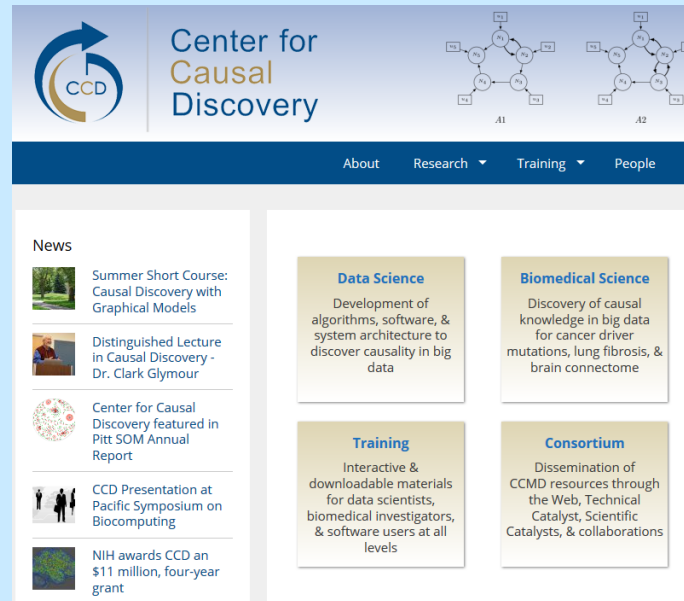


# Center for **Causal** Discovery:

## Summer Short Course/Datathon - 2016



The screenshot shows the homepage of the Center for Causal Discovery. The header features the CCD logo (a blue and yellow circular arrow) and the text "Center for Causal Discovery". To the right of the header are two causal diagrams labeled A1 and A2. Below the header is a navigation bar with links for "About", "Research", "Training", and "People". The main content area is divided into several sections:

- News:** A vertical list of five news items, each with a small image and a title. The items are: "Summer Short Course: Causal Discovery with Graphical Models", "Distinguished Lecture in Causal Discovery - Dr. Clark Glymour", "Center for Causal Discovery featured in Pitt SOM Annual Report", "CCD Presentation at Pacific Symposium on Biocomputing", and "NIH awards CCD an \$11 million, four-year grant".
- Data Science:** A box with the title "Data Science" and the text "Development of algorithms, software, & system architecture to discover causality in big data".
- Biomedical Science:** A box with the title "Biomedical Science" and the text "Discovery of causal knowledge in big data for cancer driver mutations, lung fibrosis, & brain connectome".
- Training:** A box with the title "Training" and the text "Interactive & downloadable materials for data scientists, biomedical investigators, & software users at all levels".
- Consortium:** A box with the title "Consortium" and the text "Dissemination of CCMD resources through the Web, Technical Catalyst, Scientific Catalysts, & collaborations".

June 13-18, 2015

Carnegie Mellon University

# Outline

## Models $\rightarrow$ Data

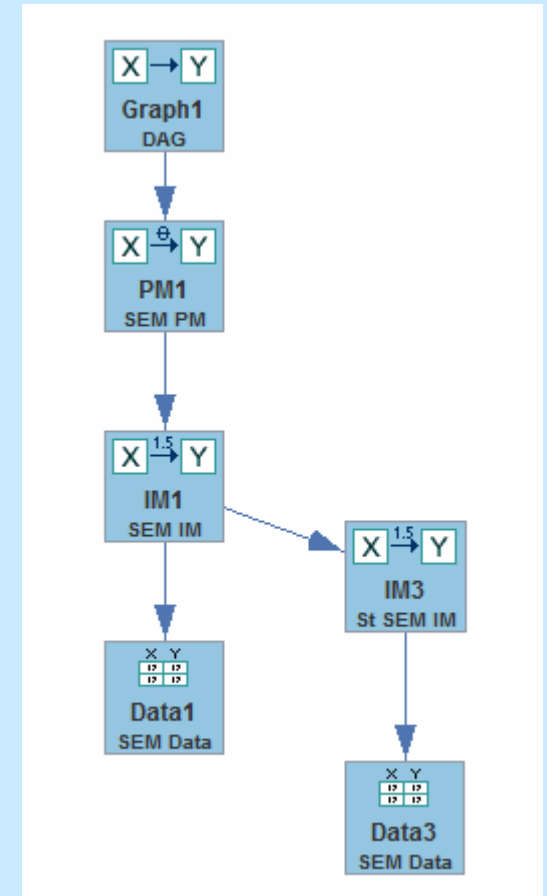
- 1) Representing/Modeling Causal Systems
- 2) Estimation and Model fit
- 3) Hands on with Real Data

## Models $\leftarrow$ 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

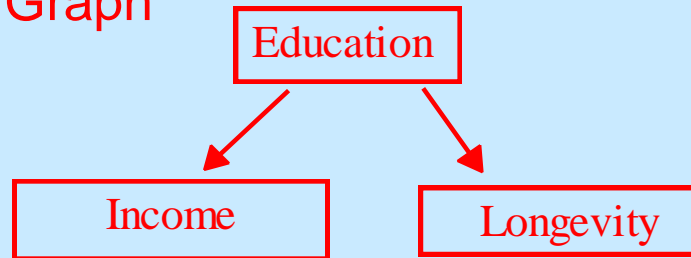


# Tetrad Demo & Hands-On

# 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:

$$\text{Education} := \varepsilon_{\text{Education}}$$

$$\text{Income} := \beta_1 \text{Education} + \varepsilon_{\text{Income}}$$

$$\text{Longevity} := \beta_2 \text{Education} + \varepsilon_{\text{Longevity}}$$

Generalized SEM Equations:

$$\text{Education} := \varepsilon_{\text{Education}}$$

$$\text{Income} := \beta_1 \text{Education}^2 + \varepsilon_{\text{Income}}$$

$$\text{Longevity} := \beta_2 \ln(\text{Education}) + \varepsilon_{\text{Longevity}}$$

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

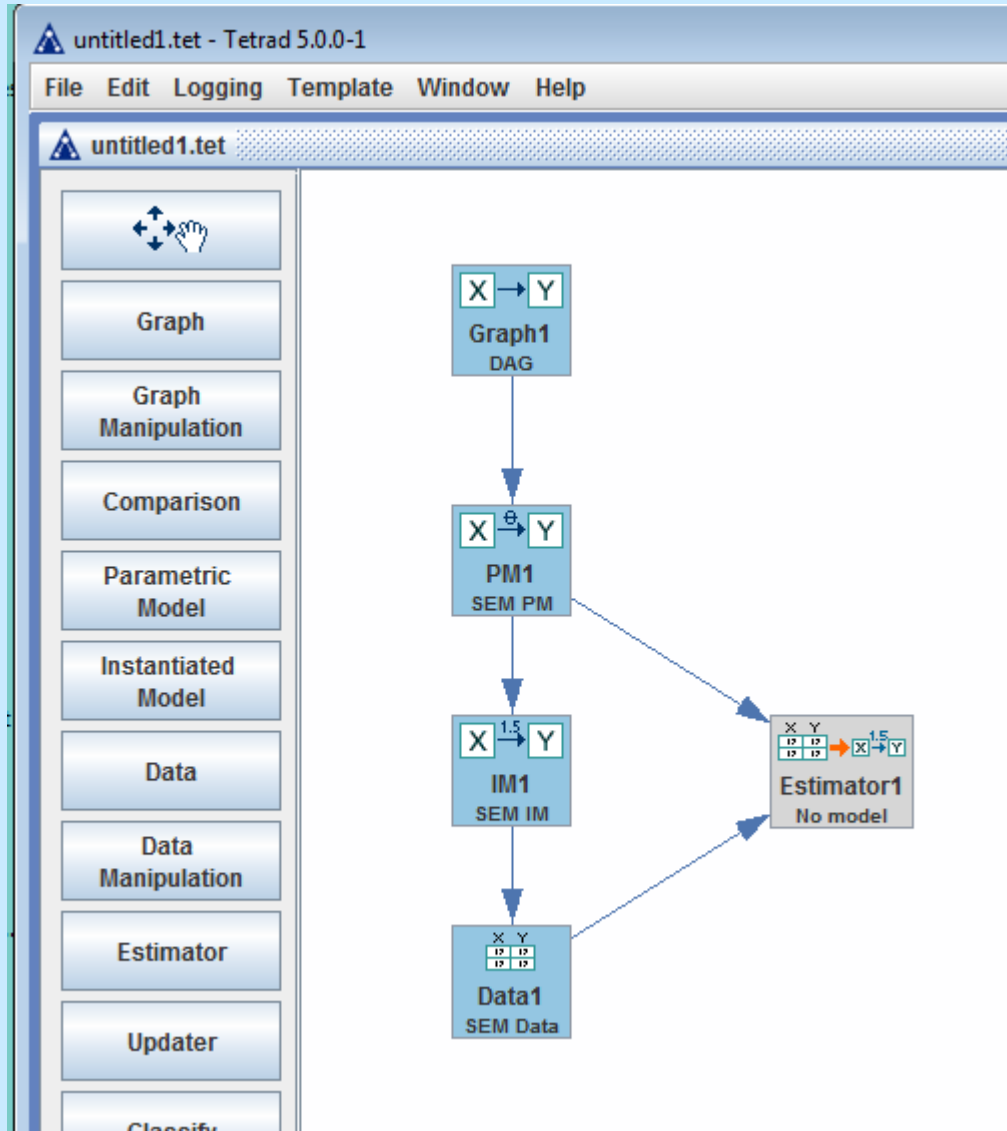
$$P(\varepsilon_{\text{ed}}, \varepsilon_{\text{Income}}, \varepsilon_{\text{Income}}) \sim U(0, 1)$$

# 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”.



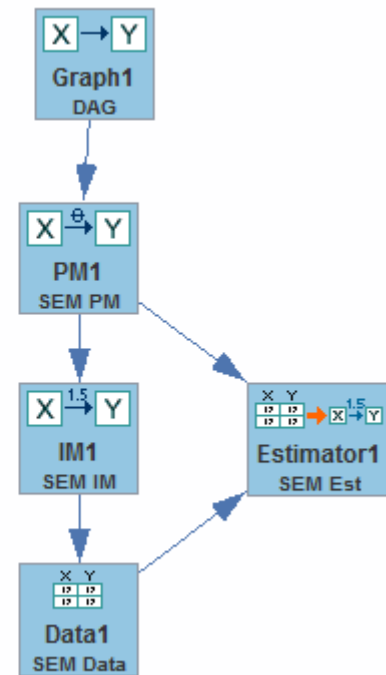
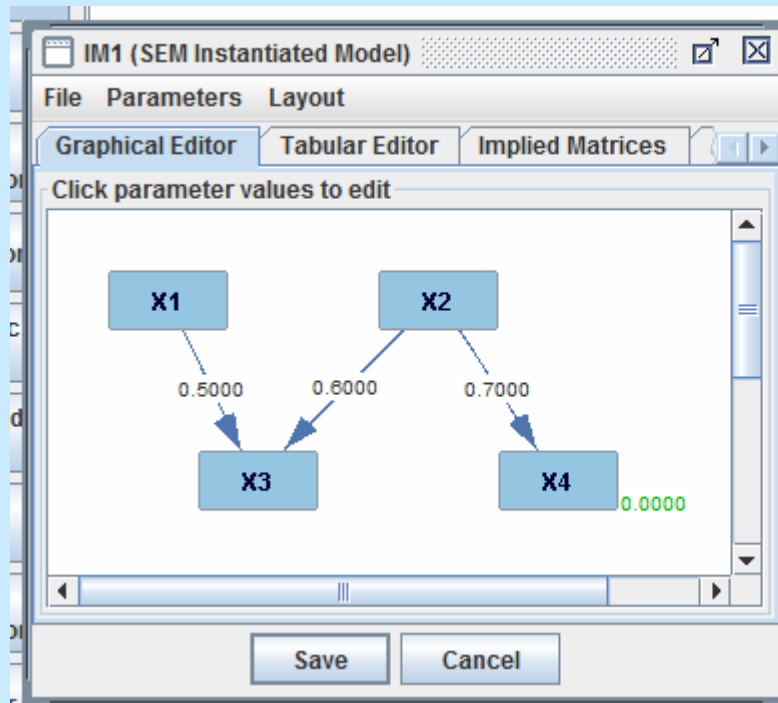
# Estimation



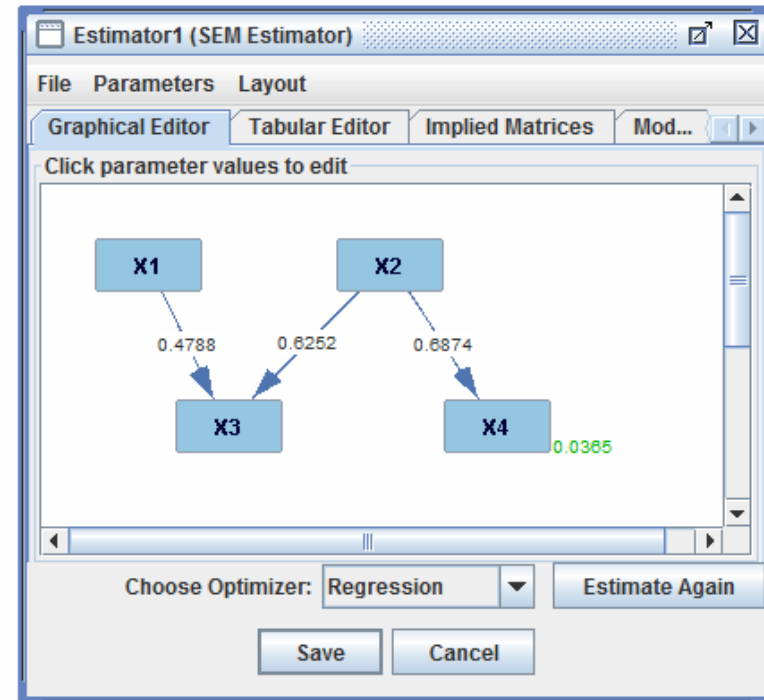
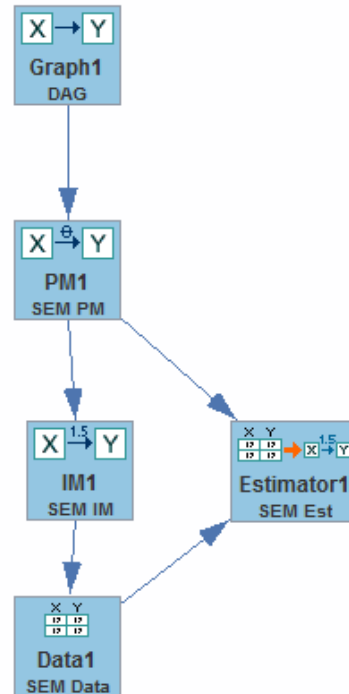
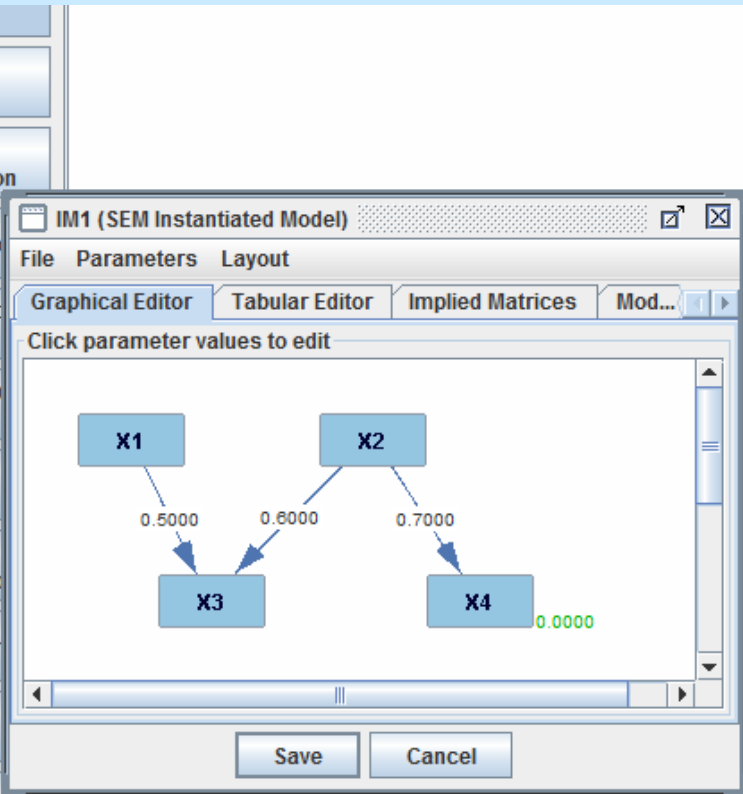


# Tetrad Demo and Hands-on

- 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.,  $\beta_{X1 \rightarrow X3} = 0$

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

The image displays three windows from a structural equation modeling (SEM) software interface:

- IM1 (SEM Instantiated Model) - Graphical Editor:** Shows a path diagram with variables X1, X2, X3, and X4. Path coefficients are: X1 to X3 (0.5000), X2 to X3 (0.6000), and X2 to X4 (0.7000). The coefficient for X3 to X4 is 0.0000.
- Graphical DAG (Directed Acyclic Graph):** A flowchart showing the relationships between model components: Data1 (SEM Data) feeds into IM1 (SEM IM), which feeds into PM1 (SEM PM), which feeds into Graph1 (DAG). Graph1 also feeds into Estimator1 (SEM Est).
- Estimator1 (SEM Estimator) - Model Statistics:** A table showing the null hypothesis for T and P that the parameter is zero. The first row is highlighted in red.

From	To	Type	Value	SE	T	P
X1	X3	Edge Coef.	0.4788	0.0334	14.3388	0.0000
X2	X4	Edge Coef.	0.6874	0.0303	22.7157	0.0000
X2	X3	Edge Coef.	0.6252	0.0316	19.7897	0.0000
X1	X1	Std. Dev.	0.9769	0.0427	22.3513	0.0000
X2	X2	Std. Dev.	1.0326	0.0477	22.3510	0.0000
X3	X3	Std. Dev.	1.0302	0.0475	22.3510	0.0000
X4	X4	Std. Dev.	0.9877	0.0436	22.3513	0.0000
X1	X1	Mean	-0.0320	0.0309	-1.0375	0.2998
X2	X2	Mean	-0.0233	0.0326	-0.7143	0.4752
X3	X3	Mean	0.0070	0.0415	0.1696	0.8654
X4	X4	Mean	0.0365	0.0384	0.9486	0.3431

# 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,

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

p-value =  $p(f(\text{Deviation}(\Sigma_{ml}, S)) \geq 5.7137 \mid \text{Model correctly specified})$

The image displays a sequence of software windows for Structural Equation Modeling (SEM). On the left, the 'IM1 (SEM Instantiated Model)' window shows a path diagram with variables X1, X2, X3, and X4. Path coefficients are 0.5000 (X1 to X3), 0.6000 (X2 to X3), 0.7000 (X2 to X4), and 0.0000 (X3 to X4). Below the diagram are 'Save' and 'Cancel' buttons.

In the center, a vertical flow of icons represents the workflow: 'Graph1 DAG' (Diagrammatic Approach Graph) points to 'PM1 SEM PM' (Path Model), which points to 'IM1 SEM IM' (Instantiated Model), which points to 'Data1 SEM Data' (Data). A box labeled 'Estimator1 SEM Est' (Estimator) is also shown, with arrows pointing to it from the 'PM1', 'IM1', and 'Data1' icons.

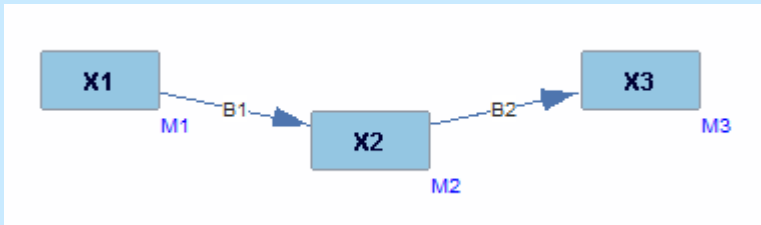
On the right, the 'Estimator1 (SEM Estimator)' window is open, showing fit statistics in a red-bordered box: Degrees of Freedom = 3, Chi Square = 5.7137, P Value = 0.1264, and BIC Score = -15.0095. Below this, a text box explains the chi-square test assumptions and formula:  $d = m(m+1)/2 - d$ , where  $d$  is the number of linear coefficients, variance terms, and error covariance terms that are not fixed in the model. At the bottom, there is a 'Choose Optimizer:' dropdown set to 'Regression' and an 'Estimate Again' button.

# Coefficient inference vs. Model Fit

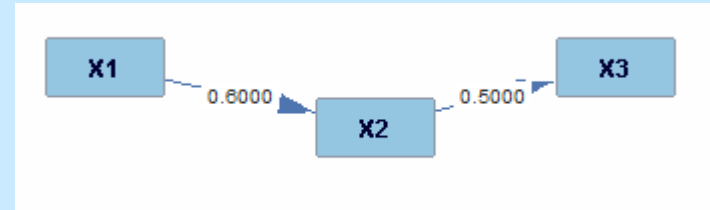
	coefficient $\hat{\beta}_{x_1 \rightarrow x_3}$	Model fit $\chi^2_{df}$
	Null: $\beta_{x_1 \rightarrow x_3} = 0$	Null: Model is correctly specified
p-value < .05	Can reject 0 Significant edge	Can reject correct specification, Model not correctly specified
p-value > .05	Can't reject 0, insignificant edge	Can't reject correct specification, model <i>may be</i> correctly specified

# Model Fit

Specified Model



True Model



Implied Covariance Matrix

	<b>X1</b>	<b>X2</b>	<b>X3</b>
X1	1		
X2	$\beta_1$	1	
X3	$\beta_1 * \beta_2$	$\beta_2$	1

Population Covariance Matrix

	<b>X1</b>	<b>X2</b>	<b>X3</b>
X1	1		
X2	.6	1	
X3	.3	.5	1

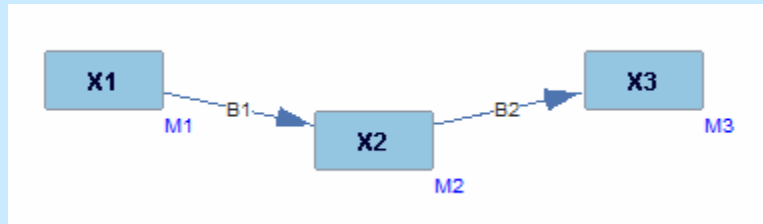
$$\hat{\beta}_1 = r_{X1,X2} = \sim .6$$

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

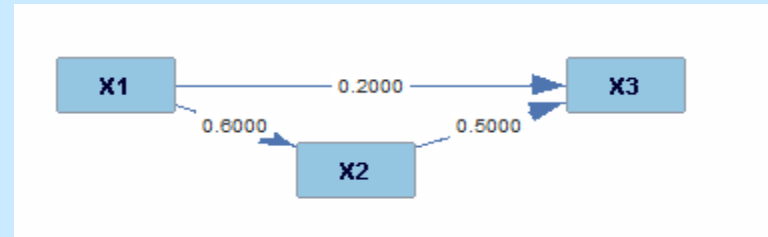
$$\hat{\beta}_1 \hat{\beta}_2 = \sim .3 = \rho_{X1,X3}$$

# Model Fit

Specified Model



True Model



Implied Covariance Matrix

	<b>X1</b>	<b>X2</b>	<b>X3</b>
X1	1		
X2	$\beta_1$	1	
X3	$\beta_1 * \beta_2$	$\beta_2$	1

Population Covariance Matrix

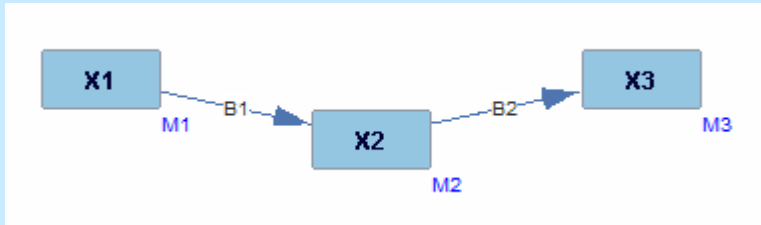
	<b>X1</b>	<b>X2</b>	<b>X3</b>
X1	1		
X2	.6	1	
X3	.5	.5	1

Unless  $r_{X1,X3} = r_{X1,X2} r_{X2,X3}$

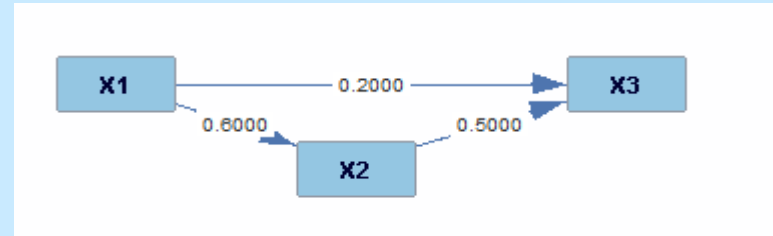
Estimated Covariance Matrix  $\neq$  Sample Covariance Matrix

# Model Fit

Specified Model



True Model



Implied Covariance Matrix

	<b>X1</b>	<b>X2</b>	<b>X3</b>
X1	1		
X2	$\beta_1$	1	
X3	$\beta_1 * \beta_2$	$\beta_2$	1

Population Covariance Matrix

	<b>X1</b>	<b>X2</b>	<b>X3</b>
X1	1		
X2	.6	1	
X3	.32	.5	1

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

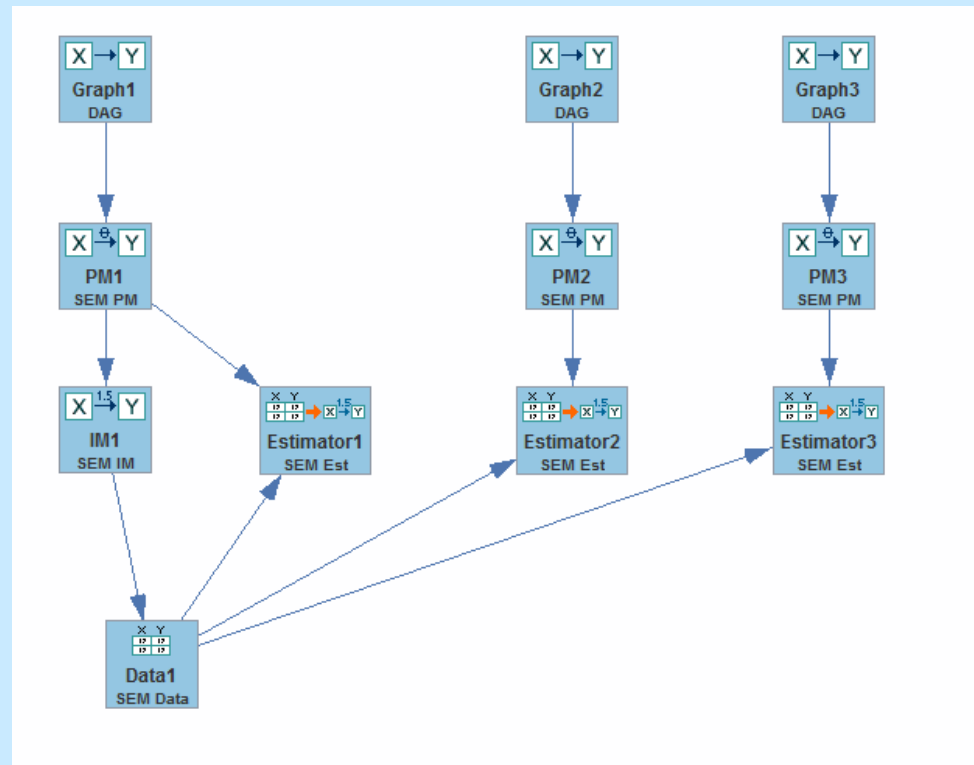
$$\rho_{X1,X3} = \rho_{X1,X2} \rho_{X2,X3}$$

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

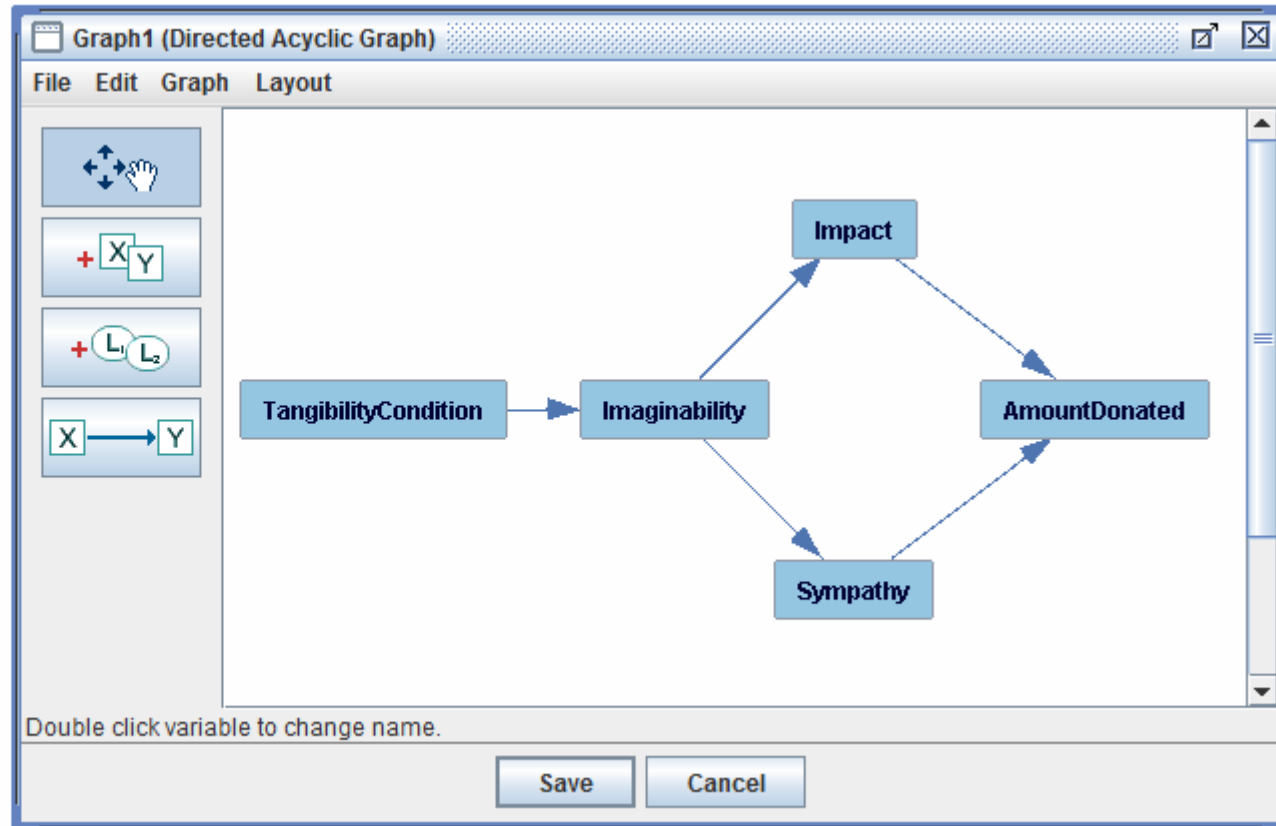
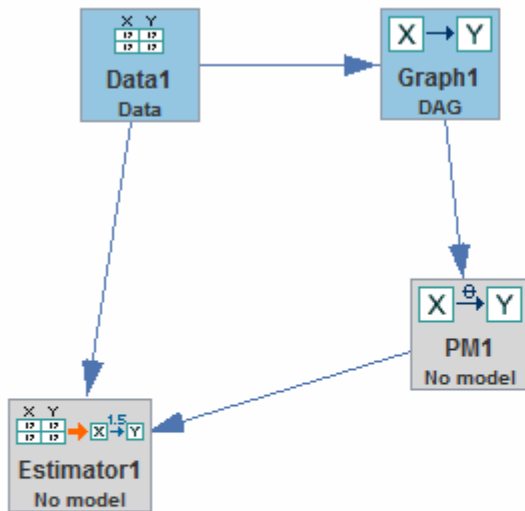
*What influences giving? Sympathy? Impact?*

*"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

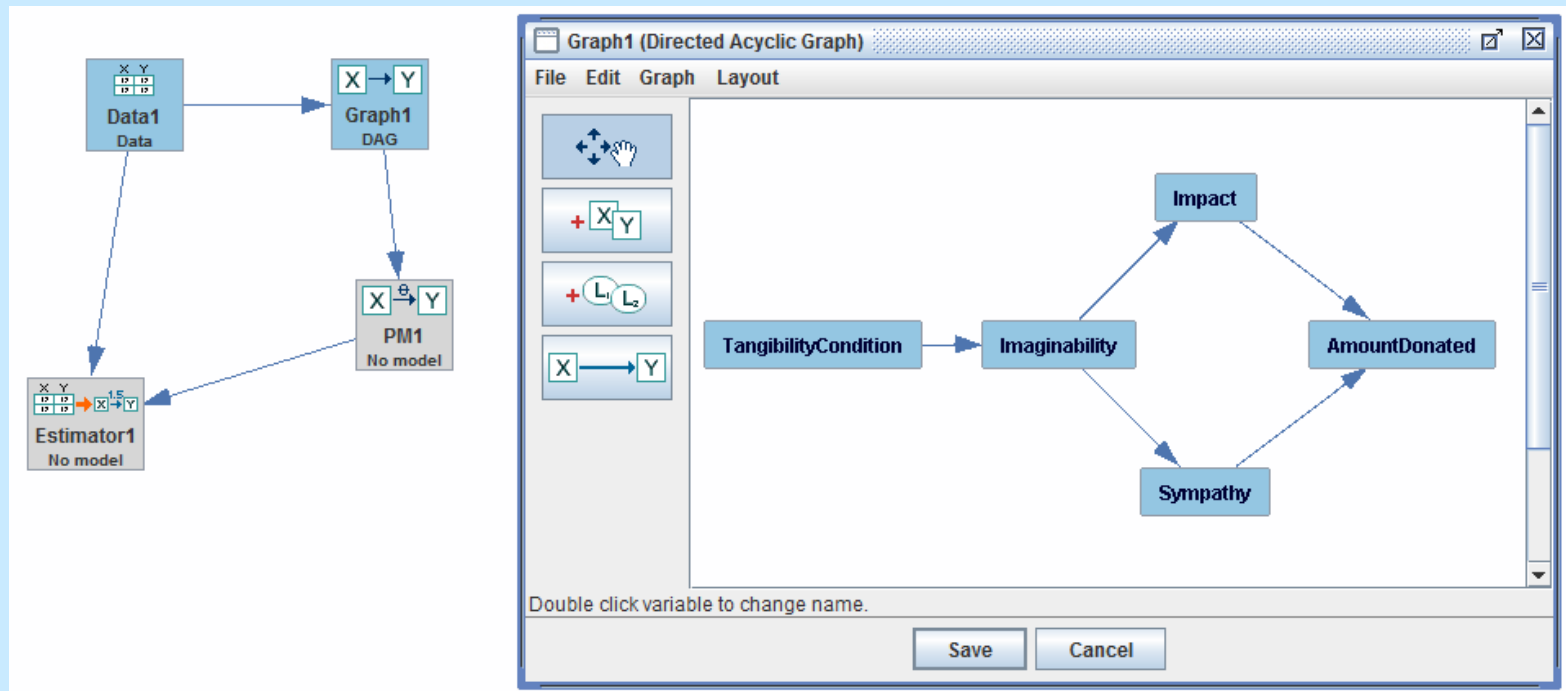
TangibilityCondition	[1,0]	Randomly assigned experimental condition
Imaginability	[1..7]	How concrete scenario I
Sympathy	[1..7]	How much sympathy for target
Impact	[1..7]	How much impact will my donation have
AmountDonated	[0..5]	How much actually donated

# 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

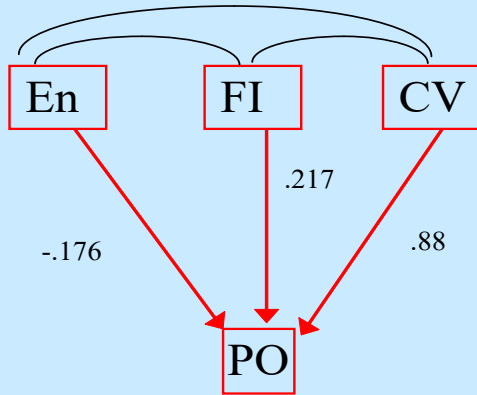
## *Does Foreign Investment in 3<sup>rd</sup> World Countries inhibit Democracy?*

Timberlake, M. and Williams, K. (1984). Dependence, political exclusion, and government repression: Some cross-national evidence. *American Sociological Review* 49, 141-146.

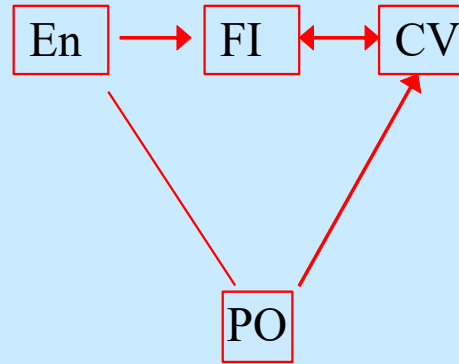
N = 72

- PO degree of political exclusivity
- CV lack of civil liberties
- EN energy consumption per capita (economic development)
- FI level of foreign investment

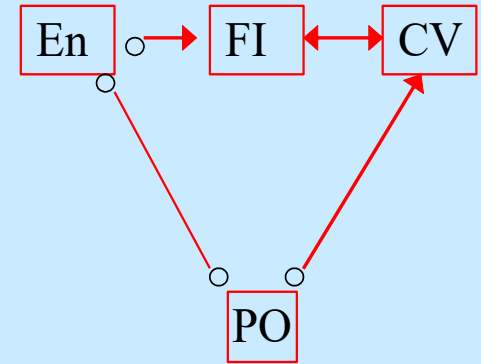
# Case Study: Foreign Investment *Alternative Models*



Regression

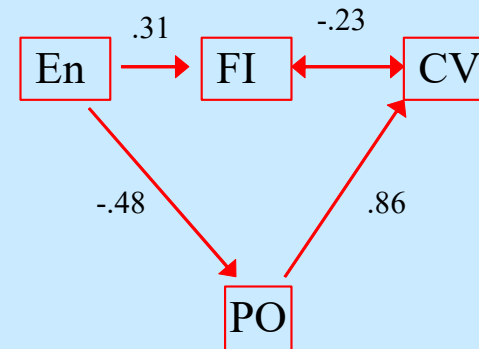


Tetrad - PC



Tetrad - FCI

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.



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