .3162278      1.270777    2.729223
           2 |       5.25   .4787136      4.146085    6.353915
-------------+------------------------------------------------
it8          |
           1 |        3.4    .678233      1.835992    4.964008
           2 |       5.25        .25      4.673499    5.826501
-------------+------------------------------------------------
it9          |
           1 |          4     .83666      2.070659    5.929341
           2 |        5.5         .5      4.346998    6.653002
--------------------------------------------------------------
```

Fig. 12.6 STATA output of cluster analysis using the centroid method (examp12-1.log)

We now consider the results of running these input files, comparing the SAS and the STATA outputs. The results from SAS are shown in Fig. 12.5, and those from STATA in Fig. 12.6.

The STATA output is listed in Fig. 12.6.

After providing standard statistics of the variables used for the classification analysis, each step at each level of the hierarchy is shown. For example, at the first step (i.e., when eight clusters are considered as shown in the output with the value 8 in the NCL column), products 7 and 12 are the least dissimilar and are placed together in a cluster called CL8. In the next step, when seven clusters are

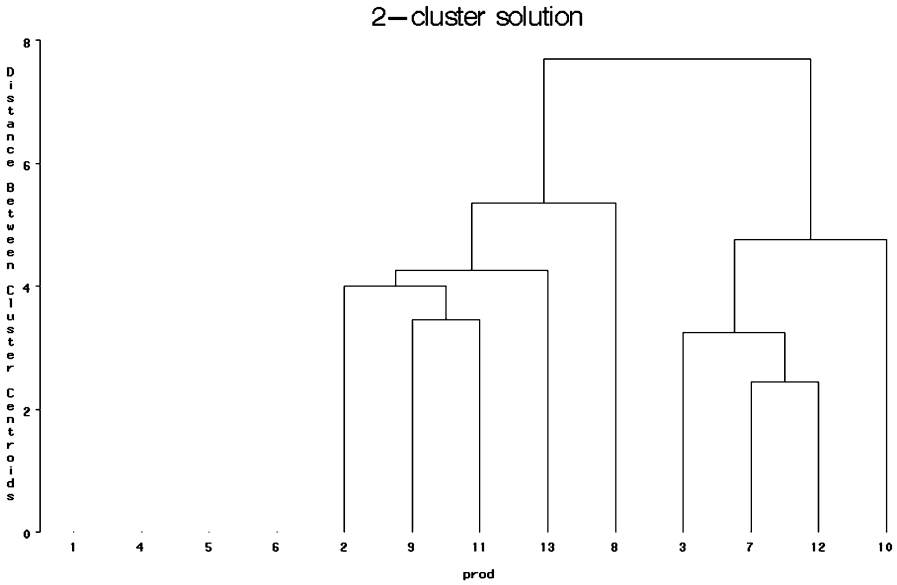


Fig. 12.7 Dendrogram of cluster analysis using the centroid method—SAS (examp12-1)

considered, product number 3 is the most similar to the newly formed cluster CL8. This process continues until only two clusters are formed.

The dendrogram corresponding to this analysis (SAS output) is shown in Fig. 12.7, where the entire hierarchy appears.

Products identified by numbers 1, 4, 5, and 6 appear as dots on the graph and were not classified because values were missing for some of the variables on these products. This final classification and the corresponding data are printed in the output sorted by cluster, and the same data can be found in the SAS work file "clus2" by clicking in the SAS menu bar on "solutions/Analysis/Interactive Data Analysis" and then by selecting the SAS library "WORK" and the SAS Data Set clus2 (the name indicated in the SAS command to create that file). That file is shown in Fig. 12.8. Note that it is possible to print the file using the File/Print menu option.

The STATA-generated dendrogram is identical (Fig. 12.9), except for the observations with missing data that were deleted from the analysis and that consequently do not appear on the graph.

It remains to interpret the grouping found statistically. For that purpose, it is useful to calculate the means of the variables by cluster, i.e., the values of the centroids at the final solution of two clusters. This was requested in the last line of the STATA input commands "mean it1-it9, over(ClusterGroupC)" in Fig. 12.4. In SAS, this can be done easily by adding the commands that appear at the bottom of Fig. 12.10.

The presentation of the results is somewhat different in SAS, as shown in Fig. 12.11. The means of all the variables are listed for cluster 1 first and then for cluster 2.

	f2	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Nom
f3		prod	it1	it2	it3	it4	it5	it6	it7	it8	it9	CLUSTER	CLUSNAME
■	1	1	1	6	2	5	.	5	5	5	5	.	
■	2	4	2	5	2	.	5	.	5	5	2	.	
■	3	5	6	.	6	2	2	2	2	2	2	.	
■	4	6	2	5	2	4	.	5	5	5	6	.	
■	5	7	1	6	1	6	4	6	6	6	6	1	CL3
■	6	12	1	6	1	6	3	6	6	5	4	1	CL3
■	7	3	1	6	1	5	6	5	5	5	6	1	CL3
■	8	10	3	5	3	4	5	3	4	5	6	1	CL3
■	9	9	4	5	4	1	3	1	2	5	6	2	CL2
■	10	11	3	4	3	1	3	3	2	4	4	2	CL2
■	11	2	6	4	6	1	2	2	1	4	4	2	CL2
■	12	13	4	4	4	4	1	4	2	3	5	2	CL2
■	13	8	4	3	4	1	1	3	3	1	1	2	CL2

Fig. 12.8 Example of 2-cluster solution with centroid method—SAS (examp12-1)

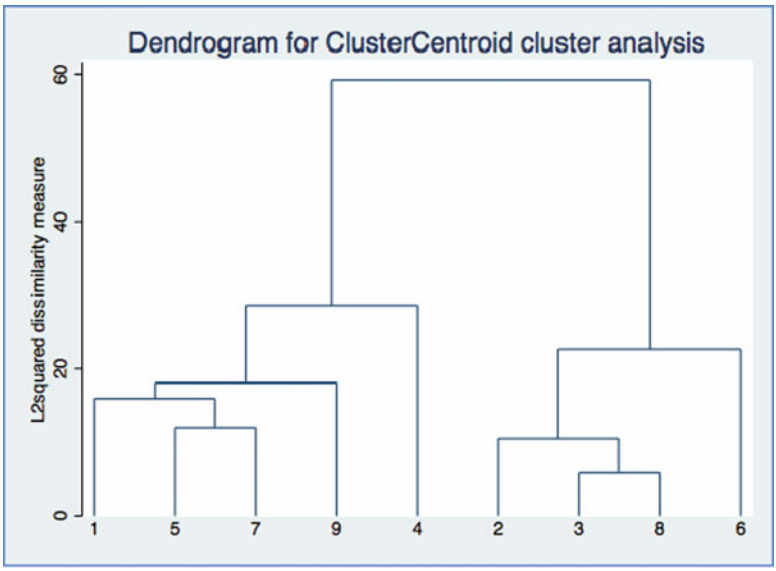


Fig. 12.9 Dendrogram of cluster analysis using the centroid method—STATA (examp12-1)

```
proc means; by cluster;
        var it1-it9;
run;
```

Fig. 12.10 Commands for calculating means of clustering variables by cluster—SAS

```
------------------------------------------------ CLUSTER=1 ---------------------------------
Variable     N           Mean         Std Dev         Minimum         Maximum
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
it1          4       1.5000000       1.0000000       1.0000000       3.0000000
it2          4       5.7500000       0.5000000       5.0000000       6.0000000
it3          4       1.5000000       1.0000000       1.0000000       3.0000000
it4          4       5.2500000       0.9574271       4.0000000       6.0000000
it5          4       4.5000000       1.2909944       3.0000000       6.0000000
it6          4       5.0000000       1.4142136       3.0000000       6.0000000
it7          4       5.2500000       0.9574271       4.0000000       6.0000000
it8          4       5.2500000       0.5000000       5.0000000       6.0000000
it9          4       5.5000000       1.0000000       4.0000000       6.0000000
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

------------------------------------------------ CLUSTER=2 ---------------------------------
Variable     N           Mean         Std Dev         Minimum         Maximum
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
it1          5       4.2000000       1.0954451       3.0000000       6.0000000
it2          5       4.0000000       0.7071068       3.0000000       5.0000000
it3          5       4.2000000       1.0954451       3.0000000       6.0000000
it4          5       1.6000000       1.3416408       1.0000000       4.0000000
it5          5       2.0000000       1.0000000       1.0000000       3.0000000
it6          5       2.6000000       1.1401754       1.0000000       4.0000000
it7          5       2.0000000       0.7071068       1.0000000       3.0000000
it8          5       3.4000000       1.5165751       1.0000000       5.0000000
it9          5       4.0000000       1.8708287       1.0000000       6.0000000
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

Fig. 12.11 Output of means by cluster

```
/*   Examp11-2-Cluster-ward.sas
*/
option ls=120;
data data1;
infile 'c:\SAMD\Chapter12\Examples\product.dat';
input prod rad it1 it2 it3 it4 it5 it6 it7 it8 it9;
if it1=9 then it1=.;
if it2=9 then it2=.;
if it3=9 then it3=.;
if it4=9 then it4=.;
if it5=9 then it5=.;
if it6=9 then it6=.;
if it7=9 then it7=.;
if it8=9 then it8=.;
if it9=9 then it9=.;

Proc cluster simple noeigen method=ward rmsstd rsquare nonorm out=tree;
id prod;
var it1-it9;
proc tree data=tree out=clus2 nclusters=2;
id prod;
copy it1-it9;
proc sort; by cluster;
proc print; by cluster;
var prod it1-it9;
Title '2-cluster solution';
run;
proc means; by cluster;
       var it1-it9;
run;
```

Fig. 12.12 SAS commands for cluster analysis using Ward's method (examp12-2.sas)

The variables that show the largest differences can help interpret the meaning of the groups. In that sense, cluster analysis is purely exploratory as there is no a priori theory needed to discover how the observations can be grouped by similarity to each other.

12.2.2 Example of Clustering with Ward's Method

The commands in SAS for Ward's method are the same as in the previous example, except for replacing "method=centroid" with "method=ward" in the proc cluster command line. An example is shown in Fig. 12.12.

Similarly, the commands in STATA are shown in Fig. 12.13. The commands are identical, except for the name "wardslinkage" that specifies the Ward method.

The output of the Ward's method example is given in Fig. 12.14. After the basic statistics for the variables used in the cluster analysis have been listed, the formation of the clusters at each step is shown and the final solution is given with the values of the variables for each observation listed by cluster. The centroids, i.e., the mean of each variable for each cluster, are then given with the standard deviation (as well as the minimum and the maximum).

The STATA output is shown in Fig. 12.15.

Finally, the dendrogram is shown in Fig. 12.16 with up to two clusters, as instructed in the input commands on the proc tree line.

```
infile prod rad it1-it9 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter12_ClusterA/PRODUCT.DAT", clear
replace it1=. if it1==9
replace it2=. if it2==9
replace it3=. if it3==9
replace it4=. if it4==9
replace it5=. if it5==9
replace it6=. if it6==9
replace it7=. if it7==9
replace it8=. if it8==9
replace it9=. if it9==9
cluster wardslinkage it1-it9, name(ClusterWard)
cluster dendrogram ClusterWard
cluster generate ClusterGroupW = groups(2), name(ClusterWard) ties(error)
mean it1-it9, over(ClusterGroupW)
```

Fig. 12.13 STATA commands for cluster analysis using Ward's method (examp12-2.do)

As for the centroid method, the STATA dendrogram is identical (Fig. 12.17), except for the observations with missing values that have been deleted.

12.2.3 Examples of K-Means Analysis

Finally, we present examples of K-means analysis using the "FASTCLUS" procedure in SAS and "Kmeans" in STATA.

The "FASTCLUS" procedure is illustrated with input commands in SAS in Fig. 12.18. The SAS command is "proc fastclus" and the maximum number of clusters is chosen with the "maxclusters=2" command. In this particular example, the observations with missing values have been deleted.

The equivalent analysis in STATA is requested in Fig. 12.19.

The method is specified by the "cluster kmeans" command, followed by the variables to be used. The number of clusters is indicated by the "k(2)" command. The variable "ClusterKmeans" contains the cluster assignment of the observations.

Figure 12.20 lists the output. The results indicate the composition of the two clusters and the cluster means on each of the variables used, similar to the output from the other methods.

The STATA output of the K-means method is listed in Fig. 12.21.

12.3 Evaluation and Interpretation of Clustering Results

Because cluster analysis techniques are exploratory and are not founded on statistical theory, authors reporting any cluster solution or the reader of cluster analyses found in the literature should evaluate very carefully the method used. A number of issues should be addressed in such reports.

The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

Variable	Mean	Std Dev	Skewness	Kurtosis	Bimodality
it1	3.0000	1.7321	0.1856	-0.6429	0.2633
it2	4.7778	1.0929	-0.1885	-1.2322	0.3101
it3	3.0000	1.7321	0.1856	-0.6429	0.2633
it4	3.2222	2.2236	0.0808	-2.0806	0.4041
it5	3.1111	1.6915	0.3772	-0.5056	0.2809
it6	3.6667	1.7321	0.1031	-0.9603	0.2799
it7	3.4444	1.8782	0.3294	-1.5570	0.3677
it8	4.2222	1.4814	-1.3745	2.1862	0.4276
it9	4.6667	1.6583	-1.4332	2.2999	0.4444

Root-Mean-Square Total-Sample Standard Deviation = 1.715039

Cluster History

NCL	Clusters Joined		FREQ	RMS STD	SPRSQ	RSQ	BSS	Tie
8	7	12	2	0.5774	0.0142	.986	3	
7	9	11	2	0.8165	0.0283	.958	6	
6	3	CL8	3	0.7454	0.0331	.924	7	
5	2	CL7	3	0.9623	0.0504	.874	10.667	
4	CL5	13	4	1.0585	0.0641	.810	13.583	
3	CL6	10	4	1.0000	0.0803	.730	17	
2	CL4	8	5	1.2156	0.1084	.621	22.95	
1	CL2	CL3	9	1.7150	0.6213	.000	131.58	

CLUSTER=.

Obs	prod	it1	it2	it3	it4	it5	it6	it7	it8	it9
1	1	1	6	2	5	.	5	5	5	5
2	4	2	5	2	.	5	.	5	5	2
3	5	6	.	6	2	2	2	2	2	2
4	6	2	5	2	4	.	5	5	5	6

CLUSTER=1

Obs	prod	it1	it2	it3	it4	it5	it6	it7	it8	it9
5	7	1	6	1	6	4	6	6	6	6

Fig. 12.14 SAS output of cluster analysis using Ward's method (examp12-2.lst)

```
     6     12      1      6      1      6      3      6      6      5      4
     7      3      1      6      1      5      6      5      5      5      6
     8     10      3      5      3      4      5      3      4      5      6
---------------------------------------------- CLUSTER=2 ------------------------------
   Obs    prod    it1    it2    it3    it4    it5    it6    it7    it8    it9
     9      9      4      5      4      1      3      1      2      5      6
    10     11      3      4      3      1      3      3      2      4      4
    11      2      6      4      6      1      2      2      1      4      4
    12     13      4      4      4      4      1      4      2      3      5
    13      8      4      3      4      1      1      3      3      1      1

---------------------------------------------- CLUSTER=. ------------------------------
Variable     N            Mean         Std Dev         Minimum         Maximum
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
it1          4       2.7500000       2.2173558       1.0000000       6.0000000
it2          3       5.3333333       0.5773503       5.0000000       6.0000000
it3          4       3.0000000       2.0000000       2.0000000       6.0000000
it4          3       3.6666667       1.5275252       2.0000000       5.0000000
it5          2       3.5000000       2.1213203       2.0000000       5.0000000
it6          3       4.0000000       1.7320508       2.0000000       5.0000000
it7          4       4.2500000       1.5000000       2.0000000       5.0000000
it8          4       4.2500000       1.5000000       2.0000000       5.0000000
it9          4       3.7500000       2.0615528       2.0000000       6.0000000
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff

---------------------------------------------- CLUSTER=1 ------------------------------
Variable     N            Mean         Std Dev         Minimum         Maximum
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
it1          4       1.5000000       1.0000000       1.0000000       3.0000000
it2          4       5.7500000       0.5000000       5.0000000       6.0000000
it3          4       1.5000000       1.0000000       1.0000000       3.0000000
it4          4       5.2500000       0.9574271       4.0000000       6.0000000
it5          4       4.5000000       1.2909944       3.0000000       6.0000000
it6          4       5.0000000       1.4142136       3.0000000       6.0000000
it7          4       5.2500000       0.9574271       4.0000000       6.0000000
it8          4       5.2500000       0.5000000       5.0000000       6.0000000
it9          4       5.5000000       1.0000000       4.0000000       6.0000000
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff

---------------------------------------------- CLUSTER=2 ------------------------------
Variable     N            Mean         Std Dev         Minimum         Maximum
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
it1          5       4.2000000       1.0954451       3.0000000       6.0000000
it2          5       4.0000000       0.7071068       3.0000000       5.0000000
```

Fig. 12.14 (continued)

```
it3          5       4.2000000       1.0954451       3.0000000       6.0000000
it4          5       1.6000000       1.3416408       1.0000000       4.0000000
it5          5       2.0000000       1.0000000       1.0000000       3.0000000
it6          5       2.6000000       1.1401754       1.0000000       4.0000000
it7          5       2.0000000       0.7071068       1.0000000       3.0000000
it8          5       3.4000000       1.5165751       1.0000000       5.0000000
it9          5       4.0000000       1.8708287       1.0000000       6.0000000
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```

Fig. 12.14 (continued)

```
. infile prod rad it1-it9 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter12_ClusterA/PRODUCT.DAT", clear
(13 observations read)

…

. cluster wardslinkage it1-it9, name(ClusterWard)

. cluster dendrogram ClusterWard

. cluster generate ClusterGroupW = groups(2), name(ClusterWard) ties(error)

. mean it1-it9, over(ClusterGroupW)

Mean estimation                     Number of obs    =         9

            1: ClusterGroupW = 1
            2: ClusterGroupW = 2

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
it1          |
           1 |        4.2   .4898979      3.070293    5.329707
           2 |        1.5         .5      .3469979    2.653002
-------------+------------------------------------------------
it2          |
           1 |          4   .3162278      3.270777    4.729223
           2 |       5.75        .25      5.173499    6.326501
-------------+------------------------------------------------
it3          |
           1 |        4.2   .4898979      3.070293    5.329707
           2 |        1.5         .5      .3469979    2.653002
-------------+------------------------------------------------
it4          |
           1 |        1.6         .6      .2163975    2.983602
           2 |       5.25   .4787136      4.146085    6.353915
-------------+------------------------------------------------
it5          |
           1 |          2   .4472136      .9687236    3.031276
           2 |        4.5   .6454972      3.011481    5.988519
-------------+------------------------------------------------
it6          |
           1 |        2.6    .509902      1.424164    3.775836
           2 |          5   .7071068      3.369409    6.630591
-------------+------------------------------------------------
it7          |
           1 |          2   .3162278      1.270777    2.729223
           2 |       5.25   .4787136      4.146085    6.353915
-------------+------------------------------------------------
it8          |
           1 |        3.4    .678233      1.835992    4.964008
           2 |       5.25        .25      4.673499    5.826501
-------------+------------------------------------------------
it9          |
           1 |          4     .83666      2.070659    5.929341
           2 |        5.5         .5      4.346998    6.653002
--------------------------------------------------------------
```

Fig. 12.15 STATA output of cluster analysis using Ward's method (examp12-2.log)

12.3.1 Determining the Number of Clusters

The determination of the number of clusters is a critical choice that, unfortunately, cannot be inferred from the analysis. In hierarchical methods, the stopping rule is fairly ad hoc and in the nonhierarchical method presented in this chapter, the choice must be done a priori. Although we can consider some guiding measures, these are not without problems and the best argument for the choice of the number of clusters is probably the one based on the interpretability of the resulting clusters.

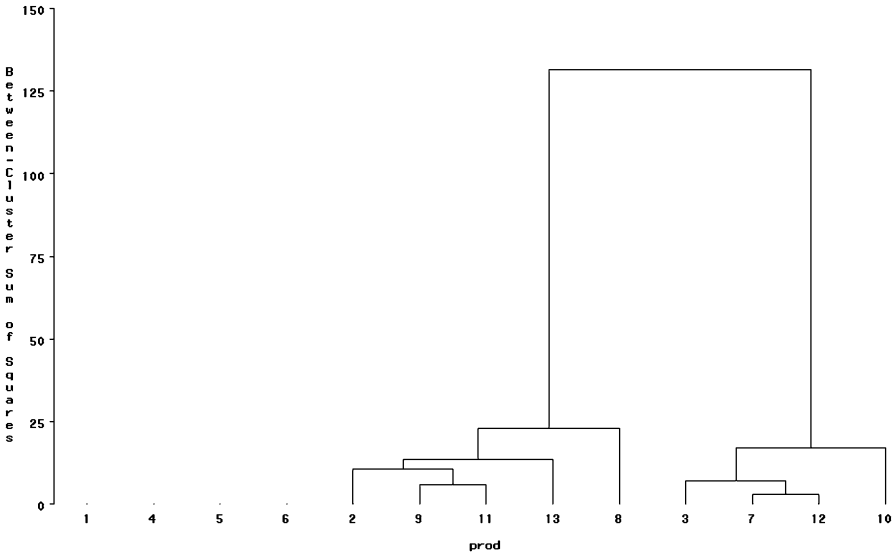


Fig. 12.16 Dendrogram from Ward's method—SAS

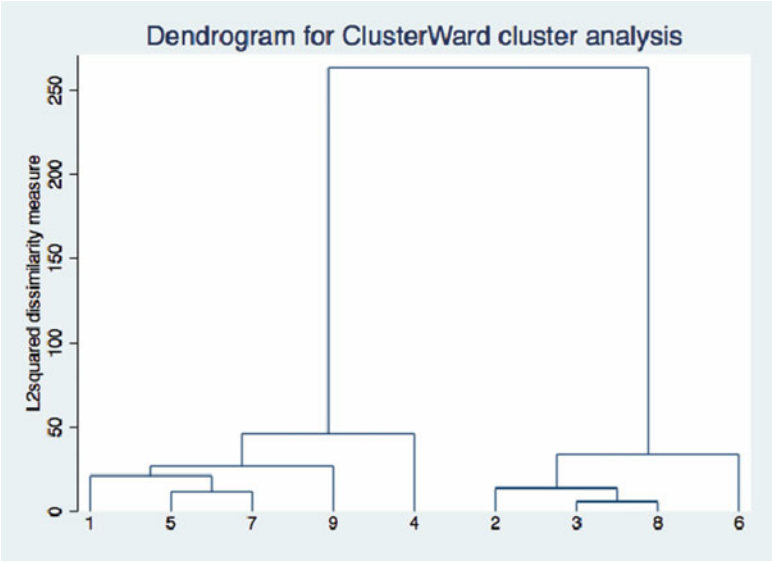


Fig. 12.17 Dendrogram from Ward's method—STATA

```
/*   Examp11-3-FASTCLUS.sas
*/
option ls=120;
data data1;
infile 'c:\SAMD\Chapter12\Examples\product.dat';
input prod rad it1 it2 it3 it4 it5 it6 it7 it8 it9;
if it1=9 then it1=.;
if it2=9 then it2=.;
if it3=9 then it3=.;
if it4=9 then it4=.;
if it5=9 then it5=.;
if it6=9 then it6=.;
if it7=9 then it7=.;
if it8=9 then it8=.;
if it9=9 then it9=.;
if it1=. then delete;
if it2=. then delete;
if it3=. then delete;
if it4=. then delete;
if it5=. then delete;
if it6=. then delete;
if it7=. then delete;
if it8=. then delete;
if it9=. then delete;
Proc fastclus radius=0 replace=full maxclusters=2 maxiter=50 list distance;
       id prod;
       var it1-it9;
run;
```

Fig. 12.18 SAS commands for cluster analysis using "FASTCLUS" (examp12-3.sas)

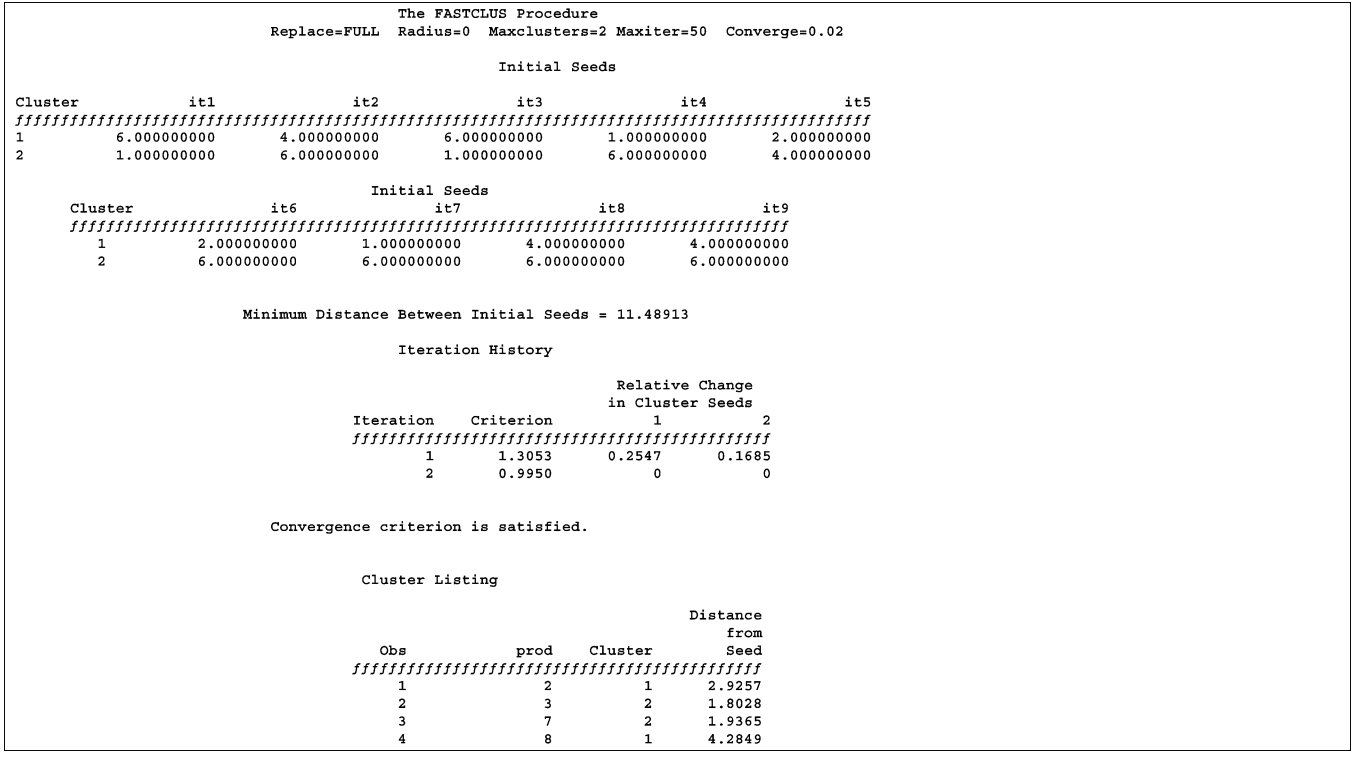
```
infile prod rad it1-it9 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter12_ClusterA/PRODUCT.DAT", clear
replace it1=. if it1==9
replace it2=. if it2==9
replace it3=. if it3==9
replace it4=. if it4==9
replace it5=. if it5==9
replace it6=. if it6==9
replace it7=. if it7==9
replace it8=. if it8==9
replace it9=. if it9==9
cluster kmeans it1-it9, k(2) measure(L2) name(ClusterKmeans) start(krandom)
mean it1-it9, over(ClusterKmeans)
```

Fig. 12.19 STATA commands for cluster analysis using "Kmeans" (examp12-3.do)

12.3.2 Size, Density, and Separation of Clusters

One criterion to assess the quality of a solution is that each cluster must contain a sufficient number of observations. A lone observation is more probably an outlier and it may be difficult to describe it based on theory-based distinguishing features. In general, a balance in the size of clusters may be ideal, although it is by no means a necessary condition for a meaningful grouping of observations.

In principle, the density and separation of the clusters are more critical because this discrimination is the reason for choosing cluster analysis. Density refers to how similar the observations within a group are (i.e., the within-group variance). Separation refers to the spread or how different observations across groups are (i.e., the between-group variance). Consequently, the least we would anticipate from a cluster solution is that the groups are statistically different on the variables



The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=2 Maxiter=50 Converge=0.02

Initial Seeds

Cluster	it1	it2	it3	it4	it5
1	6.000000000	4.000000000	6.000000000	1.000000000	2.000000000
2	1.000000000	6.000000000	1.000000000	6.000000000	4.000000000

Initial Seeds

Cluster	it6	it7	it8	it9
1	2.000000000	1.000000000	4.000000000	4.000000000
2	6.000000000	6.000000000	6.000000000	6.000000000

Minimum Distance Between Initial Seeds = 11.48913

Iteration History

Iteration	Criterion	Relative Change in Cluster Seeds 1	2
1	1.3053	0.2547	0.1685
2	0.9950	0	0

Convergence criterion is satisfied.

Cluster Listing

Obs	prod	Cluster	Distance from Seed
1	2	1	2.9257
2	3	2	1.8028
3	7	2	1.9365
4	8	1	4.2849

Fig. 12.20 SAS output for cluster analysis using FASTCLUS (examp12-3.lst)

		5	9	1	3.4000
		6	10	2	3.5707
		7	11	1	2.1817
		8	12	2	2.6926
		9	13	1	3.1559

Criterion Based on Final Seeds = 0.9950

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	5	1.2156	4.2849		2	7.6948
2	4	1.0000	3.5707		1	7.6948

Statistics for Variables

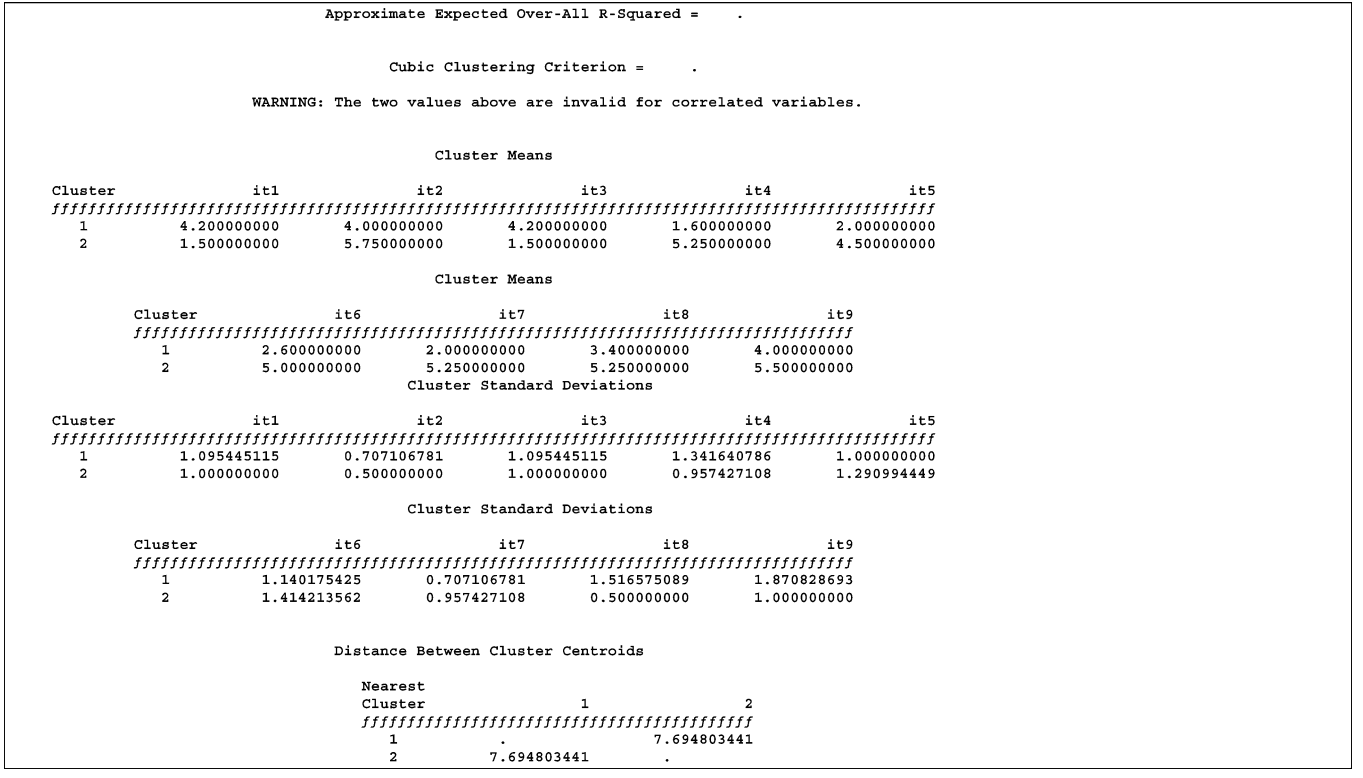
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
it1	1.73205	1.05560	0.675000	2.076923
it2	1.09291	0.62678	0.712209	2.474747
it3	1.73205	1.05560	0.675000	2.076923

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
it4	2.22361	1.19224	0.748455	2.975433
it5	1.69148	1.13389	0.606796	1.543210
it6	1.73205	1.26491	0.533333	1.142857
it7	1.87824	0.82375	0.831693	4.941520
it8	1.48137	1.19224	0.433228	0.764377
it9	1.65831	1.55839	0.227273	0.294118
OVER-ALL	1.71504	1.12828	0.621301	1.640621

Pseudo F Statistic = 11.48

Fig. 12.20 (continued)



Approximate Expected Over-All R-Squared = .

Cubic Clustering Criterion = .

WARNING: The two values above are invalid for correlated variables.

Cluster Means

Cluster	it1	it2	it3	it4	it5
1	4.200000000	4.000000000	4.200000000	1.600000000	2.000000000
2	1.500000000	5.750000000	1.500000000	5.250000000	4.500000000

Cluster Means

Cluster	it6	it7	it8	it9
1	2.600000000	2.000000000	3.400000000	4.000000000
2	5.000000000	5.250000000	5.250000000	5.500000000

Cluster Standard Deviations

Cluster	it1	it2	it3	it4	it5
1	1.095445115	0.707106781	1.095445115	1.341640786	1.000000000
2	1.000000000	0.500000000	1.000000000	0.957427108	1.290994449

Cluster Standard Deviations

Cluster	it6	it7	it8	it9
1	1.140175425	0.707106781	1.516575089	1.870828693
2	1.414213562	0.957427108	0.500000000	1.000000000

Distance Between Cluster Centroids

Nearest Cluster	1	2
1	.	7.694803441
2	7.694803441	.

Fig. 12.20 (continued)

```
. infile prod rad it1-it9 using
"/users/fblgatignon/Documents/WORK_STATA/SAMD/Chapter12_ClusterA/PRODUCT.DAT", clear
(13 observations read)

…

. cluster kmeans it1-it9, k(2) measure(L2) name(ClusterKmeans) start(krandom)

. mean it1-it9, over(ClusterKmeans)

Mean estimation                     Number of obs    =         9

            1: ClusterKmeans = 1
            2: ClusterKmeans = 2

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
it1          |
           1 |         1.5         .5      .3469979    2.653002
           2 |         4.2   .4898979      3.070293    5.329707
-------------+------------------------------------------------
it2          |
           1 |        5.75        .25      5.173499    6.326501
           2 |           4   .3162278      3.270777    4.729223
-------------+------------------------------------------------
it3          |
           1 |         1.5         .5      .3469979    2.653002
           2 |         4.2   .4898979      3.070293    5.329707
-------------+------------------------------------------------
it4          |
           1 |        5.25   .4787136      4.146085    6.353915
           2 |         1.6         .6      .2163975    2.983602
-------------+------------------------------------------------
it5          |
           1 |         4.5   .6454972      3.011481    5.988519
           2 |           2   .4472136      .9687236    3.031276
-------------+------------------------------------------------
it6          |
           1 |           5   .7071068      3.369409    6.630591
           2 |         2.6    .509902      1.424164    3.775836
-------------+------------------------------------------------
it7          |
           1 |        5.25   .4787136      4.146085    6.353915
           2 |           2   .3162278      1.270777    2.729223
-------------+------------------------------------------------
it8          |
           1 |        5.25        .25      4.673499    5.826501
           2 |         3.4    .678233      1.835992    4.964008
-------------+------------------------------------------------
it9          |
           1 |         5.5         .5      4.346998    6.653002
           2 |           4     .83666      2.070659    5.929341
--------------------------------------------------------------
```

Fig. 12.21 STATA output for cluster analysis using K-means method (examp12-3.log)

used to perform the cluster analysis. In practice, although it is not a bad idea to perform such an analysis using a MANOVA, the results tend to be highly significant and the diagnostic value is small. Moreover, the results from a MANOVA can be misleading because it does not indicate whether the distribution of any variable used for clustering is bimodal or multimodal. In fact, for example, a variable distributed according to a normal distribution may lead to the formation of two groups (low and high) when clustering the observations on that variable. The means of the two groups are likely to be significantly different from each other. Nevertheless, this does not mean that the observations in each group are sampled from a different distribution.

12.3.3 Tests of Significance on Variables Other than Those Used to Create Clusters

The best method for "validating" the clustering solution consists in verifying that the clusters differ on variables that are not used in the clustering process. These variables typically concern differences the researcher expects from such groups but do not characterize the groups per se, i.e., they do not contribute to their definition. For example, consumers can be segmented on the basis of demographics and psychographics, and once groups are formed based on these descriptive variables, it can be verified whether each group differs in terms of specific purchase behavior.

In order to run a MANOVA in SAS on variables that are not used in the cluster analysis using the clusters to define groups, you need to create an id variable for the observations and sort the cluster output file by that id before merging the cluster variable with all the data.

This can be done by the following:

1. Inserting a line to create an id variable in the data set that you read/input. After the input statement, insert a line.

   idobs=_N_;

   This creates a variable called "idobs" that will take a value of 1 to 300 corresponding to each respondent in the sample.

2. Before running a MANOVA, sort the cluster file and merge the two data sets. For example, if "clus2" is the name you have assigned to the cluster output using the "out=clus2" command:

   proc sort data=clus2;
   by idobs;
   Data merged; ("merged" is the name of the new combined data set)
   merge old clus2; ("old" is the name of the data set you created for the input data)

Once this is done, you can specify a MANOVA with any variables from the data that were originally read.

The commands are straightforward in STATA, combining the clustering commands shown in this chapter with the MANOVA commands explained in Chap. 2:

```
cluster kmeans it1-it6, k(2) measure(L2) name(ClusterKmeans) start(krandom)
mean it1-it6, over(ClusterKmeans)
manova it7-it9=ClusterKmeans
mean it7-it9, over(ClusterKmeans)
```

12.3.4 Stability of Results

Given that the results of a cluster analysis are rather exploratory and could vary depending on the method, it is best to verify the stability of the results. This can be done with a split sample procedure where the analysis is performed on the two subsamples and the researcher can check that the interpretation of the clusters remains the same, i.e., that the results are stable. Also, because it is difficult to justify one method versus another on theoretical grounds, it is reasonable to perform the analysis using different procedures so that the biases inherent in each method (e.g., tendency to cluster around seed points) can be better evaluated. Performing such an analysis is good practice since it demonstrates the necessary stability of the results; however, it does not guarantee that the clustering solution corresponds to "real" groups that present "real" differences.

12.4 Assignment

Perform a cluster analysis using the data contained in the survey (SURVEY.ASC) described in Appendix C (Chap. 14). You should identify an appropriate segmentation scheme for that sample of individuals so that these individuals are grouped with relatively homogeneous psychographic profiles.

Bibliography

Basic Technical Readings

Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a data set: An information-theoretic approach. *Journal of the American Statistical Association, 98*(463), 750–763.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Application Readings

Askegaard, S., & Madsen, T. K. (1998). The local and the global: Exploring traits of homogeneity and heterogeneity in European food cultures. *International Business Review, 7*(6), 549–568.

Calantone, R. J., & Di Benedetto, C. A. (2007). Clustering product launches by price and launch strategy. *The Journal of Business and Industrial Marketing, 22*(1), 4–19.

DeSarbo, W. S., & De Soete, G. (1984). On the use of hierarchical clustering for the analysis of nonsymmetric proximities. *Journal of Consumer Research, 11*(1), 601–610.

Hall, E. H., Jr., & St. John, C. H. (1994). A methodological note on diversity measurement. *Strategic Management Journal, 15*(2), 153–168.

Helsen, K., & Green, P. E. (1991). A computational study of replicated clustering with an application to market segmentation. *Decision Sciences, 22*, 1124–1141.

Helsen, K., Jedidi, K., & DeSarbo, W. S. (1993). A new approach to country segmentation utilizing multinational diffusion patterns. *Journal of Marketing, 57*(4), 60–71.

Hultink, E. J., Griffin, A., Robben, H. S. J., & Hart, S. (1998). In search of generic launch strategies for new products. *International Journal of Research in Marketing, 15*, 269–285.

Kale, S. H. (1995). Grouping euroconsumers: A culture-based clustering approach. *Journal of International Marketing, 3*(3), 35–48.

Kumar, V., Ganesh, J., & Echambadi, R. (1998). Cross-national diffusion research: What do we know and how certain are we? *Journal of Product Innovation Management, 15*(3), 255–268.

Oliver, R. L., & Anderson, E. (1995). Behavior- and outcome-based sales control systems: Evidence and consequences of pure-form and hybrid governance. *Journal of Personal Selling & Sales Management, 15*(4), 1–15.

Sethi, S. P. (1971). Comparative cluster analysis for world markets. *Journal of Marketing Research, 8*(3), 348–354.

Sexton, D. E., Jr. (1974). A cluster analytic approach to market response functions. *Journal of Marketing Research, 11*(1), 109–114.

Srivatsava, R. K., Leone, R. P., & Shocker, A. D. (1981). Market structure analysis: Hierarchical clustering of products based on substitution in use. *Journal of Marketing, 45*(3), 38–48.

Steenkamp, J.-B. E. M. (2001). The role of national culture in international marketing research. *International Marketing Review, 18*(1), 30–44.

Vandermerwe, S., & L'Huillier, M.-A. (1989). Euro-consumers in 1992. *Business Horizons, 32*(1), 34–40.

Völckner, F., & Sattler, H. (2007). Empirical generalizability of consumer evaluations of brand extensions. *International Journal of Research in Marketing, 24*, 149–162.

Chapter 13
Analysis of Similarity and Preference Data

Similarity data in management research are typically collected in order to understand the underlying dimensions determining perceptions of stimuli such as brands or companies. One advantage of such data is that it is cognitively easier for respondents to provide subjective assessments of the similarity between objects than to rate these objects on a number of attributes that they may not even be aware of. Furthermore, when asking respondents to rate objects on attributes, the selection of the attributes proposed may influence the results while, in fact, it is not clear that these attributes are the relevant ones. In multidimensional scaling, the methodology allows you to infer the structure of perceptions. In particular, the researcher is able to make inferences regarding the number of dimensions that are necessary to fit the similarity data. In this chapter, we first describe the type of data collected to perform multidimensional scaling and we then present metric and nonmetric methods of multidimensional scaling. Multidimensional scaling explains the similarity of objects such as brands. We then turn to the analysis of preference data, where the objective is to model and explain preferences for objects. These explanations are based on the underlying dimensions of preferences that are discovered through the methodology.

13.1 Proximity Matrices

The input data for multidimensional scaling correspond to proximity or distance measures. Several types of measures exist, especially metric versus nonmetric and conditional versus unconditional.

13.1.1 Metric Versus Nonmetric Data

The data that serve as input to similarity analysis can be metric or nonmetric. Metric measures of proximity are ratio scales where zero indicates perfect similarity of two

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_13, © Springer Science+Business Media New York 2014

objects. The scale measures the extent to which the objects differ from each other. This measure of dissimilarity between objects is used as input to the method that consists in finding the underlying dimensions that discriminate between the objects to reproduce the dissimilarities (or similarities) between objects. In effect, these measures are distance measures (dissimilarity) or proximity measures (similarity), and the objective is to generate a map that shows the underlying distances between the objects.

Nonmetric data also reflect these proximity measures; however, only information about the rank order of the distances is available. As discussed in Chap. 1, special care must be taken with such data because most standard statistics such as means, standard deviations, and correlations are inappropriate.

13.1.2 Unconditional Versus Conditional Data

With unconditional data, all entries in the rows and columns are comparable, i.e., each stimulus is ranked relative to *all* other stimuli in the matrix (a number from 1 to $n(n-1)/2$ for nonmetric data).

If only the entries within a particular row are comparable, i.e., each of the n column stimuli is ranked relative to one row stimulus (a number from 1 to n for nonmetric data), the data are said to be conditional. In this case, the data matrix consists of $n-1$ objects ranked in terms of similarity relative to the row stimulus. Unconditional data are frequent, even though it is less cognitively complex for respondents to provide conditional data.

13.1.3 Derived Measures of Proximity

It should be noted that it may be possible to derive distance measures from data consisting of the evaluation of stimuli on attributes. However, it is not clear what attributes should be used and why other relevant ones may be missing. Furthermore, if the objective is to assess the underlying dimensions behind these attributes, multidimensional scaling will use the computed proximities as input and will ignore some of the information contained in the original attribute-level information. Consequently, information is lost when using such a procedure as compared to other procedures, for example principal component analysis. We, therefore, recommend that you reserve multidimensional scaling for direct measures of similarity rather than similarity measures derived from attribute-level data.

13.1.4 Alternative Proximity Matrices

Apart from the different categories of proximity data discussed in Sects. 13.1.1–13.1.3, the data matrix can take several specific forms.

13.1.4.1 Symmetric (Half) Matrix – Missing Diagonal (=0)

When dealing with distance measures, it is clear that the distance between objects *A* and *B* is the same as the distance between objects *B* and *A*. Therefore, when concerned with pure distance or proximity data, the full data are contained in half of the matrix, where the rows and the columns denote the objects and the cells represent the distance between these two objects. This matrix is symmetric. Furthermore, the diagonal represents the distance between an object and itself and, consequently, the elements of the diagonal are zeros (often they are not even included in the input).

13.1.4.2 Nonsymmetric Matrix – Missing Diagonal (=0)

In some cases, the matrix may not be symmetric. This is the case with confusion data, which consists in having each cell represent the frequency with which object *i* is matched with object *j* (for example with Morse codes, the percentage of times that a code of a particular letter is understood to be some other letter) or 1 minus that percentage. The greater the confusion, the greater the similarity between the two objects.

13.1.4.3 Nonsymmetric Matrix – Diagonal Present (≠0)

In the case of confusion data, the diagonal may not be zeros because a particular stimulus (e.g., a letter) may not be recognized all the time.

13.2 Problem Definition

In defining the problem, we consider nonmetric dissimilarity measures among *N* stimuli. We follow the definitions used in the KYST algorithm.

Let the table or matrix of dissimilarity (input data) be represented by

$$\underset{N\times N}{\Delta} = \{\delta(j,k)\} \tag{13.1}$$

where $\delta(j,k)$ is the dissimilarity between objects *j* and *k*, Δ is symmetric, and the diagonal cells are zero ($\delta(j,j) = 0$, for all *j*'s).

Although we do not know the dimensions of perceptions underlying these distance measures, let us assume that there are r such dimensions and that the stimuli are rated on these dimensions. Let $\underset{r\times 1}{\mathbf{x}_j}$ be the vector of coordinates of object j in the r-dimensional space. If, indeed, we knew these values $\underset{r\times 1}{\mathbf{x}_j}$ and r, then we would be able to compute the Euclidean distance between each pair of objects j and k:

$$\underset{1\times 1}{d^2}(j,k) = \left(\mathbf{x}_j - \mathbf{x}_k\right)'\left(\mathbf{x}_j - \mathbf{x}_k\right) = \sum_{\ell=1}^{r}\left(x_{j\ell} - x_{k\ell}\right)^2 \tag{13.2}$$

The problem is then defined as finding the x_j's such that the computed distances $d^2(j,k)$'s for all pairs are the closest to the actual dissimilarities $\delta\,(j,\,k)$'s.

13.2.1 Objective Function

Because the input data about the dissimilarities are not metric, the basic concept used here is to transform the rank-ordered dissimilarities through a monotonic function

$$f\left(\delta_{jk}\right) = d_{jk} \tag{13.3}$$

To reproduce the original dissimilarity data, the calculated Euclidean distance should lead to a rank order of these similarities as close as possible to the original or, equivalently, there should be a monotonic transformation of the rank-ordered dissimilarities that are as similar as possible to the computed distances. The differences between the monotonic transformation of the rank-ordered dissimilarities and the calculated dissimilarities are the error in the fit for each pair i,j:

$$f\left(\delta_{jk}\right) - d_{jk} \tag{13.4}$$

which, for all the pairs, gives the function to minimize:

$$\sum_j \sum_k \left[f\left(\delta_{jk}\right) - d_{jk}\right]^2 \tag{13.5}$$

This quantity in Eq. (13.5) is divided by a scaling factor, usually $\left(\Sigma_j \Sigma_k\, {d_{jk}}^2\right)$, in order to interpret the objective function relative to the distance values:

$$\frac{\Sigma_j \Sigma_k \left[f\left(\delta_{jk}\right) - d_{jk}\right]^2}{\textit{scale factor}} \tag{13.6}$$

13.2.2 Stress as an Index of Fit

Equation (13.6) provides the basis of the measure or index of fit of the model at the optimal level. This measure is called the stress and is obtained as:

$$Stress = \sqrt{\frac{\sum_{M=1}^{MM} [DIST(M) - DHAT(M)]^2}{\sum_{M=1}^{MM} [DIST(M) - DBAR]^2}} \tag{13.7}$$

where M = index for each object pair from 1 to MM ($=N^2$), $DIST(M)$ = computed distances from the solution of $\mathbf{x}_j$'s, $DHAT$ = predicted distances obtained from the monotonic regression of $DIST$ on the rank-ordered dissimilarity data, $DBAR$ = arithmetic average of the values of variable DIST.

The denominator enables the comparison across solutions with a different number of dimensions r.

Equation (13.7) can be rewritten as

$$Stress = \sqrt{\frac{\sum_{M=1}^{MM} \left[d_M - \hat{d}_M\right]^2}{\sum_{M=1}^{MM} \left[d_M - \overline{d}\right]^2}} \tag{13.8}$$

where

$$d_M = DIST(M)$$

$$\hat{d}_M = \hat{\beta}_0 + \hat{\beta}_1 \delta_M \tag{13.9}$$

$$\overline{d} = \frac{1}{MM} \sum_{M}^{MM} d_M \tag{13.10}$$

It is clear from Eqs. (13.7) and (13.8) that a stress of 0 indicates a perfect fit.

13.2.3 Metric

The discussion in Sects. 13.2.1 and 13.2.2 assumed Euclidean distance measures:

$$d_{ij} = \left[\sum_{k=1}^{r} \left(x_{ik} - x_{jk}\right)^2\right]^{1/2} \tag{13.11}$$

This is the most commonly used metric. However, it is also possible to use the Minkowski p-metric:

$$d_{ij}(p) = \left[\sum_{k=1}^{r} \left| x_{ik} - x_{jk} \right|^{p} \right]^{1/p} \quad p \geq 1 \tag{13.12}$$

The easiest case to interpret is for $p = 1$, which represents the city block metric. For $p = 2$, it is the Euclidean distance.

These different distance measures correspond to different ways of combining the information across the dimensions. They reflect differences in how perceptions are processed on individual dimensions to arrive at the perceived similarities/dissimilarities.

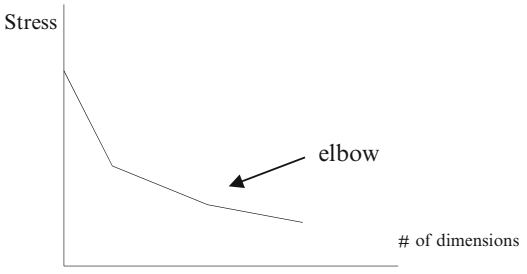
13.2.4 Minimum Number of Stimuli

A minimum number of data points (distances) are needed to be able to derive a space that can reproduce the distances. This number has been empirically assessed to be between four and six objects per dimension. Even though the researcher does not know a priori the number of dimensions, this means that a significant number of objects are needed to implement the methodology successfully. However, because the most typical solutions involve two or three dimensions, 12–18 objects should be sufficient in most cases.

13.2.5 Dimensionality

Because the number of dimensions r is not known a priori and because the solution for the x_j's depends on the number of dimensions, the dimensionality must be inferred from the results obtained for different values of r. Three criteria can be used together: the stress levels under different dimensionality assumptions, the stability of the results, and the interpretability of these solutions.

The goodness of fit or stress values can be plotted as a function of the number of dimensions (scree plot) to identify the elbow where adding dimensions produces little marginal gain in stress levels:



The stability of the results is typically assessed by splitting the sample into two and verifying that the results are similar for each subsample.

The interpretability of the results concerns the meaning of the dimensions of perception uncovered by the procedure. Although subjective, the ability to interpret the dimensions is essential for the research to be meaningful.

13.2.6 Interpretation of MDS Solution

The interpretation of the dimensions is mostly the fruit of the researcher's expertise. However, this expertise can benefit from a complementary data analysis when the objects have also been rated on a number of attributes (although this does considerably lengthen the task for the respondents). This analysis consists of property (attribute) fitting procedures. Three possible analyses are available:

(a) Maximum *r* procedure: This is based on the bivariate correlation coefficient of each attribute with a particular dimension. A high value of the correlation indicates a strong linear relationship between that attribute and the dimension. Consequently, this attribute would contribute significantly to the identification of the dimension.
(b) Monotone multiple regressions: A combination of attributes can explain the dimension in a nonlinear fashion. The R^2's provide a measure of the explanatory power.
(c) Property Fitting (PROFIT): This analysis provides for the possibility of non-monotonous relationships. The objective is to obtain a fit so that the stimulus projections are correlated with the scale.

13.2.7 The KYST Algorithm

Finding a solution, as described above, involves finding an initial configuration from which to start an iterative process and then determining the process by which to move from one iteration to the next.

Step 1: Finding initial configuration
Assume that the coordinates x_j's are centered at the origin (the means are zero). Let the n objects be identified by their coordinates in the p-dimensional space:

$$\underset{r\times n}{X} = \left(x_1, x_2, \cdots x_j, \cdots x_n\right) \tag{13.13}$$

$$\underset{n\times n}{B} = X'X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_j' \\ \vdots \\ x_n' \end{pmatrix} \left(x_1 x_2 \cdots x_j \cdots x_n\right) \tag{13.14}$$

$$= \begin{bmatrix} x_1'x_1 & x_1'x_2 & \cdots & x_1'x_n \\ \vdots & & \ddots & \\ x_n'x_1 & \cdots & \cdots & x_n'x_n \end{bmatrix} \tag{13.15}$$

$$x_j'x_k = \sum_{l=1}^{\gamma} x_{jl}x_{kl} \tag{13.16}$$

The principal component decomposition of Δ can provide the initial configuration with r eigenvectors or orthogonal dimensions.

Step 2: Improving configuration
In this step, the gradient of the stress provides the direction in which the solution should be changed to improve its value. For that purpose, the disparities between the actual dissimilarities and the predicted dissimilarities computed from the current iteration solution are calculated and the stress S is computed according to Eqs. (13.7) or (13.8). The gradient is computed from the changes in the stress from one iteration to the next relative to the changes in the coordinate values from the prior to the current iteration:

$$\frac{\partial S}{\partial x_{tn}} \quad \text{for } t = 1 \ldots r \tag{13.17}$$

The coordinate values x_{tj}'s are then modified in the direction of the gradient.

13.3 Individual Differences in Similarity Judgments

One way to recognize individual differences in perceptions is to allow all m subjects to share a common space, but to permit each individual to weight differently the dimensions of this space (which corresponds to the stretching and shrinking of the axes). This assumption is reflected in the INDSCAL algorithm.

Consequently, we denote the matrix of dissimilarities between objects for individual i as

$$\Delta^{(i)} = \left\{\delta^{(1)}(j,k)\right\} \quad \text{for } i = 1 \ldots m \tag{13.18}$$

where m is the number of individuals.

Each individual has a different weight for each dimension. These weights are represented by the diagonal matrix.

Let

$$\underset{r\times r}{\mathbf{W}^{(i)}} = diag\left\{w_t^{(i)}\right\} \tag{13.19}$$

The problem consists now in finding not only the coordinates of points in the common space but also the weights of each dimension for each individual so as to reproduce as much as possible the original dissimilarities:

$$\delta^{(i)^2}(j,k) \approx \left(\mathbf{x}_j - \mathbf{x}_k\right)' \mathbf{W}^{(i)} \left(\mathbf{x}_j - \mathbf{x}_k\right) \tag{13.20}$$

Wold's nonlinear iterative least squares procedure is used where, at each iteration, either $\mathbf{x}$ or $\mathbf{W}^{i)}$ is fixed to the last iteration estimate.

13.4 Analysis of Preference Data

In this section we no longer refer to modeling for the purpose of understanding the underlying dimensions of *perceptions*. Now the objective is to represent preferences for some stimuli over others.

Preferences follow from two basic models. One model predicts that more of any dimension is always preferred to less. This is the vector model of preferences. The other model assumes that "the more the better" is true only up to a certain point, after which too much is as bad as not enough. This assumption corresponds to the ideal point model of preferences.

13.4.1 Vector Model of Preferences

MDPREF is a model that derives the space where stimuli are represented in terms of preferences, as well as the individual differences in preference. Individuals are represented in a preference space by different vectors. Each vector is defined so that the projections of the stimuli (brands) on this vector correspond to this individual's preferences such that the more the projection falls in the direction of the vector, the more the stimulus is preferred. The stimuli are represented in the space by points such that the projections on the individual vector correspond as closely as possible to the stated preferences. In MDPREF, both the individual vectors and the stimuli points are inferred simultaneously from the preference data.

13.4.2 Ideal Point Model of Preferences

PREFMAP differs in two major ways from MDPREF. First, in MDPREF the individual vectors of preferences and the stimuli points are derived simultaneously from the preference data, but this is not the case in PREFMAP, where the stimuli configuration is provided externally. This configuration is obtained from the methods to derive a perceptual map from similarity data, which we described in Sects. 13.2 and 13.3. The results of KYST or INDSCAL can be used as input into this analysis of preferences.

The second way PREFMAP differs from MDPREF is that PREFMAP offers two models of preferences, a vector model as well as ideal point models. The vector model is similar to the model described above in the context of MDPREF. However, as already noted, the difference is that the stimuli points are externally supplied. The interpretation of the individual vectors is also similar to what is described above. However, the interpretation of the stimuli configuration is more easily accomplished, since the configuration corresponds to perceptions and not preferences. The joint space for representing perceptions and preferences also facilitates the interpretation of the individual vectors, since the dimensions are those derived from the perceptual analysis.

The ideal point model of preferences with PREFMAP is such that preferences for an individual are also represented as a point in the perceptual space. The preferences for stimuli are such that the most preferred are the stimuli that are the closest in that space to the point representing the individual ideal preference. The further away the stimuli are from the ideal point, the less preferred they are. PREFMAP derives the ideal point for each individual that best represents that person's preference. It should be noted that the vector model is a particular case of the ideal point model where the ideal point is located at infinity.

13.5 Examples

Examples of the various algorithms described above are now given using the PC-MDS software.

13.5.1 Example of KYST

Rank-ordered measures of dissimilarity between brands are the major input of KYST. The example input file is shown in Fig. 13.1.

The first line of the input file contains three numbers. The first number is the number of stimuli (here, 10 brands). The second number and the third number are for the number of replications and the number of groups (usually 1 each).

```
  10 1 1
  (9f3.0)
 22
 13 26
 01 25 36
 31 32 23 16
 44 18 14 02 30
 04 24 40 35 17 05
 07 27 38 42 19 06 34
 09 28 39 41 21 08 33 45
 37 20 15 03 29 43 12 10 11
sama
salt
semi
self
sibi
siro
sono
sold
suli
susi
```

Fig. 13.1 Example of PC-MDS input file for KYST (examp13-1.dat)

The second line is the format (Fortran-style) in which the data will be read.

The data matrix is then shown with 9 rows and 9 columns from the bottom half of a symmetric matrix without the diagonal (assumed to be zeros).

Finally, the stimuli (here, brands) labels are written on separate lines.

The output of KYST with this particular problem is shown in Fig. 13.2.

A two-dimensional solution was requested during the interactive dialog while running the software by indicating a minimum and a maximum number of dimensions of 2. The output shows the results by providing the stress obtained from that solution (a stress value of 0.266) and the coordinates in that two-dimensional space for the ten brands. The Shepard diagram represents the plot of the pairs of brands with the actual dissimilarity data on the y axis and the computed distances (before and after transformation through monotone regression). This shows how well the model replicates each of the pairs of stimuli. The plot of the brands in the two-dimensional space is shown, where the brands are numbered in the order of input. The interpretation can be inferred from the knowledge about the brands according to the attributes that appear to discriminate among these brands along the two dimensions found (here, an economy and a performance dimension). An example of PROFIT analysis to help interpret the meaning of the dimensions is shown next.

Nonmetric multidimensional scaling can be performed using STATA with the command "mdsmat" with specific parameters. The same data (ranking dissimilarity matrix) as above are read from the Excel spreadsheet Examp13-1.xlsx. The commands are shown in Fig. 13.3.

The data matrix is entered as a full symmetric matrix with "0" on the diagonal. Number 1 represents the pair that is the least dissimilar and so on. This matrix is shown in Fig. 13.4, although other options to input the matrix data are possible. The option "shape(x)" is used where x could be, for example, "upper" to indicate a row or column vector of dimension $n(n + 1)/2$ for the half-matrix cells, including the

```
                    K Y S T   MULTIDIMENSIONAL  SCALING
    WRITTEN BY  JOSEPH B. KRUSKAL, FOREST W. YOUNG, WITH JUDITH SEERY
                               PC-MDS  VERSION


ANALYSIS TITLE: KYST Rankings
DATA IS READ FROM FILE: ex_kystr.dat
OUTPUT FILE IS: ex_kystr.out


INPUT PARAMETERS:

MAXIMUM DIMENSIONS                                    2
MINIMUM DIMENSIONS                                    2
DIMENSION DECREMENT                                   1
MINIMUM STRESS                                         .01000
SCALE FACTOR GRADIENT                                  .00000
STRESS STEP RATIO                                      .99900
MAXIMUM ITERATIONS                                   50
COSINE OF ANGLE BETWEEN GRADIENTS                      .66000
AVERAGE COSINE OF ANGLE                                .66000
NUMBER OF PRE-ITERATIONS                              1
THE NUMBER OF DATA POINTS TO BE FIXED IS:              0
EUCLIDEAN DISTANCE
STRESS FORMULA 1
TIES PRIMARY
LOWER HALF MATRIX
NOT BLOCK DIAGONAL
DIAGONAL ABSENT
SPLIT BY DECK
TORSCA INITIAL CONFIGURATION
NO WEIGHTS AFTER DATA
MONOTONE MODEL
ASCENDING DATA
ALL PLOTS OF FINAL CONFIGURATION
ALL SCATTER PLOTS OF DIST VS DHAT
ROTATE FINAL CONFIG. COORDINATES


PARAMETERS:  10  1  1
TITLE:   (9f3.0)


DATA FOR RECORD:    10
 .37E+02 .20E+02 .15E+02 .30E+01 .29E+02 .43E+02 .12E+02 .10E+02 .11E+02

ON THE SHEPARD DIAGRAM THE ORIGINAL DATA (DATA) ARE PLOTTED;
ON THE Y AXIS AND DISTANCES (DIST,0) AND ESTIMATED DISTANCES
(DHAT,X) ON THE X AXIS. A ; INDICATES TWO VALUES ARE PLOTTED
ON TOP OF EACH OTHER AND A > INDICATES POINT NUMBERS GREATER
THAN 50.  IDENTIFIERS FOR THE CONFIGURATION PLOT IN 2 DIMENSIONS ARE:

    *****IDENTIFICATION KEY FOR PLOTS WITH IDENTIFIED POINTS*****

PT #   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
CHAR   1   2   3   4   5   6   7   8   9   A   B   C   D   E   F

PT #  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30
CHAR   G   H   I   J   K   L   M   N   O   P   Q   R   S   T   U

PT #  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45
CHAR   V   W   X   Y   Z   +   /   =   *   &   $   @   ]   -   <

PT #  46  47  48  49  50
CHAR   (   )   "   #   '


TITLE: KYST Rankings

INITIAL CONFIGURATION COMPUTATION   NO. PTS.=  10    DIM=   2
```

Fig. 13.2 PC-MDS output of KYST (examp13-1.out)

```
 STRESS STARTING TO INCREASE  BEST VALUE ACHIEVED ON PRE-ITERATION NUMBER  0


 THE BEST INITIAL CONFIGURATION OF  10 POINTS IN   2 DIMENSIONS
 HAS A STRESS OF    .401.  STRESS FORMULA 1 WAS USED.

 TITLE: KYST Rankings

 HISTORY OF COMPUTATION:
 N=  10 THERE ARE    45 DATA VALUES, SPLIT INTO   1 LIST(S).
 DIMENSION(S) =   2

 MINIMUM WAS ACHIEVED

 THE FINAL CONFIGURATION HAS BEEN ROTATED TO PRINCIPAL COMPONENTS.

 THE FINAL CONFIGURATION OF  10 POINTS IN  2 DIMENSIONS HAS STRESS OF  .266
 FORMULA 1 WAS USED.  THE FINAL CONFIGURATION APPEARS:

       1      2
  1 -1.007  -.210
  2   .162   .728
  3   .194  -.726
  4  -.992   .175
  5  -.030  -.009
  6  1.036   .854
  7   .715   .020
  8  1.055  -.830
  9  -.586  1.012
 10  -.546 -1.014

 DATA GROUP(S)
 SERIAL  COUNT STRESS REGRESSION COEFFICIENTS (FROM DEGREE 0 TO MAX OF 4)
     1     45   .266  ASCENDING

 ********************************************************************

 DIST AND DHAT VERSES DATA FOR   2 DIMENSION(S)
 STRESS =    .2662


              .    .5095.    .9675.   1.4255.   1.8835.   2.3415.
             .2805     .7385    1.1965    1.6545    2.1125    2.5705
            *.****.****.****.****.****.****.****.****.****.****.*
     47.20 ..                                                    .. 47.20
     45.41 ..                                                 X  .. 45.41
     43.61 ..                                              0 X0  .. 43.61
     41.82 ..              0        X                      X     .. 41.82
     40.03 ..              0        X                            .. 40.03
S    38.24 ..             0         X           0                .. 38.24
H    36.44 ..              0        X  0                         .. 36.44
E    34.65 ..              0        X       0                    .. 34.65
P    32.86 ..           0           X      0                     .. 32.86
A    31.07 ..                0      X                            .. 31.07
R    29.27 ..                   0   X0                           .. 29.27
D    27.48 ..            0         X          0                 .. 27.48
     25.69 ..                     0X  0                         .. 25.69
     23.90 ..              0      X                            .. 23.90
     22.10 ..           0         X   0                        .. 22.10
     20.31 ..                   0  X           0               .. 20.31
D    18.52 ..              0      X0                           .. 18.52
I    16.73 ..           0    0    X                            .. 16.73
A    14.93 ..            0        X                            .. 14.93
G    13.14 ..                     0X          0                .. 13.14
R    11.35 ..                      X     0        0            .. 11.35
A     9.56 ..                     0X     0                     ..  9.56
M     7.76 ..                      X     0           0         ..  7.76
      5.97 ..                      X       0                   ..  5.97
      4.18 ..              0       X         0                 ..  4.18
      2.39 ..                     0X                 0        ..  2.39
       .59 ..  X                                              ..   .59
     -1.20 ..                                                 .. -1.20
            *.****.****.****.****.****.****.****.****.****.****.*
             .    .5095.    .9675.   1.4255.   1.8835.   2.3415.
             .2805     .7385    1.1965    1.6545    2.1125    2.5705
```

Fig. 13.2 (continued)

```
CONFIGURATION PLOT  DIMENSION   2 (Y-AXIS) VS. DIMENSION   1 (X-AXIS)
KYST Rankings
       .*....*....*....*....*....*....*....*....*....*....*....*....*
 3.000**                                                           ** 3.000
 2.769**                                                           ** 2.769
 2.538**                                                           ** 2.538
 2.308**                                                           ** 2.308
 2.077**                                                           ** 2.077
 1.846**                                                           ** 1.846
 1.615**                                                           ** 1.615
 1.385**                                                           ** 1.385
 1.154**                                                           ** 1.154
  .923**                          9          6                     **  .923
  .692**                               2                           **  .692
  .462**                                                           **  .462
  .231**                        4                                  **  .231
  .000**-----------------------------5----7------------------------**  .000
 -.231**                      1                                    ** -.231
 -.462**                                                           ** -.462
 -.692**                               3                           ** -.692
 -.923**                          A          8                     ** -.923
-1.154**                                                           **-1.154
-1.385**                                                           **-1.385
-1.615**                                                           **-1.615
-1.846**                                                           **-1.846
-2.077**                                                           **-2.077
-2.308**                                                           **-2.308
-2.538**                                                           **-2.538
-2.769**                                                           **-2.769
-3.000**                                                           **-3.000
       .*....*....*....*....*....*....*....*....*....*....*....*....*
       .   -3.3333.  -2.0000.   -.6667.    .6667.   2.0000.   3.3333.
     -4.0000    -2.6667   -1.3333     .0000    1.3333    2.6667    4.0000
```

Fig. 13.2 (continued)

```
import excel "/Users/gatignon/Documents/WORK STATA/SAMD/Chapter13 MDS/examp13-1.xlsx",
sheet("Sheet1") clear
mkmat A-J , matrix(disimdata)
pause on
mdsmat disimdata, shape(full) names(sama salt semi self sibi siro sono sold suli susi)
loss(stress) transform(monotonic) config
mdsshepard
graph save "/Users/gatignon/Documents/WORK STATA/SAMD/Chapter13 MDS/mdsshepard",
replace
graph export "/Users/gatignon/Documents/WORK STATA/SAMD/Chapter13 MDS/mdsshepard.pdf",
replace
mdsconfig
graph save "/Users/gatignon/Documents/WORK STATA/SAMD/Chapter13 MDS/mdsconfig",
replace
graph export "/Users/gatignon/Documents/WORK STATA/SAMD/Chapter13 MDS/mdsconfig.pdf",
replace
```

Fig. 13.3 STATA commands for nonmetric MDS (examp13-1_Mac.do)

diagonals, or "uupper" to indicate a row or column vector of dimension $n(n-1)/2$ for the half-matrix cells excluding the diagonals. Note that "upper" and "uupper" would be replaced by "lower" and "llower" to indicate that the entries correspond to the lower half rather than the upper half of the matrix.

Fig. 13.4 Dissimilarity rank data for nonmetric MDS in examp13-1.xlsx

	A	B	C	D	E	F	G	H	I	J	K
1	0	22	13	1	31	44	4	7	9	37	
2	22	0	26	25	32	18	24	27	28	20	
3	13	26	0	36	23	14	40	38	39	15	
4	1	25	36	0	16	2	35	42	41	3	
5	31	32	23	16	0	30	17	19	21	29	
6	44	18	14	2	30	0	5	6	8	43	
7	4	24	40	35	17	5	0	34	33	12	
8	7	27	38	42	19	6	34	0	45	10	
9	9	28	39	41	21	8	33	45	0	11	
10	37	20	15	3	29	43	12	10	11	0	
11											

The command "mkmat A-J , matrix(disimdata)" converts the variables A through J into a matrix that will be called "disimdata." The "pause" command is used in order to save the graph before another one is produced. Typing "q" on the STATA command line and pressing the return key resumes the progress of the analysis.

The options "loss(stress) transform(monotonic)" correspond to the specification of nonmetric MDS. An alternative is to replace these two options by "method (nonmetric)" so that the full command line is equivalent: "mdsmat disimdata, shape (full) names(sama salt semi self sibi siro sono sold suli susi) method(nonmetric) config."

The option "names (*namelist*)" provides labels for the items being compared in the matrix. The name list should contain the same number of labels as the dimension of the matrix, separated by spaces.

Figure 13.5 displays the Shepard diagram.

The map with the configuration solution is shown in Fig. 13.6.

The log of the analysis is shown in Fig. 13.7 with the numerical values of the coordinates of each object (the ten brands analyzed as plotted in Fig. 13.6). These coordinates can then be added to a data set on evaluations of the objects on a number of variables (attributes) in order to calculate their correlations. These correlations provide important input for interpreting the meaning of the dimensions inferred from the MDS analysis.

13.5.2 Example of INDSCAL

In INDSCAL, the data for several individuals are analyzed. The input file of an example is shown in Fig. 13.8.

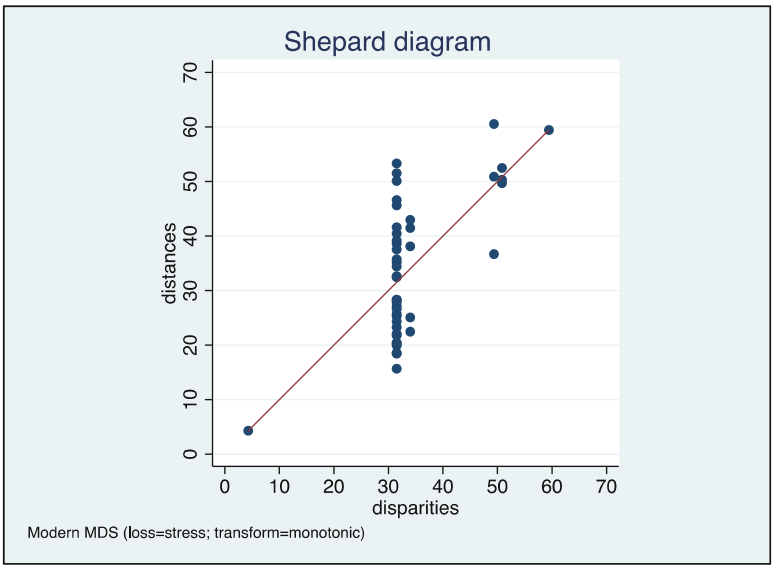


Fig. 13.5 STATA Shepard diagram for nonmetric MDS

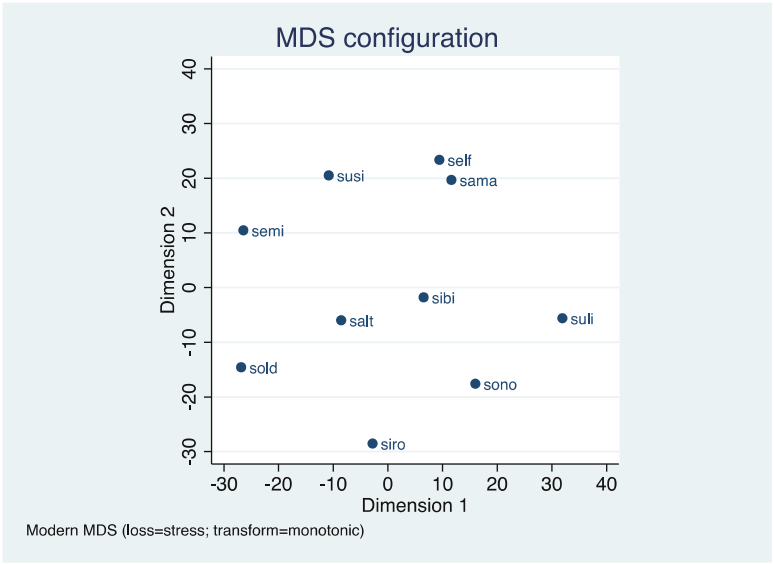


Fig. 13.6 STATA configuration map for nonmetric MDS

The first line of the input file contains the following information:

– Number of ways of the data (3-way data: # of brands x # of brands x # of subjects, indicating that the # of brands x # of brands matrix is repeated as many times as there are subjects)

```
. mdsmat disimdata, shape(full) names(sama salt semi self sibi siro sono sold suli
susi) loss(stress) transform(monotonic) config

Iteration 1t:   stress =  .40639834
Iteration 1c:   stress =  .33551015
Iteration 2t:   stress =  .32753733
Iteration 2c:   stress =  .29603901
…
Iteration 69t:   stress =  .26086989
Iteration 69c:   stress =  .26086989

Modern multidimensional scaling
    dissimilarity matrix: disimdata

    Loss criterion: stress = raw_stress/norm(distances)
    Transformation: monotonic (nonmetric)

                                                    Number of obs    =         10
                                                    Dimensions       =          2
    Normalization: principal                        Loss criterion   =     0.2609

Configuration in 2-dimensional Euclidean space (principal normalization)

    Category |         dim1          dim2
    ---------+----------------------------
        sama |      11.6145       19.6876
        salt |      -8.5461       -5.9820
        semi |     -26.4431       10.4797
        self |       9.4129       23.3648
        sibi |       6.5306       -1.7838
        siro |      -2.8052      -28.5256
        sono |      15.9945      -17.5831
        sold |     -26.8466      -14.5755
        suli |      31.9140       -5.6048
        susi |     -10.8256       20.5226
    -------------------------------------
```

Fig. 13.7 STATA nonmetric MDS results

- Maximum number of dimensions (2 in this example)
- Minimum number of dimensions (2 in this example)
- Type of input data (a value of 2 means a lower-half dissimilarity matrix without diagonal; other possibilities include a value of 1 for a lower-half similarity matrix without diagonal)
- Maximum number of iterations (25 were defined in this example)

The remaining codes on this first line correspond to more advanced options.

The second line contains a number for each way (brands and subjects). The first number indicates the number of subjects and the other two numbers give the number of stimuli.

The third line shows the format (Fortran-style) in which the data will be inputted.

The dissimilarity data are then shown for each individual (it is good practice to show the subject number first, although, as indicated by the format statement, this number is not read in).

Finally, the objects labels (brand names) are listed, one per line.

The results of INDSCAL are shown in Fig. 13.9.

The output, under the title "History of Computation," shows the fit measure at each iteration. Because INDSCAL is a metric model, the fit measure is the correlation between the input dissimilarity data and the predicted dissimilarity from the model parameter values at that iteration. The value of 0.999 obtained in the example is excellent.

```
   3   2   2   2  25   1   0   1   0   0  '12345677'  0   0   1   0  .001
   4 10 10
(2X,9F5.2)
01 4.88
01 4.07 0.93
01 5.33 0.62 1.27
01 2.89 1.99 1.24 2.47
01 0.51 5.38 4.56 5.83 3.39
01 3.67 1.37 0.44 1.69 0.94 4.16
01 5.40 0.61 1.34 0.13 2.53 5.90 1.77
01 5.38 0.59 1.33 0.13 2.51 5.88 1.76 0.02
01 0.69 5.56 4.73 5.99 3.57 0.19 4.32 6.06 6.05
02 5.65
02 6.37 2.98
02 7.84 3.52 1.54
02 3.28 2.38 3.97 5.16
02 0.63 6.10 6.58 8.08 3.77
02 6.74 3.95 0.99 1.87 4.70 6.86
02 7.42 2.78 1.48 0.77 4.57 7.70 2.17
02 7.36 2.71 1.47 0.84 4.51 7.65 2.19 0.07
02 1.18 6.18 6.35 7.87 3.93 0.65 6.54 7.55 7.51
03 4.34
03 5.08 2.45
03 6.22 2.92 1.19
03 2.51 1.84 3.27 4.20
03 0.49 4.67 5.21 6.37 2.88
03 5.44 3.25 0.80 1.42 3.90 5.49
03 5.84 2.30 1.13 0.64 3.69 6.03 1.68
03 5.79 2.24 1.12 0.69 3.64 5.98 1.70 0.06
03 0.95 4.71 4.98 6.16 2.99 0.54 5.20 5.87 5.83
04 2.42
04 4.86 2.89
04 5.63 3.56 0.80
04 1.27 1.17 3.86 4.59
04 0.34 2.33 4.63 5.41 1.25
04 5.68 3.79 0.90 0.60 4.73 5.43
04 4.91 2.78 0.46 0.78 3.83 4.70 1.15
04 4.84 2.71 0.47 0.85 3.76 4.64 1.20 0.07
04 0.96 2.04 4.06 4.85 1.19 0.64 4.84 4.16 4.10
sama
salt
semi
self
sibi
siro
sono
sold
suli
susi
```

Fig. 13.8 Example of PC-MDS input file for INDSCAL (examp13-2.dat)

Under the title "Normalized A Matrices," matrix 1 lists the individual weights for each of the four individuals. Matrix 2 lists the coordinates of the objects in the common object space.

The individual weights shown in matrix 1 are plotted along the two dimensions in the first plot. Plot no. 2 represents the brands corresponding to the coordinates listed in matrix 2.

Metric MDS is now illustrated using STATA. The data from each of the four individuals analyzed above have been entered in a separate spreadsheet for each individual. Therefore, the data from one individual are represented by a full symmetric dissimilarity matrix such as the one shown in Fig. 13.10.

The STATA commands for running a metric MDS are shown in Fig. 13.11.

The results contain the coordinates of the objects on the two dimensions, as shown in Fig. 13.12.

The perceptual map of the brands is displayed in a graph reproduced in Fig. 13.13.

```
                              I N D S C A L
                        INDIVIDUAL  DIFFERENCES  SCALING
                     BY DR. J. D. CARROLL AND JIH JIE CHANG
                              PC-MDS VERSION

ANALYSIS TITLE:              INDSCAL Example
DATA IS READ FROM FILE:      ex_inds.dat
OUTPUT FILE IS:              ex_inds.out


INDIFF- INDIVIDUAL DIFFERENCES ANALYSIS USING CANONICAL DECOMPOSITION
OF   3 WAY TABLE IN   2 DIMENSIONS

TITLE: INDSCAL Example
*******************************************************************************

PARAMETERS


NF        DIMENSION OF SOLUTION                                  2

N         NO. OF WAYS OR MATRICES                                3
MAXDIM    MAXIMUM NO. OF DIMENSIONS                              2
MINDIM    MINIMUM NO. OF DIMENSIONS                              2
IRDATA    TYPE OF DATA INPUT                                     2
ITMAX     MAXIMUM NO. OF ITERATIONS                             25
ISET      OPTION TO SET MATRIX 2 EQUAL TO MATRIX 3               1
IOY       SELECT SIMULTANEOUS SOLUTION                           0
IDR       CORRELATIONS FOR EACH SUBJECT                          1
ISAM      SOLVE FOR ALL MATRICES                                 0
IPUNSP    PUNCH SCALAR PRODUCT MATRICES                          0
IRN       RANDOM NUMBER GENERATOR START SET              12345677
CRIT      CRITERION FOR QUITTING ITERATION                    .001
IVEC      MATRIX OR VECTOR FORM FOR DATA                         0
IP        OUTPUT NORMALIZED A-MATRIX                             0
IA        PRINT ORIGINAL DATA MATRICES                           1
IS        PRINT INTERMEDIATE ITERATIVE MATRICES                  0


MATRIX SIZES     4  10  10
*******************************************************************************

   *****IDENTIFICATION KEY FOR PLOTS WITH IDENTIFIED POINTS*****

PT #   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
CHAR   1   2   3   4   5   6   7   8   9   A   B   C   D   E   F

PT #  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30
CHAR   G   H   I   J   K   L   M   N   O   P   Q   R   S   T   U

PT #  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45
CHAR   V   W   X   Y   Z   +   /   =   *   &   $   @   [   ?   <
```

Fig. 13.9 Output example for INDSCAL (examp13-2.out)

```
PT #  46  47  48  49  50
CHAR   (   )   "   ;   ]

POINT NUMBERS ABOVE 50 IDENTIFIED AS  > MULTIPLE POINTS IDENTIFIED AS  #


SUBJECT    1
    4.88
    4.07     .93
    5.33     .62    1.27
    2.89    1.99    1.24    2.47
     .51    5.38    4.56    5.83    3.39
    3.67    1.37     .44    1.69     .94    4.16
    5.40     .61    1.34     .13    2.53    5.90    1.77
    5.38     .59    1.33     .13    2.51    5.88    1.76     .02
     .69    5.56    4.73    5.99    3.57     .19    4.32    6.06    6.05

SUBJECT    4
    2.42
    4.86    2.89
    5.63    3.56     .80
    1.27    1.17    3.86    4.59
     .34    2.33    4.63    5.41    1.25
    5.68    3.79     .90     .60    4.73    5.43
    4.91    2.78     .46     .78    3.83    4.70    1.15
    4.84    2.71     .47     .85    3.76    4.64    1.20     .07
     .96    2.04    4.06    4.85    1.19     .64    4.84    4.16    4.10

INITIAL A MATRICES

MATRIX 1
 1      1.0000      1.0000      1.0000      1.0000
 2      1.0000      1.0000      1.0000      1.0000

MATRIX 2
 1       .4257      -.0724      -.1040       .4653      -.1853
        -.3849      -.0541       .3826      -.0469      -.3351

 2       .3026       .1942      -.3516      -.2383       .2954
         .3221       .3436      -.4229       .1126      -.3603


MATRIX 3
 1       .4448       .3780       .4900       .0394      -.4308
        -.2456      -.2815      -.4792      -.4867       .2676

 2      -.2278      -.4010      -.2592      -.1818       .3562
        -.1681       .1906      -.4663      -.3248       .2688


HISTORY OF COMPUTATION

ITERATION        CORRELATIONS BETWEEN
                  Y(DATA) AND YHAT         (R**2)            (1-R**2)
     0               -.021067             .000444            .999556
     1                .953993             .910103            .089897
     2                .984229             .968707            .031293
     3                .986800             .973774            .026226
     4                .990679             .981445            .018555
     5                .995783             .991585            .008415
     6                .998820             .997641            .002359
     7                .999428             .998857            .001143
     8                .999591             .999182            .000818
     9                .999690             .999380            .000620
********************************************************************************

EQUATE MATRIX 2 AND MATRIX 3, ITERATE AGAIN

INITIAL A MATRICES

MATRIX 1
 1      -.1499      -.1080      -.1020      -.0334
 2      -.0194       .1066       .1212       .2540

MATRIX 2
```

Fig. 13.9 (continued)

```
 1      1.1527      -.6224      -.4095      -.8760       .1087
        1.3216      -.2866      -.8801      -.8722      1.3638

 2       .3719       .1729      -.2487      -.3310       .2871
         .3238      -.3798      -.2185      -.2089       .2312


MATRIX 3
 1      1.1527      -.6224      -.4095      -.8760       .1087
        1.3216      -.2866      -.8801      -.8722      1.3638

 2       .3719       .1729      -.2487      -.3310       .2871
         .3238      -.3798      -.2185      -.2089       .2312

HISTORY OF COMPUTATION

ITERATION         CORRELATIONS BETWEEN
                   Y(DATA) AND YHAT          (R**2)             (1-R**2)
     0                -.795407              .632673             .367327
     1                 .999731              .999463             .000537

INDSCAL Example

NORMALIZED A MATRICES

MATRIX   1
 1        1.03187     -.05535
 2         .73697      .36435
 3         .69485      .41314
 4         .21598      .85421

MATRIX   2
 1         .41044      .41167
 2        -.22162      .19138
 3        -.14581     -.27529
 4        -.31193     -.36649
 5         .03871      .31787
 6         .47060      .35844
 7        -.10205     -.42043
 8        -.31338     -.24193
 9        -.31057     -.23122
10         .48561      .25599

MATRIX   3
 1         .41044      .41167
 2        -.22162      .19138
 3        -.14581     -.27529
 4        -.31193     -.36649
 5         .03871      .31787
 6         .47060      .35844
 7        -.10205     -.42043
 8        -.31338     -.24193
 9        -.31057     -.23122
10         .48561      .25599

MATRIX   1

SUMS OF PRODUCTS
 1        2.13736      .68297
 2         .68297     1.03618

SUM OF SQUARES =       3.17353

MATRIX   2

SUMS OF PRODUCTS
 1        1.00000      .77684
 2         .77684     1.00000

SUM OF SQUARES =       2.00000

MATRIX   3

SUMS OF PRODUCTS
 1        1.00000      .77684
 2         .77684     1.00000
```

Fig. 13.9 (continued)

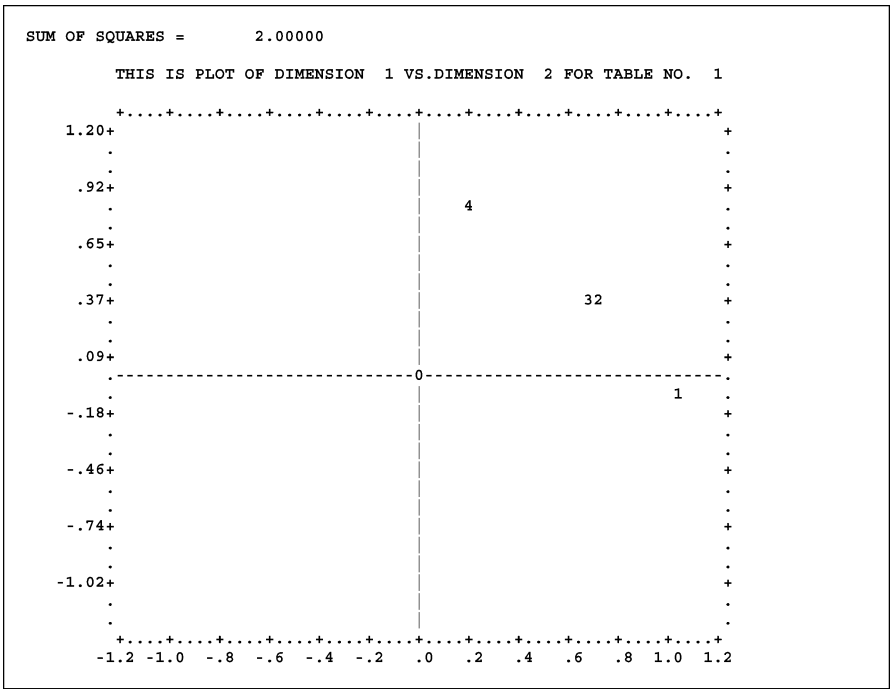


Fig. 13.9 (continued)

13.5.3 Example of PROFIT (Property Fitting) Analysis

In the example shown in Fig. 13.14, we use the configuration (coordinates) obtained from the KYST analysis described earlier in Sect. 13.5.1. (It is possible to use the output configuration of other models such as INDSCAL) . The relationships of the two dimensions corresponding to these perceptions of the ten brands with five characteristics of the brands (i.e., weight, design, volume, maximum frequency, and power) are analyzed in this run of PROFIT. Therefore, the ratings of these brands on these characteristics are matched as well as possible with the ratings obtained from the KYST configuration. Each characteristic is represented in the perceptual space by a vector so that the fit with the perceptions of the brands is maximized. For rating data on the properties (brand characteristics), the correlation between these ratings and the projection of the brand perceptions on that vector is maximized.

The input file shown in Fig. 13.14 provides the information needed to run the program. The first line of input indicates the basic parameters of the problem. The first number (1 in Fig. 13.14) indicates that a linear relationship between properties and perceptions will be evaluated. The second number (10 in Fig. 13.14) indicates

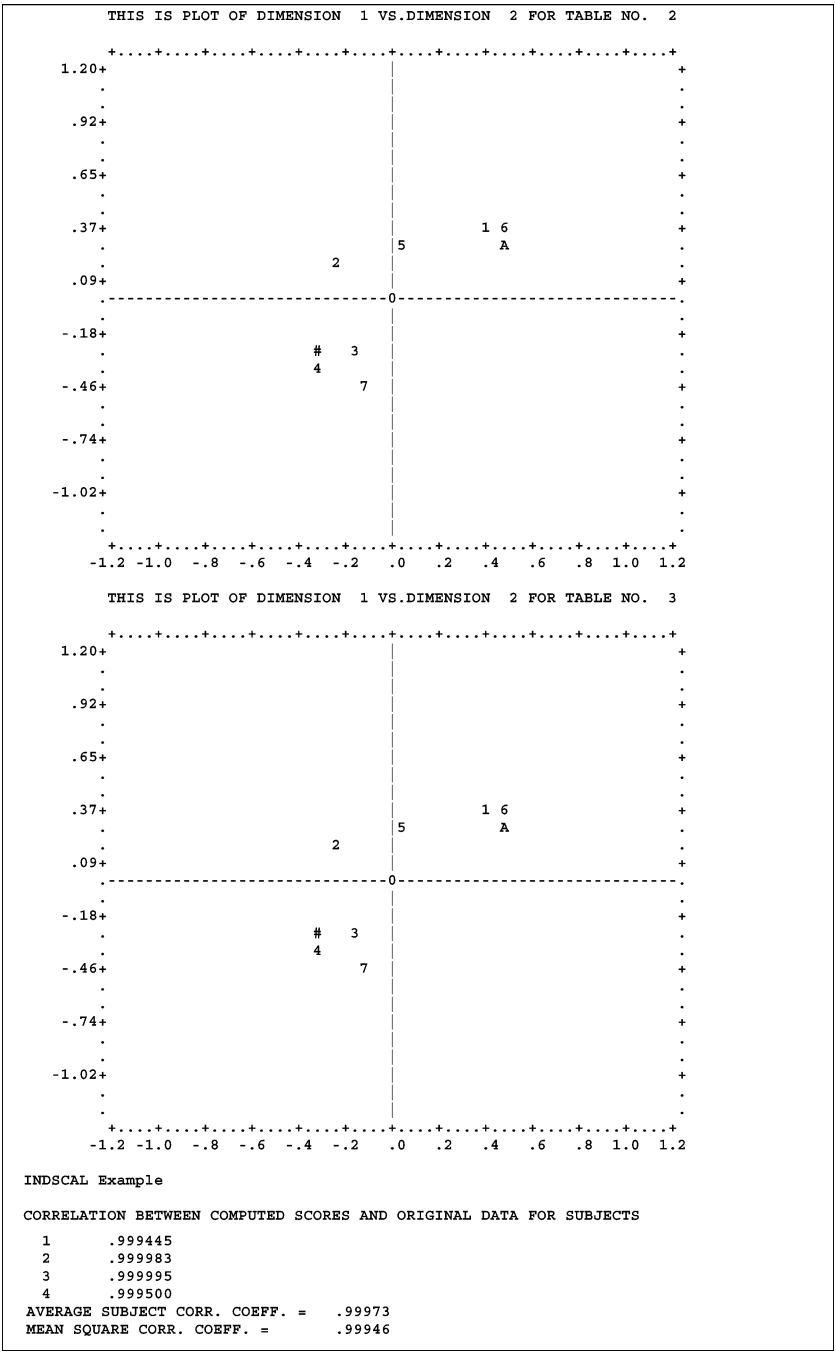


Fig. 13.9 (continued)

	A	B	C	D	E	F	G	H	I	J	K
1	0	5.65	6.37	7.84	3.28	0.63	6.74	7.42	7.36	1.18	
2	5.65	0	2.98	3.52	2.38	6.1	3.95	2.78	2.71	6.18	
3	6.37	2.98	0	1.54	3.97	6.58	0.99	1.48	1.47	6.35	
4	7.84	3.52	1.54	0	5.16	8.08	1.87	0.77	0.84	7.87	
5	3.28	2.38	3.97	5.16	0	3.77	4.7	4.57	4.51	3.93	
6	0.63	6.1	6.58	8.08	3.77	0	6.86	7.7	7.65	0.65	
7	6.74	3.95	0.99	1.87	4.7	6.86	0	2.17	2.19	6.54	
8	7.42	2.78	1.48	0.77	4.57	7.7	2.17	0	0.07	7.55	
9	7.36	2.71	1.47	0.84	4.51	7.65	2.19	0.07	0	7.51	
10	1.18	6.18	6.35	7.87	3.93	0.65	6.54	7.55	7.51	0	
11											

Fig. 13.10 Data example for metric MDS in STATA (examp13-2-02.xlsx)

```
import excel "/Users/gatignon/Documents/WORK STATA/SAMD/Chapter13 MDS/Examp13-2-
02.xlsx", sheet("Sheet1") clear
mkmat A-J , matrix(disimdata)
mdsmat disimdata, shape(full) names(sama salt semi self sibi siro sono sold suli susi)
method(modern) config
```

Fig. 13.11 STATA commands for metric MDS (examp13-2.do)

```
. mdsmat disimdata, shape(full) names(sama salt semi self sibi siro sono sold suli
susi) method(modern) config
(loss(stress) assumed)
(transform(identity) assumed)

Iteration 1:   stress =  .00047418
Iteration 2:   stress =  .00043706
Iteration 3:   stress =  .00043194
Iteration 4:   stress =  .00043107
Iteration 5:   stress =   .0004309
Iteration 6:   stress =  .00043086
Iteration 7:   stress =  .00043085
Iteration 8:   stress =  .00043085

Modern multidimensional scaling
    dissimilarity matrix: disimdata

    Loss criterion: stress = raw_stress/norm(distances)
    Transformation: identity (no transformation)

                                                    Number of obs     =         10
                                                    Dimensions        =          2
    Normalization: principal                        Loss criterion    =     0.0004

Configuration in 2-dimensional Euclidean space (principal normalization)

    Category |          dim1          dim2
    ---------+----------------------------
        sama |        4.3397        0.2065
        salt |       -0.9941        2.0654
        semi |       -1.9626       -0.7499
        self |       -3.4722       -0.4374
        sibi |        1.3175        1.4853
        siro |        4.6043       -0.3630
        sono |       -2.1168       -1.7247
        sold |       -3.0755        0.2229
        suli |       -3.0222        0.2693
        susi |        4.3820       -0.9744
    ----------------------------------------
```

Fig. 13.12 STATA output for metric MDS (examp13-2.log)

the number of stimuli (brands). The third number (2 in Fig. 13.14) shows the number of dimensions in the perceptual space used as input. The fourth number (5 in Fig. 13.14) is the number of properties to be analyzed. The other numbers correspond to more advanced options.

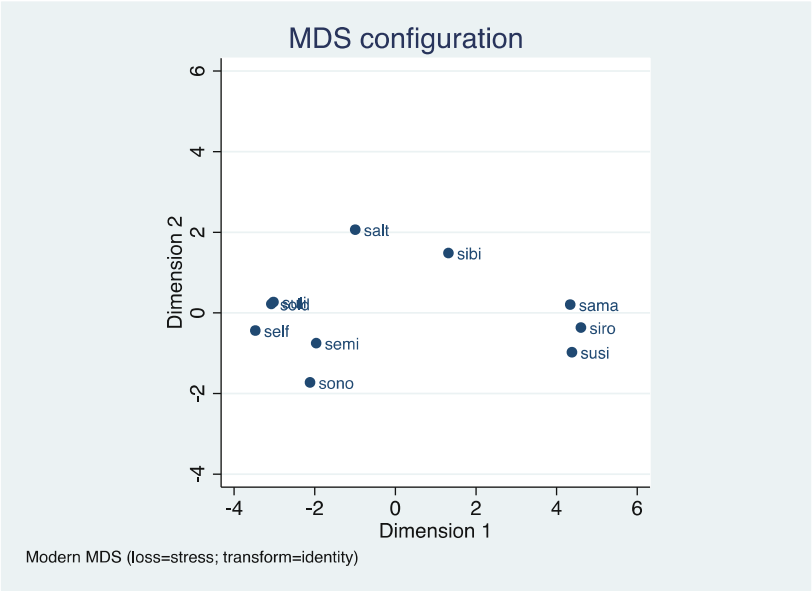


Fig. 13.13 Perceptual map from metric MDS

```
1 10 2 5 0    0    2   0.0
(2X,2F7.3)
 1 -1.007  -.210
 2   .162   .728
 3   .194  -.726
 4  -.992   .175
 5  -.030  -.009
 6  1.036   .854
 7   .715   .020
 8  1.055  -.830
 9  -.586  1.012
10  -.546 -1.014
(2X,10F3.0)
Weight
01 10 12 17 15 11 10 10 17 10 15
Design
02 08 09 07 05 09 03 03 07 03 06
Volume
03 30 37 50 60 35 50 70 50 50 40
Max Frequency
04 25 25 30 40 25 20 20 30 25 20
Power
05 10 30 80 90 20 10 90 70 20 70
sama
salt
semi
self
sibi
siro
sono
sold
suli
susi
```

Fig. 13.14 Example of PC-MDS input file for PROFIT (examp13-3.dat)

The second line is the format (Fortran-style) in which the data for the stimuli (brands) coordinates are read. Then follow the perception coordinates, one line for each stimulus (brand). In this example, the stimulus number (1 − 10) is shown in order to better visualize the input, but this information is not read by the program, since the format above indicates that the first two columns are skipped ("2X"). After the perceptual coordinates, the data on the properties are shown. First, the format in which the data are to be read is indicated in the usual Fortran-style format. Then, for each of the properties, the label of the property is shown on one line and on a separate line the values of the property on all ten stimuli are shown. The first number indicates the property number but is not used, as shown by the format of the input, which, as noted above, skips the first two columns of data. Finally, the last ten lines correspond to the labels of the ten stimuli, in this case the names of the brands.

Figure 13.15 shows the output of the PROFIT analysis. First, for each property, the correlations between the original and the fitted vectors are shown, followed by the corresponding plot of the stimuli.

The last graph in Fig. 13.15 shows the perceptions of the stimuli (the ten brands) numbered from 1 to 9, plus the letter A to represent the tenth brand. The points labeled B to F represent the end points of the property vectors that maximize the correlation with the projections of the brands on this vector with the original property values. Note that the vectors have been added in the figure and do not appear on the original computer output. B represents the weight property, C the design, D the volume, E the maximum frequency, and F the power of the brands.

```
                              P R O F I T
                         PROPERTY FITTING ANALYSIS
              PROGRAM WRITTEN BY DR. J. D. CARROLL AND JIH JIE CHANG
                              PC-MDS VERSION

ANALYSIS TITLE: Profit test
 DATA IS READ FROM FILE: ex_prof.dat
 OUTPUT FILE IS: ex_prof.out
 LANA (REGRESSION OPTION):                                          1
 N    NO. OF STIMULI   (400 MAX)                                   10
 K    NO. OF DIMENSIONS (10 MAX)                                    2
 M    NO. OF PROPERTIES (60 MAX)                                    5
 IRX  0 = N X K INPUT; 1 = K X N INPUT                              0
 IWGT 0 = RATIO OF ERROR VAR. TO TRUE VAR. (USUAL OPTION)           0
      1 = RATIO OF MEAN SQ. SUCCESSIVE DIFFERENCE TO VARIANCE
IPLOT 0 = PROPERTIES ONLY                                           2
      1 = PLOT PROPERTIES AND FUNCTIONS
      2 = DO ALL PLOTS
 BCO (FLOATING POINT NUMBER FOR NON LINEAR REG.)                    0.


 DATA FOR RECORD:     1
 -.10E+01-.21E+00


 DATA FOR RECORD:    10
 -.55E+00-.10E+01

 LINEAR REGRESSION
```

Fig. 13.15 Output example of PROFIT analysis (examp13-3.out)

```
NORMALIZED CONFIGURATION
  1        -1.0071       .1619       .1939      -.9921      -.0301
            1.0359       .7149      1.0549      -.5861      -.5461

  2         -.2100       .7280      -.7260       .1750      -.0090
             .8540       .0200      -.8300      1.0120     -1.0140


COVARIANCE MATRIX
  1       5.4020    -.0008
  2       -.0008    4.6028

X*(X''X) INVERSE
  1       -.1864     .0300     .0359    -.1837    -.0056     .1918     .1323
           .1953    -.1085    -.1011
  2       -.0457     .1582    -.1577     .0380    -.0020     .1856     .0044
          -.1803     .2198    -.2203

          PROPERTY  1


INTERMEDIATE SUMS BEFORE SQUARING:       .2484     -2.9634

SSQ =       8.84321   XL =       2.97375

ORIGINAL VALUES ON PROPERTY  1
          10.0000     12.0000     17.0000     15.0000     11.0000
          10.0000     10.0000     17.0000     10.0000     15.0000


PROJECTIONS ON FITTED VECTORS
           .1251      -.7119       .7397      -.2573       .0065
          -.7645       .0398       .9152     -1.0574       .9648



 PLOT OF ORIGINAL (X-AXIS) VERSUS OBTAINED (Y-AXIS) FOR PROPERTY VECTOR NO.  1

        +.........+.........+.........+.........+.........+.........+
  1.066+                                                           +
       .                                          A                .
       .                                                       8   .
   .819+                                                           +
       .                                                       3   .
       .                                                           .
   .572+                                                           +
       .                                                           .
       .                                                           .
   .324+                                                           +
       .                                                           .
       .   1                                                       .
   .077+   7                                                       +
       .           5                                               .
       .                                                           .
  -.170+                                                           +
       .                                          4                .
       .                                                           .
  -.417+                                                           +
       .                                                           .
       .                                                           .
  -.664+                                                           +
       .   6               2                                       .
       .                                                           .
  -.911+                                                           +
       .                                                           .
       .   9                                                       .
 -1.159+                                                           +
        +.........+.........+.........+.........+.........+.........+
     9.6510.2910.9311.5712.2212.8613.5014.1414.7815.4216.0716.7117.35


CORRELATION BETWEEN ORIGINAL AND FITTED VECTORS FOR PROPERTY   1 IS:

         R =     .713 , RSQ =     .509
```

Fig. 13.15 (continued)

```
          PROPERTY  2


INTERMEDIATE SUMS BEFORE SQUARING:      -.5318     -1.2280

SSQ =       1.79091   XL =       1.33825

ORIGINAL VALUES ON PROPERTY  2
           8.0000      9.0000      7.0000      5.0000      9.0000
           3.0000      3.0000      7.0000      3.0000      6.0000


PROJECTIONS ON FITTED VECTORS
            .5929      -.7324       .5891       .2337       .0202
          -1.1953      -.3025       .3424      -.6957      1.1475


 PLOT OF ORIGINAL (X-AXIS) VERSUS OBTAINED (Y-AXIS) FOR PROPERTY VECTOR NO.  2

        +.........+.........+.........+.........+.........+.........+
  1.265+                                                           +
       .                              A                            .
       .                                                           .
   .978+                                                           +
       .                                                           .
       .                                                           .
   .692+                                                           +
       .                                        3        1         .
       .                                                           .
   .406+                                                           +
       .                                        8                  .
       .                      4                                    .
   .119+                                                           +
       .                                                      5    .
       .                                                           .
  -.167+                                                           +
       .   7                                                       .
       .                                                           .
  -.453+                                                           +
       .                                                           .
       .                                                           .
  -.740+   9                                                   2   +
       .                                                           .
       .                                                           .
 -1.026+                                                           +
       .                                                           .
       .   6                                                       .
 -1.312+                                                           +
        +.........+.........+.........+.........+.........+.........+
      2.70 3.25 3.80 4.35 4.90 5.45 6.00 6.55 7.10 7.65 8.20 8.75 9.30


CORRELATION BETWEEN ORIGINAL AND FITTED VECTORS FOR PROPERTY   2 IS:

         R =     .404 , RSQ =     .163
```

Fig. 13.15 (continued)

This graph indicates that the *Y* dimension (dimension 2) is closely related to Weight and Power and also, although not as strongly, to Design (the higher the values of the properties, the lower the perceptions on that dimension). The *X* dimension (dimension 1) reflects more the volume, which appears to be negatively correlated with the maximum frequency. Therefore, in general, the higher the perceptual value on dimension 1, the higher the volume but the lower the maximum frequency. It should be noted that these correlations can be used to help the

```
          PROPERTY  3


INTERMEDIATE SUMS BEFORE SQUARING:      5.2440      1.5567

SSQ =      29.92298   XL =       5.47019

ORIGINAL VALUES ON PROPERTY  3
          30.0000     37.0000     50.0000     60.0000     35.0000
          50.0000     70.0000     50.0000     50.0000     40.0000


PROJECTIONS ON FITTED VECTORS
          -1.0252       .3624      -.0207      -.9013      -.0314
           1.2361       .6910       .7751      -.2739      -.8121


 PLOT OF ORIGINAL (X-AXIS) VERSUS OBTAINED (Y-AXIS) FOR PROPERTY VECTOR NO.  3

         +.........+.........+.........+.........+.........+.........+
   1.349+                                                            +
        .                               6                           .
        .                                                           .
   1.073+                                                            +
        .                                                           .
        .                                                           .
    .796+                               8                           +
        .                                                       7   .
        .                                                           .
    .520+                                                           +
        .                                                           .
        .             2                                             .
    .244+                                                           +
        .                                                           .
        .                                                           .
   -.033+          5                    3                         +
        .                                                           .
        .                                                           .
   -.309+                               9                          +
        .                                                           .
        .                                                           .
   -.586+                                                          +
        .                                                           .
        .                 A                                         .
   -.862+                                                 4         +
        .                                                           .
        .   1                                                       .
  -1.138+                                                           +
         +.........+.........+.........+.........+.........+.........+
      28.0031.6735.3339.0042.6746.3350.0053.6757.3361.0064.6768.3372.00


CORRELATION BETWEEN ORIGINAL AND FITTED VECTORS FOR PROPERTY   3 IS:

         R =     .348 , RSQ =     .121
```

Fig. 13.15 (continued)

interpretation of the dimensions. However, the dimensions and the properties do not coincide perfectly. For example, although the vectors B and F are particularly close to axis *Y*, axis *X* is not very close to either vector D or E. Consequently, the property fitting analysis will not be as useful in interpreting the *X* axis as it will be for the *Y* axis.

Similar information can be obtained with STATA. First, a spreadsheet should be created containing the information on the properties and the coordinates on the

```
          PROPERTY  4

INTERMEDIATE SUMS BEFORE SQUARING:     -2.7142      -.9683

SSQ =       8.30432   XL =       2.88172

ORIGINAL VALUES ON PROPERTY  4
          25.0000     25.0000     30.0000     40.0000     25.0000
          20.0000     20.0000     30.0000     25.0000     20.0000


PROJECTIONS ON FITTED VECTORS
           1.0191      -.3971       .0613       .8756       .0314
          -1.2626      -.6801      -.7147       .2120       .8551


 PLOT OF ORIGINAL (X-AXIS) VERSUS OBTAINED (Y-AXIS) FOR PROPERTY VECTOR NO.  4

         +.........+.........+.........+.........+.........+.........+
  1.133+                                                             +
       .                1                                            .
       .                                                             .
   .854+   A                                                     4   +
       .                                                             .
       .                                                             .
   .575+                                                             +
       .                                                             .
       .                                                             .
   .297+                                                             +
       .                9                                            .
       .                                                             .
   .018+                5             3                              +
       .                                                             .
       .                                                             .
  -.261+                                                             +
       .                2                                            .
       .                                                             .
  -.540+                                                             +
       .                                                             .
       .   7                         8                               .
  -.819+                                                             +
       .                                                             .
       .                                                             .
 -1.098+                                                             +
       .                                                             .
       .   6                                                         .
 -1.377+                                                             +
         +.........+.........+.........+.........+.........+.........+
      19.0020.8322.6724.5026.3328.1730.0031.8333.6735.5037.3339.1741.00


CORRELATION BETWEEN ORIGINAL AND FITTED VECTORS FOR PROPERTY   4 IS:

          R =      .360 , RSQ =      .130
```

Fig. 13.15 (continued)

dimensions from the configuration obtained from MDS. Figure 13.16 shows such a file.

Figure 13.17 provides the commands to compute the correlations among the dimensions obtained from the nonmetric MDS and the five vectors of properties.

The correlations that result from this analysis are shown in Fig. 13.18, as well as a plot for one of the dimension-attribute combinations (i.e., Dim1-Weight).

```
          PROPERTY  5


INTERMEDIATE SUMS BEFORE SQUARING:      3.5137    -26.3463

SSQ =     706.47380   XL =      26.57957

ORIGINAL VALUES ON PROPERTY  5
          10.0000     30.0000     80.0000     90.0000     20.0000
          10.0000     90.0000     70.0000     20.0000     70.0000


PROJECTIONS ON FITTED VECTORS
            .0750      -.7002       .7453      -.3046       .0049
           -.7096       .0747       .9622     -1.0806       .9329


 PLOT OF ORIGINAL (X-AXIS) VERSUS OBTAINED (Y-AXIS) FOR PROPERTY VECTOR NO.  5

         +.........+.........+.........+.........+.........+.........+
  1.064+                                                             +
       .                                          8                  .
       .                                          A                  .
   .815+                                                             +
       .                                                 3           .
       .                                                             .
   .565+                                                             +
       .                                                             .
       .                                                             .
   .315+                                                             +
       .                                                             .
       .                                                             .
   .066+   1                                                     7   +
       .          5                                                  .
       .                                                             .
  -.184+                                                             +
       .                                                         4   .
       .                                                             .
  -.434+                                                             +
       .                                                             .
       .                                                             .
  -.683+   6             2                                          +
       .                                                             .
       .                                                             .
  -.933+                                                             +
       .                                                             .
       .          9                                                  .
 -1.183+                                                             +
         +.........+.........+.........+.........+.........+.........+
       6.0013.3320.6728.0035.3342.6750.0057.3364.6772.0079.3386.6794.00


CORRELATION BETWEEN ORIGINAL AND FITTED VECTORS FOR PROPERTY   5 IS:

         R =     .563 , RSQ =     .317
```

Fig. 13.15 (continued)

13.5.4 Example of MDPREF

The first row in the input file shown in Fig. 13.19 defines:

- The number of rows in the data matrix, or number of subjects (here, 5)
- The number of columns in the data matrix, or number of stimuli (here, 10 brands)

```
TABLE 1.  THE MAXIMUM CORRELATION BETWEEN THE PROPERTY
          AND THE PROJECTIONS ON FITTED VECTOR

           RHO          PROPERTY
   1          .7133    Weight
   2          .4035    Design
   3          .3484    Volume
   4          .3602    Max Frequency
   5          .5630    Power

TABLE 2.  DIRECTION COSINES OF FITTED VECTORS
           IN NORMALIZED SPACE

                     DIMENSION

VECTOR    1         2
     1       .0835    -.9965
     2      -.3974    -.9176
     3       .9587     .2846
     4      -.9419    -.3360
     5       .1322    -.9912

TABLE 3.  COSINE OF ANGLES BETWEEN VECTORS

VECTOR:      1       2       3       4
    2       .881
    3      -.203   -.642
    4       .256    .683   -.999
    5       .999    .857   -.155    .209
```

Fig. 13.15 (continued)

- Number of dimensions (here, 2)
- Number of dimensions to be plotted (here, 2)
- A code to normalize by subtracting the row mean (=1) or to normalize and divide by the standard deviation (=2)
- A dummy code to normalize subject vectors (=1; 0 otherwise)

The second line defines the format in which the preference data are read, followed by the data themselves. The first number of each row is the subject number, which is not read by the program, as indicated by the format statement starting with "2X." For each row (subject), the ten numbers indicate the values given by the subject to each of the ten brands.

The lines in Fig. 13.19 are used for the labels of the subjects and then of the stimuli.

In Fig. 13.20, the first graph in the output file maps the subject vectors starting at the origin with the end point at the location of the number corresponding to the subject. The second graph maps the stimuli according to the subjects' preferences, while the third graph shows both the subject vectors and the stimuli points at the same time. This plot of the brands should be carefully interpreted, given that it does not correspond to perceptual data but is derived solely from input on preferences.

In the graphs shown in Fig. 13.20, the vectors have been added to the original output. The projections of the stimuli on a particular subject vector indicate the preferences of that individual subject. For example, subject 1 (indicated by the letter B in the figure) has a preference for brands 3 (SEMI) and 7 (SONO). Subject

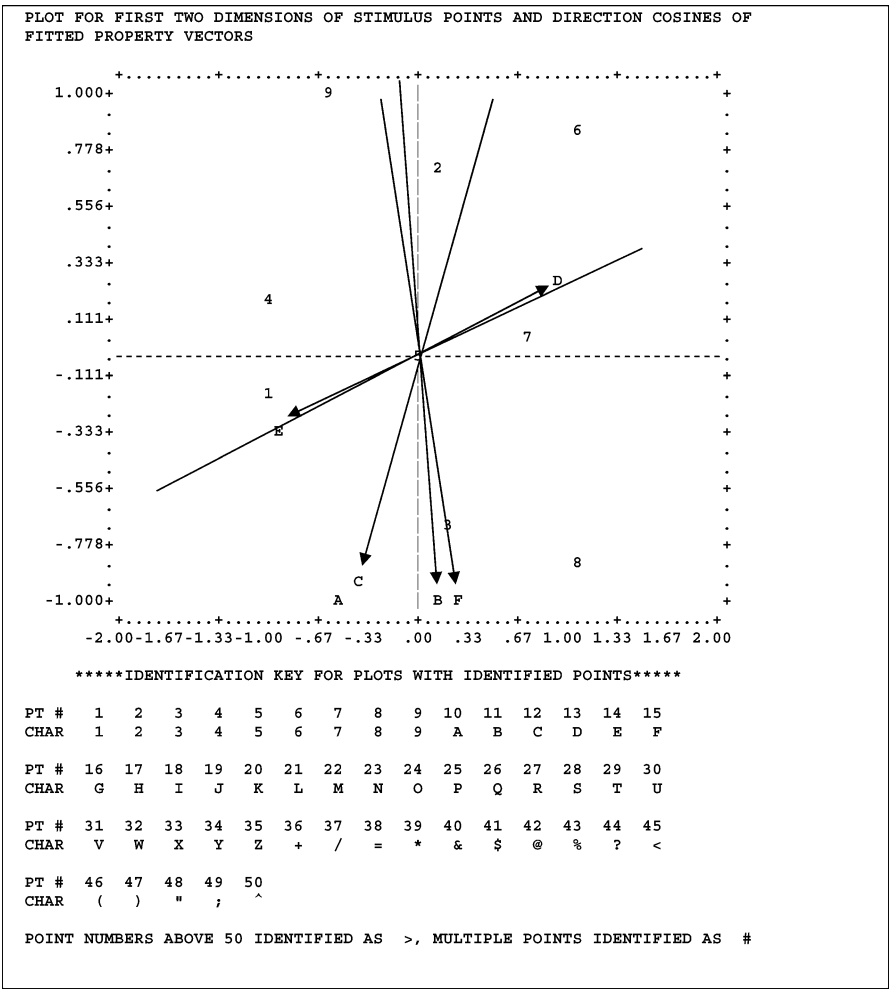


Fig. 13.15 (continued)

	A	B	C	D	E	F	G	H	I
1	Brand	Dim1	Dim2	Weight	Design	Volume	MaxFreq	Power	
2	sama	11.6145	19.6876	10	8	30	25	10	
3	salt	-8.5461	-5.982	12	9	37	25	30	
4	semi	-26.4431	10.4797	17	7	50	30	80	
5	self	9.4129	23.3648	15	5	60	40	90	
6	sibi	6.5306	-1.7838	11	9	35	25	20	
7	siro	-2.8052	-28.5256	10	3	50	20	10	
8	sono	15.9945	-17.5831	10	3	70	20	90	
9	sold	-26.8466	-14.5755	17	7	50	30	70	
10	suli	31.914	-5.6048	10	3	50	25	20	
11	susi	-10.8256	20.5226	15	6	40	20	70	
12									

Fig. 13.16 Data in Excel for correlation analysis for use in STATA (examp13-3.xlsx)

```
import excel "/Users/gatignon/Documents/WORK STATA/SAMD/Chapter13 MDS/examp13-3.xlsx",
sheet("Sheet1") firstrow clear
correlate Dim1-Power
pause on
twoway (scatter Weight Dim1), title(Weight vs. Dim1)
pause
twoway (scatter Design Dim1), title(Design vs. Dim1)
pause
twoway (scatter Volume Dim1), title(Volume vs. Dim1)
pause
twoway (scatter MaxFreq Dim1), title(MaxFreq vs. Dim1)
pause
twoway (scatter Power Dim1), title(Power vs. Dim1)
pause
twoway (scatter Weight Dim2), title(Weight vs. Dim2)
pause
twoway (scatter Design Dim2), title(Design vs. Dim2)
pause
twoway (scatter Volume Dim2), title(Volume vs. Dim2)
pause
twoway (scatter MaxFreq Dim2), title(MaxFreq vs. Dim2)
pause
twoway (scatter Power Dim2), title(Power vs. Dim2)
```

Fig. 13.17 STATA Commands for correlation analysis and plots (examp13-3.do)

```
. correlate Dim1-Power
(obs=10)

             |     Dim1     Dim2   Weight   Design   Volume  MaxFreq    Power
-------------+---------------------------------------------------------------
        Dim1 |   1.0000
        Dim2 |  -0.0000   1.0000
      Weight |  -0.7767   0.3692   1.0000
      Design |  -0.4462   0.3721   0.2789   1.0000
      Volume |   0.1487  -0.3152   0.1066  -0.7417   1.0000
     MaxFreq |  -0.1323   0.4635   0.5635   0.1880   0.1914   1.0000
       Power |  -0.3272   0.2501   0.6686  -0.1777   0.6746   0.4063   1.0000
```

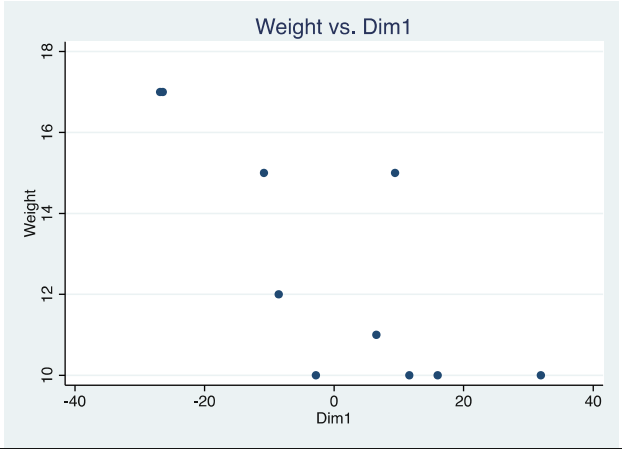


Fig. 13.18 STATA output for correlation analysis and plots

5 (letter F in the figure) prefers brand 10 (SUSI; indicated by the letter A) and then brands 1 (SAMA) and 6 (SIRO), both confounded on the map and represented by the "#" sign. The least preferred brands for this subject are brands 2 (SALT), 8 (SOLD), and 9 (SULI), these last two brands being confounded on the map and represented by the # sign in the lower right quadrant.

Preference analysis using the MDPREF model can also be done using XLSTAT. Because the data and command structures are similar for MDPREF and PREFMAP, we present the use of XLSTAT after the illustration of a PREFMAP analysis using PC-MDS.

```
5 10 2 2 1 0
(2X,10F3.0)
01 41 39 62 47 46 40 68 43 43 26
02 70 38 47 28 59 70 46 28 28 67
03 30 72 95 78 58 25 84 81 81 02
04 30 83 84 76 66 24 73 81 82 00
05 78 16 18 00 41 84 17 00 00 87
subj1
subj2
subj3
subj4
subj5
sama
salt
semi
self
sibi
siro
sono
sold
suli
susi
```

Fig. 13.19 Example of PC-MDS input file for MDPREF (examp13-4.dat)

```
                                M D P R E F
                    MULTIDIMENSIONAL ANALYSIS OF PREFERENCE DATA
                PROGRAM WRITTEN BY DR. J. D. CARROLL AND JIH JIE CHANG
                                PC - MDS VERSION

 ANALYSIS TITLE: MDPref example
 DATA IS READ FROM FILE: mdprf_t.dat
 OUTPUT FILE IS: mdprf_t.out


 NP (NO. OF VECTORS (SUBJECTS))                                    5
 NS (NO. OF POINTS  (STIMULI))                                    10
 NF (NO. OF DIMENSIONS)                                            2
 NFP (NO. OF DIMENSIONS PLOTTED)                                   2

 IREAD  1=NP X NS SCORE MATRIX WITH ROW MEAN SUBTRACTED            1
        2=SAME AS 1 WITH SCORES DIVIDED BY ROW S. D.

 NORP   0=NORMALIZE SUBJ. VECTORS                                  0
        1=DO NOT

 INPUT FORMAT = (2X,10F3.0)
 DATA FOR RECORD:     1
```

Fig. 13.20 Output example for MDPREF (examp13-4.out)

```
 .41E+02 .39E+02 .62E+02 .47E+02 .46E+02 .40E+02 .68E+02 .43E+02 .43E+02 .26E+02


DATA FOR RECORD:     5
 .78E+02 .16E+02 .18E+02 .00E+00 .41E+02 .84E+02 .17E+02 .00E+00 .00E+00 .87E+02

                      MEAN OF THE RAW SCORES  (BY SUBJECT)

        45.5000     48.1000     60.6000     59.9000     34.1000

                      FIRST SCORE MATRIX   (SUBJECT  BY  STIMULUS)

   1     -4.5000     -6.5000     16.5000      1.5000       .5000     -5.5000
         22.5000     -2.5000     -2.5000    -19.5000

   2     21.9000    -10.1000     -1.1000    -20.1000     10.9000     21.9000
         -2.1000    -20.1000    -20.1000     18.9000

   3    -30.6000     11.4000     34.4000     17.4000     -2.6000    -35.6000
         23.4000     20.4000     20.4000    -58.6000

   4    -29.9000     23.1000     24.1000     16.1000      6.1000    -35.9000
         13.1000     21.1000     22.1000    -59.9000

   5     43.9000    -18.1000    -16.1000    -34.1000      6.9000     49.9000
        -17.1000    -34.1000    -34.1000     52.9000

                      CROSS PRODUCT MATRIX OF SUBJECTS

   1   1266.5000   -511.5000   2419.0000   1961.5000  -1913.5000

   2   -511.5000   2754.9000  -3957.6000  -3985.9000   5421.9000

   3   2419.0000  -3957.6000   8640.4000   8247.6000  -9382.6010

   4   1961.5000  -3985.9000   8247.6000   8286.9000  -9282.8990

   5  -1913.5000   5421.9000  -9382.6010  -9282.8990  11630.9000

                      CORRELATION MATRIX OF SUBJECTS

   1      1.0000      -.2738       .7313       .6055      -.4986

   2      -.2738      1.0000      -.8112      -.8342       .9578

   3       .7313      -.8112      1.0000       .9747      -.9359

   4       .6055      -.8342       .9747      1.0000      -.9455

   5      -.4986       .9578      -.9359      -.9455      1.0000

                      CROSS PRODUCT MATRIX OF STIMULI

   1   4257.4400  -2026.0600  -2578.3600  -2957.7600    436.5401   4857.7400
      -2005.6600  -3181.0600  -3210.9600   6408.1400

   2  -2026.0600   1135.4400   1144.1400   1380.7400   -126.9600  -2323.7600
        753.8400   1556.4400   1579.5400  -3073.3600

   3  -2578.3600   1144.1400   2296.8400   1582.4400    -57.2600  -3008.0600
       1769.5400   1740.1400   1764.2400  -4653.6600

   4  -2957.7600   1380.7400   1582.4400   2131.0400   -400.6601  -3347.4600
       1277.1400   2257.7400   2273.8400  -4197.0600

   5    436.5401   -126.9600    -57.2600   -400.6601    210.6400    453.8401
       -110.5600   -379.9601   -373.8600    348.2401

   6   4857.7400  -2323.7600  -3008.0600  -3347.4600    453.8401   5556.0400
      -2326.3600  -3611.7600  -3647.6600   7397.4400

   7  -2005.6600    753.8400   1769.5400   1277.1400   -110.5600  -2326.3600
       1522.2400   1322.8400   1335.9400  -3538.9600

   8  -3181.0600   1556.4400   1740.1400   2257.7400   -379.9601  -3611.7600
       1322.8400   2434.4400   2455.5400  -4594.3600
```

Fig. 13.20 (continued)

```
   9  -3210.9600   1579.5400   1764.2400   2273.8400   -373.8600  -3647.6600
       1335.9400   2455.5400   2477.6400  -4654.2600

  10   6408.1400  -3073.3600  -4653.6600  -4197.0600    348.2401   7397.4400
      -3538.9600  -4594.3600  -4654.2600  10557.8400

                      ROOTS OF THE FIRST SCORE MATRIX

      30298.1700   1799.9580    417.2757     50.7452     13.4566

                     PROPORTION OF VARIANCE ACCOUNTED FOR BY EACH FACTOR

            1           2           3           4           5
           .9300       .0552       .0128       .0016       .0004

                     CUMULATIVE PROPORTION OF VARIANCE ACCOUNTED FOR

            1           2           3           4           5
           .9300       .9852       .9980       .9996      1.0000

                      SECOND SCORE MATRIX  (SUBJECT  BY  STIMULUS)

   1       -.0914      -.0220       .5479      -.0747       .1387      -.1508
            .4424      -.0662      -.0610      -.6630

   2        .4172      -.2344      -.0098      -.3569       .1138       .4534
            .0036      -.3786      -.3794       .3712

   3       -.3292       .1409       .3360       .1897       .0086      -.3888
            .2637       .2091       .2128      -.6428

   4       -.3536       .1610       .2882       .2241      -.0125      -.4107
            .2243       .2444       .2477      -.6129

   5        .3972      -.2029      -.1602      -.2976       .0631       .4460
           -.1191      -.3193      -.3215       .5143

POPULATION MATRIX (VECTORS)

FACTOR

   1        .6395       .7688

   2       -.9089       .4171

   3        .9821       .1882

   4        .9961       .0877

   5       -.9872       .1593

NORMALIZED STIMULUS MATRIX (POINTS)

FACTOR

   1       -.3717       .1903

   2        .1771      -.1759

   3        .2445       .5093

   4        .2519      -.3067

   5       -.0307       .2059

   6       -.4262       .1583

   7        .1883       .4189

   8        .2729      -.3131

   9        .2758      -.3087

  10       -.5820      -.3783
```

Fig. 13.20 (continued)

```
STIMULUS MATRIX (STRETCHED BY SQ. ROOT OF THE EIGENVALUES)

FACTOR

   1    -64.6951      8.0733

   2     30.8328     -7.4634

   3     42.5580     21.6070

   4     43.8507    -13.0122

   5     -5.3419      8.7374

   6    -74.1819      6.7176

   7     32.7685     17.7715

   8     47.4974    -13.2833

   9     48.0113    -13.0977

  10   -101.2998    -16.0502

     *****IDENTIFICATION KEY FOR PLOTS WITH IDENTIFIED POINTS*****

PT #   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
CHAR   1   2   3   4   5   6   7   8   9   A   B   C   D   E   F

PT #  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30
CHAR   G   H   I   J   K   L   M   N   O   P   Q   R   S   T   U

PT #  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45
CHAR   V   W   X   Y   Z   +   /   =   *   &   $   @   %   ?   <

PT #  46  47  48  49  50
CHAR   (   )   "   ;   @

POINT NUMBERS ABOVE 50 IDENTIFIED AS  >, MULTIPLE POINTS IDENTIFIED AS  #

IN JOINT SPACE PLOTS, THE FIRST  10 POINTS ARE STIMULI AND THE NEXT   5
ARE VECTOR (SUBJECT) END POINTS.
```

Fig. 13.20 (continued)

13.5.5 Example of PREFMAP

In the example provided in Fig. 13.21, the external source of the perceptual space configuration has been taken from the INDSCAL run. The first line of input in that file allows the user to define the various parameters concerning the data and the analysis to be done:

- The number of stimuli (here, 10 brands)
- The number of dimensions of the externally supplied perceptual space (here, 2)
- The number of subjects for which preferences are being modeled (here, 5)
- A code to indicate that the higher the score of a brand in the data, the higher the preference for that brand (code = 1) or that the higher the score, the lower the preference (code = 0); in the example, preferences are decreasing with the ratings and, therefore, a code 0 has been entered

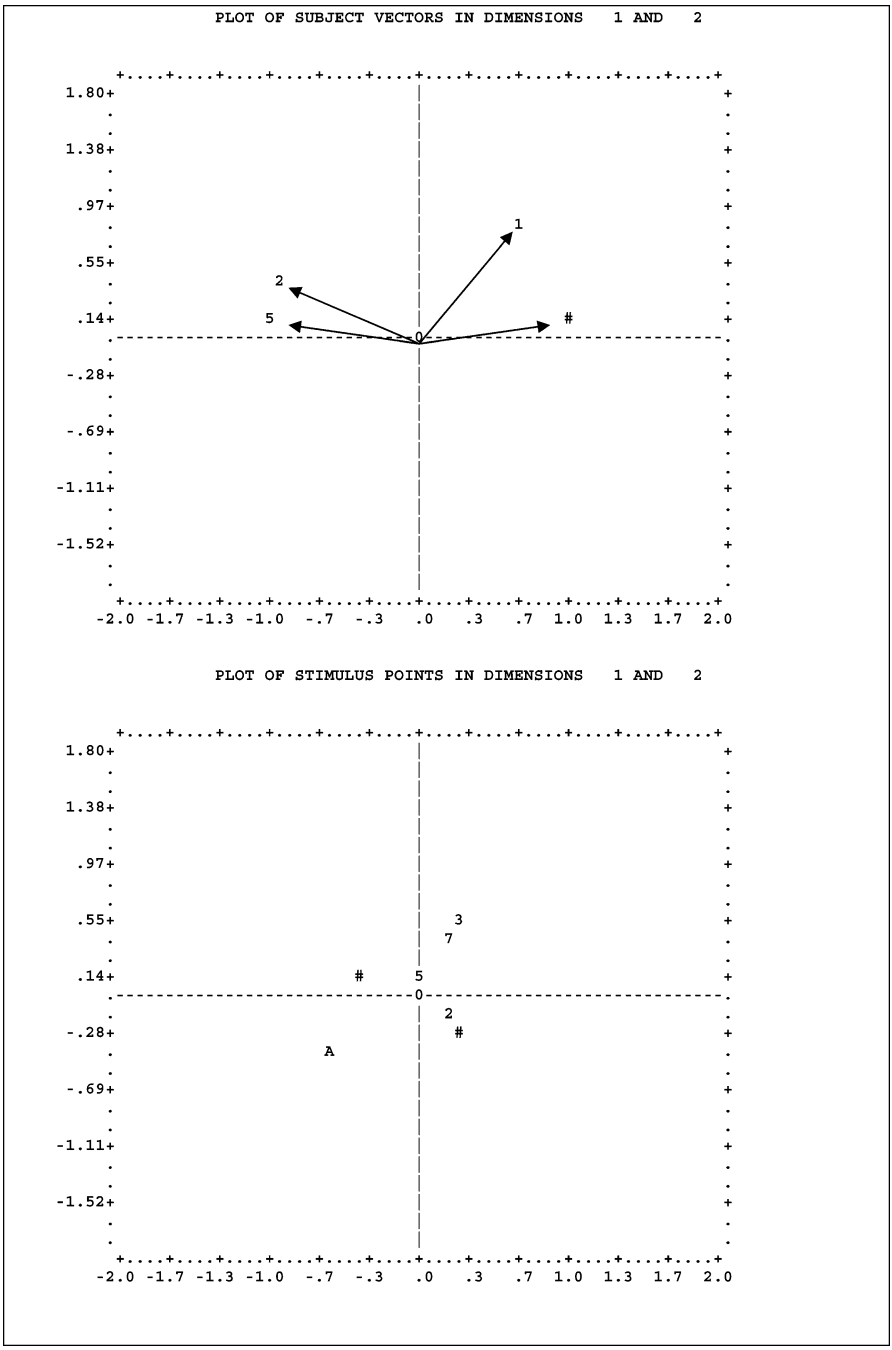


Fig. 13.20 (continued)

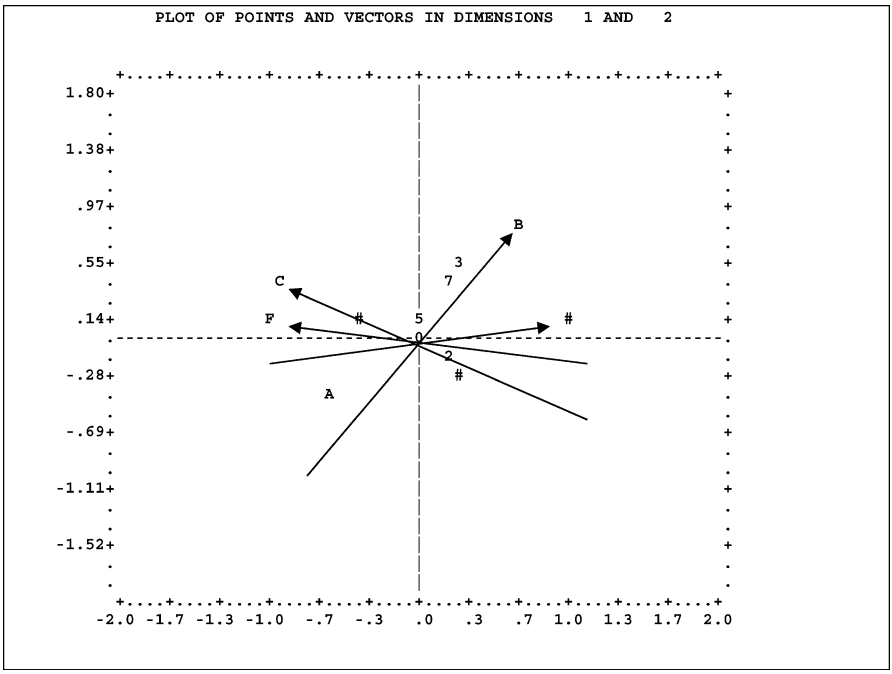


Fig. 13.20 (continued)

```
  10   2   5   0   1   0   2   4   0   1   1  15   0   0  1
 (2X,2F8.3)
  1   0.410   0.412  sama
  2  -0.222   0.191  salt
  3  -0.146  -0.275  semi
  4  -0.312  -0.366  self
  5   0.039   0.318  sibi
  6   0.471   0.358  siro
  7  -0.102  -0.420  sono
  8  -0.313  -0.242  sold
  9  -0.311  -0.231  suli
 10   0.486   0.256  susi
(2X,10F4.0)
01 059 061 038 053 054 060 032 057 057 074
02 030 062 053 072 041 030 054 072 072 033
03 070 028 005 022 042 075 016 019 019 098
04 070 017 016 024 034 076 027 019 018 100
05 022 084 082 100 059 016 083 100 100 013
sama
salt
semi
self
sibi
siro
sono
sold
suli
susi
SUBJ1
SUBJ2
SUBJ3
SUBJ4
SUBJ5
```

Fig. 13.21 Example of PC-MDS input file for PREFMAP (examp13-5.dat)

These numbers are followed by additional codes corresponding to advanced setting options.

The second line of input gives the format in which the coordinates in the perceptual space will be read. Then follow these coordinates for the ten stimuli (brands). Note that, given the format provided, the stimulus number (the first number on each of the lines for the coordinates) is not read by the program.

Then follows the format in which the preference data will be read. These preference data correspond to those described for the input of MDPREF. Therefore, the preference ratings of the ten brands are shown for each of the four subjects studied. Finally, the stimuli labels (brand names) are indicated.

Running the file contained in Fig. 13.21 leads to the results shown in Fig. 13.22. Phase 1 corresponds to the general unfolding model where the axes may be rotated differently for each subject and where each subject can weight each axis differently. Although the different rotation of the axes makes this model difficult to visualize, it is in fact the most versatile. It should be noted that there is one more point for subjects than there are subjects. This last point corresponds to the average preference (average ratings) across all the subjects.

Phase 2 corresponds to the weighted unfolding model wherein all subjects share the same configuration without rotation but each subject is allowed to weight each dimension differently. The preferences of each subject are shown by the person's ideal point in that common perceptual space.

In Phase 3, each subject uses the same perceptual space configuration with no axis rotation and no differential weighting of the dimensions.

Finally, Phase 4 corresponds to the vector model of preferences, similar to MDPREF, except for the fact that the perceptual configuration is externally provided. The figure here shows an example from the INDSCAL analysis.

The plot resulting from the analysis of Phase 3 provides the ideal points of the five subjects, as well as that of the average subject. This plot shows that subject 4 (represented by the letter D) prefers brands 2 (SALT), 3 (SEMI), 8 (SOLD), and 9 (SULI) (the closest to the subject's ideal brand). This fits the preference data used as input, where these brands have a low score value (most preferred).

For the vector model of preferences, the last graph shows the end points of the individual vectors. The vectors drawn in Fig. 13.22 have been added to the original output, showing the differences in preferences across individuals according to the projections of the stimuli on their respective vectors. For example, the projections of the brands on the vectors of subjects 2 (C) and 5 (F) indicate that brands 1 (SAMA), 6 (SIRO), and 10 (SUSI; indicated by the letter A on the plot) are the

```
                    P R E F M A P
MDSCALING VIA A GENERALIZATION OF COOMBS UNFOLDING MODEL
        BY DR. J. D. CARROLL AND JIH JIE CHANG
                    PC - MDS VERSION
```

Fig. 13.22 Output example for PREFMAP (examp13-5.out)

```
ANALYSIS TITLE: Prefmap example
 DATA IS READ FROM FILE: prefv_t.dat
 OUTPUT FILE IS: prefv_t.out

 ***************************************************************************

  N        NO. OF STIMULI                                          10
  K        NO. OF DIMENSIONS                                        2
  NSUB     NO. OF SUBJECTS                                          5
  ISV    0=SMALL SCALE VALUE REPRESENTS GREATER PREF.               0
  NORS   1=NORMALIZE SCALE VALUES                                   1
  IRX    0=STIMULUS COORDINATES N BY K, OR 1 = K BY N               0
  IPS      STARTING PHASE                                           2
  IPE      ENDING PHASE                                             4
  IRWT   1=READ IN WEIGHTS,  0=NO WEIGHTS READ IN                   0
  LFITSW   HOW D**2 IS RELATED TO SCALE VALUES                      1
         0=LINEARLY,
         1=MONOTONE WITH NO TIES,
         2=BLOCK MONOTONE WITH ORDERING IN BLOCKS
         3=BLOCK MONOTONE WITH EQUALITY IN BLOCKS
  IAV    0=AVERAGE SUBJECTS COMPUTED ONCE FOR ALL PHASES,           1
         1=CALCULATE EACH PHASE
  MAXIT    MAXIMUM ITERATIONS, WHEN 0 IT IS SET TO 15              15
  ISHAT  0=USE SCALE VALUES FROM PREVIOUS PHASE,                    0
         1=USE ORIG VALUES
  IPLOT  0=AVERAGE SUBJECTS,                                        0
         1=AVERAGE SUBJECTS & SUBJECT FUNCTIONS,
         2=ALL PLOTS
  CRIT     CRITERIA FOR STOPPING MONOTONE FIT                     .0010

 ***************************************************************************

    *****IDENTIFICATION KEY FOR PLOTS WITH IDENTIFIED POINTS*****

 PT #   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
 CHAR   1   2   3   4   5   6   7   8   9   A   B   C   D   E   F

 PT #  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30
 CHAR   G   H   I   J   K   L   M   N   O   P   Q   R   S   T   U

 PT #  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45
 CHAR   V   W   X   Y   Z   +   /   =   *   &   $   @   %   ?   <

 PT #  46  47  48  49  50
 CHAR   (   )   "   #   @

 POINT NUMBERS ABOVE 50 IDENTIFIED AS  >, MULTIPLE POINTS IDENTIFIED AS  ;


 POINTS 1 TO 10 ARE STIMULI AND POINTS  11 TO   15 ARE IDEAL POINTS

 VARIABLE FORMAT (STIMULUS COORDINATES)  =  (2X,2F8.3)

 ORIGINAL CONFIGURATION  (X MATRIX)


     1    .41000    .41200
     2   -.22200    .19100
     3   -.14600   -.27500
     4   -.31200   -.36600
     5    .03900    .31800
     6    .47100    .35800
     7   -.10200   -.42000
     8   -.31300   -.24200
     9   -.31100   -.23100
    10    .48600    .25600

 VARIABLE FORMAT  (SCALE VALUES) = (2X,10F4.0)

 PHASE  2

 X MATRIX, (INPUT CONFIGURATION AFTER NORMALIZATION)
   1       .4100      -.2220      -.1460      -.3120       .0390       .4710
          -.1020      -.3130      -.3110       .4860
   2       .4120       .1910      -.2750      -.3660       .3180       .3580
```

Fig. 13.22 (continued)

```
         -.4200      -.2420      -.2310       .2560

PHASE  2

SUBJECT   1

SCALE VALUES BEFORE NORMALIZATION FOR SUBJECT    1
      59.00000     61.00000     38.00000     53.00000     54.00000     60.00000
      32.00000     57.00000     57.00000     74.00000

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
        .12645       .18265      -.46364      -.04215      -.01405       .15455
       -.63224       .07025       .07025       .54794

BEGIN ITERATION ON MONOTONE FIT
END OF ITERATION, REACHED CRITERION

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.13961      -.72085       .81888      3.30984     -1.91886

(CORRELATION)=       .99947

SIGNED DSQ, (SIGNED DISTANCE SQUARED FROM STIMULI TO IDEAL)
        .22438       .36144      -.24263      -.05777      -.00483       .39385
       -.62257       .19123       .20465       .46720

*****************************************************************************

SUBJECT     1

COORDINATES OF IDEAL POINT WITH RESPECT TO OLD AXES
             .10889         .21338

IMPORTANCES OF NEW AXES
            3.30984       -1.91886

*****************************************************************************

SUBJECT   2

SCALE VALUES BEFORE NORMALIZATION FOR SUBJECT    2
      30.00000     62.00000     53.00000     72.00000     41.00000     30.00000
      54.00000     72.00000     72.00000     33.00000

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       -.41725       .19243       .02096       .38295      -.20767      -.41725
        .04001       .38295       .38295      -.36009

BEGIN ITERATION ON MONOTONE FIT
END OF ITERATION, REACHED CRITERION

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.17419     -1.21683      -.04638      1.55564       .18543

(CORRELATION)=       .99931

SIGNED DSQ, (SIGNED DISTANCE SQUARED FROM STIMULI TO IDEAL)
        .01582       .58556       .47845       .81375       .19976       .01999
        .43334       .79621       .79036       .01719

*****************************************************************************

SUBJECT     2

COORDINATES OF IDEAL POINT WITH RESPECT TO OLD AXES
             .39110         .12507

IMPORTANCES OF NEW AXES
            1.55564         .18543

*****************************************************************************

SUBJECT   3
```

Fig. 13.22 (continued)

```
SCALE VALUES BEFORE NORMALIZATION FOR SUBJECT    3
      70.00000     28.00000      5.00000     22.00000     42.00000     75.00000
      16.00000     19.00000     19.00000     98.00000

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
        .32920      -.12264      -.37008      -.18719       .02797       .38299
       -.25174      -.21946      -.21946       .63042

BEGIN ITERATION ON MONOTONE FIT
END OF ITERATION, REACHED CRITERION

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.15664       .48421       .24583      1.52976       .03537

(CORRELATION)=       .99909

SIGNED DSQ, (SIGNED DISTANCE SQUARED FROM STIMULI TO IDEAL)
       1.02841       .48159       .36244       .37805       .56841      1.12542
        .33497       .40634       .40792      1.12735

*******************************************************************************

SUBJECT     3

COORDINATES OF IDEAL POINT WITH RESPECT TO OLD AXES
            -.15826       -3.47505

IMPORTANCES OF NEW AXES
            1.52976         .03537

*******************************************************************************

SUBJECT   4

SCALE VALUES BEFORE NORMALIZATION FOR SUBJECT    4
      70.00000     17.00000     16.00000     24.00000     34.00000     76.00000
      27.00000     19.00000     18.00000    100.00000

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
        .32845      -.25376      -.26474      -.17686      -.06701       .39437
       -.14390      -.23179      -.24277       .65801

BEGIN ITERATION ON MONOTONE FIT
END OF ITERATION, REACHED CRITERION

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.12124       .76701      -.06093      1.39435      -.18300

(CORRELATION)=       .99917

SIGNED DSQ, (SIGNED DISTANCE SQUARED FROM STIMULI TO IDEAL)
        .59310      -.01946       .02106      -.00538       .09456       .72572
        .02999       .00097       .00104       .77492

*******************************************************************************

SUBJECT     4

COORDINATES OF IDEAL POINT WITH RESPECT TO OLD AXES
            -.27504        -.16648

IMPORTANCES OF NEW AXES
            1.39435        -.18300

*******************************************************************************

SUBJECT   5

SCALE VALUES BEFORE NORMALIZATION FOR SUBJECT    5
      22.00000     84.00000     82.00000    100.00000     59.00000     16.00000
      83.00000    100.00000    100.00000     13.00000

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       -.40706       .16783       .14929       .31619      -.06398      -.46269
```

Fig. 13.22 (continued)

```
        .15856       .31619       .31619      -.49051

BEGIN ITERATION ON MONOTONE FIT

AVERAGE SUBJECT

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
        .02086       .05391      -.12695       .02712      -.11193       .08604
       -.19693       .06852       .06852       .11084

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.13147      -.34136       .18293      1.57832      -.26520

(CORRELATION)=       .99884

SIGNED DSQ, (SIGNED DISTANCE SQUARED FROM STIMULI TO IDEAL)
        .14262       .16575       .00004       .14458       .00735       .20777
       -.08546       .18859       .18933       .22325

********************************************************************************

SUBJECT     6

COORDINATES OF IDEAL POINT WITH RESPECT TO OLD AXES
             .10814          .34488

IMPORTANCES OF NEW AXES
            1.57832        -.26520

********************************************************************************

PHASE  3

X MATRIX, (INPUT CONFIGURATION AFTER NORMALIZATION)
  1       .5151      -.2789      -.1834      -.3920       .0490       .5917
         -.1281      -.3932      -.3907       .6106
  2       .2122       .0984      -.1416      -.1885       .1638       .1844
         -.2163      -.1246      -.1190       .1318

PHASE  3

SUBJECT   1

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
        .13300       .28639      -.33440      -.14939      -.09641       .28639
       -.71466       .10653       .10653       .37602

BEGIN ITERATION ON MONOTONE FIT
END OF ITERATION, REACHED CRITERION

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.31127      -.83629      1.99854      2.36724

(CORRELATION)=       .99951

SIGNED DSQ, (SIGNED DISTANCE SQUARED FROM STIMULI TO IDEAL)
        .16681       .24310      -.44543      -.11724      -.11944       .27404
       -.74493       .06110       .06892       .24625

********************************************************************************

SUBJECT     1

COORDINATES OF IDEAL POINT WITH RESPECT TO OLD AXES
             .17664          .42212

IMPORTANCES OF NEW AXES
            2.36724       -2.36724

********************************************************************************

SUBJECT   2

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
```

Fig. 13.22 (continued)

```
       -.39970       .17073       .04090       .39919      -.21554      -.39693
        .04090       .37870       .37870      -.39693

BEGIN ITERATION ON MONOTONE FIT
END OF ITERATION, REACHED CRITERION

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.11185      -.87503      -.23917       .85102

(CORRELATION)=       .99876

SIGNED DSQ, (SIGNED DISTANCE SQUARED FROM STIMULI TO IDEAL)
       -.10586       .48661       .41405       .69670       .10530      -.08470
        .34614       .70038       .69633      -.05521

*******************************************************************************

SUBJECT     2

COORDINATES OF IDEAL POINT WITH RESPECT TO OLD AXES
             .51410        -.14052

IMPORTANCES OF NEW AXES
             .85102        -.85102

*******************************************************************************

SUBJECT   3

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       .40689      -.14069      -.27377      -.22496      -.05375       .50402
      -.27377      -.22496      -.22496       .50596

BEGIN ITERATION ON MONOTONE FIT

SUBJECT   4

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       .37192      -.22229      -.22229      -.22229      -.12725       .50470
      -.19190      -.22229      -.22229       .55396

BEGIN ITERATION ON MONOTONE FIT

SUBJECT   5

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       -.40780       .17542       .15478       .33307      -.06671      -.46796
        .15478       .30461       .30461      -.48483

BEGIN ITERATION ON MONOTONE FIT

AVERAGE SUBJECT

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       .02402       .05482      -.13502       .03375      -.10786       .08389
      -.19180       .06715       .06715       .10390

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.12947      -.27321       .36106       .98483

(CORRELATION)=       .99713

SIGNED DSQ, (SIGNED DISTANCE SQUARED FROM STIMULI TO IDEAL)
        .13869       .16465      -.00178       .14122       .00755       .20210
      -.08713       .18528       .18606       .21666

*******************************************************************************

SUBJECT     6

COORDINATES OF IDEAL POINT WITH RESPECT TO OLD AXES
             .13871         .18331

IMPORTANCES OF NEW AXES
```

Fig. 13.22 (continued)

```
              .98483        -.98483

*******************************************************************************

STIMULI COORDINATES
DIMENSION         1              2
STIMULI
  1              .51509         .21217
  2             -.27890         .09836
  3             -.18342        -.14162
  4             -.39197        -.18848
  5              .04900         .16376
  6              .59172         .18436
  7             -.12814        -.21629
  8             -.39323        -.12463
  9             -.39071        -.11896
 10              .61057         .13183


COORDINATES OF IDEAL POINTS

DIMENSION         1              2
SUBJECTS
  1              .17664         .42212
  2              .51410        -.14052
  3             -.24813         .22016
  4             -.33261        -.04437
  5            -7.43571        2.44479
  6              .13871         .18331

SUBJECT    6 IS THE AVERAGE SUBJECT

WEIGHTS OF AXES
DIMENSION         1              2
SUBJECTS
  1             2.36724       -2.36724
  2              .85102        -.85102
  3              .89211        -.89211
  4              .89720        -.89720
  5             -.04765         .04765
  6              .98483        -.98483

SUBJECT    6 IS THE AVERAGE SUBJECT
```

Fig. 13.22 (continued)

preferred ones. These correspond indeed to the lowest scores (most preferred) in the input data.

We now illustrate the use of these analyses with XLSTAT. The spreadsheet containing the data should be composed of two worksheets, one (sheet 1) containing the preference data and the other (sheet 2) containing the perceptual coordinates in the space configuration obtained from preliminary MDS analysis. Fig. 13.23 shows the preference data for the ten brands (rows). The columns correspond to the responses of individuals or segments of individuals (in this example, we use consumer segments). When there are many individuals, it is best to group them into subgroups that have similar patterns of preferences. The analysis is then done at the cluster or segment level. Consequently, the data matrix should have N rows for the N groups (individuals or clusters) and J columns for the J alternatives (brands) that are being evaluated in terms of preferences.

```
STIMULI AND IDEAL POINTS:  Prefmap example

          .*....*....*....*....*....*....*....*....*....*....*....*....*....*
     1.50**                                |                                **
     1.38**                                |                                **
     1.27**                                |                                **
     1.15**                                |                                **
     1.04**                                |                                **
      .92**                                |                                **
      .81**                                |                                **
      .69**                                |                                **
      .58**                                |                                **
      .46**                                |  B                             **
      .35**                                |                                **
      .23**                             D  | G     16                       **
      .12**                             2  |5       A                       **
      .00**-------------------------E----0-------------------------------**
     -.12**                        ;  3   |       C                        **
     -.23**                        4   7  |                                **
     -.35**                               |                                **
     -.46**                               |                                **
     -.58**                               |                                **
     -.69**                               |                                **
     -.81**                               |                                **
     -.92**                               |                                **
    -1.04**                               |                                **
    -1.15**                               |                                **
    -1.27**                               |                                **
    -1.38**                               |                                **
    -1.50**                               |                                **
          .*....*....*....*....*....*....*....*....*....*....*....*....*....*
          .   -1.6667.  -1.0000.   -.3333.    .3333.   1.0000.   1.6667.
         -2.0000   -1.3333    -.6667     .0000     .6667    1.3333    2.0000

 PHASE  4

 X MATRIX, (INPUT CONFIGURATION AFTER NORMALIZATION)
   1       .5151      -.2789      -.1834      -.3920       .0490       .5917
          -.1281      -.3932      -.3907       .6106
   2       .2122       .0984      -.1416      -.1885       .1638       .1844
          -.2163      -.1246      -.1190       .1318

 PHASE  4

 SUBJECT   1

 S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
        -.20366      -.29138       .40907       .08173       .08173      -.29138
         .70881      -.10177      -.10177      -.29138

 BEGIN ITERATION ON MONOTONE FIT
 END OF ITERATION, REACHED CRITERION

 BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
 QUADRATIC, LINEAR, THEN A CONSTANT TERM)
         .00010       .11482     -1.90631

 (CORRELATION)=       .87383

 PROJECTIONS ON THE FITTED VECTOR
        -.18082      -.11495       .13034       .16458      -.16052      -.14845
         .20820       .10076       .09525      -.09489

 SUBJECT   2

 S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
         .42686      -.16703      -.06027      -.37874       .21519       .40565
        -.06027      -.37874      -.37874       .37609

 BEGIN ITERATION ON MONOTONE FIT
 END OF ITERATION, REACHED CRITERION

 BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
 QUADRATIC, LINEAR, THEN A CONSTANT TERM)
        -.00002       .68450       .33335
```

Fig. 13.22 (continued)

```
(CORRELATION)=       .99939

PROJECTIONS ON THE FITTED VECTOR
        .55599      -.20768      -.22691      -.43493       .11575       .61271
       -.20991      -.40810      -.40336       .60666

SUBJECT   3

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       -.39168       .14174       .26475       .23034       .05320      -.50049
        .26475       .23034       .23034      -.52329

BEGIN ITERATION ON MONOTONE FIT
END OF ITERATION, REACHED CRITERION

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
        .00004      -.48253      -.86681

(CORRELATION)=       .99356

PROJECTIONS ON THE FITTED VECTOR
       -.43592       .04971       .21295       .35533      -.16692      -.44889
        .25131       .30015       .29398      -.41216

SUBJECT   4

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       -.37048       .22038       .22038       .22038       .12371      -.50450
        .20456       .22038       .22038      -.55519

BEGIN ITERATION ON MONOTONE FIT
END OF ITERATION, REACHED CRITERION

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
       -.00000      -.80876       .07031

(CORRELATION)=       .98856

PROJECTIONS ON THE FITTED VECTOR
       -.49478       .28637       .17047       .37417      -.03463      -.57353
        .10893       .38095       .37894      -.59685

SUBJECT   5

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
        .41885      -.17783      -.15883      -.32247       .06547       .47127
       -.15883      -.30594      -.30594       .47426

BEGIN ITERATION ON MONOTONE FIT

AVERAGE SUBJECT

S (VECTOR OF SCALE VALUES, E.G. PREFERENCES)
       -.06377      -.07260       .11979       .05698      -.00084      -.06240
        .11541      -.01085      -.01085      -.07087

BETA VALUES  (IN THE MOST GENERAL CASE THERE ARE (2K + K(K-1)/2 + 1) TERMS -
QUADRATIC, LINEAR, THEN A CONSTANT TERM)
        .00002       .04867      -.43837

(CORRELATION)=       .81988

PROJECTIONS ON THE FITTED VECTOR
       -.15404      -.12854       .12052       .14408      -.15736      -.11795
        .20083       .08047       .07512      -.06366

STIMULI COORDINATES
DIMENSION        1              2
STIMULI
  1             .51509         .21217
  2            -.27890         .09836
  3            -.18342        -.14162
  4            -.39197        -.18848
  5             .04900         .16376
```

Fig. 13.22 (continued)

```
  6            .59172         .18436
  7           -.12814        -.21629
  8           -.39323        -.12463
  9           -.39071        -.11896
 10            .61057         .13183


 STIMULI AND IDEAL POINTS:  Prefmap example

        .*....*....*....*....*....*....*....*....*....*....*....*....*....*
   1.50**                                                                  **
   1.38**                                                                  **
   1.27**                                                                  **
   1.15**                                                                  **
   1.04**                                                                  **
    .92**                                                                  **
    .81**                                                                  **
    .69**                                                                  **
    .58**                                                                  **
    .46**                                                                  **
    .35**                                                                  **
    .23**                                                                  **
    .12**                                                                  **
    .00**------------------------------------------------------------------**
   -.12**                                                                  **
   -.23**                                                                  **
   -.35**                                                                  **
   -.46**                                                                  **
   -.58**                                                                  **
   -.69**                                                                  **
   -.81**                                                                  **
   -.92**                                                                  **
  -1.04**                                                                  **
  -1.15**                                                                  **
  -1.27**                                                                  **
  -1.38**                                                                  **
  -1.50**                                                                  **
        .*....*....*....*....*....*....*....*....*....*....*....*....*....*
        .   -1.6667.  -1.0000.   -.3333.    .3333.   1.0000.   1.6667.
       -2.0000   -1.3333    -.6667     .0000     .6667    1.3333    2.0000


DIRECTION COSINES OF FITTED SUBJECT VECTORS
                  DIMENSION
SUBJECT       1         2
     1       .0601    -.9982
     2       .8991     .4378
     3      -.4864    -.8737
     4      -.9962     .0866
     5       .9488     .3160
 AVG R       .1103    -.9939

        CORRELATION (PHASE)                    F RATIO (PHASE)

        R1       R2       R3       R4       F1       F2       F3       F4

DF                                           5 4      4 5      3 6      2 7
SUBJ
 1       .000     .999    1.000     .874     .000 1177.124 1000.000   11.304
 2       .000     .999     .999     .999     .000  904.412  803.281 2889.408
 3       .000     .999     .998     .994     .000  682.663  642.385  269.136
 4       .000     .999    1.000     .989     .000  752.687 1000.000  150.310
 5       .000     .999    1.000    1.000     .000  765.898 1000.000 1000.000
AVG      .000     .999     .997     .820     .000  537.822  347.263    7.177

        F RATIO (BETWEEN PHASE)
        F12      F13      F14      F23      F24      F34
DF      1 4      2 4      3 4      1 5      2 5      1 6
SUBJ
 1       .000     .000     .000    -.409  554.674 1000.000
 2       .000     .000     .000    3.997    -.309   -3.077
 3       .000     .000     .000    3.491   15.060   18.817
 4       .000     .000     .000   -2.871   31.812 1000.000
 5       .000     .000     .000   -3.122   -1.820 1000.000
AVG      .000     .000     .000    7.348  350.915  337.465
```

Fig. 13.22 (continued)

```
ROOT MEAN SQUARE
PHASE
 1        .000
 2        .999
 3        .999
 4        .972


    AN F - VALUE OF 1000.0 IN THE ABOVE TABLE INDICATES
    A POSSIBLE DIVISION BY ZERO. I.E. R IS VERY CLOSE TO 1.00
```

Fig. 13.22 (continued)

	A	B	C	D	E	F	G
1	Segment	Seg1	Seg2	Seg3	Seg4	Seg5	
2	sama	41	70	30	30	78	
3	salt	39	38	72	83	16	
4	semi	62	47	95	84	18	
5	self	47	28	78	76	0	
6	sibi	46	59	58	66	41	
7	siro	40	70	25	24	84	
8	sono	68	46	84	73	17	
9	sold	43	28	81	81	0	
10	suli	43	28	81	82	0	
11	susi	26	67	2	0	87	
12							

Fig. 13.23 Preference data in Excel worksheet for MDPREF analysis in XLSTAT (examp13-4.xlsx, sheet 1)

As noted above, the perception coordinates should be on sheet 2 of this same spreadsheet. The coordinates of the perceptions of each object (alternative) are entered in the rows, and the columns correspond to the number of perceptual dimensions needed. Figure 13.24 shows the coordinates that were found as a solution to reflect the dissimilarities obtained from the MDS analysis. These could be obtained more directly from the principal component analysis of perceptions on a number of attributes. Therefore, the number of rows here in sheet 2 corresponds precisely to the number of rows in the preference data entered in sheet 1.

Figure 13.25 shows the XLSTAT dialog box for defining the problem parameters corresponding to PREFMAP or MDPREF.

Although "Quadratic" has been selected as the "model" in the dialog box shown in Fig. 13.25, the program also analyzes the vector model, and based on the assessment of both models, it reports out on the best option for each segment or individual. A mixture of vector and ideal point preference models can result when some segments' preferences are best represented by a vector model and other segments are best represented by a quadratic model. In this particular example,

Fig. 13.24 Perceptual data in Excel worksheet for MDPREF analysis in XLSTAT (examp13-4.xlsx, sheet 2)

	A	B	C	D
1		Dim1	Dim2	
2	sama	11.6145	19.6876	
3	salt	-8.5461	-5.982	
4	semi	-26.4431	10.4797	
5	self	9.4129	23.3648	
6	sibi	6.5306	-1.7838	
7	siro	-2.8052	-28.5256	
8	sono	15.9945	-17.5831	
9	sold	-26.8466	-14.5755	
10	suli	31.914	-5.6048	
11	susi	-10.8256	20.5226	
12				

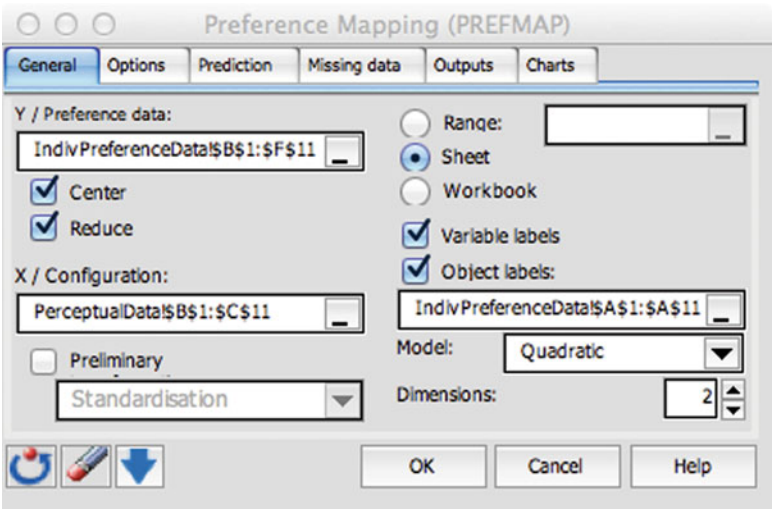


Fig. 13.25 XLSTAT dialog box for MDPREF

however, it is the vector model that best fits each of the five segments, as can be seen from the partial output displayed in Fig. 13.26.

The lines in Fig. 13.26 correspond to the vector of preference for each segment. This map of perceptions and preferences shows, for example, that segments 2 and 5 have similar preferences for "susi," "self," "sama," and "semi." In contrast, segments 1, 3, and 4 prefer brands "siro" and "sono." The order of preferences for each segment is also provided in the output shown in the top part of the figure (i.e., in the table above the map).

Model selection:

Y	Model	Point type	Dim1	Dim2
Seg1	Vector	-		
Seg2	Vector	-		
Seg3	Vector	-		
Seg4	Vector	-		
Seg5	Vector	-		

Analysis of variance:

Y	DF	Sum of squares	Mean squares	R^2	F	Pr > F
Seg1	2	0.516	0.258	0.057	0.213	0.813
Seg2	2	0.067	0.034	0.007	0.026	0.974
Seg3	2	0.300	0.150	0.033	0.121	0.888
Seg4	2	0.364	0.182	0.040	0.148	0.865
Seg5	2	0.142	0.071	0.016	0.056	0.946

Model coefficients:

Y	Intercept	Dim1	Dim2
Seg1	0.000	0.004	-0.013
Seg2	0.000	-0.003	0.004
Seg3	0.000	0.000	-0.010
Seg4	0.000	0.001	-0.011
Seg5	0.000	-0.003	0.006

Model predictions:

Objects	Seg1	Seg2	Seg3	Seg4	Seg5
sama	-0.200	0.046	-0.202	-0.208	0.088
salt	0.040	0.000	0.061	0.057	-0.012
semi	-0.244	0.113	-0.106	-0.147	0.146
self	-0.256	0.066	-0.239	-0.251	0.118
sibi	0.050	-0.025	0.018	0.027	-0.031
siro	0.349	-0.104	0.292	0.317	-0.170
sono	0.290	-0.112	0.179	0.215	-0.159
sold	0.071	0.016	0.150	0.133	-0.010
suli	0.206	-0.108	0.056	0.098	-0.132
susi	-0.306	0.110	-0.209	-0.242	0.161

Fig. 13.26 XLSTAT output of preference analysis (examp13-4, “PREFMAP” sheet)

Preference scores from 0 to 1:

Objects	Seg1	Seg2	Seg3	Seg4	Seg5
sama	0.161	0.702	0.071	0.077	0.779
salt	0.527	0.498	0.566	0.544	0.478
semi	0.094	1.000	0.251	0.184	0.953
self	0.075	0.792	0.000	0.000	0.868
sibi	0.543	0.389	0.484	0.491	0.420
siro	1.000	0.036	1.000	1.000	0.000
sono	0.910	0.000	0.788	0.821	0.035
sold	0.575	0.569	0.733	0.677	0.484
suli	0.781	0.017	0.556	0.616	0.115
susi	0.000	0.987	0.056	0.016	1.000

Ranks of the preference scores:

Objects	Seg1	Seg2	Seg3	Seg4	Seg5
sama	4	7	3	3	7
salt	5	5	7	6	5
semi	3	10	4	4	9
self	2	8	1	1	8
sibi	6	4	5	5	4
siro	10	3	10	10	1
sono	9	1	9	9	2
sold	7	6	8	8	6
suli	8	2	6	7	3
susi	1	9	2	2	10

Objects sorted by increasing preference order:

Seg1	Seg2	Seg3	Seg4	Seg5
susi	sono	self	self	siro
self	suli	susi	susi	sono
semi	siro	sama	sama	suli
sama	sibi	semi	semi	sibi
salt	salt	sibi	sibi	salt
sibi	sold	suli	salt	sold
sold	sama	salt	suli	sama
suli	self	sold	sold	self
sono	susi	sono	sono	semi
siro	semi	siro	siro	susi

Fig. 13.26 (continued)

Percentage of satisfied assessors for each object:

Object	%
sama	40%
salt	60%
semi	40%
self	40%
sibi	60%
siro	60%
sono	60%
sold	80%
suli	60%
susi	40%

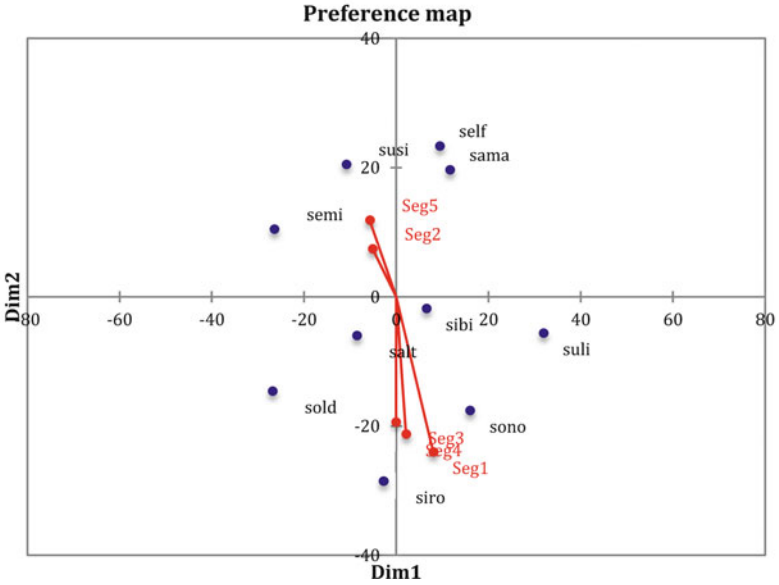


Fig. 13.26 (continued)

13.6 Assignment

Collect proximity data about a set of brands of your choice (limit the number of brands to ten maximum) and determine the dimensions used in the perception of these brands. Gather data about characteristics of these brands to help you interpret the underlying perceptual dimensions. For these same brands, obtain preferences of the respondents in order to develop a map of subject preferences and stimuli.

Bibliography

Basic Technical Readings

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology, 31*, 607–49.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage Publications.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*(24), 390–398.

Application Readings

Bijmolt, T. H. A., & Wedel, M. (1999). A comparison of multidimensional scaling methods for perceptual mapping. *Journal of Marketing Research, 36*, 277–285.

Cooper, L. G. (1983). A review of multidimensional scaling in marketing research. *Applied Psychological Measurement, 7*(4), 427–450.

DeSarbo, W. S., Young, M. R., & Rangaswamy, A. (1997). A parametric multidimensional unfolding procedure for incomplete nonmetric preference/choice set data in marketing research. *Journal of Marketing Research, 34*(4), 499–516.

DeSarbo, W. S., & De Soete, G. (1984). On the use of hierarchical clustering for the analysis of nonsymmetric proximities. *Journal of Consumer Research, 11*, 601–610.

Green, P. E. (1975). Marketing applications of MDS: Assessment and outlook. *Journal of Marketing, 39*, 24–31.

Green, P. E., & Carmone, F. J. (1989). Multidimensional scaling: An introduction and comparison of nonmetric unfolding techniques. *Journal of Marketing Research, 6*, 330–341.

Helsen, K., & Green, P. E. (1991). A computational study of replicated clustering with an application to market segmentation. *Decision Sciences, 22*, 1124–1141.

Johnson, R. M. (1971). Market segmentation: A strategic management tool. *Journal of Marketing Research, 8*, 13–18.

Malhotra, N. K., Jain, A. K., Patil, A., Pinson, C., & Lan, W. (2010). Consumer cognitive complexity and the dimensionality of multidimensional scaling configurations. *Review of Marketing Research, 7*, 199–253.

Neidell, L. A. (1969). The use of nonmetric multidimensional scaling in marketing analysis. *Journal of Marketing, 33*, 37–43.

Sexton, D. E., Jr. (1974). A cluster analytic approach to market response functions. *Journal of Marketing Research, 11*, 109–114.

Srivatsava, R. K., Leone, R. P., & Shocker, A. D. (1981). Market structure analysis: Hierarchical clustering of products based on substitution in use. *Journal of Marketing, 45*(Summer), 38–48.

Stuart, T. E., & Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal, 17*, 21–38.

Chapter 14
Appendices

14.1 Appendix A: Rules in Matrix Algebra

14.1.1 Vector and Matrix Differentiation

$$\frac{\partial \mathbf{a}'\mathbf{v}}{\partial \mathbf{v}} = \mathbf{a} \tag{14.1}$$

$$\frac{\partial \mathbf{v}'\mathbf{A}\mathbf{v}}{\partial \mathbf{v}} = (\mathbf{A} + \mathbf{A}')\mathbf{v} \tag{14.2}$$

14.1.2 Kronecker Products

$$\mathbf{A} \otimes \mathbf{B} \tag{14.3}$$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{14.4}$$

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{bmatrix} \tag{14.5}$$

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} \tag{14.6}$$

14.1.3 Determinants

$$|\mathbf{A}| = \prod_{i=1}^{P} \lambda_i \tag{14.7}$$

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0_14, © Springer Science+Business Media New York 2014

where λ_i represents the eigenvalues of matrix $\mathbf{A}$ and P the dimensionality of that matrix.

$$|\mathbf{A}| = |\mathbf{A}||\mathbf{B}| \tag{14.8}$$

14.1.4 Trace

$$tr(\mathbf{ABC}) = tr(\mathbf{ACB}) = tr(\mathbf{CAB}) = tr(\mathbf{BCA}) = tr(\mathbf{CBA}) \tag{14.9}$$

14.2 Appendix B: Statistical Tables

14.2.1 Cumulative Normal Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.05	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9027	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

(continued)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

14.2.2 Chi-Square Distribution

ν	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995
1	0.00004	0.0002	0.001	0.004	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	28.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	38.29	44.64	50.98	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49

14.2.3 F Distribution

ν_1 = Degrees of freedom for the numerator

ν_2	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

ν_1 = Degrees of freedom for the numerator

ν_2	10	12	15	20	30	40	50	60	∞
1	241.88	243.91	245.95	248.01	250.10	251.14	252.20	252.20	254.19
2	19.40	19.41	19.43	19.45	19.46	19.47	19.48	19.48	19.49
3	8.79	8.74	8.70	8.66	8.62	8.59	8.57	8.57	8.53
4	5.96	5.91	5.86	5.80	5.75	5.72	5.69	5.69	5.63
5	4.74	4.68	4.62	4.56	4.50	4.46	4.43	4.43	4.37
6	4.06	4.00	3.94	3.87	3.81	3.77	3.74	3.74	3.67
7	3.64	3.57	3.51	3.44	3.38	3.34	3.30	3.30	3.23
8	3.35	3.28	3.22	3.15	3.08	3.04	3.01	3.01	2.93
9	3.14	3.07	3.01	2.94	2.86	2.83	2.79	2.79	2.71
10	2.98	2.91	2.85	2.77	2.70	2.66	2.62	2.62	2.54
15	2.54	2.48	2.40	2.33	2.25	2.20	2.16	2.16	2.07
20	2.35	2.28	2.20	2.12	2.04	1.99	1.95	1.95	1.85
25	2.24	2.16	2.09	2.01	1.92	1.87	1.82	1.82	1.72
30	2.16	2.09	2.01	1.93	1.84	1.79	1.74	1.74	1.63
40	2.08	2.00	1.92	1.84	1.74	1.69	1.64	1.64	1.52
50	2.03	1.95	1.87	1.78	1.69	1.63	1.58	1.58	1.45
70	1.97	1.89	1.81	1.72	1.62	1.57	1.50	1.50	1.36
100	1.93	1.85	1.77	1.68	1.57	1.52	1.45	1.45	1.30
∞	1.83	1.75	1.67	1.57	1.46	1.39	1.34	1.31	1.30

14.3 Appendix C: Description of Data Sets

The data sets described below can be downloaded from the Web at http://faculty.insead.edu/hubert-gatignon

Three different kinds of information, which correspond to typically available data about markets, are provided for analysis: industry, panel, and survey data. In addition, scanner data is provided for a product category in the form typically available in actual practice.

The industry data set includes aggregate product and market data for all of the brands sold in each time period. This type of information is often provided by market research services, trade and business publications, and trade associations to all of the firms competing in an industry. The other two data sets contain information collected from a sample of consumers. The first, panel data, is gathered from a group of consumers who have agreed to periodically record their brand perceptions, preferences, and purchase behavior. This information is often purchased by advertisers from syndicated research services and is useful for tracking changes in consumer behavior over time. The second, survey data, is collected by questionnaire or personal interview from a large group of consumers. Surveys are often conducted by advertising agencies (such as DDB Needham Worldwide, N. W Ayer, and others), by survey research companies, and by the advertisers themselves. These surveys typically measure a broad range of consumer characteristics, including attitudes, interests, values, and lifestyles. This information is especially useful for selecting target audiences and designing creative appeals.

The MARKSTRAT® market simulation program was used to create the industry and panel data sets. The survey data set was developed separately to conform to this environment. We first describe the MARKSTRAT® environment and the characteristics of the industry. We then present the three types of data provided with this book and discuss the contents of each data set.

14.3.1 The MARKSTRAT® Environment

To understand the industry in which competing firms operate, the reader must be familiar with two general dimensions of the MARKSTRAT® environment: (1) the structure of the industry in terms of the products, competition, and market characteristics, and (2) the marketing decisions that each firm can make over time. The discussion that follows concentrates on those aspects that are most relevant to advertising planning decisions.

14.3.1.1 Competition and Market Structure

In the MARKSTRAT® environment, five firms compete in a single market with a number of brands. Each firm starts out with a set of brands and has the ability to initiate research and development (R&D) projects to create new brands. If an R&D project is successful, then the sponsoring firm has the option of bringing the new product to the market. The firm can then modify the product marketed under a given brand name (i.e., a product improvement) or a new product can be introduced with a new brand name.

Product Characteristics

The generic products in this industry are consumer durable goods comparable to electronic entertainment products. They are called Sonites. Because these products are durable, each customer will usually purchase only one unit over a long period of time. Consequently, there are no issues of repeat purchase, brand loyalty, or brand switching in this market.

The products are characterized by five physical attributes: (1) weight (in kilograms), (2) design (measured on a relative scale), (3) volume (in cubic decimeters), (4) maximum frequency (in kilohertz), and (5) power (in watts). Not all attributes are equally important to consumers. Different consumer segments have different preferences for these product characteristics, although the preferences are expressed in terms of brand image rather than purely physical characteristics. Industry research has shown that consumers' brand evaluations in this market are a function of their perceptions of the brands on three general dimensions, related to some degree to the five physical characteristics listed above that define the product. The first and most important characteristic is the perceived price of the product. Next, consumers consider the product's power (wattage). Finally, they evaluate the product's design (aesthetic value). Although less important than the other dimensions, the product's design helps consumers to differentiate among the various competing brands. The design attribute is measured on a scale from 1 to 10 by expert judges, although consumers' perceptions may vary from these "rational" expert evaluations. To form an overall evaluation of each brand, consumers compare their brand's performance on each dimension with their preferences for a certain "ideal level" on each of these dimensions.

Because of the durability of the Sonite product and the importance of the purchase, the consumer decision process tends to follow a "high involvement" hierarchy. Measures of brand awareness, perceptions, preferences, and purchase intentions are, therefore, particularly relevant to the advertising decisions.

Consumer Segments

The consumer market for Sonites is composed of five segments with distinguishable preferences. Segment 1 consists of the "buffs," or experts in the product category. They are innovators and have high standards and requirements in terms of the technical quality of the product. Segment 2 is composed of "singles" who are relatively knowledgeable about the product but somewhat price sensitive. "Professionals" are found in segment 3. They are demanding in terms of product quality and are willing to pay a premium price for that quality. "High earners" constitute segment 4, exclusive of "professionals." These individuals are also relatively price insensitive. However, in general, they are not as educated as the professionals, and are not particularly knowledgeable about the product category. They buy the product mostly to enhance their social status. The fifth and last segment covers all consumers who cannot be grouped with any of the other four segments. They have used the product less than consumers in other segments and are considered to be late adopters of this product category. Given that this group is defined as a residual, it is difficult to characterize the members in terms of demographics or lifestyle.

Although the preferences of the five consumer segments may change over time, the composition of each segment does not. Consequently, the survey data collected in the eighth time period (described in Sect. 3.3 below) also describe consumers during the previous seven periods.

Distribution Structure

Sonites are sold through three different distribution channels. The three channels vary in terms of the proportion of the product that they sell (relative to their total product sales) and the types of clientele that they attract. Each channel carries all brands of Sonites, but the potential number of distributors within each channel and the characteristics of that channel are different. Channel 1 is made up of 3,000 specialty retail stores. These stores provide specialized services to customers, and the bulk of their sales comes from Sonites. Channel 2 consists of 35,000 electric appliance stores. These stores carry Sonite products only as an addition to their main product lines. Channel 3 represents the 4,000 department stores that exist in the MARKSTRAT® world. These stores sell a broad range of products, including clothing, furniture, housewares, and appliances.

14.3.2 Marketing Mix Decisions

A product's marketing mix reflects the marketing strategy for the brand. A brand's attributes will influence how the brand is positioned and to whom it is marketed. Its price will affect the advertising budget and the brand image. Its distribution will

determine where the brand is advertised, and so on. In this section we review the four main marketing mix variables—price, sales force, advertising, and product—that characterize brands in the MARKSTRAT® environment.

14.3.2.1 Price

Each brand of Sonite has a recommended retail price. These prices are generally accepted by the distribution channels and are passed on to consumers. The different consumer segments defined in the earlier section are more or less sensitive to price differences across brands. A segment's price sensitivity (price "elasticity") also depends on the selection of products offered to that segment and on the other marketing mix variables.

14.3.2.2 Sales Force

The two most important aspects of a firm's sales force are its size and its assignment to the three channels of distribution. Each salesperson carries the entire line of brands produced by his or her company. When a firm changes the number of salespeople it assigns to a particular channel, this is likely to affect the availability or distribution coverage of the firm's brands.

14.3.2.3 Advertising

Each brand of Sonite is advertised individually. Firms in this industry do not practice umbrella or generic (product category) advertising. However, advertising of specific brands can increase the total market demand for Sonites or affect Sonite demand in one or more segments.

Advertising can serve a number of communication purposes. It can be used to increase top-of-mind brand awareness and inform consumers about a brand's characteristics. Research has revealed that advertising expenditures are strongly positively related to brand awareness. Advertising can also have a substantial persuasive effect on consumers. Advertising can be used to position or reposition a brand so that the brand's image is more closely aligned with consumers' needs.

In addition, it is clear that advertising plays an important competitive role. One cannot consider a brand's advertising in isolation. Instead, the relative "share of voice"—the ratio of a brand's advertising expenditures to the total industry's advertising expenditures—is a better predictor of consumers' purchase behavior than absolute advertising expenditures.

Table 14.1 Names of brands marketed during each period

Firm	Brand	Period of availability
1	SALT	0–6
1	SAMA	0–6
2	SELF	0–5[a]
2	SELT	3–6
2	SEMA	4–6
2	SEMI	0–6
2	SEMU	4–6
3	SIBI	0–6
3	SICK	4–6
3	SIRO	0–3[a]
3	SIRT	4–6
4	SODA	2–6
4	SOLD	0–6
4	SONO	0–5[a]
5	SULI	0–6
5	SUSI	0–6

[a]Indicates a discontinued brand

14.3.2.4 Products

The database reports information on all of the brands of Sonites that were marketed by firms during an 8-year time period. The names of the brands sold during this period are listed in Table 14.1. This table also lists the periods during which each brand was available. Note that some of the brands were introduced after the first time period and/or were discontinued before the last (eighth) period.

The brands of Sonites are named to facilitate identification of the marketing firm. The second letter of each brand name is a vowel that corresponds to one of the five competing firms. All the brands sold by Firm I have an "A" as the second letter of the name, such as SAMA. "E" corresponds to firm 2, "I" to firm 3, "O" to firm 4, and "U" to firm 5.

During the eight time periods, each firm has the opportunity to design new products and market a portfolio of different brands. In response to consumer or market pressures, companies may change the physical characteristics of each brand over time. Information about brands and their attributes is provided in the industry data set, as described in Table 14.1.

14.3.3 Survey

A mail survey of a group of 300 consumers was conducted in the eighth (last and most recent) time period. The survey collected a variety of consumer information including demographic data, psychographics, information on product and brand purchase behavior, decision processes, and media habits. These data are

particularly useful for segmentation analysis, which is an important precursor to selecting a target market, generating advertising copy appeals, and media selection. A list of the variables from the questionnaire and the coding scheme for the items are provided in Tables 14.2 and 14.3, respectively.

14.3.4 Indup

The industry data set provides two types of performance information for each brand and time period: sales figures (in units and dollars) and market share (based on unit and dollar sales). The data set also includes information on the values of the marketing mix variables for each competing brand. The data describe each brand's price, advertising expenditures, sales force size (for each channel of distribution), and physical characteristics (i.e., the four Ps). Finally, the data set reports the variable cost of each brand in each time period. Note that this cost is not the actual current production cost, as this information is typically not available for each competing brand. The reported cost figures reflect the basic cost of production that can be estimated for a given first batch of 100,000 units at the time the brand was introduced. A list of the variables in the industry data set is given in Table 14.4.

14.3.5 Panel

The panel data set provides information that, in many ways, complements the data in the industry data set. Panel data are available at the level of the individual market segment rather than at the total market level. The panel data set includes information on the size of each segment (in unit sales of Sonites) and the market share for each brand with each segment. The data set also provides the results of a panel questionnaire with items related to advertising communication such as brand awareness and brand perceptions, and preferences. Specific variables for each consumer segment include the extent of brand name awareness, preferences in terms of the ideal levels of the three most important attributes (price, power, and design), brand perceptions on the same three attributes, and brand purchase intentions. Finally, the data set reports the shopping habits of each segment in the three channels of distribution. A summary of these variables is provided in Table 14.5.

14.3.6 Scan

The SCAN.DAT file contains a simulated sample of scanner data, similar to the refrigerated orange juice data set used in Fader and Lattin (1993); Fader, Lattin, and

Table 14.2 Survey questionnaire and scale type

Number	Abbreviation	Question	Scale
Demographics			
1	Age	Age	Continuous
2	Marital	Marital status	Categorical
3	Income	Total household income	Categorical
4	Education	Education	Categorical
5	HHSize	Household size	Continuous
6	Occupation	Occupation	Categorical
7	Location	Geographic location of household	Categorical
Psychographics			
8	TryHairdo	I often try the latest hair styles.	Likert[a]
9	LatestStyle	I usually have one or more pieces of clothing that are of the latest fashion.	Likert
10	DressSmart	An important part of my life and activities is dressing smartly.	Likert
11	BlondsFun	I really do believe that blondes have more fun.	Likert
12	LookDif	I want to look a little different from others.	Likert
13	LookAftract	Looking attractive is important in keeping your wife/husband.	liked
14	GrocShop	I like shopping.	Likert
15	LikeCooking	I love to cook and frequently do.	Likert
16	ClothesFresh	Clothes should be dried outdoors in the fresh air.	Likert
17	WashHands	It is very important for people to wash their hands before each meal.	Likert
18	Sporting	I would rather go to a sporting event than a dance.	Likert
19	LikeColors	I like bright, splashy colors.	Likert
20	FeelAffract	I like to feel attractive.	Likert
21	TooMuchSex	There is too much emphasis on sex today.	Likert
22	Social	I do more things socially than do most of my friends.	Likert
23	LikeMaid	I would like to have a maid to do the housework.	Likert
24	ServDinners	I like to serve unusual dinners.	Likert
25	SaveItems	I save items from newspapers and magazines.	Likert
26	LivingRoom	The living room is my favorite room.	Likert
27	LoveEat	I love to eat.	Likert
28	SpiritualVal	Spiritual values are more important than material things.	Likert
29	Mother	If it was good enough for my parents, it is good enough for me.	Likert
30	ClassicMusic	Classical music is more interesting than popular music.	Likert
31	Children	I try to arrange my home for my children's convenience.	Likert
32	Appliances	I enjoy having the latest technology.	Likert
33	CloseFamily	Our family is a close-knit group.	Likert
34	LoveFamily	There is a lot of love in our family	Likert
35	TalkChildren	I spend a lot of time with my children talking about their activities, friends, and problems.	Likert

(continued)

Table 14.2 (continued)

Number	Abbreviation	Question	Scale
36	Exercise	Everyone should take walks, bicycle, garden, or otherwise exercise several times a week.	Likert
37	LikeMyself	I like what I see when I look in the mirror.	Likert
38	PersonalAppear	I care about my personal appearance.	Likert
39	MedCheckup	You should have a medical checkup at least once a year.	Likert
40	EveningHome	I would rather spend a quiet evening at home than go out to a party.	Likert
41	TripWorld	I would like to take a trip around the world.	Likert
42	Homebody	I am a homebody.	Likert
43	LondonParis	I would like to spend a year in London or Paris.	Likert
44	Comfort	I furnish my home for comfort, not for style.	Likert
45	Ballet	I like classical ballet.	Likert
46	Parties	I like parties where there is lots of music and talk.	Likert
47	FoulLanguage	People should not use foul language in public.	Likert
48	BrightFun	I like things that are bright, fun, and exciting.	Likert
49	Seasoning	I enjoy spicy foods.	Likert
50	ThreeDTV	If I had to choose, I would rather have a 3D television than a new computer.	Likert
51	Sloppy	If I look sloppy, I do not feel good about myself.	Likert
Purchase behavior			
52	Smoke	How often do you smoke?	0–7
53	Gasoline	How much gasoline do you use?	0–7
54	Headache	How often do you use headache remedies?	0–7
55	Whiskey	How much whiskey do you drink?	0–7
56	Bourbon	How much bourbon do you drink?	0–7
57	FastFood	How often do you eat at fast-food restaurants?	0–7
58	Restaurants	How often do you eat at restaurants with table service?	0–7
59	OutForDinner	How often do you go out for dinner?	0–7
60	OutForLunch	How often do you go out for lunch?	0–7
61	RentVideo	How often do you rent movies?	0–7
62	Catsup	How often do you use catsup?	0–7
Purchase decision process			
63	KnowledgeSon	How much do you know about the product category of Sonites?	Likert
64	PerceiveDif	How large a difference do you perceive between various brands of Sonites?	Likert
65	BrandTrust	When purchasing (or considering purchasing) a Sonite, do you prefer to buy a brand that you know and trust or to try a new brand?	Likert
66	CategMotiv	What is your primary reason or motivation for purchasing (or considering purchasing) a Sonite (any brand in the product category)?	Categorical
67	BrandMotiv	What is your primary reason or motivation for purchasing (or considering purchasing) a particular brand of Sonite?	Categorical

(continued)

Table 14.2 (continued)

Number	Abbreviation	Question	Scale
68	OwnSonite	Do you currently own a Sonite?	0/1
69	NecessSonite	Do you feel that owning a Sonite is a necessity?	0/1
70	Otherinflnc	If you were to purchase a Sonite, would you make the decision about which brand to purchase by yourself or with the help of others?	Categorical
71	DecisionTime	If you were to purchase a Sonite, would you make the decision about which brand to purchase before going to the retail store, or would you wait until you were in the store to decide?	Categorical
Media habits			
72	ReadWomen	I read women's magazines.	0/1
73	ReadDoItYourself	I read do-it-yourself magazines.	0/1
74	ReadFashion	I read fashion magazines.	0/1
75	ReadMenMag	I read men's magazines.	0/1
76	ReadBusMag	I read business and financial magazines.	0/1
77	ReadNewsMag	I read news magazines.	0/1
78	ReadGIMag	I read general interest magazines.	0/1
79	ReadYouthMag	I read youth magazines.	0/1
80	ReadNwspaper	I read the newspaper.	0/1
81	WtchDayTV	I watch television during the day time.	0/1
82	WtchEveTV	I watch television early evening news.	0/1
83	WtchPrmTV	I watch television during prime time.	0/1
84	WtchLateTV	I watch late-night television.	0/1
85	WtchWkEndTV	I or my children watch children's programs on television during the weekend.	0/1
86	WtchModFamTV	I watch Modern Family regularly.	0/1
87	WtchBigBangTV	I watch The Big Bang Theory regularly.	0/1
88	WtchMeetMotherTV	I watch How I Met Your Mother regularly.	0/1
89	WtchSimpsonsTV	I watch The Simpsons regularly.	0/1
90	WtchNCISTV	I watch NCIS (Naval Criminal Investigative Service) regularly.	0/1
91	WtchGreyTV	I watch Grey's Anatomy regularly.	0/1
92	WtchMadMenTV	I watch Mad Men regularly.	0/1
93	WtchDancingTV	I watch Dancing with the Stars regularly.	0/1
94	WtchAbbeyTV	I watch Downton Abbey regularly.	0/1
95	WtchBowlTV	I watch the Super Bowl each year.	0/1

[a]Likert items are scaled from 1 = Disagree to 7 = Agree

Table 14.3 Coding of variables

Variable	Category	Code
Question #2		
Marital status	Married	1
	Widowed	2
	Divorced	3
	Separated	4
	Single	5
Question #3		
Household income	Less than $20,000	1
	$20,000–$39,999	2
	$40,000–$59,999	3
	$60,000–$79,999	4
	$80,000–$99,999	5
	$100,000–$119,999	6
	$120,000–$139,999	7
	$140,000–$159,999	8
	$160,000–$179,999	9
	$180,000–$199,999	10
	$200,000–$219,999	11
	$220,000 and over	12
Question #4		
Education level	Did not attend school	1
	Graduated from elementary school	2
	Went to secondary school for less than 4 years	3
	Graduated from secondary school or trade school	4
	Some college, Jr. college, or technical school	5
	Graduated from college	6
	Have postgraduate degree	7
Question #6		
Occupation	Legislators, senior officials, and managers	1
	Professionals	2
	Technicians and associate professionals	3
	Clerks	4
	Service workers and shop and market sales workers	5
	Skilled agricultural and fishery workers	6
	Craft and related trade workers	7
	Plant and machine operators and assemblers	8
	Elementary occupations	9
	Armed forces	0
Question #7		
Location	New York City	1
	Los Angeles	2
	Chicago	3
	Philadelphia	4
	San Francisco	5
	Boston	6
	Detroit	7

(continued)

Table 14.3 (continued)

Variable	Category	Code
	Dallas	8
	Washington, DC	9
	Houston	10
	Cleveland	11
	Atlanta	12
	Pittsburgh	13
	Miami	14
	Minneapolis–St. Paul	15
	Seattle–Tacoma	16
	Tampa–St. Petersburg	17
	St. Louis	18
	Denver	19
	Sacramento–Stockton	20
Question #66		
Category purchase motivation	To solve (remove) a problem	1
	To avoid having a problem	2
	To replace another Sonite	3
	For sensory stimulation	4
	For intellectual stimulation	5
	For social approval	6
	To enhance my self-esteem	7
Question #67		
Brand purchase motivation	To solve (remove) a problem	1
	To avoid having a problem	2
	Because of dissatisfaction with my current brand	3
	For sensory stimulation	4
	For intellectual stimulation	5
	For social approval	6
	To enhance my self-esteem	7
Question #70		
Decision making	By myself (individually)	1
	With the help of others (as a group)	2
Question #71		
Decision timing	Before going to the store	1
	In the store	2
Coding for other variables		
Questions	*Scale*	
8–51	Disagree 1 2 3 4 5 6 7 Agree	
63–65		
52–62	Never/none 0 1 2 3 4 5 6 7 Very often/a lot	
68 and 69	0 = No; 1 = Yes	
72–95		

Table 14.4 Variables in industry-level database

Abbreviation	Variable
Period	Period number
Firm	Firm number
Brand	Brand name
Price	Price
Adver	Advertising expenditures
Char0l	Product characteristic #1: Weight (kg)
Char02	Product characteristic #2: Design (Index)
Char03	Product characteristic #3: Volume (dM3)
Char04	Product characteristic #4: Maximum frequency (kHz)
Char05	Product characteristic #5: Power (W)
Salesmenl	Number of salesmen-channel 1
Salesmen2	Number of salesmen-channel 2
Salesmen3	Number of salesmen-channel 3
Cost	Average unit cost of initial batch
Dist0l	Number of distributors-channel 1
Dist02	Number of distributors-channel 2
Dist03	Number of distributors-channel 3
UnitSales	Total sales in units
DolSales	Total sales in dollars
UnitShare	Market share (based on units)
DolShare	Market share (based on dollars)
AdShare	Advertising share (share of voice)
RelPrice	Relative price (price relative to average market price)

Table 14.5 Variables in panel database

Abbreviation	Variable
Period	Period number
Segment	Segment number
SegSize	Segment size (unit sales in segment)
Ideal01	Ideal value of price (for each segment)
ldeaI02	Ideal value of power (for each segment)
IdeaI03	Ideal value of design (for each segment)
Brand	Brand name
Awareness	Percentage of segment aware of the brand
Intent	Purchase intent (for each brand and segment)
Shop01	Percentage of segment shopping in channel 1
Shop02	Percentage of segment shopping in channel 2
Shop03	Percentage of segment shopping in channel 3
Perc01	Perception of price (for each brand)
Perc02	Perception of power (for each brand)
Perc03	Perception of design (for each brand)
Dev01	Deviation from ideal price (for each brand in each segment)
Dev02	Deviation from ideal power (for each brand in each segment)
Dev03	Deviation from ideal design (for each brand in each segment)
Share	Segment share (for each brand)

Little (1992); and Hardie, Johnson, and Fader (1992). These articles (listed in the Bibliography section of Chap. 8) give a full description of a similar data set. The six brands along with their brand id codes are as follows:

1	Brand 1
2	Brand 2
3	Brand 3
4	Brand 4
5	Brand 5
6	Brand 6

This "SCAN.DAT" data file is set up for the estimation of the standard Guadagni and Little (G&L, 1983) multinomial logit (MNL) model of brand choice, including their "loyalty" variable. The value of the smoothing constant used to calculate the loyalty variable is set to 0.8, and the loyalty variable is initialized using purchase information for weeks 1 through 52.

In this data set, the number of choice alternatives varies over time (due to shopping at different stores, stock-outs, etc.). Rather than having a single record per purchase occasion, we have as many records as they are choice alternatives at one purchase occasion of a consumer.

The format of SCAN.DAT is as follows:

- Panelist id
- Week of purchase
- A dummy variable indicating whether this record is associated with the brand chosen
- The number of brands available (records) associated with this purchase occasion
- The brand id of this record
- Regular shelf price for this brand
- Any price reduction for this brand on this purchase occasion (price paid = price − price cut)
- A dummy variable indicating the presence of a feature ad for this brand
- The value of the Guadagny and Little loyalty variable for this brand (on this purchase occasion)
- A brand-specific constant/dummy for brand 1
- A brand-specific constant/dummy for brand 2
- A brand-specific constant/dummy for brand 3
- A brand-specific constant/dummy for brand 5
- A brand-specific constant/dummy for brand 6

Therefore, given that there is no dummy variable for brand 4 (a private label), this brand becomes the reference brand.

The LIMDEP file "examp8-2.lim"and the STATA file "examp8-2.do" in Chap. 8 contain sample commands for reading this data set with LIMDEP and STATA, respectively.

Index

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4614-8594-0,