Intro to Mediation Analysis with Structural Equation Modeling
MENAS Workshop Series

Alberto Stefanelli

2022-03-18

Alberto

- PhD Candidate at KU Leuven, Belgium
- Previously...
  - MA at Central European University
- My research:
  - Radial Beliefs
  - Polarization
  - Liberal Values
  - Methods: Causality, experimental and semi-experimental design, SEM etc.
- Contact: alberto.stefanelli@kuleuven.be
- Website: www.albertostefanelli.com
- Twitter: `@sergsagara`

Your turn

- Name?
- Research interests?
- Previous experience with SEM?
- Previous experience with R?
- Why are you taking this workshop?

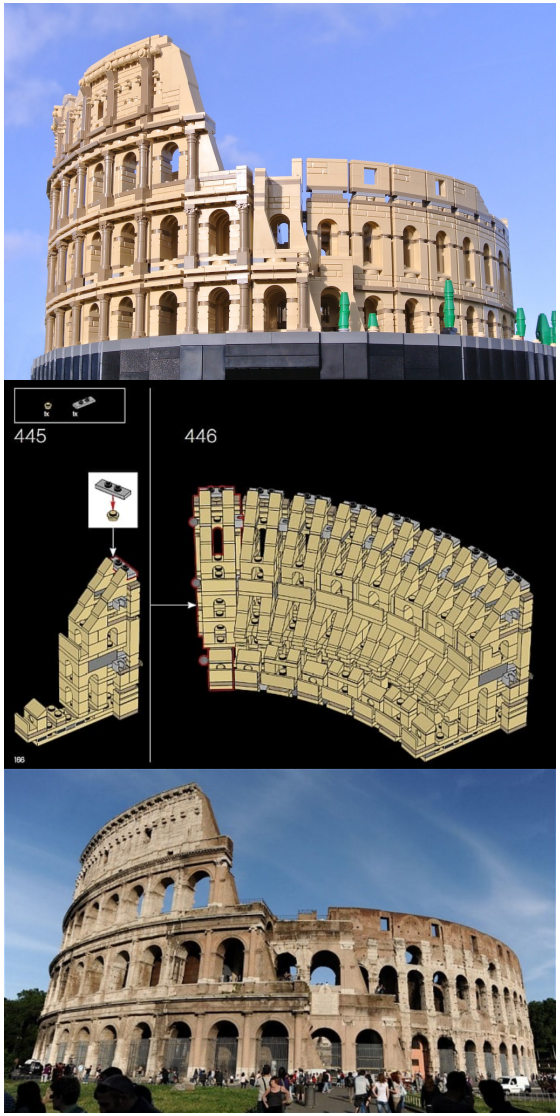
Suggested readings

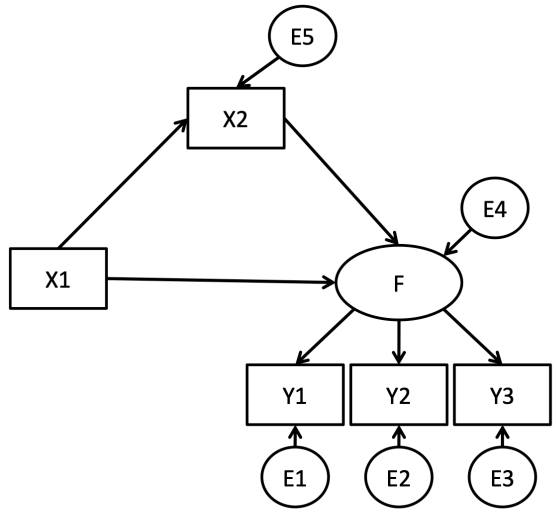
1. Kline, R.B. (2016).Principles and practices of structural equation modeling. Guilford: New York.
2. Zhao, Xinshu et al. (2010). "Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis." Journal of Consumer Research 37 (2): 197–206.
3. Brown, T.A. (2015). Confirmatory Analysis for Applied Research.

What we are going to cover

1. 15.15 - 16.20
   - Into to SEM. What it is ?
   - Estimation basics
   - Path modelling
   - Exercise
   - Equivalent models
2. 16.30 - 17.00
   - Fit MIMIC and mediation modesl in R.
   - Wrap up

What is SEM (1)?





	x1	x2	x3	x4
x1	0.652			
x2	0.470	0.694		
x3	0.384	0.423	0.771	
x4	0.455	0.501	0.410	0.687

What is SEM (2)?

- Multivariate analytical technique: to gain insights in the relations between multiple variables
- Example: how are education, income, and threat related?

What is SEM (3)?

Rather than single equations, *systems of equations* are modelled

- OLS:

$$Y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \epsilon$$

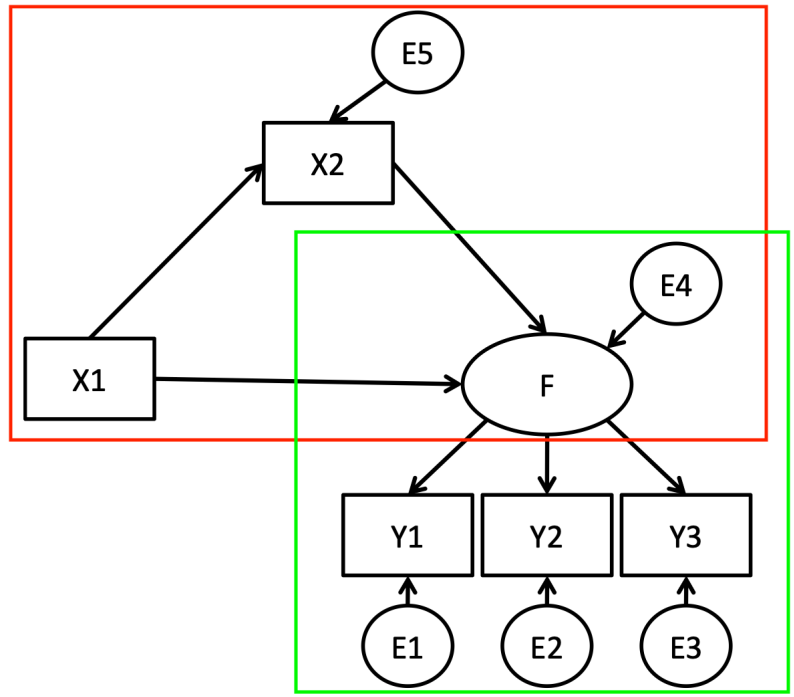
- SEM:

$$\begin{aligned} Y1 &= \lambda 1 F + \epsilon 1 \\ Y2 &= \lambda 2 F + \epsilon 2 \\ Y3 &= \lambda 3 F + \epsilon 3 \\ F &= \beta 1 X1 + \beta 2 X2 + \epsilon 4 \\ X2 &= \beta 3 X1 + + \epsilon 5 \end{aligned}$$

What is SEM (4)?

Typically consists of two components that represent different sets of equations (full SEM):

1. Measurement model: relations between indicators and hypothetically observed constructs, that cannot be measured directly
   - Latent variable estimation: concept measured by multiple indicators that contain measurement error (random & non-random)
2. Structural model: effects between variables.
   - Can be direct, indirect, conditional...



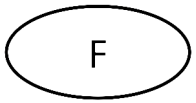
Advantages of SEM?

- Estimate relations between latent variables instead of between unreliable indicators
- Test of complex relationships between variables (e.g., multiple DVs, mediation, moderation, multi-group)
- Handle different types of variables in the same model (i.e., nominal, ordinal, continuous)
- Can handle equality constrains between different parameters of the model
- Extensive set of fit measure to evaluate model performance
- Can be expanded "easily" to more complex data structures (e.g., multilevel modelling, latent growth modelling)

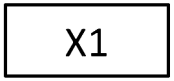
Graphical Notation

Graphically visualize the relationship (paths) between the variables included in the model.

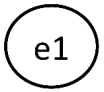
Types of variables



Latent or unobserved variable



Manifest or observed variable

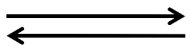


Stochastic / Measurement error

Types of relationships

Uni-directional relation / Effect

Correlation

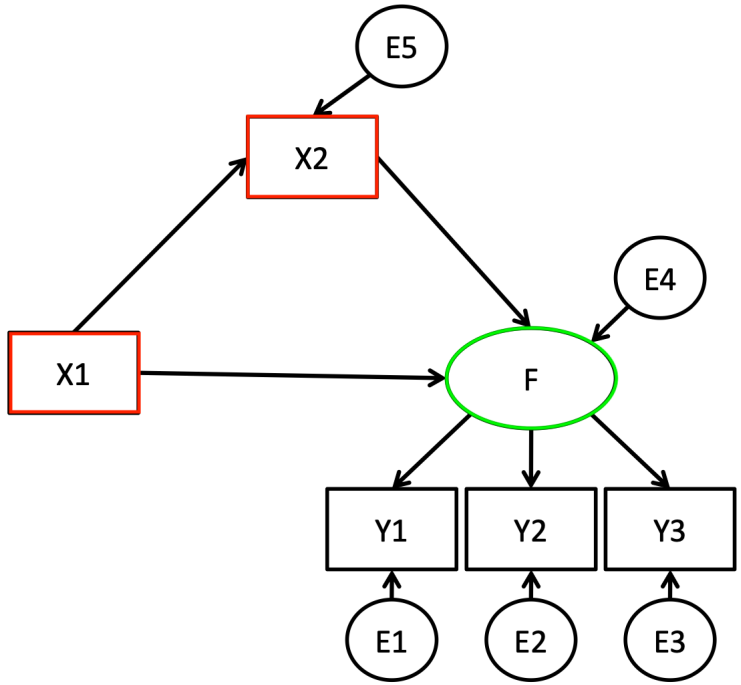


Feedback relation

No relation

SEM Notation

1. Exogenous Variables: means "from without (the outside)". Variable that are not explained by the model for which causes are unknown as far as the model is concerned.
2. Endogenous Variables: means "from within". Variables that are explained by the model and have at least one cause Usually placed in the left side of the diagram.



Model estimation (1)

Instead of raw data (rows=observations), SEM uses the variance covariance matrix to estimate model parameters.

Egalitarianism	Age	Education
2	36	18
2	26	15
3	69	18
2	77	15
4	27	13
3	32	12
3	19	13
2	28	17
2	49	16
2	57	16

	Egalitarianism	Age	Education
Egalitarianism	0.50	-4.00	-0.83
Age	-4.00	396.67	18.11
Education	-0.83	18.11	4.46

Model estimation (2)

Parameter estimation:

- SEM comes down to estimate the parameters so that they approach the sample observed variance-covariance matrix S as good as possible
- This means that every set of parameters implies a certain variance-covariance (and mean structure) matrix Σ that is estimated by the model.
- Distance between the two matrices is evaluated based on a fit function

Most frequently used is Maximum Likelihood Estimation:

- $F_{ML} = ln|S| - ln|\Sigma| + tr[(S)(\Sigma^{-}1)] - p$ with p equal to the order of the input matrix
- If $S = \Sigma$
  - the logs of the determinants are equal
  - and $|S|X|\Sigma|$ is the identity matrix with trace p
  - the fit function is equal to 0

Model estimation (3)

The fit function is minimized by means of an iterative procedure

- Point of departure: set of starting values for the parameters
- Step 1: Evaluate the fit function
- Step 2: Modify parameters in the direction of a better fit (till convergence)
- Step 3: Evaluate the improvement in the fit function (till convergence)

Model fit evaluation

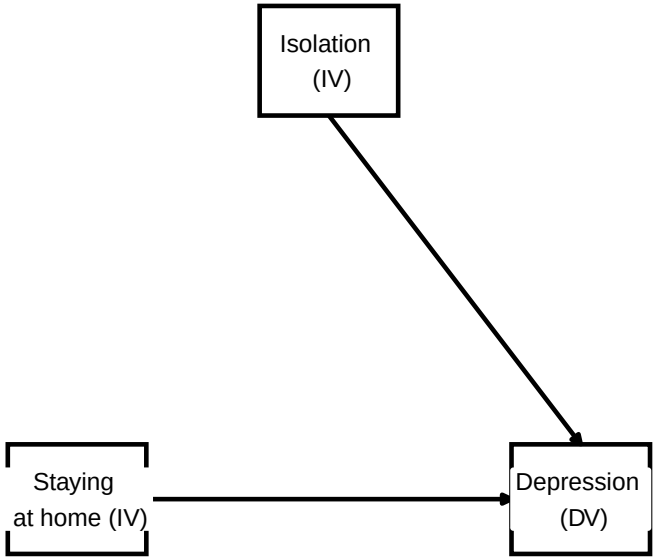
- Chi-square 'goodness of fit' test
  - Tests whether the assumed linear structure holds in the population
  - If chi-square value is statistically significant, reject the model
  - Degrees of freedom: $p(p+1)/2$
  - sensitive for large sample sizes and deviations from multivariate normality
- Standardized Mean Square Residual (SRMR): square root of the average discrepancy between implied and observed covariance
  - Between 0 and 1, smaller values indicating a better fit
  - Rule of thumb: $<.08$ indicates acceptable fit
- Comparative Fit Index (CFI)
  - $CFI = \frac{(\chi^2 - df)_{NullModel} - (\chi^2 - df)_{EstimatedModel}}{(\chi^2 - df)_{NullModel}}$
  - Evaluates the fit of your model compared to the independence model
  - Range 0-1; higher values indicating a better fit
  - Rule of thumb: $>.9$ indicates acceptable fit
- There are others (e.g. RMSEA, TLI) but they are similar to the one presented

Path modelling

Path analysis is a form of multiple regression analysis that is used to examine the relationships between variables. It is useful to decompose the effect of a variable into different 'path' or 'components' with the goal of understanding how different variables influence our outcome(s) of interest.

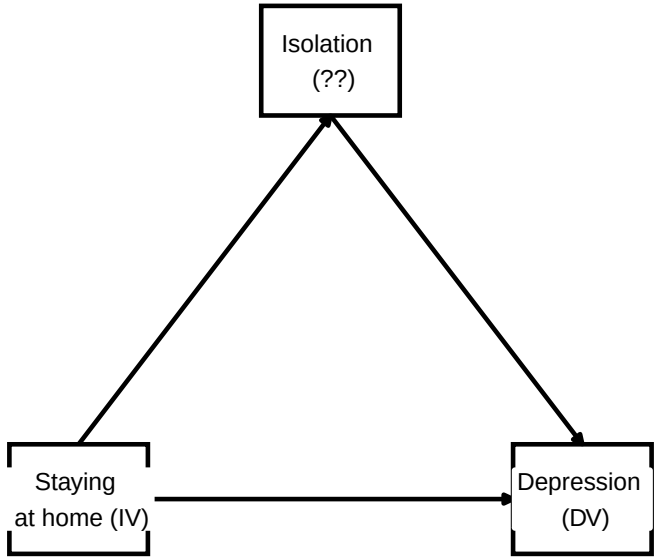
Traditional regression analysis

- 1 DV
- 1 or more IVs



Path modelling

- Multiple DVs
- 1 or more IVs
- 1 or more mediators

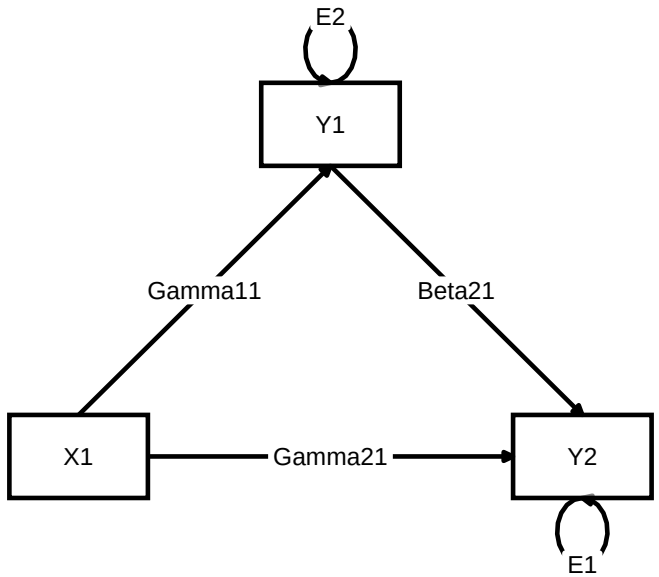


Correspondence between graphical and mathematical notation

$$Y1 = \gamma_{11} X1 + \epsilon 1$$
$$Y2 = \gamma_{21} X1 + \beta_{21} Y1 + \epsilon 2$$

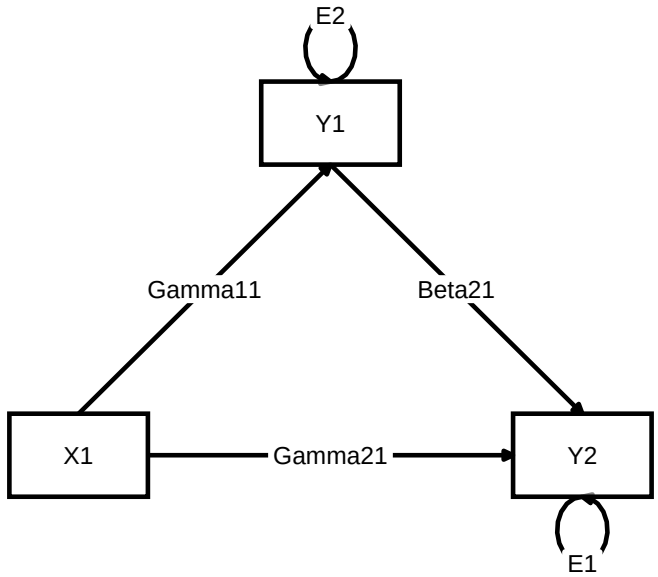
$$\begin{bmatrix} Y1 \\ Y2 \end{bmatrix} = \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix} [X1] + \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} Y1 \\ Y2 \end{bmatrix} + \begin{bmatrix} \epsilon 1 \\ \epsilon 2 \end{bmatrix}$$

$$Y = \Gamma X + \beta Y + E$$



Effect decomposition

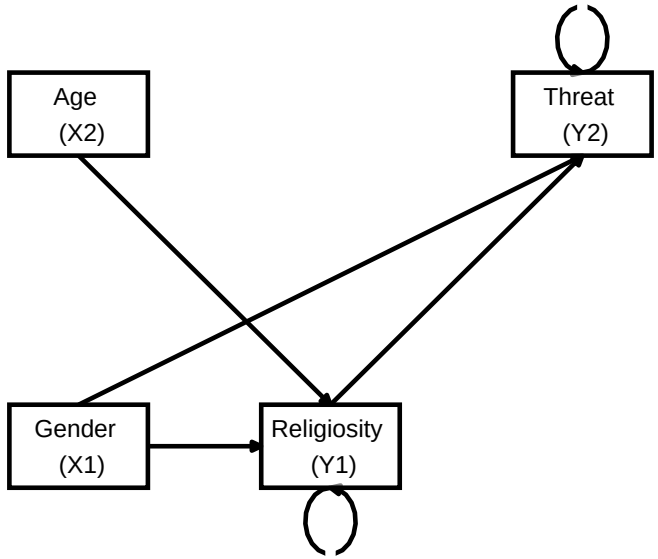
1. Direct effect γ_{21}
2. Indirect effect $\gamma_{11} \times \beta_{21}$
3. Total effect: Direct effect γ_{21} + Indirect effect $\gamma_{11} \times \beta_{21}$



Exercise (1):

Tasks:

1. Exogenous variables?
2. Endogenous variables?
3. Write out the set of equations [skip the matrix notation]
4. Count the number of information present in the sample-implied var-cov matrix
5. Count the number of parameters to estimate
6. Which paths are not estimated ?
7. Write out the indirect effects



Solutions (1)

1. Exogenous variables: Gender, Age
2. Endogenous variables: Threat, Religiosity
3. Pieces of info: 10

$$\sigma(X_{1...4})$$
$$cov(X1, X2)$$
$$cov(X1, X3)$$
$$cov(X1, X4)$$
$$cov(X2, X3)$$
$$cov(X2, X4)$$
$$cov(X3, X4)$$

Solutions (2)

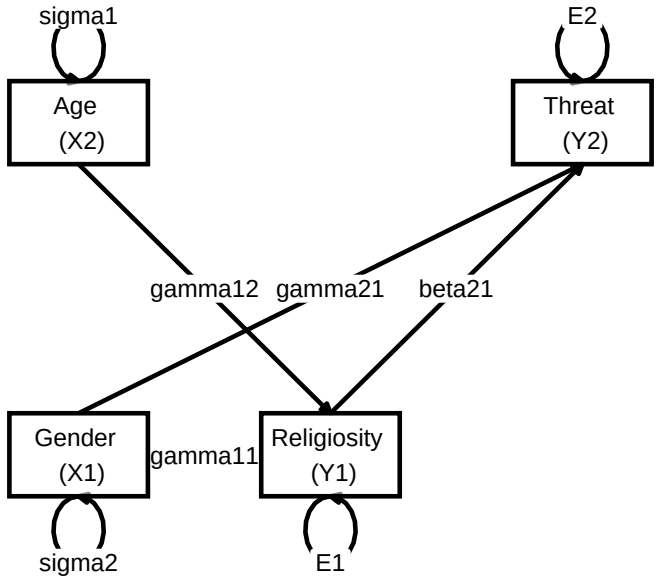
4. Parameters to estimate: 8

$$\sigma_{age}, \sigma_{gender}$$

$$\epsilon_1, \epsilon_2$$

$$\beta_{21}$$

$$\gamma_{11}\gamma_{21}\gamma_{12}$$



5. Equations:

$$Religiosity = \gamma_{11} Gender + \gamma_{12} Age + \epsilon_1$$

$$Threat = \gamma_{21} Gender + \beta_{21} Religiosity + \epsilon_2$$

6. Paths not estimated:
   $\beta_{32} Age, \gamma_{22} Age$
7. Indirect effects:

$$\gamma_{12} Age \times \beta_{21} Religiosity, \gamma_{11} Gender \times \beta_{21} Religiosity$$

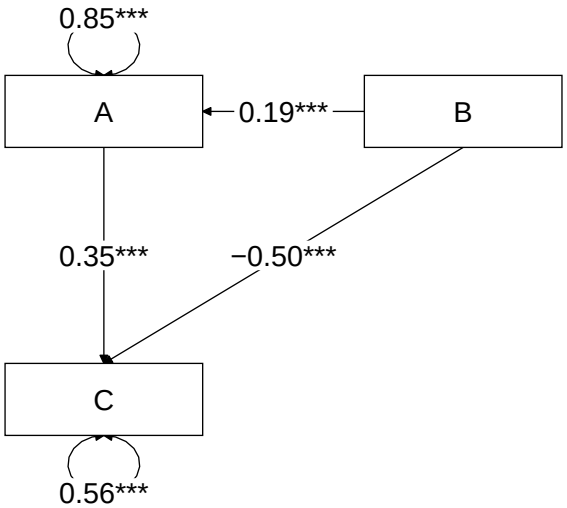
Assumptions

1. Relations among the variables must be linear
2. Errors
   - Constant variance (Homoscedasticity)
   - Covariances among the disturbance terms are zero (equivalent to the assumption of uncorrelated errors among the predictor variables in regression)
   - Normality of residuals
3. Multivariate normality of the included variables
4. No specification error (aka, no omitted variables)

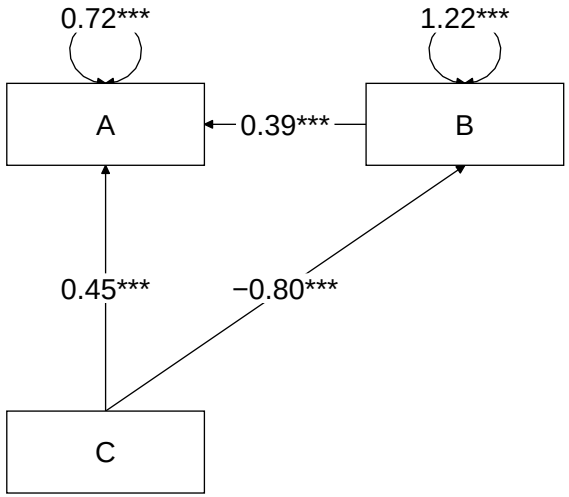
Equivalent models (1)

Sometimes path analysis is called "causal modelling." However, the word "causal" refers to an assumption of the model rather than an intrinsic property of path analysis. To make a causal claim, the model needs to be causally identified.

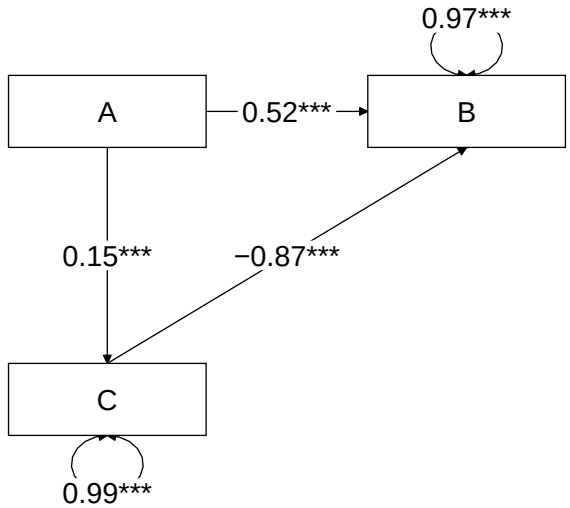
Model: Mediator A



Model: Mediator B



Model: Mediator C



Equivalent models (2)

Model implied variance-covariance for Model: Mediator A

	A	B	C
A	0.92	0.36	0.14
B	0.36	1.86	-0.81
C	0.14	-0.81	1.01

Model implied variance-covariance for Model: Mediator B

	A	B	C
A	0.92	0.36	0.14
B	0.36	1.86	-0.81
C	0.14	-0.81	1.01

Model implied variance-covariance for Model: Mediator C

	A	B	C
A	0.92	0.36	0.14
B	0.36	1.86	-0.81
C	0.14	-0.81	1.01

Equivalent models (3)

Two or more models, with the same variables are equivalent if:

1. Same conditional independence relationships
2. Have the same fit
3. Have the same number of degrees of freedom
4. Are saturated

Equivalent models can not be distinguished in statistical ways. Theory is your best friend !!