Overview of SEM
What is it, when should I use it, and where did it come from?

Steven M. Boker

Department of Psychology
University of Virginia

Introduction to SEM
Psyc–8501–001

Outline

- Review of the syllabus.
- A little history.
- What is Structural Equation Modeling?
- When should I use SEM?

Review of the Syllabus

- Prerequisites.
  - First year graduate sequence in statistics.
  - Multivariate Analysis or its equivalent.
- Student Evaluation.
  - Article in APA style.
  - You must be first author, it must be new work, and it must include an SEM analysis comparing at least two models.
- Textbook(s)
  - Basics of Structural Equation Modeling by Maruyama.
  - A First Course in Structural Equation Modeling by Raykov and Marcoulides.
- Organization.
  - Preparatory learning.
  - Practical demonstrations with the students' data.

Tracing a Brief History of SEM

- 1896 — Pearson: Correlation
- 1904 — Spearman: Factor analysis.
- 1918 — Sewall Wright: Path coefficients.
- 1920 — Sewall Wright: Path analysis.
- 1934 — Sewall Wright. Components of correlation
- Time passes.
- 1966 — Duncan “rediscovers” Sewall Wright.
- 1970 — Jöreskog proposes LISREL.
- 1984 — McArdle and McDonald’s RAM model.
- 1990s — RAMpath, Mx, Amos, EQS, Ramona, Mplus.

Pearson and Spearman

- Karl Pearson (1896) proposed the correlation coefficient.

$$r_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

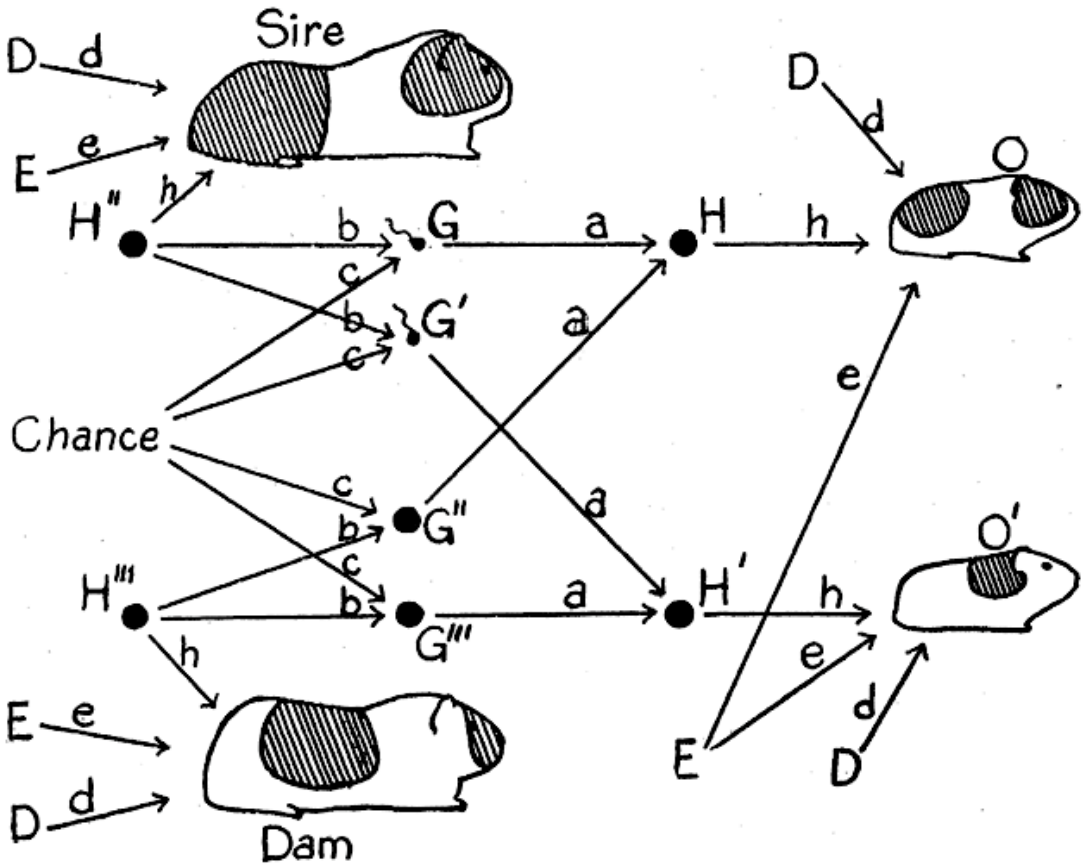
- Spearman (1904b) built on Pearson correlations and developed a method of factor analysis.
  - He used this to method to argue for a general factor of intelligence (Spearman, 1904a).

Sewall Wright (1918)

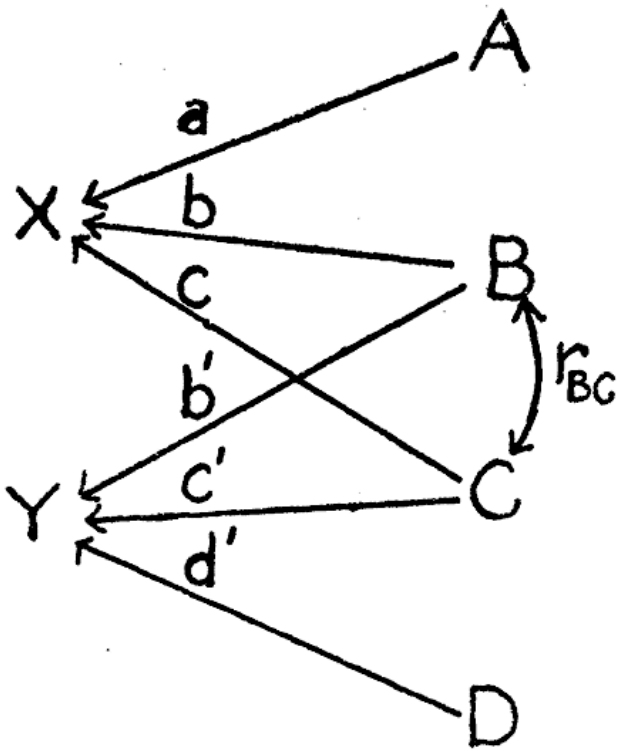
- Wright (1918) first proposed the method of *path coefficients* which was based on partial correlations.
  - He used this to method to show that while there was a general factor for size in rabbits, there were unique factors for skull and leg bone size.
  - A variant of factor analysis.

Sewall Wright (1920)

- Wright (1920) proposed the basis of structural equation modeling in a paper on heredity in guinea–pigs.
- The first statement of path analysis comes in this paper: “The correlation between two variables can be shown to equal the sum of the products of the chains of path coefficients along all of the paths by which they are connected.”
- The first path diagrams also were published here.



The first published path diagram (Wright, 1920)



"Diagram illustrating two effects(XY) which are determined in part by the same correlated causes (BC)." (Wright, 1920)

Wright (1920) on components of correlation

It can be shown that the squares of the path coefficients measure the degree of determination by each cause. If the causes are independent of each other, the sum of the squared path coefficients is unity. If the causes are correlated, terms representing joint determination must be recognized. The complete determination of X in figure 6 by factor *A* and the correlated factors *B* and *C*, can be expressed by the equation:

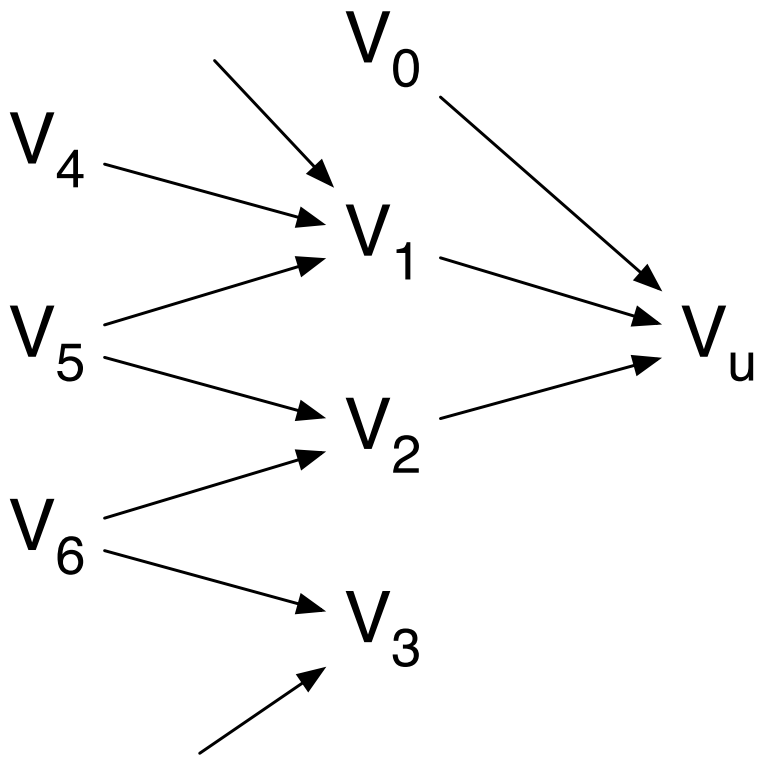
$$a^2 + b^2 + c^2 + 2bcr_{BC} = 1 \quad (1)$$

Sewall Wright (1934)

- Wright (1934) proposed a set of tracing rules that such that, “Any correlation between variables in a network of sequential relations can be analyzed into contributions from all of the paths (direct or through common factors) by which the two variables are connected, such that the value of each contribution is the product of the coefficients pertaining to the elementary paths.”

Sewall Wright (1934)

- He proposed that the method of path analysis was completely general, “The correlation is thus analyzed into contributions from all of the paths in the diagram passing through each factor of one of the variables.”
- “The solution for the path coefficients . . . need to be multiplied by the proper ratio of standard deviations to give Pearson’s formulae for the partial regression coefficients.”
- Here he also introduced the notion of latent variables as representing true scores.



An example path diagram (after Wright, 1934)

Time Passes . . .

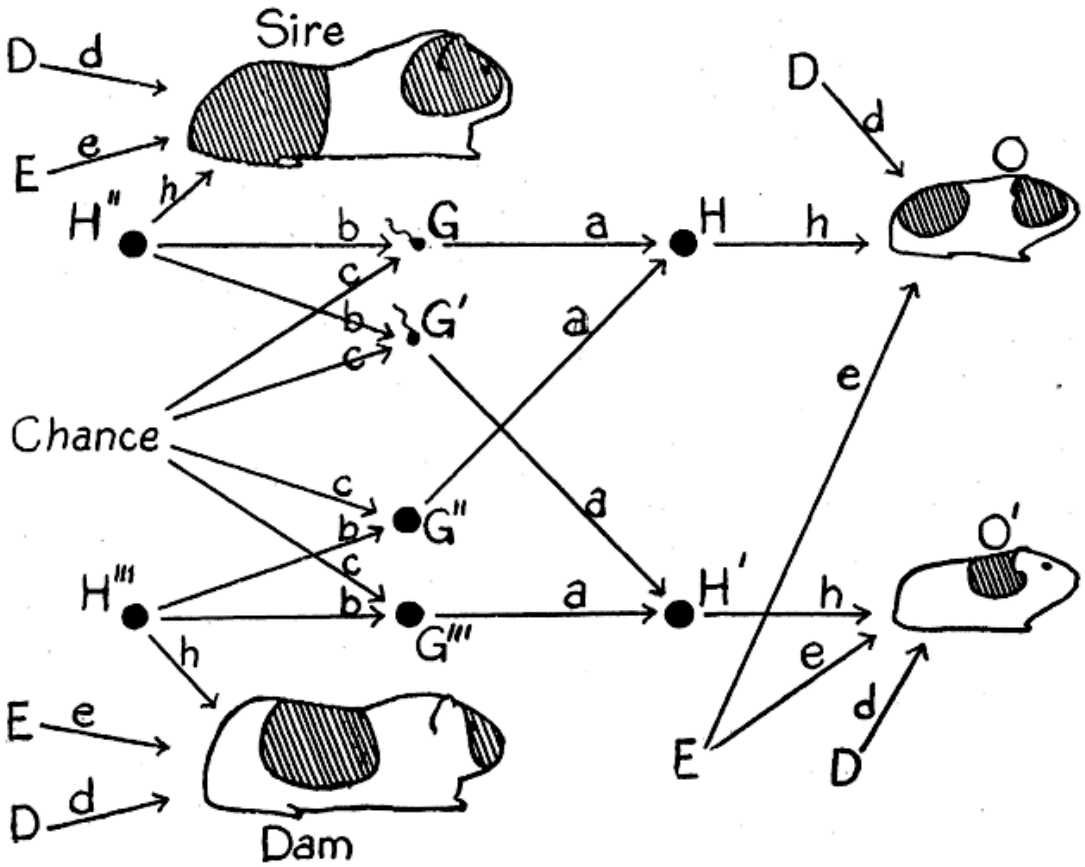
- For 30 years the method of path coefficients took a back seat to ANOVA, multiple regression, factor analysis and other statistical methods.
- George Link wrote about his hikes with Wright, “In the winter Benjie, Sewall, Henry, and I often walked from Tremont to Michigan City and back along the beach where the sand and snow offered excellent facilities for execution of biologic diagrams, curves, and equations of all sorts.”

Duncan (1966)

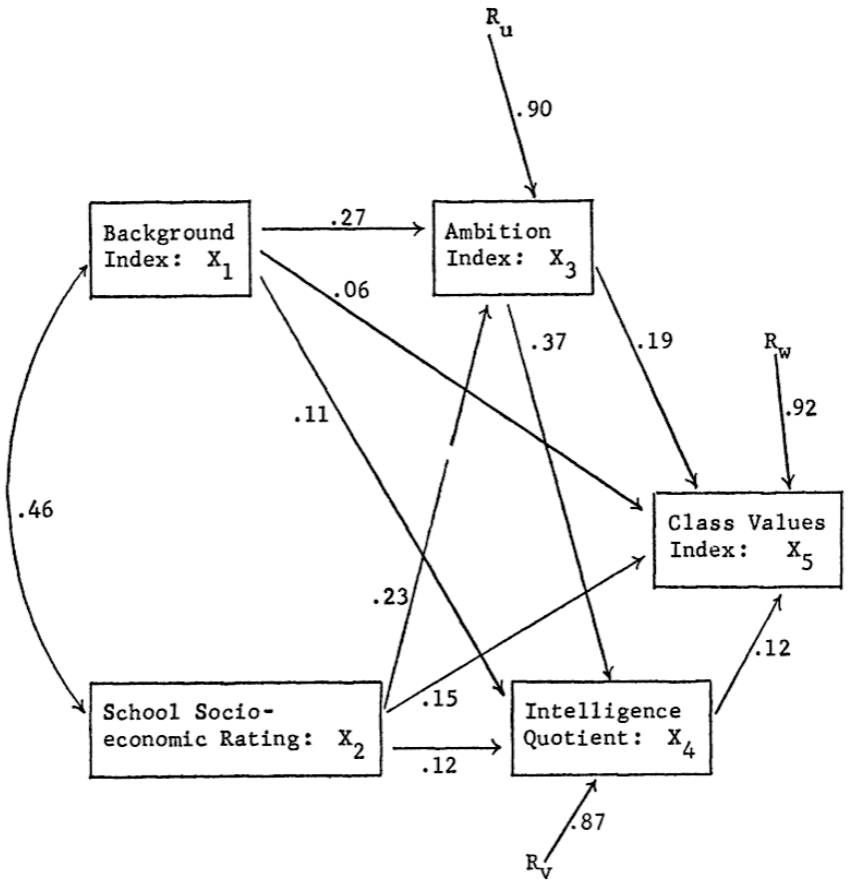
- Duncan brought path analysis to the attention of social sciences researchers in a widely cited overview (1966).
- Duncan realized that methods of estimation ("the inverse problem") for what would soon be called structural equation models were lacking.
- He also recognized the need for a principled system of graphic representation these models.

Duncan (1966)

"...the diagrammatic representation of such a system is of great value in thinking about its properties. A word of caution is necessary, however. Causal diagrams are appearing with increasing frequency in sociological publications. Most often these have some kind of pictorial or mnemonic function without being isomorphic with the algebraic and statistical properties of the postulated system of variables—or, indeed without having a counterpart in any clearly specified system of variables at all."



Clearly specified variables? (Wright, 1920)



An example path diagram (cited by Duncan, 1966)

Goldberger (1972)

- Goldberger published a review (1972) in order to "redress economists' neglect of the work of Sewall Wright."
- Goldberger helped bring together the psychometric community (including Jöreskog), the econometric community (including Wold) and the sociological community (including Duncan) to discuss structural equation models.
- At the first of these meetings (in 1970) Jöreskog presented the basis for LISREL.

Jöreskog (1973)

- Karl Jöreskog's LInear Structural RELations model was the foundation for the explosion in SEM.
- Dag Sörbom programmed the first version of LISREL.
- A partial list of SEM packages (in alphabetical order)
  - Amos
  - Calis
  - EQS
  - OpenMx
  - Mplus
  - Mx
  - Ramona
  - sem
  - SEPATH

McArdle & McDonald (1984)

- McArdle & McDonald proposed a general method for the analysis of moment structures that was based on Sewall Wright's path tracing rules.
- McArdle noticed that there was a general rule by which Wright's tracing rules could be simplified.
- Part of this algorithm required finding the limit of a power series which McDonald noticed had a relatively simple solution.

RAMpath (McArdle & Boker, 1990)

- At the time, we were using either costly mainframes or Apple IIs.
- In 1982, I developed a sparse matrix algorithm that calculated expected covariances without taking an inverse, but was not guaranteed to converge.
- In order solve the non-convergence problem, in 1984, I took McArdle's tracing rules and created a linked list solution to quickly test for convergence, the basis of RAMpath and the graphical interface for Mx.
- RAMpath was the first software to automatically display path diagrams and decompose the components of covariance using tracing rules.

Some Recent Developments

- In 1987 McArdle rewrites RAMit as PROC CALIS for SAS, the first release of the RAM matrix solution.
- Mx (Neale, 1994) pioneered full information maximum likelihood, nonlinear constraints, and multiple groups.
- Amos (Arbuckle, 1997) was the first SEM package to incorporate automatic path model display.
- MPlus (Muthén, 1998) was the first to fit mixture distributions.
- MxWindows (Neale, Boker, Xie, & Maes, 1999) allowed path model fitting from a graphical user interface.
- OpenMx (Boker et al., 2009) is a full featured open source SEM package that runs under R.

What is Structural Equation Modeling?

- The general linear model.
- Principle components and factor analysis.
- Covariance algebra.
- Latent and manifest variables.
- Confirmatory factor analysis and the measurement model.
- Partitioning variance and components of covariance.
- Model expectations.
- Model fit and model selection with nested comparisons.
- Parameter standard errors.
- Parameter constraints.
- Multigroup models.
- Full information maximum likelihood and missing data.
- Latent variables for measurement of change and dynamics.

The general linear model

$$
\begin{aligned}
y_{i1} &= b_{01} + b_{11}x_{i1} + b_{21}x_{i2} + e_{i1} \\
y_{i2} &= b_{02} + b_{12}x_{i1} + b_{22}x_{i2} + e_{i1}
\end{aligned}
$$

$$
\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{N1} & y_{N2} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} \end{bmatrix} \cdot \begin{bmatrix} b_{01} & b_{02} \\ b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} + \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \\ \vdots & \vdots \\ e_{N1} & e_{N2} \end{bmatrix}
$$

$$
\mathbf{Y} = \mathbf{XB} + \mathbf{E}
$$

Principal components and exploratory factor analysis

- Principal components decomposes a data matrix into orthogonal components.
- Principal components assumes that measurement error is zero.
- Exploratory factor analysis attempts to estimate loadings for (possibly correlated) factors that are “pure” or “error free” constructs by partitioning a data matrix into common (shared) and unique parts.
- Both of these methods can estimate whether a parsimonious explanation can account for most of the data.

Covariance Algebra

$$
\begin{aligned}
\frac{1}{N}\mathbf{Z}\mathbf{Z}' &= \frac{1}{N}(\mathbf{A}\cdot\mathbf{X}+\mathbf{U})\cdot(\mathbf{A}\cdot\mathbf{X}+\mathbf{U})' \\
\mathbf{R} &= \frac{1}{N}(\mathbf{A}\cdot\mathbf{X})(\mathbf{A}\cdot\mathbf{X})' + \frac{1}{N}\mathbf{U}\cdot(\mathbf{A}\cdot\mathbf{X})' + \\
&\quad \frac{1}{N}(\mathbf{A}\cdot\mathbf{X})\cdot\mathbf{U}' + \frac{1}{N}\mathbf{U}\cdot\mathbf{U}' \\
\mathbf{R} &= \frac{1}{N}(\mathbf{A}\cdot\mathbf{X}\cdot\mathbf{X}'\cdot\mathbf{A}') + 0 + 0 + \frac{1}{N}\mathbf{U}\cdot\mathbf{U}' \\
\mathbf{R} &= \mathbf{A}\frac{1}{N}(\mathbf{X}\cdot\mathbf{X}')\cdot\mathbf{A}' + \mathbf{U}^2 \\
\mathbf{R}-\mathbf{U}^2 &= \mathbf{A}\cdot\mathbf{C}_{\mathbf{XX}}\cdot\mathbf{A}' \\
\mathbf{R}-\mathbf{U}^2 &= \mathbf{A}\cdot\mathbf{L}\cdot\mathbf{A}'
\end{aligned}
$$

Latent and manifest variables

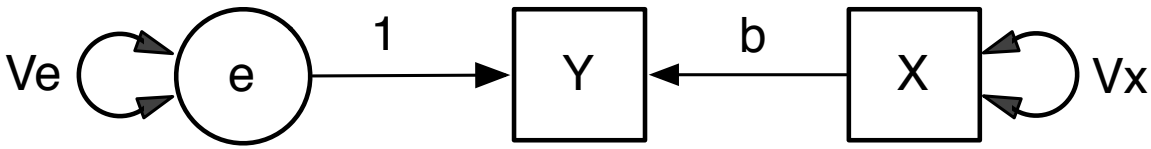
- Latent variables are unmeasured.
- So, we infer latent variables from a model.
- The latent variables relationship to measured (manifest variables) is called the “measurement model”.
- Latent variables can either be a construct that represents either shared variance or unique variance.
- One type of latent variable is a factor.

Confirmatory factor analysis and the measurement model

- Confirmatory factor analysis is a way of performing factor analysis with statistical tests and measurement constraints.
- Factor loadings tell us what a factor (latent variable) is measuring.
- With confirmatory factor analysis, we make statistical tests by constraining loadings and/or covariances between factors.
- We can estimate covariances or regression coefficients between latent variables, but we still don't have factor scores.

Partitioning variance and components of covariance

- The covariance or variance between any two variables in a path diagram can be partitioned into a list of components of covariance.
- This can be done via tracing rules that can be solved in the general case.
- We will learn the tracing rules and how they relate to the covariance algebra.



Model expectations

$$\mathcal{E}(\mathbf{C}_{xx}) = \mathbf{F}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{S}((\mathbf{I} - \mathbf{A})^{-1})'\mathbf{F}'$$

where if there are p manifest variables and q total variables (manifest plus latent),

- $\mathbf{I}$ is a $q \times q$ identity matrix.
- $\mathbf{A}$ is a $q \times q$ matrix of asymmetric elements (arrows) in the path diagram.
- $\mathbf{S}$ is a $q \times q$ matrix of symmetric elements (variances and covariances) in the path diagram.
- $\mathbf{F}$ is a $p \times q$ filter matrix.

This works because $(\mathbf{I} - \mathbf{A})^{-1}$ is the solution to the geometric series $(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \mathbf{A}^4 + \cdots$

Model fit and model selection

- An SEM model with a given set of parameters predicts an expected covariance matrix.
- Our data gives us an observed covariance matrix.
- We can calculate the difference between the observed and expected covariance matrix and call it the "model fit".
- Model parameters are generally estimated by minimizing the difference between the expected and observed covariance matrices.
- There are many ways to calculate the difference between two matrices and so there are many types of model fit.
- We want to understand how well the model fits given how parsimonious the model is, so the difference is often penalized for model complexity.

Parameter standard errors

- Parameter estimates are dependent on the chosen model as well as the particular data sample.
- A different data sample may give different parameter estimates.
- Parameter standard errors give an estimate of the width of the distribution of parameters if we were to draw many samples from the same population and refit the model.
- Bootstrapping is commonly used to make these estimates, although there are also analytic methods.

Parameter constraints

- Sometimes we want to constrain a parameter to be a particular value.
- Perhaps we know that a parameter must be 0.5 due to some prior knowledge.
- Sometimes we want to constrain a parameter to be within certain bounds.
- For instance, variances should always be positive.
- Sometimes we want to constrain a parameter to be a nonlinear function.
- Modern SEM packages allow all of these constraints.

Multigroup models

- Suppose we want to fit a model to two groups, perhaps males and females.
- Perhaps we want to know whether a particular parameter value is the same in the two groups.
- We can use standard errors of parameters to see if the parameter distributions overlap.
- But much better, we can fit a two group model and constrain the parameters to be equal or free and see what the difference in fit is.

Full information maximum likelihood and missing data

- Suppose some rows in our data matrix are not complete.
- Perhaps some questions on a questionnaire were skipped by some people.
- What do we do? Delete those rows? Impute data?
- One way to deal with this problem is with full information maximum likelihood (FIML).
- FIML calculates the likelihood of the data independently for each row, so each row contributes to the overall model fit using the part of the data that exists.
- This is a very efficient way of dealing with data that are missing at random.

Latent variables for measurement of change and dynamics

- There are many SEM techniques for the measurement of change and dynamics.
- We will cover a few such as Latent Difference Scores, Latent Growth Curves, and Latent Differential Equations.
- We don't have much time for each type, but I hope to give you a head start on doing this type of modeling.
- There will be a Dynamical Systems Analysis class next fall that will go into these models in more detail.

When to use Structural Equation Modeling?

- You must have
  - Multiple variables or occasions per person.
  - Many observations or people.
  - Several theories to compare.
- You might have
  - Latent constructs that can not be directly measured.
  - Complicated data that theoretically should have a more simple explanation.
  - Complex or nonrandom sampling.
  - Complex patterns of missingness.
  - Constraints to place on regression coefficients.
  - Multiple groups within the population.
  - Time dependence in longitudinal data.
  - Cohort sequential data.

Next Week

- Review of the basics of matrix algebra.

Arbuckle, J. L. (1997). *Amos user's guide. version 3.6*. Chicago: SPSS.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2009). *Openmx: Multipurpose software for statistical modeling.* (University of Virginia, Department of Psychology, Box 400400, Charlottesville, VA 22904. http://openmx.psyc.virginia.edu)

Duncan, O. D. (1966). Path analysis: Sociological examples. *The American Journal of Sociology*, *72*(1), 1–16.

Goldberger, A. S. (1972). Structureal equation methods in the social sciences. *Econometrica*, *40*(6), 979–1001.

Jöreskog, 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: Seminar.

McArdle, J. J., & Boker, S. M. (1990). *Rampath*. Hillsdale, NJ: Lawrence Erlbaum.

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the Reticular Action Model for moment structures. *British Journal of Mathematical and Statistical Psychology*, *87*, 234–251.

Muthén, B. O., L. K. & Muthén. (1998). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Neale, M. C. (1994). *Mx: Statistical modeling.* (Box 710 MCV, Richmond, VA 223298: Department of Psychiatry. 2nd Edition)

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). *Mx: Statistical modeling.* (Box 126 MCV, Richmond, VA 23298: Department of Psychiatry, 5th Edition)

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, *187*, 253–318.

Spearman, C. (1904a). General intelligence objectively

determined and measured. *American Journal of Psychology*, *15*, 201–293.

Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Wright, S. (1918). On the nature of size factors. *The Annals of Mathematical Statistics*, *3*, 367–374.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea–pigs. *Proceedings of the National Academy of Sciences*, *6*, 320–332.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*, 161–215.