Chapter 1

Introduction

The wisest mind has something yet to learn.

—George Santayana

Overview

- Plan of the book
- Notation
- Computer programs for SEM
- Statistical journeys
- Family values
- Family history

Plan of the book

- Part I (Fundamental Concepts):
  - Chap. 1 — Introduction
  - Chap. 2 — Review of regression basics, statistical tests, bootstrapping
  - Chap. 3 — Data screening and preparation, score reliability and validity
  - Chap. 4 — Overview of core SEM techniques, SEM computer programs

Plan of the book

- Part II (Core Techniques):
  - Chap. 5 — Introduction to path analysis
  - Chap. 6 — Details of path analysis, basic rationale of model testing
  - Chap. 7 — Confirmatory factor analysis measurement models, analysis of nonnormal data
  - Chap. 8 — Analysis of models with both structural and measurement components

Plan of the book

Part III (Advanced Techniques, Avoiding Mistakes):

- Chap. 9 — Analysis of nonrecursive structural models
- Chap. 10 — Analysis of means, latent growth models
- Chap. 11 — Multiple-sample SEM
- Chap. 12 — How to fool yourself with SEM
- Chap. 13 — Brief overviews of multilevel SEM, estimation of curvilinear or interactive effects of observed or latent variables

Notation

- As with other statistical techniques, there is no "gold standard" for notation in SEM
- The symbol set associated with the LISREL program is probably the most widely used in books and journal articles about SEM
- It features many double-subscripted Greek letters (e.g., ϕ_{31}, Λ_{31}, etc.) and matrix algebra
- This can be confusing to follow unless one has memorized the entire system
- See Schumaker and Lomax (1996, chap. 11) for a concise review of LISREL notation and a matrix approach to modeling

Notation

- A minimum number of alphabetic characters are used here to represent various aspects of SEM, such as observed versus latent variables
- Also, to avoid double notation, a distinction is not usually made between population values and sample statistics
- It is assumed that readers are already aware of this difference and know that sample data can only be considered as estimates of population values

Computer programs for SEM

- Computer programs are necessary tools for the conduct of SEM
- About 30 years ago, LISREL was essentially the only widely-available SEM program
- However, the situation is now very different as there are many other choices of SEM computer programs, including:

Amos, CALIS, EQS, Mplus, Mx, RAMONA, SEPATH, among others

- For this reason, the presentation of concepts here is *not* linked to a particular SEM computer program
- Instead, essential principles of SEM that users of *any* computer tool must understand are emphasized

Computer programs for SEM

- SEM computer programs were traditionally difficult to use for two reasons: They

  1. usually required users to generate a lot of rather arcane syntax for each analysis

  2. tended to be available only on mainframe computers (i.e., batch-file-type programming, command-line user interfaces)

- However, statistical software programs for personal computers based on graphical user interfaces can be much easier to use, and this includes modern programs for SEM

Computer programs for SEM

- In some SEM computer programs, such as Amos, EQS, LISREL, and Mx Graph, the user can specify the model by drawing it on the screen

- The computer then translates the figure into lines of code, which are then run to generate the output

- Thus, the user need not know very much about the writing of program code in order to conduct a sophisticated type of analysis

- This also means that the importance of highly technical programming skills for the conduct of SEM is likely to diminish even further

Computer programs for SEM

- For researchers who have a good understanding of the fundamental concepts of SEM, the increasing ease of use of computer tools is a boon
- However, low-effort programming could encourage the use of SEM in uninformed or careless ways
- *It is thus more important than ever to be familiar with the conceptual and statistical bases of SEM!*
- Computer programs, however easy to use, should be only the tools of your knowledge and not its master

Computer programs for SEM

- Steiger (2001): The emphasis on the ease of use of computer tools can give beginners the false impression that SEM itself is easy
- That is, to beginners it may appear that all one has to do is draw the model on the screen and let the computer take care of everything else
- However, the reality is that things can and do often go wrong in SEM

Computer programs for SEM

- Specifically, beginners often quickly discover that analyses fail because of technical problems, including
  - computer system crashes
  - terminated program runs with many error messages
  - uninterpretable output

Computer programs for SEM

- The kinds of problems just mentioned happen in part because real research problems can be very technical
- The availability of user-friendly SEM computer tools cannot change this fact
- Accordingly, one of the goals of the book is to point out the kinds of things that can go wrong in SEM and how to deal with them
- For the same reason, not all examples of SEM analyses described in the book are free of problems

Statistical journeys

- Kühnel (2001) noted that learning about SEM has the by-product that students must deal with many fundamental issues of methods and general statistical concepts
- One of these issues is measurement
- Specifically, one cannot analyze a structural equation model with latent variables that represent hypothetical constructs without thinking about how those constructs are to be measured

Statistical journeys

- Measurement theory is too often neglected nowadays in undergraduate and graduate degree programs in psychology (e.g., Frederich, Buday, & Kerr, 2000) and related areas
- Many undergraduate psychology programs no longer even offer courses in measurement
- *However, correct use of SEM requires strong knowledge about measurement*

Statistical journeys

- SEM is generally a priori, which means that the researcher must specify a model in order to conduct the analysis
- The model's specification must have some basis, whether it be
  - theory
  - results of previous studies
  - educated guesses that reflect the researcher's domain knowledge and experience

Statistical journeys

- The emphasis on the testing of a whole model may also be a kind of antidote against the over-reliance on statistical tests of individual hypotheses
- That is, increased use of model-fitting techniques, such as SEM, can be part of the reform of data analysis methods in the behavioral sciences (Kline, 2004)
- Best practices in SEM also require that the researcher consider alternative models that may explain the same data equally well

Family values

- The term “SEM” does not designate a single statistical technique but instead refers to a family of related procedures

- Shared characteristics of techniques in the SEM family include:

  1. SEM is a priori and requires researchers to think in terms of models, but a priori does not mean exclusively confirmatory—many applications of SEM are a blend of exploratory and confirmatory analyses

Family values

- Shared characteristics include:

  2. The explicit representation of the distinction between observed and latent variables is characteristic of many structural equation models

     a. However, it is not necessary to have in your models latent variables that represent hypothetical constructs

     b. There is more than one type of latent variable, each of which reflects different assumptions about the relation between observed and unobserved variables

Family values

- Shared characteristics include:

  2. The explicit representation of the distinction between observed and latent variables

     c. Latent variables in SEM can represent a wide range of phenomenon, such as characteristics of persons or variables (e.g., method effects)

     d. The observed-latent distinction also provides a way to take account of imperfect score reliability

Family values

- Shared characteristics include:

  3. The basic statistic in SEM is the covariance—however, it is possible to analyze other types of data, such as means

     a. The technique of SEM incorporates the types of analyses traditionally associated with ANOVA including between-groups and within-groups (e.g., repeated measures) mean comparisons

     b. It is also possible in SEM to test for group mean differences on latent variables, something that is not really feasible in standard ANOVA

Family values

- Shared characteristics of techniques in the SEM family include:

4. The technique of SEM is not just for nonexperimental (correlational) data—instead, it is a very flexible analytical tool that can be applied to data from experiments, too

5. Many standard statistical procedures, including the analysis of variance (ANOVA) and its multivariate counterpart (MANOVA), can be viewed as special cases of SEM

Family values

- Shared characteristics include:

  6. SEM is *still* generally a large-sample technique—some guidelines about sample size requirements:

     a. With $N < 100$, almost any type of SEM analysis may be untenable unless a very simple model is evaluated

     b. For descriptive purposes, sample sizes less than 100 could be considered “small”

     c. Between 100 and 200 subjects—a “medium” sample size—is a better minimum, but model complexity must be considered

Family values

- Shared characteristics include:

    6. Some guidelines about sample size requirements:

        d. Sample sizes that exceed 200 cases could be considered "large," but again model complexity must be considered

        e. Breckler (1990) found a median sample size of about *N* = 200 across 72 published studies

Family values

- Shared characteristics include:

    7. It is possible to test many different types of effects for statistical significance in SEM, but the role of statistical tests in the overall analysis may be less important compared with more traditional techniques

Extended latent variable families

- Latent variables analyzed in SEM are generally assumed to be continuous
- There are other techniques that specialize in the analysis of models with categorical latent variables
- The levels of a categorical latent variable are called *classes*, and they represent a mixture of subpopulations where membership is not known but is inferred from the data (i.e., the observed variables)
- In other words, a goal of the analysis is identify the nature and number of latent classes

Extended latent variable families

- The technique of *latent class analysis* is a type of factor analysis but for categorical observed and latent variables (e.g., Hagenaars & McCutcheon, 2002)
- A special type of latent class factor model that represents the shift from one of two different states, such as from nonmastery to mastery of a skill, is a *latent transition model*
- There are also analogues of techniques, such as multiple regression, for the analysis of categorical latent variables
- In *latent class regression*, a criterion is predicted by estimated class membership and other variables that covary with class membership

Extended latent variable families

- In contrast to standard regression techniques for continuous variables, the predictors in latent class regression can be a mix of continuous, discrete, or count variables
- Also, the criterion can be continuous, discrete, or a repeated measures variable
- It is also not assumed in latent class regression that the same predictor model holds for all cases

Extended latent variable families

- Until recently, SEM for the analysis of continuous latent variables and techniques like those just mentioned for the analysis of categorical latent variables were viewed as relatively distinct
- However, this view is changing because of recent attempts to express all latent variable models within a common mathematical framework (e.g., Bartholomew, 2002)

Extended latent variable families

- For example, B. Muthén (2001) describes the analysis of mixture models with latent variables that may be continuous or categorical
- When both are present in the same model, the analysis is basically a SEM conducted across different inferred subpopulations
- The Mplus computer program can analyze all basic kinds of SEM models and mixture models, too
- Some computer tools, such as Mplus, blur the distinction between SEM and techniques for categorical latent variables, such as latent class analysis and latent class regression

Family history

- Because SEM is a collection of related techniques, it does not have a single source
- Part of its origins date to the turn of the last century with the development of what we now call exploratory factor analysis, usually credited to Charles Spearman

Family history

- Only a few years later, Sewall Wright, a geneticist, developed the basics of path analysis
- Path analysis was later introduced by various authors to researchers in other disciplines such as sociology, economics, and later, psychology
- An annotated bibliography by Wolfle (2003) traces the introduction of path analysis to the social sciences

Family history

- These early measurement (factor analysis) and structural (path analysis) approaches were integrated in the early 1970s in the work of K. G. Jöreskog, J. W. Keesling, and D. E. Wiley

- Bentler (1980) referred to this body of work as the JKW model

- One of the first widely-available computer programs able to analyze models based on the JKW framework (now called SEM) was LISREL, developed by K. G. Jöreskog and D. Sörbom in the 1970s

Family history

- The 1980s and 1990s witnessed the development of many more computer programs and a rapid expansion of the use of SEM techniques in more and more areas
- Some of these areas include developmental psychology, behavioral genetics, sports medicine, education, and public health, to name just a few (e.g., Hershberger, 2003)

Family history

- The increase in the use of SEM during this time (1980s onward) could be described as exponential
- *However, there are some troubling aspects about the way SEM is often applied that limit its potential*
- These shortcomings in the way is SEM is used in practice are addressed throughout the book

References

Bartholomew, D. J. (2002). Old and new approaches to latent variable modeling. In G. A. Marcoulides and I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 1-13). Mahwah, NJ: Lawrence Erlbaum.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31*, 419-456.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, *107*, 260-273.

Frederich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology*, *27*, 248-257.

Hagenaars, J. A., & McCutcheon, A.L. (Eds.). (2002). *Applied latent class analysis*. Cambridge, MA: Cambridge University Press.

Hershberger, S. L. (2003). The growth of structural equation modeling: 1994-2001. *Structural Equation Modeling*, *19*, 35-46.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, D.C.: American Psychological Association.

Kühnel, S. (2001). The didactical power of structural equation modeling. In R. Cudeck, S. du Toit, and D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 79-96). Lincolnwood, IL: Scientific Software International.

Muthén, B. (2001). Latent variable mixture modeling (pp. 1-33). In G. A. Marcoulides and R. E. Schumaker (Eds.), *New developments and techniques in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.

Schumaker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association*, *96*, 331-338.

Wolfle, L. M. (2003). The introduction of path analysis to the social sciences, and some emergent themes: An annotated bibliography. *Structural Equation Modeling*, *10*, 1-34.