Introduction to Structural Equation Modeling

Lecture Notes

Daniel J. Bauer & Patrick J. Curran

May 10-12, 2021

CenterStat

CenterStat.org

Table of Contents

Chapter 4: Confirmatory Factor Analysis

Chapter 5: Structural Equation Models with Latent Variables

Appendix: Maximum Likelihood Estimation

Definitions: Symbols

N	Total sample size
i	An individual sample unit, where $i = 1, 2, 3, \ldots N$
n_i	Number of observed variable values for case i; e.g., $n_i = 4$ if case i has observed data for four of five variables in the model)
k	The number of unique population means, variances, and covariances for the set of observed variables in a model
t	The number of parameters to be estimated for a model
p	The number of endogenous variables in a model
q	The number of exogenous variables in a model
r	The number of latent variables (factors) in a model
L	Likelihood of a model (used in reference to maximum likelihood estimation)
ℓ	Log-likelihood of a model (used in reference to maximum likelihood estimation)
T	Test statistic comparing a hypothesized structural equation model to the saturated model (i.e., test of overall model fit), chi-square distributed with $k - t$ degrees of freedom under certain assumptions and conditions; may be subscripted to specify a particular model (e.g., by h or b to specify hypothesized or baseline model, respectively).
y	A single endogenous observed variable (a.k.a. an outcome or criterion variable)
x	A single exogenous observed variable (a.k.a. a predictor, covariate, or independent variable)
η	Eta; a single latent variable
μ	Mu; population mean for a single observed variable
σ^2	Sigma-squared; population variance for a single observed variable
$E(\cdot)$	Expected value operator; returns the long-run average or population mean; e.g., $E(y) = \mu$
$VAR(\cdot)$	Variance operator; returns the population variance; e.g., $VAR(y) = \sigma^2$
$\bar{y}$, $\bar{x}$	Sample mean for y and x, respectively.
s^2	Sample variance for a single observed variable (may be subscripted by x or y for specificity)

$\mathbf{y}$ A p-length vector of endogenous observed variables (a.k.a. outcomes or criterion variables)

$\mathbf{x}$ A q-length vector of exogenous observed variables (a.k.a. predictors, covariates, or independent variables)

$\mathbf{z}$ Combined $(p + q)$-length vector of $\mathbf{y}$ and $\mathbf{x}$ ($\mathbf{y}$ stacked on top of $\mathbf{x}$ in a single vector)

$\boldsymbol{\eta}$ Eta; r-length vector of latent variables (latent factors)

$\boldsymbol{\theta}$ Theta; a t-length vector consisting of all the parameters in a given structural equation model

$\hat{\boldsymbol{\theta}}$ Theta-hat; a t-length vector consisting of all the sample estimates for the parameters of a given structural equation model

$\boldsymbol{\mu}$ Mu; population mean vector for a set of observed variables (may be subscripted by $\mathbf{x}$ or $\mathbf{y}$ to make specific to exogenous or endogenous variables)

$\boldsymbol{\Sigma}$ Sigma; population covariance matrix for a set of observed variables (may be subscripted by $\mathbf{xx}$, $\mathbf{yy}$, or $\mathbf{xy}$ to specifically reference covariances of exogenous variables, endogenous variables, or exogenous with endogenous variables, respectively)

$\boldsymbol{\mu}(\boldsymbol{\theta})$ The model-implied population mean vector for a set of observed variables, implied by a structural equation model with parameters $\boldsymbol{\theta}$ (where $\boldsymbol{\mu}$ may be subscripted by $\mathbf{x}$ or $\mathbf{y}$ to make specific to exogenous or endogenous variables)

$\boldsymbol{\Sigma}(\boldsymbol{\theta})$ The model-implied population covariance for a set of observed variables, implied by a structural euation model with parameters $\boldsymbol{\theta}$ (where $\boldsymbol{\Sigma}$ may be subscripted by $\mathbf{xx}$, $\mathbf{yy}$, or $\mathbf{xy}$ to specifically reference model-implied covariances of exogenous variables, endogenous variables, or exogenous with endogenous variables, respectively)

$\mathbf{m}$ The sample mean vector for a set of observed variables (may be subscripted by $\mathbf{x}$ or $\mathbf{y}$ to make specific to exogenous or endogenous variables)

$\mathbf{S}$ The sample covariance matrix for a set of observed variables (may be subscripted by $\mathbf{xx}$, $\mathbf{yy}$, or $\mathbf{xy}$ to specifically reference covariances of exogenous variables, endogenous variables, or exogenous with endogenous variables, respectively)

$\boldsymbol{\mu}\left(\hat{\boldsymbol{\theta}}\right)$ The sample estimated model-implied mean vector for a set of observed variables implied by a structural equation model with parameters $\boldsymbol{\theta}$ (where $\boldsymbol{\mu}$ may be subscripted by $\mathbf{x}$ or $\mathbf{y}$ to make specific to exogenous or endogenous variables)

$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ The sample estimated model-implied population covariance for a set of observed variables implied by a structural euation model with parameters $\boldsymbol{\theta}$ (where $\boldsymbol{\Sigma}$ may be subscripted by $\mathbf{xx}$, $\mathbf{yy}$, or $\mathbf{xy}$ to specifically reference model-implied covariances of exogenous variables, endogenous variables, or exogenous with endogenous variables, respectively)

$\boldsymbol{\alpha}$ Alpha; vector of intercepts for endogenous latent variables (in full SEM with latent variables) or endogenous observed variables (in path analysis / simultaneous equation model)

$\boldsymbol{\Gamma}$ Gamma; matrix of regression slopes capturing causal effects of observed exogenous variables on latent variables (in full SEM) or endogenous observed variables (in path analysis / simultaneous equation model)

$\mathbf{B}$ Beta; matrix of regression slopes capturing causal effects of latent variables on other latent variables (in full SEM with latent variables) or of endogenous observed variables on other endogenous observed variables (in path analysis / simultaneous equation model)

$\boldsymbol{\zeta}$ Zeta; vector of residuals (or disturbances) for latent variables (in full SEM with latent variables) or endogenous observed variables (in path analysis / simultaneous equation model)

$\boldsymbol{\Psi}$ Psi; matrix of residual variances and covariances for latent variables (in full SEM with latent variables) or endogenous observed variables (in path analysis / simultaneous equation model); i.e., $VAR(\boldsymbol{\zeta}_i) = \boldsymbol{\Psi}$

$\boldsymbol{\nu}$ Nu; vector of residuals for observed indicators of latent factors in full SEM with latent variables

$\boldsymbol{\Lambda}$ Lambda; matrix of factor loadings (regression slopes) relating latent factors to observed indicator variables; individual elements denoted λ

$\boldsymbol{\varepsilon}$ Epsilon; vector of residuals for observed indicator variables after accounting for the influence of latent factors

$\boldsymbol{\Theta}$ Theta; matrix of residual variances and covariances for observed indicator variables after accounting for the influence of the latent factors; i.e., $VAR(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta}$

h^2 Communality, the proportion of variance in an observed variable that is accounted for by the latent variables in the model (equivalent to to a multiple R^2 from a regression of the indicator on all related latent factors); used in factor analysis to gauge the quality of indictor variables.

Definitions: Terminology

Exogenous Variables: Variables that are not expressed as a function of other variables; their causes are "outside" the system of variables under study. Also referred to as independent variables.

Endogenous Variables: Variables that are expressed as a function of one or more other variables; their causes are "inside" the system of variables under study. Also referred to as dependent variables.

Observed v. Latent Variable: An observed variable is one for which you directly obtain sample values; a latent variable is one for which there are (usually) no sample realizations and which is inferred from patterns of association among the observed variables.

Saturated Model: A model with no restrictions and no degrees of freedom, usually referring to a model where all means, variances, and covariances are distinct parameters to be estimated.

Hypothesized Model: The model of interest, reflecting restrictions based on theoretical hypotheses about how variables are related to one another.

Baseline Model: Classically defined as a model in which the only parameters are means and variances but no covariances (assuming independence of all variables in the model), but sometimes defined as a model in which only the exogenous variables are allowed to covary but the endogenous variables are not. Used in computing relative goodness of fit indices.

Path Analysis / Simultaneous Equation Model: Structural equation model consisting of only observed variables.

Recursive v. Non-Recursive Model: A recursive model meets two conditions: (1) No correlated residuals; (2) only unidirectional effects. If either of these conditions is not met, the model is non-recursive. Can be useful to determine whether the model is recursive or not for purposes of evaluating model identification.

Total Effect: Sum of all possible paths (direct and indirect) leading from a predictor to an outcome.

Direct Effect: The part of the effect of one variable on another that does not pass through any intervening variables (represented as a single one-headed arrow in a path diagram, e.g., X -> Y).

Indirect Effect: The part of the effect of one variable on another that passes through one or more intervening variables (e.g., in X -> M -> Y, the effect of X on Y passes through M, representing an indirect effect) and which is equal to the sum of all specific indirect effects.

Specific Indirect Effects: The part of the effect of one variable on another that passes through one specific intervening variables; useful when considering models with multiple mediators and breaking an indirect effect into unique components associated with each mediator.

Exploratory Factor Analysis: A factor analysis model in which all observed indicator variables load on all factors (i.e., all factors effect all indicators). Rotation methods are typically used to try to achieve an interpretable structure in which each indicator has a high loading on one and only one latent factor (also known as simple structure). Factors may be allowed to correlate (oblique rotation) or not (orthogonal rotation) but they do not have causal effects on one another.

Confirmatory Factor Analysis: A factor analysis model in which restrictions are imposed so that indicators only load on certain factors and not others. These restrictions obviate the need to use rotation to enhance the interpretation of the factors and (ideally) over-identify the model to allow for formal tests of the adequacy of the model. Factors may be allowed to correlate or not but they do not have causal effects on one another.

Structural Equation Model: Allows for both observed and latent variables and where variables of either type can either covary or have causal effects on one another. Simplifies to a simultaneous equation model if no latent variables are included; simplifies to a confirmatory factor analysis if all observed variables are expressed solely as functions of latent variables and the latent variables have no causal relationships on each other.

Multiple Groups Model: A model fit simultaneously to each group in a stratified sample (e.g., boys versus girls). Interest centers on which parts of the model are the same across groups versus different, usually evaluated by imposing and testing equality constraints on specific parameters of the model.

Latent Curve Model: A structural equation model for longitudinal data in which the repeated measures are treated as indicators of latent growth factors that define the underlying trajectory of change over time for the individuals in the population (e.g., latent intercepts and slopes for a linear growth trajectory).

Chapter 1
Introduction, Background, and Multiple Regression

1.1 Introduction

Objectives

- Review advantages of the structural equation model (SEM)
- Describe the organization of our workshop

CenterStat 1.3

What is the Structural Equation Model?

- The structural equation model (or SEM) is a general analytic framework that subsumes a large number of models
  - e.g., the *t*-test, ANOVA, ANCOVA, MANOVA, MANCOVA, multiple regression, and factor analysis can all be cast as an SEM
- But allows for many critically important extensions:
  - estimating models with multiple dependent variables
  - testing complex mediating mechanisms
  - estimating latent variables to account for measurement error
  - testing invariance of models across groups
  - estimating latent factors with binary or ordinal indicators
  - modeling growth trajectories with repeated measures

CenterStat 1.4

What are Advantages of the SEM?

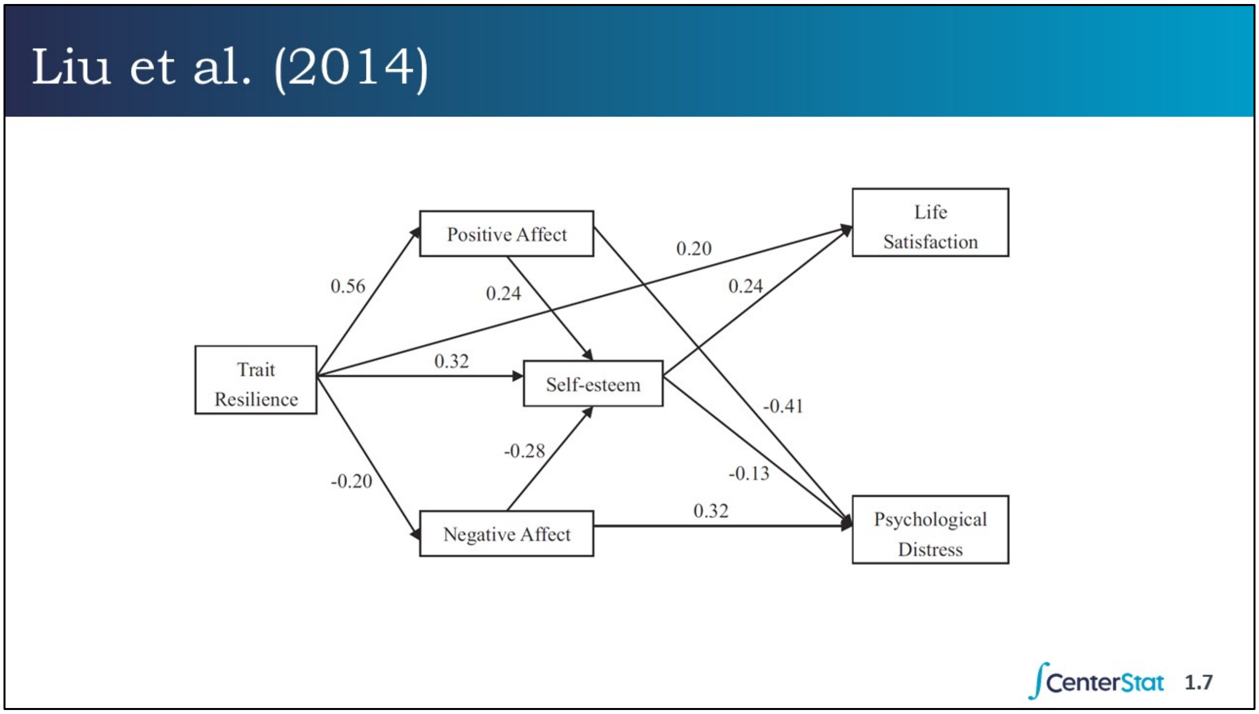
- Highly flexible in allowing various parameter structures and tests
- Allows for many types of simultaneous tests that otherwise must be done in a step-wise fashion
  - e.g., testing complex mediation in a single model
- Offers many alternative methods of estimation
  - e.g., continuous & non-normally distributed, or binary, ordinal, count data
- Provides powerful and rigorous tests of individual stability and change in longitudinal data

CenterStat 1.5

What are Disadvantages of the SEM?

- Models can be complicated and wander away from theory
- Challenging to decide if a model adequately "fits" the data
- Radically different models can fit data precisely the same
- Philosophical blind spot in that the SEM takes the *absence* of evidence as *support* for the hypothesized model
- Large *N* is needed for stable model estimation, but how large is large enough is hard to know
- Like any modeling framework, the SEM cannot "fix" poor sampling or weak measurement

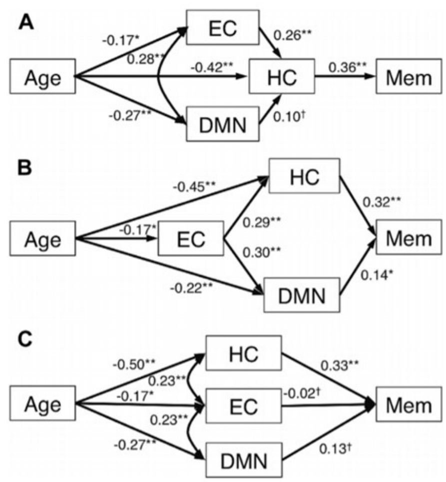
CenterStat 1.6



Liu, Y., Wang, Z., Zhou, C., & Li, T. (2014). Affect and self-esteem as mediators between trait resilience and psychological adjustment. *Personality and Individual Differences, 66*, 92-97.

Abstract from manuscript: The primary purpose of the current study was to examine the potential sequential mediation effects of affect and self-esteem on the association between trait resilience and psychological adjustment, as indexed by life satisfaction and psychological distress. A total of 412 undergraduate students completed a packet of questionnaires that assessed trait resilience, positive and negative affect, self-esteem, life satisfaction and psychological distress. Mediation analyses showed that self-esteem mediated the relation between trait resilience and life satisfaction. In addition, positive affect, negative affect, and self-esteem were found to intervene between trait resilience and psychological distress. Furthermore, the sequential mediation effects of affect–self-esteem on the relations between trait resilience and life satisfaction as well as psychological distress were confirmed. Results are discussed in light of previous findings. Limitations of the study and suggestions for future research are briefly discussed.

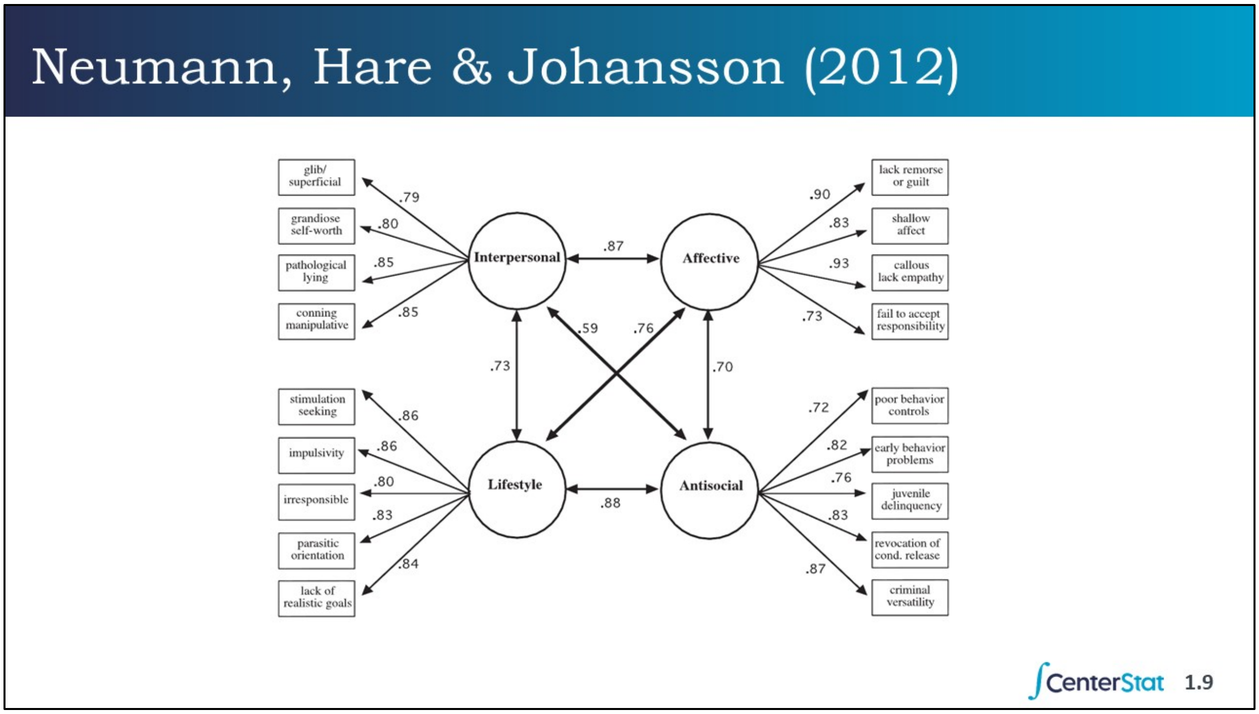
Ward et al. (2015)



CenterStat 1.8

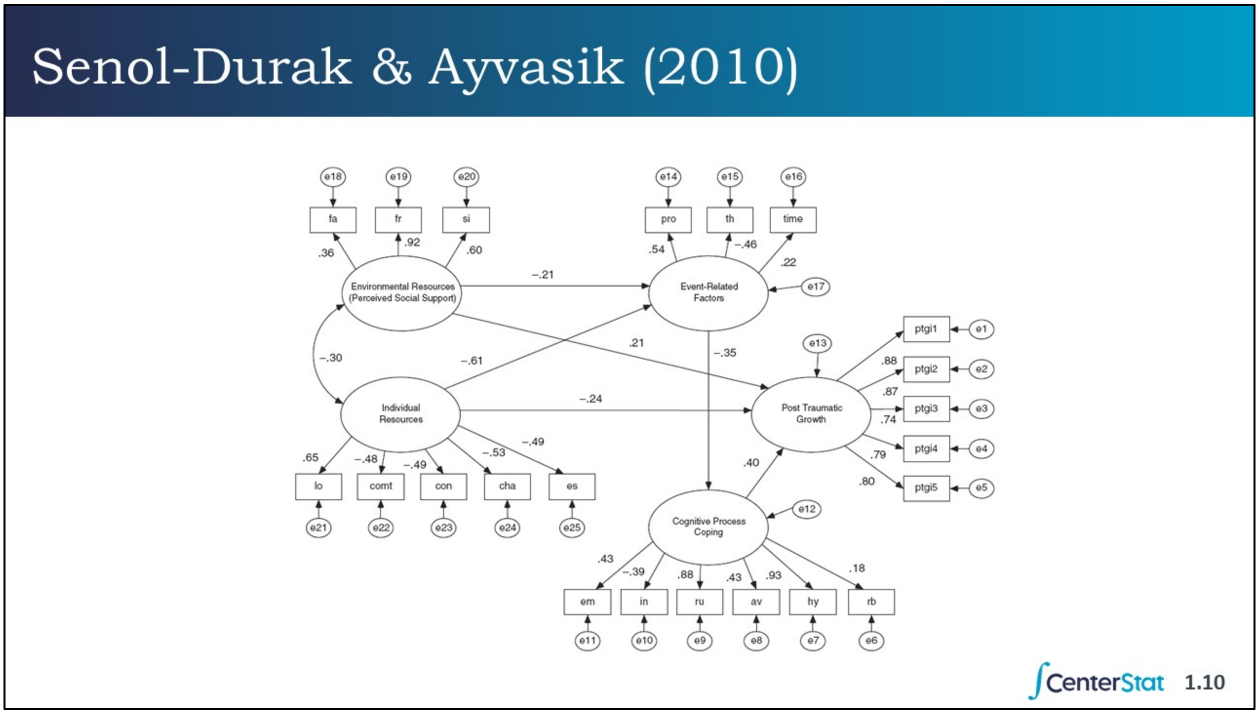
Ward, A.M., Mormino, E.C., Huijbers, W., Schultz, A.P., Hedden, T., & Sperling, R.A. (2015). Relationships between default-mode network connectivity, medial temporal lobe structure, and age-related memory deficits. *Neurobiology of Aging, 36*, 265-272.

<u>Abstract from manuscript</u>: Advanced aging negatively impacts memory performance. Brain aging has been associated with shrinkage in medial temporal lobe structures essential for memory, including hippocampus and entorhinal cortex, and with deficits in default-mode network connectivity. Yet, whether and how these imaging markers are relevant to age-related memory deficits remains a topic of debate. Using a sample of 182 older (age 74.6 +/- 6.2 years) and 66 young (age 22.2 +/- 3.6 years) participants, this study examined relationships among memory performance, hippocampus volume, entorhinal cortex thickness, and default-mode network connectivity across aging. All imaging markers and memory were significantly different between young and older groups. Each imaging marker significantly mediated the relationship between age and memory performance and collectively accounted for most of the variance in age-related memory performance. Within older participants, default-mode connectivity and hippocampus volume were independently associated with memory. Structural equation modeling of cross-sectional data within older participants suggest that entorhinal thinning may occur before reduced default-mode connectivity and hippocampal volume loss, which in turn lead to deficits in memory performance.



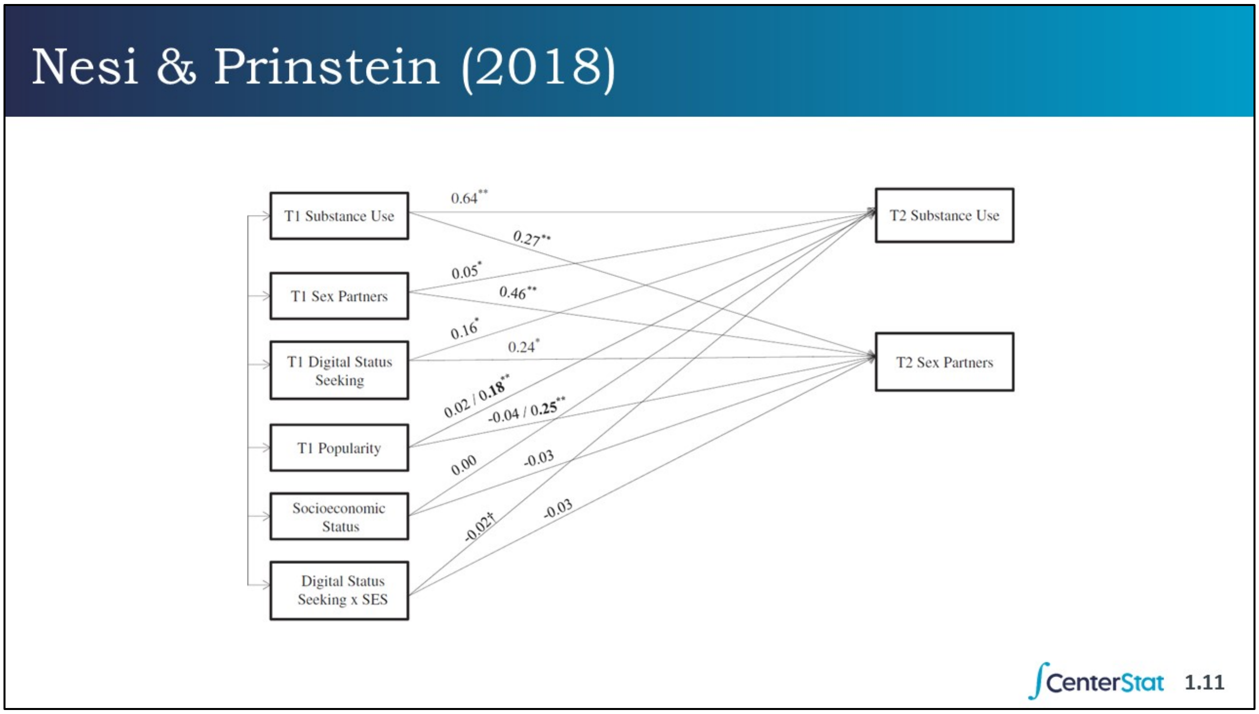
Neumann, C.S., Hare, R.D. & Johansson, P.T. (2012). The Psychopathy Checklist-Revised (PCL-R), low anxiety, and fearlessness: a structural equation modeling analysis. *Personality Disorders: Theory, Research, and Treatment, 4,* 129-137.

Abstract from manuscript: The current study employed a large representative sample of violent male offenders within the Swedish prison system to examine the factor structure of the PCL-R and the latent variable relations between the PCL-R items and clinical ratings of low trait anxiety and trait fearlessness (LAF). Consistent with previous research, confirmatory factor analysis (CFA) revealed strong support for the four-factor model of psychopathy (Interpersonal, Affective, Lifestyle, and Antisocial). Also, a series of CFAs revealed that the LAF items could be placed on any of the PCL-R factors without any changes in model fit. Finally, structural equation modeling results indicated that a PCL-R superordinate factor was able to account for most of the variance of a separate LAF factor. Taken together, the results indicate that if low anxiety and fearlessness, as measured via clinical ratings, are part of the psychopathy construct they are comprehensively accounted for by extant PCL-R items.



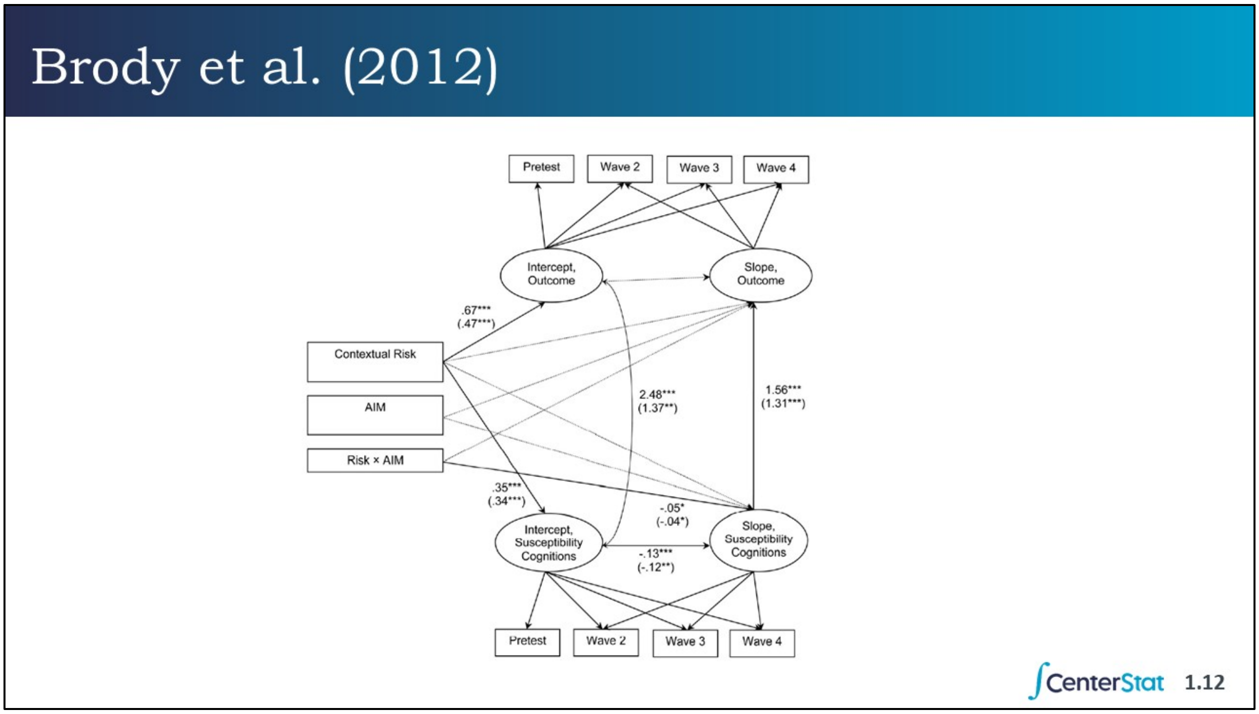
Senol-Durak, E., & Ayvasik, H.B. (2010). Factors associated with posttraumatic growth among the spouses of myocardial infarction patients. *Journal of Health Psychology, 15,* 85-95.

Abstract from manuscript: To clarify the rationale behind Posttraumatic Growth (PTG), a model by Schaefer and Moos describes the relative contribution of environmental resources, individual resources, event related factors, cognitive processing and coping (CPC) on PTG. In the present study, this model was tested with the spouses of myocardial infarction patients with data from various hospitals in Turkey. A structural equation model revealed that neither individual nor environmental resources had indirect effects on PTG through the effect of event-related factors and CPC, while they showed direct effects on PTG. The findings were discussed in the context of the theoretical model.



Nesi, J., & Prinstein, M.J. (2018). In search of likes: Longitudinal associations between adolescents' digital status seeking and health-risk behaviors. *Journal of Clinical Child & Adolescent Psychology*, *Published Online First,* 1-9.

This study introduces a new construct—digital status seeking—which reflects a set of behaviors made possible by the social media environment. Digital status seeking is defined as the investment of significant effort into the accumulation of online indicators of peer status and approval. The concurrent validity of this construct was examined, as well as the longitudinal implications of digital status seeking for adolescents' engagement in health-risk behaviors. A school-based sample of 716 participants (Mage = 16.01 at Time 1; 54.2% female) participated at 2 time points, 1 year apart. Sociometric nomination procedures were used to assess digital status seeking and peer status. Participants self-reported indices of social media use, peer importance, and risky behavior engagement (substance use, sexual risk behavior). For a subset of participants, social media pages were observationally coded for status indicators (i.e., likes, followers) and status-seeking behaviors. Adolescents with greater reputations of digital status seeking reported more frequent social media use, desire for popularity, belief in the importance of online status indicators, and use of strategies to obtain these indicators. Multiple group path analyses indicated that for both genders, digital status seekers engaged in higher levels of substance use and sexual risk behavior 1 year later. Moderation of this effect by race/ethnicity and socioeconomic status was explored. This novel, multimethod investigation reveals digital status seeking as an important construct for future study and offers preliminary evidence for the unique role of social media experiences in contributing to adolescent adjustment.



Brody, G. H., Yu, T., Chen, Y., Kogan, S. M., & Smith, K. (2012). The Adults in the Making program: long-term protective stabilizing effects on alcohol use and substance use problems for rural African American emerging adults. *Journal of Consulting and Clinical Psychology*, *80*, 17-28.

<ins>Abstract from manuscript</ins>: Objective: This report addresses the long-term efficacy of the Adults in the Making (AIM) prevention program on deterring the escalation of alcohol use and development of substance use problems, particularly among rural African American emerging adults confronting high levels of contextual risk. Method: African American youths (M age, pretest = 17.7 years) were assigned randomly to the AIM (n = 174) or control (n = 173) group. Past 3-month alcohol use, past 6-month substance use problems, risk taking, and susceptibility cognitions were assessed at pretest and at 6.4, 16.6, and 27.5 months after pretest. Pretest assessments of parent–child conflict, affiliations with substance-using companions, and perceived racial discrimination were used to construct a contextual risk factor index. Results: A protective stabilizing hypothesis was supported; the long-term efficacy of AIM in preventing escalation of alcohol use and substance use problems was greater for youths with higher pretest contextual risk scores. Consistent with a mediation-moderation hypothesis, AIM-induced reductions over time in risk taking and susceptibility cognitions were responsible for the AIM × contextual risk prevention effects on alcohol use and substance use problems. Conclusions: Training in developmentally appropriate protective parenting processes and self-regulatory skills during the transition from adolescence to emerging adulthood for rural African Americans may contribute to a self-sustaining decreased interest in alcohol use and a lower likelihood of developing substance use problems.

What The Next Three Days Hold…

- Review of matrix algebra
- Review of multiple regression
- Multiple regression as a structural equation model
- Path analysis (*aka* simultaneous equations model)
- Confirmatory factor analysis (CFA)
- Structural equations model (SEM)
- Topics not covered due to limited time:
  - Multiple group SEMs
  - SEMs with binary and ordinal measures
  - Latent curve model (LCM)

CenterStat 1.13

Summary

- SEM subsumes and extends upon many commonly used models
- SEM is a flexible modeling framework that offers opportunities to test complex, multivariate hypotheses
- To express SEMs we need some familiarity with matrix algebra

CenterStat 1.14

1.2 A Brief Review of Matrix Algebra

Objectives

- Introduce the building blocks of matrix algebra in terms of scalars, vectors and matrices
- Define core mathematical operations for matrices
- Provide sufficient understanding of basic concepts to allow for the navigation of matrix notation in the SEM

CenterStat 1.16

Why Do We Need Matrices?

- Multivariate models compactly expressed as matrices
  - particularly SEM -- many variables and many parameters
- Understanding matrices allows us to better understand core statistical model that underlies the SEM
  - also allows us to navigate technical resources; e.g., textbooks, computer manuals, online resources, etc.
- Many computer programs are set up in matrix form (LISREL, Mx)
- All programs list warnings and errors in terms of matrices
  - e.g., "psi matrix is not positive definite", or "Hessian matrix cannot be inverted" or "defined model is inadmissible due to a singularity in sigma"
- Finally, matrices make the world a happier place

CenterStat 1.17

Matrices We Already Know

- A matrix is simply a doubly-ordered arrangement of numbers
- One example is raw data matrix
  - data matrix has n-rows (one for each observation) and p-columns (one for each variable)
  - any data organized in a spreadsheet is a matrix
- Another example is correlation matrix
  - we are all familiar with the correlation matrix in which ones are on the diagonal and all bivariate correlations are on the off-diagonal
- Even a calendar is a matrix
  - the rows are weeks and the columns are days

CenterStat 1.18

Scalars

- A scalar is simply an "ordinary" number
- Scalars are typically denoted by a lower-case italicized letter
- The algebra of scalars is arithmetic, and arithmetic provides the rules for operating on scalars

$$a = -6$$

$$b = 5$$

$$a + b = -6 + 5 = -1$$

CenterStat 1.19

Matrix

- A matrix is a doubly-ordered arrangement of scalars
- The rows represent one set of categories
- The columns represent the other set of categories
- Matrices are usually denoted by bold capital letters

$$\mathbf{A} = \begin{bmatrix} 6 & 1 & 3 & 5 \\ -4 & 2 & 9 & 11 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} -5 & 3 \\ 2.6 & 43 \end{bmatrix}$$

CenterStat 1.20

Order of a Matrix

- The *order* of a matrix refers to the number of rows and columns. For example, this matrix is of order "3 by 4:"

$$\mathbf{A}_{(3\times4)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}$$

- In general, matrices are of order $r \times c$. The first number always denotes row & the second column
  - Thus, the mnemonic $r \times c$.

CenterStat 1.21

The Elements of a Matrix

- A matrix is composed of scalars called *elements*
- Element in row i and column j of matrix $\mathbf{A}$ is designated a_{ij} where the first subscript designates *row* and the second *column*
- Two matrices are equal if and only if they are of the same order and all corresponding elements are equal

$$\mathbf{A} = \begin{bmatrix} 3 & 4 & 8 \\ 9 & 4 & 0 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 3 & 4 & 8 \\ 9 & 4 & 0 \end{bmatrix} \qquad \mathbf{A} = \mathbf{B}$$

CenterStat 1.22

The Transpose of a Matrix

- The transpose of a matrix is formed by interchanging the elements in each row and column.
- The transpose of matrix $\mathbf{A}$ $(r \times c)$ is designated $\mathbf{A}'$ $(c \times r)$.
- The first row simply becomes the first column, the second row the second column, and so on.

$$\mathbf{A}_{(r \times c)} = \begin{bmatrix} 6 & 2 & 4 \\ 8 & 1 & 0 \end{bmatrix} \qquad \mathbf{A}'_{(c \times r)} = \begin{bmatrix} 6 & 8 \\ 2 & 1 \\ 4 & 0 \end{bmatrix}$$

CenterStat 1.23

Symmetric Matrices

- A matrix is considered *symmetric* if the matrix is equal to the transpose of the same matrix: $\mathbf{A} = \mathbf{A}'$
- Note that a matrix must be *square* to be symmetric
  - a square matrix has an equal number of rows and columns; e.g., $c=r$
- Important cases: Covariance and correlation matrices

$$\mathbf{R}_{(3\times3)} = \begin{bmatrix} 1 & .2 & .4 \\ .2 & 1 & .3 \\ .4 & .3 & 1 \end{bmatrix} \qquad \mathbf{R}'_{(3\times3)} = \begin{bmatrix} 1 & .2 & .4 \\ .2 & 1 & .3 \\ .4 & .3 & 1 \end{bmatrix}$$

- Given the redundancy of elements above and below the diagonal, typically only the lower triangle is shown

CenterStat 1.24

Diagonal Matrices

- A particular kind of symmetric matrix is a diagonal matrix
  - Contains non-zero values on the diagonal, zeros everywhere else
- In many latent variable models, the residual covariance matrix will be diagonal
  - reflects assumption that all common variance among the observed variables is explained by the underlying latent factors

CenterStat 1.25

Diagonal Matrices

- For example, define a sample covariance matrix to be

$$\mathbf{S} = \begin{bmatrix} s_{11} & & & \\ s_{21} & s_{22} & & \\ s_{31} & s_{32} & s_{33} & \\ s_{41} & s_{42} & s_{43} & s_{44} \end{bmatrix}$$

- Then the diagonal matrix is simply

$$\mathbf{D} = diag(\mathbf{S}) = \begin{bmatrix} s_{11} & & & \\ 0 & s_{22} & & \\ 0 & 0 & s_{33} & \\ 0 & 0 & 0 & s_{44} \end{bmatrix}$$

CenterStat 1.26

The Identity Matrix

- An important diagonal matrix is the identity matrix
  - Ones on the diagonal and zeros everywhere else
- The identity matrix plays the same role in matrix algebra as the number 1 in arithmetic

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

CenterStat 1.27

Vectors: A Particular Type of Matrix

- Vectors are matrices that have one row or one column
- A vector of order $r \times 1$ is a *column vector*
- A vector of order $1 \times c$ is a *row vector*
- Vectors are usually denoted by a lower-case bold letter

$$\mathbf{x}_{(r\times 1)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} \qquad \mathbf{x}'_{(1\times c)} = \begin{bmatrix} x_1 & x_2 & \cdots & x_c \end{bmatrix}$$

- The prime (′) represents the "transpose" of the vector

CenterStat 1.28

Matrix Operations: Addition

- For addition, matrices must be of the same order
- Addition of two matrices is accomplished by adding corresponding elements, e.g., $c_{ij} = a_{ij} + b_{ij}$

$$\begin{bmatrix} 3 & 1 & 5 \\ 2 & 4 & 6 \end{bmatrix} + \begin{bmatrix} 2 & 4 & 7 \\ 8 & 11 & 4 \end{bmatrix} = \begin{bmatrix} 5 & 5 & 12 \\ 10 & 15 & 10 \end{bmatrix}$$

- Matrix addition is:
  - *commutative*: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
  - *associative*: $\mathbf{A} + \mathbf{B} + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$
- The resulting matrix has the same dimensions as the component matrices
  - e.g., $(r \times c) + (r \times c) = (r \times c)$

CenterStat 1.29

Matrix Operations: Subtraction

- For subtraction, matrices must be of the same order
- Subtraction of two matrices is accomplished by subtracting corresponding elements, e.g., $c_{ij} = a_{ij} - b_{ij}$

$$\begin{bmatrix} 3 & 1 & 5 \\ 2 & 4 & 6 \end{bmatrix} - \begin{bmatrix} 2 & 4 & 7 \\ 8 & 11 & 4 \end{bmatrix} = \begin{bmatrix} 1 & -3 & -2 \\ -6 & -7 & 2 \end{bmatrix}$$

- As with addition, the resulting matrix is of the same dimensions as the two component matrices:
  - e.g., $(r \times c) - (r \times c) = (r \times c)$

CenterStat 1.30

Matrix Operations: Multiplication

- ***Rule 1***: Only matrices of form $(p \times q) * (q \times k)$ are conformable for multiplication
  - The number of columns in the premultiplier must equal the number of rows in the post multiplier
- ***Rule 2***: The product matrix will have the following order:
  - $\mathbf{A}_{(p \times q)}\mathbf{B}_{(q \times k)} = \mathbf{C}_{(p \times k)}$
- ***Rule 3***: c_{ij} represents an element in row i, column j of the product matrix and results from the product of row i of the premultiplier with column j of the post multiplier
  - $c_{ij}=a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{iq}b_{qj}$
- Easier to see an example....

CenterStat 1.31

These rules are a bit tedious. As the next slide shows, you're just adding up the product of each element of the first row with the first column, and then the first row with the second column, and so on.

Matrix Operations: Multiplication

$$\mathbf{A}_{(2\times3)}\mathbf{B}_{(3\times2)} = \mathbf{C}_{(2\times2)} = \begin{bmatrix} 2 & 4 & 1 \\ 3 & 0 & 4 \end{bmatrix}\begin{bmatrix} 1 & 3 \\ 2 & 0 \\ 4 & 2 \end{bmatrix} = \begin{bmatrix} 14 & 8 \\ 19 & 17 \end{bmatrix}$$

$$c_{1,1} = (2)(1)+(4)(2)+(1)(4) = 14$$

$$c_{1,2} = (2)(3)+(4)(0)+(1)(2) = 8$$

$$c_{2,1} = (3)(1)+(0)(2)+(4)(4) = 19$$

$$c_{2,2} = (3)(3)+(0)(0)+(4)(2) = 17$$

CenterStat 1.32

In scalar algebra, order of multiplication is not important: $ab = ba$. In matrix algebra, order of multiplication is critical: in general, $\mathbf{AB} \neq \mathbf{BA}$.

Matrix Operations: Division via Inverses

- Division is not defined for matrix operations, but may be accomplished by multiplication by the *inverse* of a matrix
- Remember that multiplying by the inverse of a scalar is the same as dividing by that same scalar:

$$\frac{20}{5} = 20\left(\frac{1}{5}\right) = 20(5^{-1}) = 4$$

- Similarly, we can multiply one matrix by another inverted matrix produce a desired outcome:

$$\mathbf{AX}^{-1} = \mathbf{Z}$$

- But to invert a matrix, we must first calculate its *determinant*

CenterStat 1.33

Inverting matrices is a critical component of nearly all statistical models.

The Determinant of a Matrix

- To invert a matrix, say $\mathbf{S}$, must first obtain its determinant, denoted $|\mathbf{S}|$
- The determinant is a single scalar value that represents the generalized variance of an entire matrix
  - The larger the determinant, the greater the generalized variance
- Determinants can only be computed for square matrices
  - e.g., correlation and covariance matrices
- We commonly compute determinants for different types of covariance matrices within the SEM

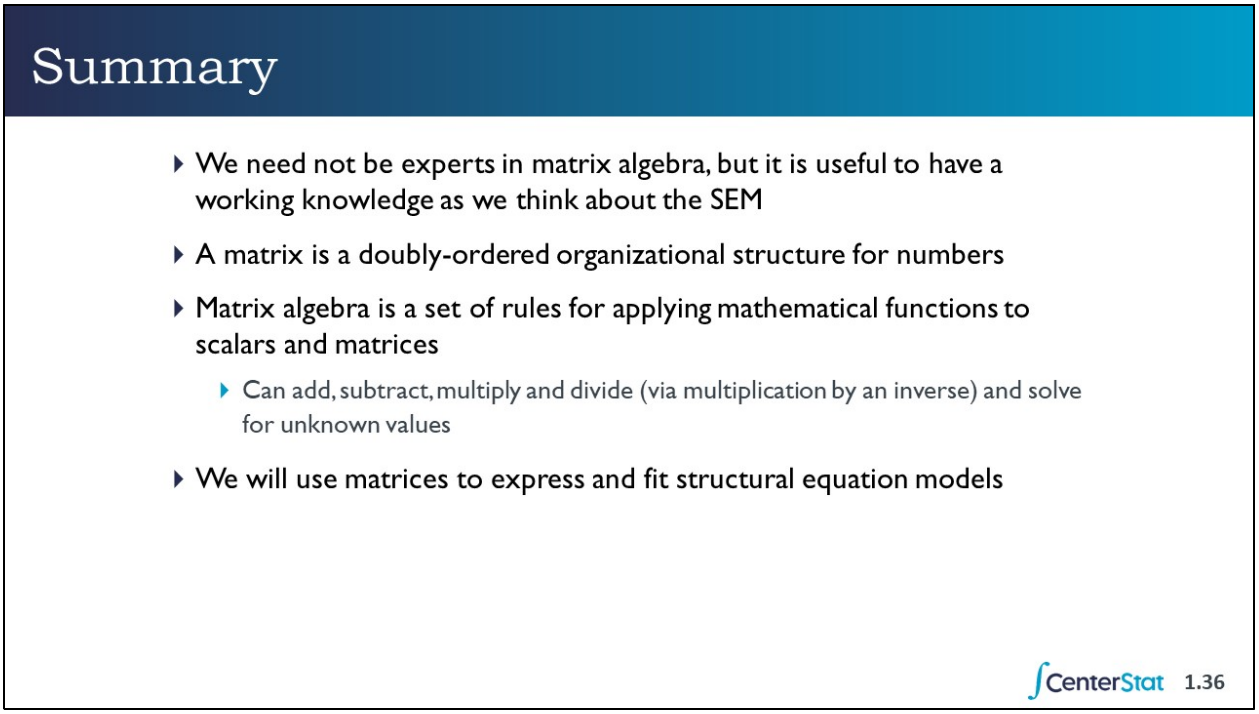
CenterStat 1.34

Warning: The notation for a determinant in matrix algebra looks just like the notation for an absolute value in scalar algebra, but they are not in any way related to one another.

The Determinant of a Matrix

- A determinant equal to zero reflects that there is no generalized variance for the matrix
  - there is a linear dependency among columns or rows (multicollinearity) or there is a zero or negative variance on the diagonal
- A matrix with a zero determinant cannot be inverted
  - Such a matrix is said to be non-positive definite (i.e., NPD)
- But inverting a matrix is often a necessary calculation in many SEMs and thus a zero determinant can pose problems in the SEM
  - e.g., the notorious error message *"psi matrix is not positive definite"*
- This is often a significant error, and steps must be taken to find and eliminate the problem

CenterStat 1.35


Summary
We need not be experts in matrix algebra, but it is useful to have a working knowledge as we think about the SEM
A matrix is a doubly-ordered organizational structure for numbers
Matrix algebra is a set of rules for applying mathematical functions to scalars and matrices
Can add, subtract, multiply and divide (via multiplication by an inverse) and solve for unknown values
We will use matrices to express and fit structural equation models
CenterStat
1.36

1.3 Review of Multiple Regression

Objectives

- Provide a brief review of the multiple regression model
- Present an example using real data to model predictors of adolescent deviant peer affiliations

CenterStat 1.38

Multiple Regression

- Multiple regression maps a single dependent (or criterion) variable onto one or more independent (or predictor) variables
  - we "regress" the dependent variable on the set of independent variables
- The DV is (typically) assumed continuous and normally distributed and the IVs can take on any distributional form
- There are *joint tests* of all the predictors taken together
  - e.g., omnibus F-test, multiple R-squared
- There are *unique tests* of each individual predictor above and beyond all other predictors
  - e.g., raw and standardized betas, squared semi-partial correlations

CenterStat 1.39

Multiple Regression

- Estimation centers on selecting parameter values that minimize the sum of the squared residuals
- Consider a one-predictor regression

$$y_i = \alpha + \gamma_1 x_{1i} + \zeta_i$$

- The DV is denoted y_i, the IV x_{1i}, the intercept is α, and the residual is ζ_i
- We assume ζ_i is normally distributed with a mean of zero and variance ψ This is expressed as:

$$\zeta_i \sim N(0, \psi)$$

CenterStat 1.40

Note that here we are using SEM notation to define the regression model. More commonly, the regression model is written as $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$. The use of SEM notation will facilitate extension to more complex models.

Multiple Regression

- The model for the *observed* value of y_i is

$$y_i = \alpha + \gamma_1 x_{1i} + \zeta_i$$

- We can also express the model for the *predicted* value of y_i as

$$\hat{y}_i = \alpha + \gamma_1 x_{1i}$$

- The only difference is the lack of the residual. The residual is thus the difference between the *observed* and *predicted* values of the outcome and is expressed as

$$\zeta_i = y_i - (\alpha + \gamma_1 x_{1i}) = y_i - \hat{y}_i$$

- We want these residuals to be as small as possible

CenterStat 1.41

A Hypothetical Example

- Consider a hypothetical example where we are studying the relation between academic achievement and juvenile delinquency in a sample 250 children
  - hypothesize a negative relation between achievement and delinquency
- The one-predictor regression model is defined as

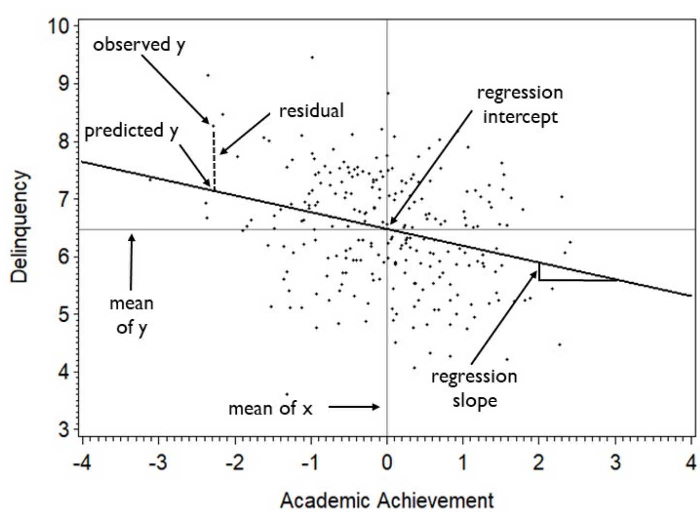
$$delinq_i = \alpha + \gamma_1 acadach_i + \zeta_i$$

- Delinquency is the *dependent variable* and academic achievement is the *independent variable*
- We can better see key parts of the model in a scatter plot

CenterStat 1.42

This is a simple simulated example intended to demonstrate the key features of a one-predictor regression model. We will turn to a real data application in just a few pages.

A Hypothetical Example



CenterStat 1.43

Least Squares Estimation

- The term "least squares" reflects that we are selecting parameter estimates that minimize the sum of the squared residuals
- Consider again our one-predictor regression: $y_i = \alpha + \gamma_1 x_{1i} + \zeta_i$
- The sum of the squared residuals is thus given as

$$\sum_{i=1}^{N} \zeta_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - \alpha - \gamma_1 x_i)^2$$

- This defines a function linking all possible values of the regression parameters to the corresponding sums of squares

CenterStat 1.44

Least Squares Estimation

- The derivations for the least squares estimates can be found elsewhere, but the end result is easy to define
- For a one predictor regression, the intercept and slope are

$$\hat{\gamma}_1 = \frac{s_{xy}}{s_x^2} \qquad \hat{\alpha} = \bar{y} - \gamma_1 \bar{x}$$

- s_x^2 is the variance of x, s_{xy} is the covariance between x and y, and $\bar{x}$ and $\bar{y}$ are the mean of x and y, respectively
- No other values of the intercept and slope will provide a smaller sum of squared residuals than these estimates

CenterStat 1.45

Specification: Two Predictors

- We can extend the model to allow for two predictors:

$$y_i = \alpha + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \zeta_i$$

- The regression coefficients are now *partial* coefficients
  - γ_1 represents the relation between x_1 and y above and beyond the effects of x_2 -- i.e., holding x_2 constant
  - γ_2 represents the relation between x_2 and y above and beyond the effects of x_1 -- i.e., holding x_1 constant
- Can differentiate unique effects from joint effects
  - Unique effect of x_1 on y is γ_1; unique effect of x_2 on y is γ_2
  - Joint effect of x_1 and x_2 on y is captured in the multiple R^2

CenterStat 1.46

Multiple Regression: Extensions

- The regression model can be extended in a number of ways
  1. A large number of predictors can be included
  2. Predictors can be entered individually or in blocks
  3. Categorical predictors can be incorporated via coding variables
  4. Can include interactions between two or more predictors
  5. Can include non-linear effects of a single predictor
- We can compactly define any variant of the regression model in matrix terms

CenterStat 1.47

Multiple Regression in Matrix

- For a regression model with q predictors

$$
\begin{aligned}
y_i &= \alpha + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \cdots + \gamma_q x_{qi} + \zeta_i \\
&= \alpha + \sum_{q=1}^{Q} \gamma_q x_{qi} + \zeta_i \\
&= \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i
\end{aligned}
$$

- This is precisely the same model but is expressed in a more compact form

CenterStat 1.48

For the regression model it is more common to use p to reference the number of predictors; however, to be consistent with SEM notation, we will use q instead. Later, we will use p to indicate the number of outcome variables in the model (where $p = 1$ for the multiple regression model).

Multiple Regression in Matrix

- More specifically, one outcome and q predictors the model is:

$$y_i = \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i \qquad VAR(\zeta_i) = \psi$$

- Each component is defined as follows:

y_i is a scalar for the outcome

$\mathbf{x}_i$ is a $q \times 1$ vector of predictors

α is a scalar for the single regression intercept

$\boldsymbol{\gamma}$ is a $q \times 1$ vector of regression slopes

ζ_i is scalar of errors (i.e., residuals, or disturbances)

ψ is a scalar residual variance for the outcome

CenterStat 1.49

Example: Deviant Peer Affiliations

- To demonstrate the multiple regression model, we will estimate a series of models using a real example
- We will fit a series of regression models to examine predictors of deviant peer affiliations
  - associating with deviant peers is an important training ground for other high-risk adolescent behaviors
- We use data from the Adolescent and Family Development Project (AFDP) housed at Arizona State University
  - We are indebted to Dr. Laurie Chassin for generously sharing these data

CenterStat 1.50

Example: Deviant Peer Affiliations

- We draw our data from a much larger study that followed children and their families over many years
- AFDP designed to study parental alcoholism as a risk factor for substance use and other psychopathology
- Our subsample consists of 316 adolescents interviewed at the initial wave of assessment
  - Adolescent age ranged from 10 to 16 (mean=12.7)
  - 54% (n=170) were male
  - 53% (n=166) were biological & custodial children of an alcoholic parent (abuse or dependent)

CenterStat 1.51

Example: Deviant Peer Affiliations

- Deviant peer affiliations (***peer***) is self report on peer substance use and tolerance of use
- Child of alcoholic (***coa***) is parent report of alcoholism diagnosis where 0=non-alcoholic, 1=alcoholic
- Self-identified gender (***gen***) where 0=girl & 1=boy
- Age (***age***) measured in years at assessment
- Stress (***stress***) is self report of uncontrollable negative life stressful events
- Emotionality (***emotion***) is self report of temperamental emotional expressiveness
- Negative affect (***negaff***) is self report of depression and anxiety

CenterStat 1.52

Example: Deviant Peer Affiliations

- Begin with summary statistics for all predictors and the criterion

Variable	N	Mean	Std Dev	Minimum	Maximum
coa	316	0.53	0.50	0	1.0
gen	316	0.54	0.49	0	1.0
age	316	12.73	1.45	10	16
emotion	316	2.03	0.50	1.07	3.45
stress	316	0.94	0.68	0	3.74
negaff	316	2.88	0.96	1.42	5.76
peer	316	0.39	0.54	0	2.57

CenterStat 1.53

Example: Deviant Peer Affiliations

- We start with simplest possible regression by examining the relation between deviant peer affiliations and COA status
  - this is precisely the same as a two-group t-test
- The regression equation is given as

$$peer_i = \alpha + \gamma_1 coa_i + \zeta_i$$

$$\zeta_i \sim N(0, \psi)$$

- We can estimate this model using the 316 sample cases.

CenterStat 1.54

Example: Deviant Peer Affiliations

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.45	2.45	8.68	0.0035
Error	314	88.63	0.28		
Corrected Total	315	91.08			

Root MSE	0.53	R-Square	0.02
Dependent Mean	0.39	Adj R-Sq	0.02
Coeff Var	136.11		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Semi-partial Corr Type II
Intercept	1	0.29	0.04	6.86	<.0001	0	.
coa	1	0.17	0.05	2.95	0.0035	0.16	0.02

CenterStat 1.55

Example: Deviant Peer Affiliations

- We expanded this model to include the effects of *gen* and *age*

$$peer_i = \alpha + \gamma_1 coa_i + \gamma_2 gen_i + \gamma_3 age_i + \zeta_i$$

- Now we will get two types of tests:
  - The *joint* test of all three predictors as reflected in multiple R-squared
  - The *unique* test of each predictor above-and-beyond the other two predictors as reflected in the partial regression coefficients

CenterStat 1.56

Example: Deviant Peer Affiliations

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	17.88	5.96	25.41	<.0001
Error	312	73.19	0.23		
Corrected Total	315	91.08			

Root MSE	0.48	R-Square	0.19
Dependent Mean	0.39	Adj R-Sq	0.18
Coeff Var	124.09		

CenterStat 1.57

We could compute an F-test of the change in r-squared associated with the inclusion of gender and age, but we do not demonstrate this here.

Example: Deviant Peer Affiliations

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Semi-partial Corr Type II
Intercept	1	-1.61	0.25	-6.41	<.0001	0	.
coa	1	0.20	0.05	3.83	0.0002	0.19	0.03
gen	1	-0.04	0.05	-0.87	0.3844	-0.04	0.001
age	1	0.15	0.01	7.94	<.0001	0.40	0.16

CenterStat 1.58

Example: Deviant Peer Affiliations

- Finally, we can expand this to include *emotion*, *stress*, and *negaff*

$$peer_i = \alpha + \gamma_1 coa_i + \gamma_2 gen_i + \gamma_3 age_i + \\ \gamma_4 emote_i + \gamma_5 stress_i + \gamma_6 negaff_i + \zeta_i$$

- We again get a single joint test of R-squared and unique tests of each of the six predictors

CenterStat 1.59

Example: Deviant Peer Affiliations

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	24.73	4.12	19.19	<.0001
Error	309	66.35	0.21		
Corrected Total	315	91.08			

Root MSE	0.46	R-Square	0.27
Dependent Mean	0.39	Adj R-Sq	0.25
Coeff Var	118.72		

CenterStat 1.60

Example: Deviant Peer Affiliations

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Semi-partial Corr Type II
Intercept	1	-1.93	0.27	-7.15	<.0001	0	.
coa	1	0.13	0.05	2.46	0.0143	0.12	0.01
gen	1	-0.02	0.05	-0.50	0.6149	-0.02	0.001
age	1	0.14	0.01	7.57	<.0001	0.37	0.13
emotion	1	0.03	0.05	0.51	0.6076	0.02	0.001
stress	1	0.11	0.04	2.52	0.0123	0.14	0.01
negaff	1	0.10	0.03	3.62	0.0003	0.19	0.03

CenterStat 1.61

Summary of Findings

- Model regressed deviant peer affiliations on parent alcoholism, gender, age, stress, emotionality, and negative affect
- Ultimately accounted for 27% of the variance
- Significant unique effects of parental alcoholism, age, stress, and negative affect
  - higher levels of each predictor associated with greater peer affiliation
- No significant unique effects of gender or emotionality

CenterStat 1.62

Summary

- Multiple regression model is a powerful and flexible analytic method widely used in social, behavioral, and health sciences
- Can conduct individual and joint tests of relation between predictors and outcome
- DV is assumed continuous and normal, but make no assumptions about distribution of IVs
- Not shown here, but can extend to include polynomial terms and higher-order multiplicative interactions

CenterStat 1.63

1.4 Linear Regression as a Structural Equation Model

Objectives

- Recast the multiple regression as an SEM
- Introduce path diagrams and path tracing rules
- Re-estimate peer example as an SEM

CenterStat 1.65

Regression as an SEM

- It might seem on the surface that the regression and SEM are very different, but regression is a specific type of SEM
- Many core concepts of regression apply to more complex SEMs
  - observed and implied moment structures, unique vs joint tests, raw vs standardized estimates, residual variance and R^2
- There is thus much benefit to redefining the regression model as a specific form of SEM to use as a jumping-off point for the more general models yet to come

CenterStat 1.66

Modeling Steps in the SEM

1. Specification:
   - What is the model? Predictors? Outcomes? Mediators? Factors?
2. Identification:
   - Is it possible to obtain unique estimates for all model parameters?
3. Estimation:
   - How do we obtain the "best" estimates of model parameters?
4. Evaluation:
   - How well does the model fit the data?
5. Potential re-specification:
   - Can the model be improved? How are modifications identified?
6. Interpretation:
   - Which effects are significant? Which are substantively meaningful?

CenterStat 1.67

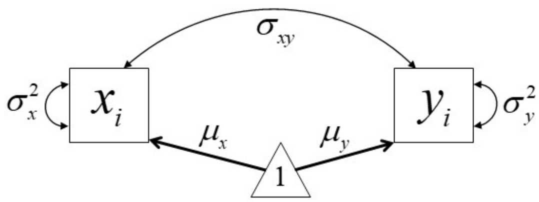
Model Specification & Path Diagrams

- Often model specification done using path diagrams
  - when constructed correctly, path diagrams isomorphic with equations
- There are many ways to define path diagrams, and here we use the most standard methods seen in practice
- There are 5 specific types of shapes used in path diagrams
  1) A rectangle (or square) denotes a measured variable
  2) A circle represents an unmeasured (or latent) variable
  3) A straight one-headed arrow represents a regression coefficient (or factor loading)
  4) A curved two-headed arrow represents a variance or a covariance
  5) A triangle represents a constant value of 1 for means or intercepts

CenterStat 1.68

Path Diagrams

- Let's start with a simple diagram representing the relation between two observed variables

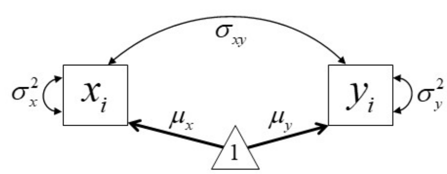


- σ_x^2 is the population variance of x
- σ_y^2 is the population variance of y
- σ_{xy} is the population covariance between x and y
- μ_x is the population mean of x
- μ_y is the population mean of y

CenterStat 1.69

Path Diagrams

- We can more compactly organize these statistics in matrix form



- The population covariance matrix and mean vector is given as

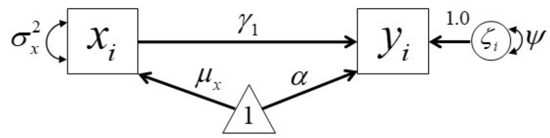
$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_y^2 & \\ \sigma_{xy} & \sigma_x^2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}$$

- We will rely heavily on these expressions as we move forward

CenterStat 1.70

Path Diagrams

- We can make a simple modification where we re-express the covariance between x and y as the regression of y on x



- Note how this precisely corresponds to the regression equation

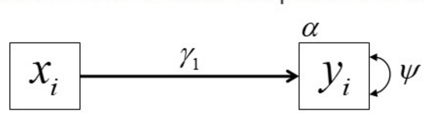
$$y_i = \alpha + \gamma_1 x_{1i} + \zeta_i \qquad \zeta_i \sim N(0, \psi)$$

- Only difference from prior slide is relation between x and y is defined as a *regression coefficient* and mean and variance of y is now an *intercept* and *residual*, respectively

CenterStat 1.71

Path Diagrams

- As models become more complex, the fully articulated path diagrams can become cluttered and tedious
- It is common to make certain simplifications in practice
- We usually omit the variance of the predictor, the triangle representing means and intercepts, and the latent residual
  - the intercept is instead listed above the dependent variable



- We will use variations of this diagram throughout the class, but you will see other expressions across the literature

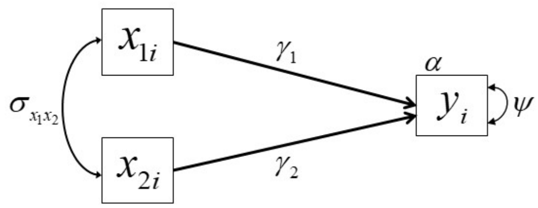
CenterStat 1.72

Path Diagrams

- We can easily extend the diagram for two predictors
- Consider the regression of y_i on two predictors x_1 and x_2

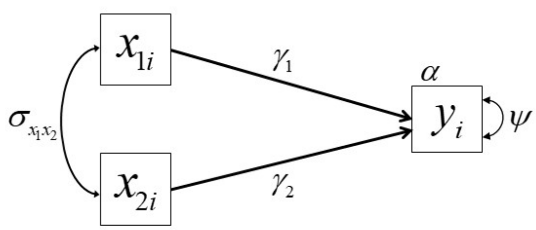
$$y_i = \alpha + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \zeta_i$$

- The corresponding path diagram is



CenterStat 1.73

Path Diagrams



$\sigma_{x_1x_2}$ is the covariance between $x_{1\mathrm{i}}$ and $x_{2\mathrm{i}}$

γ_1 is the unique effect between $x_{1\mathrm{i}}$ and y_i controlling for $x_{2\mathrm{i}}$

γ_2 is the unique effect between $x_{2\mathrm{i}}$ and y_i controlling for $x_{1\mathrm{i}}$

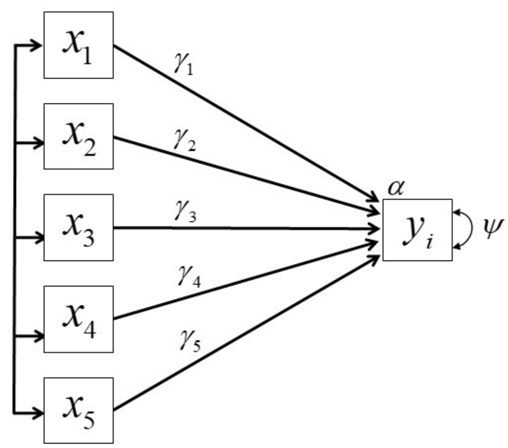
α is the mean of y_i when both predictors equal zero

ψ is the variance in y_i not explained by $x_{1\mathrm{i}}$ and $x_{2\mathrm{i}}$

CenterStat 1.74

Path Diagrams

- The diagram can be expanded to include any number of IVs
  - covariances among all predictors are shown as a single attached line



CenterStat 1.75

Specification: Multiple Regression

- We have control over model specification in terms of selection of dependent variable, predictors, and relations among predictors
  - e.g., main effects, interactions, polynomial effects
- However, other than predictors and outcome, no further decisions are possible
  - the single dependent variable is regressed on all predictors
  - all predictors correlate with one another
- The structural specification of the regression is thus limited
  - but in the next chapter we will see that there is substantial control we can exert over the specification of more complex path models

CenterStat 1.76

Specification: Multiple Regression

- A key advantage of path diagrams is that they can help derive the *model-implied moment structure* of the observed variables
  - "moments" are means, variances and covariances of set of variables
- The specification of a given model implies a specific form of means, variances, and covariances among the observed variables
  - one particular specification of relations among a set of variables will result in one corresponding moment structure, whereas another specification will result in another moment structure
- Model-implied moment structures are fundamental in all SEMs and we consider these more closely here

CenterStat 1.77

Moment Structures

- To begin, consider four types of covariance and mean structures
  - these are the first two "moments" of a multivariate distribution

1. Population moments denoted $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$
2. Sample moments denoted $\mathbf{S}$ and $\mathbf{m}$
3. Population model-implied moments denoted $\mathbf{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$
4. Sample model-implied moments denoted $\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$

- We'll talk about 1 and 3 here, and revisit 2 and 4 in a moment.

CenterStat 1.78

Population Moments

- For one outcome and q-predictors (as in the regression), there exists a population covariance matrix and mean vector denoted:

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|cccc} \sigma_y^2 & & & & \\ \hline \sigma_{x_1 y} & \sigma_{x_1}^2 & & & \\ \sigma_{x_2 y} & \vdots & \sigma_{x_2}^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \sigma_{x_q y} & \sigma_{x_q x_1} & \sigma_{x_q x_2} & \cdots & \sigma_{x_q}^2 \end{array}\right] \qquad \boldsymbol{\mu} = \left[\begin{array}{c} \mu_y \\ \hline \mu_{x_1} \\ \mu_{x_2} \\ \vdots \\ \mu_{x_q} \end{array}\right]$$

- In SEM, we list the DVs first followed by the IVs,
  - in multiple regression, we only have one DV; later SEMs will not

CenterStat 1.79

The lines are drawn in the matrix and vector simply to offer a visual guide when ordering the DV listed first and the IVs second. These become more useful later when we consider multiple DVs.

Population Model-Implied Moments

- Next we consider the moment matrix *implied by the model*
  - this is similar to y and $\hat{y}$ in regression, but at the level of the covariance matrix and mean vector
- To begin, we define the regression model in the usual way:

$$y_i = \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i$$

- Next, we gather the intercept, regression parameters, and residual variance together in a single vector:

$$\boldsymbol{\theta}' = \begin{bmatrix} \alpha & \gamma_1 & \gamma_2 & \cdots & \gamma_q & \psi \end{bmatrix}$$

CenterStat 1.80

Population Model-Implied Moments

- This vector always represents "the model"

$$\boldsymbol{\theta}' = \begin{bmatrix} \alpha & \gamma_1 & \gamma_2 & \cdots & \gamma_q & \psi \end{bmatrix}$$

- The covariance and mean structures that are implied by the vector of parameters are denoted $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$
- Each particular form of SEM has its own unique expression for the model implied mean and covariance structure
- This can be derived from matrix algebra, but is more clearly seen using *path tracing rules*

CenterStat 1.81

Path Tracing Rules

- Path analysis invented by Sewall Wright in 1920's and 1930's
  - path analysis is the focus of Chapters 2 and 3
- Wright laid the foundation for *path tracing rules*
  - formal guidelines that derive the moment structures implied by model
  - instead of matrix algebra, use set rules to trace through path diagram
- Motivating goal is to compute the sum of all pathways so that you can:

1. leave from and return to a variable to reconstruct the *variance*, or
2. leave one variable and travel to another to reconstruct the *covariance*

- There are three core rules that allow us to do this

CenterStat 1.82

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

A free PDF version of this book can be found here: http://davidakenny.net/doc/cc_v1.pdf

Path Tracing Rules: Covariance Structure

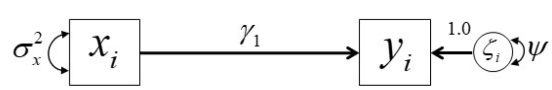
- Rule 1: When you begin a trace backward from a variable using a single headed arrow, you can proceed backward any number of times; upon reaching a variable, you can do one of two things:
  1. span a double-headed arrow and stop
  2. span a double-headed arrow and proceed the trace forward
- Rule 2: Once you start forward you may move forward any number of times but, but after moving forward you may not move backwards or span a double-headed arrow
- Rule 3: For any given trace you can only pass through a variable once
  - thus no loops are allowed

CenterStat 1.83

Note that the double-headed arrow can either be a covariance or a variance and you can only span a double-headed arrow **one time** in any given trace.

Path Tracing Rules

- Let's consider the one-predictor regression
  - we start with covariance structure and consider means in a moment



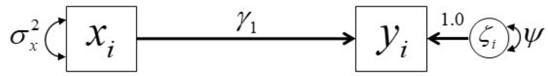
- To derive the model-implied variance of x, we start at x and cross the span back to x

$$\sigma_x^2(\theta) = \sigma_x^2$$

- This nicely highlights that we're not imposing any structure on the variance of the predictor and it simply takes on the population value

CenterStat 1.84

Path Tracing Rules



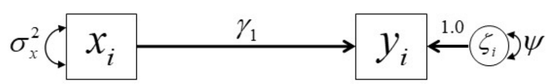
- To derive the model-implied covariance between x and y we start at y, trace backwards, and span the double-headed arrow of x :

$$\sigma_{yx}(\theta) = \gamma_1 \sigma_x^2$$

- Note that, not coincidentally, we can solve for γ_1 such that $\gamma_1 = \dfrac{\sigma_{yx}}{\sigma_x^2}$
- This is the OLS estimate of the regression coefficient

CenterStat 1.85

Path Tracing Rules



- To derive the model-implied variance of y, we start at y, trace backwards and span the double-headed arrow of x and return whence we came; to this we add the residual trace resulting in

$$\sigma_y^2(\theta) = \gamma_1 \sigma_x^2 \gamma_1 + 1\psi 1 = \gamma_1^2 \sigma_x^2 + \psi$$

- The first term is the *explained* variance, the second term is the *residual* variance, and the sum is the *total* variance

CenterStat 1.86

Path Tracing Rules

- We can gather these all together in the covariance matrix that is implied by the model

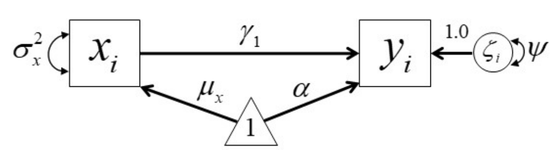
$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \gamma_1^2\sigma_x^2 + \psi & \\ \gamma_1\sigma_x^2 & \sigma_x^2 \end{bmatrix}$$

- The important thing to see here is that we expressing the population covariance matrix of our two measures entirely in terms of parameters that define the model
- This concept plays a fundamental role through all of SEM
  - we "repackage" the moment structure in terms of model parameters

CenterStat 1.87

Path Tracing Rules: Means

- We can return to the comprehensive diagram to derive the model-implied mean structure

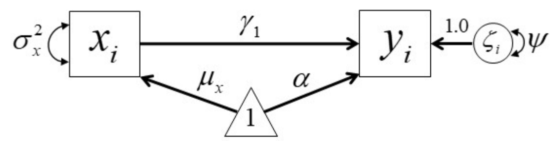


- There is only one tracing rule for mean structures: you begin at the triangle, can only move forward, and cannot span a double-headed arrow
- As you proceed down the trace, you simply pick up the relevant coefficients as you proceed

CenterStat 1.88

The value of 1.0 on the path for the disturbance is often assumed to exist but is not explicitly included in the diagram, and we adopt this convention from here forward.

Path Tracing Rules: Means

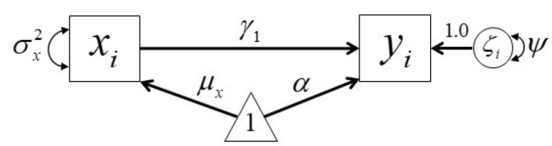


- The model-implied mean of x is simply $\mu_x(\theta) = \mu_x$
  - as with the model implied variance of x, this highlights that we are not imposing any structure on the exogenous predictor
- The model-implied mean of y involves the sum of two traces

$$\mu_y(\theta) = \alpha + \mu_x \gamma_1$$

CenterStat 1.89

Model-Implied Moment Structures



- Gathering everything together, the model implied moment structure for a one-predictor regression is

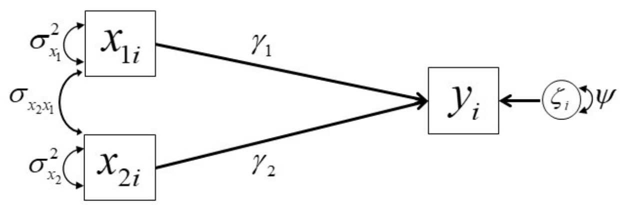
$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \gamma_1^2 \sigma_x^2 + \psi & \\ \gamma_1 \sigma_x^2 & \sigma_x^2 \end{bmatrix} \qquad \boldsymbol{\mu}(\boldsymbol{\theta}) = \begin{bmatrix} \alpha + \mu_x \gamma_1 \\ \mu_x \end{bmatrix}$$

- We will return repeatedly to structures such as these all week

CenterStat 1.90

Path Tracing Rules

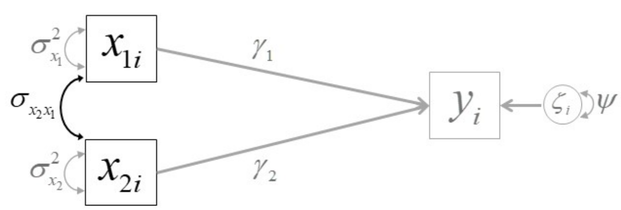
- Let's consider a slightly more complicated two-predictor model



- Potential traces are more involved because we must also consider the covariance between the two predictors
  - note we add the double-headed arrows on the predictors back into the diagram to show the variance; we don't usually include these in practice
- We won't show all possible traces, but let's examine a few

CenterStat 1.91

Path Tracing Rules



- As a simple starting point, the covariance between the two predictors is easy: we start at x_1 and span to x_2
  - there is only one span possible to get from x_1 to x_2

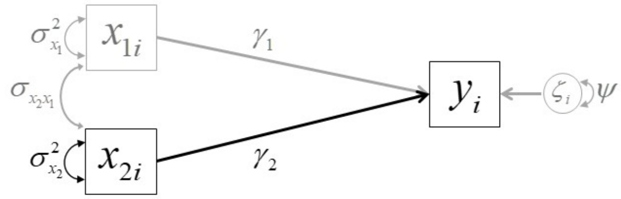
$$\sigma_{x_2x_1}(\theta) = \sigma_{x_2x_1}$$

  - this highlights that we are not imposing structure on the predictors

CenterStat 1.92

We lighten the paths and symbols simply to highlight each individual trace under consideration.

Path Tracing Rules



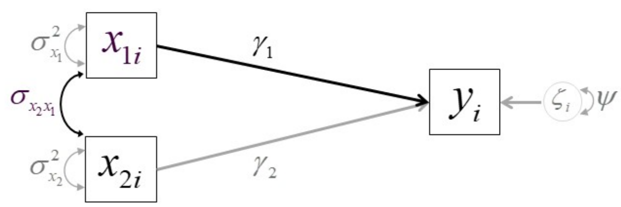
- Let's do a harder one: the covariance between y and x_2
- First, we can start at y, track back to x_2 and span the double-arrow to itself resulting in:

$$\gamma_2\sigma^2_{x_2}$$

- But there's a second trace we can also make…

CenterStat 1.93

Path Tracing Rules



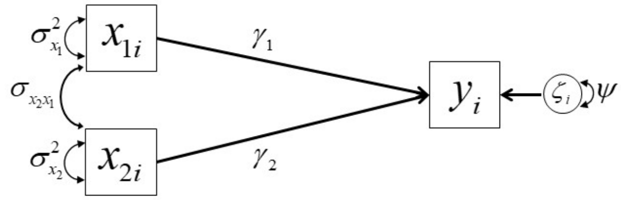
- We can **also** start at y, track back to x_1, and span the double-arrow to x_2 resulting in:

$$\gamma_1\sigma_{x_2x_1}$$

- We need to add this to the prior trace

CenterStat 1.94

Path Tracing Rules

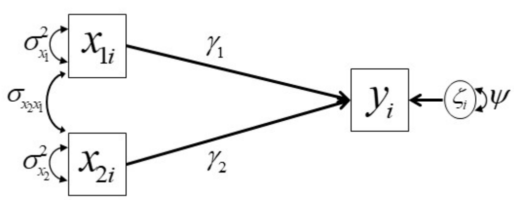


- The sum of the two possible traces construct the model-implied covariance between y and x_2:

$$\sigma_{y,x_2}(\theta) = \gamma_2\sigma^2_{x_2} + \gamma_1\sigma_{x_2x1}$$

CenterStat 1.95

Path Tracing Rules



The variance of y is harder still: there are five traces possible

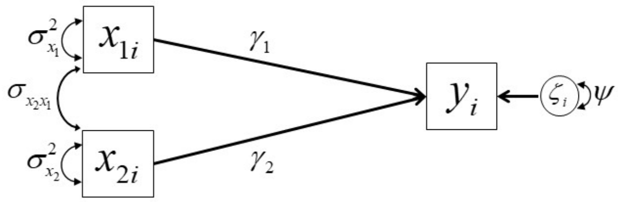
1. y to x_1 and back: $\gamma_1\sigma^2_{x_1}\gamma_1 = \gamma_1^2\sigma^2_{x_1}$
2. y to x_2 and back: $\gamma_2\sigma^2_{x_2}\gamma_2 = \gamma_2^2\sigma^2_{x_2}$
3. y to x_1 to x_2 and back: $\gamma_1\sigma_{x_2x_1}\gamma_2$
4. y to x_2 to x_1 and back: $\gamma_2\sigma_{x_2x_1}\gamma_1$

   (3 and 4) $= 2\gamma_1\gamma_2\sigma_{x_2x_1}$
5. y to ζ and back: $1\psi 1 = \psi$

CenterStat 1.96

We don't lighten each individual trace for the variance given how many sub-traces there are.

Path Tracing Rules



- Putting all of these together, we get:

$$\sigma_y^2(\theta) = \gamma_1^2\sigma_{x_1}^2 + \gamma_2^2\sigma_{x_2}^2 + 2\gamma_1\gamma_2\sigma_{x_1x_2} + \psi$$

- This expresses the model-implied variance of y in terms of the covariance structure among the IVs and the model parameters

CenterStat 1.97

Path Tracing Rules

- Tracing rules can be used to derive every element of the covariance matrix among the IVs and DV
- But rules get wickedly complex to track with many predictors
  - particularly for more advanced SEMs yet to come
- Fortunately, the matrix fairy gives these all-in-one expression
  - knowing these expressions is not fundamental to using them in practice, but it is good to at least be familiar with these
  - we thus show them throughout the week when needed
- However, a key advantage of tracing rules is to help establish *model identification*, and it is to this we now turn

CenterStat 1.98

Identification

- *Model identification* refers to the extent to which there is sufficient known information to infer unknown values
- Consider a simple equation with one unknown: $x - 5 = 6$
  - can uniquely solve equation because one unknown value (x) and two known values (5 and 6)
- But consider a second equation with two unknowns: $x - y = 6$
  - cannot uniquely solve equation because insufficient information
- This analogy reflects core issue of *identification*

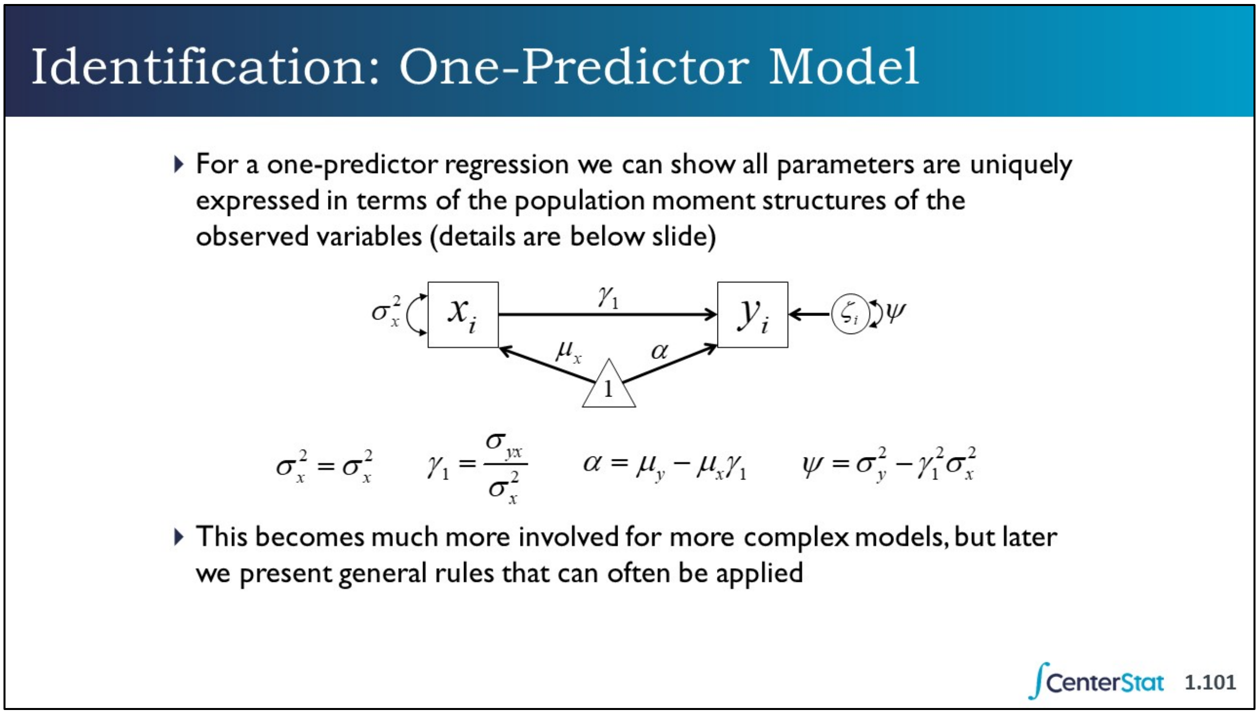
CenterStat 1.99

Model Identification

- Over-identified: there is more observed information than the number of parameters to be estimated
  - this is the case for many path analysis models and nearly all SEMs
- Just-identified: there is the same amount of observed information as the number of parameters to be estimated
  - this is always the case for the multiple regression model
- Under-identified: there is less observed information than the number of parameters to be estimated
  - this is the case that gives you a royal headache
- It is easy to show that regression model is *always* just-identified

CenterStat 1.100

In later chapters we will also describe a number of identification rules that exist and serve as shortcuts for establishing the identification of various types of SEMs.



Although we will never need to establish identification for a regression model in practice (because all regression models are *just identified*) it is helpful to consider identification in regression to set a foundation for moving to more complex path models presented in the following chapter. We can do this in four steps.

First, the model-implied variance of the predictor x is easy; this is simply $\sigma_x^2(\theta) = \sigma_x^2$.

Second, because the covariance between x and y is $\sigma_{yx}(\theta) = \gamma_1\sigma_x^2$, we can solve for γ_1 as a function of the population moments $\gamma_1 = \sigma_{yx} / \sigma_x^2$ thus showing the regression coefficient is identified.

Third, we already know the model-implied variance of y is $\sigma_y^2(\theta) = \gamma_1\sigma_x^2\gamma_1 + 1\psi 1 = \gamma_1^2\sigma_x^2 + \psi$ and simple algebra gives us $\psi = \sigma_y^2 - \gamma_1^2\sigma_x^2$ showing that the residual variance is identified.

Finally, we see that the mean of x is μ_x and the intercept of y is α . Thus the model-implied mean of y is $\mu_y = \alpha + \mu_x\gamma_1$ and simple algebra gives the intercept as $\alpha = \mu_y + \mu_x\gamma_1$ establishing that the intercept is identified.

These steps can be generalized to standard multiple regression models with any number of predictors and always result in just identified models.

Summary Thus Far

- We're covering lots of ground, so let's briefly review thus far
- The first step in defining the regression model as an SEM is *specification* of the model structure
  - what variable is DV, which are IVs, what are relations among IVs
- Can express specification of model as a path diagram
  - certain rules allow derivation of model equations from diagram
  - additional rules allow for derivation of model-implied moment structure
- Establishing model-implied moments allows for *identification* step
  - can each parameter be uniquely expressed using observed information
- Once established, specification and identification allows us to move forward to the next step: *estimation*

CenterStat 1.102

Estimation

- Core topic that relates to all we discuss this week is *estimation*
  - how do we obtain our parameter estimates from sample data?
- Three characteristics of an estimator:
  - <u>unbiased</u>: if we were to repeat our study an infinite number of times, the mean of the sample estimates would equal the population value
  - <u>consistent</u>: as the sample size approaches infinity, the sample estimate approaches the population value
  - <u>efficient</u>: no other estimator has a smaller sampling error for the estimate
- The estimator used to fit most SEMs is maximum likelihood (ML)
  - ML provides estimates for model parameters that have the highest likelihood of giving rise to the sample data

CenterStat 1.103

Model Estimation

- ML estimation based on same moment structures from earlier, but we now add the sample counterparts
- Recall population and model implied moment structures were

$$\left.\begin{matrix}\boldsymbol{\Sigma}\\ \boldsymbol{\mu}\end{matrix}\right\}\ \text{population moments} \qquad \left.\begin{matrix}\boldsymbol{\Sigma}(\boldsymbol{\theta})\\ \boldsymbol{\mu}(\boldsymbol{\theta})\end{matrix}\right\}\ \text{population model-implied moments}$$

- We add two **sample** counterparts for covariances and means

$$\left.\begin{matrix}\mathbf{S}\\ \mathbf{m}\end{matrix}\right\}\ \text{sample moments} \qquad \left.\begin{matrix}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\\ \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})\end{matrix}\right\}\ \text{sample model-implied moments}$$

  - These are precisely the same structures, but based on the observed sample of data

CenterStat 1.104

Sample Observed Moments

- For the same set of observed variables there exists a sample covariance matrix and mean vector denoted:

$$\mathbf{S} = \left[\begin{array}{c|cccc} s_y^2 & & & & \\ \hline s_{x_1 y} & s_{x_1}^2 & & & \\ s_{x_2 y} & s_{x_2 x_1} & s_{x_2}^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ s_{x_q y} & s_{x_q x_1} & s_{x_q x_2} & \cdots & s_{x_q}^2 \end{array}\right] \qquad \mathbf{m} = \left[\begin{array}{c} \bar{y} \\ \hline \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_q \end{array}\right]$$

- These are sample estimates of the corresponding population values calculated in the usual way

CenterStat 1.105

Guidelines within the matrix and vector are sometime used to delineate the dependent variables from the independent variables, and we do this here. In regression there is only a single DV, but there will be multiple DVs as we move to more complex models.

Sample Model-Implied Moments

- Recall the *population* vector of model parameters is $\boldsymbol{\theta}$
- The *sample* vector of model parameters is $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}}' = \begin{bmatrix} \hat{\alpha} & \hat{\gamma}_1 & \hat{\gamma}_2 & \cdots & \hat{\gamma}_q & \hat{\psi} \end{bmatrix}$$

  - these are our usual sample estimates of regression parameters, but we describe momentarily how we obtain these in the SEM framework

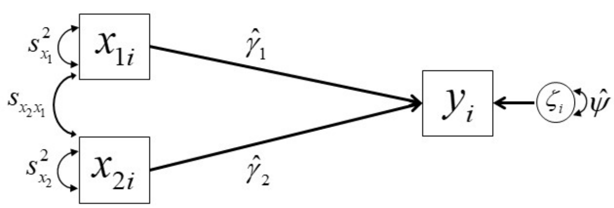
- Using the same path tracing rules described earlier, we can derive the model-implied covariance matrix and mean vector based on *sample estimates* of the model parameters:

$$\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right) \qquad \boldsymbol{\mu}\left(\hat{\boldsymbol{\theta}}\right)$$

CenterStat 1.106

Sample Model-Implied Moments

- Just as one example, consider two predictor example but now with sample estimates instead of population values



- The model-implied sample variance of y is thus

$$\sigma^2_y\left(\hat{\theta}\right) = \hat{\gamma}_1^2 s^2_{x_1} + \hat{\gamma}_2^2 s^2_{x_2} + 2\hat{\gamma}_1\hat{\gamma}_2 s_{x_1x_2} + \hat{\psi}$$

- All other matrix elements can be derived similarly

CenterStat 1.107

Model Estimation

- Goal to select values for model parameters based on sample data so that *model-implied* covariance matrix and mean vector are as close as possible to *sample* covariance matrix and mean vector
  - *jointly* minimize difference between observed and implied moments:

$$\min\left(\mathbf{S}-\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)\right) \qquad \min\left(\mathbf{m}-\boldsymbol{\mu}\left(\hat{\boldsymbol{\theta}}\right)\right)$$

- Note the similarity to estimation in the regression model

$$\min\left(y_i-\hat{y}_i\right)$$

- Analytically these procedures are different, but conceptually we are trying to do the very same thing as in OLS regression

CenterStat 1.108

Maximum Likelihood Estimation

- ML discrepancy function optimized to give smallest possible differences between observed and model-implied mean and covariance structures
- Two general approaches:
  1. **Sufficient-statistic ML** (SSML) estimation is based solely on the observed covariance matrix and mean vector
     - assumes complete-case data and normally distributed DVs
  2. **Full-information ML** (FIML) estimation is based on whatever data are observed for each individual case
     - allows for partially missing data and alternative methods for addressing non-normal distributions and nested data structures
- For complete and normal data, SSML and FIML are identical

CenterStat 1.109

See the Appendix in these notes for more details about ML estimation. Also, note that the assumption of continuous and normal distributions applies only to the dependent variables in the model, not the independent (or predictor) variables. For various reasons, some historical and others practical, the SEM was first formulated assuming using *joint* multivariate normality for all IVs and DVs. However, under broad conditions, this joint assumption of normality yields the same maximum likelihood estimates as the assumption of normality applied to just the DVs, even when the IVs are not normally distributed (Jöreskog, 1973; see also Bollen, 1989, pp. 126-128). Thus, the distinction is typically of little practical importance and we may include non-normally distributed predictors in our regression models without concern even when using the joint normality assumption to obtain the MLEs.

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). New York: Academic Press.

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

Beneficial Properties of ML

- Can be used to fit a wide variety of models
- ML is asymptotically unbiased, consistent, and maximally efficient
- Estimates are asymptotically normally distributed, providing basis for inference tests
- Combinations of ML estimates are themselves ML estimates
- Relative fit of competing models can be compared using likelihood ratio tests (chi-square difference tests)
- FIML can accommodate partially missing and non-normal data

CenterStat 1.110

Steps in ML Estimation

- SSML and FIML approach estimation in a broadly similar way
- First, initial values for parameter estimates selected
  - called *start values*
- Next, the likelihood computed and estimates updated
  - called an *iteration*
- The likelihood computed again, and this continues until difference between two successive likelihood values is sufficiently small
  - means ML has *converged*
- Fit statistics, parameter estimates, and standard errors retained from final step

CenterStat 1.111

A problem sometimes encountered in more complex SEMs is that ML “failed to converge” and no solution is provided.

Maximum Likelihood Estimation

- Upon successful completion, ML provides sample estimates of population parameters
  - can be scaled in raw metric, standardized, or partially-standardized
- It also provides standard errors for each estimate
  - this allows our usual inferential tests of parameters
- Finally, it provides information about overall model fit that is critical to all of the SEM
- All of this information is considered in *model evaluation* that we consider next

CenterStat 1.112

Model Evaluation

- In SEM, model evaluation typically involves multiple components
  - How well does the model fit the data?
  - Which estimates are significant?
  - How well can I predict the outcome?

CenterStat 1.113

The Fundamental Hypothesis

- To evaluate fit, we must first define our null hypothesis
- In SEM, the *fundamental hypothesis* states that the population and model-implied population moment structure are equal:

$$H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta}), \;\; \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$$

- Can equivalently state hypothesis as

$$H_0 : \mathbf{\Sigma} - \mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{0}, \;\; \boldsymbol{\mu} - \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{0}$$

  - this highlights the "null" (or "zero") hypothesis
- This hypothesis is stated *at the level of the population*
  - but we must evaluate this based on *sample* data

CenterStat 1.114

Note that this is a different notion of "fit" than what is captured by R^2 (which measures strength of prediction). Here we are solely concerned with reproducing the values of the means, variances, and covariances of y and $\mathbf{x}$, regardless of the strength of prediction. Of course, both aspects of fit are important and should be considered when evaluating the model.

Evaluation: Model Fit

- In linear regression model, it is always true that

$$\mathbf{S} = \mathbf{\Sigma}(\hat{\boldsymbol{\theta}}) \qquad \mathbf{m} = \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$$

- The just-identified regression model uses the same number of parameters to exactly reproduce the observed moments
  - The model is said to be "saturated"
  - The model necessarily (and trivially) fits perfectly
- Most SEMs, however, are over-identified, resulting in testable restrictions on the model-implied moments
  - We will thus explore this in greater detail in Chapter 3

CenterStat 1.115

Evaluation: Tests of Parameters

- Two kinds of tests are commonly conducted with multiple regression and with SEM in general
  - Tests of individual parameters
  - Joint tests of multiple parameters

CenterStat 1.116

Tests of Individual Parameters

- A property of maximum likelihood is that the parameter estimates are asymptotically normally distributed
- For a parameter θ, we can thus test the null that $H_0 : \theta = 0$ using the large-sample z-test (called a Wald test) where:

$$z = \frac{\hat{\theta}}{SE_{\hat{\theta}}}$$

- Alternatively, we can compute a confidence interval as $\hat{\theta} \pm z_{crit} SE_{\hat{\theta}}$
- More common in regression to use the t-distribution, but in the SEM we almost always assume large *N* and use a z-distribution

CenterStat 1.117

This z-test is the large sample equivalent of the t-test that is commonly implemented when fitting regression models via OLS. The t-distribution converges to the z-distribution as N increases and become functionally indistinguishable at about 25 degrees of freedom. Likewise, the z-type confidence interval is the large sample equivalent of the t-type confidence interval.

Joint Parameter Tests

- We can also test multiple parameters simultaneously
- For instance, suppose we have the model

$$y_i = \alpha + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \zeta_i$$

- We may wish to test whether our predictors jointly explain any variance in y using the null hypothesis

$$H_0 : \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- This null hypothesis can be tested using a multivariate Wald test or a likelihood ratio test
  - this can also be construed as a test of multiple R^2

CenterStat 1.118

This null hypothesis is equivalent to the null hypothesis for the test of R^2 in the traditional multiple regression model. However, whereas an F-test is typically used to evaluate this null hypothesis when using OLS, with ML one would implement a Wald test or likelihood ratio test. The multivariate Wald χ^2 test is the large-sample equivalent to the F-test.

Model Interpretation

- For interpretation, we usually focus on
  - Raw parameter estimates
  - Standardized parameter estimates
  - R^2, explained variance in outcomes

CenterStat 1.119

Interpretation: Raw Estimates

- The raw regression coefficients are in the metric of the predictor and criterion measures
  - e.g., a one-unit change in x is associated with a γ-unit change in y
- Often useful to standardize the coefficients to reflect a standardized metric for x and y
  - e.g., a one standard deviation change in x is associated with a γ^*-unit change in y

CenterStat 1.120

Partially Standardized Estimates

- Sometimes we may want to partially standardize effects
  - If x or y has an intrinsically meaningful metric
  - If x represents a coding variable for a categorical predictor
- Example: x is a binary predictor
  - Standardizing on y only, we obtain the standardized mean difference between group $x = 1$ versus $x = 0$ (i.e., in standard deviation units of y)
- Example: y is annual income
  - Standardizing on x only, we obtain the expected dollar increase in y per standard deviation unit change in x

CenterStat 1.121

Variance Explained: R²

- In multiple regression, often want to know how much variance has been explained in our outcome, y, by the predictors $\mathbf{x}$:

$$\rho^2 = \frac{VAR(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i)}{VAR(\alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i)} = \frac{explained}{total}$$

- Sample estimate is R^2
- Ranges from 0 to 1 and indicates proportion of variance explained jointly by the set of predictors
- Also used extensively in factor analysis and SEM
  - Sometimes goes by another name, "communality"
- Let's return briefly to our deviant peer affiliation final model

CenterStat 1.122

Example: Deviant Peer Affiliations

- Recall our final model estimated as an OLS regression was

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	24.73	4.12	19.19	<.0001
Error	309	66.35	0.21		
Corrected Total	315	91.08			

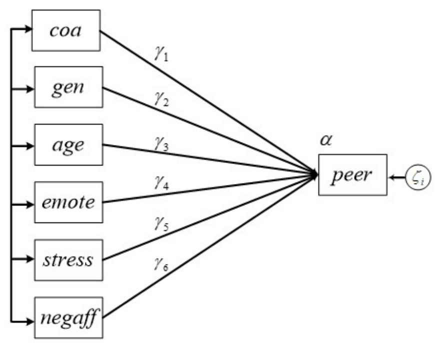
Root MSE	0.46	R-Square	0.27
Dependent Mean	0.39	Adj R-Sq	0.25
Coeff Var	118.72		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Semi-partial Corr Type II
Intercept	1	-1.93	0.27	-7.15	<.0001	0	.
coa	1	0.13	0.05	2.46	0.0143	0.12	0.01
gen	1	-0.02	0.05	-0.50	0.6149	-0.02	0.001
age	1	0.14	0.01	7.57	<.0001	0.37	0.13
emotion	1	0.03	0.05	0.51	0.6076	0.02	0.001
stress	1	0.11	0.04	2.52	0.0123	0.14	0.01
negaff	1	0.10	0.03	3.62	0.0003	0.19	0.03

CenterStat 1.123

Example: Deviant Peer Affiliations

- We can use a path diagram for the first step of **specification**



- We also know from earlier path tracing rules that all regression models are **just identified**, so that completes step 2

CenterStat 1.124

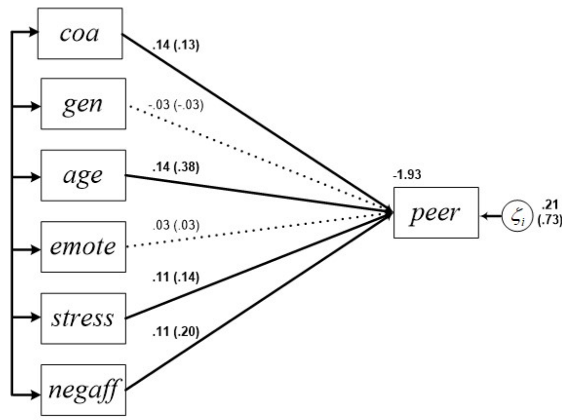
Example: Deviant Peer Affiliations

- For step 3, we will use ML for **estimation**
  - given there are no missing data, SSML and FIML are identical here
- Although ML provides identical parameter estimates to OLS, the residual variance and standard errors will be slightly different
  - OLS includes *df* in error variance estimate where ML does not
- For **evaluation** there are no tests of overall model fit because the regression model is just-identified and thus fits perfectly
  - but we can examine regression coefficients and multiple r-squared
- There is typically no **respecification** step in OLS regression
  - although residual plots could be examined for omitted interactions, etc.
- Finally, we will approach **interpretation** as guided by theory

CenterStat 1.125

Example: Deviant Peer Affiliations

- Our final regression model as estimated as an SEM is below
  - solid lines significant & dashed non-sig; raw coefficients first, then standardized (*coa*, *gen*, & *age* partially standardized and others fully)



CenterStat 1.126

Summary

- This section covered all core elements of the SEM
  - specification
  - identification
  - estimation
  - evaluation
  - respecification
  - interpretation
- We revisit this process in nearly every remaining chapter
- Following the above, we see multiple regression is an SEM
- Despite strengths, there are also limitations to the regression, many of which are easily addressed in the full SEM

CenterStat 1.127

1.5 Limitations of the Multiple Regression Model

Objectives

- Review advantages of regression model but also note assumptions and limitations
- Introduce the basic concept of mediation
- Describe tests of mediation using the regression model
- Establish the need for a path analysis model

CenterStat 1.129

Advantages of Regression Model

- Provides individual tests of the unique contribution of each predictor net all other predictors
- Provides joint tests of sets of predictors and estimates proportion of explained variance via R^2
- Can include continuous or categorical predictors
- Can include interactions and power terms

CenterStat 1.130

Assumptions of the Regression Model

- The traditional form of the multiple regression model is

$$y_i = \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i \qquad VAR(\zeta_i) = \psi$$

Assumptions:

- All effects of $\mathbf{x}$ on y are linear
- Residuals have constant variance ψ (homoscedasticity)
- Residuals are normally distributed
- Residuals are independent
- A single outcome variable, y
- Only direct effects from $\mathbf{x}$ to y
- All variables measured without error

- We can overcome many of these assumptions by embedding and expanding the regression model within the SEM

CenterStat 1.131

Overcoming Limitations of Regression

- SEM allows for latent variables to account for measurement error; Chapters 4 and 5
- SEM can allow for differences in mean and variance structures across groups (e.g., heteroscedasticity); Chapter 6
- SEM can be generalized to discrete outcomes; Chapter 7
- SEMs can be used to model dependent observations (e.g., in repeated measures); Chapter 8
- Next we'll see how SEMs extend beyond the regression model to allow multiple DVs and indirect effects / mediation; Chapters 2 and 3
  - We will first extend from regression to path analysis then from path analysis to full SEM with latent variables

CenterStat 1.132

Multiple Dependent Variables: Mediation

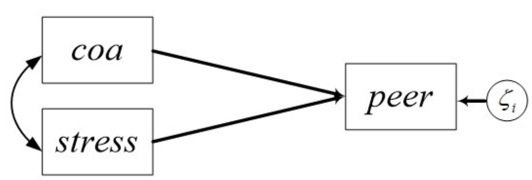
- The regression model evaluates the question:
  - *What is the relation between the predictor and the outcome above and beyond all other predictors?*
- However, theory might posit a different question:
  - *How does one variable influence another that in turn influences a third?*
- This is the central question in *mediation*
- Baron & Kenny (1986): *"In general, a given variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the criterion".* *(p1176)*
- A mediator thus attempts to assess *why* or *how* one variable is related to another; sometimes called an *intervening variable*

CenterStat 1.133

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.

Mediation

- Returning to our prior example, consider two models
- The first examines the relation between *coa* and *peer* net the effects of *stress*

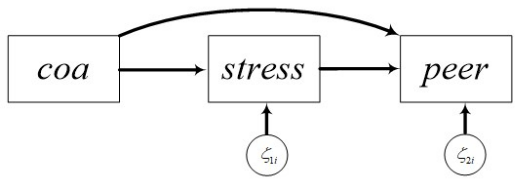


CenterStat 1.134

This represents a simple two-predictor multiple regression model.

Mediation

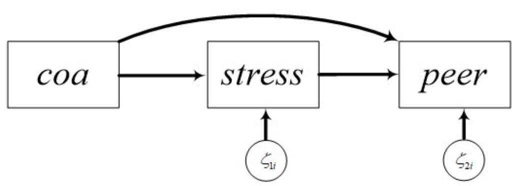
- The second dictates that the relation between *coa* and *peer* is partially due to stress



  - parental alcoholism may increase environmental stress that in turn increases deviant peer affiliations (the ***indirect*** effect)
  - parental alcoholism may increase deviant peer affiliations via other mechanisms net of environmental stress (the ***direct*** effect)

CenterStat 1.135

Mediation



- This simple mediating model requires two dependent variables
  - *stress* regressed on *coa*; and *peer* regressed on *stress* & *coa*
  - these cannot be obtained from a single regression model

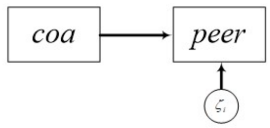
CenterStat 1.136

This represents a simple two-predictor multiple regression model.

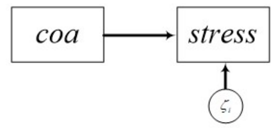
Estimate Separate Regression Models

- A well-established method for testing mediation is to estimate separate regression models (Baron & Kenny, 1986)
- For our example, we would estimate two of three models:

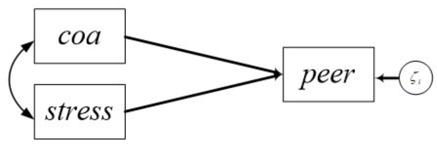
(1)



(2)



(3)



CenterStat 1.137

Estimate Separate Regression Models

- This approach works well for simple mediation models
  - e.g., one mediating variable and one dependent variable
- However, it becomes more challenging for more complex models
- This strategy also precludes obtaining a single omnibus test of model fit
  - each component regression model is saturated (i.e., all possible parameters are estimated), so cannot test overall hypothesized model
- The path analysis model will allow for these omnibus tests, as well as several other things not available within the multiple regression framework
  - path analysis is a type of SEM with only observed variables

CenterStat 1.138

Summary

- Multiple regression is a powerful analytic method and is useful in a variety of applied research settings
  - joint and unique tests of regression coefficients are possible
  - predictors can enter the model as polynomials or interactions
- But key limitation of regression model is that it only allows for one outcome variable at a time
  - thus statistical model may not optimally correspond to theoretical model
- For some simpler models, can estimate separate regressions and piece results together to test mediation, but many limitations
- More general path models estimated within the SEM will overcome many of these limitations

CenterStat 1.139

Chapter Summary

- SEM subsumes and extends many familiar statistical models
- To better understand the statistical aspects of SEM, it is helpful to be familiar with matrix algebra
- Can also define regression as a specific form of SEM
  - express model as a path diagram, derive model-implied moment structure, use ML estimation
- SEM-based regression precisely same as OLS regression, but now embedded within the broader SEM framework
- We next turn to a generalization of the regression model called path analysis (or simultaneous equations model) that permits multiple dependent variables and indirect effects

CenterStat 1.140

Chapter 2
Path Analysis: Part I

2.1 The Path Analysis Model

Objectives

- Introduce the general concept of the path analysis model
- Demonstrate different types of path analysis models
  - recursive and non-recursive
- Define the model

CenterStat 2.3

Path Analysis Models

- Path analysis models expand the multiple regression model to allow for
  - more than one criterion measure
  - omnibus tests of fit of restricted models
    - e.g., not all paths need be estimated as is done in regression
  - formal tests of direct and indirect (or mediated) effects
  - feedback loops among multiple variables (e.g., bi-directional effects)
- Go by several different names in the literature
  - simultaneous equations models (e.g., econometrics)
  - SEMs with observed measures (e.g., behavioral sciences)

CenterStat 2.4

Definitions

- Exogenous variable: a variable that is not expressed as a function of other variables; its causes are "outside" of the system
  - also called an independent variable (IV)
  - in path diagrams, no single-headed arrows point to it
- Endogenous variable: a variable that is expressed as a function of one or more other variables; its causes are "within" the system
  - also called a criterion variable or dependent variable (DV)
  - in path diagrams, at least one single-headed arrow points to it
- Disturbance: the residual (or unexplained) variance of an endogenous (or dependent) variable
  - the part that is unrelated to the predictors

CenterStat 2.5

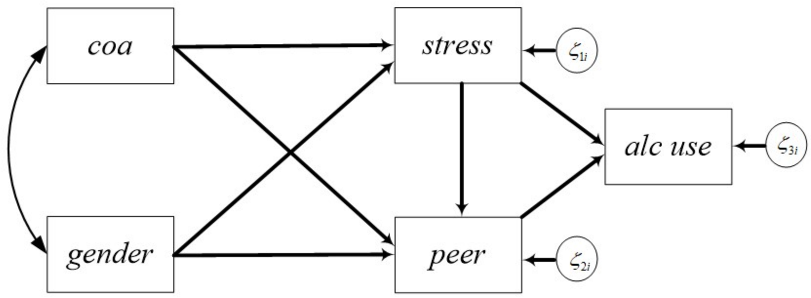
Recursive vs. Non-Recursive Models

- There are two types of path models: recursive and non-recursive
  - Their definitions are somewhat non-intuitive
- A recursive model is defined as having
  1. no correlated residuals, and
  2. only unidirectional effects
- A non-recursive model is defined as having
  1. correlated residuals, and/or
  2. feedback loops (e.g., bi-directional effects)
- This distinction has little bearing on how we estimate or interpret these models, but helps in communicating model structure and establishing identification

CenterStat 2.6

Example: Recursive

- No correlated residuals, bi-directional effects, or feedback loops

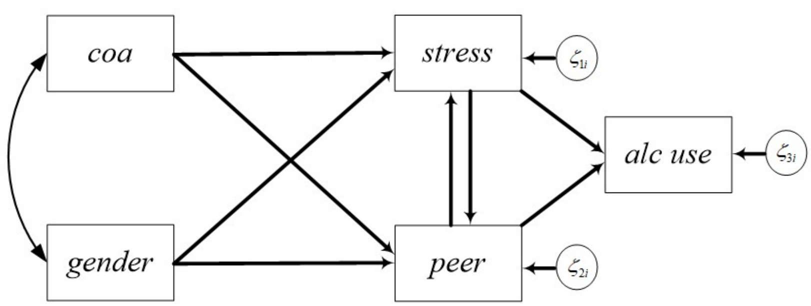


CenterStat 2.7

We again assume an implicit value of 1.0 is associated with each disturbance path.

Example: Non-Recursive

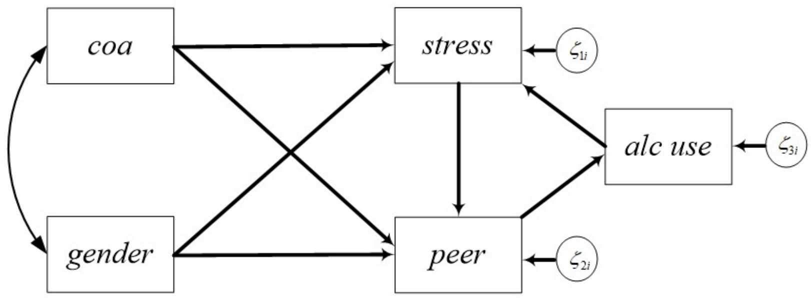
- A bi-directional effect between *stress* and *peer* make this model non-recursive



CenterStat 2.8

Example: Non-Recursive

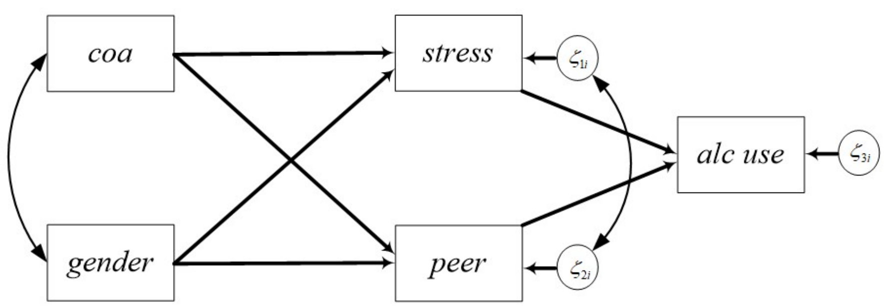
- Note the feedback loop among the three dependent variables



CenterStat 2.9

Example: Non-Recursive

- No bi-directional or feedback loops, but two of the residuals are allowed to covary



CenterStat 2.10

Modeling Steps in the SEM

1. Specification:
   - What is the model? Predictors? Outcomes? Mediators?
2. Identification:
   - Is it possible to obtain unique estimates for all model parameters?
3. Estimation:
   - How do we obtain the "best" estimates of model parameters?
4. Evaluation:
   - How well does the model fit the data?
5. Potential re-specification:
   - Should I modify the model?
6. Interpretation:
   - Which effects are significant? Which are substantively meaningful?

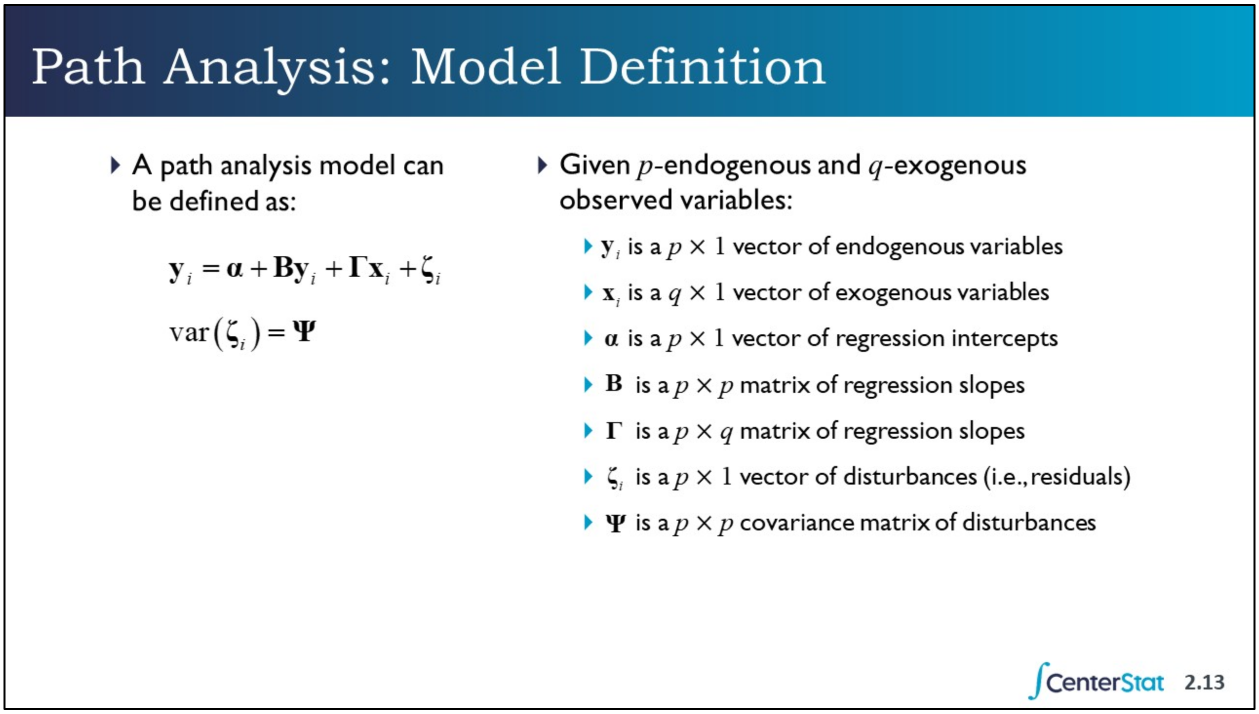
CenterStat 2.11

These are the same steps that we described for the multiple regression model, with the exception that interpretation may involve computing and making inferences for both direct and indirect effects of predictors.

Path Analysis: Model Specification

- Multiple regression has only one endogenous variable that is regressed on all exogenous variables
  - no restrictions are placed on the model, model is "saturated"
- The path analysis model is more complex in that there are multiple endogenous variables and one or more paths between exogenous and endogenous variables are typically fixed to zero
  - Thus, typically *not* saturated
- We must introduce some new matrices to account for these additional features of the path analysis model

CenterStat 2.12



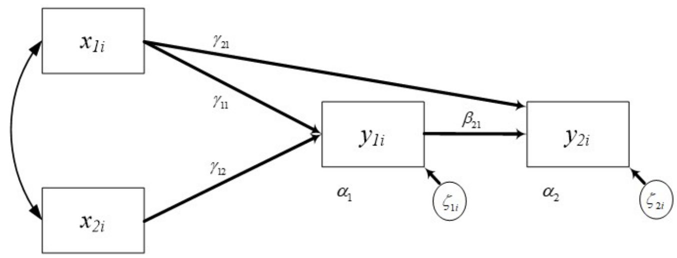
Recall from Chapter 1 that the regression model is

$$y_i = \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i \text{ where } VAR(\zeta_i) = \psi$$

Notice that the path analysis extends the scalar y from the regression model into a vector of endogenous variables $\mathbf{y}$. Additionally, because there are multiple endogenous variables, the effects of the exogenous predictors are now contained in a full matrix $\boldsymbol{\Gamma}$ (each row holds the regression coefficients for the set of exogenous variables predicting a given endogenous variable). Endogenous variables are also allowed to causally influence other endogenous variables through the new matrix $\mathbf{B}$. Finally, the disturbances of the endogenous variables have a covariance matrix $\boldsymbol{\Psi}$, as opposed to the single residual variance ψ of the regression model.

Path Analysis: Model Matrices

- Consider a simple path model with two exogenous variables ($q=2$) and two endogenous variables ($p=2$)
  - note that the path from x_2 to y_2 is fixed to zero



- Let's examine the individual elements in matrix expression for this model

Path Analysis: Model Matrices

- We will begin with the general matrix expression for the path analysis model:

$$\mathbf{y}_i = \boldsymbol{\alpha} + \mathbf{B}\mathbf{y}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i$$

- We then express the matrices to define our model:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & 0 \end{bmatrix} \begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} + \begin{bmatrix} \zeta_{1i} \\ \zeta_{2i} \end{bmatrix}$$

- And we finish with the final expressions for the model:

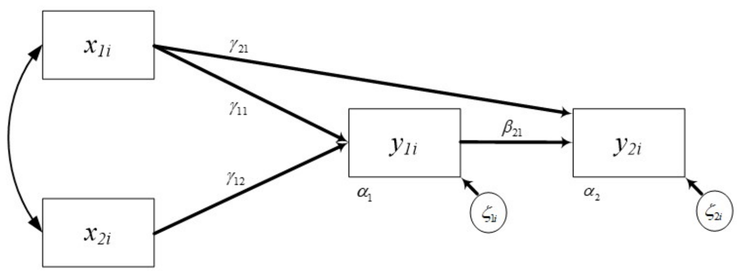
$$\begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} = \begin{bmatrix} \alpha_1 + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \zeta_{1i} \\ \alpha_2 + \beta_{21}y_{1i} + \gamma_{21}x_{1i} + \zeta_{2i} \end{bmatrix}$$

  - This is why this model is sometimes called "simultaneous equations"

Path Analysis: Model Matrices

- We can map the simultaneous equations back onto the path diagram

$$\begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} = \begin{bmatrix} \alpha_1 + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \zeta_{1i} \\ \alpha_2 + \beta_{21}y_{1i} + \gamma_{21}x_{1i} + \zeta_{2i} \end{bmatrix}$$



CenterStat 2.16

Summary

- The path analysis model is a generalization of multiple regression
  - Can include multiple endogenous variables
  - Can include causal chains and feedback loops
  - Can impose restrictions on one or more parameters and empirically evaluate the resulting impact, e.g., set a parameter to zero, or set two parameters to be equal
- To identify and estimate the path analysis model, we must first better understand the observed and model-implied mean and covariance structures

CenterStat 2.17

2.2 Mean and Covariance Structures

Objectives

- Define the population covariance matrices and mean vectors for the exogenous and endogenous variables
- Define the model-implied covariance matrices and mean vectors for the exogenous and endogenous variables

CenterStat 2.19

Mean and Covariance Structure

- In Chapter 1 we defined a model for one DV and q-IVs

$$y_i = \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i$$

- The population covariance matrix and mean vector were:

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|cccc} \sigma_y^2 & & & & \\ \hline \sigma_{x_1 y} & \sigma_{x_1}^2 & & & \\ \sigma_{x_2 y} & \vdots & \sigma_{x_2}^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \sigma_{x_q y} & \sigma_{x_q x_1} & \sigma_{x_q x_2} & \cdots & \sigma_{x_q}^2 \end{array}\right] \qquad \boldsymbol{\mu} = \left[\begin{array}{c} \mu_y \\ \hline \mu_{x_1} \\ \mu_{x_2} \\ \vdots \\ \mu_{x_q} \end{array}\right]$$

- We are going to expand this to include more than one DV and allow some DVs to predict other DVs

CenterStat 2.20

Mean and Covariance Structure

- Similar to regression model, path analysis model implies a specific structure for means, variances, and covariances of $\mathbf{x}$ and $\mathbf{y}$
  - reflects a theoretical hypothesis for how these variables relate to one another in the population
- Exploring the moment structure implied by the model is necessary to establish identification and to estimate the model
  - primary focus usually on covariances
  - often no structure placed on means, with exceptions of multiple group models and growth curve models

CenterStat 2.21

Covariance Matrices

- The joint covariance matrix for the endogenous and exogenous variables is denoted $VAR(\mathbf{y}_i, \mathbf{x}_i) = \mathbf{\Sigma}$ that is partitioned into four sub-matrices

$$\mathbf{\Sigma} = \left[\begin{array}{c|c} \mathbf{\Sigma_{yy}} & \mathbf{\Sigma_{yx}} \\ \hline \mathbf{\Sigma_{xy}} & \mathbf{\Sigma_{xx}} \end{array}\right]$$

  - $\mathbf{\Sigma_{yy}}$ is the covariance matrix of the endogenous variables (or DVs)
  - $\mathbf{\Sigma_{xx}}$ is the covariance matrix of the exogenous variables (or IVs)
  - $\mathbf{\Sigma_{yx}} = \mathbf{\Sigma'_{xy}}$ are the covariances of endogenous with exogenous variables
- Note direct parallel with regression model shown earlier
  - here we have multiple y's whereas in regression we only had one $VAR(\mathbf{y}_i, \mathbf{x}_i) = \mathbf{\Sigma}$

CenterStat 2.22

As we noted in Chapter 1, the lines denoted within the matrix are simply visual guides to demarcate the multiple endogenous variables and the multiple exogenous variables.

Joint Covariance Matrix

- To build up $\boldsymbol{\Sigma}$, we begin with the sub-matrix for the endogenous variables

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{yy}} & \\ \hline & \end{array}\right] = \left[\begin{array}{cccc|c} \sigma^2_{y_1} & & & & \\ \sigma_{y_2 y_1} & \sigma^2_{y_2} & & & \\ \vdots & \vdots & \ddots & & \\ \sigma_{y_p y_1} & \sigma_{y_p y_2} & \cdots & \sigma^2_{y_p} & \\ \hline & & & & \\ & & & & \end{array}\right]$$

CenterStat 2.23

Joint Covariance Matrix

- We then bring in the sub-matrix for the exogenous variables

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{yy}} & \\ \hline & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{array}\right] = \left[\begin{array}{cccc|cccc} \sigma^2_{y_1} & & & & & & & \\ \sigma_{y_2 y_1} & \sigma^2_{y_2} & & & & & & \\ \vdots & \vdots & \ddots & & & & & \\ \sigma_{y_p y_1} & \sigma_{y_p y_2} & \cdots & \sigma^2_{y_p} & & & & \\ \hline & & & & \sigma^2_{x_1} & & & \\ & & & & \sigma_{x_2 x_1} & \sigma^2_{x_2} & & \\ & & & & \vdots & \vdots & \ddots & \\ & & & & \sigma_{x_q x_1} & \sigma_{x_q x_2} & \cdots & \sigma^2_{x_q} \end{array}\right]$$

CenterStat 2.24

Joint Covariance Matrix

- Finally, we bring in the sub-matrix relating the exogenous and endogenous variables

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{yy}} & \\ \hline \boldsymbol{\Sigma}_{\mathbf{xy}} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{array}\right] = \left[\begin{array}{cccc|cccc} \sigma^2_{y_1} & & & & & & & \\ \sigma_{y_2 y_1} & \sigma^2_{y_2} & & & & & & \\ \vdots & \vdots & \ddots & & & & & \\ \sigma_{y_p y_1} & \sigma_{y_p y_2} & \cdots & \sigma^2_{y_p} & & & & \\ \hline \sigma_{x_1 y_1} & \sigma_{x_1 y_2} & \cdots & \sigma_{x_1 y_p} & \sigma^2_{x_1} & & & \\ \sigma_{x_2 y_1} & \sigma_{x_2 y_2} & \cdots & \sigma_{x_2 y_p} & \sigma_{x_2 x_1} & \sigma^2_{x_2} & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \\ \sigma_{x_q y_1} & \sigma_{x_q y_2} & \cdots & \sigma_{x_q y_p} & \sigma_{x_q x_1} & \sigma_{x_q x_2} & \cdots & \sigma^2_{x_q} \end{array}\right]$$

CenterStat 2.25

Mean Vectors

- We must similarly consider the means of the variables
- The "joint" mean vector for both endogenous and exogenous variables is denoted

$$E(\mathbf{y}_i, \mathbf{x}_i) = \boldsymbol{\mu}$$

- This vector can likewise be partitioned as

$$\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_{\mathbf{y}} \\ \hline \boldsymbol{\mu}_{\mathbf{x}} \end{array}\right]$$

  - $\boldsymbol{\mu}_{\mathbf{y}}$ is the mean vector for the endogenous variables
  - $\boldsymbol{\mu}_{\mathbf{x}}$ is the mean vector for the exogenous variables

CenterStat 2.26

Mean Vectors

- Expanding the elements of these vectors, we have

$$\boldsymbol{\mu}_{\mathbf{y}} = \begin{bmatrix} \mu_{y_1} \\ \mu_{y_2} \\ \vdots \\ \mu_{y_p} \end{bmatrix} \qquad \boldsymbol{\mu}_{\mathbf{x}} = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \vdots \\ \mu_{x_q} \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{y}} \\ \hline \boldsymbol{\mu}_{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \mu_{y_1} \\ \vdots \\ \mu_{y_p} \\ \hline \mu_{x_1} \\ \vdots \\ \mu_{x_q} \end{bmatrix}$$

- As before, this is just like the regression model in which the single DV appears first and is followed by the IVs
  - the only difference here is that there are multiple DVs

CenterStat 2.27

Model-Implied Moment Structures

- We have defined the population covariance and mean vector:

$$\boldsymbol{\Sigma} \qquad \boldsymbol{\mu}$$

- As we did with the regression model, we must now define the mean vector and covariance matrix implied by the path analysis

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) \qquad \boldsymbol{\mu}(\boldsymbol{\theta})$$

- Keep in mind we are still working at the level of the *population*

CenterStat 2.28

Model-Implied Covariance Structure

- Earlier we saw that the joint covariance matrix as

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{yy} & \\ \hline \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{array}\right]$$

- We can now define the model-implied covariance matrix as

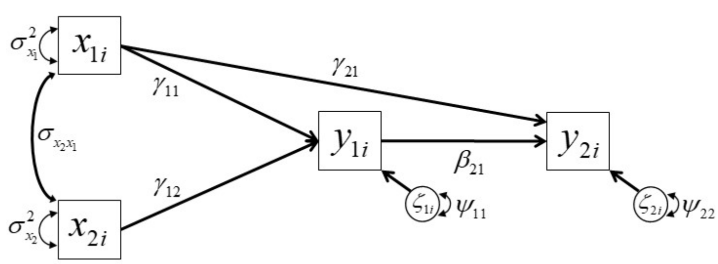
$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{yy}(\boldsymbol{\theta}) & \\ \hline \boldsymbol{\Sigma}_{xy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{xx}(\boldsymbol{\theta}) \end{array}\right]$$

- We can build each partition using matrix algebra (which is elegant), but this is much more clear using path tracing rules

CenterStat 2.29

Model-Implied Covariance Structure

- We'll start with a rather simple path analysis model



- We're imposing one structural restriction on the model
  - we are not allowing x_2 to directly predict y_2

CenterStat 2.30

Model-Implied Covariance Structure

- This model consists of four variables
  - two are endogenous and two are exogenous
- The four-by-four model-implied covariance matrix is

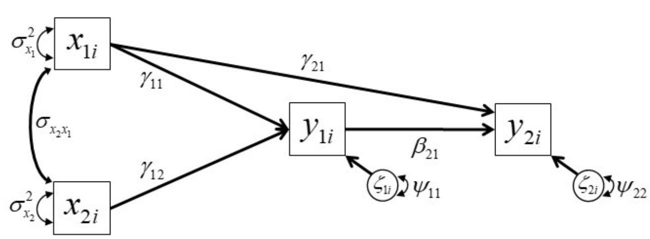
$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{yy}}(\boldsymbol{\theta}) & \\ \hline \boldsymbol{\Sigma}_{\mathbf{xy}}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{\mathbf{xx}}(\boldsymbol{\theta}) \end{array}\right] = \left[\begin{array}{cc|cc} \sigma^2_{y_1}(\boldsymbol{\theta}) & & & \\ \sigma_{y_2 y_1}(\boldsymbol{\theta}) & \sigma^2_{y_2}(\boldsymbol{\theta}) & & \\ \hline \sigma_{x_1 y_1}(\boldsymbol{\theta}) & \sigma_{x_1 y_2}(\boldsymbol{\theta}) & \sigma^2_{x_1}(\boldsymbol{\theta}) & \\ \sigma_{x_2 y_1}(\boldsymbol{\theta}) & \sigma_{x_2 y_2}(\boldsymbol{\theta}) & \sigma_{x_2 x_1}(\boldsymbol{\theta}) & \sigma^2_{x_2}(\boldsymbol{\theta}) \end{array}\right]$$

- We can use tracing rules to build every cell of the matrix

CenterStat 2.31

Model-Implied Covariance Structure

- We will first consider the covariance between y_2 and x_1

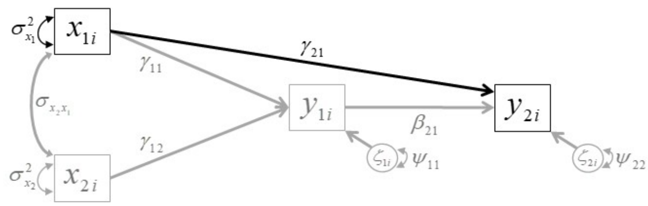


- There are three possible traces

CenterStat 2.32

Model-Implied Covariance Structure

- The first trace involves the direct path between y_2 and x_1



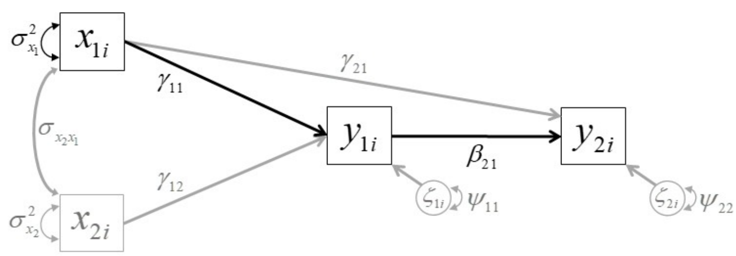
- This results in $\gamma_{21}\sigma^2_{x_1}$

CenterStat 2.33

We again use shading to highlight each segment in the trace.

Model-Implied Covariance Structure

- The next involves the path between y_2 and x_1 via y_1

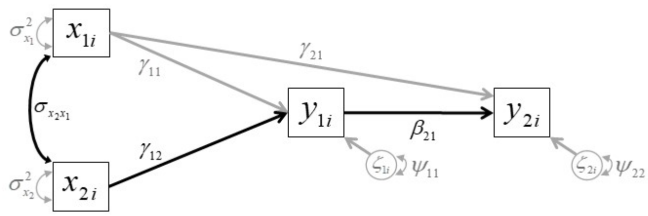


- This results in $\beta_{21}\gamma_{11}\sigma^2_{x_1}$

CenterStat 2.34

Model-Implied Covariance Structure

- The third involves the path between y_2 and x_1 via y_1 and x_2

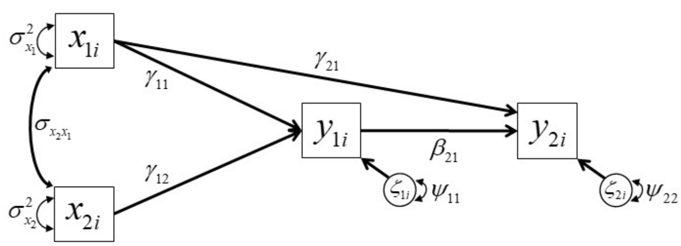


- This results in $\beta_{21}\gamma_{12}\sigma_{x_2x_1}$
- As before, we must sum all three to provide the model-implied value of the covariance between y_2 and x_1

CenterStat 2.35

Model-Implied Covariance Structure

- Putting the three traces together results in



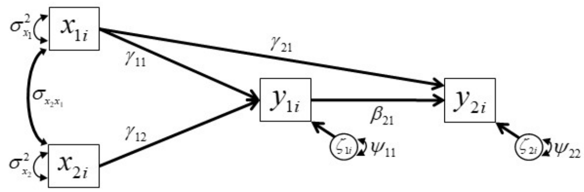
$$\sigma_{x_1y_2}(\boldsymbol{\theta}) = \gamma_{21}\sigma^2_{x_1} + \beta_{21}\gamma_{11}\sigma^2_{x_1} + \beta_{21}\gamma_{12}\sigma_{x_2x_1}$$

- This is the element in the 3rd row and 2nd column of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$

CenterStat 2.36

Model-Implied Covariance Structure

- Next, let's consider the covariance between y_2 and x_2

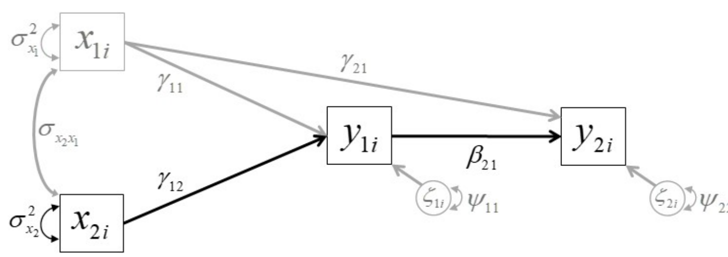


- This covariance must be **entirely** captured through the *indirect effects* given there is no direction relation between y_2 and x_2
  - this is an important part of how we will empirically test the structure of the model relative to the characteristics of the observed data

CenterStat 2.37

Model-Implied Covariance Structure

- The first trace is

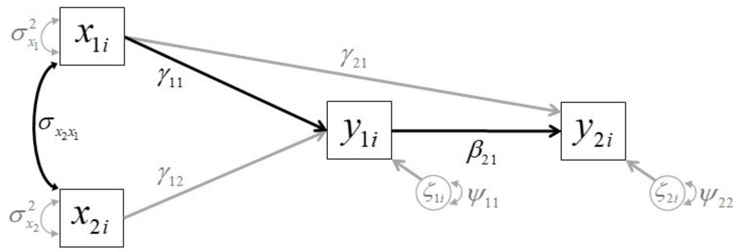


- This results in $\beta_{21}\gamma_{12}\sigma^2_{x_2}$

CenterStat 2.38

Model-Implied Covariance Structure

- The second trace is

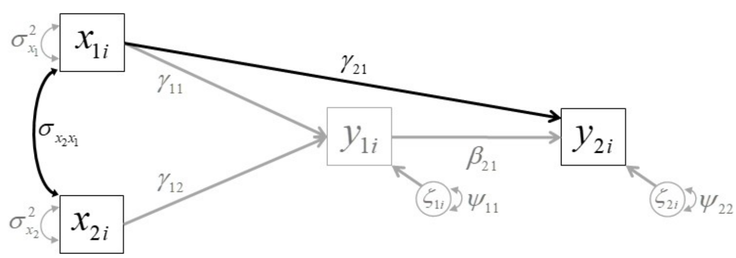


- This results in $\beta_{21}\gamma_{11}\sigma_{x_2x_1}$

CenterStat 2.39

Model-Implied Covariance Structure

- The final trace is

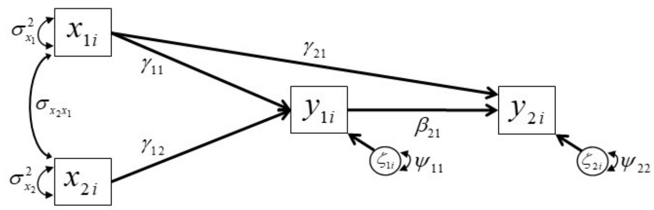


- This results in $\gamma_{21}\sigma_{x_2x_1}$
- We again sum the three traces to provide the model-implied covariance between y_2 and x_2

CenterStat 2.40

Model-Implied Covariance Structure

- The model-implied covariance between y_2 and x_2 is



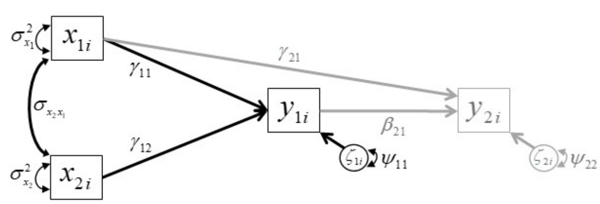
$$\sigma_{x_2y_2}(\theta) = \beta_{21}\gamma_{12}\sigma^2_{x_2} + \beta_{21}\gamma_{11}\sigma_{x_2x_1} + \gamma_{21}\sigma_{x_2x_1}$$

- This is the element in the 4th row and 2nd column of $\mathbf{\Sigma}(\boldsymbol{\theta})$

CenterStat 2.41

Model-Implied Covariance Structure

- Finally, we can consider the variance of y_1



- Interestingly, this is exactly the same as for the regression model

$$\sigma^2_{y_1}(\boldsymbol{\theta}) = \gamma^2_{11}\sigma^2_{x_1} + \gamma^2_{12}\sigma^2_{x_2} + 2\gamma_{11}\gamma_{12}\sigma_{x_2x_1} + \psi_{11}$$

- This is the element in the first row and first column of $\mathbf{\Sigma}(\boldsymbol{\theta})$

CenterStat 2.42

Model-Implied Covariance Structure

- We can do this for all of the elements of the matrix

$$\boxed{\sigma_{y_1}^2(\boldsymbol{\theta})} = \gamma_{11}^2\sigma_{x_1}^2 + \gamma_{12}^2\sigma_{x_2}^2 + 2\gamma_{11}\gamma_{12}\sigma_{x_2x_1} + \psi_{11}$$

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \left[\begin{array}{cc|cc} \boxed{\sigma_{y_1}^2(\boldsymbol{\theta})} & & & \\ \sigma_{y_2y_1}(\boldsymbol{\theta}) & \sigma_{y_2}^2(\boldsymbol{\theta}) & & \\ \hline \sigma_{x_1y_1}(\boldsymbol{\theta}) & \boxed{\sigma_{x_1y_2}(\boldsymbol{\theta})} & \sigma_{x_1}^2(\boldsymbol{\theta}) & \\ \sigma_{x_2y_1}(\boldsymbol{\theta}) & \boxed{\sigma_{x_2y_2}(\boldsymbol{\theta})} & \sigma_{x_2x_1}(\boldsymbol{\theta}) & \sigma_{x_2}^2(\boldsymbol{\theta}) \end{array}\right]$$

$$\boxed{\sigma_{x_1y_2}(\boldsymbol{\theta})} = \gamma_{21}\sigma_{x_1}^2 + \beta_{21}\gamma_{11}\sigma_{x_1}^2 + \beta_{21}\gamma_{12}\sigma_{x_2x_1}$$

$$\boxed{\sigma_{x_2y_2}(\theta)} = \beta_{21}\gamma_{12}\sigma_{x_2}^2 + \beta_{21}\gamma_{11}\sigma_{x_2x_1} + \gamma_{21}\sigma_{x_2x_1}$$

CenterStat 2.43

Model-Implied Covariance Structure

- However, this path tracing process rapidly becomes very complicated in even modest-sized models
  - not uncommon for a single element to contain a dozen terms
- Fortunately, these element-wise values can be compactly and elegantly captured in a small number of matrix expressions
  - see Bollen (1989, pages 85-88) for a crystal-clear treatment of this
- Mechanics of how this is done not necessary for using in practice, but good to understand general concepts when moving to identification, estimation, and evaluation
- We can briefly be impressed with the elegance of these two matrix expressions...

CenterStat 2.44

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Model-Implied Covariance Structure

- The full matrix expressions for the model-implied covariance and mean structures for the general path model are:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \left[\begin{array}{c|c} (\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1'} & \\ \hline \boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B})^{-1'} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{array}\right]$$

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \left[\begin{array}{c} (\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\mu}_{\mathbf{x}}) \\ \hline \boldsymbol{\mu}_{\mathbf{x}} \end{array}\right]$$

- No matter how complex the model, these expressions compute all possible path traces to result in the model-implied matrices
  - this is entirely done behind-the-scenes, but it's pretty neat to see

CenterStat 2.45

See Bollen (1989, pages 80-88) for an incredibly clear derivation of these expressions.

Summary

- Path model implies a specific structure for means, variances, and covariances of both endogenous and exogenous variables
- Determining the model-implied covariance matrix and mean vector enables us to both identify and estimate the model
- As we already know, goal of estimation will be to obtain parameter estimates that minimize the difference between the moment structure observed in sample and that implied by model

CenterStat 2.46

2.3 Model Identification and Estimation

Objectives

- Discuss model identification for path analysis models
- Describe identification rules
- Define maximum likelihood estimator for path analysis model
- Give sufficient statistic ML estimator
- Describe other estimators for path analysis model

CenterStat 2.48

Model Identification

- Recall from Chapter 1 that an identified model is one for which there is enough available information to obtain a unique estimate for each parameter
  - A model can be **under**-identified, **just**-identified, or **over**-identified
- Formally, identification is established by showing that the parameters of the model are unique functions of the population moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
  - Demonstrated this for regression model in Chapter 1; requires use of matrix algebra or path tracing rules
- Algebraic approach is tedious and error-prone for more complex models, so tend to rely on identification rules instead
  - some rules are sufficient but not necessary; others are necessary but not sufficient

CenterStat 2.49

Model Identification: The t-rule

- The t-rule simply states that there must be more means, variances, and covariances than there are parameters to be estimated
- For p-endogenous and q-exogenous variables, the unique number of means, variances and covariances is

$$k = \frac{(p+q)(p+q+1)}{2} + (p+q)$$

- For t-unique elements (i.e., free parameters) in $\boldsymbol{\theta}$, the t-rule states that the value of t must be less than or equal to k.
- This condition is necessary but <u>not</u> sufficient

CenterStat 2.50

We will also see later that the difference between k and t represents the degrees of freedom for the chi-square test of overall model fit.

Model Identification: The Null B Rule

- Recall the equation for the path analysis model is

$$\mathbf{y}_i = \boldsymbol{\alpha} + \mathbf{B}\mathbf{y}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i \qquad \text{var}(\boldsymbol{\zeta}_i) = \boldsymbol{\Psi}$$

  - earlier we noted that when $\mathbf{B} = \mathbf{0}$ there are no relations among the DVs
- The Null B rule states that when $\mathbf{B} = \mathbf{0}$ (i.e., the "null B" or "zero beta"), this is *sufficient* to establish model identification
- This even holds when one or more disturbances are correlated
  - at extreme, all possible disturbances can be correlated and model is still identified as long as no regression coefficients among any endogenous variables
  - indeed, under these conditions the model is simply a multivariate regression and is thus just identified

CenterStat 2.51

Model Identification: The Recursive Rule

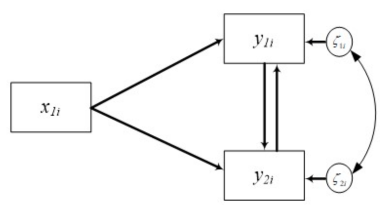
- For the recursive rule to apply, two conditions must hold
1. matrix of regression coefficients for the endogenous variables (i.e., $\mathbf{B}$) can be arranged to be lower triangular
   1. i.e., there are no feedback loops or bi-directional effects
2. covariance matrix of residuals (i.e., $\boldsymbol{\Psi}$) must be diagonal
   1. i.e., there are no correlated residuals
- Recall that these two conditions define a model to be recursive
- In terms of path diagrams, all single-headed arrows emanate from exogenous variables and span from left to right and there are no curved arrows between disturbances
- This is a sufficient (but not necessary) condition

CenterStat 2.52

There are additional strategies available for establishing the identification of more complicated model structures. One important example is the *rank and order* condition. We do not detail this method here, but see Bollen (1989, pages 98-103) for a clear discussion of this method.

The Common Sense Rule

- Although less formal (and admittedly made up by the two of us), under-identification can often be avoided by using common sense
- Just as matter can neither be created nor destroyed, we can only infer structures based on sufficient information
- For example, this model implies three forms of relation between the endogenous variables, yet only one covariance is observed



- This model is not algebraically identified
  - But the commonsense rule would suggest this would be the case

CenterStat 2.53

Model Estimation

- Once we have determined that the model is identified, we next need to obtain estimates for the parameters
- Thus far we have discussed model parameters, mean vectors, and covariance matrices strictly at level of population
- We never have access to this information, and must instead estimate values based on sample data
- Goal is to obtain the best estimates for model parameters given observed data

CenterStat 2.54

Sample Vector of Model Parameters

- Recall parameters that define a particular model are in vector $\boldsymbol{\theta}$
  - this is "the model"
- For example, population elements of $\boldsymbol{\theta}$ might be

$$\boldsymbol{\theta}' = \begin{bmatrix} \alpha_1 & \alpha_2 & \gamma_{11} & \gamma_{21} & \beta_{21} & \beta_{12} & \psi_{11} & \psi_{22} \end{bmatrix}$$

- Sample estimates of this same vector are then denoted

$$\hat{\boldsymbol{\theta}}' = \begin{bmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 & \hat{\gamma}_{11} & \hat{\gamma}_{21} & \hat{\beta}_{21} & \hat{\beta}_{12} & \hat{\psi}_{11} & \hat{\psi}_{22} \end{bmatrix}$$

- We can obtain these sample estimates using maximum likelihood
  - precisely as we discussed in Chapter 1

CenterStat 2.55

ML Estimation

- We can choose from two approaches to ML estimation
  - if DVs are continuously and normally distributed with no missing data, we can use **sufficient statistics** ML
  - if DVs are continuously distributed but are potentially non-normally distributed or have missing data, we can use **full information** ML
  - SSML requires covariance matrix & mean vector; FIML requires raw data
- As before, both approaches are identical when no missing data and DVs are continuously and normally distributed

∫CenterStat 2.56

See Appendix for details on ML estimation.

ML Estimation

- Given generality of FIML, we typically use this whenever working with raw data files
  - FIML default estimator in most SEM software packages
  - but must use SSML if only have covariance matrix and mean vector as unit of analysis
- Regardless of approach, ML proceeds iteratively to obtain sample estimates of $\boldsymbol{\theta}$ that minimize the differences between the observed and model implied moment structures

$$\min\left(\mathbf{S}-\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)\right) \qquad \min\left(\mathbf{m}-\boldsymbol{\mu}\left(\hat{\boldsymbol{\theta}}\right)\right)$$

∫CenterStat 2.57

ML Estimation

- Other minimization procedures represent alternative estimators
  - e.g., unweighted, generalized, or weighted least squares
- But ML has many desirable properties for many types of SEM
  - but in Chapter 7 we move to alternative estimators for discrete DVs
- There remain the important steps of *evaluation* and *respecification*, but first we return to our model of deviant peer affiliation

CenterStat 2.58

Summary

- Typically, in more complex SEMs, model identification is not achieved using covariance algebra
- Instead, general rules can be applied to derive necessary and sufficient conditions to establish model identification
  - don't apply to all SEMs, but to many
- Following identification, we again use SSML or FIML to estimate unknown model parameters
  - precisely same as described in Chapter 1

CenterStat 2.59

2.4 Demonstration: Deviant Peer Affiliations

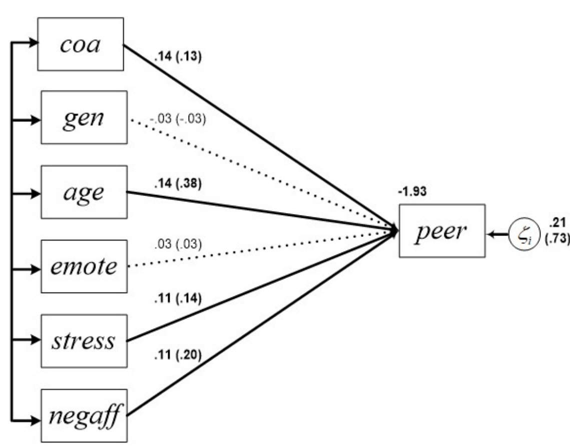
Objectives

- Return to our deviant peer example
- Expand our final regression model to a path model
- Examine resulting coefficients

CenterStat 2.61

Example: Deviant Peer Affiliations

- In Chapter 1, we tested a model that regressed one endogenous variable on six exogenous variables with direct effects



CenterStat 2.62

But Regression Inconsistent With Theory

- But theory dictates that a more complex etiological model holds
- More specifically, the *stress and negative affect* hypothesis predicts:
  - parental alcoholism increases uncontrollable life stressful events
  - increased stress in turn increases negative affect
  - increased negative affect in turn increases deviant peer affiliations
- Further, the *biological propensity* hypothesis predicts:
  - the biological basis underlying alcoholism increases negative emotionality
  - increased emotionality in turn increases negative affect
  - negative affect in turn influences deviant peer affiliations

CenterStat 2.63

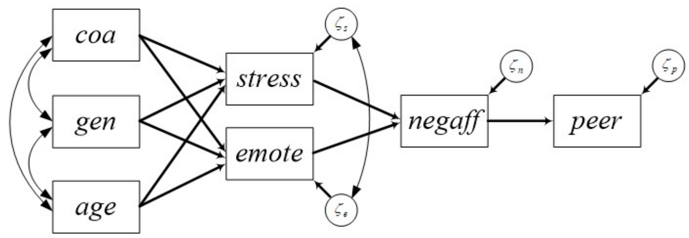
Direct vs. Indirect Effects

- Theory thus posits two *indirect* effects of alcoholism on peer
  - alcoholism to stress to negative affect to deviant peer affiliations
  - alcoholism to emotionality to negative affect to deviant peer affiliations
- Yet the regression model only estimates direct effects
  - there is thus a disjoint between the theoretical and statistical models

CenterStat 2.64

A More Theoretically Consistent Model

- We will begin by expressing a statistical model that more closely corresponds to the theoretical model



- Note that these are *precisely* the same variables from the prior model, but here we are imposing a radically different structure

CenterStat 2.65

There are several key changes from the regression model:

First, the path analysis model is defined by three exogenous variables (age, gender, and coa) and four endogenous variables (stress, emotionality, negative affect, and deviant peer affiliations)

Second, in the regression model all six predictors had a direct influence on peer affiliations, but now only negative affect does

Third, it is predicted that parental alcoholism will influence peer affiliations, but this effect will be fully mediated via stress, emotionality, and negative affect

Finally, whereas the regression model tested all possible paths between the exogenous and endogenous variables, here we have imposed restrictions that can be empirically tested, e.g., eight potential direct paths have been fixed to zero.

Model Identification

- Recall that an identified model is one for which there is enough available information to obtain a unique estimate for each parameter
  - A model can be **under**-identified, **just**-identified, or **over**-identified
- Formally, identification is established by showing that the parameters of the model are unique functions of the population moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
  - Demonstrated this for regression model in Chapter 1; requires use of matrix algebra or path tracing rules
- Algebraic approach is tedious and error-prone for more complex models, so tend to rely on identification rules instead
  - some rules are sufficient but not necessary; others are necessary but not sufficient

CenterStat 2.66

Identification of Hypothesized Path Model

- We must first consider whether the model is identified
  - the necessary but not sufficient condition of the t-rule is met
- That is, there are $k = 35$ pieces of observed information:

$$k = \frac{(p+q)(p+q+1)}{2} + (p+q) = \frac{(4+3)(4+3+1)}{2} + (4+3) = 35$$

- There are thus $t = 27$ freely estimated parameters
  - 7 variances; 7 means; 4 covariances; 9 regression coefficients
  - $t < k$ and the t-rule is met
- So, we know we at least have enough observed information to estimate the parameters of our hypothesized model

CenterStat 2.67

Identification of Hypothesized Path Model

- Interestingly, we do not meet sufficient conditions to establish identification using our other rules
  - we can't apply null $\mathbf{B}$ rule because we have endogenous effects
  - we can't apply recursive rule because we have correlated disturbances
- We partially meet the recursive rule given the matrix of coefficients for endogenous variables is lower triangular

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 & 0 \\ 0 & 0 & \beta_{43} & 0 \end{bmatrix} \begin{bmatrix} stress \\ emote \\ negaff \\ peer \end{bmatrix}$$

- But we don't fully meet this rule because we estimated the covariance between the residuals of stress and emotion

CenterStat 2.68

Identification of Hypothesized Path Model

- However, we can use one final rule we have not discussed to establish identification for this particular model: *Rank and order conditions*
  - given their complexity we did not present these conditions, but see citations earlier in chapter
- Applying these rank and order conditions here, we can establish that this model is identified
- The model is also identified by the "common sense rule"
  - specifically, stress and emotionality are not allowed to relate to one another without correlating the residual variance
  - this is reasonable and theoretically sound

CenterStat 2.69

See Bollen (1989) pages 98-104 for a detailed description of rank and order conditions for identification.

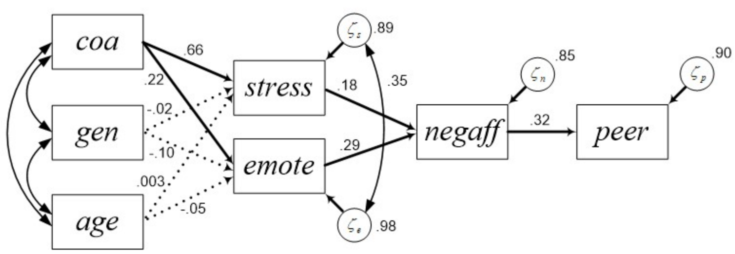
Estimation

- Given that we have established model identification, we can now proceed to model estimation
- We use raw-data ML here
  - because there are no missing data, direct or raw-data ML and sufficient statistic-based ML provide identical results
- We also use standard software defaults for number of iterations, convergence criterion, and start values
- No problems were encountered with model estimation and we can proceed to examining the ML parameter estimates and standard errors

CenterStat 2.70

Estimated Parameters for Path Model

- The path diagram below shows standardized coefficients
  - the effects from the 3 exogenous variables are "partially" standardized where the endogenous variable is standardized but the exogenous is not
  - the effects from the 3 endogenous variables are fully standardized
- Solid lines are $p<.05$ and dashed lines are $p>.05$



CenterStat 2.71

Model-Implied Moments

- We have not yet discussed how to evaluate the "fit" of the hypothesized model to the observed data
  - we address this in detail in the next chapter
- However, to anticipate this topic, we can tie back to our discussion of path tracing rules to examine the covariance matrix that was observed in the sample, $\mathbf{S}$, and the covariance matrix that is implied by the model, $\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$
- This comparison highlights the extent to which our hypothesized structural model is accurately reproducing the moment structure observed in the sample
  - we do not show mean structures because these are saturated and are thus perfectly reproduced

CenterStat 2.72

Model-Implied Moments: Raw Metric

- The observed (top) and model-implied (bottom) matrices are:

$\mathbf{S}$

	STRESS	EMOTION	NEGAFF	PEER	COA	AGE	GEN
STRESS	0.463						
EMOTION	0.125	0.252					
NEGAFF	0.183	0.170	0.917				
PEER	0.088	0.036	0.162	0.288			
COA	0.112	0.029	0.039	0.044	0.249		
AGE	-0.019	-0.059	0.210	0.307	-0.055	2.095	
GEN	-0.003	-0.010	-0.048	-0.022	0.002	-0.070	0.249

$\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$

	STRESS	EMOTION	NEGAFF	PEER	COA	AGE	GEN
STRESS	0.463						
EMOTION	0.125	0.252					
NEGAFF	0.183	0.170	0.917				
PEER	0.032	0.030	0.162	0.288			
COA	0.112	0.029	0.044	0.008	0.249		
AGE	-0.019	-0.059	-0.037	-0.007	-0.055	2.095	
GEN	-0.003	-0.010	-0.006	-0.001	0.002	-0.070	0.249

- discrepancies are highlighted in boxed cells

CenterStat 2.73

The model-implied covariance matrix represents the elements of $\Sigma(\hat{\boldsymbol{\theta}})$ that would be obtained using path tracing rules (or matrix algebra) based on sample estimates of all parameters. Fortunately, any off-the-shelf SEM package will provide these values as part of the model output.

Model-Implied Moments: Raw Metric

- It is simple to compute the differences between these matrices $\mathbf{S}-\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$

	STRESS	EMOTION	NEGAFF	PEER	COA	AGE	GEN
STRESS	0.000						
EMOTION	0.000	0.000					
NEGAFF	0.000	0.000	0.000				
PEER	0.055	0.006	0.000	0.000			
COA	0.000	0.000	-0.004	0.036	0.000		
AGE	0.000	0.000	0.247	0.314	0.000	0.000	
GEN	0.000	0.000	-0.042	-0.021	0.000	0.000	0.000

- Each cell is difference (or "residual") between corresponding elements of the observed and model-implied covariance matrices
  - but discrepancies are in raw metric of the covariances
  - can standardize these to see relative magnitude

CenterStat 2.74

Model-Implied Moments: Standardized

- The standardized discrepancies between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ are

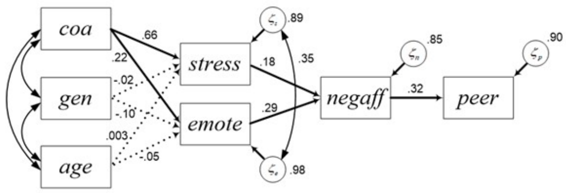
	STRESS	EMOTION	NEGAFF	PEER	COA	AGE	GEN
STRESS	0.000						
EMOTION	0.000	0.000					
NEGAFF	0.000	0.000	0.000				
PEER	**2.626**	0.394	0.000	0.000			
COA	0.000	0.000	-0.159	**2.375**	0.000		
AGE	0.000	0.000	**3.130**	**6.681**	0.000	0.000	
GEN	0.000	0.000	-1.541	-1.383	0.000	0.000	0.000

- Values exceeding ±2.0 are taken to indicate a large discrepancy between observed and model-implied covariances
  - the model is having trouble reproducing these values given the structure
- We can compare these discrepancies to our path model

CenterStat 2.75

There are two types of rescaled residuals generally available: *standardized* and *normalized*. The values in this slide are actually computed as *normalized* residuals, although we refer to these using the more typical terminology of standardized residual. See Bollen (1989) pages 257-262 for complete details.

Largest Standardized Residuals



- Largest standardized residual is between ***age*** and ***peer*** (6.68)
  - suggests effect of age not fully mediated by intervening variables
- Other large values are between *age* and *negaff* (3.13), *coa* and *peer* (2.35), and *peer* and *stress* (2.63)
- These residuals imply covariances not well reproduced by structure of model
  - we examine how this impacts model fit in the next chapter

CenterStat 2.76

Summary

- Examination of sample parameter estimates indicated these were consistent with theory
  - alcoholism predicted stress which predicted negative affect which predicted deviant peer affiliations
  - alcoholism predicted emotionality which predicted negative affect which predicted deviant peer affiliations
- However, we cannot interpret these results with confidence given that we have not yet formally evaluated the overall fit of the model
- In the next chapter we will consider ways that we can evaluate overall model fit and examine potential sources of model misfit

CenterStat 2.77

Chapter Summary

- Path analysis extends regression to multiple endogenous variables and allows for restrictions to be imposed on model structure
- Path models must be shown to be identified prior to estimation
  - only then can we consider methods for estimating model parameters from our sample data
- Maximum likelihood is the most widely used estimator in SEM
  - raw-data or direct ML is based on the raw data and allows for partially missing data
  - sufficient statistic ML is based on the sample means and covariances; yields equivalent MLEs under complete data
- But we must next consider how to carefully and thoughtfully evaluate the goodness-of-fit of our hypothesized model to our observed sample data

CenterStat 2.78

Chapter 3
Path Analysis: Part II

3.1 Assessing Model Fit

Objectives

- Define the likelihood ratio chi-square test
- Describe the need for additional indices of fit
- Define relative goodness-of-fit indices
- Define absolute goodness-of-fit indices

CenterStat 3.3

Assessing Model Fit

- There is not one clear and universally accepted strategy for assessing the adequacy of fit of SEMs
- Model evaluation something of an art
- No single number that will reflect the fit or misfit of a model
- Further, no "golden rules" about cut-offs for fit indices that will show a good fitting model
- We will build a more holistic understanding about all aspects of model fit and make a fully informed decision about the general adequacy of the hypothesized model

CenterStat 3.4

How best to establish model fit has been a controversial topic in the SEM literature for decades.

The Chi-Square Test Statistic

- The most ubiquitous measure of fit is the likelihood ratio (or "chi-square") test statistic
- Test statistic evaluates the fundamental null hypothesis of the SEM that

$$H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta}), \ \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$$

or, equivalently

$$H_0 : \mathbf{\Sigma} - \mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{0}, \ \boldsymbol{\mu} - \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{0}$$

- In words: the model-implied means, variances, and covariances are equal to the population means, variances, and covariances

CenterStat 3.5

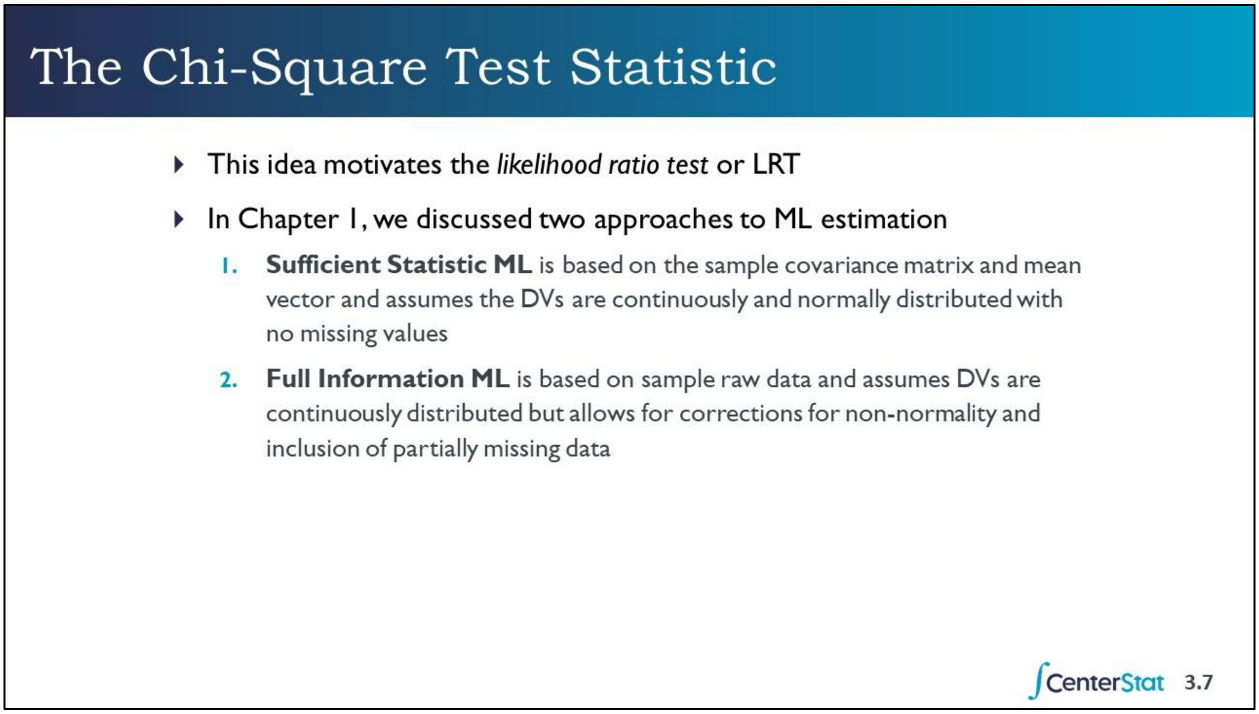
The Chi-Square Test Statistic

- The null hypothesis:

$$H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta}), \ \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$$

- Test null by comparing likelihood values obtained under hypothesized model versus a saturated model
  - Saturated model imposes no restrictions on means or covariances
  - Maximum Likelihood estimates are simply $\mathbf{m}$ and $\mathbf{S}$
- If null true, imposing hypothesized moment structure should not significantly reduce likelihood of observing sample data relative to the saturated model

CenterStat 3.6



To compute the likelihood ratio chi-square test we must literally fit two models, the hypothesized model and the saturated model. This can be done using either SSML or FIML. Years back we had to do this by hand (don't we sound like a couple of old men?), but now all software packages estimate these two models automatically and compute the chi-square test for us. However, it is important to realize that two separate models are being estimated. As such, it sometimes happens that the hypothesized model converges but the saturated model does not (e.g., when there is complete missing data for some cells of the covariance matrix). When this occurs, you can usually "turn off" estimation of the saturated model. You will then at least get estimates for the hypothesized model, but you will not obtain a chi-square test of overall model fit.

The Chi-Square Test Statistic

- LRT is computed somewhat differently for SSML & FIML, but both result in a sample statistic denoted T that reflects fit of hypothesized model relative to saturated model
- In large samples when all assumptions are met, T is distributed as a central chi-square with $df = k - t$
  - k is total number of moments and t is number of estimated parameters
  - This is often called the "model chi-square"

CenterStat 3.8

Inference with the Chi-Square Test of Fit

- For both SSML and FIML, T-statistic evaluates joint null hypothesis

$$H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta}),\ \ \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$$

- A significant T-statistic (e.g., p<.05) indicates it is <u>improbable</u> that test statistic of this value or larger would be found if null true
  - We reject the null, and thus reject the hypothesized model
- A non-significant T-statistic (e.g., p>.05) indicates it is <u>probable</u> that test statistic of this value or larger would be found if null true
  - We fail to reject the null, and thus fail to reject the hypothesized model

CenterStat 3.9

Inference with the Chi-Square Test of Fit

- But here's a terribly weird thing about the SEM
- In nearly all of statistics, rejecting null is taken as evidence of support for hypothesized model
  - The null hypothesis is a statement of no effect, and the **presence** of refuting evidence is required to imply that the model is correct
- But in SEM, accepting null is taken as evidence of support for hypothesized model
  - The null hypothesis is a statement of no effect (e.g., the model is correct) and the **absence** of refuting evidence implies the model is correct
- This is just one of many concerns that have been raised about using the T-statistic to evaluate the fit of SEMs

CenterStat 3.10

In most testing situations we hope to reject the null hypothesis because our "preferred" model is represented in the alternative hypothesis. However, here our "preferred" model is represented in the null hypothesis, the rejection of which indicates the preferred model does *not* fit the sample data.

The Chi-Square Test Statistic

- Another concern about the T-statistic relates to statistical power
- By definition, *all* models are wrong
  - That's what makes them models
- Models that are even trivially incorrect can be rejected given a sufficiently large sample size
- Similarly, a poorly fitting model can be accepted because there is insufficient power to detect the misspecification
  - We are rewarded for having low power
- Because of these issues, numerous alternative indicators of fit have been proposed to augment (or even replace) the T-statistic

CenterStat 3.11

Relative Goodness-of-Fit Indices

- Relative goodness-of-fit indices represent the improvement in fit of the hypothesized model relative to some baseline model
- Typical baseline model includes a mean and variance for all measured variables, but assumes all covariances to be zero
  - Some software programs allow exogenous predictors to covary
- This baseline model is highly restricted
  - Literally states no variable is related to any other variable
- Fit of hypothesized model computed relative to baseline model
  - The improvement in fit moving from the baseline to the tested model

CenterStat 3.12

Some software packages define the baseline model as one in which all measured variables are uncorrelated while other packages define the baseline model as one in which the exogenous variables are allowed to correlate with one another, but all endogenous variables are uncorrelated. The resulting differences in fit indices from these different baseline models are often trivial.

Relative Goodness-of-Fit Indices

- To compute indices we need test statistic and *df* for both hypothesized and baseline models
- Denote T-statistic and df for hypothesized model as T_h and df_h
- Denote T-statistic and df for baseline model as T_b and df_b
- Can then define variety of relative goodness-of-fit indices based on these four pieces of information

CenterStat 3.13

Relative Goodness-of-Fit Indices

- The oldest relative fit index is the Tucker-Lewis Index

$$TLI = \frac{(T_b / df_b) - (T_h / df_h)}{(T_b / df_b) - 1}$$

- Another example is the Comparative Fit Index

$$CFI = \frac{(T_b - df_b) - (T_h - df_h)}{(T_b - df_b)}$$

- And a final example is the Incremental Fit Index

$$IFI = \frac{T_b - T_h}{T_b - df_h}$$

CenterStat 3.14

Relative Goodness-of-Fit Indices

- Most relative fit indices fall between 0 and 1.0, although several can exceed 1.0 due to sampling variability
- Sampling distributions of these fit indices are unknown
  - Thus, no formal statistical tests are possible
  - Subjective decisions are needed about what value reflects acceptable fit
- For many years, values over .90 were taken to indicate good model fit
  - has recently crept up to .95
- But no universally accepted cut-points exist
  - Value that denotes "acceptable" model fit can vary widely from paper to paper and journal to journal

CenterStat 3.15

There are methods proposed to "bootstrap" a sampling distribution for a fit index; see

Bollen, K.A., & Stine, R.A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research, 21*, 205-229.

Absolute Goodness-of-Fit Indices

- One key concern about the relative measures of fit is that the baseline model is absurdly restricted
- Absolute goodness-of-fit indices do not rely on an arbitrary baseline model
- The most widely used of these is the root mean squared error of approximation, or RMSEA
- Two advantages of the RMSEA are
  - It does not rely on an arbitrary baseline independence model
  - The sampling distribution is known so that confidence intervals can be constructed and inferential tests conducted

CenterStat 3.16

The RMSEA

- The RMSEA is quite simple to compute

$$RMSEA = \sqrt{\frac{T_h - df_h}{df_h(N-1)}}$$

- RMSEA has lower bound of 0 and theoretically no upper bound
  - Sample values falling below zero are fixed at zero
- Confidence intervals can be constructed, but cut-points persist
  - Values less than .05 reflect excellent fit; less than .08 moderate fit; and values over .10 poor fit
- However, these cut-points are no better informed than those used for the relative fit indices

CenterStat 3.17

Standardized Root Mean Square Residual

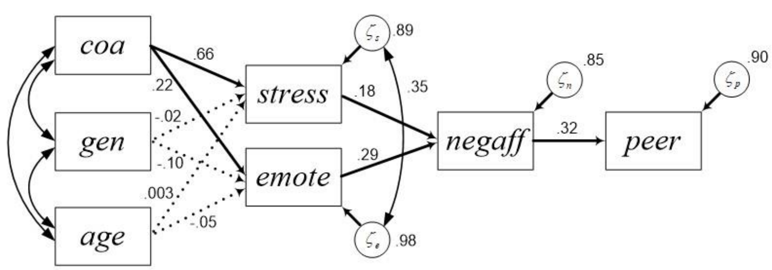
- Another absolute fit index is standardized root mean square residual
- SRMR expresses the difference between the observed and model-implied covariance matrices in a standardized metric
- SRMR values less than .08 often regarded as good fit, but there's no empirical evidence to support this value
- SRMR is not penalized for the number of model parameters, so tends to be better for less restricted models
- Whereas the RMSEA has a known sampling distribution the SRMR is more a descriptive measure

CenterStat 3.18

Remember that we first encountered elementwise standardized residuals in the deviant peer affiliation demonstration in Chapter 2. The SRMR is simply taking a mean of all of these standardized residuals.

Example: Deviant Peer Affiliations

- At the end of the last chapter, we estimated this path model and obtained these standardized estimates:



- However, we did not establish whether this model adequately fit the observed sample data -- we will do this now

CenterStat 3.19

Example: Deviant Peer Affiliations

- To begin, rarely if ever would interpret parameter estimates prior to establishing adequacy of model fit
- However, we did this here strictly for pedagogical purposes
  - We wanted to establish model estimation prior to model evaluation
- In practice, almost always first establish appropriate model fit prior to moving to parameter interpretation
  - We will model this strategy in later chapters
- Let's return to model from end of Chapter 2

CenterStat 3.20

Example: Deviant Peer Affiliations

- Key indices suggests that fit of model to sample data was very poor:
  - $\chi^2(8) = 81.17, p < .0001$
  - CFI=.67; TLI=.29; RMSEA=.17; SRMR=.09
- Indeed, every single index drastically falls short of even the most liberal criteria indicating appropriate model fit
- The implication is that model does extremely poor job of reproducing observed sample covariance matrix
  - In Chapter 2 we identified large residuals between the observed and implied covariance matrix, and this misfit is being identified here
- We next consider strategies for where to go in a situation like this

CenterStat 3.21

Summary

- Model evaluation is an extremely contentious issue
- There are many fit indices from which to choose
  - Some software packages offer several dozen options
- Best practice is to report the T-statistic, df, and p-value, one or two relative fit indices, the RSMEA, and the SRMR
  - All of these should jointly indicate appropriate model fit
  - If one type of index is aberrant from the rest, additional work is needed to fully understand why this might be
- Only by taking all possible information into account can a full sense of the adequacy of model fit be established

CenterStat 3.22

3.2 Model Comparisons

Objectives

- Discuss concept of model comparison
- Define nested models
- Define likelihood ratio test (a.k.a., "chi-square difference test")
- Comment on the comparison of non-nested models

CenterStat 3.24

Model Comparison

- Often challenging to derive an accurate *a priori* specification for a path model, particularly with multiple endogenous variables
- Want to balance theory, parsimony, and fit
  - Optimally reproduce the observed moments with as few parameters as possible while staying true to the underlying substantive theory
- Adding more parameters improves fit of model to data
  - But adding more parameters simultaneously decreases parsimony
  - At the extreme, estimating all possible parameters will perfectly reproduce the observed moment structure, just as it does in regression

CenterStat 3.25

Model Comparison

- What do we do, then, when our initial model does not adequately fit the observed data?
  - This brings us to model comparison
- When making model comparisons we seek to find model that best reproduces observed data
- We hope to find a good-fitting model, but this is not sole motivation for doing this
- We also strive for a *properly specified* model so we can obtain valid sample estimates of the parameters and standard errors

CenterStat 3.26

Model Comparison

- For example, if a path truly exists but is omitted from the model, this can significantly bias other estimated paths
  - Values of the other parameters change to adjust to the misspecification
- Model comparisons can be guided by theory or by sample data
  - First consider modifications with respect to theory
  - Then describe methods for letting sample data guide our decisions

CenterStat 3.27

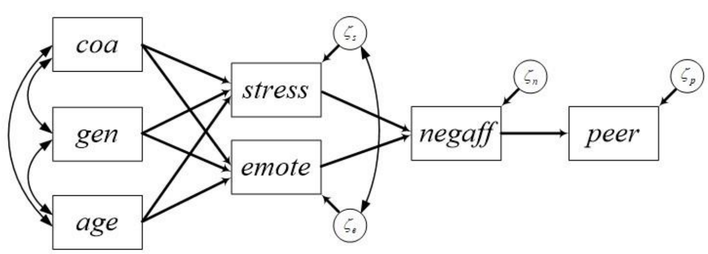
Nested Models

- Key motivating question is whether one model fits appreciably better than another
  - e.g., does the addition of three parameters significantly improve model fit
- When models are "nested" we can test whether one model fits data significantly better than the other
- Two models are nested if simpler (or more parsimonious) model can be derived by imposing constraints on more complex model
  - Constraints could be setting two or more parameters to be equal, or setting one or more parameters to be equal to zero
- Models are <u>not</u> nested if the simpler model has both restrictions *and* additions relative to the comparison model

CenterStat 3.28

Example: Deviant Peer Affiliations

- For example, consider our estimated model from Chapter 2



- Note that eight regression parameters are fixed to zero
  - *coa*, *gen*, & *age* to *negaff* and *peer*; and *stress* & *emote* to *peer*
  - (we are not considering reversed directionality here)

CenterStat 3.29

Example: Deviant Peer Affiliations

- Model fit is poor, but might hypothesize that too restrictive to require effect of *coa* on *negaff* and *peer* be entirely mediated by *stress* and *emote*
- Instead, might consider expanding model to include two additional effects: *coa* predicting *negaff*, and *coa* predicting *peer*
  - Revised model would be *more complex* than original model because it requires *more parameters*
- This raises question: *Is the addition of two regression parameters worth the added complexity?*
  - In other words, does the improvement in model fit exceed what we would expect by chance alone?

CenterStat 3.30

Nested Models

- To test differences in model fit must first establish models are nested
- Denote original model Model A and revised model Model B
- Model A is more restricted compared to Model B
  - e.g., Model A has two fewer parameters than does Model B
- For Model A to be nested within Model B, we must be able to impose constraints on Model B to define Model A
  - We can do this here: if the two additional regression parameters in Model B are fixed to zero then we obtain Model A
- Given that the models are nested, we can use a likelihood ratio test to test the difference in their fit to the data

CenterStat 3.31

The Likelihood Ratio Test for Nested Models

- Likelihood ratio chi-square test of overall model fit described earlier is actually a comparison between two nested models
  - Any hypothesized model is nested within the saturated model
  - Measures the relative decrement in fit brought about by imposing structure in the hypothesized model
- Can similarly use likelihood ratio test (LRT) to compare the difference in fit between *any* pair of nested models
- Sometimes called *chi-square difference test*, since simplest computation is to take difference in T statistics for two models

CenterStat 3.32

The Likelihood Ratio Test for Nested Models

- The fit of Models A and B is reflected in their T statistics

$$T_A \sim \chi^2\left(df_A = k - t_A\right) \qquad T_B \sim \chi^2\left(df_B = k - t_B\right)$$

- Difference in fit is $T_\Delta = T_A - T_B$
  - This is the likelihood ratio test statistic, or chi-square difference
- In large samples, under certain assumptions, T_Δ is distributed as a central χ^2 with degrees of freedom equal to $df_A - df_B$, or

$$T_\Delta \sim \chi^2\left(df_\Delta = df_A - df_B\right)$$

- The LRT tests the null that the two models fit equally well

$$H_0 : \boldsymbol{\Sigma}(\boldsymbol{\theta}_A) = \boldsymbol{\Sigma}(\boldsymbol{\theta}_B),\ \ \boldsymbol{\mu}\left(\boldsymbol{\theta}_A\right) = \boldsymbol{\mu}\left(\boldsymbol{\theta}_B\right)$$

CenterStat 3.33

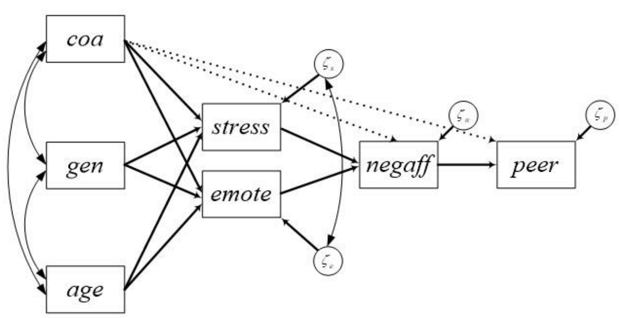
Inferences with The Likelihood Ratio Test

- There are two possible outcomes of the test
  1. LRT is *non-significant* and we fail to reject null hypothesis.
     - Implies that the additional parameters of Model B do <u>not</u> significantly improve the fit of the model and we should retain Model A
  2. LRT is *significant* and we reject null hypothesis.
     - Implies that the additional parameters in Model B <u>do</u> significantly improve the fit of the model and we should retain Model B
- Note that like any inferential test, this is affected by sample size
  - Excessively low power could miss an effect that is really there
  - Excessively high power could identify an effect that is trivial in size

CenterStat 3.34

Example: Deviant Peer Affiliations

- We will test the improvement in model fit with the addition of two direct effects: from *coa* to *negaff*, and from *coa* to *peer*
  - The original model ("A") is denoted with solid paths, and the two additional paths are shown with dashed paths ("B")



CenterStat 3.35

Example: Deviant Peer Affiliations

- The test statistic for the original model (Model A) is

$$T_A = 81.17, \ \left(df_A = 8\right)$$

- The test statistic for the modified model (Model B) is

$$T_B = 74.32, \ \left(df_B = 6\right)$$

- Although clearly an improvement in model fit, must formally test if improvement is greater than would be expected by chance

CenterStat 3.36

Example: Deviant Peer Affiliations

- The difference between the two test statistics is

$$T_\Delta = T_A - T_B = 81.17 - 74.32 = 6.85$$

- The p-value from a χ^2 distribution with $df_\Delta = 2$ is $p = .033$
- Because LRT less than $\alpha = .05$, we reject null and conclude that the two additional parameters do result in significant improvement in model fit
- However, given the large sample size, the improvement is quite modest and the overall model fit remains extremely poor
  - $\chi^2(6) = 74.32, p < .0001$
  - CFI=.71; TLI=.12; RMSEA=.19; SRMR=.08

CenterStat 3.37

Non-Nested Model Comparisons

- The comparison of non-nested models is more difficult because LRT is not available
- Cannot simply compare likelihood values of models, since a model with more parameters will tend to fit better
- Information criteria takes the model -2 log-likelihood and imposes a penalty related to number of model parameters to reward parsimony
  - e.g., Akaike information criterion (AIC) and Bayes information criterion (BIC)
- Usually scaled so that the best model is the one with the lowest value for the information criterion (e.g., minimum BIC)
- But most model comparisons in SEM are between nested models

CenterStat 3.38

Summary

- We can use a likelihood ratio test (chi-square difference test) to formally compare the fit of two nested models
  - The LRT cannot be used for models that are not nested
- The LRT can test a single parameter or multiple parameters
- A significant LRT implies that additional parameters improve model fit
- However, the LRT requires *a priori* specification of both models
  - e.g., We must determine exactly which additional parameters to include
- In complex models, may be difficult to determine where misfit lies
- We can turn to modification indices to address this issue

CenterStat 3.39

3.3 Model Respecification and Modification Indices

Objectives

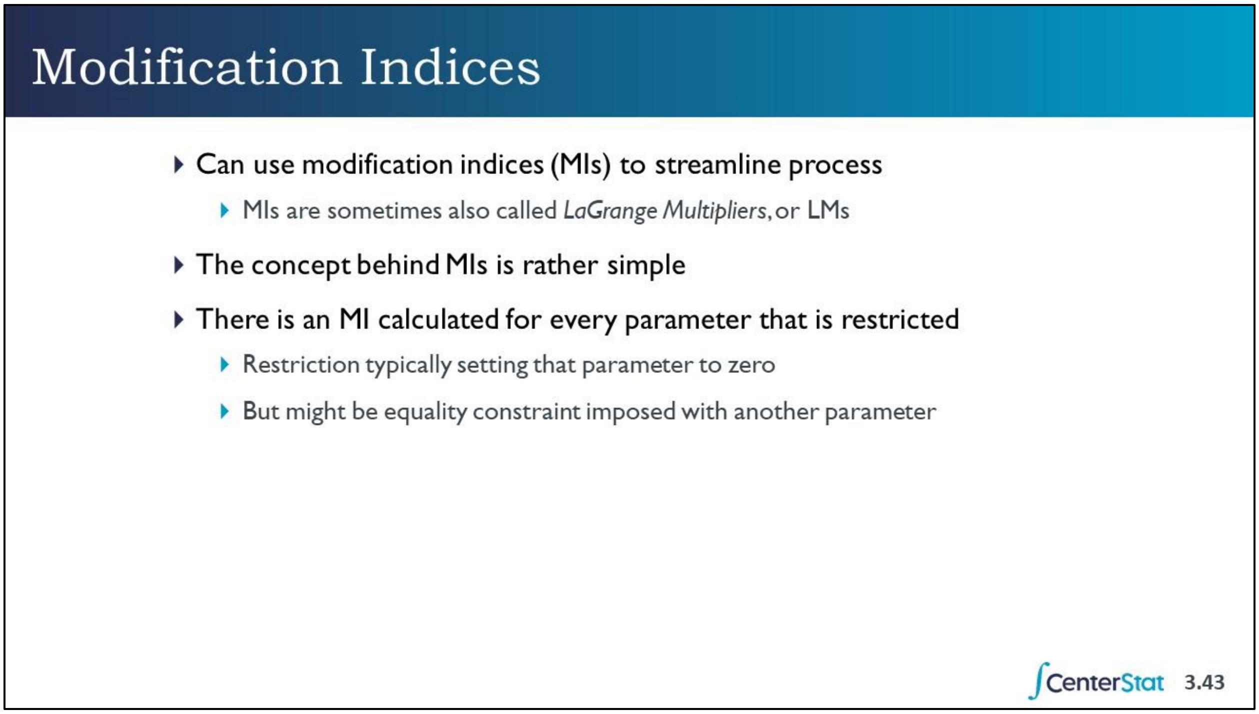
- Differentiate model comparison from model respecification
- Introduce modification indices
- Compare these to likelihood ratio tests
- Make recommendations for use in practice

CenterStat 3.41

Modification Indices

- A distinct advantage of model comparison using the LRT is that we can conduct *a priori* multi-parameter tests
  - e.g., Compare a model with and without direct effects
- This is a powerful tool to test specific hypotheses about the structure of our model
- But less useful when considering more general approach to testing model misspecification
  - e.g., Might be dozens of parameters fixed to zero and inefficient to use separate LRTs for every restriction

CenterStat 3.42



For those of you who are gluttons for punishment, the MI for parameter k is given as

$$MI_k = \left[\partial\ell\left(\hat{\theta}_r\right) / \partial\hat{\theta}_k\right]^2 \left[\mathbf{I}^{-1}\left(\hat{\theta}_r\right)\right]_{kk}$$

The MI value is the change in the likelihood ratio chi-square test of overall model fit that we would expect to observe by removing the restriction on parameter k (where r is the original model). Note that the matrix $\mathbf{I}$ in this expression represents the *Fisher information matrix* (and not the *identity matrix* as was first introduced in Chapter 1). Perhaps (only slightly) more intuitively, the inverse of the Fisher information matrix is the estimated sampling variance-covariance matrix for the parameters, or $\mathbf{I}^{-1}\left(\hat{\theta}_r\right) = V\hat{A}R\left(\hat{\theta}_r\right)$, which quantifies uncertainty in the parameter estimates. The standard errors for the model estimates are in fact plucked from the diagonal of $V\hat{A}R\left(\hat{\theta}_r\right)$ Thus the first bracketed term captures expected change in the log-likelihood with the removal of the restriction whereas the second bracketed term accounts for the sampling variability in this parameter. Together they provide an estimate of the expected change in the overall fit of the model. For further details, see pp. 293-295 of

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

The important thing to see here is that only the restricted model is estimated, as reflected in the term $\ell\left(\hat{\theta}_r\right)$. This means that we can get an MI estimate of the improvement in fit associated with removing any one restriction in our model without ever estimating a second model (in contrast to an actual likelihood ratio test, which would require estimating both the model with and without the restriction).

Modification Indices

- MI indicates expected change in model fit (or LRT value) that would result if the restriction on that parameter was removed
  - If restriction is parameter equal to zero, MI is the expected change in model fit that would result if that parameter were freely estimated
- It is "expected" change because MIs are based on model derivatives and are not *actual* likelihood ratio tests
  - But typically very close estimates

CenterStat 3.44

Modification Indices

- Each MI is a test statistic distributed with $df = 1$
- The critical value defining the upper 5% of the central chi-square distribution for $df = 1$ is 3.84
  - Thus, any MI exceeding 3.84 is "significant" at $p < .05$
- However, it is often recommended that values greater than 3.84 be used to indicate significant MIs
  - First, we are conducting multiple tests (oftentimes a very high number of tests), so some informal alpha-correction is useful
  - Second, it protects against excessive power and identifying statistically significant yet substantively meaningless effects

CenterStat 3.45

Modification Indices

- As with everything else, there is no golden rule for cut-offs
- Many programs provide an expected parameter change estimate (EPC)
- MI gives expected change in model LRT with freeing of constraint, EPC gives expected value that parameter would obtain
  - e.g., EPC for a regression coefficient is the estimate of the value that parameter would obtain if freed
- EPC estimate is useful in considering both improvement in model fit and the resulting value of the newly added parameter

CenterStat 3.46

Modification Indices

- Instead of using a pre-defined cut-point to identify meaningful MIs we recommend examining relative size for a given analysis
  - e.g., Are there one or two large MIs, and the rest are small
  - Or are there many large MIs that only gradually get smaller and smaller
- Once the most important MI is identified, only this one parameter should be freed and the model re-estimated
  - Reason is that all MIs will change given the inclusion of the additional parameter, and these must be re-estimated for the new model
  - Further, one misspecification may be reflected in multiple MIs, so freeing a single parameter can make other MIs go away

CenterStat 3.47

Modification Indices

- Finally, only release parameters based on MIs if this is truly consistent with theory
- MIs are sample size dependent, so larger samples will reflect larger MIs even though the resulting parameters are trivial
- In sum, MIs can provide valuable evidence about potential sources of misspecification, but these must be used with great care, thoughtfulness, and with close respect to theory

CenterStat 3.48

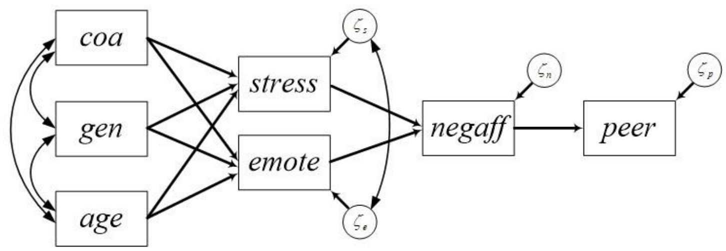
Modification Indices

- However, MIs are not without their cost or danger
- MIs are entirely data driven
  - Nothing more than first-order partial derivatives of fit function that are devoid of theory
- MIs are taken just one parameter at a time
  - Means that a given misspecification in model might appear in multiple places via MIs, but not known where "true" misspecification resides

CenterStat 3.49

Example: Deviant Peer Affiliations

- We used the LRT to test the improvement in fit of the model when we included two direct effects from *coa*
  - improvement was significant, but overall model still fit poorly
- We will return to the original model (without the direct effects from *coa*) and consult MIs for suggestions for respecification



CenterStat 3.50

Model-Implied Moments: Standardized

- Recall from Chapter 2 that the standardized residuals between the observed and model-implied covariance matrices were:

	STRESS	EMOTION	NEGAFF	PEER	COA	AGE	GEN
STRESS	0.000						
EMOTION	0.000	0.000					
NEGAFF	0.000	0.000	0.000				
PEER	**2.626**	0.394	0.000	0.000			
COA	0.000	0.000	-0.159	**2.375**	0.000		
AGE	0.000	0.000	**3.130**	**6.681**	0.000	0.000	
GEN	0.000	0.000	-1.541	-1.383	0.000	0.000	0.000

- We already know that largest MIs will likely be associated with the largest covariance residuals

CenterStat 3.51

Example: Deviant Peer Affiliations

- Recall that the model fit was very poor:
  - $\chi^2(8)=81.17, p<.0001$
  - RMSEA=.17
  - CFI=.67
  - SRMR=.09
  - TLI=.29
- From this model we can get a complete listing of MIs
  - Here we semi-arbitrarily restrict MIs to be equal to 10 or more

CenterStat 3.52

Note that these are exactly as expected based on the magnitude of the covariance residuals.

Example: Deviant Peer Affiliations

negaff on *age*:	MI=11.89	EPC=.119
peer on *age*:	MI=42.54	EPC=.129
negaff with *age*:	MI=10.88	EPC=.236
peer with *age*:	MI=44.54	EPC=.275

- Note overlap (two are regressions and two are covariances)
  - Clearly chasing the same misspecification
- Also remember that in Chapter 2 our two largest standardized residuals were between *age* and *peer*, and *age* and *negaff*
  - Highlights exactly what MIs are doing here
- Most reasonable first step is to free regression of *peer* on *age*
  - Should reduce model LRT by approximately 43, which is highly significant

CenterStat 3.53

In most software packages the EPCs are given in the raw metric of the observed variables; however, it is possible to obtain standardized EPCs as well.

Example: Deviant Peer Affiliations

- We re-estimate the model including the regression of *peer* on *age*
  - $\chi^2(7) = 34.37, p < .0001$
  - CFI=.88
  - TLI=.70
  - RMSEA=.11
  - SRMR=.06
- Drastic improvement in model fit, but model still does not fit well

CenterStat 3.54

Example: Deviant Peer Affiliations

- MI predicted model LRT would be 81.17-42.54=38.63 and the actual value was 34.37
  - MI was a good, but not exact, estimate of the LRT
- EPC predicted newly estimated parameter would be equal to .129, and actual value was .132
  - Again, EPC is close but not perfect
- We now want to compute MIs from this model to determine if any further changes to model necessary

CenterStat 3.55

Example: Deviant Peer Affiliations

- There were five MIs exceeding 10 in the modified model
  - *negaff* on *age*: MI=11.89 EPC=.119
  - *peer* on *stress* MI=13.38 EPC=.149
  - *peer* on *coa* MI=11.85 EPC=.185
  - *peer* with *coa* MI=11.85 EPC=.046
  - *negaff* with *age* MI=10.88 EPC=.236
- Can again see we're chasing around the same misspecification
- Considering options, conclude that *peer* on *coa* is most defensible
  - Reflects direct effect of COA above and beyond the two indirect effects
- We will next add this regression coefficient

CenterStat 3.56

Example: Deviant Peer Affiliations

- We re-estimate the model including the regression of *peer* on *coa*
  - $\chi^2(6) = 22.27, p < .0001$
  - CFI=.93
  - TLI=.79
  - RMSEA=.09
  - SRMR=.04
- The MI predicted the model LRT would be 34.37-11.85=22.52 and the obtained value was 22.27
  - But the model still only attains borderline fit
- The MIs for this model are
  - *negaff* on *age*: MI=11.89 EPC=.119
  - *negaff* with *age* MI=10.88 EPC=.236

CenterStat 3.57

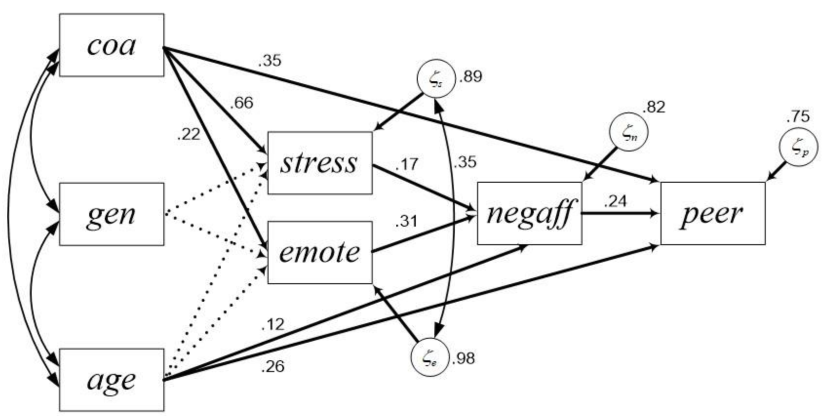
Example: Deviant Peer Affiliations

- We free the regression of *age* on *negaff* and the model fit is
  - $\chi^2(5)=10.16, p=.07$
  - CFI=.98
  - TLI=.92
  - RMSEA=.06
  - SRMR=.03
- Not only is this an appropriately fitting model, but there are no more MIs that are greater than 10
- We conclude that this is our final best fitting model

CenterStat 3.58

Example: Deviant Peer Affiliations

- Below is the final path diagram with standardized estimates



CenterStat 3.59

Example: Deviant Peer Affiliations

- We can draw the following substantive conclusions
  - COA leads to higher stress, emotionality, and deviant peer affiliations
  - Age leads to higher negative affect and higher deviant peer affiliations
  - Stress leads to greater negative affect
  - Emotionality leads to greater negative affect
  - Greater negative affect leads to higher deviant peer affiliations
- There is thus empirical support for our two indirect paths:
  - *coa* to *stress* to *negaff* to *peer*
  - *coa* to *emote* to *negaff* to *peer*
  - However, significant links in a mediational chain are not sufficient to infer mediation -- we must explicitly test this in our model

CenterStat 3.60

Summary

- LRT allows for test of *a priori* determined restrictions on the model
- MIs estimate improvement in model fit associated with freeing each constraint in model; EPCs estimate value that parameter will obtain
- MI of 3.84 is significant at $p = .05$, but higher values often used to account for multiple tests and elevated power
- One parameter is freed and then the MIs are recalculated
- MIs are strictly data driven – must be used with great caution
- Parameter should only be freed based on an MI if it is reasonable and fully justified by theory
  - And modifications based on MIs should be fully described to a reader

CenterStat 3.61

3.4 Testing Direct and Indirect Effects

Objectives

- Motivate why we desire tests of mediation
- Differentiate direct and indirect effects
- Describe how these are estimated
- Demonstrate their use in the peer data set

CenterStat 3.63

Mediation

- In Chapter 1 we noted the definition from Baron & Kenny (1986)
  - *"In general, a given variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the criterion." (p1176)*
- In other words, a mediator explains *why* or *how* one variable exerts an effect on another
- In example, attempting to better understand relation between parental alcoholism and deviant peer affiliations
- Initial regression model suggested that being child of an alcoholic increases probability of affiliation with deviant peers
  - But this does not offer insights as to *why* this effect might exist

CenterStat 3.64

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.

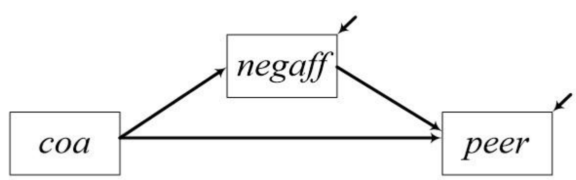
Mediation

- Expanded regression model to path model to consider two potential mediating mechanisms
  - *coa* to *stress* to *negaff* to *peer*
  - *coa* to *emotion* to *negaff* to *peer*
- Final model indicated significant links within each mediating chain
- However, well known that significant links are necessary but not sufficient to infer significant mediation

CenterStat 3.65

Mediation

- Must conduct additional tests to test extent to which each of our two pathways mediates the influence of *coa* to *peer*
- We begin with a simple one-mediator model
- Consider model that posits negative affect is sole mediator between *coa* and *peer*



CenterStat 3.66

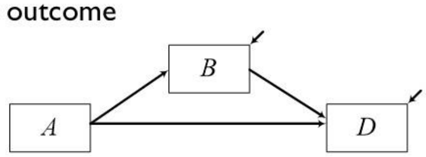
Mediation

- Might hypothesize that being child of an alcoholic increases negative affect that in turn increases peer deviance
  - In other words, negative affect mediates the relation between parental alcoholism and deviant peer affiliations
- Although a significant relation between *coa* and *negaff*, and between *negaff* and *peer* infers mediation, this is not sufficient
- We must formally test the entire mediating pathway

CenterStat 3.67

Direct and Indirect Effects

- As a starting point, consider a simple model with one predictor, one mediator, and one outcome

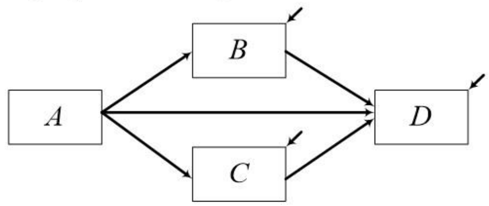


- A *direct effect* is the direct influence of one variable on another without any other intervening variables
  - Here there is one direct effect: A to D
- An *indirect effect* is the influence of one variable on another as mediated via one or more intervening variables
  - Here there is one indirect effect: A to B to D

CenterStat 3.68

Direct and Indirect Effects

- But consider a slightly more complex model:



- Here there remains just one direct effect (A to D), but now there are *two* indirect effects: A to B to D and A to C to D
- We must define and organize these various effects so that these may be tested in a formal way

CenterStat 3.69

Direct and Indirect Effects

- Define ***total effect*** as sum of all possible paths (direct and indirect) leading from predictor to outcome
  - In our example, this would be the one direct and two indirect paths
- Define ***total indirect effect*** as sum of all possible indirect paths leading from predictor to outcome
  - In our example, this would be the two indirect paths
- Define ***specific indirect effect*** as any single indirect path leading from predictor to outcome
  - In our example, each indirect path would be a specific indirect effect
- Can conduct formal tests of all of these effects, the specific selection of which will depend on the questions posed in a given substantive application

CenterStat 3.70

Direct and Indirect Effects

- The matrix equations involved in testing total, direct, indirect, and specific indirect are complex and we need not present these here
  - See Bollen (1987) and Bollen (1989) pp.376-389 for details
- However, although tedious in their details, the matrix manipulations are ultimately quite straightforward
  - We use matrix algebra to compute products of coefficients, sums of products of coefficients, and standard errors for each effect
- It is easy to gain a conceptual understanding of what we are attempting to do here by returning to our hypothetical mediational model

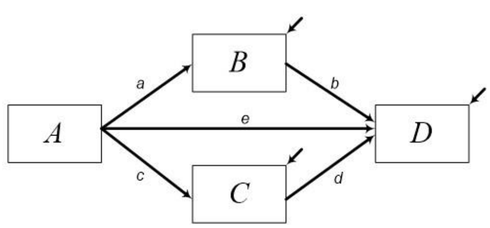
CenterStat 3.71

Bollen, K.A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology, 17*, 37-69.

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

Direct and Indirect Effects

- We can define our effects in terms of the relevant pathways



- The direct effect is simply e
- The first specific indirect effect is the product of a and b
- The second specific indirect effect is the product of c and d
- The total indirect effect is the sum of the two specific indirect effects
- The total effect is the sum of the total indirect effect & the direct effect
- Next, we need standard errors for all of these effects

CenterStat 3.72

Indirect Effects: Delta Method

- Simple to compute *point estimate* for indirect effect
  - Just a compound parameter that is product of coefficients in chain
- Much more complex to conduct *inferential tests* of effect
- Traditional approach is to use *delta method* (or *Sobel method*) to analytically derive appropriate standard errors (SEs)
  - See Bollen (1987) for a comprehensive review of this topic

CenterStat 3.73

An excellent discussion of different methods for testing indirect effects can be found in:

MacKinnon, D.P. (2008). *Introduction to Statistical Mediation Analysis*. Taylor & Francis: NY.

Indirect Effects: Delta Method

- But several problems with delta method
  1. It is only a first-order approximation and SEs can be inexact
  2. Assumes indirect effects have normal sampling distributions but this is typically not the case
  3. Basing inferential test on traditional critical ratio assumes symmetric sampling distribution that is also typically not the case
- Bootstrapping methodologies overcome these limitations

CenterStat 3.74

Indirect Effects: Bootstrap

- Central concept to understand in bootstrapping is *resampling*
  - Models are fit repeatedly to multiple random draws from original sample
  - Tests are based on the empirical distribution of the repeated samples
- Typical steps:
  1. Randomly draw bootstrap sample from original sample with replacement
  2. Fit model to bootstrap sample and retain indirect effect
  3. Repeat large number of times -- maybe 500 or 1000 times
  4. Create empirical distribution of all obtained estimates

CenterStat 3.75

Indirect Effects: Bootstrap

- Can use SD of bootstrapped estimates as SE for critical ratio, but not recommended because z-test assumes a symmetric distribution
  - Indirect effects involve products of estimates and tend to be asymmetrically distributed
- Best practice is to use bootstrap confidence intervals
  - Based on 2.5% and 97.5% values of bootstrapped distribution
  - Can be asymmetric

CenterStat 3.76

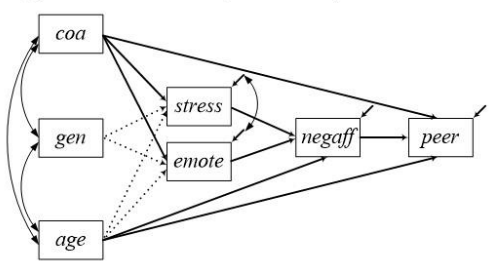
Indirect Effects: Bootstrap

- Advantages of bootstrap approach over delta method
  1. Does not make unrealistic assumption of normality for sampling distribution
  2. Allows for asymmetric CIs
  3. Is a very broad approach that can be applied in many different circumstances (e.g., fit indices, mixture models, etc.)
- But also potential disadvantages
  1. Computationally intensive fitting hundreds or thousands of individual models
  2. Each individually fitted model can encounter its own estimation problems
  3. In many cases, virtually no difference between bootstrap and delta tests
- Bootstrap CIs are considered the gold standard for testing indirect effects and should be used in practice when possible

CenterStat 3.77

Example: Deviant Peer Affiliations

- Let's return yet again to our final path analysis model

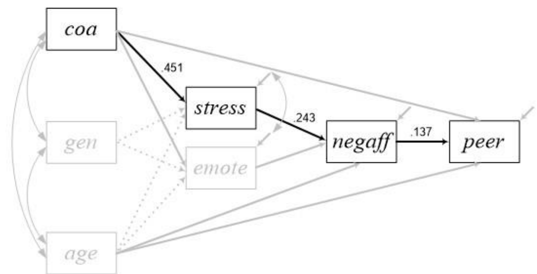


- We are interested in testing three components of mediation
  - The specific indirect effect of *coa* on *peer* via *stress* and *negaff*
  - The specific indirect effect of *coa* on *peer* via *emote* and *negaff*
  - The total effect of *coa* on *peer* as the sum of the direct & indirect effects

CenterStat 3.78

Example: Deviant Peer Affiliations

- We'll begin with the first specific indirect effect
  - Raw coefficients are shown on the paths

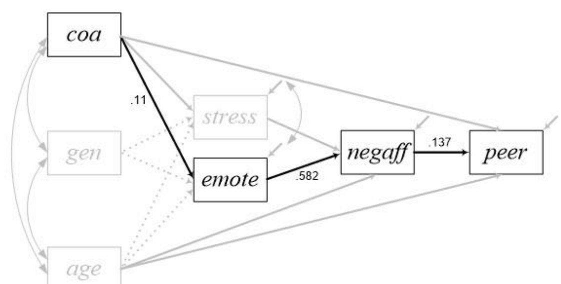


- The first **specific indirect effect** is (.451)(.243)(.137)=.015
  - 95% bootstrap CI is .003 and .033 and is significant (CI does not contain 0)

CenterStat 3.79

Example: Deviant Peer Affiliations

- Now we'll consider the second specific indirect effect

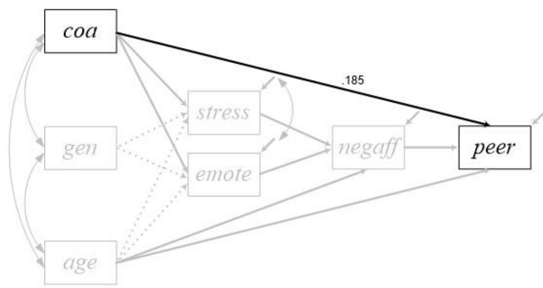


- The second **specific indirect effect** is (.11)(.582)(.137)=.009
  - 95% bootstrap CI is 0 and .21 and is *not* significant (CI contains 0)

CenterStat 3.80

Example: Deviant Peer Affiliations

- We next consider the single direct effect, net all other effects



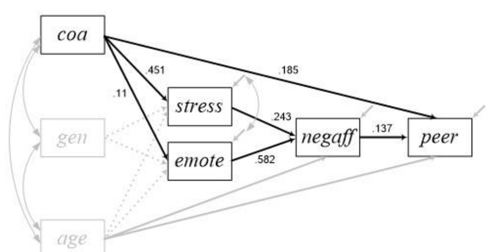
- The **direct effect** is .185
  - 95% bootstrap CI is .09 and .28 and is significant

CenterStat 3.81

We could have also used the usual ML standard error for the direct effect because it is not a compound parameter and can thus be reasonably assumed to follow a normal distribution. However, here we use the bootstrap CIs to remain consistent with the testing of the other effects.

Example: Deviant Peer Affiliations

- Finally, we can piece the various effects together



- The **total indirect effect** is (.015)+(.009)=.024
  - 95% bootstrap CI is .008 and .047 and is significant
- The **total effect** is (.015)+(.009)+(.185)=.209
  - 95% bootstrap CI is .107 and .303 and is significant

CenterStat 3.82

Example: Deviant Peer Affiliations

- Conclusions from these analyses are:
  1. Effect of parental alcoholism on deviant peer affiliation mediated by increases in stress and subsequent increases in negative affect
  2. No support for hypothesis that effect of parental alcoholism on deviant peer affiliations mediated by emotionality
- These tests of total, direct, and indirect effects provide a much more nuanced understanding of the relation between the exogenous and endogenous variables in the model

CenterStat 3.83

The implication here is that disrupting any of the nodes in the chain from parental alcoholism to stress to negative affect could potentially protect against increases in deviant peer affiliations. In contrast, there is no evidence that the greater emotionality of children of alcoholics plays any role in their increased affiliations with deviant peers.

Summary

- The path analysis model can assess direct or indirect effects
- Significant links in a mediating chain are not sufficient to infer mediation
  - The full indirect effect must be formally tested
- Effects can be broken down into direct, indirect, and total
  - Indirect effects can be further broken down to specific indirect effects
- Indirect effects estimated by multiplying raw coefficients in the chain
  - Confidence intervals obtained from bootstrap procedure
- Effect decomposition is a significant strength of the path model

CenterStat 3.84

Chapter Summary

- Prior to interpreting parameters, we must first establish the adequacy of overall model fit
- We can conduct formal model comparisons to test the improvement in fit of one model relative to another
- We can use MIs to identify specific parameters that could be added to improve model fit
- Once fit is established, we can test total, direct, total indirect, and specific indirect effects to examine mediation

CenterStat 3.85

3.5 Self-Study: Assumptions

Objectives

- Describe assumptions of maximum likelihood estimation
- Describe assumptions of the path analysis model itself
- Highlight assumption of perfect reliability as motivation to move to latent variable models

CenterStat 3.87

Assumptions

- Care must be taken when discussing assumptions of a "model" versus assumptions of an "estimator"
- That is, we can lay out a particular path analysis model, but then consider a host of alternative methods of estimation, each of which invokes different underlying assumptions for proper usage
- We will first focus on assumptions underlying the maximum likelihood estimator and then assumptions underlying the more general path analysis model itself

CenterStat 3.88

Assumptions

- There are four key assumptions underlying ML estimation that we will consider here:
  1. Sufficiently large sample size
  2. Independence of residuals
  3. Multivariate normality of residuals
  4. Missing data are missing at random
- And there are two additional key assumptions underlying the path model itself that we will consider here:
  1. Linearity among the observed measures
  2. Perfect reliability
- We'll explore each of these in turn

CenterStat 3.89

Sufficient Sample Size

- Maximum likelihood is derived under asymptotic conditions
  - It assumes infinite sample size, but we never have infinite sample sizes
- We must have a sufficiently large sample size for two reasons
  1. We need a large sample size to obtain a properly converged solution and stable sample estimates
     - Small sample sizes can lead to problems in optimization, non-positive definite matrices, etc.
  2. We need a large sample size to ensure that the test statistic follows the assumed probability distribution
     - Small sample size can result in inflated test statistics and increased Type I error

CenterStat 3.90

Sufficient Sample Size

- Yet how large is large enough is inherently unknowable, so we strive to obtain as large a sample as is reasonably possible
  - "Sufficient" sample size can depend on things such as quality of measurement and complexity of the model
  - Thus, no firm recommendations are possible
- Note that statistical power is *not* an assumption of the estimator
- That is, we need a sufficient sample size to detect a given effect with some degree of probability
  - We can estimate precisely what sample size is needed to detect a given effect size with a given probability, but this is not an *assumption*

CenterStat 3.91

Independence of Residuals

- As in regression, ML assumes residuals are independent
  - No two residuals any more or less similar than any other two
- However, in some designs this assumption might not hold
  - e.g., Hierarchical nesting of multiple children within a classroom or multiple time points nested within an individual
- When independence violated, both test statistics and standard errors can be biased, sometimes significantly so
  - It is also difficult (if not impossible) to disaggregate between- and within-group (or between- and within-person) effects
- If independence assumption is violated, additional analytic steps are necessary such as multilevel analysis or growth models

CenterStat 3.92

Normality of Residuals

- In regression model residuals assumed normally distributed, but no distributional assumptions made about exogenous variables
  - This is sometimes called *conditional normality* because it is the distribution of the dependent variable conditioned on the predictors
- Same distributional assumption holds for path model, but is expanded so that residuals are *multivariate* normally distributed
  - It is multivariate because there are multiple endogenous variables
- Motivation for assumption is that ML likelihood is derived under multivariate normality
  - That is, it only considers the first two distributional moments -- the mean vector and the covariance matrix

CenterStat 3.93

Normality of Residuals

- Often difficult to unambiguously determine when data are sufficiently non-normal to be of concern
- One challenge is that univariate normality is necessary but not sufficient to infer multivariate normality
  - Thus, examining univariate distributions is helpful, but is not definitive
- Another challenge is that, although there are numerical tests for univariate and multivariate normality, these are often over-powered and almost always identify non-normality
- Often, though not definitive, careful graphical analysis can be most useful when examining distributions of dependent variables
- There are two issues to consider when thinking about normality

CenterStat 3.94

Normality of Residuals

- First, we do *not* make any distributional assumptions about exogenous variables
  - These may be normal, skewed, kurtotic, binary, ordinal -- go nuts
- Second, if the residuals are continuous but non-normal, two specific problems can arise
  1. The likelihood ratio test statistic is too large, thus increasing the likelihood of falsely rejecting a correct null

     i.e., conclude the model is incorrect when it is not
  2. The standard errors of the parameters are too small, thus increasing the likelihood of falsely rejecting a correct null

     i.e., conclude there is an effect when there is not

CenterStat 3.95

Normality of Residuals

- To assess the conditional normality assumption we can ask:
  - Do the univariate distributions appear non-normal?
  - Do departures from normality appear severe?
  - Do bivariate scatter plots appear linear?
- Chapter 5 describes ways to account for non-normality in SEMs
- As a final note, the normality assumption is invoked for *continuously distributed* variables
  - Chapter 7 discusses methods for fitting SEMs to *discrete* variables

CenterStat 3.96

Missing Data are Missing at Random

- There are three types of missing data
  - Missing Completely At Random (MCAR): any missing case is strictly random and unrelated to any other variables in or out of the model
  - Missing At Random (MAR): any missing case can be related to variables in the model, but not to variables omitted from the model
  - Missing Not At Random (MNAR): any missing case is related to variables that are omitted from the model
- Raw-data ML assumes MAR (which subsumes MCAR)
- Much more complex approaches are needed for MNAR
- It is likely that most data structures in the behavioral sciences are MAR, so this is not an unreasonable assumption to invoke

CenterStat 3.97

Linearity

- Sufficient sample size, independent residuals, and normally distributed residuals are all assumptions of the ML estimator
- Linearity is an assumption of the path analysis model itself
- More specifically, we assume that all relations among observed measures are linear in form
- It is possible to introduce non-linear relations among predictors and an outcome but these still enter the model linearly
  - e.g., multiplicative interactions or quadratic effects
- Violation of this assumption leads to a misspecified model

CenterStat 3.98

Here we are primarily concerned with nonlinear relationships among continuous observed variables. Similar concerns could be expressed with respect to discrete observed variables (e.g., binary or ordinal indicators), whose relationships are intrinsically nonlinear. For discrete variables we must consider alternative approaches for specifying and estimating nonlinear SEMs.

Perfect Reliability

- A second assumption of the path model is perfect reliability
  - This assumption has significant implications for all of the models we've discussed thus far, and for where we will go from here
- More specifically, assume all observed variables measured without error
  - All observed variance is true variance to be modeled
- Within behavioral, health, & social sciences, this assumption is absurd
  - Most (if not all) constructs assessed with some degree of measurement error
- Yet both regression and path analysis models assume all observed variance to be true variability of the construct
- There are two rather insidious problems associated with this

CenterStat 3.99

Perfect Reliability

- First, it can be shown that an estimated regression coefficient is bounded as function of reliability of predictor
- For example, consider a one predictor sample regression model:

$$y_i = \hat{\alpha} + \hat{\gamma}_1 x_i + \hat{\zeta}_i$$

- Denote the population reliability of the exogenous variable as ρ
  - Reliability is the ratio of true score variance divided by the total variance
  - Thus, reliability can only equal one when there is no error variance
- It can be shown that: $E(\hat{\gamma}_1) = \gamma_1 \rho$
  - On average, sample estimate is limited to the product of the population coefficient and the reliability of the exogenous predictor

CenterStat 3.100

Perfect Reliability

- Sometimes called *attenuation due to unreliability*
  - Sample estimate of coefficient is smaller than corresponding population value in presence of unreliability
- There are three ways to avoid this attenuation
  1. Use perfectly reliable predictors, but this is unrealistic
  2. Correct for unreliability by fixing residual variance to some estimate in path model, but tends to perform poorly and is not recommended
  3. Construct multiple indicator latent factors to empirically estimate and remove error variance and thus disattenuate the coefficients, and this is what we'll do next

CenterStat 3.101

Perfect Reliability

- There is a second consequence of unreliability to consider
- Nearly all attention in literature focused on unreliability in *predictors*
- However, we also assume perfect reliability in the *endogenous* or *dependent* variables
  - even if these do not predict other dependent variables
- Unreliability in the endogenous variables does not bias parameter estimates, but does increase the standard errors and thus reduces statistical power

CenterStat 3.102

Perfect Reliability

- It is occasionally recommended that unreliability in endogenous variables is not problematic and can be ignored
  - This is not true, and we must carefully consider the role of unreliability in both exogenous and endogenous measured variables
- We will next explore a variety of latent variable models that directly address problems with variables measured with error

CenterStat 3.103

Summary

- Both ML and the path model invoke a number of assumptions necessary to support the validity of the approach
- The four key assumptions invoked by ML are sufficient sample size, independence of residuals, normality of residuals, and MAR
- The two key assumptions invoked by the path model are linearity and perfect reliability
- Different assumptions can be addressed to varying degrees
- Important to our work here, the assumption of perfect reliability is extremely unrealistic, and this can both drive down parameter estimates and drive up standard errors
  - We will next move to a fully latent variable model to address this issue

CenterStat 3.104

Chapter 4
Confirmatory Factor Analysis

4.1 Exploratory Factor Analysis

Objectives

- Discuss and define "latent variables" as they are used in everyday thinking, science, and statistics
- Briefly describe and demonstrate the exploratory factor analysis model
- Differentiate between exploratory and confirmatory factor analysis

CenterStat 4.3

Latent Variables

- In everyday life, latent variables are invoked all the time:
  - "I'm happy today"
  - "Bob is a very efficient worker"
  - "She's introverted"
- Speaking in terms of latent variables allows us to describe properties and relationships that generalize beyond more specific observable behaviors
  - "I've smiled a lot today"
  - "Bob's quarterly revenue generation is at the 93rd percentile"
  - "She rarely attends large social functions"

CenterStat 4.4

Latent Variables

- In science, latent variables are referred to by many names:
  - Unmeasured variables
  - Factors
  - Unobserved variables
  - Constructs
  - True scores
  - Unobserved heterogeneity
- Latent variables play key roles in many psychological, social science, and health science theories:
  - Self esteem
  - Socioeconomic status
  - Quality of life

CenterStat 4.5

Definitions

- Bollen (2002) reviews alternative formal definitions of latent variables
- Local independence definition
  - Latent variables explain associations among observed variables (e.g., "depression" explains association between mood and well being)
- Nondeterministic function definition
  - A latent variable cannot be expressed as a direct, deterministic function of the observed variables
- Sample realizations definition (most general)
  - A variable for which there is no sample realization for at least some observations in a given sample

CenterStat 4.6

Bollen, K.A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605-634.

Uses of Latent Variables in Statistics

- Latent variables have wide use in statistics, though they may not always be referred to as such
  - Factors in factor analysis
  - Latent variables in structural equation models
  - Basis curves in latent curve analysis
  - Latent classes (class/cluster/mixture models)
  - Latent traits (IRT)
  - Underlying propensity variables for censored/limited dependent variables
  - Missing data
  - Random effects in a multilevel or mixed model

CenterStat 4.7

Some statisticians look askance at latent variable models, without realizing that they in fact use models that include latent variables quite regularly.

Uses of Latent Variables in Statistics

- Latent variable models used to
  - Understand the structure of a set of observed measures and abstract general relationships from these measures
  - Obtain score estimates for the latent variables for other analyses
  - Directly model latent variables as predictors or outcomes
- The factor analysis model provides an ideal entry point into latent variable modeling, en route to the full SEM

CenterStat 4.8

Goals of Factor Analysis

- Determine the underlying structure behind a set of observed measures
  - May or may not be a theoretical model for this structure at the outset
- Data reduction
  - Conceptual – can many observed variables be thought of in terms of a few underlying traits?
  - Practical – can we reduce the variables under consideration from a high number of observed variables to a relatively low number of latent variables?
- Scale Development, Scoring & Assessment
  - Determine core set of items that best reflect underlying factors
  - Develop scale scores for use in subsequent models, or for decision making purposes

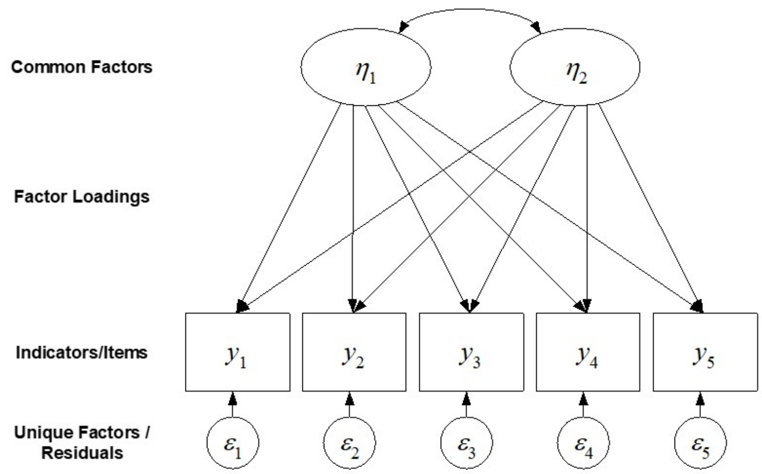
CenterStat 4.9

Exploratory and Confirmatory Models

- EFA and CFA share the same basic model for how latent variables relate to observed variables, but differ in emphasis
- EFA is primarily data-driven
  - Purpose is to empirically derive the number and nature of latent factors
  - All observed variables (indicators) regressed on all factors and the regression slopes (factor loadings) interpreted to determine what factors represent
- CFA is primarily theory-driven
  - Emphasis is on formally testing a theoretical factor structure
  - Number and nature of latent variables specified by theory
  - Relationships between indicators and latent variables specified by theory
- Useful to briefly review EFA before turning to CFA

CenterStat 4.10

Blueprint of an EFA Model



CenterStat 4.11

Note that all indicators/items are regressed on all factors and the unique factors are assumed to be uncorrelated.

Common Factors

- Common factors, or η, are common causes for two or more indicator variables, inducing the covariance/correlation between the indicators.
  - Depression is a common cause of mood and well being ratings
  - Mood and well being ratings are correlated because they are both caused by depression
- Sufficiently many common factors must be estimated to explain all of the covariance/correlation among the observed measures
  - If indicators are still correlated after accounting for a single common factor, there must be a second common factor that explains this remaining correlation, or a third, etc.

CenterStat 4.12

Common Factors

- Key question in EFA is how many common factors are needed to explain the associations among the observed measures

Pearson Correlation Coefficients

	y1	y2	y3	y4	y5	y6	y7
y1	1.00						
y2	0.55	1.00					
y3	0.41	0.46	1.00				
y4	0.27	0.31	0.19	1.00			
y5	0.15	0.21	0.11	0.49	1.00		
y6	0.25	0.26	0.15	0.62	0.44	1.00	
y7	0.26	0.27	0.18	0.74	0.48	0.56	1.00

- This correlation matrix was generated from two common factors, where the common factors have a correlation of .43

CenterStat 4.13

Note that the lower left block contains non-zero correlations due to the fact that the two factors are correlated with one another. Nevertheless, the upper left and lower right triangles of the correlation matrix have higher correlations, indicative that they share a common factor.

Specific Factors

- Independent variance, unique to the specific indicator and unexplained by the common factors, is contained in ε.
- The "uniqueness" term ε is actually composed of two parts, random measurement error and specificity (specific factor variability in true scores)
  - These two parts are typically inseparable, however, and are normally pooled together in any analysis

CenterStat 4.14

Error and specificity can, to some extent, be separated in higher-order factor model or bi-factor models, which we will discuss later.

Communality

- The communality of an item, in effect an R^2 for the regression of the indicator on the set of common factors, is defined as

$$h^2 = 1 - \frac{VAR(\varepsilon)}{VAR(y)}$$

- Items with high communalities are thought to be "good" items, as most of their variance reflects the influence of the common factors

CenterStat 4.15

The Factor Model

- Each observed variable regressed on one or more latent (or "common") factors

$$y_{1i} = \nu_1 + \lambda_{11}\eta_{1i} + \lambda_{12}\eta_{2i} + \cdots + \lambda_{1r}\eta_{ri} + \varepsilon_{1i}$$

- Note that the factor model is identical in form to a multiple regression model, but where predictors are *unobserved*
- ν designates an intercept term, λ a factor loading, η a common factor, ε the residual, and r the number of common factors
  - often, assumed that all variables (observed and latent) centered (have means of zero) so intercept term ν omitted from model
  - we include intercept here for generality since we later consider CFAs and SEMs with means

CenterStat 4.16

The Factor Model

- A set of multiple observed variables is analyzed, so that we have

$$\begin{aligned}
y_{1i} &= \nu_1 + \lambda_{11}\eta_{1i} + \lambda_{12}\eta_{2i} + \cdots + \lambda_{1r}\eta_{ri} + \varepsilon_{1i} \\
y_{2i} &= \nu_2 + \lambda_{21}\eta_{1i} + \lambda_{22}\eta_{2i} + \cdots + \lambda_{2r}\eta_{ri} + \varepsilon_{2i} \\
&\vdots \\
y_{pi} &= \nu_p + \lambda_{p1}\eta_{1i} + \lambda_{p2}\eta_{2i} + \cdots + \lambda_{pr}\eta_{ri} + \varepsilon_{pi}
\end{aligned}$$

- Given this system of equations, it is much more compact to write the model in vector/matrix form

CenterStat 4.17

The Factor Model in Matrix

- The matrix expression for this model is

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{pi} \end{pmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_p \end{pmatrix} + \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1r} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pr} \end{pmatrix} \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \\ \vdots \\ \eta_{ri} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{pi} \end{pmatrix}$$

- Or, more simply

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

Where $\mathbf{y}_i$, $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}_i$ are $p \times 1$ vectors, $\boldsymbol{\eta}_i$ is an $r \times 1$ vector, and $\boldsymbol{\Lambda}$ is a $p \times r$ factor pattern matrix (and $r < p$)

CenterStat 4.18

Recall in Chapter 1 where we defined our usual univariate regression as $y_i = \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i$; you can see the direct parallel between the regression model and the factor model defined above.

Model-Implied Moment Structure

- The factor analysis model implies a specific structure for the means and covariances of the observed measures
- We could derive these using path tracing rules as we did earlier, but it is more compact to show matrix expressions

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad \Rightarrow \quad \begin{aligned} \boldsymbol{\mu}(\boldsymbol{\theta}) &= \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha} \\ \boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \end{aligned}$$

- $\boldsymbol{\theta}$ is the vector of estimated model parameters.
- $\boldsymbol{\nu}$ is the vector of item intercepts
- $\boldsymbol{\Lambda}$ is the matrix of factor loadings
- $\boldsymbol{\alpha}$ is the mean vector of the factors
- $\boldsymbol{\Psi}$ is the covariance matrix of the factors
- $\boldsymbol{\Theta}$ is the covariance matrix of the residuals

CenterStat 4.19

These model-implied moments are important as they are used in fitting and testing the model.

Covariance Structure Analysis

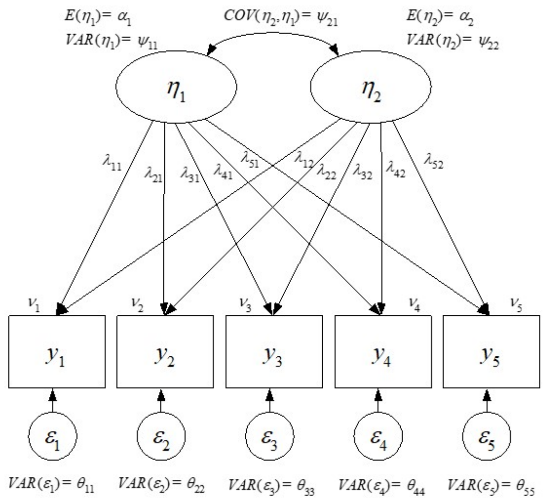
$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad \Rightarrow \quad \begin{aligned} \boldsymbol{\mu}(\boldsymbol{\theta}) &= \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha} \\ \boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \end{aligned}$$

- The mean structure is typically saturated (i.e., the observed means are perfectly reproduced), so this is often of little interest.
  - Mean structure often omitted entirely
  - Mean structure is important for some later models, however
- The correlation/covariance structure is of primary interest
  - EFA usually proceeds from analysis of correlation matrix (covariance matrix of standardized $\mathbf{y}$)
  - Goal is to reproduce correlation/covariance matrix with $r < p$ factors

CenterStat 4.20

Note that we could use precisely the same path tracing rules presented in Chapter 1 to recreate each mean, variance and covariance among the set of measured variables.

Model Matrices



$$\boldsymbol{\alpha}=\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} \quad \boldsymbol{\Psi}=\begin{pmatrix}\psi_{11} & \\ \psi_{21} & \psi_{22}\end{pmatrix}$$

$$\boldsymbol{\nu}=\begin{pmatrix}\nu_1\\ \nu_2\\ \nu_3\\ \nu_4\\ \nu_5\end{pmatrix} \quad \boldsymbol{\Lambda}=\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\\ \lambda_{31} & \lambda_{32}\\ \lambda_{41} & \lambda_{42}\\ \lambda_{51} & \lambda_{52}\end{pmatrix}$$

$$\boldsymbol{\Theta}=\begin{pmatrix}\theta_{11} & & & & \\ 0 & \theta_{22} & & & \\ 0 & 0 & \theta_{33} & & \\ 0 & 0 & 0 & \theta_{44} & \\ 0 & 0 & 0 & 0 & \theta_{55}\end{pmatrix}$$

CenterStat 4.21

Note diagonal structure of error covariance matrix, reflecting assumption of local independence. Typically both factors and items are standardized, so that mean structure can be omitted, factor loadings are interpretable like standardized regression coefficients, and ψ_{11} is in a correlation metric.

Conducting EFA

- Determine number of latent factors
  - Usually based on inspection of *eigenvalues* and *scree plot*
- "Rotate" factor pattern matrix Λ at chosen number of factors
  - Goal is usually to maximize *simple structure*
  - Can use *oblique* rotation (allow factors to correlate) versus *orthogonal* rotation (factors assumed uncorrelated)
- Eliminate problematic items and refit model (optional)
  - Items with large cross-loadings or "orphans" with low communalities
  - Excessively redundant items (otherwise may "hijack the factor")
- Inspect final factor pattern matrix to determine meaning of latent factors
  - Latent factors usually named by reference to items with highest loadings

CenterStat 4.22

An entire literature exists for each of these steps that we do not pretend to do justice to in this brief review.

Example: Holzinger-Swineford Data

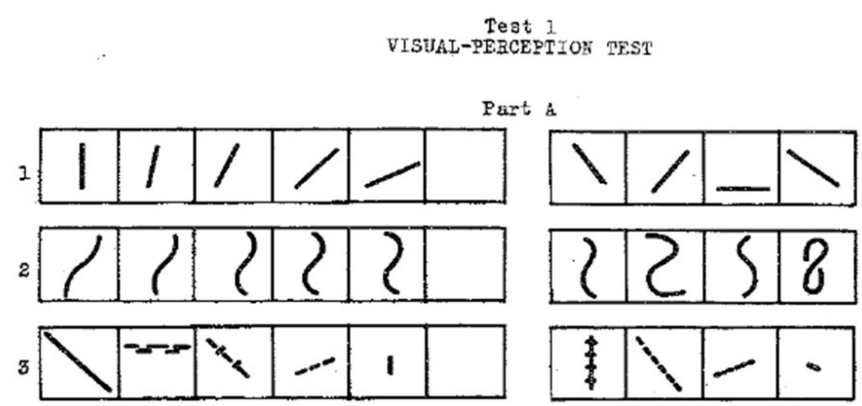
- Holzinger & Swineford (1939) collected data on mental tests to evaluate a theoretical model of intelligence
- We will not fit their bi-factor model -- we focus on a subset of the mental tests with a theoretically simpler factor structure.
- Here we consider 9 tests, with a putative 3-factor structure of visual ability, verbal ability & mental speed.
- N=301, 7th and 8th grade students, between 11-16 years of age

CenterStat 4.23

The raw data is contained in the original monograph:

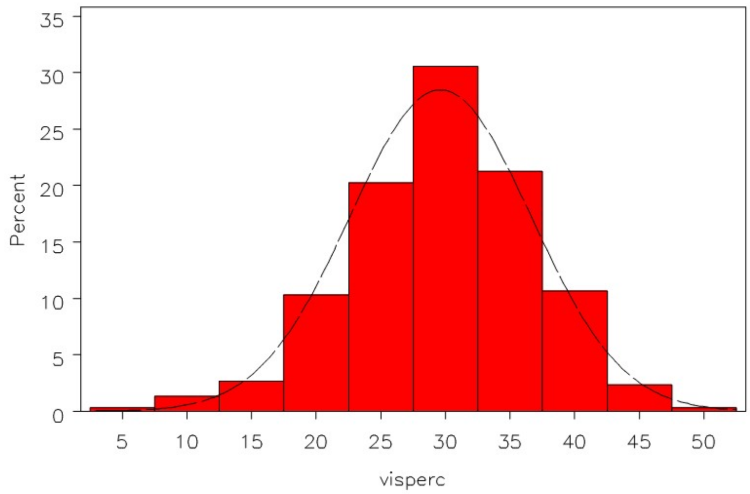
Holzinger, K.J. & Swineford, F. (1939). *A Study in Factor Analysis: The Stability of a Bi-Factor Solution.* Supplementary Educational Monographs: University of Chicago.

Visual Perception Test: visperc



CenterStat 4.24

Visual Perception Test: visperc



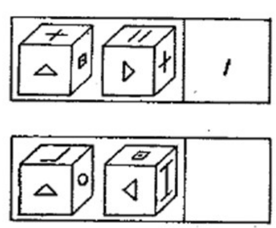
CenterStat 4.25

Note that these are continuous measures and most have reasonably nice distributions.

Cubes

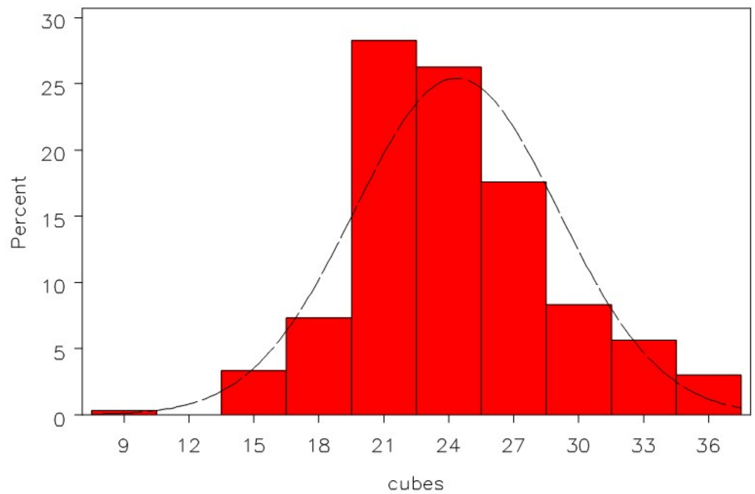
If a pair of drawings can be two pictures of the same cube, put a "1" in the blank square.

If the pair of drawings must be pictures of two different cubes, put a "2" in the square.



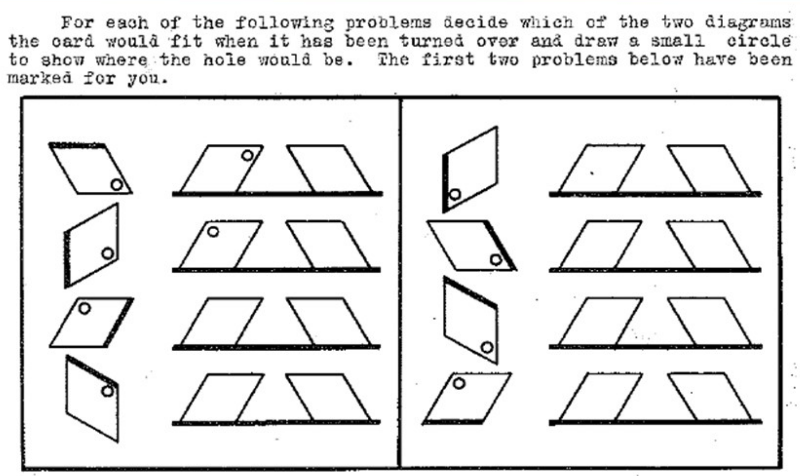
CenterStat 4.26

Cubes



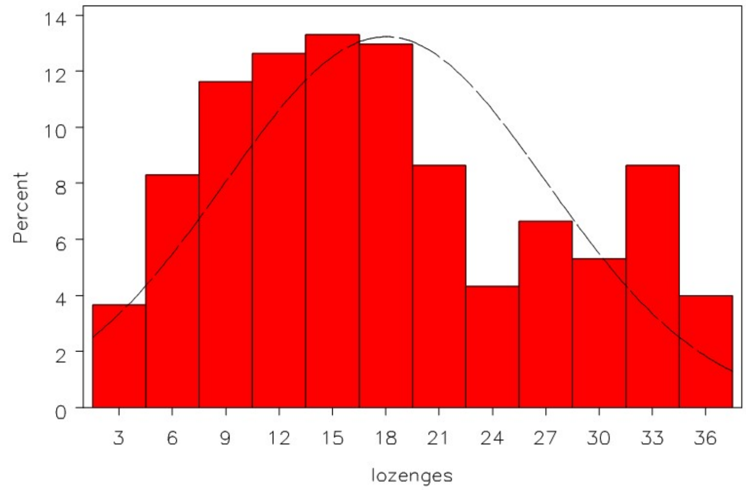
CenterStat 4.27

Lozenges



CenterStat 4.28

Lozenges



CenterStat 4.29

This distribution is not quite so pretty, but not too bad.

Paragraph Comprehension: parcomp

It was a cold and stormy December day. A blinding snow swept down upon the city from the north. A stream of afternoon shoppers hurried along the street with heads bent against the wintry blast.

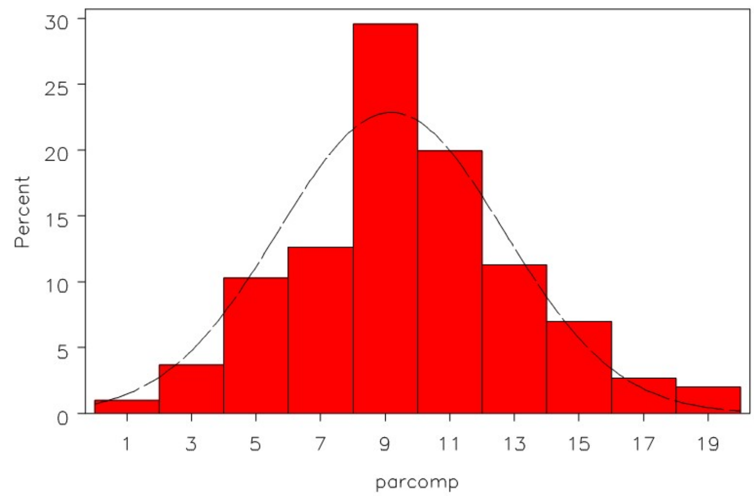
1. What season of the year was it? ____________

2. Was the sun shining? ____________

3. Underline the word that tells how the shoppers walked. (a)slowly (b)leisurely (c)gayly (d)rapidly (e)proudly

CenterStat 4.30

Paragraph Comprehension: parcomp



CenterStat 4.31

Sentence Completion: sencomp

Draw a line under the word which belongs at the end of the sentence.

1. He could not speak their language but made himself understood by
   sounds smiles signs threats

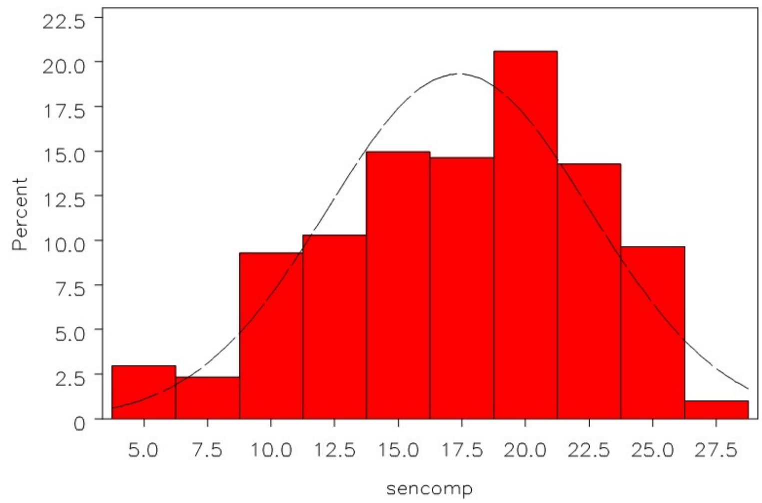
2. He could do almost everything well, because he was so
   clever brave happy tall

3. A person may have much money and yet not be able to obtain good
   clothes health dinners jewelry

4. At home she does not understand how to make
   difficulties quarrels arrangements labor

CenterStat 4.32

Sentence Completion: sencomp



CenterStat 4.33

Word Meaning: wordmean

Directions. In each of the exercises below, you should read the sample sentence or expression, then read the five words or phrases following it, and select the one whose meaning is most nearly like the meaning of the word which is underlined in the sentence. Draw a line under the word thus selected and place its number in the parenthesis at the right. Notice the example:

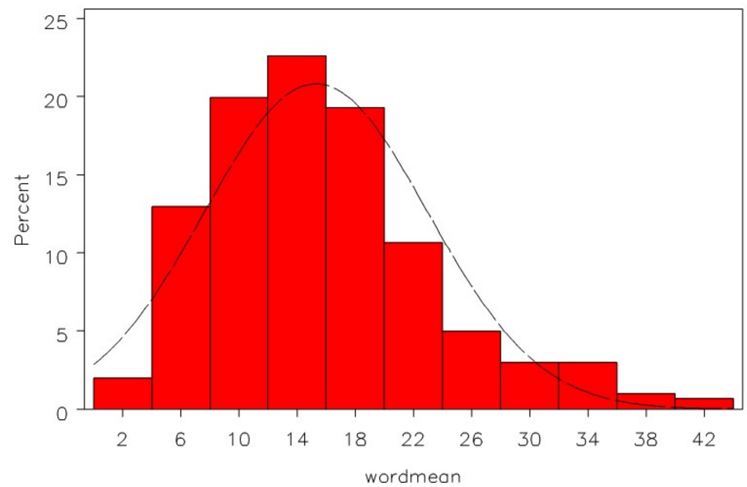
Example. They will invite him to go.
(1)tell (2)command (3)ask (4)forbid (5)select (3)

1. They adapt themselves well.
(1)please (2)adjust (3)carry (4)conduct (5)consider ( )

2. The apparatus is expensive.
(1)appliance (2)clothing (3)reading matter (4)engine (5)experiment ( )

CenterStat 4.34

Word Meaning: wordmean



CenterStat 4.35

Some skew here…

Addition

Test 10
ADD

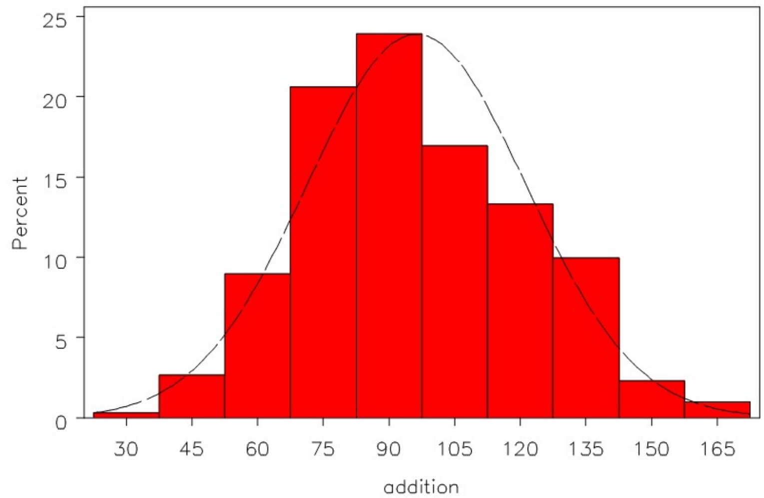
Name________________________

2	6	4	9	2	3	1	1	3	7	6	2	5	8	7	2
1	3	3	0	7	5	8	6	6	2	1	2	2	0	1	4
7	0	9	3	5	1	6	4	0	1	5	3	8	4	8	4
5	8	4	3	8	7	4	2	9	2	3	4	1	9	5	6
5	9	5	2	7	6	7	9	1	4	5	4	8	0	8	3
0	8	1	5	3	6	8	1	5	2	7	4	9	5	7	1
2	0	7	4	1	2	7	3	3	1	7	8	2	9	2	0
4	2	9	4	3	8	0	2	7	9	5	2	0	7	3	7

Timed test: 2 minutes

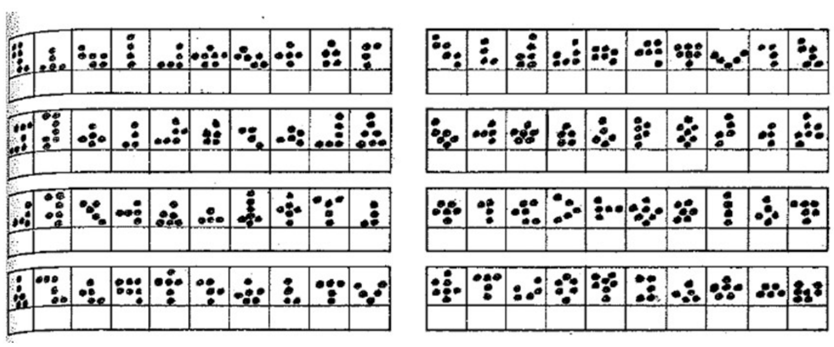
CenterStat 4.36

Addition



CenterStat 4.37

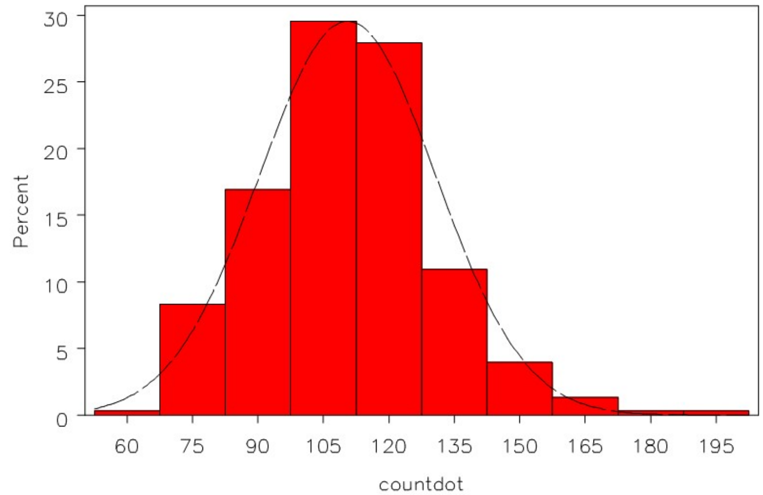
Counting Groups of Dots: countdot



Timed test: 4 minutes

CenterStat 4.38

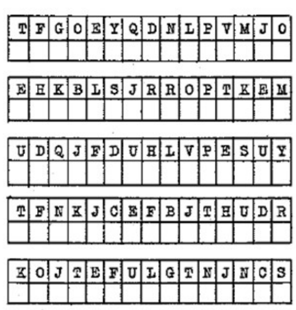
Counting Groups of Dots: countdot



CenterStat 4.39

Straight and Curved Capitals: sccaps

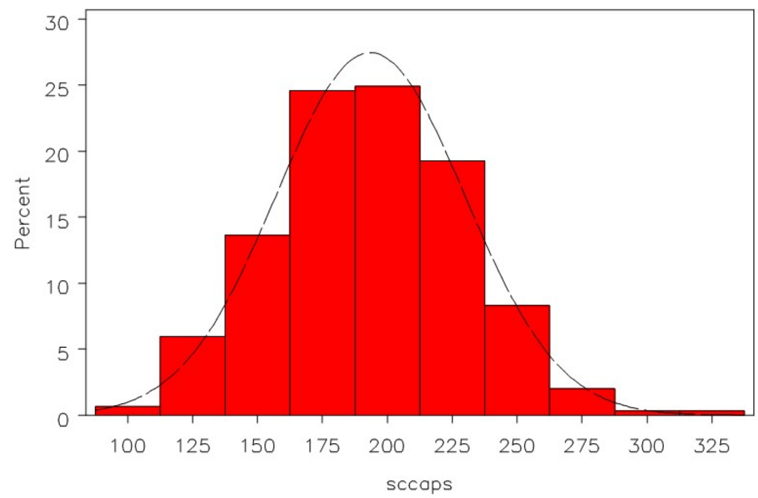
- Draw a "–"under each letter composed of all straight lines; if curved lines draw an "o"



Timed test: 3 minutes

CenterStat 4.40

Straight and Curved Capitals: sccaps



CenterStat 4.41

Descriptive Statistics

Simple Statistics

Variable	*N*	*Mean*	*Std Dev*	*Sum*	*Minimum*	*Maximum*
visperc	301	29.61462	7.00459	8914	4.00000	51.00000
cubes	301	24.35216	4.70980	7330	9.00000	37.00000
lozenges	301	18.00332	9.04784	5419	2.00000	36.00000
parcomp	301	9.18272	3.49235	2764	0	19.00000
sencomp	301	17.36213	5.16189	5226	4.00000	28.00000
wordmean	301	15.29900	7.66922	4605	1.00000	43.00000
addition	301	96.27575	25.05927	28979	30.00000	171.00000
countdot	301	110.54153	20.25230	33273	61.00000	200.00000
sccaps	301	193.46844	36.32946	58234	100.00000	333.00000

- Note that the test scores are on quite different scales, particularly addition, countdot & sccaps

CenterStat 4.42

The difference in scales here is not of much concern, since in EFA items are first standardized. But this can be important in CFA where items are generally not standardized. Very different scales can cause numerical problems for model fitting algorithms.

Associations Among Tests

Pearson Correlation Coefficients, N = 301

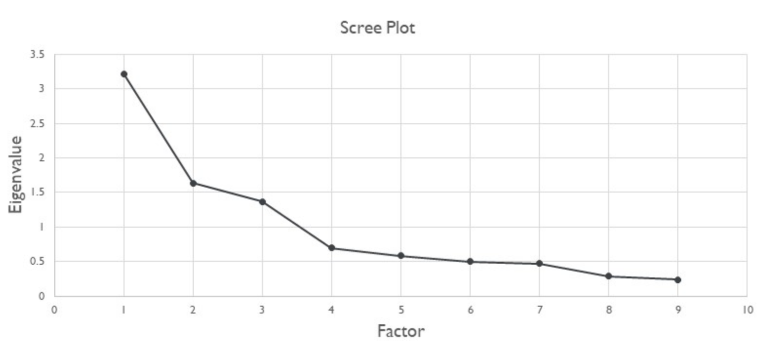
	visperc	*cubes*	*lozenges*	*parcomp*	*sencomp*	*wordmean*	*addition*	*countdot*	*sccaps*
visperc	**1.00000**								
cubes	**0.29735**	**1.00000**							
lozenges	**0.44067**	**0.33985**	**1.00000**						
parcomp	0.37271	0.15293	0.15864	**1.00000**					
sencomp	0.29344	0.13939	0.07720	**0.73317**	**1.00000**				
wordmean	0.35677	0.19253	0.19766	**0.70448**	**0.71996**	**1.00000**			
addition	0.06686	-0.07567	0.07193	0.17383	0.10204	0.12110	**1.00000**		
countdot	0.22393	0.09228	0.18601	0.10690	0.13867	0.14961	**0.48676**	**1.00000**	
sccaps	0.39034	0.20604	0.32865	0.20785	0.22747	0.21416	**0.34065**	**0.44902**	**1.00000**

- Putative 3 factor structure to the 9 measures, should be reflected in higher correlations within blocks (although some other correlations are large too)

CenterStat 4.43

In particular, *sccaps* seems to have a large correlation with some other items.

Selecting Optimal Number of Factors



- "Bend" in scree plot suggests four factors, but fourth Eigenvalue < 1.0
- Conclude three factor solution ideal for these data

CenterStat 4.44

EFA of H-S Data: Oblique Promax Rotation

Rotated Factor Pattern (Standardized Regression Coefficients)

	Factor1	Factor2	Factor3
visperc	0.22421	**0.64587**	0.08483
cubes	-0.00456	**0.75725**	-0.18493
lozenges	-0.10457	**0.79722**	0.07962
parcomp	**0.89968**	0.00664	0.00443
sencomp	**0.92410**	-0.06046	-0.00464
wordmean	**0.87278**	0.06636	-0.01378
addition	0.04244	-0.24609	**0.86118**
countdot	-0.05612	0.07347	**0.82406**
sccaps	0.01269	**0.38165**	**0.60020**

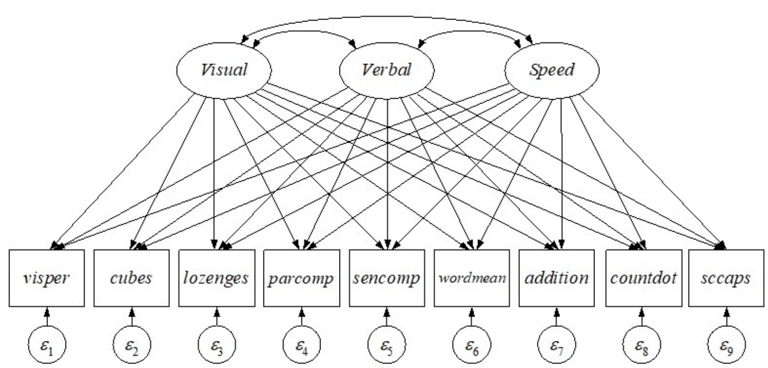
Inter-Factor Correlations

	Factor1	Factor2	Factor3
Factor1	1.00000	0.27748	0.21117
Factor2	0.27748	1.00000	0.21672
Factor3	0.21117	0.21672	1.00000

CenterStat 4.45

Here we have bolded all factor loadings above .3 to better discern the factor structure. Given the pattern of loadings, Factor 1 can be interpreted as a Verbal factor, Factor 2 as a Visual factor, and Factor 3 as a Speed factor. The three factors are all positively correlated, as one would expect. Note that simple structure was not fully achieved here (*sccaps* has a moderate cross-loading on the Visual factor).

EFA of H-S Data With 3 Factors



- Note that all items load on all factors
- Variables are standardized in analysis, so no means shown

CenterStat 4.46

Summary

- Latent variables are commonly used in everyday thinking, scientific theories, and a variety of statistical models
- EFA is an inductive procedure used to identify the number and nature of the latent factors underlying a set of observed measures
- For H-S data, the hypothesized 3 factor solution emerged from the data, but with one unexpected cross-loading item
- Given that there is a theoretical model, however, might be better to adopt a deductive, hypothesis-testing approach from the start
- This is where CFA comes in

CenterStat 4.47

4.2 Confirmatory Factor Analysis

Objectives

- Further differentiate between exploratory and confirmatory factor analysis
- Discuss model identification, estimation, and re-specification
- Demonstrate CFA

CenterStat 4.49

Confirmatory Factor Analysis

- Theory-driven: goal is to test a theoretical model that specifies the number and nature of the latent factors behind a set of correlated observed measures
- Identified through restrictions on parameters (e.g., loadings)
- Number of latent factors determined by theory
- Factor pattern matrix is restricted by analyst to reflect theory
  - e.g., to test a model that stipulates that depression will split between two correlated factors of mood-related and somatic symptoms
  - rotation of factor pattern matrices is not required
- Attention paid to <u>fit</u> of the model
  - possible to test fit because model is (typically) over-identified

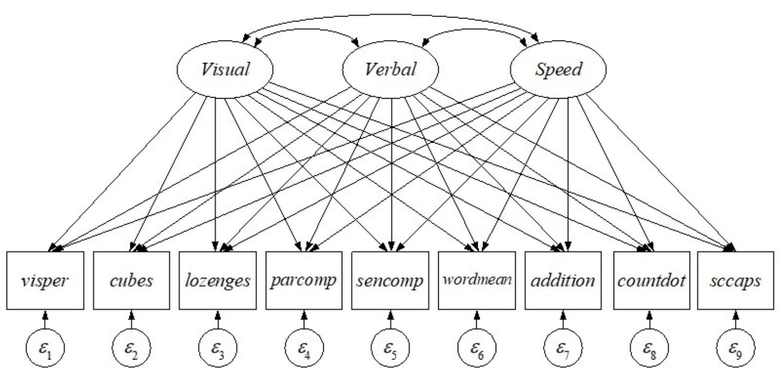
CenterStat 4.50

CFA of H-S Data

- The Holzinger-Swineford analysis is an example where CFA might be preferable to EFA
  - Theory suggests 3 factors
  - Theory suggests 3 items should load on each factor
- By fitting CFA, can formally test whether the data conforms to the theoretical model or not.

CenterStat 4.51

EFA of H-S Data With 3 Factors

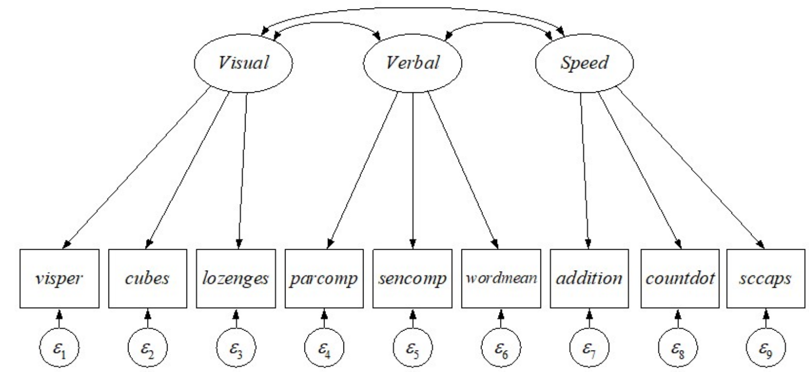


- All items load on all factors

CenterStat 4.52

For contrast, here is the EFA that we originally fit to the Holzinger & Swineford (1939) data.

CFA of H-S Data With 3 Factors



- Loadings restricted so that items load only on hypothesized factors

CenterStat 4.53

Now here is a path diagram of the CFA model. Note the highly restricted factor pattern matrix reflecting the assumption of simple structure.

EFA v. CFA

- The EFA and CFA models are the same for the H-S data except for restrictions on the factor pattern matrix

$$\text{EFA:} \begin{pmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} \\ \lambda_{51} & \lambda_{52} & \lambda_{53} \\ \lambda_{61} & \lambda_{62} & \lambda_{63} \\ \lambda_{71} & \lambda_{72} & \lambda_{73} \\ \lambda_{81} & \lambda_{82} & \lambda_{83} \\ \lambda_{91} & \lambda_{92} & \lambda_{93} \end{pmatrix} \qquad \text{CFA:} \begin{pmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & 0 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{52} & 0 \\ 0 & \lambda_{62} & 0 \\ 0 & 0 & \lambda_{73} \\ 0 & 0 & \lambda_{83} \\ 0 & 0 & \lambda_{93} \end{pmatrix}$$

CenterStat 4.54

Aside from the factor pattern matrices, the EFA and CFA models are identical for this data. Note the restriction of many loadings to zero in the CFA factor pattern matrix that were estimated in the EFA factor pattern matrix.

Model Identification

- In conducting CFA, one must ensure that the model is identified, i.e., that there is a unique solution for the parameter estimates
- This is done by imposing restrictions on the model
- In EFA, all items load on all factors, so model is not over-identified and there is no ability to test the model
- In CFA, model is identified through two types of constraints
  - Constraints that set the scale of the latent variables
  - Other restrictions, for instance on the factor pattern matrix

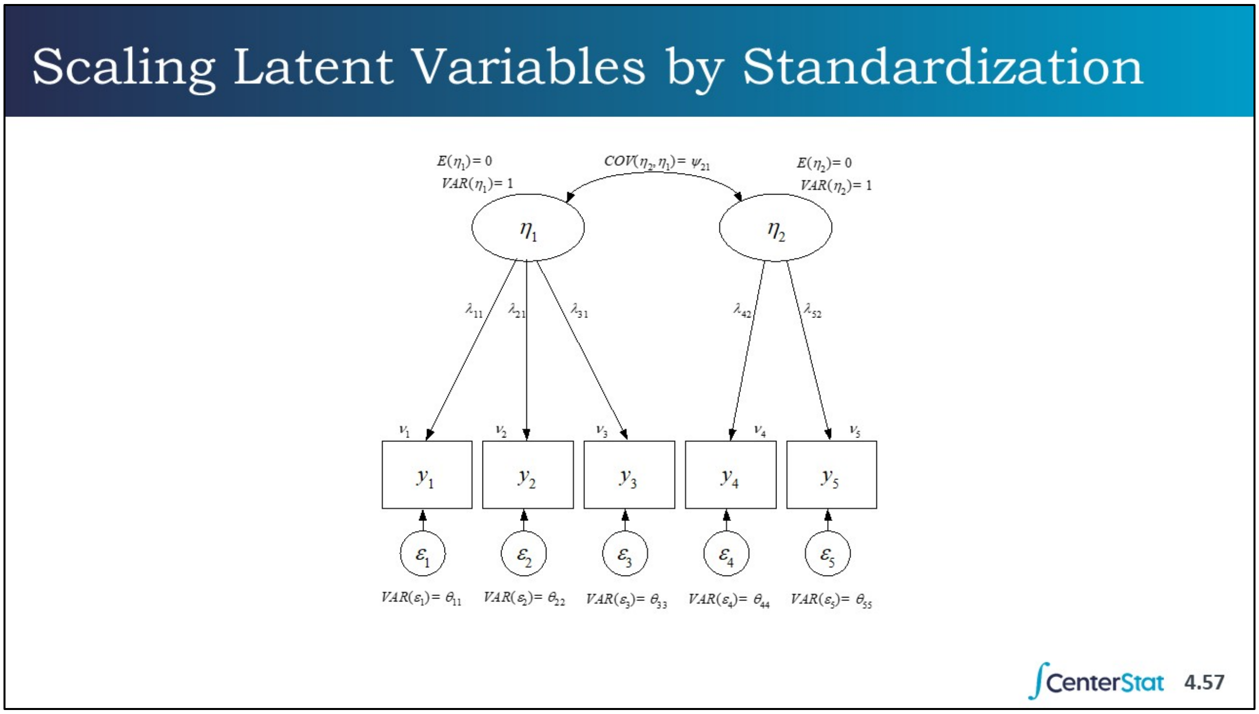
CenterStat 4.55

Setting the Scale of Latent Variables

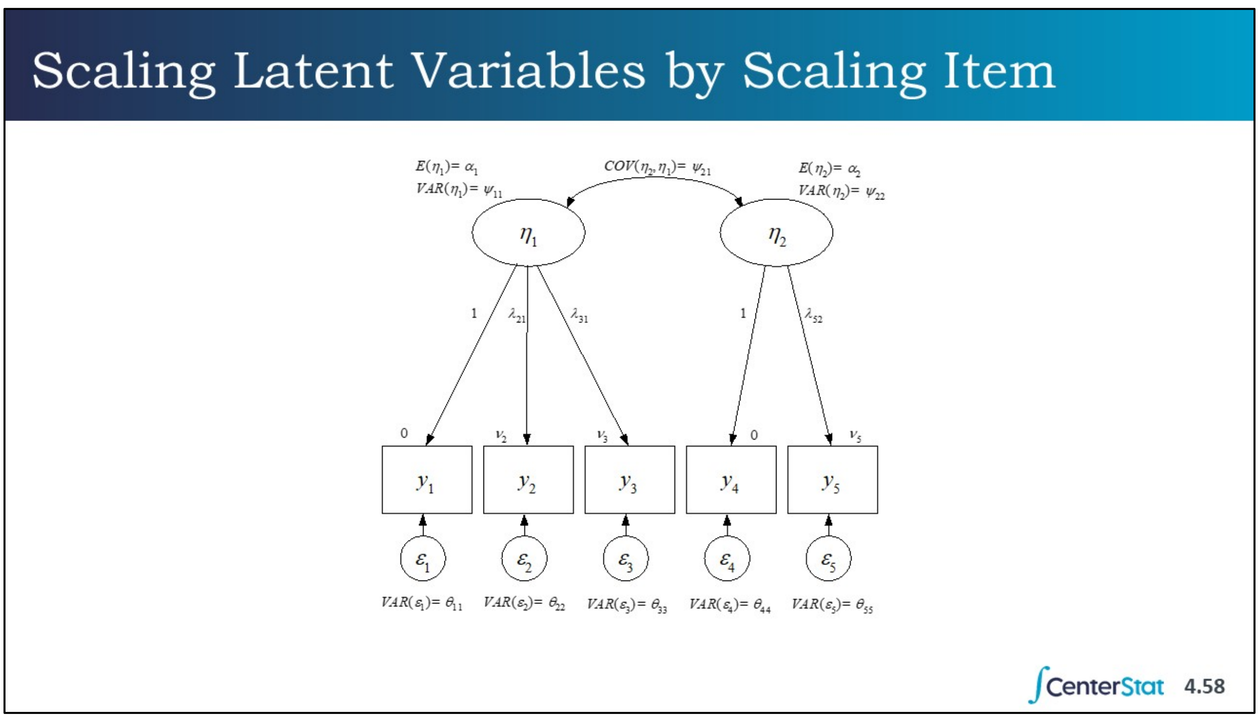
- Since latent factors are unobserved, must set *location* and *metric*
- One popular option is to set the means and variances of the latent variables to zero and one, respectively
  - Puts the latent variables on a standardized scale
  - Allows for estimation of all factor loadings (and intercepts)
  - Same scaling option that is used in EFA
- Another option is to give each latent variable the same metric as one of the observed variables. The intercept and loading for the "scaling item" are set to zero and one, respectively.
  - Mean and variance of latent variable now estimated
  - The latent variable has the mean and metric of the scaling item

CenterStat 4.56

Other scaling constraints can also be considered, but these are the two most common.



Note that in this path diagram the factors have been scaled by setting their means and variances to zero and one. The factor covariance is then also interpretable as a factor correlation.



In contrast, in this diagram, the mean and variance of each factor is estimated. This is possible because the metric of the latent variables has been defined to match the metric of the scaling indicators y_1 and y_4. The intercept and loading of each scaling indicator has been set to zero and one, respectively.

Identification Conditions

- Additional restrictions are typically necessary to identify the CFA
- Common restrictions include
  - Assuming no (or few) cross-loading items
  - Assuming no (or few) covariances between residuals
- Model identification can be determined by
  - Matrix algebra (hard)
  - *t*-rule (necessary)
  - Three-indicator rule (sufficient)
  - Two-indicator rule (sufficient)
- We review these rules under the assumption the mean structure is saturated
  - i.e., as many means and intercepts are estimated as are observed
  - This is typical for CFA

CenterStat 4.59

t-Rule

- Based on simple notion that there must be more observed pieces of information than parameters estimated
- Recall that model-implied moment structure is

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha} \qquad \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$$

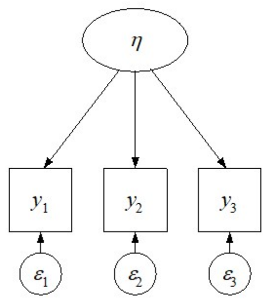
- The population moments are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
- There are p unique elements in $\boldsymbol{\mu}$ and $p(p+1)/2$ unique elements in $\boldsymbol{\Sigma}$
- Therefore, if t is the number of freely estimated parameters in θ then a necessary (but not sufficient) condition for identification is

$$t \le p(p+1)/2 + p$$

CenterStat 4.60

Three-Indicator Rule

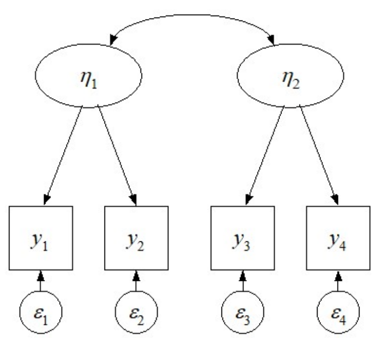
- A sufficient condition for identification is for there to be three indicators with non-zero loadings for one factor, and Θ diagonal (local independence).
- None of the indicators cross-loads with another factor (if present)



CenterStat 4.61

Two-Indicator Rule

- A sufficient condition for identification is for there to be two indicators per factor, at least two factors, diagonal Θ, no cross-loadings, and non-zero correlation between factors



CenterStat 4.62

Example: CFA for H-S Data

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

$$\begin{pmatrix} visper_i \\ cubes_i \\ lozenges_i \\ parcomp_i \\ sencomp_i \\ wordmean_i \\ addition_i \\ countdot_i \\ sccaps_i \end{pmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \nu_5 \\ \nu_6 \\ \nu_7 \\ \nu_8 \\ \nu_9 \end{pmatrix} + \begin{pmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & 0 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{52} & 0 \\ 0 & \lambda_{62} & 0 \\ 0 & 0 & \lambda_{73} \\ 0 & 0 & \lambda_{83} \\ 0 & 0 & \lambda_{93} \end{pmatrix} \begin{pmatrix} visual_i \\ verbal_i \\ speed_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \\ \varepsilon_{5i} \\ \varepsilon_{6i} \\ \varepsilon_{7i} \\ \varepsilon_{8i} \\ \varepsilon_{9i} \end{pmatrix}$$

CenterStat 4.63

Let us consider the identification of the CFA model we seek to fit to the Holzinger & Swineford (1939) data. Note that the factor pattern matrix already contains many restrictions.

Example: CFA for H-S Data

- Setting scale of latent factors (standardized):

$$E(\boldsymbol{\eta}_i) = \boldsymbol{\alpha} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \quad COV(\boldsymbol{\eta}_i) = \boldsymbol{\Psi} = \begin{pmatrix} 1 & & \\ \psi_{21} & 1 & \\ \psi_{31} & \psi_{32} & 1 \end{pmatrix}$$

- Assuming local independence

$$COV(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta} = DIAG(\theta_{11}, \theta_{22}, \theta_{33}, \theta_{44}, \theta_{55}, \theta_{66}, \theta_{77}, \theta_{88}, \theta_{99})$$

CenterStat 4.64

Here we have scaled the latent factors by setting their means and variances to zero and one, respectively. We've also imposed restrictions on the covariance matrix for the residuals consistent with the local independence assumption (that the covariance matrix be diagonal).

Identification of H-S CFA

- *t*-rule:
  - t = 9 intercepts + 9 loadings + 9 residual variances + 3 factor correlations = 30 parameters
  - p = 9 observed variables

$$t \le \frac{p(p+1)}{2} + p \qquad 30 \le \frac{9(9+1)}{2} + 9 \qquad 30 \le 54$$

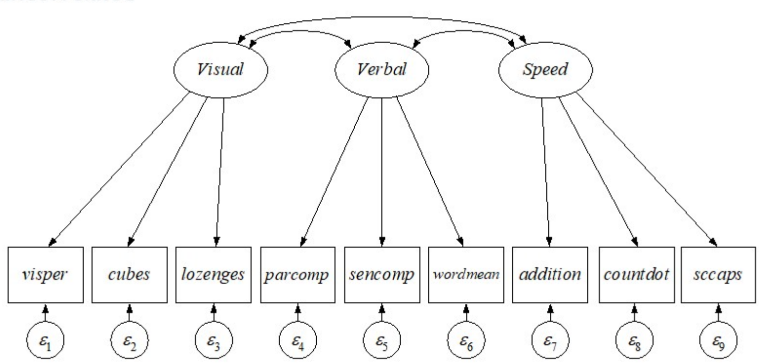
  - Model meets necessary (but not sufficient) condition for identification
  - 54 – 30 = 24 degrees of freedom for chi-square test of model fit

CenterStat 4.65

Counting up all the to-be-estimated parameters and comparing the number of unique first- and second-order moments (means and covariances) we obtain a difference of 24. Thus the model meets the *t*-rule, and the chi-square test of model fit will have 24 degrees of freedom.

Identification of H-S CFA

- 3-indicator rule:
  - Each factor has 3 indicators, none cross-load, and uniquenesses are uncorrelated

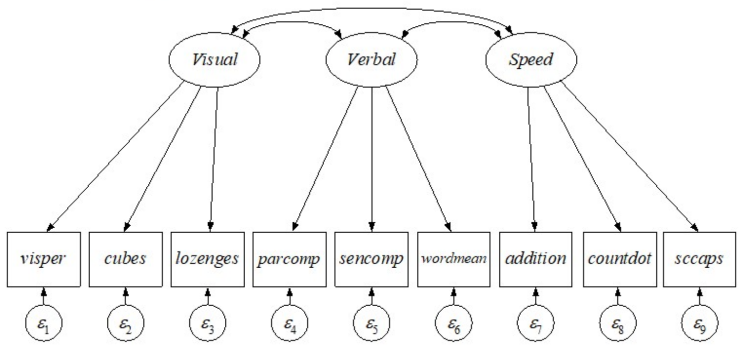


CenterStat 4.66

Each factor has three indicators, so the 3-indicator rule is satisfied here.

Identification of H-S CFA

- 2-indicator rule:
  - Even if only 2 indicators per factor, would be identified, since correlated factors, no cross loadings, independent uniquenesses



CenterStat 4.67

The model would likewise be identified if there were but two indicators per factor, given inter-factor correlations.

Model Estimation

- Given identification of model, can now turn to estimation
- As with simultaneous equation models, multiple estimators exist
  - Maximum likelihood
  - Ordinary least squares
  - Generalized least squares
  - Weighted least squares
- As before, we will concentrate on maximum likelihood, for which there are sufficient-statistic and raw-data variants
  - recall that SSML and FIML are equivalent when DVs are continuously and normally distributed and there are no missing data

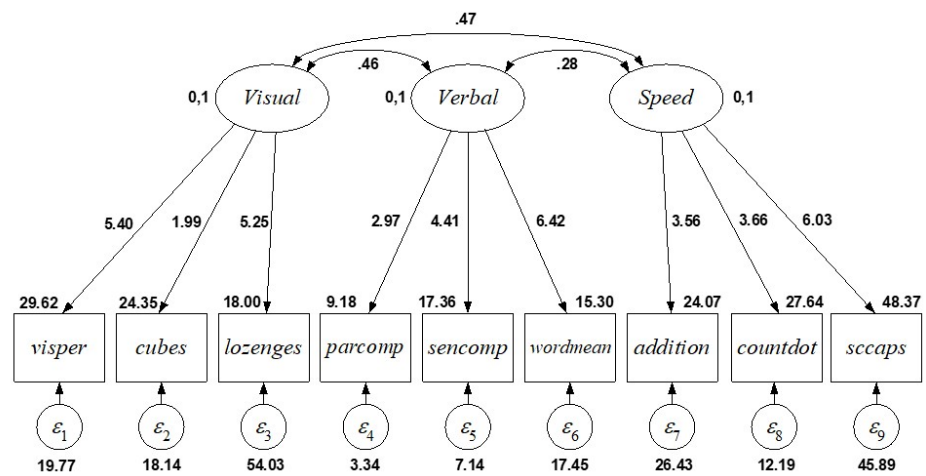
CenterStat 4.68

H-S CFA: Overall Model Fit

- As with simultaneous equation models, rely on chi-square test as well as fit statistics such as CFI, TLI, RMSEA, SRMR
- For 3-Factor model on H-S data, obtain
  - $\chi^2(24) = 85.31, p < .0001$
  - CFI = .93, TLI = .90
  - RMSEA = .092 CI_{90} = (0.071, 0.114)
  - SRMR = .060
- The hypothesized model is rejected, and fit indices generally reflect poor model fit
  - We will examine the parameter estimates despite this, although we might normally proceed to model respecification

CenterStat 4.69

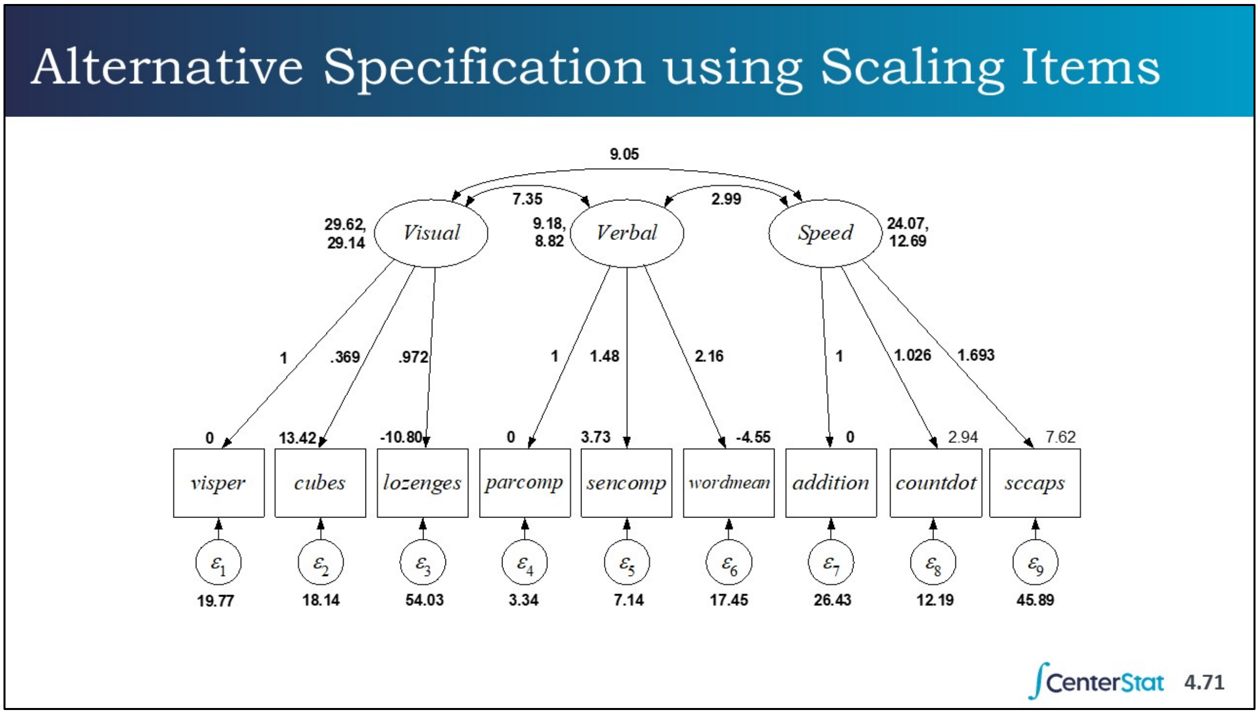
ML Estimates for H-S CFA



CenterStat 4.70

The estimates shown here are interpreted differently than in EFA. In EFA, both factors and variables are standardized, so loadings are interpreted as standardized regression coefficients, intercepts are all zero, and residual variances are equal to 1-communality. Here, only the latent variables are standardized, not the observed variables, so factor loadings are interpreted as raw regression coefficients and intercepts are non-zero. The estimates are not directly comparable across items, given that the items are on different scales. We will proceed to a standardized solution shortly.

Given the discrepant scales of the observed variables, the scores for the three speed indicators were rescaled prior to the analysis by dividing by four. Estimation is often quicker and more stable when the observed variables have similar standard deviations.



Here the first indicator per factor has been chosen as the scaling item. The factor loadings and intercepts of the remaining two indicators on each factor are then interpreted in reference to the scaling item.

This scaling option is more useful when indicators are measured with the same range (e.g., all items are scored from 0 to 100). In that case, factor loadings over one indicate items with stronger relations to the latent factor than the scaling item and intercepts greater than zero indicate items with higher means than the scaling item (e.g., easier tests).

In this case, given the discrepant scales of the indicators, the raw parameter estimates are somewhat difficult to interpret.

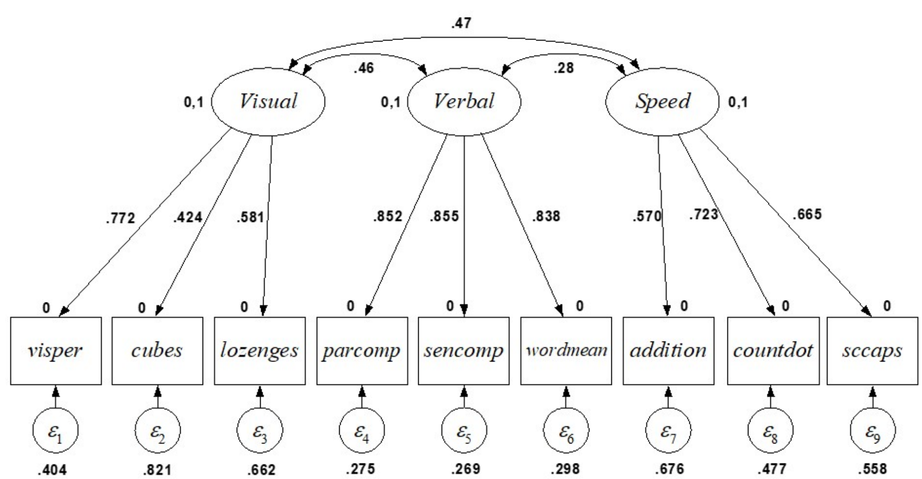
Evaluation of Parameter Estimates

- Parameter estimates interpreted in terms of direction and significance, but relative magnitude sometimes difficult to assess if observed variables are on different scales
- Compute standardized parameter estimates to aid evaluation
  - factor loadings: same as computing standardized regression coefficients
  - factor covariances: rescaling to correlation metric
  - item residual variances: rescaled to unexplained variance
- Can also compute communality (h^2) to determine amount of indicator variance accounted for by factors
  - sometimes labeled R^2 in computer output
  - in standardized solution, equivalent to $1 - \theta$ (1 - residual variance)

CenterStat 4.72

When we refer to a standardized solution, this is a solution in which both the observed and latent variables are scaled to have means and variances of zero and one. Some software programs standardize differently, or provide multiple standardization options that may be useful in different settings.

Standardized Solution for H-S CFA



CenterStat 4.73

Note that these estimates are now in the same scale that is customary for EFA estimates.

Model Respecification

- Given poor fit of a model, may seek to make changes to achieve adequate fit
- Changes may be based on theory, or may be based empirically on examination of modification indices
  - Modification index is expected change in chi-square with relaxation of a given parameter constraint
- Simulation studies have shown that over-reliance on modification indices will often lead to incorrect respecifications
- However, can simultaneously inspect modification indices and consider whether suggested respecification is consistent with theory

CenterStat 4.74

Examining mean and covariance residuals can also be helpful, but often provides redundant information with modification indices, so we focus on the latter here.

Model Respecification

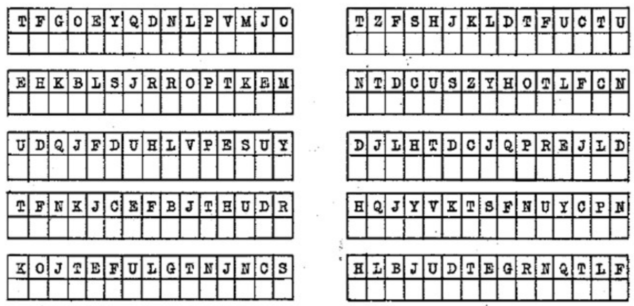
- For H-S 3-Factor model, find two large modification indices:
  - M.I. = 36.47 to include a cross-loading of sccaps on the visual factor
  - M.I. = 34.12 to allow residuals to covary for countdot and addition
- Do these suggested paths make sense to add to the model?
- Note that M.I.s provide information on one restriction in isolation of all others
  - Both M.I.s above may reflect the same misspecification
  - M.I.s aren't usually informative about major structural problems in the model (e.g., need for additional factors)

CenterStat 4.75

We will consider each large MI for this model in turn. We first consider whether the MI is plausible, and then (assuming plausibility) refit the model with the new path included and reassess model fit.

M.I. for sccaps on Visual Factor

- Draw a "–" under each letter composed of all straight lines; if curved lines draw an "o" (3 minutes)



- sccaps obviously contains a visual element, so cross-loading on visual factor might make sense, even if primarily a speed test

CenterStat 4.76

Loading sccaps on Visual Factor

$\chi^2(23) = 52.38$, $p = .0004$

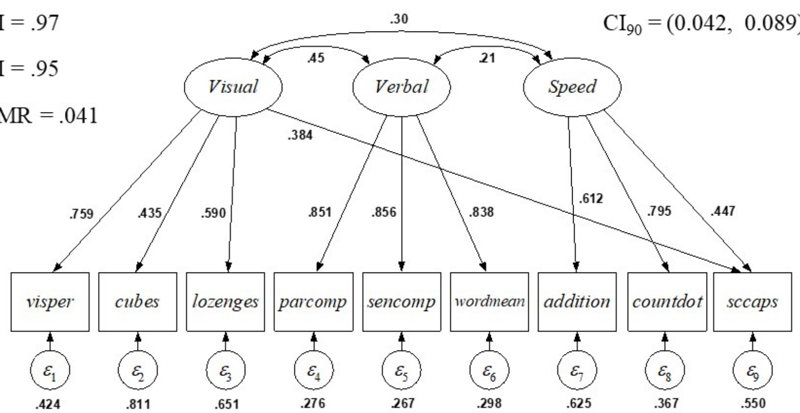
CFI = .97

TLI = .95

SRMR = .041

RMSEA = .065

$CI_{90} = (0.042,\ 0.089)$

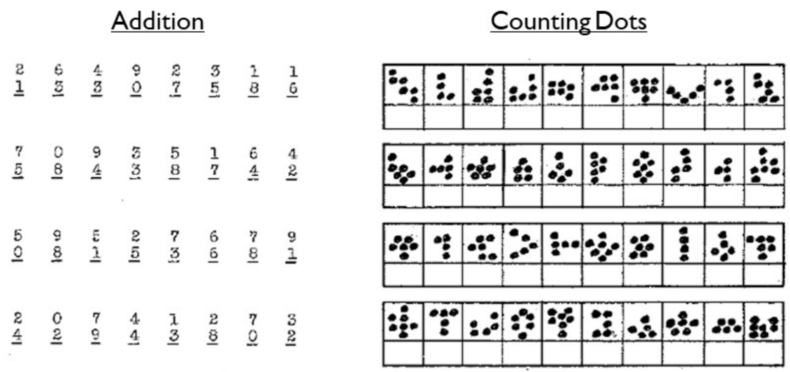


CenterStat 4.77

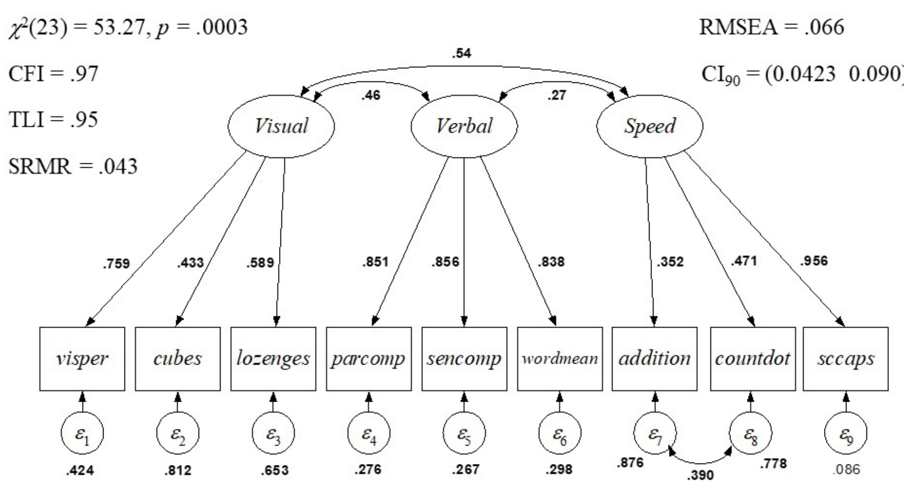
Interestingly, this is highly similar to the factor structure that EFA identified for the data. No large MIs were obtained from this model.

M.I. for countdot with addition

- To some degree, both addition and countdot involve math, so could see residual correlation as influence of math minor factor



Including Correlated Residuals



This model fits nearly identically to the prior one. Again, no large MIs were detected. Both models modifications are likely accommodating the same source of model misfit.

Model Respecification

- Ultimately, how to respecify a model is a partially subjective decision
  - Including cross-loading or correlated residual in H-S model results in OK but not great fit. In modified models, no large modification indices.
  - Hard to say which to include – we would probably include cross-loading and not correlated residual, since otherwise speed factor gets hijacked by sccaps item
  - Also makes sense to view speed tests as a combination of mental quickness and content matter ability
  - If had additional math tests, might then also include a math factor and have countdot and addition crossload on it.

CenterStat 4.80

Confirmatory or Exploratory?

- The distinction between confirmatory and exploratory analysis is seldom so sharp as it might initially seem
- EFA often conducted even though there is a theoretical model for what the factor structure *should* look like
- CFA tests a hypothesized factor structure, but if model fails to fit respecification moves the analysis in an exploratory direction.
- EFA and CFA can also be used in conjunction
  - Can partition sample, conduct EFA in first sample, delete problem items, select # factors, determine primary loadings, then conduct cross-validation analysis using CFA on hold-out sample (e.g., Sayers, Curran & Mueser, 1996).

CenterStat 4.81

Summary

- Conducting a CFA involves:
  - Model specification
    - Theory based, hypothesis driven
  - Identification
    - Scaling the factors
    - Other identification constraints
  - Estimation
    - Maximum likelihood
  - Evaluation
    - Overall model fit
    - Individual parameter estimates
  - Respecification (if necessary)
    - Moves towards exploratory analysis

CenterStat 4.82

Chapter Summary

- Discussed the use of latent variables in everyday language, science, and statistics
- Briefly described the EFA model
- Explicated the CFA model:
  - Specification, identification, estimation, evaluation, respecification
  - Reflective versus formative measurement models
  - Assumptions and issues that arise in estimation and interpretation
  - Ability to fit more complex, theoretically motivated factor models
- Next expand the CFA to include structural relations among the latent factors via the full SEM

CenterStat 4.83

4.3 Self-Study: Issues and Extensions

Objectives

- Note assumptions of the CFA model
- Discuss issues that can arise in model estimation
- Show how CFA used to specify a variety of factor structures
- Distinguish between formative and reflective factor models

CenterStat 4.85

Assumptions

- The CFA model assumes
  - Items are linearly related to the latent factors
  - Uniquenesses are uncorrelated with common factors
  - Expected values of uniquenesses are zero
  - All individuals' data can be described with one covariance matrix and one mean vector
  - Individuals are independent of one another

CenterStat 4.86

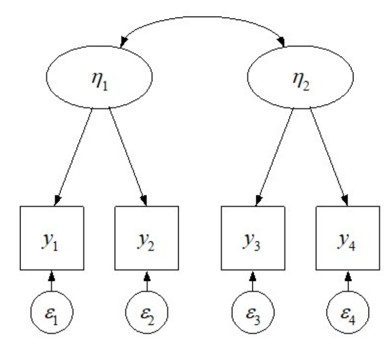
Assumptions

- Other assumptions often noted are assumptions of <u>estimators</u> rather than assumptions of the <u>model</u>
  - Normality is an assumption made by ML estimator with naïve standard errors
- Later, we will discuss how to relax some of these assumptions
  - e.g., categorical SEM and multiple groups SEM

CenterStat 4.87

Empirical Under-Identification

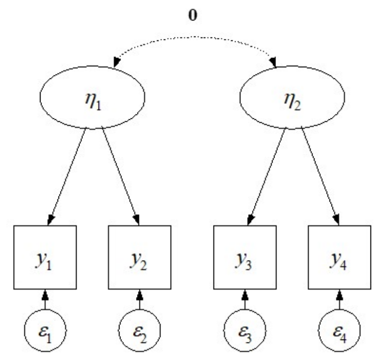
- Empirical under-identification refers to the condition where the model would normally be identified, but one or more of the parameter estimates takes on a null value (or some other value) that leads to under-identification.
- Example:
  The model to the right is identified by the two indicator rule



CenterStat 4.88

Empirical Under-Identification

- But, as the estimated covariance between factors approaches zero, the model becomes empirically under-identified
- Each factor becomes isolated from the other and fails the three-indicator rule
- Empirical under-identification for this model would also occur if any of the factor loadings approached zero



CenterStat 4.89

Heywood Cases & Improper Solutions

- Heywood cases were first noticed in EFA but can occur in CFA as well: Refers to the situation where the estimated residual variance for one or more items is negative
- The term <u>improper solution</u> (or inadmissible solution) refers to a solution that includes either negative variance estimates or implied correlations greater than one
  - e.g., $|\text{inter-factor correlation}| > 1.0$
- In general, all estimates should be in the plausible range
  - Variances must be positive
  - Correlations can't exceed 1 in absolute magnitude

CenterStat 4.90

Heywood Cases & Improper Solutions

- Often, an improper solution is taken as evidence of model misspecification
  - But Chen et al (2001) found improper solutions sometimes less common in misspecified models
- Could be due to outliers or influential cases
  - Always check the data before conducting an analysis
- Can occur if model is not identified
  - Check identification rules; consider possible empirical under-identification

CenterStat 4.91

Heywood Cases & Improper Solutions

- Could also be sampling variance
  - Improper solutions are most common at low N
  - Can test if negative variance estimate is significantly different than zero
  - If not too large, can "fix" problem by setting variance estimate to zero, or imposing inequality constraint that it cannot go below zero

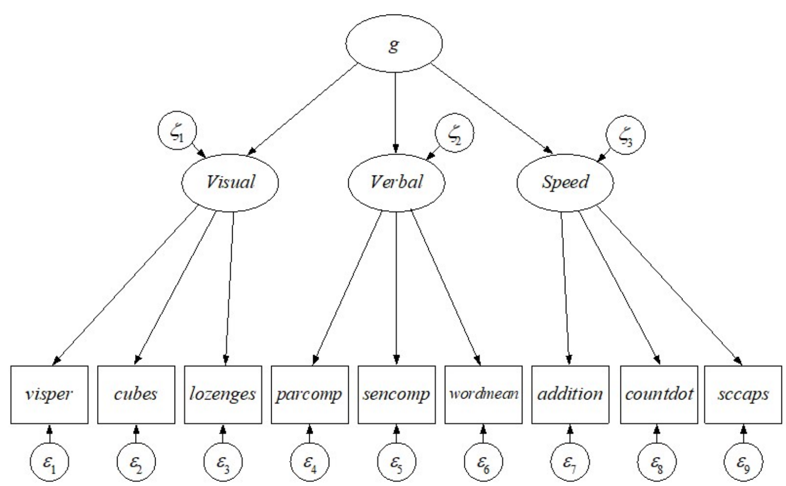
CenterStat 4.92

Advanced Applications

- So far, all the CFA models we have examined have assumed a simple, clustered factor structure
  - Each factor is defined by an exclusive set of items
  - Each item loads on only one factor
- We can, however, also use the CFA to test a variety of other model structures
  - Higher-order factor models
  - Bi-factor models
  - Models for multitrait-multimethod data

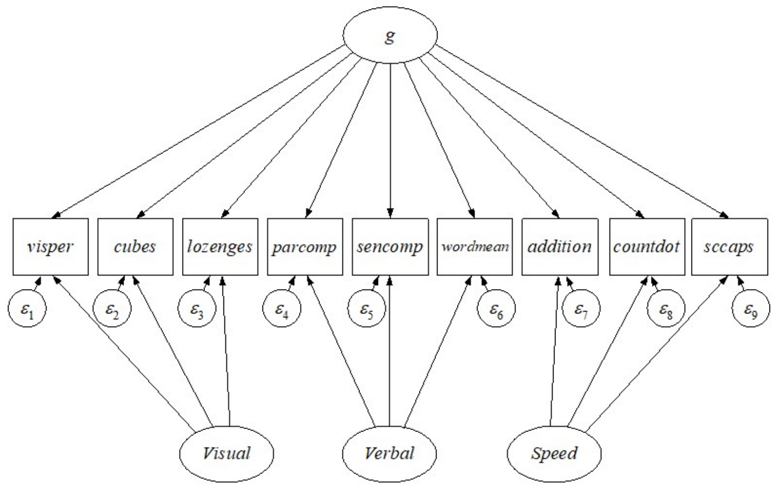
CenterStat 4.93

Higher-Order Factor Model for H-S Data



CenterStat 4.94

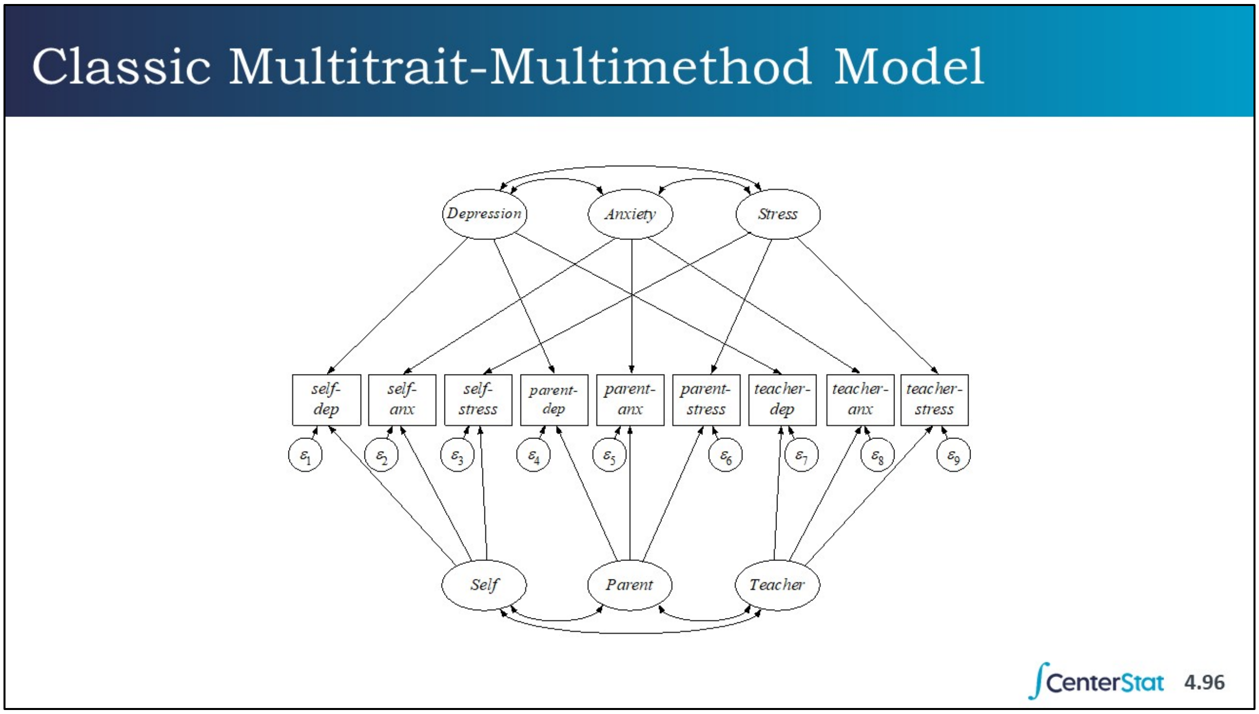
Bifactor Model for H-S Data



CenterStat 4.95

Bifactor models (sometimes referred to as hierarchical factor models) are similar to higher-order factor models in some ways and the two models are often in fact nested, permitting likelihood ratio tests to compare between the two structures (with the higher-order factor model being the more restricted null model and the bifactor model being the less parsimonious alternative model). For further details on the relationships between these two models see

Yung, Y.-F., Thissen, D. & McLeod, L.D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64,* 113-128.



This model is often referred to as the Correlated Trait/Correlated Method (CTCM) model and was the first to be broadly implemented with multitrait-multimethod data, with key references being:

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

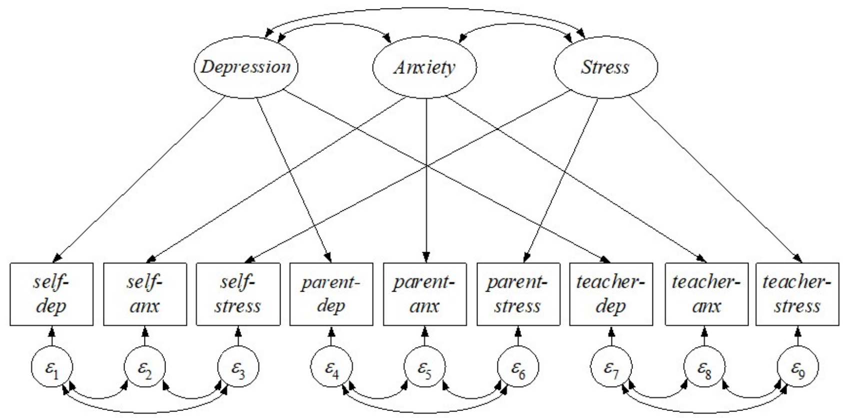
Widaman K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1–26.

Unfortunately, a well-known problem with the CTCM model is that it often fails to converge due to empirical under-identification:

Kenny D.A, Kashy D.A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165–172.

Marsh H.W, Bailey M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of the behavior of alternative models. *Applied Psychological Measurement, 15*, 47–70.

Correlated Uniqueness Model



4.97

Given the estimation difficulties associated with the CTCM model, Marsh (1989) advocated this Correlated Trait/Correlated Uniqueness (CTCU) model as an alternative approach to accounting for shared method variance. The modeling of MTMM data remains an active area of research with other modeling alternatives proposed by Eid (2000), Pohl and Steyer (2010), and Castro-Schilo, Grimm, and Widaman (2013).

Castro-Schilo, L., Grimm, K. J., & Widaman, K. F. (2013). Uncrossing the correlated trait-correlated method model for multitrait-multimethod data [Abstract]. *Multivariate Behavioral Research, 48*,152

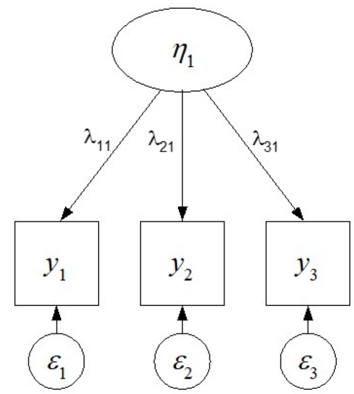
Eid M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.

Marsh H.W. (1989). Confirmatory factor analyses of multitrait-multimethod data: many problems and a few solutions. *Applied Psychological Measurement, 13*, 335–361.

Pohl S., & Steyer R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research, 45*, 45–72.

Reflective Factor Models

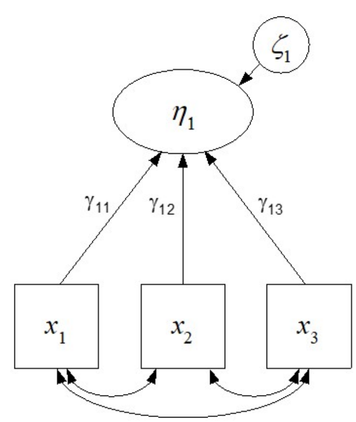
- Most factor models are reflective, meaning that the indicators are thought to be observed reflections of the latent variable
- y_1, y_2, and y_3 are referred to as *effect indicators*
- In the model shown here, a one unit change in η produces a λ_{11} unit change in y_1, a λ_{21} unit change in y_2, and a λ_{31} unit change in y_3
- That is, a change in the latent factor results in a simultaneous shift in all observed indicators of that factor



CenterStat 4.98

Formative Factor Models

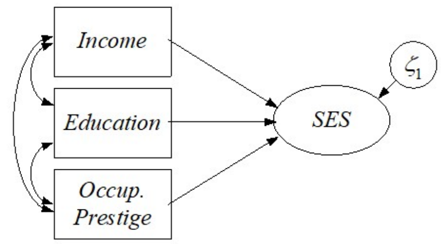
- In some cases, it may make more sense to view the indicators as causing the factor
- In this model, the latent variable is defined by a composite of the observed measures
- x_1, x_2 and x_3 are referred to as *causal indicators*
- In this model, a change in x_1 produces a γ_{11} unit change in η, but x_2 and x_3 can remain constant



CenterStat 4.99

A Formative Model: The Case of SES

- Bollen & Lennox (1991) argue that socioeconomic status is a good case for a formative model
- A one unit change in SES does not cause a simultaneous change in income, education and occupational prestige (reflective model)
- A change in income, education, or occupational prestige increases SES (formative model)



CenterStat 4.100

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

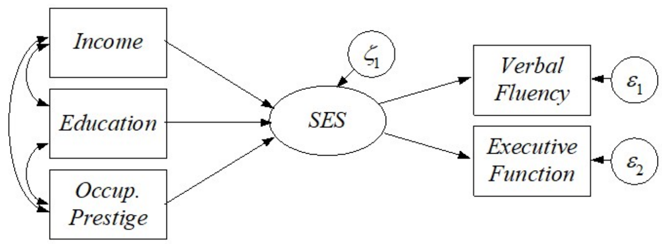
Fitting Formative Models

- The primary issue that one must consider when fitting a formative model is that the factor model isn't identified on its own
- The factor must be a cause of other observed variables for the model to be identified

CenterStat 4.101

Fitting Formative Models

- For instance, for SES we might identify the model as follows

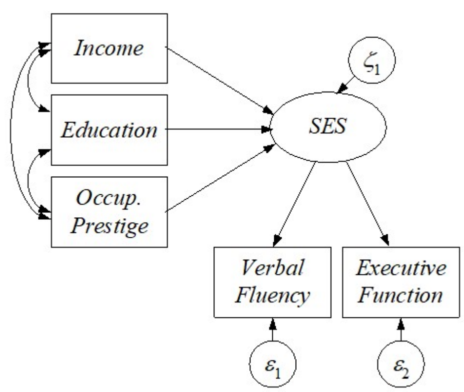


- But is this a formative model for SES with causal indicators of income, education and occupational prestige and effects of SES on verbal fluency and executive function…

CenterStat 4.102

Fitting Formative Models

…or is this a reflective model for SES (IQ?) with effect indicators of verbal fluency and executive function?



CenterStat 4.103

Summary

- The CFA makes a number of assumptions, including linearity, homogeneity, and independence
- Even if the model is identified by the 3-indicator or 2-indicator rules, empirical under-identification can still occur
- Can obtain an improper solution (e.g., Heywood case), especially at low sample sizes
- CFA not restricted to simple structure; can specify factor structure that bests conform to theory
- Should consider distinction between formative and reflective factor models when specifying model

CenterStat 4.104

Chapter 5
Structural Equation Models with Latent Variables

5.1 Introduction to Structural Equation Models

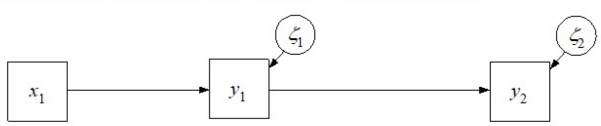
Objectives

- Introduce SEM as a flexible combination of path analysis and CFA
- Review recent SEMs in the literature

CenterStat 1.3

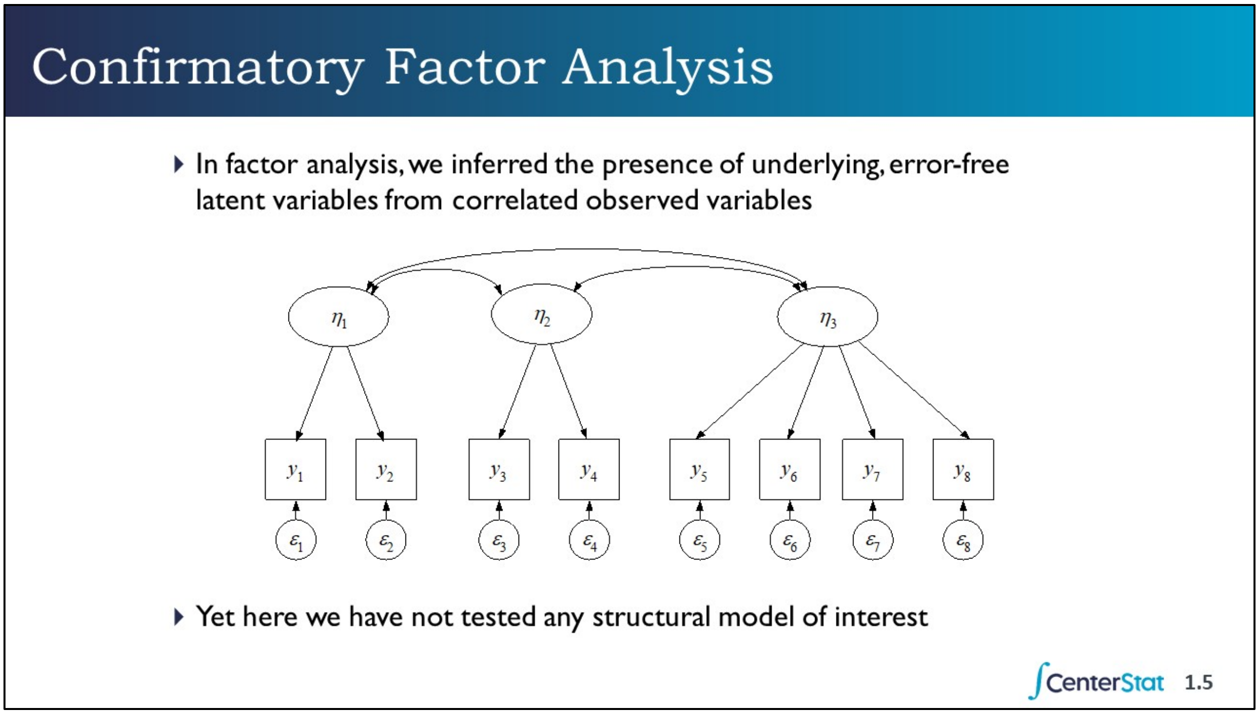
Path Analysis Models

- With path analysis (or *simultaneous equations*) models we could fit complex structural models for observed variables

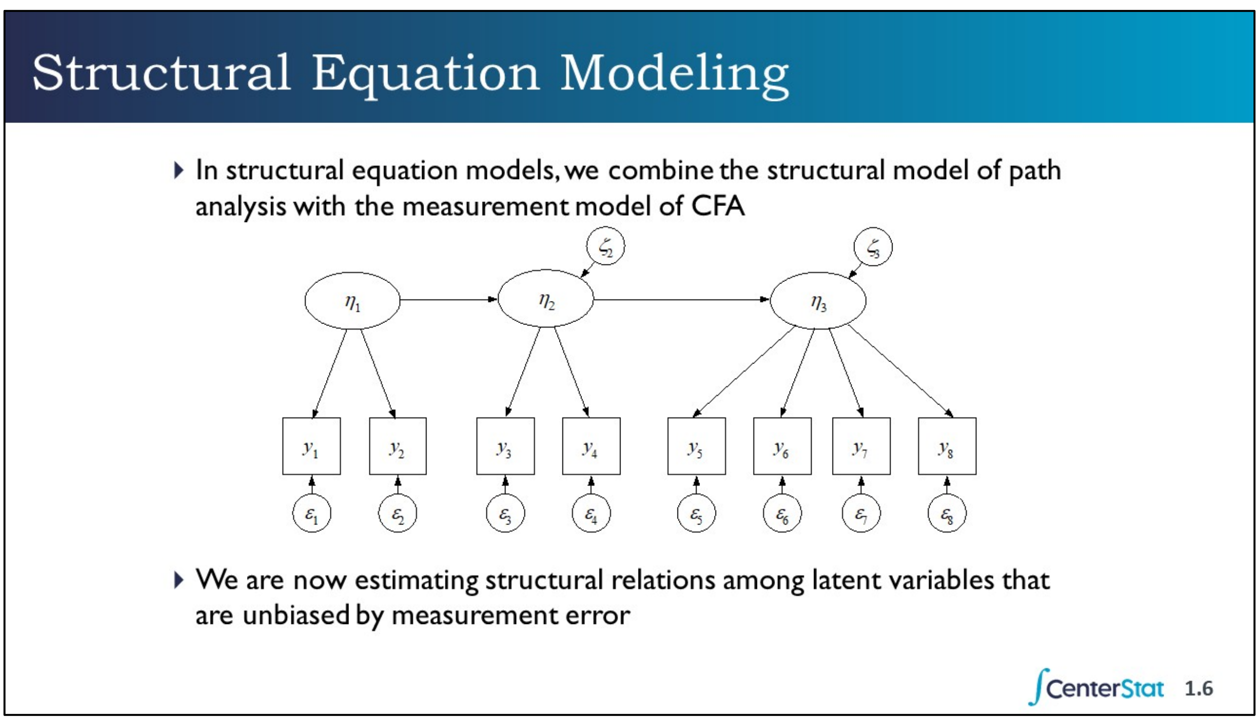


- Here, we have a relatively simple path analysis model where the effect of x is mediated by y_1.
- Although widely used, we have shown that fitting this model will almost certainly result in biased effect estimates due to failure to account for measurement error in the observed variables

CenterStat 1.4



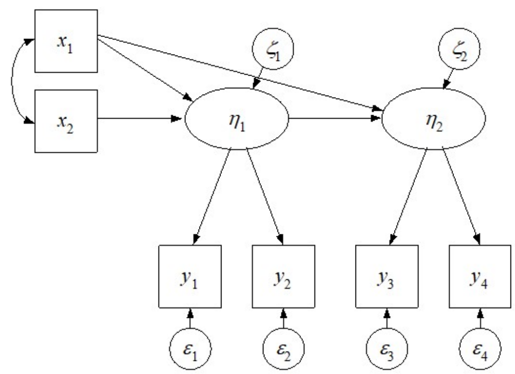
Here the covariance matrix among the factors is saturated, including non-directional associations among all latent variables.



Here a causal structure has been applied to the covariance matrix among the factors

Fixed-x, Exogenous predictors

- We can also include manifest fixed regressors, or exogenous predictors, in our models, as in the model below



CenterStat 1.7

Fixed-x, Exogenous predictors

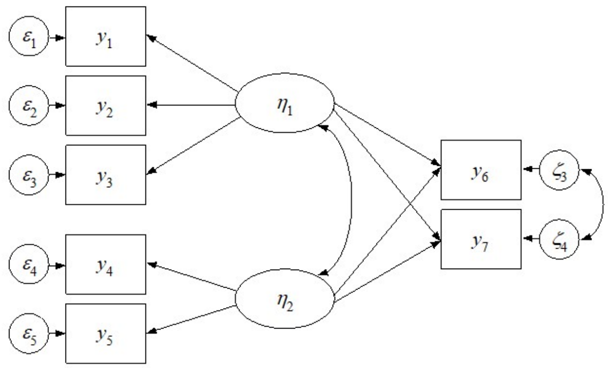
- As with the simultaneous equation model, we make no distributional assumptions about the exogenous predictors
- Exogenous predictors can even be binary
  - Gender
  - Child of an Alcoholic
- As in simultaneous equation models, however, exogenous predictors are still assumed to be measured without error. If this is not true, effect estimates may be biased.
  - Preferable to use multiple indicator latent factors as predictors whenever measurement error is likely

CenterStat 1.8

As with path analysis / simultaneous equation models, we can include nominal exogenous predictors via coding variables.

Models with Manifest Outcomes

- In some models, we may also have manifest variables as outcome variables, as in the model shown below.



CenterStat 1.9

Although ε and ζ both denote residuals in observed measures, these are labeled differently simply as a function of the role played in the model: ε denotes the residual of a measured indicator on a latent factor and ζ denotes the disturbance of an endogenous manifest variable. These only differ in notation.

Models with Manifest Outcomes

- Like simultaneous equation models, manifest outcomes are assumed to be free of measurement error
- Unlike measurement error in predictors, measurement error in outcomes does not bias structural paths
- But it does cause inflation of *standard errors* of structural paths thus resulting in lower power
- Thus, given sufficient sample size, still best to use full SEM including multiple indicator latent outcomes instead of manifest outcomes when measurement error is likely

CenterStat 1.10

Summary

- Structural equation models represent natural combination of simultaneous equation models with confirmatory factor analysis
- Full SEM very flexible, permitting complex structural models involving both observed and latent predictors and outcomes
- Latent factors permits estimation of structural effects without bias or standard error inflation due to measurement error
- We will now see that fitting SEMs is also a straightforward extension of procedures we have covered for simultaneous equation models and CFA

CenterStat 1.11

5.2 Fitting Structural Equation Models

Objectives

- Show specification of SEMs in path diagram and matrix form
- Describe model identification rules
- Describe process of model estimation, evaluation and re-specification

CenterStat 1.13

Fitting SEMs

- As with simultaneous equation models and CFA, model fitting involves the following steps
  - Specification
  - Identification
  - Estimation
  - Evaluation
  - Potential re-specification
  - Interpretation
- Fortunately, can draw on knowledge base we have built for simultaneous equation models and CFA to develop general principles for full SEM with latent variables

CenterStat 1.14

Model Specification

- As with prior models, the initial specification of the model should be heavily guided by theory.
  - Which variables are predictors, mediators, and outcomes?
  - Which variables have causal effects, and which are merely associated?
- The model can be specified graphically via a path diagram, or in equations using an expanded set of model matrices.
  - Our model matrices now combine matrices used previously with simultaneous equation models and matrices used with CFA.
- The full SEM consists of two parts:
  - The measurement model (like CFA)
  - The structural model (like path analysis)

CenterStat 1.15

Model Specification

- The *measurement model* is used to define the relationships of the indicator variables to the latent variables:

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i,$$

where

$$VAR(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta}$$

- This is identical to the measurement model of CFA

CenterStat 1.16

Model Specification

- The *structural model* is used to define the relationships among the latent variables and between the latent variables and any fixed-x exogenous predictors:

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i, \qquad \text{where} \qquad VAR(\boldsymbol{\zeta}_i) = \boldsymbol{\Psi}$$

  This is identical to the structural model from path analysis, but latent variables have replaced observed variables as outcomes

- Compare to

  Regression: $y_i = \alpha + \boldsymbol{\gamma}'\mathbf{x}_i + \zeta_i$

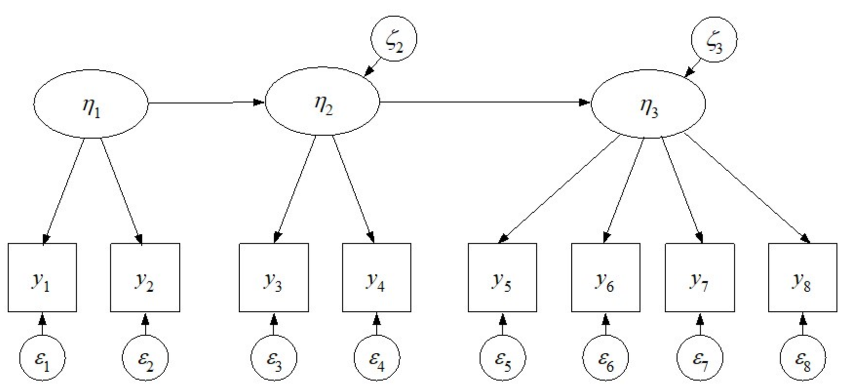
  Path Analysis: $\mathbf{y}_i = \boldsymbol{\alpha} + \mathbf{B}\mathbf{y}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i$

CenterStat 1.17

Notice that the structural model is the same as path analysis, except that the latent variables defined by the measurement model have replaced the observed variables.

Model Specification: Example

- Consider specification of the following model



CenterStat 1.18

Model Specification: Example

- Using scaling indicators to set the metric of the latent variables, the *measurement model* is

$$\begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \\ y_{5i} \\ y_{6i} \\ y_{7i} \\ y_{8i} \end{bmatrix} = \begin{bmatrix} 0 \\ \nu_2 \\ 0 \\ \nu_4 \\ 0 \\ \nu_6 \\ \nu_7 \\ \nu_8 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{63} \\ 0 & 0 & \lambda_{73} \\ 0 & 0 & \lambda_{83} \end{bmatrix} \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ \eta_{3i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \\ \varepsilon_{5i} \\ \varepsilon_{6i} \\ \varepsilon_{7i} \\ \varepsilon_{8i} \end{bmatrix}$$

where $VAR(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta} = DIAG(\theta_{11}, \theta_{22}, \cdots, \theta_{88})$

CenterStat 1.19

We'll talk more about scaling latent variables in SEM shortly. For now, we just use scaling items.

Model Specification: Example

- The *structural model* is specified as

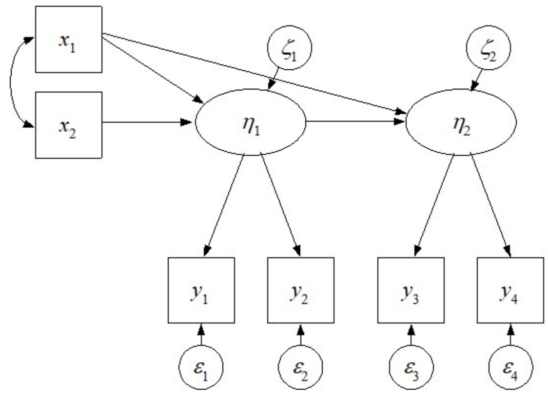
$$\begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ \eta_{3i} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ 0 & \beta_{32} & 0 \end{bmatrix} \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ \eta_{3i} \end{bmatrix} + \begin{bmatrix} \zeta_{1i} \\ \zeta_{2i} \\ \zeta_{3i} \end{bmatrix}$$

where $VAR(\boldsymbol{\zeta}_i) = \boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & & \\ 0 & \psi_{22} & \\ 0 & 0 & \psi_{33} \end{bmatrix}$

CenterStat 1.20

Model Specification: Example 2

- Let us now consider the example model with exogenous x's:



CenterStat 1.21

Model Specification: Example 2

- Using scaling indicators to set the metric of the latent variables, the *measurement model* is

$$\begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \end{bmatrix} = \begin{bmatrix} 0 \\ \nu_2 \\ 0 \\ \nu_4 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{bmatrix} \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix}$$

where $VAR(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta} = DIAG\left(\theta_{11}, \theta_{22}, \theta_{33}, \theta_{44}\right)$

CenterStat 1.22

Model Specification: Example 2

- The *structural model* is specified as

$$\begin{bmatrix}\eta_{1i}\\ \eta_{2i}\end{bmatrix}=\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}+\begin{bmatrix}0 & 0\\ \beta_{21} & 0\end{bmatrix}\begin{bmatrix}\eta_{1i}\\ \eta_{2i}\end{bmatrix}+\begin{bmatrix}\gamma_{11} & \gamma_{12}\\ \gamma_{21} & 0\end{bmatrix}\begin{bmatrix}x_{1i}\\ x_{2i}\end{bmatrix}+\begin{bmatrix}\zeta_{1i}\\ \zeta_{2i}\end{bmatrix}$$

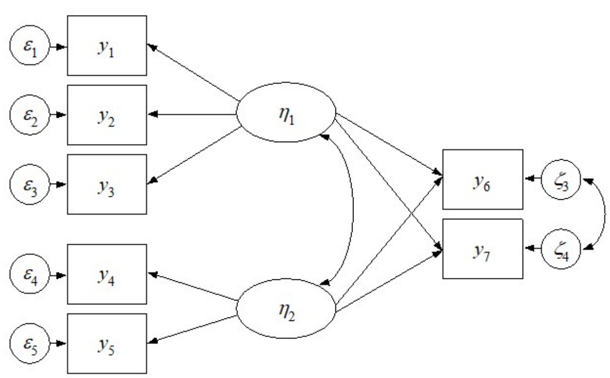
where

$$VAR(\boldsymbol{\zeta}_i)=\boldsymbol{\Psi}=\begin{bmatrix}\psi_{11} & \\ 0 & \psi_{22}\end{bmatrix}$$

CenterStat 1.23

Model Specification: Example 3

- Recall that our third example model includes two manifest outcomes:



CenterStat 1.24

Model Specification: Example 3

- Using scaling indicators to set the metric of the latent variables, the *measurement model* for the 5 indicator variables is

$$\begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \\ y_{5i} \end{bmatrix} = \begin{bmatrix} 0 \\ \nu_2 \\ \nu_3 \\ 0 \\ \nu_5 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \end{bmatrix} \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \\ \varepsilon_{5i} \end{bmatrix}$$

where $VAR(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta} = DIAG\left(\theta_{11}, \theta_{22}, \theta_{33}, \theta_{44}, \theta_{55}\right)$

CenterStat 1.25

The variables y_6 and y_7 do not appear here as they are not indicators of a latent factor. They will appear in the structural model instead as, essentially, “manifest factors”.

Model Specification: Example 3

- The *structural model* is specified as

$$\begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ y_{6i} \\ y_{7i} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 & 0 \\ \beta_{41} & \beta_{42} & 0 & 0 \end{bmatrix} \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ y_{6i} \\ y_{7i} \end{bmatrix} + \begin{bmatrix} \zeta_{1i} \\ \zeta_{2i} \\ \zeta_{3i} \\ \zeta_{4i} \end{bmatrix}$$

where $VAR(\boldsymbol{\zeta}_i) = \boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & & & \\ \psi_{21} & \psi_{22} & & \\ 0 & 0 & \psi_{33} & \\ 0 & 0 & \psi_{43} & \psi_{44} \end{bmatrix}$

CenterStat 1.26

We see y_6 and y_7 listed with the latent variables; the numbering on the coefficients and disturbances is a little unfortunate because it doesn’t match the numbering of the y variables. Such is life.

Model-Implied Moment Structure

- The measurement and structural models of full SEM are

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \qquad VAR(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta}$$

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i, \qquad VAR(\boldsymbol{\zeta}_i) = \boldsymbol{\Psi}$$

- These jointly imply specific mean and covariance structure
  - can use path tracing rules but matrix expressions are compact & elegant

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \left[\begin{array}{c} \boldsymbol{\nu} + \boldsymbol{\Lambda}(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\mu}_{\mathbf{x}}) \\ \hline \boldsymbol{\mu}_{\mathbf{x}} \end{array}\right]$$

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \left[\begin{array}{c|c} \boldsymbol{\Lambda}(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1'}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} & \\ \hline \boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B})^{-1'}\boldsymbol{\Lambda}' & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{array}\right]$$

CenterStat 1.27

There is a lot going on here, but the important thing to realize is that, just as in prior chapters, the hypothesized model implies a specific structure for the means and covariances. This could be accomplished using standard path tracing rules or the matrix algebra shown above.

Model Identification

- Like CFA, one requirement for identification is that scale of latent variables be defined through specific restrictions in model
- As before, one way to set the scale of latent factors is by fixing ν and λ to 0 and 1.0 for a scaling indicator, respectively
  - This is what we did when specifying the preceding models
- Alternatively, can set scale of latent variables by fixing α and ψ to 0 and 1.0 , respectively.
  - Unlike CFA, not always equivalent to scaling factor as a standard normal
  - If factor predicted by other variables in model, then standardizing factor *disturbances* and total variance of factor will exceed 1.0

CenterStat 1.28

One can put the factors on a standard normal scale post-estimation by requesting the standardized solution. We will see this in our example later.

Model Identification

- Aside from assigning a metric to the latent variables, other restrictions are also necessary to identify the model.
  - e.g., common to assume simple structure and local independence of items (but not required)
- Model identification can be determined by
  - Matrix algebra (hard)
  - *t*-rule (necessary but not sufficient)
  - Two-step rule (sufficient but not necessary)

CenterStat 1.29

t-Rule

- As before, *t*-rule simply stipulates that there should be more observed means and (co)variances than estimated parameters.
- More formally, if t is number of freely estimated parameters in θ then a necessary (but not sufficient) condition for identification is

$$t \leq (p+q)(p+q+1)/2+(p+q)$$

where p and q are the number of y and x variables, respectively
  - Because μ_x and Σ_{xx} are always saturated, some software programs do not count the elements in these matrices as parameters to be estimated
  - In assessing the *t*-rule, however, one should include unique elements in μ_x and Σ_{xx} when counting the number of parameters t.

CenterStat 1.30

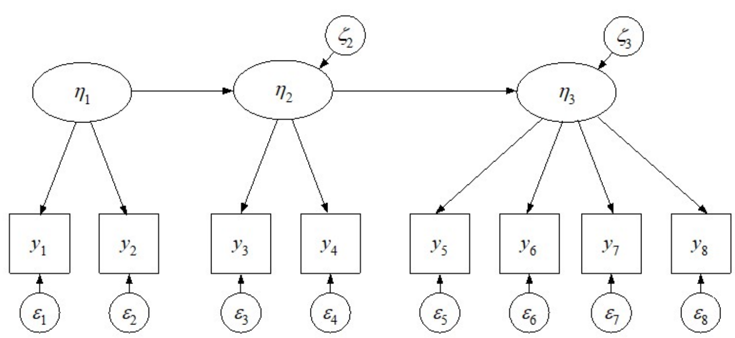
Two-Step Rule

- Step 1: Respecify model as CFA with all possible associations between factors
  - determine identification of measurement model as would do for a CFA
- Step 2: Treat all latent variables within structural model as if they were observed variables
  - determine identification of structural model as would do for path analysis model for observed variables
- Passing the two-step rule is sufficient to ensure model identification, but not necessary (i.e., some identified models fail the two-step rule)
- *Note two-step rule is a thought experiment; separate models are not actually estimated to sample data*

CenterStat 1.31

Example of Two-Step Rule

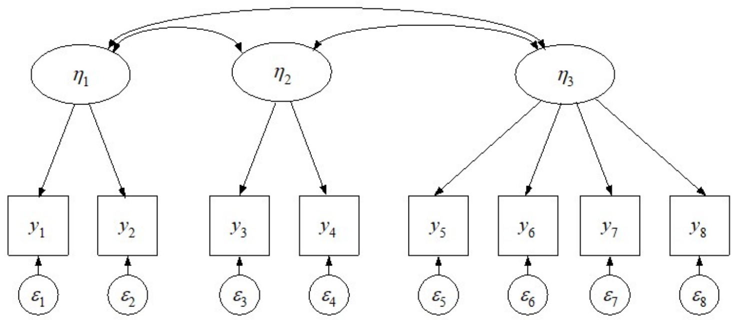
- Consider the following model



CenterStat 1.32

Example of Two-Step Rule

- The two-step rule involves first respecifying the model as a CFA with all possible associations between factors

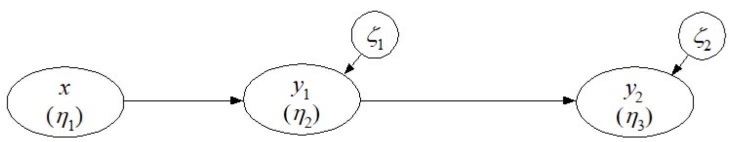


- This model is identified by the two-indicator rule.

CenterStat 1.33

Example of Two-Step Rule

- Second, we evaluate the structural model as we would a path analysis model for observed variables



- The structural model is identified by the recursive rule (the $\mathbf{B}$ matrix can be written as a lower triangular matrix and the $\boldsymbol{\Psi}$ matrix is diagonal)

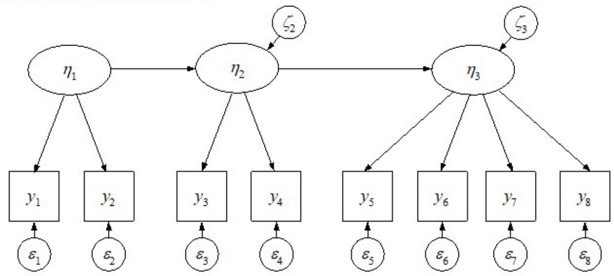
$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \qquad \boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \\ 0 & \psi_{22} \end{bmatrix}$$

CenterStat 1.34

Note that the first latent variable is now treated as if it were a fixed-x exogenous predictor, so the $\mathbf{B}$ and $\boldsymbol{\Psi}$ matrices are only 2 x 2 for the simultaneous equation model (whereas they are 3 x 3 for the SEM as actually fit to the data).

Example of Two-Step Rule

- Because this model satisfies the two-step rule, that is sufficient to say that the model is identified



- Empirical under-identification is still possible (e.g., if certain structural paths or factor loadings approach values of zero)

CenterStat 1.35

Model Estimation

- Given identification of model, we can now turn to estimation
- As with simultaneous equation models and CFA, multiple estimators exist for SEMs
  - Maximum likelihood
  - Ordinary least squares
  - Generalized least squares
  - Weighted least squares
- As before, we will concentrate on maximum likelihood
  - The only significance difference from path analysis and CFA is in the computation of the model-implied means and covariances

CenterStat 1.36

Evaluation of Parameter Estimates

- Interpretation of the measurement model is like CFA, and interpretation of the structural model is like path analysis
- Can again compute standardized parameter estimates to aid evaluation.
  - For factor loadings and causal structural paths this is akin to computing standardized regression coefficients
  - For covariance parameters, rescaling to correlation metric
  - For indicator residual variances, rescaled to variance unexplained
- Can compute communality, or h^2, for indicators
- Can compute variance explained, or R^2, for latent or manifest outcomes

CenterStat 1.37

Model Respecification

- When the model fails to fit, modification indices may suggest avenues for improvement.
- As always, must treat these with skepticism, as use of MIs is exploratory and fit can be improved by capitalizing on chance.
- MIs indicate expected improvement in fit for freeing a single parameter at a time, as opposed to larger scale model misspecifications.

CenterStat 1.38

Model Respecification

- Another strategy for locating model misfit is to isolate problems to the structural or measurement model.
  - Useful only if the structural model is not saturated.
  - Similar in concept to the two-step rule for identification.
- Model χ^2 can be decomposed into two additive components
  - one due to misfit of measurement model
  - one due to misfit of structural model
- This can help isolate potential source of model misfit

CenterStat 1.39

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411-423.

Model Respecification

- The part of the model χ^2 due to misfit in the measurement model can be obtained by refitting the SEM as a CFA with a freely estimated ψ matrix.
  - χ^2 obtained for this model is the misfit due to the measurement model where poor fit indicates a problem with the measurement model.
- The difference in χ^2 between the CFA and original SEM is the misfit due to the structural model
- Evaluate using an LRT given these are nested models
- Significant LRT indicates a problem with the structural model

CenterStat 1.40

Summary

- General SEM logical intersection of path analysis and CFA
- Specification, identification, and process of fitting and interpreting SEMs all follow straightforwardly from procedures for path analysis and CFA
- Can use two-step rule to aid in identification of complex SEMs
- Overall, there are few new complexities in the SEM that we have not already encountered

CenterStat 1.41

5.3 Example SEM

Objectives

- Provide a real-data example of the process of fitting, evaluating, re-specifying, and interpreting an SEM

CenterStat 1.43

Example: Posttraumatic Growth

- Senol-Durak & Ayvasik (2010) wished to determine what factors relate to posttraumatic growth in spouses of myocardial infarction patients.
- Sample size of 132.
- Hypotheses were that
  - Environmental Resources (ER, i.e., social support) and Individual Resources (IR) influence PostTraumatic Growth (PTG)
  - ER and IR have direct effects on PTG.
  - ER and IR also have indirect effects on PTG.
    - ER and IR influence Event-Related Factors (ERF)
    - ERF influences Cognitive Process Coping (Coping)
    - Cope influences PTG

CenterStat 1.44

It is worth noting that the sample size is rather low, especially for the complexity of the fitted model, but this is not uncommon for published applications of SEM. We use this as an example both to demonstrate a fully latent SEM but also to highlight certain alternative modeling decisions that might be considered compared to those presented in the published manuscript.

The full reference is:

Senol-Durak, E. & Ayvasik, H.B. (2010). Factors associated with posttraumatic growth among the spouses of myocardial infarction patients. *Journal of Health Psychology*, *15*, 5–95.

Abstract from manuscript: To clarify the rationale behind Posttraumatic Growth (PTG), a model by Schaefer and Moos describes the relative contribution of environmental resources, individual resources, event related factors, cognitive processing and coping (CPC) on PTG. In the present study, this model was tested with the spouses of myocardial infarction patients with data from various hospitals in Turkey. A structural equation model revealed that neither individual nor environmental resources had indirect effects on PTG through the effect of event-related factors and CPC, while they showed direct effects on PTG. The findings were discussed in the context of the theoretical model.

Example: Multiple Indicator Latent Factors

- Environmental Resources
  - Three indicators: Social support from family (*fa*), friends (*fr*), significant others (*si*)
- Individual Resources
  - Five indicators: psychological hardiness indicators: commitment (*comt*), control (*con*), challenge (*cha*); self esteem scale scores (*es*), and locus of control scale scores (*lo*)
- Event-Related Factors
  - Three indicators: subjective evaluations of prognosis (*pro*), threat to future health (*th*), and time since diagnosis (*time*)

CenterStat 1.45

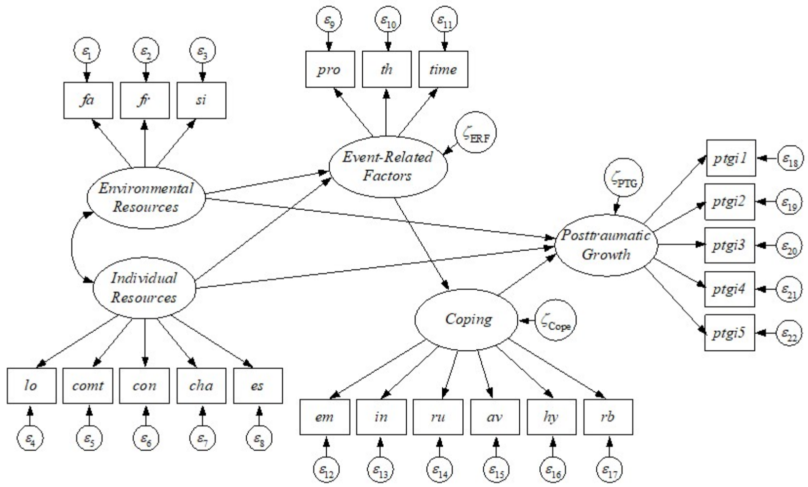
One could probably make a case that the indicators for Event-Related Factors should be causal indicators rather than effect indicators. The authors used a reflective measurement model for all factors, however, and we remain consistent with their analysis.

Example: Multiple Indicator Latent Factors

- Cognitive process coping
  - Six indicators: emotion focused coping (*em*), indirect coping (*in*), rumination (*ru*), avoidance (*av*), hyper-vigilance (*hy*), and religious belief (*rb*)
- Posttraumatic growth
  - Five indicators, subscale scores from post-traumatic growth inventory: improved relationship (*ptgi1*), new possibilities for one's life (*ptgi2*), greater appreciation of life (*ptgi3*), greater sense of personal strength (*ptgi4*), spiritual development (*ptgi5*).

CenterStat 1.46

Example: The Model



CenterStat 1.47

Example: The Model

- Like many SEM's, the model is rather large and the matrix representation of the model is unwieldy.
- The path diagram, however, nicely conveys the proposed structure and is sufficient to infer the matrix representation of the model.
  - We thus forgo showing the matrix representation here – this is provided in the text

CenterStat 1.48

The measurement model of the Post Traumatic Growth Model is:

$$\begin{bmatrix} fa_i \\ fr_i \\ si_i \\ lo_i \\ comt_i \\ con_i \\ cha_i \\ es_i \\ pro_i \\ th_i \\ time_i \\ em_i \\ in_i \\ ru_i \\ av_i \\ hy_i \\ rb_i \\ ptgi_{1i} \\ ptgi_{2i} \\ ptgi_{3i} \\ ptgi_{4i} \\ ptgi_{5i} \end{bmatrix} = \begin{bmatrix} \nu_{1i} \\ \nu_{2i} \\ \nu_{3i} \\ \nu_{4i} \\ \nu_{5i} \\ \nu_{6i} \\ \nu_{7i} \\ \nu_{8i} \\ \nu_{9i} \\ \nu_{10i} \\ \nu_{11i} \\ \nu_{12i} \\ \nu_{13i} \\ \nu_{14i} \\ \nu_{15i} \\ \nu_{16i} \\ \nu_{17i} \\ \nu_{18i} \\ \nu_{19i} \\ \nu_{20i} \\ \nu_{21i} \\ \nu_{22i} \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 & 0 \\ \lambda_{31} & 0 & 0 & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 & 0 \\ 0 & \lambda_{52} & 0 & 0 & 0 \\ 0 & \lambda_{62} & 0 & 0 & 0 \\ 0 & \lambda_{72} & 0 & 0 & 0 \\ 0 & \lambda_{82} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{93} & 0 & 0 \\ 0 & 0 & \lambda_{10,3} & 0 & 0 \\ 0 & 0 & \lambda_{11,3} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{12,4} & 0 \\ 0 & 0 & 0 & \lambda_{13,4} & 0 \\ 0 & 0 & 0 & \lambda_{14,4} & 0 \\ 0 & 0 & 0 & \lambda_{15,4} & 0 \\ 0 & 0 & 0 & \lambda_{16,4} & 0 \\ 0 & 0 & 0 & \lambda_{17,4} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{18,5} \\ 0 & 0 & 0 & 0 & \lambda_{19,5} \\ 0 & 0 & 0 & 0 & \lambda_{20,5} \\ 0 & 0 & 0 & 0 & \lambda_{21,5} \\ 0 & 0 & 0 & 0 & \lambda_{22,5} \end{bmatrix} \begin{bmatrix} \eta_{ERi} \\ \eta_{IRi} \\ \eta_{ERFi} \\ \eta_{COPE_i} \\ \eta_{PTGi} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \\ \varepsilon_{5i} \\ \varepsilon_{6i} \\ \varepsilon_{7i} \\ \varepsilon_{8i} \\ \varepsilon_{9i} \\ \varepsilon_{10i} \\ \varepsilon_{11i} \\ \varepsilon_{12i} \\ \varepsilon_{13i} \\ \varepsilon_{14i} \\ \varepsilon_{15i} \\ \varepsilon_{16i} \\ \varepsilon_{17i} \\ \varepsilon_{18i} \\ \varepsilon_{19i} \\ \varepsilon_{20i} \\ \varepsilon_{21i} \\ \varepsilon_{22i} \end{bmatrix}$$

where

$$\mathbf{\Theta} = DIAG\left(\theta_{11}, \theta_{22}, \ldots, \theta_{22,22}\right)$$

indicating that none of the residual variances are correlated.

The latent variable model is:

$$\begin{bmatrix} \eta_{ERi} \\ \eta_{IRi} \\ \eta_{ERFi} \\ \eta_{COPE_i} \\ \eta_{PTGi} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 & 0 & 0 \\ 0 & 0 & \beta_{43} & 0 & 0 \\ \beta_{51} & \beta_{52} & 0 & \beta_{54} & 0 \end{bmatrix} \begin{bmatrix} \eta_{ERi} \\ \eta_{IRi} \\ \eta_{ERFi} \\ \eta_{COPE_i} \\ \eta_{PTGi} \end{bmatrix} + \begin{bmatrix} \zeta_{ERi} \\ \zeta_{IRi} \\ \zeta_{ERFi} \\ \zeta_{COPE_i} \\ \zeta_{PTGi} \end{bmatrix}$$

where

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & & & & \\ \psi_{21} & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

indicating that only the first two latent factors are correlated.

Note that the means/intercepts and (residual) variances of the factors have been fixed to 0 and 1, respectively to scale the latent variables.

Example: Identification

- We shall scale the latent factors/residuals by setting their means and variances to 0 and 1, respectively
- Let us now see if the model passes the t-rule.
- The number of estimated parameters is 22 factor loadings + 22 intercepts + 22 residual variances + 1 factor correlation + 6 factor regression slopes = 73
- The number of observed variables is 22, so there are

$$22(22+1)/2+22=275$$

unique observed first- and second moments from which to fit the model

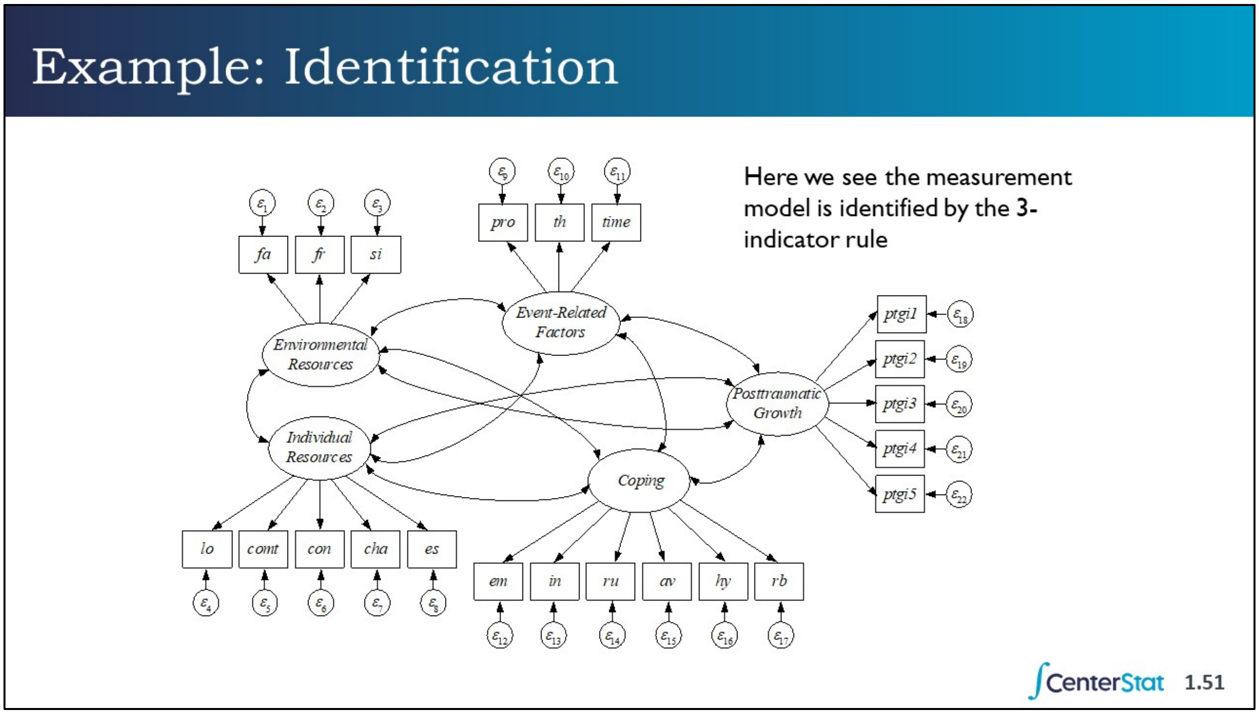
- The model thus passes the t-rule and will have $275-73=202$ degrees of freedom.

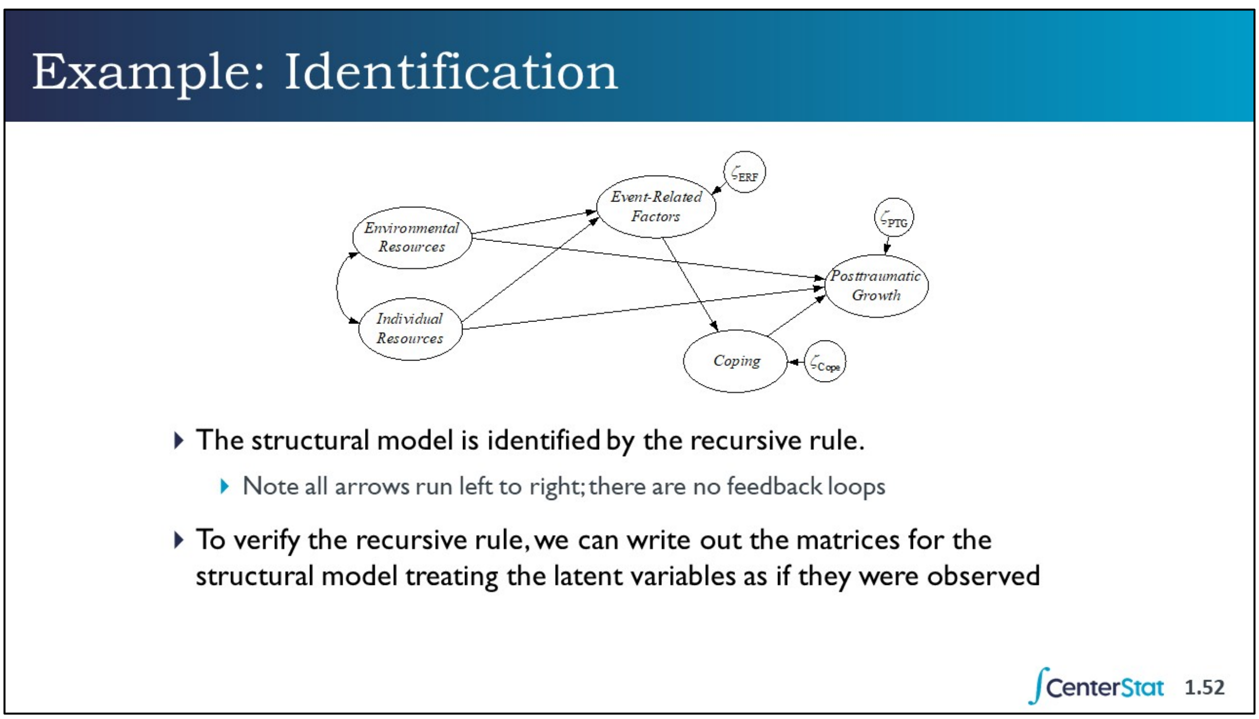
CenterStat 1.49

Example: Identification

- The t-rule is necessary but not sufficient.
- It quickly shows us that our model *may* be identified.
- Now we can turn to the two-step rule to further evaluate the identification of the model.
- Recall that the first step is to reparameterize the model as a confirmatory factor model with all possible associations among the factors to verify identification of the measurement model.
- The second step is to examine the structural model as if it were a structural model for observed variables to verify its identification.

CenterStat 1.50





No causal feedback loops or correlated disturbances are present here.

Example: Identification

- If the latent variables were observed, the simultaneous equation model would be

$$\begin{bmatrix} ERF_i \\ Coping_i \\ PTG_i \end{bmatrix} = \begin{bmatrix} \alpha_{ERF} \\ \alpha_{COPE} \\ \alpha_{PTG} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ 0 & \beta_{32} & 0 \end{bmatrix} \begin{bmatrix} ERF_i \\ Coping_i \\ PTG_i \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & 0 \\ \gamma_{31} & \gamma_{32} \end{bmatrix} \begin{bmatrix} ER_i \\ IR_i \end{bmatrix} + \begin{bmatrix} \zeta_{ERFi} \\ \zeta_{COPEi} \\ \zeta_{PTGi} \end{bmatrix}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{ERF} & & \\ 0 & \psi_{COPE} & \\ 0 & 0 & \psi_{PTG} \end{bmatrix}$$

- The recursive model is satisfied because the $\mathbf{B}$ matrix is lower triangular and the ψ matrix is diagonal.

CenterStat 1.53

Note that the Environmental and Individual Resources factors are treated as if they were exogenous x variables in this step.

Example: Fitting the Model

- Since the model is identified, we can proceed to fit it to the data
  - Caution: $N = 132$ is rather low for a model of this complexity.
- Senol-Durak & Ayvasik (2010) provided their correlation matrix, standard deviations, and means. We can therefore replicate their analysis via the summary data version of ML estimation.
- We obtain the following fit measures
  - $\chi^2(202) = 350.00, p < .0001$
  - RMSEA = .075; CI90 = (.061, .087)
  - CFI = .848; TLI = .826
  - SRMR = .101
- These fit statistics suggest mediocre fit of the model to the data.

CenterStat 1.54

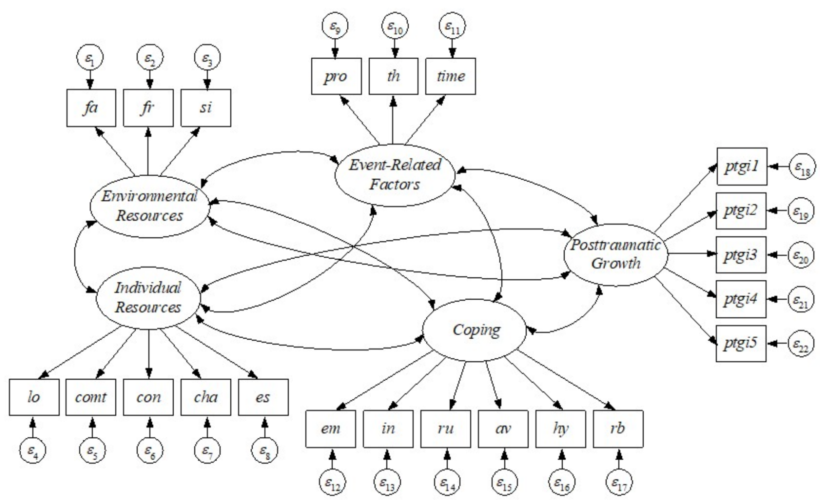
Our fit measures do not perfectly reproduce the results reported in Senol-Durak & Ayvasik (2010) but this may be due to rounding error in the correlation matrix.

Example: Model Respecification

- The largest MIs for the model are associated with residual covariances and cross-loadings. This suggests a problem with the measurement model.
- To more formally locate the source of misfit, we can respecify the model as a CFA model.
  - Poor absolute fit indicative of misfit of measurement model.
  - Significant chi-square difference test with original SEM indicative of misfit of structural model.

CenterStat 1.55

Fitting as CFA Model



CenterStat 1.56

Fitting as CFA Model

- Fitting the CFA model to test the measurement model, we obtain:
  - $\chi^2(199) = 348.64, p < .0001$
  - CFI = .846; TLI = .822
  - RMSEA = .075; CI90 = (.062, .088)
  - SRMR = .099
- Conducting the chi-square difference test between the original SEM and the CFA to test the structural model, we obtain:
  - $\Delta\chi^2(3) = 350.00 - 348.64 = 1.36, p = .71.$
- In line with the MIs, these results show that the mediocre fit is a reflection of problems with the measurement model.

CenterStat 1.57

Respecifying the Model

- Closer examination of the indicators suggests one possible respecification of the model.
- For some factors, subsets of indicators are composed of subscales from a single scale
  - Individual Resources factor includes three Psychological Hardiness subscale scores as indicators (*comt*, *con*, *cha*), but also two indicators from independent scales (*lo* and *es*).
  - Coping factor includes two indicators from the Ways of Coping Inventory (*em* and *in*), three indicators from the Impact of Event Scale (*ru*, *av*, & *hy*), and a religious beliefs score from another scale (*rb*).

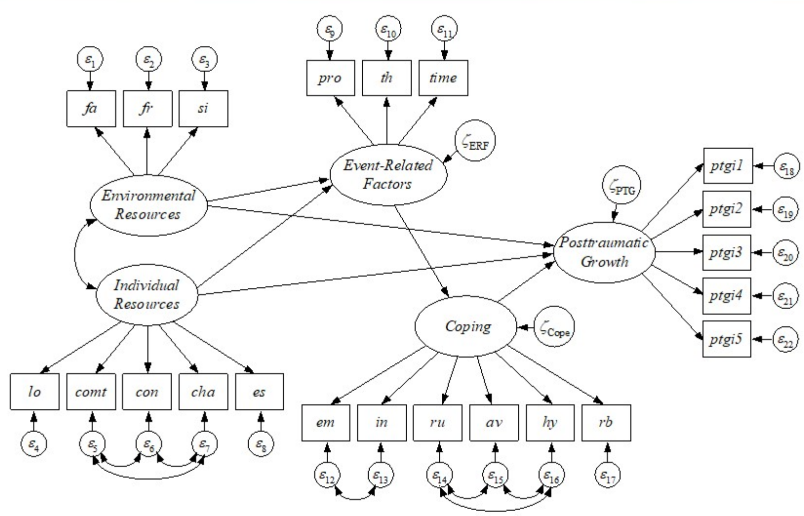
CenterStat 1.58

Respecifying the Model

- We will use a correlated uniqueness model to account for possible scale factors
  - Indicators from the same scale likely to be more highly related than indicators from different scales, above and beyond common factor
- Respecificying the model in this way is theoretically motivated but is also consistent with the MIs
  - Four of the five highest MIs are for residual covariances between subscale scores from the same scale:
    - $\theta_{hy,ru}$ **MI = 32.258**
    - $\theta_{cha,comt}$ **MI = 27.845**
    - $\theta_{in,em}$ **MI = 21.340**
    - $\theta_{in,fa}$ **MI = 14.250**
    - $\theta_{cha,com}$ **MI = 12.345**

CenterStat 1.59

Revised Model



CenterStat 1.60

Although we only identified five large MIs that are consistent with theory, we are including seven residual covariances. The reason is that the correlated residuals all relate to items that had been drawn from existing subscales, so we are generalizing this rule to all relevant items.

Revised Model

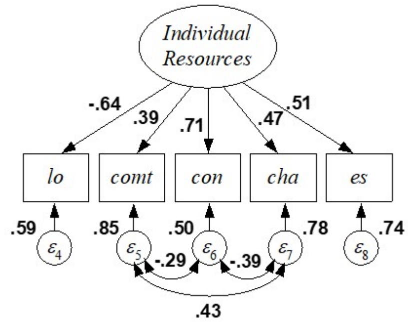
- Fit statistics for the correlated uniqueness model:
  - $\chi^2(195) = 270.28, p = .0003$
  - $\text{RMSEA} = .054; \text{CI90} = (.037, .069)$
  - $\text{CFI} = .923; \text{TLI} = .908$
  - $\text{SRMR} = .090$
- CU model fits significantly better than the original model:
  - $\Delta\chi^2(7) = 350.00 - 270.28 = 79.72, p < .0001.$
- Overall fit is borderline, especially given low power with low N
  - Some large MIs still, but suggested paths not consistent with theory
  - Especially at low N, suggested paths may reflect only chance fluctuations in data that would not replicate in another sample
- We will retain the revised, correlated uniqueness model and proceed to interpret the estimates.

CenterStat 1.61

In practice we would draw strong conclusions from these results given the poor overall fit. Also, our model differs slightly from the final published model due to alternative modeling decision points.

Interpretation

- Interpretation of the measurement model is done similarly to CFA. We will thus consider just the IR factor here as an example
  - Note that locus of control is a negative indicator of individual resources (but not clear from article how this was scored)
  - Control subscale of Psychological Hardiness Scale is negatively correlated with commitment and challenge subscales
    - unusual for a CU model & inconsistent with notion that residuals correlate due to a common "scale" factor

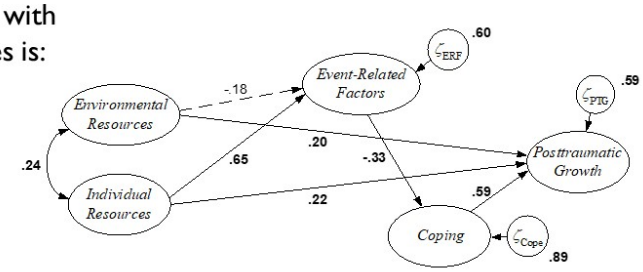


CenterStat 1.62

All reported values are standardized estimates. Although we consider only the Individual Resources factor here, the interpretation of the other factors would proceed similarly.

Interpretation

- The structural model with standardized estimates is:

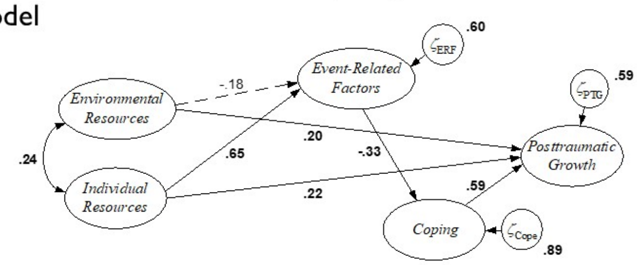


- Findings include:
  - Environmental resources (social support) and Individual Resources positively influence posttraumatic growth
  - Coping positively influences posttraumatic growth
  - Positive event related factors (better prognosis, less threat, more time from event) predict less coping

CenterStat 1.63

Total, Direct, and Indirect Effects

- Note, however, that there is a complex pattern of direct and indirect effects in the model



- To fully interpret the model, we need to evaluate total, direct and indirect effects
  - The computation of these effects and their standard errors mimics the procedures discussed earlier for path analysis

CenterStat 1.64

Effect Decomposition

- Let's first consider the effect of Event-Related Factors on Post-Traumatic Growth
- The effects of Event-Related Factors on Post-Traumatic Growth:
  - Total: $-.20, p < .05$
  - Direct: 0
  - Indirect: $-.20, p < .05$
- Interpretation: More positive event-related factors reduce the need for Coping. Since Coping positively impacts Post-Traumatic Growth, higher Event-Related Factors lead to less opportunity for Post-Traumatic Growth
  - Note direct effect of Event-Related Factors on Post-Traumatic Growth was constrained to be zero, so the total effect equals the indirect effect in this case.

CenterStat 1.65

Effect Decomposition

- Let us now consider the effects of Environmental Resources
- The effects of Environmental Resources on Post-Traumatic Growth:
  - Total: $.24, p < .05$
  - Direct: $.20, p < .05$
  - Indirect: $.04, ns$
- Environmental Resources do not significantly impact Event-Related Factors, and this is reflected in a non-significant indirect effect of Environmental Resources on Post-Traumatic Growth through Event-Related Factors and Coping.
- Environmental Resources do, however, have a significantly positive direct effect on Post-Traumatic Growth.

CenterStat 1.66

Effect Decomposition

- Last we can consider the effect decomposition for Individual Resources
- The effects of Individual Resources on Post-Traumatic Growth:
  - Total: $.09, ns$
  - Direct: $.22, p < .05$
  - Indirect: $-.13, p = .07$
- There are countervailing indirect and direct effects that result in a non-significant total effect.
  - Higher Individual Resources predict more positive Event-Related Factors (which predict less Coping and, in turn, less opportunity for Post-Traumatic Growth), leading to the negative indirect effect and null total effect

CenterStat 1.67

Summary

- The procedures we followed in fitting the SEM are very similar to those used in path analysis and CFA
- Like in path analysis, our models may include complex causal chains and mediation effects, but now including latent predictors, mediators and outcomes.
- The computation, testing, and interpretation of total, direct and indirect effects is similar to path analysis, and can aid in model interpretation.

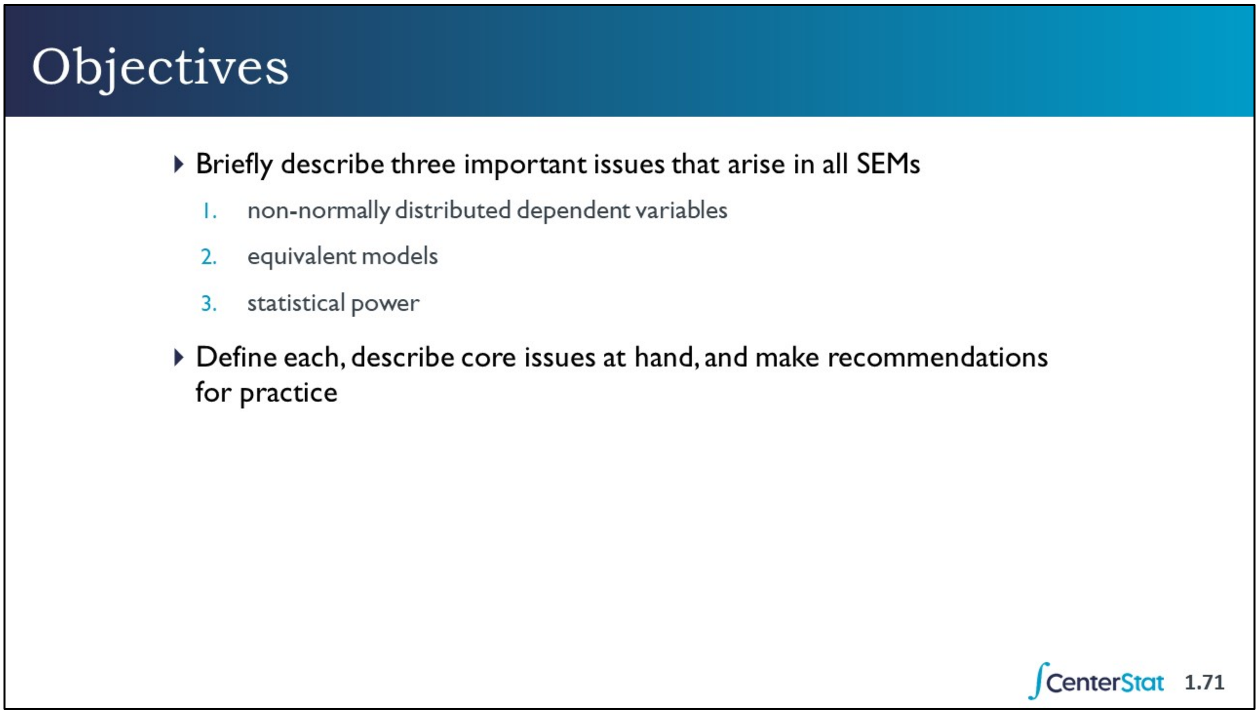
CenterStat 1.68

Chapter Summary

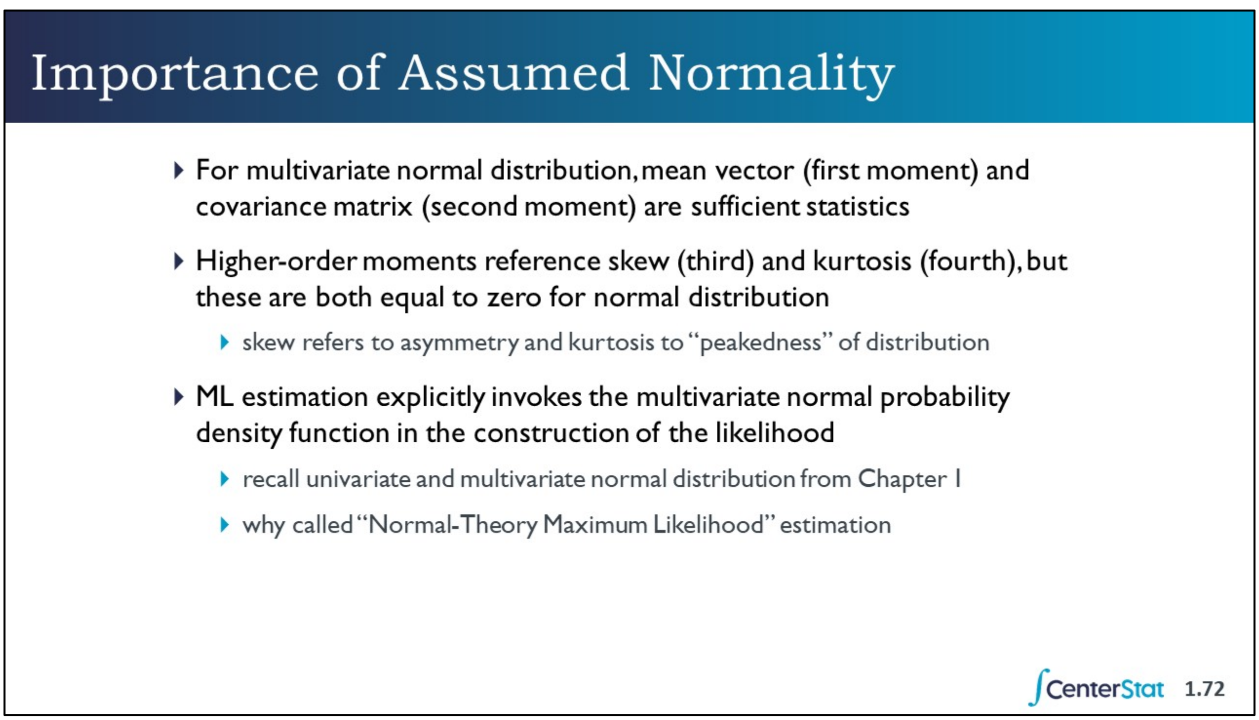
- Structural equation models with latent variables blend together the core features of path analysis and confirmatory factor analysis
- SEMs are composed of two parts
  - Measurement Model (like CFA)
  - Structural Model (like path analysis)
- Fitting and interpreting SEMs involves similar procedures to path analysis and CFA
  - Identification, estimation, evaluation, re-specification, interpretation
  - Interpretation aided by computation of total, direct, and indirect effects
- Relative to path analysis, SEM with latent variables provides the key advantage that the coefficients in the structural model are unbiased by measurement error

CenterStat 1.69

5.4 Self-Study: Additional Considerations



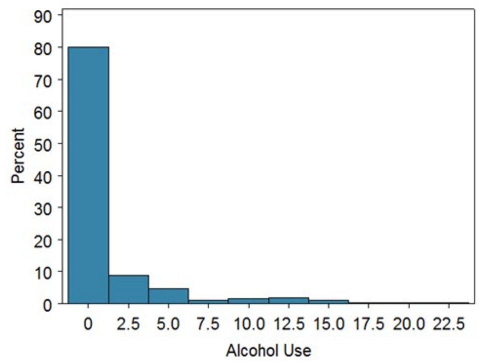
Although we review these issues in the chapter dedicated to the general SEM, normality, equivalent models, and power all equally apply to the path models and CFAs we described earlier.



Technically, kurtosis is equal to three for a normal distribution, but in nearly all applications this is re-scaled to equal zero to minimize confusion.

Non-Normal Outcomes

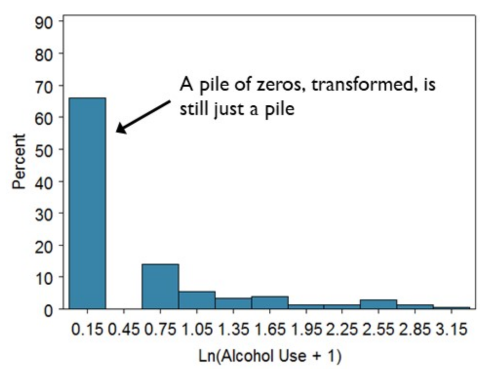
- Yet many measures in social sciences are not normally distributed
  - e.g., alcohol use as measured by mean of four items in 10-16 year olds



CenterStat 1.73

Transformations

- Sometimes nonlinear power transformations can be used to better meet the assumption of normality
  - but often transformations fail



CenterStat 1.74

Although the log transformation affected original values that were non-zero, those that were equal to zero remain clustered together. A classic work on nonlinear power transformations is:

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, 211-252.

Implications: Estimates

- Important to know implications of using normal-theory ML when observed measures are not normally distributed
  - remember assumption of normality only on *dependent* or *criterion* variables and no distributional assumptions are made about *independent* or *predictor* variables
  - predictor variables can be binary, ordinal, count, product terms, etc.
- ML *parameter estimates* are still consistent under non-normality as long as the kurtosis in the data is not excessive
  - with large N and no excessive kurtosis, still obtain good point estimates
- *Inferences* about point estimates are, however, biased

CenterStat 1.75

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62-83.

Implications: Inference

- "Naïve" ML-based SEs and chi-square test statistics will often be incorrect under non-normality
  - means and covariances no longer sufficient statistics because higher-order moments are needed to fully describe data, yet these are omitted
  - ML estimates are no longer asymptotically efficient, but SEs and test statistic are still *naïvely* computed as if they were
    - thus referred to as "naïve" because omitting relevant information
- Usually, naïve SEs are too small and naïve chi-squares are too big
  - confidence intervals around parameter estimates are too small
  - Type I error rates for parameter estimates are too large
  - models are over-rejected when using the chi-square LRT

CenterStat 1.76

West, S.G., Finch, J.F., & Curran, P.J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues and Applications*, (pp. 56-75). Newbury Park, CA: Sage.

Robust Standard Errors & Fit Statistics

- SEs for parameter estimates can be corrected for non-normality based on characteristics of the sample data
  - robust SEs also known as Satorra-Bentler or "sandwich estimator" SEs
  - terms often used interchangeably to reference the same thing
- Likelihood ratio test statistic can also be corrected for non-normality based on characteristics of the sample data
  - involves rescaling of test statistic to approximate appropriate chi-square
- Slight variations in the calculation of both robust SEs and corrected LRTs available in most major SEM packages

CenterStat 1.77

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity, 24*, 367-386.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507-514.

Practical Usage

- Simulations show these robust SEs and chi-square tests perform well under a variety of conditions
- But loss of power when using corrected versus naïve SEs and chi-square tests when data are approximately normal
- Good to examine the data for evidence of gross non-normality
  - but keep in mind univariate normality (e.g., in histograms) is not sufficient to indicate multivariate normality
  - when in doubt, do sensitivity analyses comparing naïve and robust results
- Finally, robust methods still assume *continuous* distributions
  - if distributions are binary or ordinal, need completely different approach

CenterStat 1.78

Equivalent Models

- Equivalent models defined as alternative model specifications that result in precisely the same model fit
  - assume models are fitted to same data using same measures and only the model specification is changed
- Despite identical fit, parameter estimates and standard errors can be radically different and lead to different theoretical conclusions
- For most path models, CFAs, and SEMs, can use "Lee-Hershberger replacing rules" to determine equivalent models
  - follow set of rules to organize variables in model into various "blocks" and then make modifications within a particular block
- These changes will result in a model that is mathematically identical to the original but has one or more parameters altered

CenterStat 1.79

Hershberger, S. L. (1994). The specification of equivalent models before the collection of data. In A. von Eye and C. C. Clogg (Eds.), *Latent Variables Analysis* (pp. 68–108). Thousand Oaks, CA: Sage.

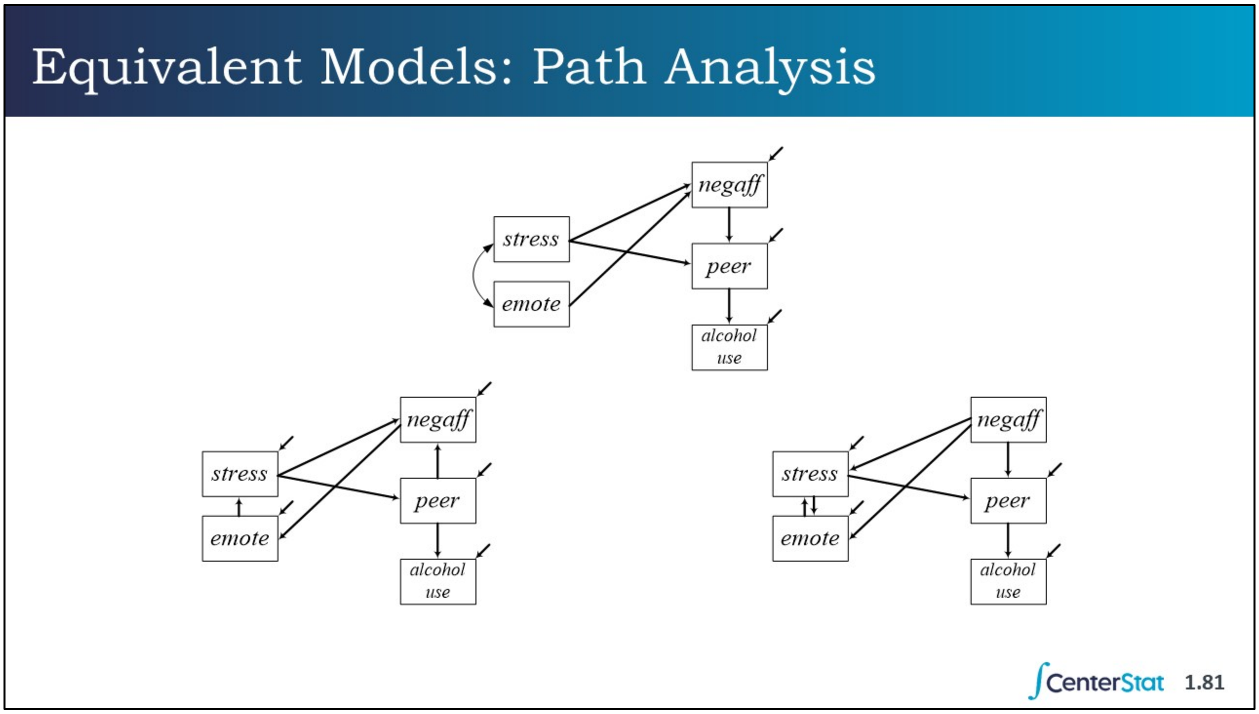
Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25*, 313–334.

Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling, 8*, 142-149.

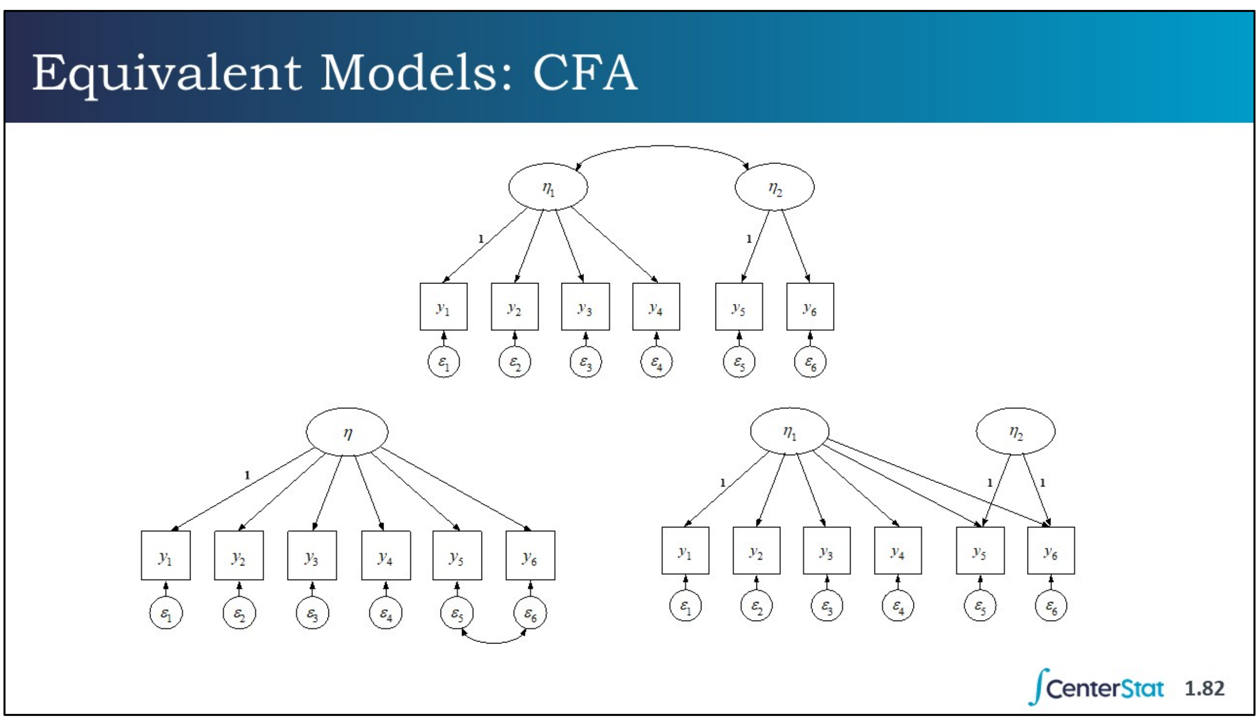
Equivalent Models

- For simple models, may be one or two equivalent models
- For more complex models, may be hundreds or even thousands of equivalent models
- Some equivalent models are logically implausible
  - e.g., reversing a regression to predict race
  - e.g., replacing a covariance with a regression predicting back in time
- But vast majority of equivalent models are typically plausible alternatives to varying degrees

CenterStat 1.80



These are all equivalent models for a version of the path model we explored in Chapter 3. Although all three models would obtain precisely equal fit, each model implies quite different causal mechanisms.

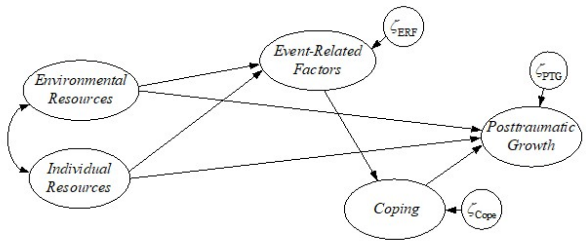


This example was drawn from

Kline, R.B. (2005). *Principles and Practice of Structural Equation Modeling*. Guilford Press. New York.

Equivalent Models: SEM Demonstration

- We more closely consider the SEM that we fit in PTG example:

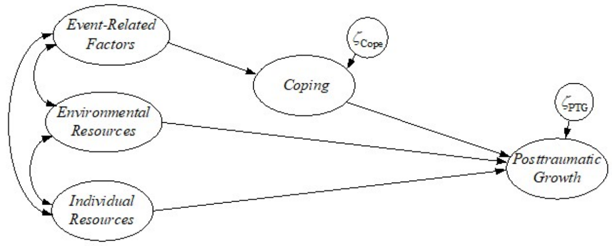


- Event-Related Factors constitute partial mediator of effects of Environmental Resources and Individual Resources, and Coping totally mediates effect of Event-Related Factors

CenterStat 1.83

Equivalent Models: SEM Demonstration

- Holding measurement model constant, there is an equivalent model that should be considered:

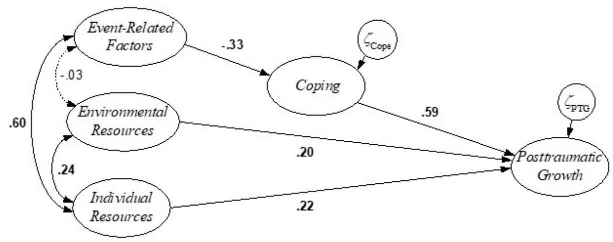


- This model is substantively quite different from original, because Event-Related Factors and Coping are not construed as partial mediators of effects of ER and IR

CenterStat 1.84

Equivalent Models: SEM Demonstration

- Re-estimating model and examining standardized estimates, this equivalent model may be more theoretically sensible



  - More positive Event-Related Factors lead to less need for coping and in turn less opportunity for post-traumatic growth
  - Environmental and Individual Resources promote post-traumatic growth

CenterStat 1.85

Equivalent Models: Recommendations

- Nothing can be done to "fix" issue
  - simply a characteristic of path models, CFAs, and SEMs
- Sharpest sword available to deal with this issue is theory
  - equivalent models are mathematically indistinguishable, so must rely on theory
- But often difficult to address equivalent models in meaningful way in published manuscript
  - may be dozens of theoretically plausible equivalent models
  - few editors will allow much space to be allocated to this topic
- Even if do not address in paper, good practice to be aware of issue and consider one or small number of equivalent models that fit equally yet may have different theoretical implications

CenterStat 1.86

Statistical Power

- Equivalent models are reparameterizations of a given model that all result in precisely the same degree of numerical fit
- A distinctly different question relates to *statistical power* for assessing fit
- Statistical power is the probability that the null hypothesis will be rejected if the null is false
  - or more colloquially, an effect will be found if an effect really exists
- But this issue is complicated by the fact that in SEM rejection of the null is taken as evidence of *lack of fit* of the tested model
  - so a false null indicates a misspecified (or incorrect) model
- So in SEM an "effect" is some degree of misspecification in the hypothesized model

CenterStat 1.87

An excellent recent overview of power in SEMs can be found in:

Lee. T., Cai, L., & MacCallum, R.C. (2012). Power Analysis for tests of structural equation models. In R. Hoyle, D. Kaplan, G. Marcoulides, & S. West (Eds.), *Handbook of Structural Equation Modeling* (pp.181-194), New York: Guilford.

Comparing Two Independent Means

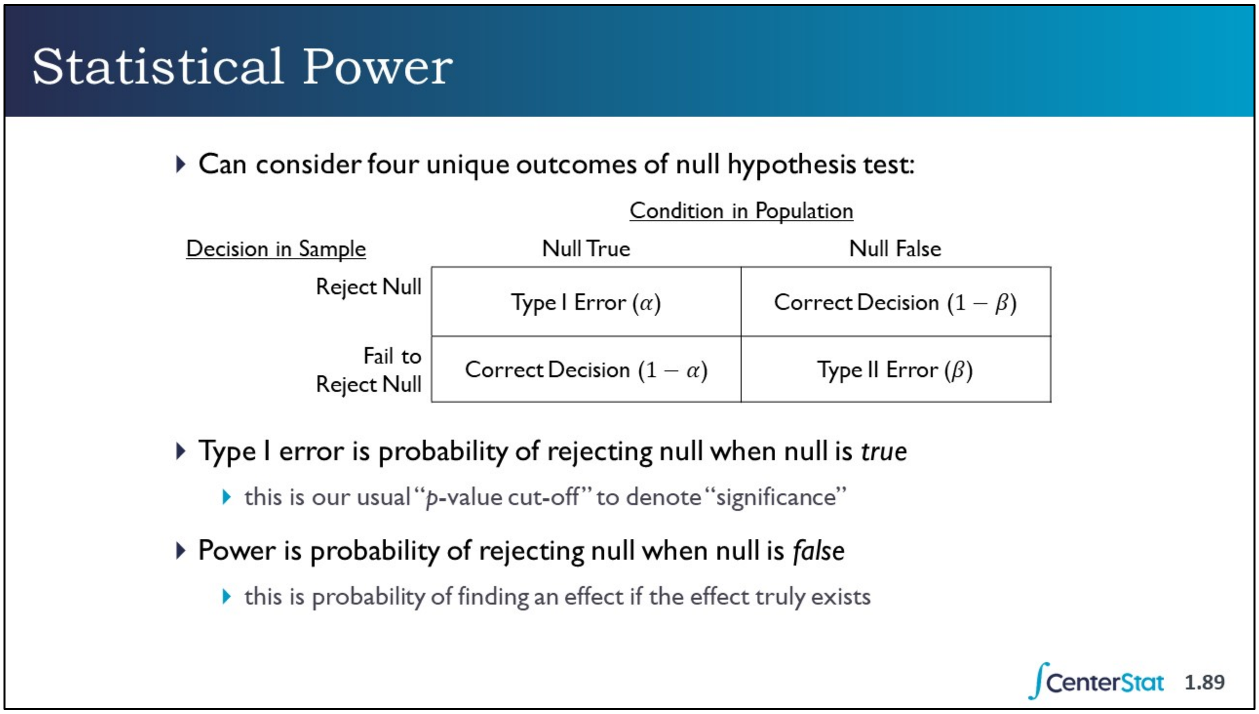
- As a simple starting point, say we are interested in testing gender differences in some measured outcome.
- Stated more formally, the null hypothesis states that the *population* means of depression for boys and girls are equal:

$$H_0 : \mu_{boys} - \mu_{girls} = 0$$

$$H_1 : \mu_{boys} - \mu_{girls} \neq 0$$

- Null and alternative hypotheses stated in terms of *population parameters* yet tests are conducted using *sample statistics*
- Goal is to make probabilistic inference about the likelihood of observing the sample data under the null hypothesis
- Must then consider correct and incorrect decisions

CenterStat 1.88



Excellent general introductions to power in the social and behavioral sciences include:

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 98-101.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Statistical Power: Relevant Components

- Any inferential test thus relies on four components
  1. Type I error rate (α)
  2. Statistical power $(1-\beta)$
  3. Sample Size (N)
  4. Effect Size (ES)
- Any of the four can be computed given the other three
- Often biggest challenge when computing power is determining ES
- In SEM there are many different ways to define null hypothesis to be tested, and thus many different types of ES to consider

CenterStat 1.90

Determining Effect Size in SEMs

- Recall fundamental hypothesis in all SEMs:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

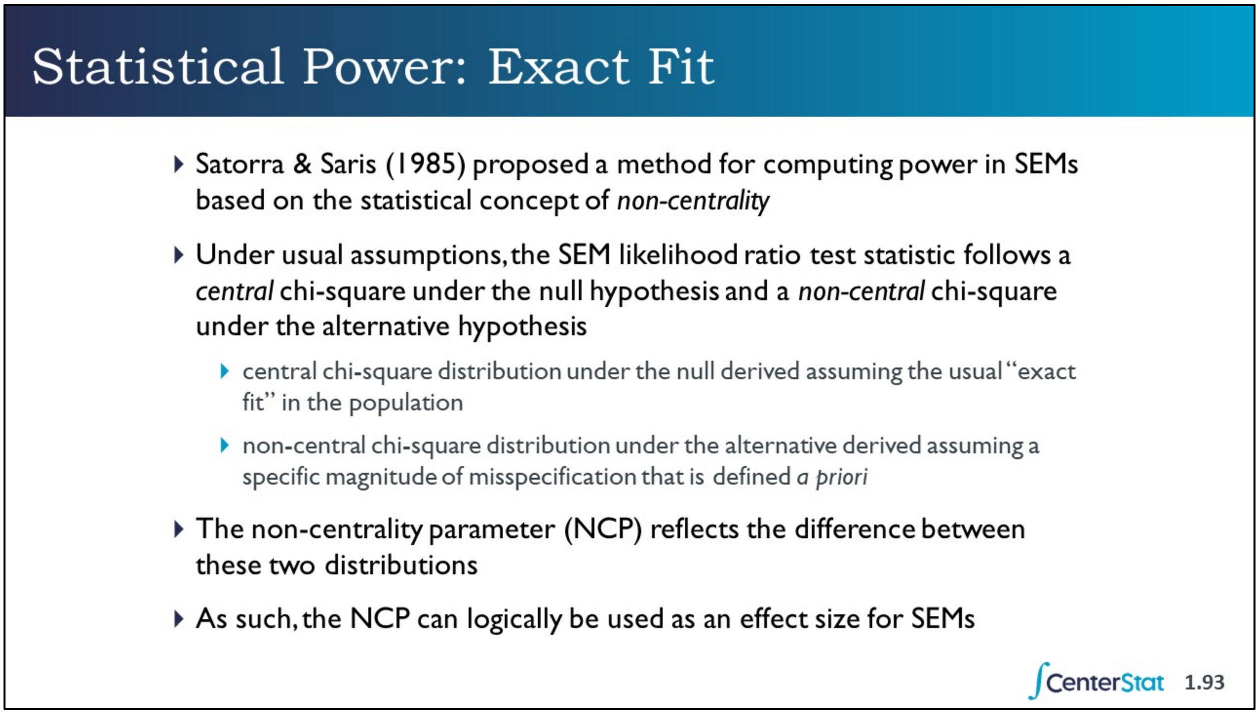
- This states that the population mean and covariance structures are *precisely equal* to those implied by the hypothesized model
  - in other words, the hypothesized model is *exactly correct*
- Thus in SEM rejection of the null hypothesis indicates that the hypothesized model is incorrect
- So when considering power in SEM, determining the effect size is tantamount to stating the degree of "incorrectness" of the hypothesized model

CenterStat 1.91

Determining Effect Size in SEMs

- In SEM, effect size can take many forms
  - the ES for a single parameter: e.g., a treatment effect or gender effect
  - the ES for a set of parameters: e.g., the joint influence of covariates
  - the ES for an entire model: e.g., the overall fit of the hypothesized model
- Different methods have been proposed for computing power in SEMs that vary in how ES is defined
- We will briefly consider three closely-related approaches

CenterStat 1.92



Albert Satorra and colleagues have made a number of significant contributions to the definition and estimation of statistical power in structural equation modeling. A few classic papers are:

Saris, W.E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. Bollen and S. Long (Eds.), *Testing Structural Equation Models* (pp. 181-204). Newbury Park: Sage.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561-582.

Satorra, A. (2003). Power of χ^2 goodness-of-fit tests in structural equation models: The case of non-normal data. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, J. J. Meulman (Eds.) *New Developments in Psychometrics* (pp. 57-68). Springer: Japan.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83-90.

Steps in Computing Power using NCP

- Satorra-Saris method straightforward to use

1. Define structure of hypothesized model of interest, including exact values of all parameters
2. Compute model-implied mean vector and covariance matrix
3. Define structure of alternative model that is nested within hypothesized model
   - e.g., parameters from hypothesized model fixed to zero or set equal
4. Fit alternative model to mean vector and covariance matrix implied by hypothesized model at a given sample size N
5. The chi-square test statistic obtained from Step #4 is the NCP
6. Referencing non-central chi-square distribution, compute power based on specific values of NCP, N, *df*, and α

CenterStat 1.94

Once the non-centrality parameter (NCP) is obtained in Step #5, it is straightforward to obtain the final estimate of power. To do so, we must compute the partial area under a non-central chi-square distribution. Critical values might be found in more advanced statistics text books, or these can easily be computed using any major statistics package. For example, the SAS code below computes power using the Satorra-Saris method for a single degree-of-freedom test to detect a misspecification of NCP=10.

```
data satorrasaris;                          *defines data step named satorrasaris;
 df=1;                                      *defines one degree-of-freedom;
 critval=3.841459;                          *sets critical value for p=.05 for central chi-square;
 ncp=10;                                    *hypothetical non-centrality parameter of 10;
 power=(1-(probchi(critval,df,ncp)));       *computes power using the non-central chisquare;
run;
proc print data=satorrasaris; run;          *prints contents of data step;
```

The output from this simple program is:

Obs	df	critval	ncp	power
1	1	3.84146	10	0.88538

So for a single degree-of-freedom and an assumed non-centrality parameter equal to 10, there is an 88.5% we will reject the null hypothesis. In other words, our likelihood ratio test has a power of .885 to detect a misspecification of this magnitude. This program can be easily modified to include more than one degree-of-freedom and any value of NCP. Further, this could be placed within a simple do-loop to compute an entire range of power spanning some minimum and maximum sample size. For more details on using this approach to create entire power curves across multiple effect sizes, see:

Muthén, B.O., & Curran, P.J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*, 371–402.

Pros and Cons of Satorra-Saris Method

- Advantages of NCP-based approach
  - under assumption of large sample size and multivariate normality, can obtain exact power estimate for a given hypothesized effect
  - hypothesized effect can involve one or multiple parameters
  - power can be easily computed using any SEM computer package
  - very useful to estimate power for a <u>specific</u> hypothesized effect
- Disadvantages of NCP-based approach
  - requires exact specification of model structure and numerical values of all parameters that are sometimes difficult to define and justify
  - assumes exact model fit under null hypothesis which does not likely hold
  - cannot be used to estimate power for a <u>general</u> misspecification somewhere within the model

CenterStat 1.95

Statistical Power: RMSEA-based

- Recall that one limit of LRT is it evaluates hypothesis of *exact* fit
- Many reasons to recast null hypothesis in terms of *close* fit
- Root Mean Squared Error of Approximation (RMSEA) provides method for conducting tests of close fit
  - RMSEA can also be used as alternative estimation method for ES

CenterStat 1.96

Statistical Power: RMSEA-based

- RMSEA expresses error of approximation per model *df*

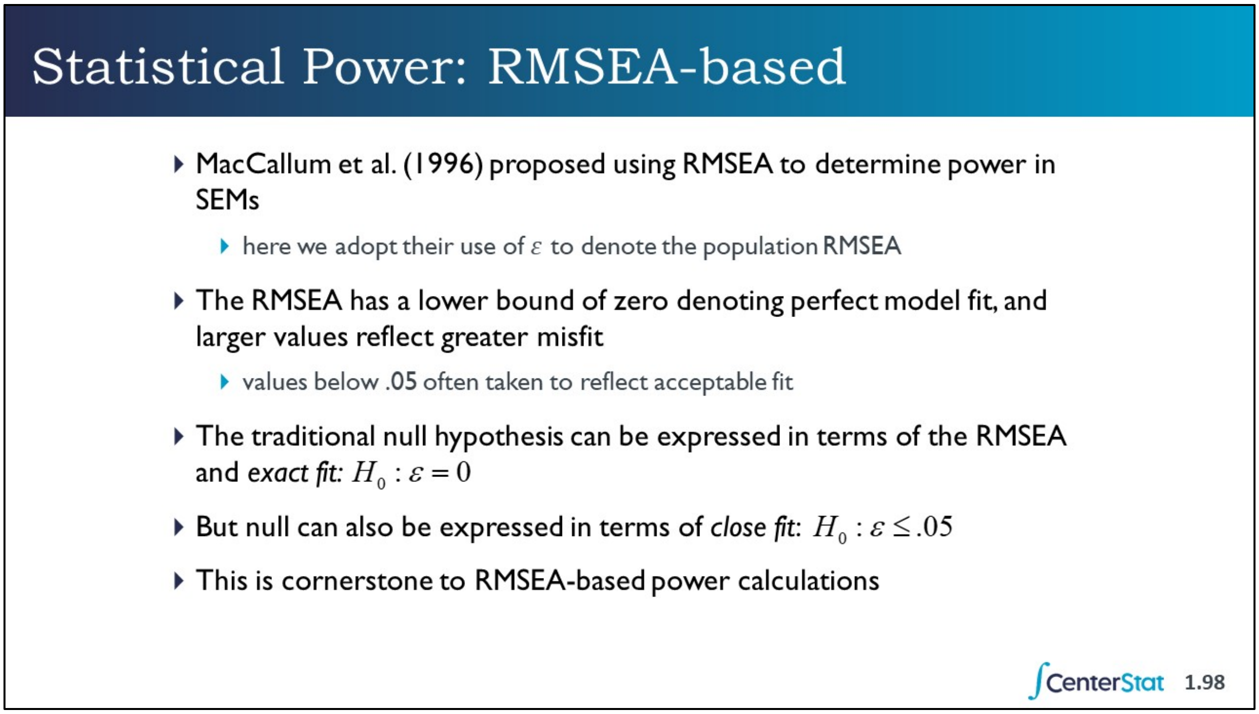
$$RMSEA = \sqrt{\frac{\chi^2 - df}{df\left(N-1\right)}}$$

- The numerator is the NCP for a misspecified model, so this too is NCP-based, but here we will consider the power of "close" fit

1.97

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*, 230-258.

Chen, F., Curran, P.J., Bollen, K.A., Kirby, J., and Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research*, 36, 462-494.



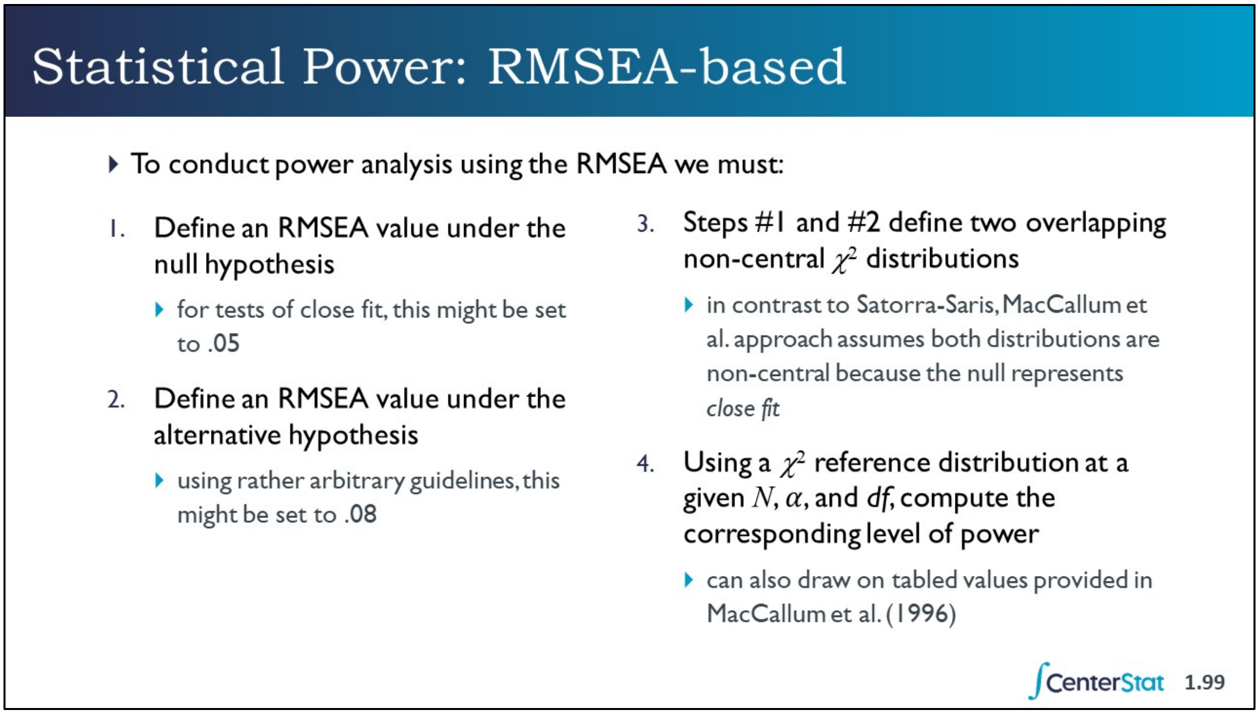
There are a number of excellent papers that explore various aspects of using the RMSEA to estimate power. See, for example:

Kim, K.H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling, 12*, 368-390.

Lee. T., Cai, L., & MacCallum, R.C. (2012). Power Analysis for tests of structural equation models. In R. Hoyle, D. Kaplan, g. Marcoulides, & S. West (Eds.), *Handbook of Structural Equation Modeling* (pp.181-194), New York: Guilford.

MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

MacCallum, R.C., Lee, T., & Browne, M.W. (2010). The issue of isopower in power analysis for tests of structural equation models. *Structural Equation Modeling, 17*, 23-41.



In many ways, the MacCallum et al. (1996) approach is similar to that of Satorra-Saris. The key difference is that the Satorra-Saris method assumes that the model obtains exact fit under the null hypothesis. As such, the distribution of the test statistic under the null is a central chi-square and the distribution under the alternative is a non-central chi-square that is shifted by the non-centrality parameter. In contrast, the MacCallum et al. approach starts with the assumption of *close fit* under the null hypothesis, thus the distributions of the test statistics under the null and alternative hypotheses are both non-central. The MacCallum et al. approach allows for the estimation of statistical power to detect a *general* misspecification somewhere within the confines of the model as opposed to defining a *specific* misspecification that results in an exact NCP as is done with Satorra-Saris. As such, it is quite easy to determine power using tables provided in MacCallum et al. (1996).

Table 2
Power Estimates for Selected Levels of Degrees of Freedom (df) and Sample Size

df and test		Sample size 100	200	300	400	500
5	Close	0.127	0.199	0.269	0.335	0.397
	Not close	0.081	0.124	0.181	0.248	0.324
	Exact	0.112	0.188	0.273	0.362	0.449
10	Close	0.169	0.294	0.413	0.520	0.612
	Not close	0.105	0.191	0.304	0.429	0.555
	Exact	0.141	0.266	0.406	0.541	0.661
15	Close	0.206	0.378	0.533	0.661	0.760
	Not close	0.127	0.254	0.414	0.578	0.720
	Exact	0.167	0.336	0.516	0.675	0.797
20	Close	0.241	0.454	0.633	0.766	0.855
	Not close	0.148	0.314	0.513	0.695	0.830
	Exact	0.192	0.400	0.609	0.773	0.882

For example, say that you wanted to compute statistical power as you planned a new study. Say that your hypothesized model was defined by *df*=15 and you want to determine the power associated with various sample sizes. Consulting Table 2 from MacCallum et al. (1996), the power of the test of *close fit* for *df*=15 would be .206 at a sample size of 100, .378 at 200, .533 at 300, .661 at 400, and .76 at 500. You would thus need a sample size greater than 500 to achieve a power of .80 for this model. It is clear from the table that power increases at a given *df* with increasing sample size, or at a given sample with increasing *df*. This information can provide valuable information in the planning and design of a new study.

Tabled Values of Power and Sample Size

- Example of partial tables drawn from MacCallum et al. (1996)

Table 2
Power Estimates for Selected Levels of Degrees of Freedom (df) and Sample Size

df and test		Sample size 100	200	300	400	500
5	Close	0.127	0.199	0.269	0.335	0.397
	Not close	0.081	0.124	0.181	0.248	0.324
	Exact	0.112	0.188	0.273	0.362	0.449
10	Close	0.169	0.294	0.413	0.520	0.612
	Not close	0.105	0.191	0.304	0.429	0.555
	Exact	0.141	0.266	0.406	0.541	0.661
15	Close	0.206	0.378	0.533	0.661	0.760
	Not close	0.127	0.254	0.414	0.578	0.720
	Exact	0.167	0.336	0.516	0.675	0.797
20	Close	0.241	0.454	0.633	0.766	0.855
	Not close	0.148	0.314	0.513	0.695	0.830
	Exact	0.192	0.400	0.609	0.773	0.882

Table 4
Minimum Sample Size to Achieve Power of 0.80 for Selected Levels of Degrees of Freedom (df)

df	Minimum *N* for test of close fit	Minimum *N* for test of not-close fit
2	3,488	2,382
4	1,807	1,426
6	1,238	1,069
8	954	875
10	782	750
12	666	663
14	585	598
16	522	547
18	472	508
20	435	474

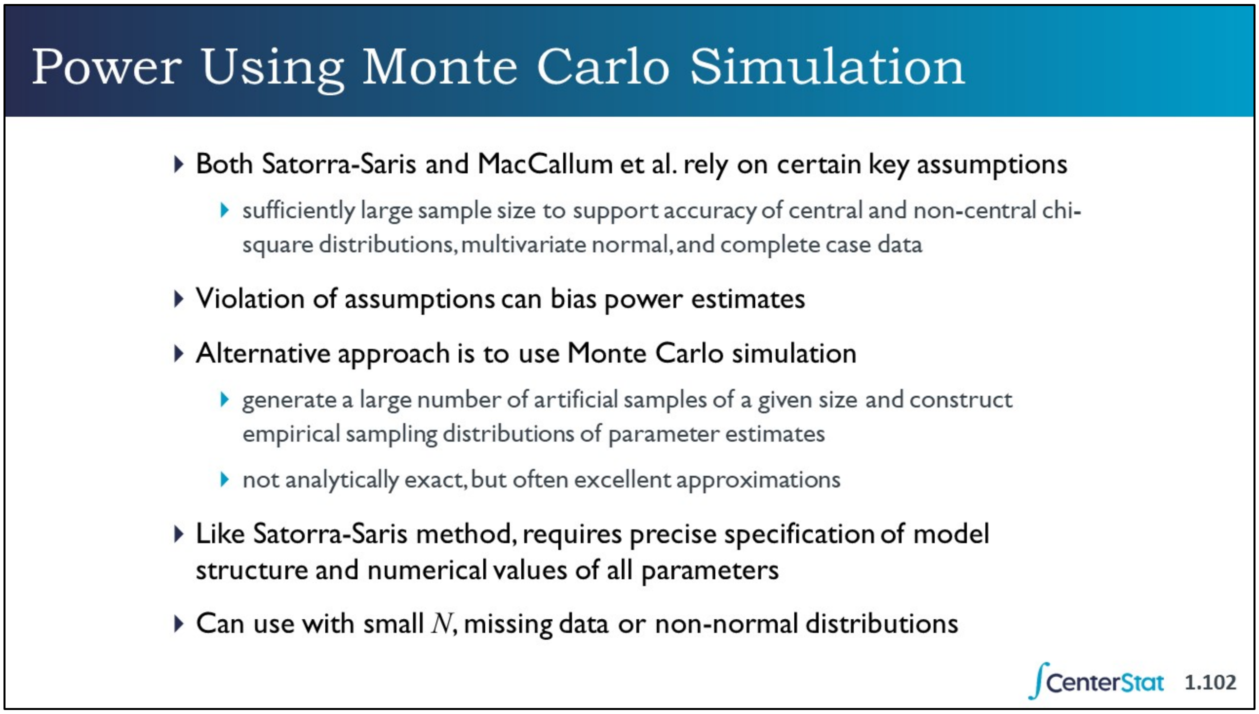
CenterStat 1.100

The table on the left allows you to determine the power that is associated with different combinations of *df* and sample size. The table on the right allows you to determine the sample size that is needed to achieve a power of .80 as a function of *df*. Both tables also include tests of *not-close fit*; see MacCallum et al. (1996) for details about how this differs from tests of *close fit*.

Pros and Cons of RMSEA-based Power

- Advantages of RMSEA-based approach
  - does not require exact specification of model or selection of precise numerical values for all parameters
  - allows for power estimation from a testing perspective of *close fit*
  - can obtain power estimates from existing tables in MacCallum et al. or easily compute custom values using any standard software package
  - same as Satorra-Saris approach for tests of exact fit
- Disadvantages of RMSEA-based approach
  - computes power for a general misspecification somewhere within the SEM and not for a specific hypothesized effect
  - RMSEA values denoting close fit not strongly established (e.g., .05)
  - like Satorra-Saris, assumes large samples and multivariate normal data

CenterStat 1.101



The Satorra-Saris and MacCallum et al. methods draw on asymptotic sampling distributions to determine the probability that the null would be rejected if the null were false. For example, as we saw earlier, we can use the MacCallum et al. method determine that the power to reject a test of close fit with *df*=15 at *N*=500 is .76. Because both the Satorra-Saris and MacCallum et al. method are based on asymptotic sampling distributions, these also require assumptions about sufficiently large sample size, multivariate normal distributions, and complete case data. If any of these conditions are violated, the resulting power estimates can be biased. To address this limitation, it is possible to use a Monte Carlo simulation approach in which we (1) artificially generate a large number of samples that are consistent with a hypothesized model; (2) to these samples we fit some alternative model that is a misspecified version of the hypothesized model; and (3) we count up the number of times the alternative model is rejected based on the LRT. This final tally represents our empirical estimate of power. Instead of relying on an asymptotic sampling distribution (as is done with the other two methods), here we build an empirical sampling distribution based on the artificial data. The advantage of this approach is that we can estimate power under conditions such as very small sample sizes, non-normally distributed data, or missing data. The disadvantage is that this approach is highly computationally intensive and remains only an approximation to the asymptotic sampling distributions. For more details, see:

Curran, P.J., Bollen, K.A., Paxton, P., Kirby, J., & Chen, F. (2002). The non-central chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research, 37*, 1-36.

Dolan, C., van der Sluis, S., & Grasman, R. (2005). A note on normal theory power calculation in SEM with data missing completely at random. *Structural Equation Modeling, 12*, 245-262.

Muthén, L.K., & Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.

Summary

- In any SEM application must consider normality of DVs, issue of equivalent models, and statistical power
- For continuous but non-normal DVs, can use robust methods of estimation although there is some loss of power
  - for DVs that are binary or ordinal, must move to nonlinear SEM
- Although there is no specific "fix" for equivalent models, good to be aware of issue and temper interpretations accordingly
- To determine power, can use Satorra-Saris method of exact fit, MacCallum et al. method of close fit, or Monte Carlo simulation
  - provides important information for determining power of existing model or when planning future studies

CenterStat 1.103

Appendix:
Maximum Likelihood

Objectives

- Establish conceptual understanding of maximum likelihood estimation
  - Estimator most commonly used with SEMs
- Begin with univariate case
- Extend to multivariate case
- Note how ML allows for missing data

CenterStat A.2

Estimation

- A core topic that relates to everything we will discuss this week is estimation
  - how do we obtain our parameter estimates from sample data?
- There are three optimal characteristics for an estimator:
  - unbiased: if we were to repeat our study an infinite number of times, the mean of the sample estimates would equal the population value
  - consistent: as the sample size approaches infinity, the sample estimate approaches the population value
  - efficient: no other estimator has a smaller sampling error for the parameter estimate

CenterStat A.3

Background on ML

- The estimator used to fit most SEMs is maximum likelihood (ML)
- ML provides estimates for model parameters that have the highest likelihood of giving rise to the sample data
- ML is a large sample estimator

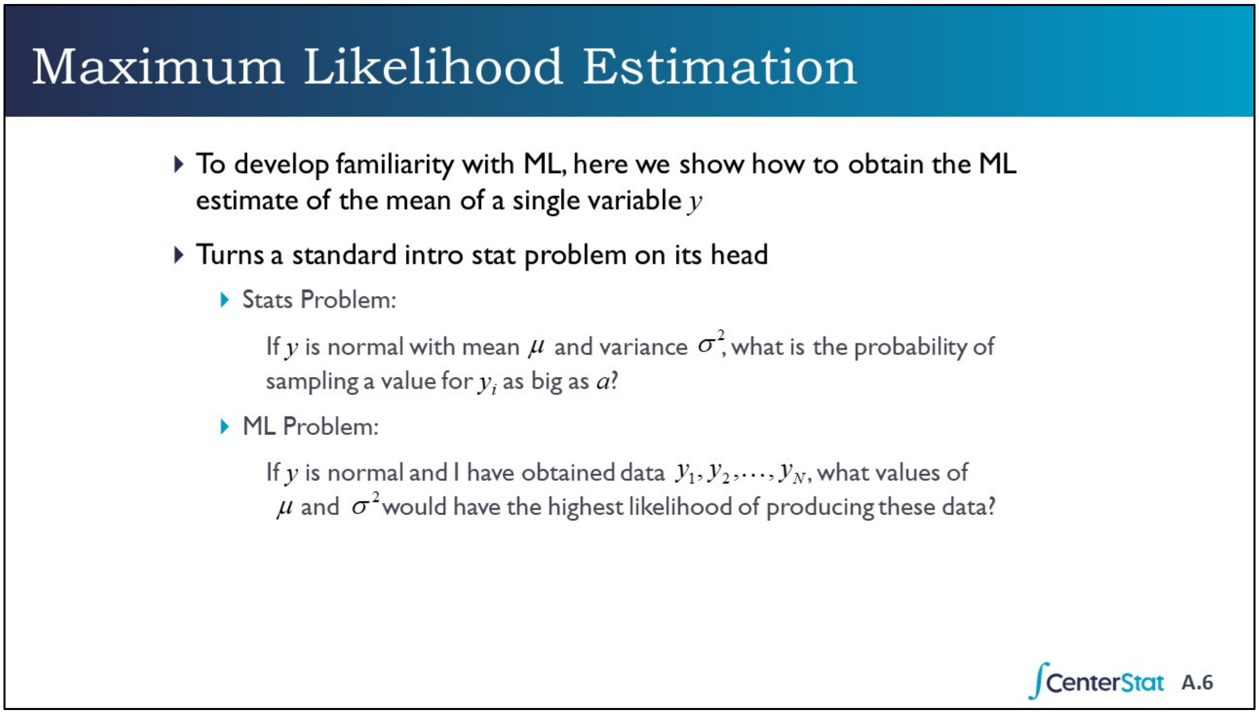
CenterStat A.4

Background on ML

- ML has many useful properties:
  - Can be used to fit a variety of models
  - Under certain assumptions, ML is asymptotically unbiased, consistent, and maximally efficient
  - Estimates are asymptotically normally distributed, providing basis for inference tests
  - Combinations of ML estimates are themselves ML estimates
  - Relative fit of competing models can be compared using likelihood ratio tests (chi-square difference tests)
  - Can accommodate partially missing data under MAR assumption

CenterStat A.5

Asymptotic means as $N \to \infty$



Example of ML: You go to a new town and see Bus #23 drive past. Assuming buses are numbered consecutively, what is the maximum likelihood estimate for the number of buses in the town's fleet? The answer is 23. If there were fewer than 23 buses in the fleet then the probability of observing Bus #23 would be zero, so there must be at least 23 buses. If there are 23 buses, then the probability of observing Bus #23 is 1/23. If there are 24 buses, the probability is 1/24. If there are 25 buses, then the probability is 1/25. The highest probability of observing Bus #23 occurs if there are 23 buses in the fleet; if there are any more buses then the probability of observing Bus #23 is reduced. Thus, 23 buses is the maximum likelihood estimate for the number of buses in the town's fleet.

Incidentally, Michael Jordan wore #23 when he was a basketball player at North Carolina (and later, much less importantly, on the Chicago Bulls).

The Probability Density Function

- Maximum likelihood requires that we state the distribution for our variables
- For instance, "y is normally distributed in the population with mean μ and variance σ^2"
- The distribution of y is then the normal curve, which is traced out by the probability density function (PDF):

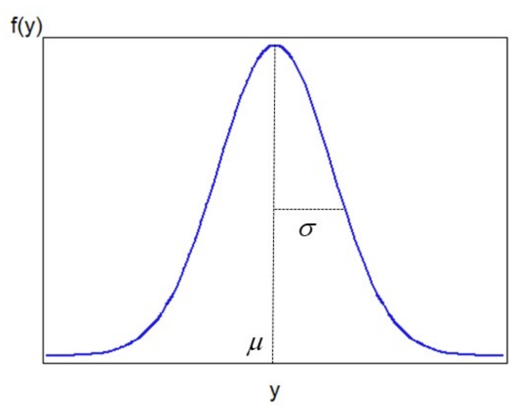
$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(y_i - \mu)^2 / \sigma^2 \right]$$

CenterStat A.7

This function looks complicated, but it simply traces out the familiar normal curve. The center of the normal curve is determined by the parameter μ (mean) and the spread is determined by the parameter σ^2 (variance). Note that, although it is designated with a Greek letter, π is the famous constant 'pi = 3.14…' and not a parameter of the distribution.

The Normal PDF

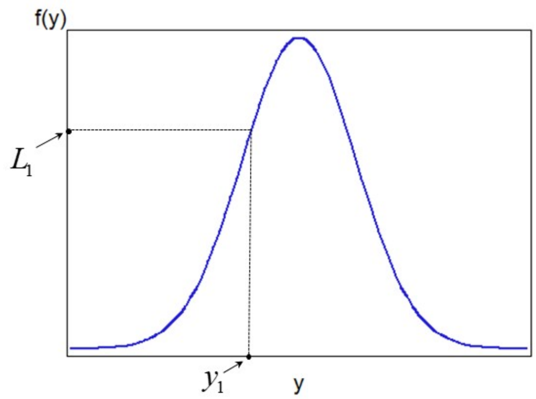
$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(y_i - \mu)^2 / \sigma^2 \right]$$



CenterStat A.8

The Likelihood

- The *likelihood* of a given observation y_i, denoted L_i, is defined as the height of the PDF at that value, or $L_i = f\left(y_i\right)$



CenterStat A.9

The Joint Likelihood

- L_i is the likelihood for a single observation y_i
- Given a set of observations, $y_1, y_2, \cdots, y_N$, want to find parameter estimates that maximize their joint occurrence
- If observations are independent, joint likelihood is the product of the individual likelihoods, or

$$L = \prod_{i=1}^{N} L_i$$

- Replacing L_i with the normal PDF, we obtain

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(y_i - \mu\right)^2 / \sigma^2\right]$$

CenterStat A.10

If the observations are not independent (e.g., data is obtained on students nested within schools) then a more complex formulation of the joint likelihood is required, such as is used with multilevel models.

Maximizing the Likelihood

- We now seek values for μ and σ^2 that maximize this likelihood

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(y_i - \mu)^2 / \sigma^2 \right]$$

- These values, denoted $\hat{\mu}$ and $\hat{\sigma}^2$ are the maximum likelihood estimates, or MLEs
- We can better see how this works by focusing on just $\hat{\mu}$, assuming that is already known σ^2

CenterStat A.11

We are assuming σ^2 to be known to simplify the presentation but, in practice, this would rarely be the case and we would simultaneously maximize the likelihood with respect to both the mean and variance.

A Simple Example: The Mean

- Suppose we take N=5 observations on y:

{3.92, 4.92, 5.40, 6.80, 7.24}

- We seek the MLE $\hat{\mu}$ that maximizes the joint likelihood of observing these five values
- Suppose we know that y is normally distributed and has variance $\sigma^2 = 1$
- We can use an iterative procedure to find $\hat{\mu}$
  - We begin by guessing that $\hat{\mu}$ will be 4
  - We revise our guess until we maximize L

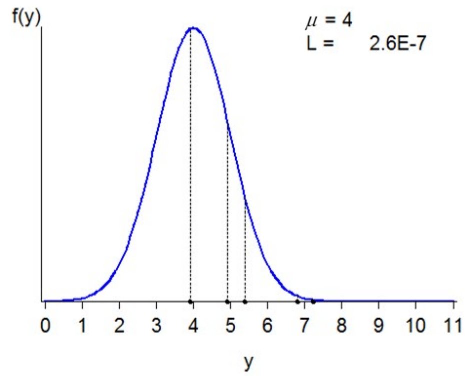
CenterStat A.12

We demonstrate the iterative approach here because that is what will generalize to the SEM. For this simple problem there is actually a closed-form mathematical solution (equivalent to the usual formula for the sample mean; i.e., sum of y divided by N), but such solutions will not be available in the SEM.

Maximizing the Likelihood

{3.92, 4.92, 5.40, 6.80, 7.24}

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(y_i - \mu)^2 / \sigma^2 \right]$$



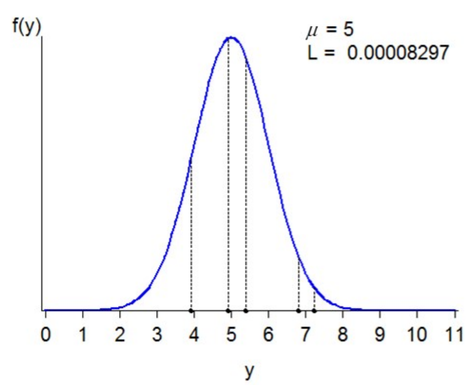
CenterStat A.13

Our initial guess of 4 isn’t very good. Notice that the center of the distribution is too far to the left, generating very low likelihood values for the observations to the right.

Maximizing the Likelihood

{3.92, 4.92, 5.40, 6.80, 7.24}

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(y_i - \mu)^2 / \sigma^2 \right]$$



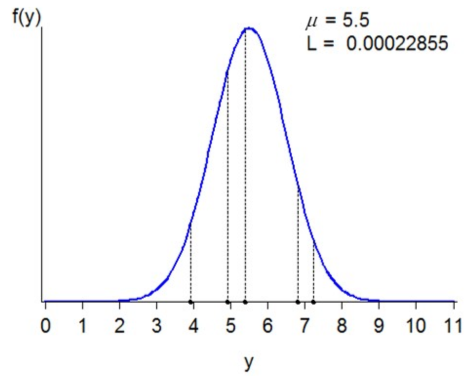
CenterStat A.14

Nudging the distribution to the right by revising our guess for the mean to be 5 we obtain a higher likelihood. But maybe we can still do better.

Maximizing the Likelihood

{3.92, 4.92, 5.40, 6.80, 7.24}

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} (y_i - \mu)^2 / \sigma^2 \right]$$



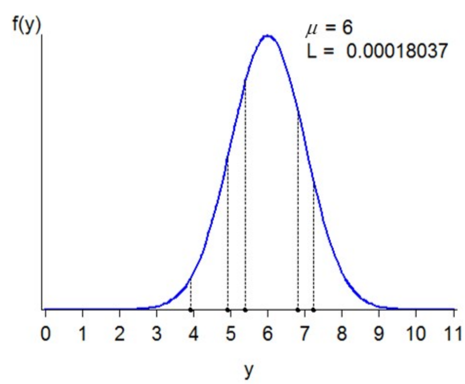
CenterStat A.15

Revising our guess for the mean to be 5.5, we obtain an even higher likelihood.

Maximizing the Likelihood

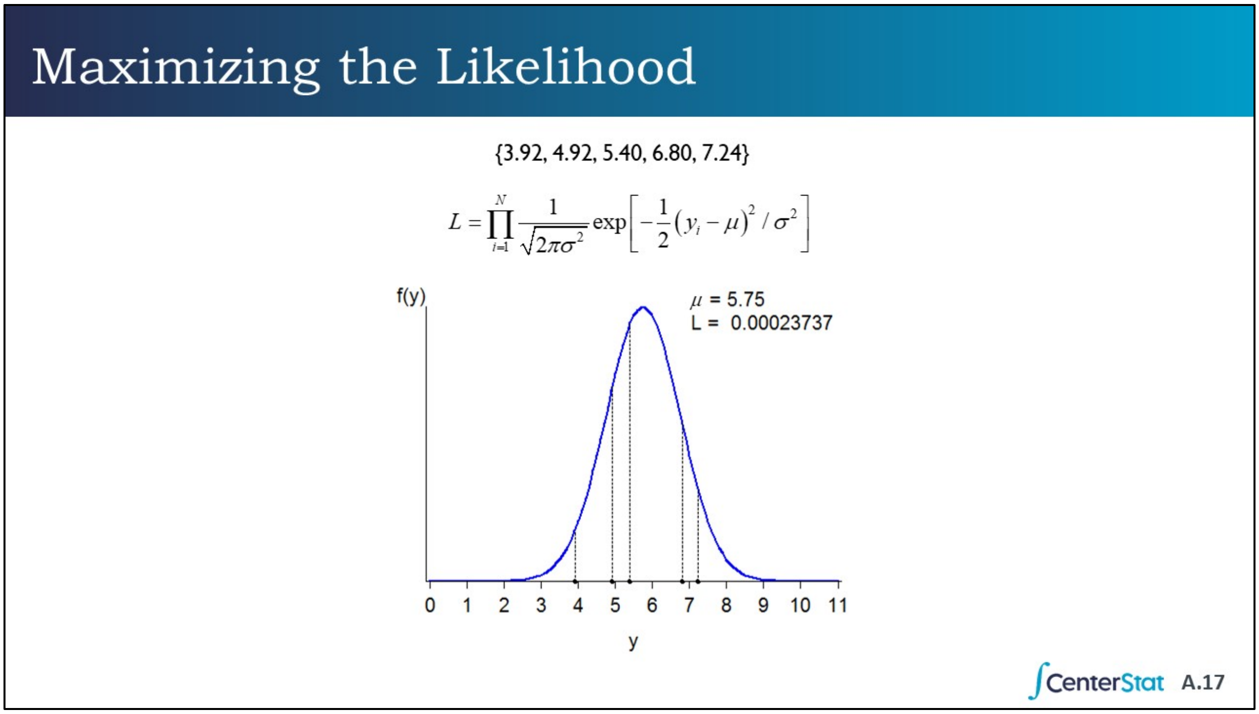
{3.92, 4.92, 5.40, 6.80, 7.24}

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} (y_i - \mu)^2 / \sigma^2 \right]$$

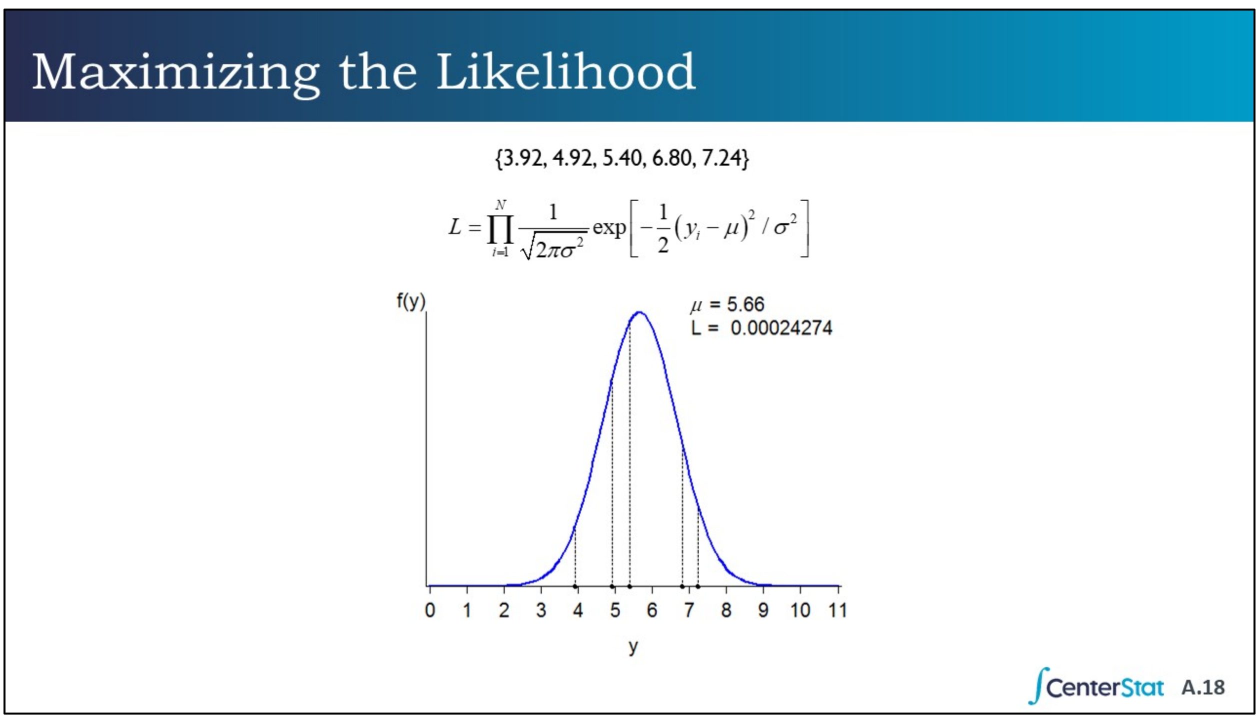


CenterStat A.16

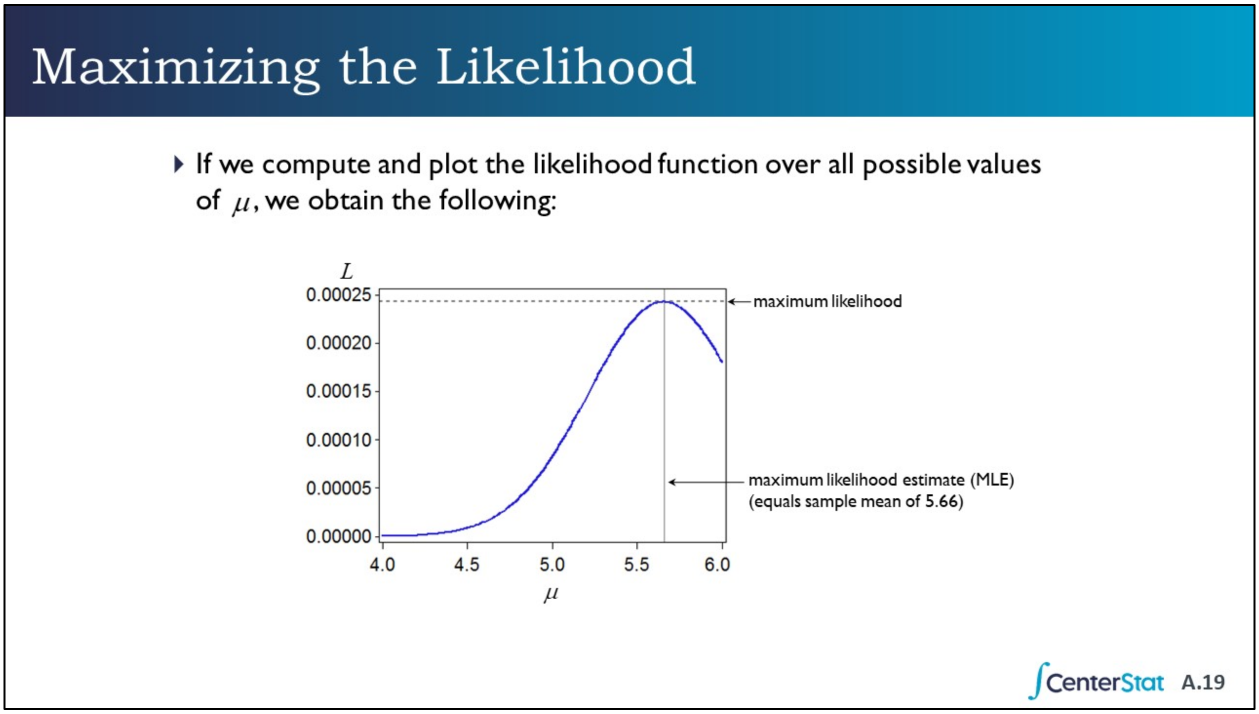
Now we’ve gone too far; a mean of 6 actually produces a lower likelihood than a mean of 5.5.



A guess midway between 5.5 and 6 yields the highest likelihood so far.



We can keep playing this guessing game until we arrive at the best possible guess for the population mean, that is, the specific value that maximizes the likelihood of the observed data. That value is the maximum likelihood estimate (MLE). In this case, the MLE for the mean is 5.66.



Here's how we can determine the MLE: The first derivative of the likelihood function with respect to μ (also known as the gradient) indicates the slope of the tangent line to the curve depicted above. The MLE is the value of μ for which the slope of the tangent line is zero, indicating that the maximum has been obtained. Thus, to obtain the MLE $\hat{\mu}$, we set the first derivative equal to zero and solve.

Sparing details, the resulting MLE for the mean is $\hat{\mu} = \frac{\sum y_i}{N}$.

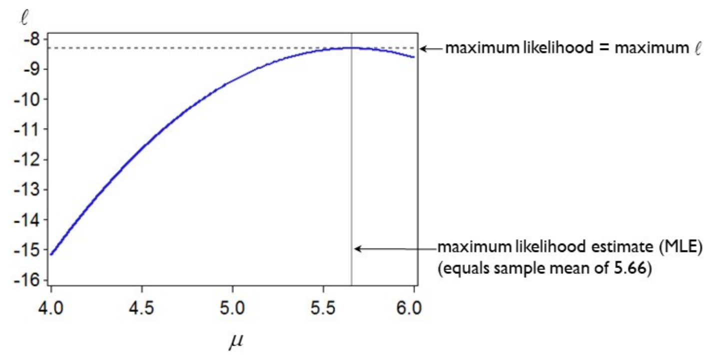
In practice, we would also need to obtain the MLE for the variance. The curve above would then become a surface with coordinates determined jointly by μ and σ^2, and we would need to find the highest point (like climbing to the top of a hill). To obtain the MLE for the variance, we would then take the partial derivative of the likelihood function with respect to σ^2, set it to zero, and solve.

The MLE for the variance is $\hat{\sigma}^2 = \frac{\sum (y_i - \hat{\mu})^2}{N}$

Notice that the MLE for the variance differs slightly from what we learned in intro stats class, where the denominator was $N-1$ to account for the lost degree of freedom associated with estimating the mean. Because it divides by N instead of $N-1$, the MLE for the variance is actually biased in small samples. The difference between N and $N-1$ becomes increasingly trivial, however, as N increases. This demonstrates that the MLE for the sample variance is asymptotically unbiased (i.e., unbiased in large samples). Although "closed form" solutions for these MLEs exist and can be obtained mathematically, for more complex models no closed form solution may exist and the iterative approach is usually required.

The Log-Likelihood

- To avoid multiplying tiny numbers, easier to use natural log of the likelihood: $\ell = \ln(L)$
- ℓ is a monotonic transformation of L, so the maximum occurs at the same value



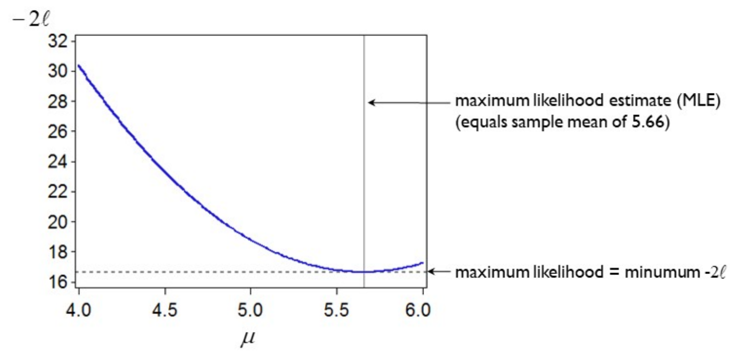
CenterStat A.20

Since each individual likelihood is between zero and one (like a probability), multiplying many of these values together to obtain the joint likelihood results in very tiny numbers, as seen in our simple demonstration even with just five observations. Taking the natural log avoids these very tiny numbers and also converts the product operator in the likelihood function into a sum operator for the log-likelihood function (shown on the next page). That is, we now take the sum of the individual log-likelihoods and maximize this sum.

Taking the log of a number between zero and one produces a negative value; hence the log-likelihood is negative.

-2 Log-Likelihood

- Even more convenient to use $-2\ell = -2\ln(L)$
- Again, a monotonic transformation, so obtain the same MLE
- Now, however, we must locate the <u>minimum</u> of the function



CenterStat A.21

By multiplying the log-likelihood by negative two we can make the value positive. Doing so also has the effect of removing the multiplier -1/2 that appears in the log-likelihood function (see below), further simplifying computations. As we will see, the value of the -2 Log-Likelihood at the MLEs also plays a role a number of fit statistics, such as likelihood ratio tests and information criteria like AIC and BIC. In some contexts, the -2 Log-Likelihood is referred to as *model deviance* (e.g., in multilevel modeling).

Simplifications

- Likelihood:

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(y_i - \mu)^2 / \sigma^2\right]$$

- Log-Likelihood:

$$\ell = \sum_{i=1}^{N} \left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2}(y_i - \mu)^2 / \sigma^2\right]$$

- -2 Log-Likelihood:

$$-2\ell = \sum_{i=1}^{N} \left[\ln(2\pi) + \ln(\sigma^2) + (y_i - \mu)^2 / \sigma^2\right]$$

CenterStat A.22

Maximizing the Likelihood

- Usually more than just one parameter to estimate
  - e.g., both mean and variance
- General procedure for finding MLEs:
  - Take first partial derivatives with respect to each parameter
    - Gives slope of the tangent line to the function, called "gradient"
  - Iteratively revise values for MLEs until partial derivatives all within tolerance of zero
    - When within tolerance, model has "converged"
  - SEs for MLEs based on how rapidly partial derivatives approach zero
    - Second derivatives, contained in "Hessian matrix"
  - Many different algorithms exist for finding MLEs
    - Fisher Scoring, Newton-Raphson, Expectation-Maximization, etc.

CenterStat A.23

There are various ways to set the tolerance for convergence, known as "convergence criteria." Before interpreting any estimates it is critical to verify that the model has converged. Many software programs will report the values of the estimates at the last iteration even if the model has not converged. These provisional estimates can be helpful for diagnosing possible problems with the model but should not be interpreted. The values output by the program are only MLEs if the model has converged.

Extending to Multivariate Models

- In the univariate case, we must estimate only one mean μ and one variance σ^2
- In the multivariate case with p variables, these expand to a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & & & \\ \sigma_{21} & \sigma_{22}^2 & & \\ \vdots & \ddots & \ddots & \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}^2 \end{bmatrix}$$

CenterStat A.24

Extending to Multivariate Models

- Univariate normal likelihood:

$$L=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(y_i-\mu\right)^2/\sigma^2\right]$$

$$-2\ell=\sum_{i=1}^{N}\left[\ln\left(2\pi\right)+\ln\left(\sigma^2\right)+\left(y_i-\mu\right)^2/\sigma^2\right]$$

Replacing scalars with matrix equivalents

- Multivariate normal likelihood for p variables:

$$L=\prod_{i=1}^{N}\frac{1}{\sqrt{\left(2\pi\right)^p\left|\boldsymbol{\Sigma}\right|}}\exp\left[-\frac{1}{2}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)\right]$$

$$-2\ell=\sum_{i=1}^{N}\left[p\ln\left(2\pi\right)+\ln\left|\boldsymbol{\Sigma}\right|+\left(\mathbf{y}_i-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)\right]$$

CenterStat A.25

Notice the critical role of matrix algebra in transitioning from the univariate to the multivariate likelihood. Recall that the determinant represents a generalized measure of variance, hence $\left|\boldsymbol{\Sigma}\right|$ replaces σ^2 in the first part of the likelihood function. Likewise, in the second part of the likelihood function, squared deviations from the mean are produced through the pre- and post-multiplication of $\left(\mathbf{y}_i-\boldsymbol{\mu}\right)$ (transposed, in the first instance). Finally, since division is no longer possible, we weight these deviations by the inverse $\boldsymbol{\Sigma}^{-1}$. Thus, the multivariate normal likelihood simply generalizes the univariate likelihood through the use of matrix algebra equivalents for the scalar values.

Missing Data

- The complete data -2 log-likelihood is

$$-2\ell = \sum_{i=1}^{N} \left[p \ln(2\pi) + \ln|\boldsymbol{\Sigma}| + (\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right]$$

- but missing data can be accommodated by using

$$-2\ell = \sum_{i=1}^{N} \left[n_i \ln(2\pi) + \ln|\boldsymbol{\Sigma}_i| + (\mathbf{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right]$$

Where

n_i is the number of variables with observed values for case i

$\boldsymbol{\mu}_i$ is the mean vector minus rows for variables missing for case i

$\boldsymbol{\Sigma}_i$ is the covariance matrix minus rows and columns for variables missing for case i

CenterStat A.26

Although this is typical notation, there is some potential to confuse n_i, the number of observations obtained for person i (with a maximum value of p), with N, the total number of people in the sample.

ML with incomplete data assumes that missing values are missing at random (MAR). MAR is satisfied if the missing data values are related only to the observed values of the variables in the model and not to any unobserved (or missing) values. We discuss the MAR assumption in greater detail in Chapter 3.

For an accessible overview of missing data mechanisms and how ML accommodates missing data (in relation to other options, such as multiple imputation), see

Enders, C.K. (2010). *Applied Missing Data Analysis.* New York: Guilford.

Moving Towards the SEM

- Multivariate normal -2 Log-Likelihood:

$$-2\ell = \sum_{i=1}^{N}\left[n_i \ln(2\pi) + \ln|\boldsymbol{\Sigma}_i| + (\mathbf{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right]$$

- Goal is simply to obtain best estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
  - No model structure
- The SEM will impose structure on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
  - e.g., "feeling sad", "loneliness", and "crying" covary because they reflect an underlying latent factor of depression
- Seek to maximize the likelihood given this structure
  - Ideally, likelihood of the data under the SEM not significantly worse than when no structure is applied, i.e., SEM accurately reproduces $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

CenterStat A.27

We will see a variety of mean and covariance structures throughout the course. Ultimately, we will also compare the likelihood of the data given the structure imposed by the specified model to the likelihood of the data when no structure is imposed to evaluate model fit. Since all parameters of the mean vector and covariance matrix are freely estimated when no structure is imposed this point of comparison is often referred to as a "saturated model."

Model-Implied Moment Structure

- Recall measurement and structural models of full SEM are

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \qquad VAR(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta}$$

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i, \qquad VAR(\boldsymbol{\zeta}_i) = \boldsymbol{\Psi}$$

- These jointly imply specific mean and covariance structure

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \left[\begin{array}{c} \boldsymbol{\nu} + \boldsymbol{\Lambda}(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\mu}_x) \\ \hline \boldsymbol{\mu}_x \end{array}\right]$$

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \left[\begin{array}{c|c} \boldsymbol{\Lambda}(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{xx}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1'}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} & \\ \hline \boldsymbol{\Sigma}_{xx}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B})^{-1'}\boldsymbol{\Lambda}' & \boldsymbol{\Sigma}_{xx} \end{array}\right]$$

CenterStat A.28

These are the most general expressions for the fully latent SEM. For a manifest variable path model, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ drop out. Similarly, if no endogenous variables relate to other endogenous variables, $\mathbf{B}$ and $\mathbf{I}$ drop out.

Model Estimation

- The normal-theory ML fitting function is again

$$-2\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left[n_i \log(2\pi) + \log|\boldsymbol{\Sigma}_i(\boldsymbol{\theta})| + (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}))' \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta})(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) \right]$$

  - The i subscripts on n, Σ, and μ are there to allow for missing data.

- An equivalent ML fitting function can be written in terms of sufficient summary statistics only

$$F_{ML}(\boldsymbol{\theta}) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + tr\left[\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right] - \log|\mathbf{S}| - p +$$
$$\left[\mathbf{m} - \boldsymbol{\mu}(\boldsymbol{\theta})\right]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\left[\mathbf{m} - \boldsymbol{\mu}(\boldsymbol{\theta})\right]$$

  - Assumes complete data

CenterStat A.29

These fit functions apply to whatever model structure is defined in vector $\boldsymbol{\theta}$.

Sidebar: The likelihood shown above is based on a joint, multivariate normal distribution for $\mathbf{y}$ and $\mathbf{x}$ where $\mathbf{y}$ is a single vector with endogenous variables listed first and exogenous variables listed second. More commonly, maximum likelihood for the linear regression model is based on the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$. The conditional likelihood formulation is appealing because it explicates the assumption of conditional normality for $\mathbf{y}$ (i.e., normally distributed residuals), and because it makes no explicit assumptions about the distribution of $\mathbf{x}$. For various reasons, some historical and others practical, the SEM was first formulated using a joint multivariate normal likelihood, so this is what we present here. Under broad conditions, this joint likelihood yields the same maximum likelihood estimates as the conditional likelihood, even when the $\mathbf{x}$ variables are not normally distributed (Jöreskog, 1973; see also Bollen, 1989, pp. 126-128). Thus, the distinction is typically of little practical importance and we may include non-normally distributed predictors in our regression models without concern even when using the joint likelihood formulation to obtain the MLEs.

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). New York: Academic Press.

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

Summary

- ML provides a principled way of obtaining sample estimates that best match the observed data
- MLEs have many useful properties
  - Asymptotic unbiasedness, consistency, efficiency, and normal sampling distributions
- Maximum of likelihood often found using iterative procedures
  - Typically minimize -2log-likelihood instead of maximizing likelihood itself
- ML can be used with multivariate models like the SEM and can accommodate partially missing data

CenterStat A.30