Harvard Catalyst Biostatistical Seminar

Neuropsychological Profiles in Alzheimer’s Disease and Cerebral Infarction: A Longitudinal MIMIC Model

An Overview of Structural Equation Modeling using Mplus

Richard N. Jones, Sc.D.
jones@hsl.harvard.edu

Institute for Aging Research, Hebrew SeniorLife
Beth Israel Deaconess Medical Center, Harvard Medical School

HSPH Kresge G2 October 5, 2011
Objective

- Introduce
  - the concepts and terminology relevant to structural equation modeling (SEM) as applied to health research

- Specific Example
  - Cognitive Epidemiology
  - Mplus software

- Emphasis
  - on a broad survey of applications, results, challenges, and opportunities

1 www.statmodel.com
Mplus and SEM Trends in Epidemiology

- Relative to the frequency with which *Cox Regression* and *Epidemiology* appear in Google Scholar...
- Citations matching *Mplus* and *Epidemiology* are increasing
- Although specific applications are decreasing
- Mplus use is increasing and applications are becoming more diverse
Mplus: General statistical analysis software good for...

- Analysis with latent variables
- Clustered and correlated data
  - Complex sampling, weighting
  - Repeated measures
  - Multicomponent variables (i.e., scales, composite outcomes)
  - Correlated observations (e.g., twins, families)
  - Multilevel contexts
- Particular strengths
  - Missing data modeling
  - Bayesian data analysis
  - Complex models
    - Joint models of change and event occurrence
    - Mixture models (population heterogeneity)
    - Longitudinal factor analysis
- Where Mplus is not strong
  - Data management
  - Graphics
Structural Equation Modeling in Epidemiologic Research

Practice of Epidemiology

Statistical Issues in Life Course Epidemiology

Bianca L. De Stavola¹, Dorothea Nitsch¹, Isabel dos Santos Silva¹, Valerie McCormack¹, Rebecca Hardy², Vera Mann¹, Tim J. Cole³, Susan Morton¹, and David A. Leon¹

¹ Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom.
² MRC National Survey of Health and Development, Department of Epidemiology, Royal Free and University College Medical School, London, United Kingdom.
³ Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, University College London, London, United Kingdom.

Received for publication September 9, 2004; accepted for publication July 25, 2005.
STRUCTURAL EQUATION MODELING IN EPIDEMIOLOGIC RESEARCH

DeStavola et al (2005), bottom panel Figure 3

FIGURE 3. Example of path diagrams. Top: path for one distal outcome (Y), one intermediate outcome (X₃), and two background variables (X₁, X₂); bottom: path for one distal outcome (Y) and a latent variable (U) measured by three proxy variables (X₁, X₂, X₃). Arrows depicting random variation for each variable are omitted for simplicity. Boxes, proxy variables; circle, latent variable.
Structural Equation Modeling (SEM) is

- A general multivariate regression modeling framework
  - General - flexible model types
  - Multivariate - multiple dependent variables
  - Regression - it’s just regression. Regression can be viewed as a special case of SEM

- SEMs often include *latent variables*
  - Continuous latent variables (i.e., factors)
  - Categorical latent variables (i.e., classes, mixtures)
### Varieties of Covariance Structure Modeling

<table>
<thead>
<tr>
<th>Continuous Latent Variables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td>Regressions among</td>
<td>Yes</td>
</tr>
<tr>
<td>Dependent or Latent Variables</td>
<td>No</td>
</tr>
<tr>
<td><strong>No</strong></td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td>Regression (Multivariate)</td>
<td>Factor Analysis</td>
</tr>
<tr>
<td>$y = \nu + \Gamma X + \epsilon$</td>
<td>$y = \nu + \Lambda \eta + \epsilon$</td>
</tr>
<tr>
<td>Path Analysis</td>
<td>Structural Equation Modeling</td>
</tr>
</tbody>
</table>
| $y = \nu + BY + \Gamma X + \epsilon$ | $y = \nu + \Lambda \eta + KX + \epsilon$
| $\eta = \alpha + B \eta + \Gamma X + \zeta$ |
SEM Prerequisites

- General linear model
  - Linear, logistic, & probit regression
  - Multivariable regression
  - Mixed effect models for longitudinal data
  - Survival and event occurrence (Cox, parametric survival)

- Missing data theory

- Factor Analysis

- Item Response Theory

- Path Analysis

- Structural Equation Modeling
Mplus Workflow

- **Introduction**

- **What you have to look forward to**

### Mplus Workflow Diagram

1. **Study Data**
   - **Preprocess**
     - **ASCII Data**
     - **Command File**
     - **Mplus**
     - **Output & Inferences**
     - Write the Paper

### Tools
- SAS, SPSS, R/S-Plus, STATA

### Data Handling
- **Clean data, handle missingness**
- **Select cases, variables**
- **Transformations**
- **Descriptive statistics**

### Data Format
- **A text file with selected data elements.** Comma delimited works best for Mplus

### Additional Information
- Also a raw text (ASCII) file
- Instructions for a single analysis
Mplus Workflow

- **Introduction:**

  - But life can be better

- **Mplus Workflow Diagram:**

  - **Study Data**
    - Preprocess
      - Clean data, handle missingness
      - Select cases, variables
      - Transformations
      - Descriptive statistics
  
  - **Analysis**
    - Integrated Mplus
      - STATA, R/MplusAutomation
        - (SAS, SPSS - roll your own)
      - Mplus analysis integrated with conventional analysis, in single syntax file

- **Output & Inferences**

- **Write the Paper**
Mplus Workflow - Weaving = Reproducible Research
Multivariate Regression

\[ y = \nu + \Gamma X + \epsilon \]
Multivariate Regression

Mplus Syntax

```
TITLE: Multivariate regression
DATA: FILE = data.dat ;
VARIABLE: NAMES = y1 y2 x1 x2 ;
MODEL: y1 y2 on x1 x2 ;
```
Confirmatory Factor Analysis

\[ y = \nu + \Lambda \eta + \epsilon \]
Confirmatory Factor Analysis

Mplus Syntax

TITLE: Confirmatory Factor Analysis
DATA: FILE = data.dat ;
VARIABLE: NAMES = y1 y2 y3 ;
MODEL: eta by y1 y2 y3 ;
Structural Equation Modeling

\[ y = \nu + \Lambda \eta + KX + \epsilon \]
\[ \eta = \alpha + B\eta + \Gamma X + \zeta \]
Structural Equation Modeling

Mplus Syntax

TITLE: Structural Equation Model
DATA: FILE = data.dat ;
VARIABLE: NAMES = y1-y6 x1 ;
MODEL: eta1 by y1-y3 ; ! measurement model for eta1
eta2 by y4-y6 ; ! measurement model for eta2
eta2 on eta1 ; ! a structural regression
eta1 on x1 ; ! an "indirect effect"
y1 on x1 ; ! a "direct effect"
Structural Equation Modeling

Mplus Syntax

```
. runmplus y1-y6 x1 , model(eta1 by y1-y3 ; eta2 by y4-y6 ; ///
    eta2 on eta1 ; eta1 on x1 ; y1 on x1 ;)
```
What Latent Variables Are

- Latent variables are mathematical abstractions that account for covariation among observed variables
- Latent variables may be continuous or categorical
- But what do they mean?
What is the Meaning Behind a Latent Variable?

- The answer depends on the
  - scientific question
  - philosophical position

- Two broad classes of latent variable (LV) applications
  - Instrumentalist
    - the LV is a mathematical abstraction
  - Realist
    - the LV exists
    - the LV reflects some unmeasurable quantity or quality that really exists in nature
    - the LV exists independently of our measurement of it

- Realist or Instrumentalist interpretations are a matter of statistical inference

---

Borsboom, Mellenbergh et al., 2003 Psychol Rev 110:203-18
Parting Words

Why use SEM?
- More easily specify analysis to answer research question
- Gain statistical power

Not sure if SEM is right for you?
- Stata Corp recently added SEM to Stata (version 12)
- ... and pitching SEM to Economists

Why use Mplus?
- Regression-based framework for exogenous variables
- Categorical dependent variables
- Categorical latent variables
- Complex sampling weights
- Survival analysis
- Bayesian data analysis
## Google Scholar Hits 1998-2011 (5 Oct 2011)

<table>
<thead>
<tr>
<th>Keyword phrase</th>
<th>NEJM</th>
<th>JAMA+</th>
<th>AJE</th>
</tr>
</thead>
<tbody>
<tr>
<td>analysis</td>
<td>10,000</td>
<td>16,000</td>
<td>4,500</td>
</tr>
<tr>
<td>logistic regression</td>
<td>1,000</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>linear regression</td>
<td>620</td>
<td>1,700</td>
<td>1,700</td>
</tr>
<tr>
<td>cox proportional</td>
<td>610</td>
<td>760</td>
<td>650</td>
</tr>
<tr>
<td>random effects</td>
<td>100</td>
<td>540</td>
<td>340</td>
</tr>
<tr>
<td>generalized estimating</td>
<td>110</td>
<td>280</td>
<td>320</td>
</tr>
<tr>
<td>factor analysis</td>
<td>90</td>
<td>190</td>
<td>170</td>
</tr>
<tr>
<td>structural equation</td>
<td>21</td>
<td>25</td>
<td>49</td>
</tr>
<tr>
<td>profile mixture OR latent class</td>
<td>13</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>item response theory</td>
<td>2</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>path analysis</td>
<td>5</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>latent growth curve</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>SAS Institute</td>
<td>440</td>
<td>2,500</td>
<td>1,400</td>
</tr>
<tr>
<td>SPSS</td>
<td>210</td>
<td>1,100</td>
<td>240</td>
</tr>
<tr>
<td>Stata</td>
<td>190</td>
<td>800</td>
<td>590</td>
</tr>
<tr>
<td>R Foundation OR R Development</td>
<td>36</td>
<td>100</td>
<td>87</td>
</tr>
<tr>
<td>Mplus</td>
<td>0</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>LISREL</td>
<td>1</td>
<td>3</td>
<td>13</td>
</tr>
</tbody>
</table>

**Note:** Hits include the text matches in the reference list.

Values greater than 100 rounded to two significant digits.