Center for **Causal** Discovery:

**Summer Short Course/Datathon - 2016**

June 13-18, 2015

Carnegie Mellon University
Outline

Models → Data
1) Representing/Modeling Causal Systems
2) Estimation and Model fit
3) Hands on with Real Data

Models ← Data
1) Markov Axiom and D-separation
2) Model Equivalence
3) Model Search
Standardized SEMs

1) Attach a SEM PM to your 3-4 variable graph

2) Attach a SEM IM to the SEM PM

3) Change the coefficient values.

4) Attach a Standardized SEM IM to the SEM PM, or the SEM IM

5) Compare the Implied Matrices
Tetrad Demo & Hands-On
Generalized SEM

1) The Generalized SEM is a generalization of the linear SEM model.
2) Allows for arbitrary connection functions
3) Allows for arbitrary distributions
4) Simulation from cyclic models supported.

Causal Graph

SEM Equations:
- Education := \( \varepsilon_{\text{Education}} \)
- Income := \( \beta_1 \text{Education} + \varepsilon_{\text{Income}} \)
- Longevity := \( \beta_2 \text{Education} + \varepsilon_{\text{Longevity}} \)

\[ P(\varepsilon_{\text{ed}}, \varepsilon_{\text{Income}}, \varepsilon_{\text{Income}}) \sim N(0, \Sigma^2) \]

Generalized SEM Equations:
- Education := \( \varepsilon_{\text{Education}} \)
- Income := \( \beta_1 \text{Education}^2 + \varepsilon_{\text{Income}} \)
- Longevity := \( \beta_2 \ln(\text{Education}) + \varepsilon_{\text{Longevity}} \)

\[ P(\varepsilon_{\text{ed}}, \varepsilon_{\text{Income}}, \varepsilon_{\text{Income}}) \sim U(0, 1) \]
Hands On

1) Create a DAG.

2) Parameterize it as a Generalized SEM.

3) In PM – select from Tools menu “show error terms”
   Click on error term, change its distribution to Uniform

4) Make at least one function non-linear

5) Make at least one function interactive

6) Save the session as “generalizedSEM”.
Estimation
Estimation
Tetrad Demo and Hands-on

1) Select Template: “Estimate from Simulated Data”

2) Build the standardized SEM IM shown below

3) Generate simulated data N=1000

4) Estimate model.

5) Save session as “Estimate1”
Estimation
Coefficient inference vs. Model Fit

Coefficient Inference: Null: coefficient = 0, e.g., $\beta_{X_1 \rightarrow X_3} = 0$

$p$-value = $p(\text{Estimated value} \hat{\beta}_{X_1 \rightarrow X_3} \geq .4788 \mid \beta_{X_1 \rightarrow X_3} = 0 \& \text{rest of model correct})$

Reject null (coefficient is “significant”) when $p$-value < $\alpha$, $\alpha$ usually = .05
Coefficient inference vs. Model Fit

Coefficient Inference: Null: coefficient = 0, e.g., $\beta_{X1 \rightarrow X3} = 0$

$p$-value = $p(\text{Estimated value } \hat{\beta}_{X1 \rightarrow X3} \geq .4788 \mid \beta_{X1 \rightarrow X3} = 0 \& \text{rest of model correct})$

Reject null (coefficient is “significant”) when $p$-value $< \alpha$, $\alpha$ usually $= .05$,

Model fit: Null: Model is correctly specified (constraints true in population)

$p$-value = $p(\text{f(Deviation}(\Sigma_{ml}, S)) \geq 5.7137 \mid \text{Model correctly specified})$
# Coefficient inference vs. Model Fit

<table>
<thead>
<tr>
<th>p-value</th>
<th>Coefficient $\hat{\beta}_{X_1 \rightarrow X_3}$</th>
<th>Model fit $\chi^2_{df}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; .05</td>
<td>Can reject 0, Significant edge</td>
<td>Can reject correct specification, Model not correctly specified</td>
</tr>
<tr>
<td>&gt; .05</td>
<td>Can’t reject 0, insignificant edge</td>
<td>Can’t reject correct specification, model <em>may be</em> correctly specified</td>
</tr>
</tbody>
</table>

Null: $\beta_{X_1 \rightarrow X_3} = 0$

Null: Model is correctly specified
Model Fit

Specified Model

True Model

Implied Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>β1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>β1*β2</td>
<td>β2</td>
<td>1</td>
</tr>
</tbody>
</table>

Population Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>.6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>.3</td>
<td>.5</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \hat{\beta}_1 = r_{X1, X2} = \sim .6 \]

\[ \hat{\beta}_2 = r_{X2, X3} = \sim .5 \]

\[ \hat{\beta}_1 \hat{\beta}_2 = \sim .3 = \rho_{X1, X3} \]
Model Fit

Specified Model

<table>
<thead>
<tr>
<th></th>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(X_2)</td>
<td>(\beta_1)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(X_3)</td>
<td>(\beta_1\beta_2)</td>
<td>(\beta_2)</td>
<td>1</td>
</tr>
</tbody>
</table>

Implied Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(X_2)</td>
<td>0.6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(X_3)</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
</tbody>
</table>

Population Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(X_2)</td>
<td>0.2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(X_3)</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
</tbody>
</table>

 Unless \(r_{X_1,X_3} = r_{X_1,X_2} r_{X_2,X_3}\)

Estimated Covariance Matrix ≠ Sample Covariance Matrix
Model Fit

Specified Model

True Model

Implied Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>β1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>β1*β2</td>
<td>β2</td>
<td>1</td>
</tr>
</tbody>
</table>

Population Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>.6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>.32</td>
<td>.5</td>
<td>1</td>
</tr>
</tbody>
</table>

Model fit: Null: Model is *correctly specified* (constraints true in population)

\[ \rho_{X1,X3} = \rho_{X1,X2} \rho_{X2,X3} \]

p-value = \( p(f(\text{Deviation}(\Sigma_{ml}, S)) \geq \chi^2 \mid \text{Model correctly specified}) \)
Tetrad Demo and Hands-on

1) Create two DAGs with the same variables – each with one edge flipped, and attach a SEM PM to each new graph (copy and paste by selecting nodes, Ctl-C to copy, and then Ctl-V to paste)

2) Estimate each new model on the data produced by original graph

3) Check p-values of:
   a) Edge coefficients
   b) Model fit

4) Save session as: “estimation2”
Charitable Giving

What influences giving? Sympathy? Impact?

"The Donor is in the Details", Organizational Behavior and Human Decision Processes, Issue 1, 15-23, C. Cryder, with G. Loewenstein, R. Scheines.

N = 94

TangibilityCondition [1,0] Randomly assigned experimental condition
Imaginability [1..7] How concrete scenario I
Sympathy [1..7] How much sympathy for target
Impact [1..7] How much impact will my donation have
AmountDonated [0..5] How much actually donated
Theoretical Hypothesis
Tetrad Demo and Hands-on

1) Load charity.txt (tabular – not covariance data)
2) Build graph of theoretical hypothesis
3) Build SEM PM from graph
4) Estimate PM, check results
Foreign Investment

Does Foreign Investment in 3rd World Countries inhibit Democracy?


\[ N = 72 \]

PO degree of political exclusivity
CV lack of civil liberties
EN energy consumption per capita (economic development)
FI level of foreign investment
Case Study: Foreign Investment  Alternative Models

There is no model with testable constraints (df > 0) that is not rejected by the data, in which FI has a positive effect on PO.

Tetrad - PC

Fit: df=2, $\chi^2=0.12$, p-value = .94
Tetrad Demo and Hands-on

1) Load tw.txt (this IS covariance data)
2) Do a regression
3) Build an alternative hypothesis, Graph - SEM PM, SEM IM
4) Estimate PM, check results
Hands On
Lead and IQ

Lead: Lead concentration in baby teeth

CIQ: child’s IQ score at 7

PIQ: Parent’s average IQ

MED: mother’s education (years)

NLB: number of live births prior to child

MAB: mother’s age at birth of child

FAB: father’s age at birth of child
Hands On
Lead and IQ

1) Load leadiq1.tet

2) Specify different hypotheses, test the model fit on each

3) See if you can find a model (without using search), that is not rejected by the data