Chapter 5

Introduction to Path Analysis

Put simply, the basic dilemma in all sciences is
that of how much to oversimplify reality.

—H. M. Blalock

Overview

- Correlation and causation
- Specification of path models
- Types of path models
- Principles of identification
- Overview of estimation options
- Maximum likelihood estimation

Correlation and causation

- A substantial correlation could indicate only a spurious association
- It is also true that an observed correlation of about zero does not preclude a true causal effect (e.g., suppression)
- In path analysis (PA), the researcher specifies a model that attempts to explain why *X* and *Y* (and other observed variables) covary

Correlation and causation

- Part of the explanation about why two variables covary may include presumed causal effects (e.g., *X* causes *Y*)
- Other parts of the explanation may reflect presumed noncausal relations, such as a spurious association
- The overall goal in PA is to estimate causal versus noncausal aspects of observed covariances

Correlation and causation

- To reasonably infer that *X* is a cause of *Y*, all of the following conditions must be met:

    1. There is time precedence, that is, *X* precedes *Y* in time

    2. The direction of the causal relation is correctly specified, that is, *X* causes *Y* instead of the reverse or that *X* and *Y* cause each other

    3. The association between *X* and *Y* does not disappear when external variables, such as common causes of both, are held constant (i.e., it is not spurious)

- It is very unlikely that all these conditions would be satisfied in a single study

Correlation and causation

- The assessment of variables at different times at least provides a measurement framework consistent with the specification of directional causal effects
- However, longitudinal designs pose many potential difficulties, such as subject attrition and the need for additional resources
- Probably because of these problems, most path analytic studies feature concurrent rather than longitudinal measurement

Correlation and causation

- When the variables are concurrently measured, it is not possible to demonstrate time precedence

- Therefore, the researcher needs a very clear, substantive rationale for specifying that *X* causes *Y* instead of the reverse or that *X* and *Y* mutually influence each other when all variables are measured at the same time

Correlation and causation

- It is only from a solid base of knowledge about theory and research that one can even begin to address these requirements for inferring causation from correlation

- It is best to take the view that just as correlation does not imply causation, statistical causal modeling does not prove causation either

- This is why Wilkinson and the Task Force on Statistical Inference (1999) emphasized that use of SEM computer programs

  "...rarely yields any results that have any interpretation as casual effects" (p. 600)

Specification of path models

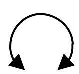
- The specification error of omitting causal variables from a path model has the same potential consequence as omitting predictors from a regression equation
- That is, estimates of causal effects of included variables may be inaccurate if there are omitted causal variables that covary with the included variables
- The direction of this inaccuracy could be either underestimation of true causal effects or overestimation, depending upon the correlations between included and excluded variables
- Underestimation probably occurs more often than overestimation

Specification of path models

- In path analysis (PA), there is only one observed measure of each construct (i.e., a single indicator)
- It is therefore crucial that measures have good psychometric characteristics
- Score reliability is especially critical because one assumption of PA is that the exogenous variables are measured without error
- The general consequence of error-prone measures of either exogenous or endogenous variables in PA is that the statistical estimates of presumed causal effects may be inaccurate

Specification of path models

- Basic elements of path models:

	Symbol	Interpretation
1.	X	Observed exogenous variable
2.	Y	Observed endogenous variable
3.	D	Unobserved exogenous variable (i.e., a disturbance)
4.		Variance of exogenous variable
5.		Covariance between a pair of exogenous variables
6.	$\rightarrow$	Presumed direct causal effect (e.g., $X \rightarrow Y$)
7.	$\rightleftarrows$	Presumed reciprocal causal effects (e.g., $Y_1 \rightleftarrows Y_2$)

Specification of path models

- An account of why two observed variables covary in a path model can reflect two kinds of presumed causal relations:

  1. Unidirectional:

     a. *Direct effects* of one variable on another—example:

     $$X \rightarrow Y$$

     b. *Indirect effects* through mediating variables—example:

     $$X \rightarrow Y_1 \rightarrow Y_2$$

Specification of path models

- Two kinds of presumed causal relations:

  2. Feedback loops: Mutual influence among variables measured at the same time including

     a. *direct feedback* involving two variables in a reciprocal relation—example:

     $$Y_1 \rightleftarrows Y_2$$

     b. *indirect feedback* involving three or more variables—example:

     $$Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_1$$

Specification of path models

- An account of why two observed variables covary with a path model can also reflect two kinds of presumed noncausal associations,

    1. unanalyzed

    2. spurious

- Recall that the specification of an unanalyzed association means that it is unknown why two variables covary

- Path models with more than one observed exogenous variable almost always assume unanalyzed associations between them (e.g., $X_1 \leftrightarrow X_2$)

Specification of path models

- It is also possible to specify unanalyzed associations between disturbances (e.g., $D_1 \leftrightarrow D_2$)

- Recall that a disturbance represents all omitted causes of the corresponding endogenous variable

- An unanalyzed association between a pair of disturbances is called a *disturbance correlation* (for standardized variables) or a *disturbance covariance* (for unstandardized variables)

- The term "disturbance correlation" is used from this point whether the variables are standardized or not

Specification of path models

- A disturbance correlation reflects the assumption that the corresponding endogenous variables share at least one common omitted cause

- Accordingly, the *absence* of the symbol for an unanalyzed association between two disturbances reflects the

  1. presumption of independence among the unmeasured causes

  2. hypothesis that the observed correlation between that pair of endogenous variables can be entirely explained by other observed variables in the model

Specification of path models

- Unlike unanalyzed associations between measured exogenous variables, the inclusion of disturbance correlations in a structural model is not so simple (elaborated later)
- It is also theoretically possible to specify an unanalyzed association between a measured exogenous variable and a disturbance (e.g., $X \leftrightarrow D$)
- Such a correlation would imply the presence of an omitted variable that causes both *X* and the corresponding endogenous variable

Specification of path models

- However, it is usually assumed that measured exogenous variables and disturbances are unrelated

- Some reasons:

  1. When multiple regression is used to estimate recursive path models (defined later), it must be assumed that the predictors (observed exogenous variables) and residuals (disturbances) are uncorrelated

  2. Assuming the independence of disturbances and observed exogenous variables permits the estimation of direct effects of the latter (e.g., $X \rightarrow Y$), holding omitted causal variables constant

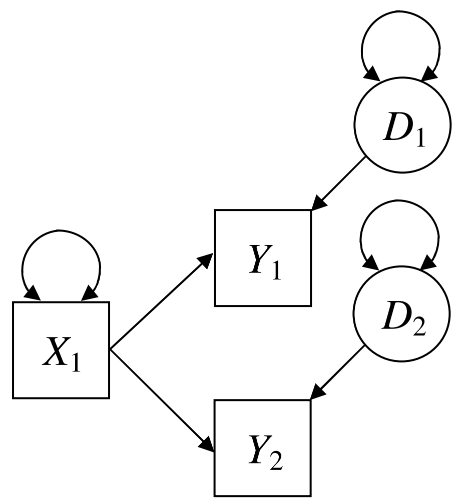
Specification of path models

- Bollen (1989) refers to the second reason just given as *pseudo-isolation*, an assumption that is probably violated in most applications of SEM
- The seriousness of violating this assumption increases with the magnitudes of the correlations between excluded and included variables

Specification of path models

- Spurious associations are the other kind of noncausal associations
- They are represented in path models by specifying common causes
- Examples of spurious associations:

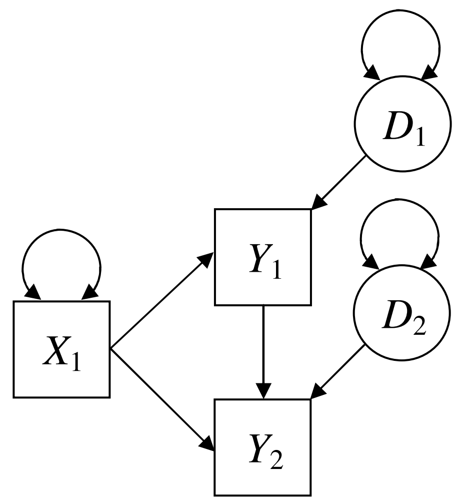
  1. The entire observed association between Y_1 and Y_2 is predicted to be spurious:



Specification of path models

- Examples of spurious associations:

  2. Only part of the observed association between Y_1 and Y_2 is predicted to be spurious, and the direct effect $Y_1 \rightarrow Y_2$ is estimated controlling for their common cause:



- Spurious associations can also involve multiple common causes

Specification of path models

- A path model can only be so complex relative to the information available to estimate it
- Specifically, the number of parameters that can be represented in a path model is limited by the number of observations
- In PA, the number of *observations* is the number of variances and unique (nonredundant) covariances among the observed variables

Specification of path models

- If v is the number of observed variables, the number of observations equals $v\,(v + 1)/2$

- The number of observations remains the same regardless of the sample size

- Adding *cases* does not increase the number of observations—only adding *variables* can do so

Specification of path models

- The number of model parameters that require a statistical estimate cannot exceed the number of observations
- A model may have fewer parameters than observations, but no more
- The difference between the number of observations and the number of its parameters are the *model degrees of freedom*, df_M
- The general rule for counting the number of parameters in a path model is:

> *The total number of variances and covariances (i.e., unanalyzed associations) of exogenous variables and direct effects on endogenous variables from other observed variables equals the number of parameters*

Specification of path models

- The status of each model parameter is free, fixed, or constrained:

    1. *Free:* to be estimated by the computer with sample data

    2. *Fixed:* specified to equal a constant such as zero

    3. *Constrained:* estimated by the computer within some restriction, but it is not fixed to equal a constant

Specification of path models

- Types of constraints on parameter estimates:

  1. *Equality:* estimates of two or more parameters are forced to be equal

  2. *Proportionality:* forces one parameter estimate to be some proportion of another (e.g., twice as large)

  3. *Nonlinear constraint:* imposes a nonlinear relation between two parameter estimates (e.g., one must be the square of another)

  4. *Inequality constraint:* forces the value of a parameter estimate to be less than or greater than a specified value (e.g., > 5.00)

Specification of path models

- It is common in SEM to test hypotheses by specifying that a previously fixed-to-zero parameter becomes a free parameter or vice-versa
- Results of such analyses may indicate whether to respecify a model by making it more complex (an effect is added) or more parsimonious (an effect is dropped)

Specification of path models

- In a multiple-sample SEM analysis, a *cross-group equality constraint* forces the computer to derive equal estimates of that parameter across all groups
- This specification corresponds to the null hypothesis that the parameter is equal in all populations from which the samples were drawn (chap. 11)

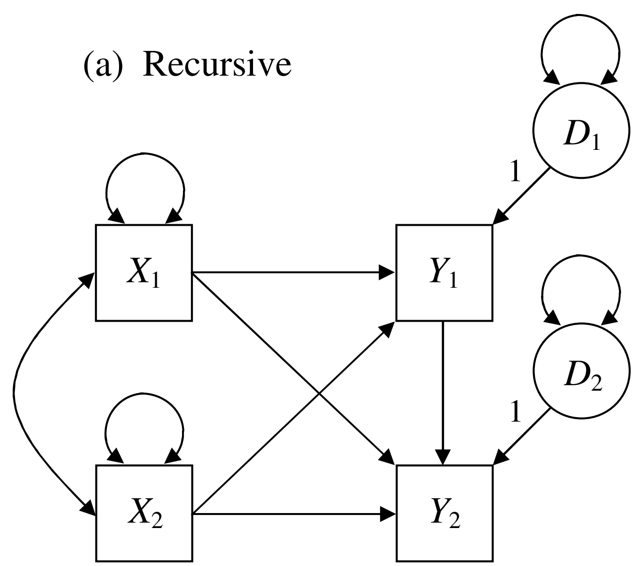
Types of path models

- *Recursive models* are the most straightforward and have two basic features:

    1. The disturbances are uncorrelated

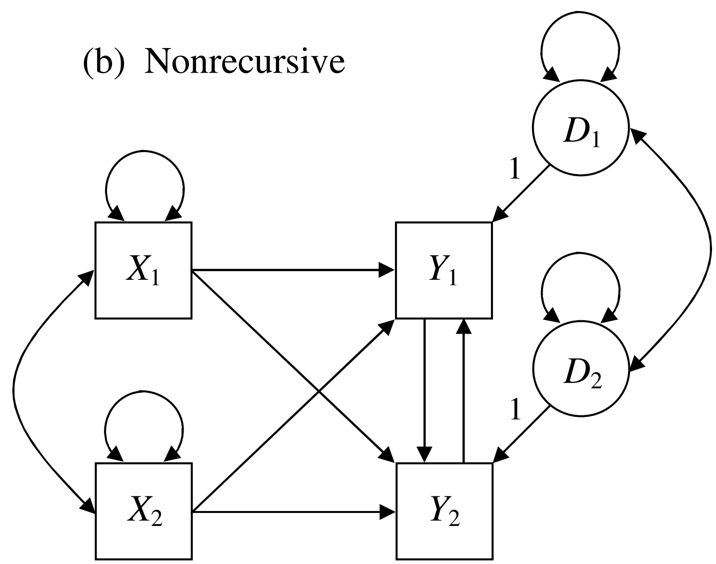
    2. All causal effects are unidirectional

- Example (Figure 5(a)):



Types of path models

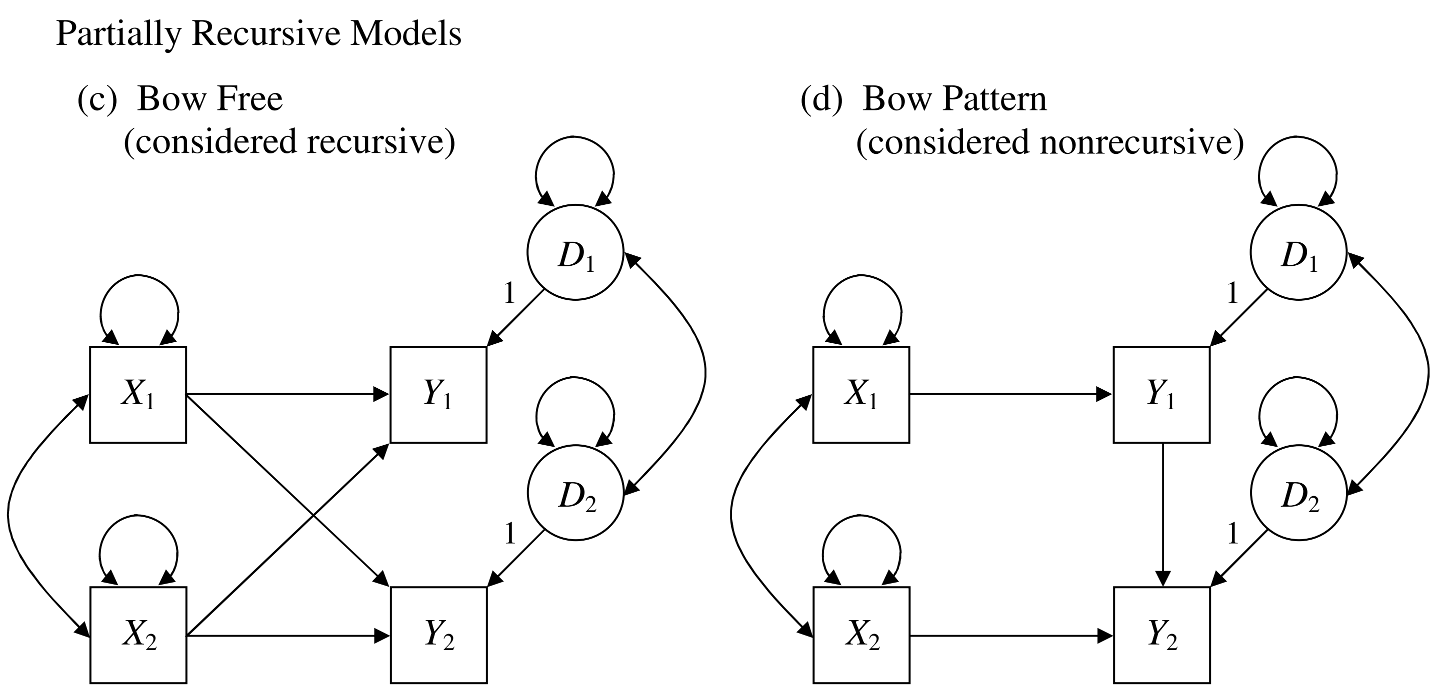
- *Nonrecursive models* have feedback loops or may have correlated disturbances
- Example (Figure 5(b)):



Types of path models

- There is another type of path model, one that has directional effects and correlated disturbances

- Examples (Figures 5.1(c) and 5.2(d)):



Types of path models

- The classification of structural models with directional effects and correlated disturbances in the SEM literature is not consistent

- Some authors call these models nonrecursive, but others use the term *partially recursive*

- But more important than the label is the distinction made in Figure 5.1—partially recursive models with a

  1. bow-free pattern of disturbance correlations can be treated in the analysis just like recursive models (e.g., Figure 5.1(c))

  2. bow pattern of disturbance correlations must be treated in the analysis as nonrecursive models (e.g., Figure 5.1(d))

Types of path models

- A *bow-free pattern* means that correlated disturbances are restricted to pairs of endogenous variables *without* direct effects between them (e.g., Figure 5.1(c))

- A *bow pattern* means that a disturbance correlation occurs *with* a direct effect between the endogenous variables (e.g., Figure 5.1(d); Brito & Pearl, 2003; Kenny, 1979)

- All ensuing references to recursive and nonrecursive models include, respectively, partially recursive models without and with direct effects among the endogenous variables

Types of path models

- Some implications of the distinction between recursive and nonrecursive path models:

  1. The assumptions of recursive models that all causal effects are unidirectional and that the disturbances are independent when there are direct effects among the endogenous variables simplify the statistical demands for their analysis—example:

     Multiple regression can be used to estimate path coefficients for direct effects and disturbance variances in recursive models

Types of path models

- Some implications of the distinction between recursive and nonrecursive path models:

    2. The same assumptions of recursive models that ease the analytical burden are also very restrictive, however—example:

        Causal effects that are not unidirectional (e.g., as in a feedback loop) cannot be represented in a recursive model

Types of path models

- Some implications of the distinction between recursive and nonrecursive path models:

   3. Nonrecursive models cannot be analyzed with standard multiple regression

   4. Nonrecursive models require more sophisticated methods and may also require additional assumptions

   5. The likelihood of a problem in the analysis of a nonrecursive model is also greater than for a recursive model

Types of path models

- Perhaps due to the difficulties just mentioned, one sees relatively few nonrecursive path models in the behavioral science literature

- But in some disciplines, especially economics, they are much more common, which suggests that the challenges of nonrecursive models can be overcome (chap. 9)

Principles of identification

- A model is said to be *identified* if it is theoretically possible to derive a unique estimate of each parameter; if not, the model is not identified
- The word "theoretically" emphasizes identification as a property of the model and not of the data
- For example, if a model is not identified, then it remains so regardless of the sample size
- Therefore, models that are not identified should be respecified; otherwise, attempts to analyze them may be fruitless

Principles of identification

- There are two basic requirements for the identification of any kind of structural equation model:

  1. There must be at least as many observations as free model parameters (i.e., $df_M \geq 0$)

  2. Every unobserved (latent) variable must be assigned a scale (metric)

- Models that violate the first requirement are not identified; specifically, they are *underidentified*

Principles of identification

- Assuming that all latent variables have been assigned a scale, it is impossible to specify a recursive path model with more parameters than observations

- In fact, recursive path models are always identified (e.g., Bollen, 1989, pp. 95-98)

- However, data-related problem can cause the analysis of a recursive path model (or other kinds of structural equation models) to fail

- For example, multicollinearity can result in *empirical underidentification* (Kenny, 1979)

- Inaccurate *start values* (initial parameter estimates) can also cause the analysis to fail (explained later)

Principles of identification

- The identification status of nonrecursive path models is more ambiguous
- This is because particular configurations of paths in a nonrecursive model can make it nonidentified even if there are as many observations as parameters and every latent variable has a scale
- There are relatively straightforward ways to determine whether some (but not all) types of nonrecursive models are identified (chap. 9)

Principles of identification

- One way to remedy an identification problem of a nonrecursive model is to add exogenous variables, but this typically can be done only before the data are collected
- Thus, it is crucial to evaluate whether a nonrecursive model is identified right after it is specified and before the study is conducted

Principles of identification

- A *just-identified* model has equal numbers of parameters and observations and is identified
- Because a nonrecursive model with equal numbers of parameters and observations is not necessarily identified, this term does not automatically apply to it
- But it does apply to recursive models with the same property
- A just-identified path model will perfectly fit the data, but it tests no particular hypothesis

Principles of identification

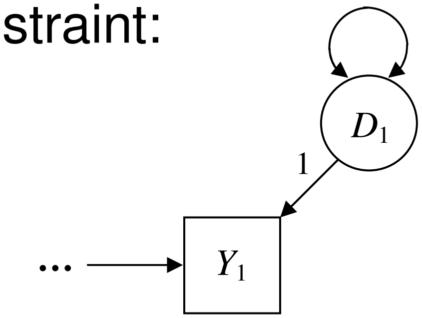
- If a path model is identified and has fewer parameters than observations, then it is *overidentified*
- Although it is possible to find a unique solution for overidentified models, it may not perfectly reproduce the data
- This characteristic has an important role in model testing, one that is elaborated later

Principles of identification

- Recall that the second general requirement for identification is that every latent variable must be assigned a scale
- This is necessary in order for the computer to be able to calculate estimates of effects that involve latent variables
- Disturbances are the only kind of latent variable in path models
- Scales for disturbances are usually assigned through a *unit loading identification* (ULI) *constraint*
- This means that the unstandardized path coefficient for the direct effect of a disturbance—also called an *unstandardized residual path coefficient*—is fixed to equal 1.0

Principles of identification

- Example of a ULI constraint:



- The specification $D_1 \rightarrow Y_1 = 1.0$ has the consequence of assigning to D_1 a scale related to that of the unexplained variance of Y_1

- Because the unstandardized residual path coefficient is here a fixed parameter (1.0), the computer needs only to estimate the variance of D_1

Principles of identification

- There is another method to scale latent variables that is discussed in Chapter 7, but it is rarely applied to disturbances in path models
- Most SEM computer programs also make it easier to specify a ULI constraint for disturbances

Overview of estimation options

- There are two basic options for analyzing recursive path models:

    1. Multiple regression (MR)

    2. Estimation with a SEM computer program, which typically offers the choice of different estimation methods

Overview of estimation options

- The most widely used estimation method in SEM is maximum likelihood (ML) estimation
- ML estimation is the default method in most SEM computer programs
- For just-identified recursive path models, MR and ML estimation yield identical estimates of direct effects (path coefficients)
- Values of path coefficients for overidentified recursive path models may be slightly different, but MR and ML estimation generally yield similar results in large samples

Overview of estimation options

- Some reasons why it is worth the effort to learn how to use a SEM computer program:

    1. There are many statistical indexes of the overall fit of the model to the data that are available in the output of SEM computer programs that are not generated by regression programs

    2. There are types of results that are automatically calculated by SEM computer programs that must be derived by hand when one uses a regression program

    3. Nonrecursive path models and other kinds of structural equation models where latent variables represent hypothetical constructs can be estimated with ML, but not MR

Maximum likelihood estimation

- ML estimates are the ones that maximize the likelihood (the continuous generalization) that the data (the observed covariances) were drawn from this population

- It is a *normal theory* method because ML estimation assumes that the population distribution for the endogenous variables is multivariate normal

- Other methods are based on different parameter estimation theories, but they are not currently used as often

- In fact, the use of an estimation method other than ML requires explicit justification (Hoyle, 2000)

Maximum likelihood estimation

- Most forms of ML estimation in SEM are *simultaneous*, which means that estimates of model parameters are calculated all at once
- For this reason, ML estimation is described as a *full-information method*
- In contrast, techniques that analyze the equation for only one endogenous variable at a time are known as *partial-information* or *limited-information methods*
- Multiple regression used to estimate a recursive path model is an example of a partial-information method
- Implications of the difference between full-information versus partial-information estimation when there is specification error are considered later (chap. 6)

Maximum likelihood estimation

- ML estimation is also typically *iterative*, which means that the computer derives an initial solution and then attempts to improve these estimates through subsequent cycles of calculations

- "Improvement" means that the overall fit of the model to the data generally becomes better from step to step

- Iterative estimation will continue until the increments of the improvement in model fit fall below a predefined minimum value

Maximum likelihood estimation

- Iterative estimation may converge to a solution quicker if the procedure is given reasonably accurate *start values*, which are initial estimates of a model's parameters

- If start values are very inaccurate, then iterative estimation may fail to converge, which means that a stable solution has not been reached

- Iterative estimation can also fail if the relative variances among the observed variables are very different; that is, the covariance matrix is ill-scaled

Maximum likelihood estimation

- Computer programs typically issue a warning message if iterative estimation is unsuccessful
- When this occurs, whatever final set of estimates were derived by the computer may warrant little confidence
- Some SEM computer programs automatically generate their own start values
- It is important to understand, however, that computer-derived start values do not always lead to converged solutions
- Although the computer's "guesses" about start values are usually pretty good, sometimes it is necessary for the researcher to provide better start values for the solution to converge (see Appendix 5.A)

Maximum likelihood estimation

- ML estimation is generally *scale free*, which means that if a variable's scale is linearly transformed, a parameter estimated for the transformed variable can be algebraically converted back to the original metric

- It is also generally *scale invariant*, which means the value of the statistical criterion minimized remains the same regardless of the scale of the observed variables

- However, these properties may not hold if a correlation matrix is analyzed instead of a covariance matrix

- That is, ML estimation generally assumes the analysis of unstandardized variables

Maximum likelihood estimation

- When a raw data file is analyzed, standard ML estimation assumes there are no missing values
- There are special forms of ML estimation available for raw data files where some observations are missing at random (MAR)—see Arbuckle (1996)

Maximum likelihood estimation

- The statistical assumptions of ML estimation include

  1. independence of the observations
  2. multivariate normality of the endogenous variables
  3. independence of the exogenous variables and disturbances
  4. the exogenous variables are measured without error
  5. correct specification of the model

Maximum likelihood estimation

- The requirement for correct specification is critical because full-information methods, such as ML estimation, tend to propagate errors throughout the whole model

- This means that specification error in one parameter can affect results for other parameters elsewhere in the model

- It is difficult to predict the specific direction and magnitude of this error propagation because it depends in part upon the relation between the incorrect parameters and other parameters

- However, the more serious the specification error, the more serious may be the resulting bias

- See Kaplan (2000, chap. 5) for more information about the statistical assumptions underlying SEM

Maximum likelihood estimation

- Path coefficients are interpreted as regression coefficients in multiple regression, which means that they control for correlations among multiple presumed causes
- In the unstandardized solution, disturbance variances are estimated in the metric of the unexplained variance of the corresponding endogenous variable
- The ratio of the unstandardized disturbance variance over the total variance of the corresponding endogenous variable is the proportion of unexplained variance

Maximum likelihood estimation

- One minus the proportion of unexplained variance is the proportion of explained variance, R^2_{smc}
- The square root of 1 − R^2_{smc} equals the *standardized residual path coefficient* for the effect of the disturbance on the endogenous variable

Maximum likelihood estimation

- In basic ML estimation, standard errors are calculated only for the unstandardized solution
- This means that results of statistical tests (i.e., ratios of parameter estimates over their standard errors) are available only for the unstandardized solution
- Note that results of statistical tests of unstandardized estimates may not apply to the corresponding standardized estimates
- Also, basic ML estimation may derive incorrect standard errors when the variables are standardized (i.e., a correlation matrix is analyzed)

Maximum likelihood estimation

- There are basically two ways to obtain estimated standard errors for the standardized solution:

    1. Some SEM computer programs, such as Amos, use bootstrapping to generate standard errors for standardized estimates (chap. 6)

    2. Use the method of constrained estimation, which is available as a user-specified option in some SEM computer programs, such as SEPATH and RAMONA (chap. 7)

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Brito, C., & Pearl, J. (2003). A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, *9*, 459-474.

Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsley and S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 465-497). New York: Academic Press

Kaplan, D. (2000). *Structural equation modeling*. Thousand Oaks, CA: Sage.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.