STATISTICAL HORIZONS

# Categorical Structural Equation Modeling

Kevin Grimm, Ph.D.

*Upcoming Seminar:*

May 6-8, 2021, Remote Seminar

# Categorical Structural Equation Modeling

Kevin J. Grimm

Statistical Horizons

# Categorical Structural Equation Modeling

- Structural equation modeling (SEM) is a powerful data analytic framework for testing theories with statistical models and data
- Statistical models can be complex
  - *Multivariate* models with multiple outcomes
  - *Direct* and *indirect* effects
  - *Latent* variables
  - *Multiple groups* to examine invariance of model parameters
  - *Longitudinal* models to examine change over time

# Categorical Structural Equation Modeling

- Historically, SEM was a *linear* modeling framework
  - Models were additive
  - Data were assumed to have a multivariate normal distribution
  - Essentially, an extension of the *linear regression modeling* framework
- Early work by Muthén (and others) extended SEM to involve binary and ordinal outcomes
  - Introduced nonlinear link functions into SEM and developed estimation approaches
  - Subsequently, extended to handle additional types of non-normal data (*count* & *nominal* outcomes)

# Day 1: Structural Equation Modeling with *Binary* Outcomes

- Introduction to structural equation modeling
- Review of logistic and probit regression in `R`
- Introduction to `Mplus` and `lavaan` notation
- Logistic and probit regression in `Mplus` and `lavaan`
- Path models with binary mediators and outcomes
- Confirmatory factor models with binary indicators
- Model fit for maximum likelihood and weighted least squares estimators

# Day 2: Structural Equation Models for Ordinal Outcomes

- Cumulative logit and probit regression models in `R`
- Cumulative logit and probit regression models in `Mplus` and `lavaan`
- Confirmatory factor models with ordinal indicators
- Latent variable path models with binary and ordinal indicators
- Multiple group analysis with binary and ordinal indicators
- Missing data handling with maximum likelihood and weighted least squares estimators

# Day 3: Longitudinal Models & Count Outcomes

- Latent growth models with binary and ordinal outcomes
- Survival analysis
- Review of count regression models in `R`
- Count regression models in `Mplus`
- Zero-inflated count regression models in `Mplus`

Introduction to *Structural Equation Modeling*

# Preliminary Steps

- Specify research question in terms of **constructs**, direct, indirect, and symmetric associations
- Choose manifestations of the constructs (appropriate representative samples, timing, etc.)
- Examine measurement properties of the manifest (observed) variables
- Examine univariate and multivariate distributions of all manifest variables

# Steps in Structural Equation Modeling

1. **Theory-Data**: Form some basic ideas merging theory and data
2. **Specification**: Form explicit hypotheses regarding the associations among variables in terms of a path model
3. **Estimation**: Use SEM programs to estimate parameters, standard errors, and various indicators of model fit
4. **Evaluation & Interpretation**: Examine the **fit** of the model, potentialy compare the fit of the proposed model to **alternative** models, and interpret the model parameters
5. **Re-evaluation & Extension**: Explore new ideas/models

# Step #1: Theory-Data

- “The purpose of statistical procedures is to assist in establishing the plausibility of a theoretical model” (Cooley, 1978)
  - SEM is a general statistical framework that allows researchers to be **explicit** about theory and how it is reflected in one’s data
- Statistical models are where theories and data **collide**
  - Statistical models invoke a particular notion of reality that may or may not match one’s theoretical ideas
- Goal is to match theory and model **as close as possible** and examine plausibility of model given the data

# Step #1: Theory-Data

- SEM is a **confirmatory** framework for testing an *a-priori* hypothesis about the structure of the data
- Requires specific expectations regarding
  - One's theory
  - How one's theory is reflected by the particular structural equation model

# Step #2: Model Specification

- Specify a model that matches the theory to be tested
- Statistical model carries a set of expectations (expected variances & covariances) to test against the observed data

# Step #2: Model Specification

- Make sure model is identified
  - There exists a unique solution (locally and globally)
  - Particularly important in latent variable models

# Step #2: Model Specification

- Model Identification
  - Determine the total degrees of freedom available from the data
  - The total degrees of freedom indicates how many parameters we can estimate
  - Total degrees of freedom is equal to the number of unique pieces of information from the data
    - For our purposes, this is the number of variances and unique covariances in the covariance matrix

# Step #3: Model Estimation

- There are several SEM programs (e.g., `Mplus`, `lavaan`, `OpenMx`, `Lisrel`, etc.) to estimate the model parameters
- **Maximum Likelihood** estimation is typically used to estimate the model's parameters
  - The maximum likelihood estimates (for multivariate normal data) are those that minimize the maximum likelihood fit function ($F_{ML}$)

$$F_{ML} = \log|\boldsymbol{\Sigma}\left(\hat{\theta}\right)| + \text{tr}\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\left(\hat{\theta}\right)\right) - \log|\mathbf{S}| - (p+q)$$

  where
    - $\boldsymbol{\Sigma}\left(\hat{\theta}\right)$ is the model-implied covariance matrix with current estimates (i.e., $\hat{\theta}$)
    - $\mathbf{S}$ is the observed covariance matrix
    - $(p+q)$ is the number of observed variables

# Step #4: Model Evaluation & Interpretation

- Examine the **fit** of the model with respect to the model's $\chi^2$ and degrees of freedom to determine if the model fits significantly worse than a **perfect fitting model**

- Examine the **fit** of the model with respect to other model fit statistics
  - Root Mean Square Error of Approximation (RMSEA)
  - Comparative Fit Index (CFI)
  - Tucker-Lewis Index (TLI)
  - Standardized Root Mean Square Residual (SRMR)

- Examine the **residual** covariance matrix

# Step #4: Model Evaluation & Interpretation

- Interpret the parameter estimates obtained from the model
  - Primarily focused on the interpretation of path coefficients
    - Direct, Indirect, & Total Effects
  - Can also discuss explained variance for outcomes

# Step #5: Re-evaluation & Extension

- Consider alternative models based on the results of the fitted model(s)
- **Important Note**
  - Must be discussed as exploratory and preliminary when the newly generated models are based on the results of previously fit models to the same data