Introduction to Structural Equation Modelling (using the 'lavaan' package in R)

Todd K. Hartman
Lecturer in Quantitative Methods
Sheffield Methods Institute

14 February 2017

What is Structural Equation Modelling?

1. Structural model

What is Structural Equation Modelling?

1. Structural model

   Standard model: X $\rightarrow$ Y

What is Structural Equation Modelling?

1. Structural model

    Standard model: X $\rightarrow$ Y

    Path analysis: X $\rightarrow$ Y $\rightarrow$ Z

What is Structural Equation Modelling?

1. Structural model

   Standard model: X $\rightarrow$ Y
   Path analysis: X $\rightarrow$ Y $\rightarrow$ Z

2. Measurement model of latent (unobserved) constructs

   Confirmatory factor analysis (CFA)

What is Structural Equation Modelling?

1. Structural model

   Standard model: X $\rightarrow$ Y
   Path analysis: X $\rightarrow$ Y $\rightarrow$ Z

2. Measurement model of latent (unobserved) constructs

   Confirmatory factor analysis (CFA)

3. Models with structural and measurement components

   Uses CFA to account for measurement error
   Yet, models directional ('causal') relationships

Common SEM Notation

□ observed variable (a.k.a., exogenous variable)

Common SEM Notation

□ observed variable (a.k.a., exogenous variable)

◯ latent (unobserved) variable (a.k.a., endogenous variable)

Common SEM Notation

□ observed variable (a.k.a., exogenous variable)

◯ latent (unobserved) variable (a.k.a., endogenous variable)

△ constant (1)

Common SEM Notation

$\square$ observed variable (a.k.a., exogenous variable)

$\bigcirc$ latent (unobserved) variable (a.k.a., endogenous variable)

$\triangle$ constant (1)

$\longrightarrow$ directional ("*causal*") relationship

Common SEM Notation

$\Box$ observed variable (a.k.a., exogenous variable)

$\bigcirc$ latent (unobserved) variable (a.k.a., endogenous variable)

$\triangle$ constant (1)

$\longrightarrow$ directional ("*causal*") relationship

$\leftrightarrow$ non-directional relationship (covariances for unstandardized solutions or correlations for standardized ones)

Benefits and Limitations of SEM

- Benefits

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships
  - Multiple outcome variables

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships
  - Multiple outcome variables
  - Account for measurement error

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships
  - Multiple outcome variables
  - Account for measurement error
- Limitations

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships
  - Multiple outcome variables
  - Account for measurement error
- Limitations
  - Requires *a priori* specification (i.e., theory)

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships
  - Multiple outcome variables
  - Account for measurement error

- Limitations
  - Requires *a priori* specification (i.e., theory)
  - ‘Large’ sample technique (maximum likelihood)

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships
  - Multiple outcome variables
  - Account for measurement error
- Limitations
  - Requires *a priori* specification (i.e., theory)
  - 'Large' sample technique (maximum likelihood)

    'Small' is $N < 100$
    'Medium' is $100 \leq N \leq 200$
    'Large' is $N > 200$
    (Depends on the complexity of model and estimator used)

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships
  - Multiple outcome variables
  - Account for measurement error
- Limitations
  - Requires *a priori* specification (i.e., theory)
  - 'Large' sample technique (maximum likelihood)

    'Small' is $N < 100$
    'Medium' is $100 \leq N \leq 200$
    'Large' is $N > 200$
    (Depends on the complexity of model and estimator used)
  - Infinite number of possible models

Benefits and Limitations of SEM

- Benefits
  - Simultaneously models a system of relationships
  - Multiple outcome variables
  - Account for measurement error
- Limitations
  - Requires *a priori* specification (i.e., theory)
  - 'Large' sample technique (maximum likelihood)

    'Small' is $N < 100$
    'Medium' is $100 \leq N \leq 200$
    'Large' is $N > 200$
    (Depends on the complexity of model and estimator used)
  - Infinite number of possible models
  - Correlation $\neq$ Causation

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)
    - Can be nominal, ordinal, or continuous

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)
   - Can be nominal, ordinal, or continuous
2. Latent variables

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)
   - Can be nominal, ordinal, or continuous
2. Latent variables
   - Correspond to hypothetical constructs or factors

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)
    - Can be nominal, ordinal, or continuous
2. Latent variables
    - Correspond to hypothetical constructs or factors
    - Observed variables used as ‘indicators’

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)
   - Can be nominal, ordinal, or continuous
2. Latent variables
   - Correspond to hypothetical constructs or factors
   - Observed variables used as ‘indicators’
   - Must be continuous (e.g., intelligence)

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)
   - Can be nominal, ordinal, or continuous
2. Latent variables
   - Correspond to hypothetical constructs or factors
   - Observed variables used as 'indicators'
   - Must be continuous (e.g., intelligence)

- Residual (error) terms for observed variables or factors as outcome variables

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)
    - Can be nominal, ordinal, or continuous
2. Latent variables
    - Correspond to hypothetical constructs or factors
    - Observed variables used as 'indicators'
    - Must be continuous (e.g., intelligence)

- Residual (error) terms for observed variables or factors as outcome variables
    - For indicators: residual is variance unexplained by hypothesized factor (e.g., random measurement error)

Two Types of Variables in SEM

1. Observed variables (a.k.a., manifest variables)
   - Can be nominal, ordinal, or continuous
2. Latent variables
   - Correspond to hypothetical constructs or factors
   - Observed variables used as 'indicators'
   - Must be continuous (e.g., intelligence)

- Residual (error) terms for observed variables or factors as outcome variables
  - For indicators: residual is variance unexplained by hypothesized factor (e.g., random measurement error)
  - For outcomes (observed or latent factors): residual is variance unexplained by their predictors

Basic Statistic is Covariance

- *Covariance* is strength of association between X and Y and their variabilities
  (unlike correlation, covariance has no upper or lower bounds)

Basic Statistic is Covariance

- *Covariance* is strength of association between X and Y and their variabilities
  (unlike correlation, covariance has no upper or lower bounds)
- Means are not analyzed in most SEMs (although it can be done)

Basic Statistic is Covariance

- *Covariance* is strength of association between X and Y and their variabilities
  (unlike correlation, covariance has no upper or lower bounds)
- Means are not analyzed in most SEMs (although it can be done)

$$cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

$$cov_{XY} = rr_{XY} SD_X SD_Y$$

Basic Statistic is Covariance

- *Covariance* is strength of association between X and Y and their variabilities
  (unlike correlation, covariance has no upper or lower bounds)
- Means are not analyzed in most SEMs (although it can be done)

$$cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

$$cov_{XY} = rr_{XY} SD_X SD_Y$$

1. Goal 1: to understand patterns of covariances among observed variables

Basic Statistic is Covariance

- *Covariance* is strength of association between X and Y and their variabilities
  (unlike correlation, covariance has no upper or lower bounds)
- Means are not analyzed in most SEMs (although it can be done)

$$cov(X, Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

$$cov_{XY} = rr_{XY} SD_X SD_Y$$

1. Goal 1: to understand patterns of covariances among observed variables
2. Goal 2: to explain as much variation as possible with model

How Does SEM Differ from Regression?

Estimated Covariance

$$cov_{XY} = rr_{XY} SD_X SD_Y$$

How Does SEM Differ from Regression?

Estimated Covariance

$$cov_{XY} = rr_{XY} SD_X SD_Y$$

Regression Estimate

$$\beta = rr_{XY}(SD_Y / SD_X)$$

SEM Equation

Structural form

$$y = \alpha + By + \Gamma x + \zeta$$

SEM Equation

Structural form

$$y = \alpha + By + \Gamma x + \zeta$$

- where y is a vector of observed **en**dogenous variables

SEM Equation

Structural form

$$y = \alpha + By + \Gamma x + \zeta$$

- where y is a vector of observed **en**dogenous variables
- x is a vector of observed **ex**ogenous variables; $cov(x) = \Phi$ is their covariance matrix

SEM Equation

Structural form

$$y = \alpha + By + \Gamma x + \zeta$$

- where y is a vector of observed **en**dogenous variables
- x is a vector of observed **ex**ogenous variables; $cov(x) = \Phi$ is their covariance matrix
- α is a vector of structural intercepts

SEM Equation

Structural form

$$y = \alpha + By + \Gamma x + \zeta$$

- where y is a vector of observed **en**dogenous variables
- x is a vector of observed **ex**ogenous variables; $cov(x) = \Phi$ is their covariance matrix
- α is a vector of structural intercepts
- B is a coefficient matrix that relates endogenous variables to each other

SEM Equation

Structural form

$$y = \alpha + By + \Gamma x + \zeta$$

- where y is a vector of observed **en**dogenous variables
- x is a vector of observed **ex**ogenous variables; $cov(x) = \Phi$ is their covariance matrix
- α is a vector of structural intercepts
- B is a coefficient matrix that relates endogenous variables to each other
- Γ is a coefficient matrix that relates endogenous variables to exogenous variables

SEM Equation

Structural form

$$y = \alpha + By + \Gamma x + \zeta$$

- where y is a vector of observed **en**dogenous variables
- x is a vector of observed **ex**ogenous variables; $cov(x) = \Phi$ is their covariance matrix
- α is a vector of structural intercepts
- B is a coefficient matrix that relates endogenous variables to each other
- Γ is a coefficient matrix that relates endogenous variables to exogenous variables
- ζ is a vector of disturbance terms; $cov(\zeta) = \Psi$ is their covariance matrix

What Sample Size Is Enough?

- *N: q rule of thumb* (maximum likelihood estimation)

What Sample Size Is Enough?

- *N: q rule of thumb* (maximum likelihood estimation)

  N: observations from the dataset

  q: model parameters

What Sample Size Is Enough?

- *N: q rule of thumb* (maximum likelihood estimation)

    N: observations from the dataset

    q: model parameters

- Ideal 20:1 (e.g., if $q = 10$, then need sample size of at least 200

What Sample Size Is Enough?

- *N: q rule of thumb* (maximum likelihood estimation)

  N: observations from the dataset

  q: model parameters

- Ideal 20:1 (e.g., if $q = 10$, then need sample size of at least 200
- Minimal 10:1 ratio (just like regression)

Uncertain Causality?

- SEM is a correlational approach; causation is driven by theory

Uncertain Causality?

- SEM is a correlational approach; causation is driven by theory
- When the direction of causality is uncertain. . .

Uncertain Causality?

- SEM is a correlational approach; causation is driven by theory
- When the direction of causality is uncertain. . .
    1. Specify model but without directionality between key variables (i.e., no causal paths)

Uncertain Causality?

- SEM is a correlational approach; causation is driven by theory
- When the direction of causality is uncertain...
  1. Specify model but without directionality between key variables (i.e., no causal paths)
  2. Specify and test alternative models with different causal directionalities (with similar results, no statistical method can identify which is correct)

Uncertain Causality?

- SEM is a correlational approach; causation is driven by theory
- When the direction of causality is uncertain...
  1. Specify model but without directionality between key variables (i.e., no causal paths)
  2. Specify and test alternative models with different causal directionalities (with similar results, no statistical method can identify which is correct)
  3. And/or include reciprocal effects to cover both possibilities (but can create problems of identification)

Identification and Model Complexity

- Concerns total number of parameters to be estimated (regardless of sample size)

Identification and Model Complexity

- Concerns total number of parameters to be estimated (regardless of sample size)
- If v is the number of observed variables, then we can estimate $v(V+1)/2$ parameters

Identification and Model Complexity

- Concerns total number of parameters to be estimated (regardless of sample size)
- If v is the number of observed variables, then we can estimate $v(V+1)/2$ parameters
- Example: suppose $v = 4$ observed variables

Identification and Model Complexity

- Concerns total number of parameters to be estimated (regardless of sample size)
- If v is the number of observed variables, then we can estimate $v(V+1)/2$ parameters
- Example: suppose $v = 4$ observed variables

  Then, the max number of parameters estimated is 10
  $4(4+1)/2 = 10$

Identification and Model Complexity

- Concerns total number of parameters to be estimated (regardless of sample size)
- If v is the number of observed variables, then we can estimate $v(V+1)/2$ parameters
- Example: suppose $v = 4$ observed variables

  Then, the max number of parameters estimated is 10
  $4(4+1)/2 = 10$

  Total number of variances (4) and covariances (6) in the data matrix (fewer can be estimated, but the max is 10)

Parameter Status

- Each model parameter can be free, fixed, or constrained

Parameter Status

- Each model parameter can be free, fixed, or constrained
  - 'Free' is estimated by the software from the data

Parameter Status

- Each model parameter can be free, fixed, or constrained
  - ‘Free’ is estimated by the software from the data
  - Fixed’ is set to a constant by the researcher (and accepted by the software regardless of the data)

Parameter Status

- Each model parameter can be free, fixed, or constrained
  - ‘Free’ is estimated by the software from the data
  - Fixed’ is set to a constant by the researcher (and accepted by the software regardless of the data)
  - ‘Constrained’ is estimated by the software with some restrictions (e.g., constrained to be equal to another parameter)

Steps for Conducting SEM

1. Specify the model (draw out hypotheses)

Steps for Conducting SEM

1. Specify the model (draw out hypotheses)
2. Estimate the model using software

Steps for Conducting SEM

1. Specify the model (draw out hypotheses)
2. Estimate the model using software
3. Evaluate the model fit (Chi-square 'badness of fit' ($>.05$), RMSEA ($<.05$), CFI ($>.95$), TLI ($>.95$))

Steps for Conducting SEM

1. Specify the model (draw out hypotheses)
2. Estimate the model using software
3. Evaluate the model fit (Chi-square 'badness of fit' ($>$.05), RMSEA ($<$.05), CFI ($>$.95), TLI ($>$.95))
4. Respecify the model (if needed using using theory)

Steps for Conducting SEM

1. Specify the model (draw out hypotheses)
2. Estimate the model using software
3. Evaluate the model fit (Chi-square 'badness of fit' ($>$.05), RMSEA ($<$.05), CFI ($>$.95), TLI ($>$.95))
4. Respecify the model (if needed using using theory)
5. Reevaluate the model fit

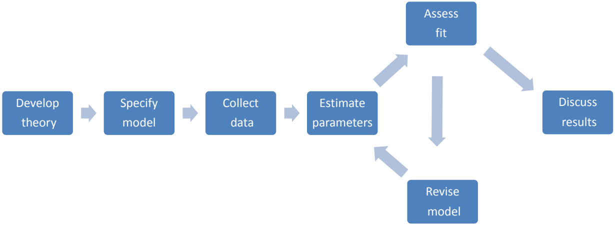
Steps for Conducting SEM

1. Specify the model (draw out hypotheses)
2. Estimate the model using software
3. Evaluate the model fit (Chi-square 'badness of fit' ($>$.05), RMSEA ($<$.05), CFI ($>$.95), TLI ($>$.95))
4. Respecify the model (if needed using using theory)
5. Reevaluate the model fit
6. Rinse and repeat

Steps for Conducting SEM

1. Specify the model (draw out hypotheses)
2. Estimate the model using software
3. Evaluate the model fit (Chi-square 'badness of fit' ($>$.05), RMSEA ($<$.05), CFI ($>$.95), TLI ($>$.95))
4. Respecify the model (if needed using using theory)
5. Reevaluate the model fit
6. Rinse and repeat
7. Interpret the results

SEM Flowchart



Practical Example: Support for Social Welfare Spending

What affects preferences for government spending on social welfare programs?

Practical Example: Support for Social Welfare Spending

What affects preferences for government spending on social welfare programs?

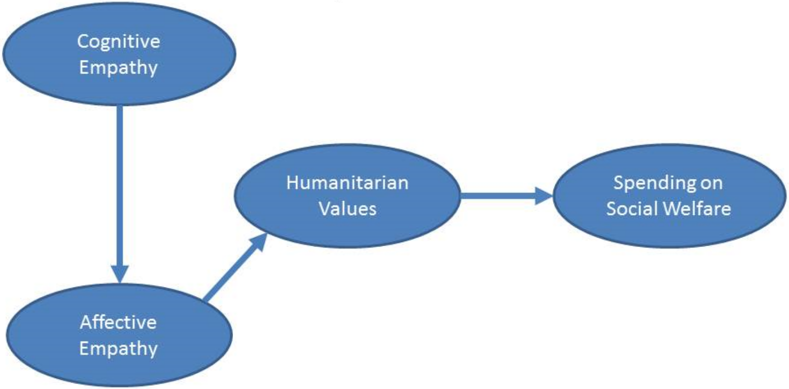
1. Attitudes toward Government
    - Ideology
    - Party Affiliation
2. Personal Experiences
    - Income
    - Generational Cohorts (Age)
3. Attitudes toward Beneficiaries
    - Racial stereotypes

Practical Example: Support for Social Welfare Spending

What affects preferences for government spending on social welfare programs?

1. Attitudes toward Government
   - Ideology
   - Party Affiliation
2. Personal Experiences
   - Income
   - Generational Cohorts (Age)
3. Attitudes toward Beneficiaries
   - Racial stereotypes
4. **Pro-Social Orientations**
   - Humanitarianism (*value* helping those in need)
   - Empathy (*ability* to understand/feel what another being is experiencing)

Practical Example: Support for Social Welfare Spending



Practical Example: Support for Social Welfare Spending

- American National Election Study 2008 - 2009 Panel Study
  - Monthly surveys with representative Internet panel
  - 1,420 to 2,665 completed interviews per wave
  - Social Spending toward Social Security, Aid to the Poor, Job Retraining, and Public Schools
  - 8 Humanitarianism Items
    - 'It is important to help one another so that the community in general is a better place.'
  - 21 Empathy Items
  - Other Demographic Controls

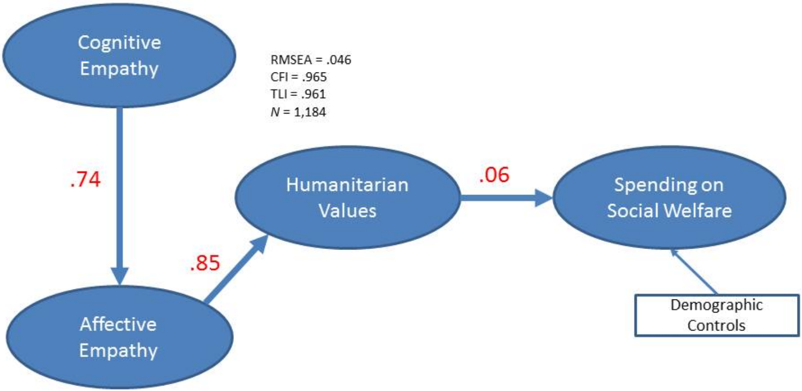
Empathy

- Interpersonal Reactivity Index (Davis (1980, 1983)
  - **Empathic Perspective-Taking**
  - **Empathic Concern**
  - Personal Distress
  - Fantasy

2-Factor CFA Model of Empathy

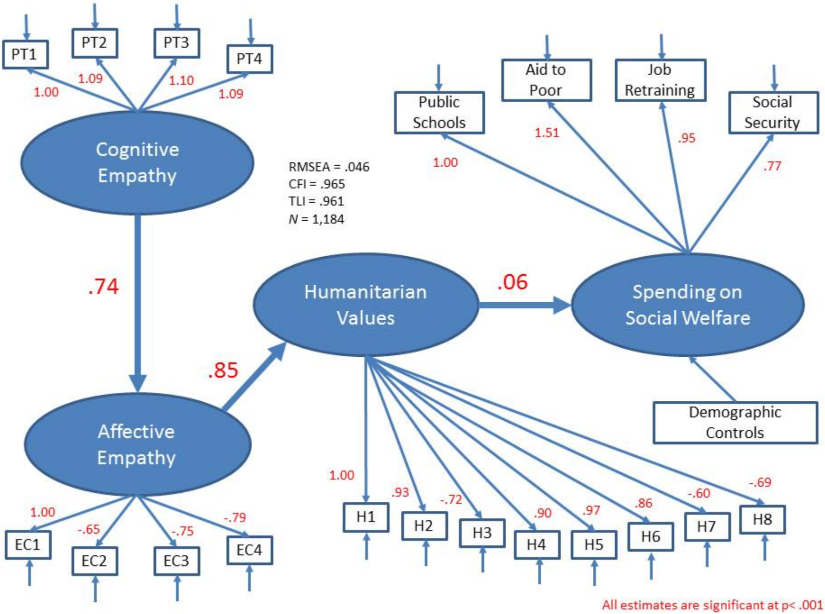
Survey Item	Cognitive Empathy	Affective Empathy
I try to look at everybody's side of a disagreement before I make a decision.	**.64**	.36
I sometimes try to understand my friends better by imagining how things look from their perspective.	**.73**	.41
I believe that there are two sides to every question and try to look at them both.	**.73**	.40
Before criticizing somebody, I try to imagine how I would feel if I were in their place.	**.69**	.38
I often have tender, concerned feelings for people less fortunate than me.	.39	**.70**
Sometimes I don't feel very sorry for other people when they are having problems. (R)	-.33	**-.59**
When I see someone being treated unfairly, I sometimes don't feel very much pity for them. (R)	-.37	**-.67**
Other people's misfortunes do not usually disturb me a great deal. (R)	-.38	**-.69**

Structural Equation Model Results

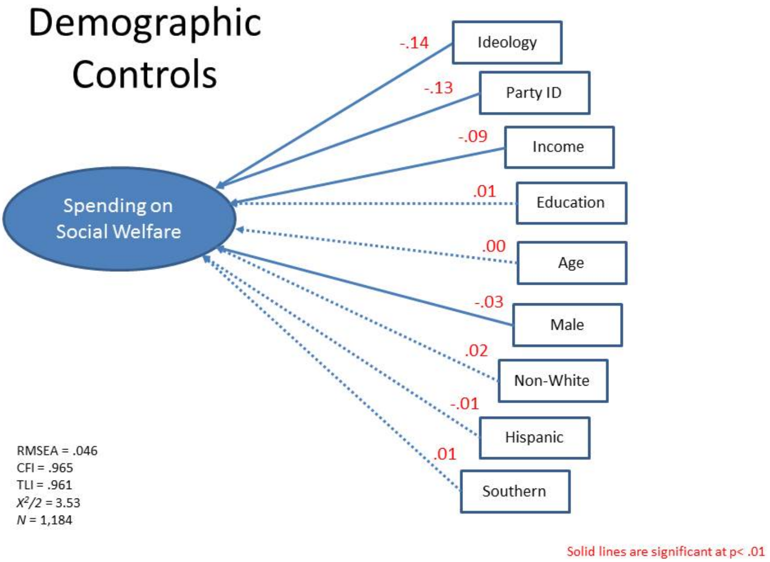


All estimates are significant at p< .001

Structural Equation Model Results



Structural Equation Model Results



Lavaan Code

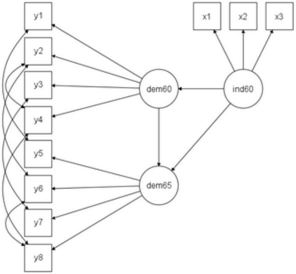
```
## Install Lavaan for SEM
install.packages("lavaan",repos = "http://cran.us.r-project.org/")

## Require Needed Packages
require(lavaan)
require(foreign)


## SEM for Empathy, Humanitarianism, and Social Spending
model <- ' # Latent Variables
          cempathy =~ ept2 + ept3 + ept5 + ept7
          aempathy =~ ec1 + ec2 + ec4 + epd4
          human =~ hu1 + hu2 + hu3 + hu4 + hu5 + hu6 + hu7 + hu8
          social =~ school15 + ss15 + poor15 + job15
          # Regressions
          aempathy ~ a*cempathy
          human ~ b*aempathy
          social ~ c*human
          social ~ ideology + party + male + age + hispanic
                        + nonwhite + education + income + south
          # Indirect Effect (a*b)
              ab := a*b
              bc := b*c
              abc := a*b*c
          # Residual Covariances
          # cempathy ~~ aempathy
        '

fit <- sem(model,
           data=anes,
           ordered=c("ept2","ept3",
                     "ept5","ept7",
                     "ec1", "ec2",
                     "ec4", "epd4",
                     "hu1","hu2",
                     "hu3","hu4",
                     "hu5","hu6",
                     "hu7","hu8"))

summary(fit)
parameterEstimates(fit)
fitMeasures(fit, c("cfi", "rmsea", "tli"))
```

The official reference to the lavaan package is the following paper:

> Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

First impression

To get a first impression of how lavaan works in practice, consider the following example of a SEM model. The figure below contains a graphical representation of the model that we want to fit.




```
model <- '
  # latent variables
    ind60 =~ x1 + x2 + x3
    dem60 =~ y1 + y2 + y3 + y4
    dem65 =~ y5 + y6 + y7 + y8
  # regressions
    dem60 ~ ind60
    dem65 ~ ind60 + dem60
  # residual covariances
    y1 ~~ y5
    y2 ~~ y4 + y6
    y3 ~~ y7
    y4 ~~ y8
    y6 ~~ y8
'
fit <- sem(model,
           data=PoliticalDemocracy)
summary(fit)
```

lavaan in R

lavaan is (relatively) easy and intuitive lavaan is (relatively) easy and intuitive

- lavaan in R is free (as in beer!)

lavaan in R

lavaan is (relatively) easy and intuitive

- lavaan in R is free (as in beer!)
- Strong online support/community

lavaan in R

lavaan is (relatively) easy and intuitive

- lavaan in R is free (as in beer!)
- Strong online support/community
- Compact, readable R commands

lavaan in R

lavaan is (relatively) easy and intuitive

- lavaan in R is free (as in beer!)
- Strong online support/community
- Compact, readable R commands
- Constant development of latest methods

lavaan in R

lavaan is (relatively) easy and intuitive

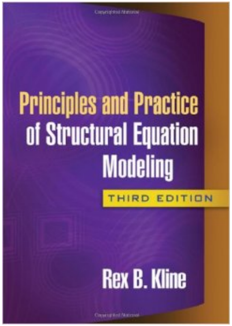
- lavaan in R is free (as in beer!)
- Strong online support/community
- Compact, readable R commands
- Constant development of latest methods
- Full support for categorical data!

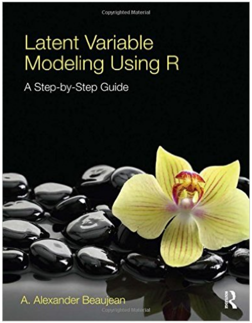
lavaan in R

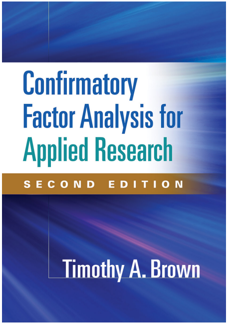
lavaan is (relatively) easy and intuitive

- lavaan in R is free (as in beer!)
- Strong online support/community
- Compact, readable R commands
- Constant development of latest methods
- Full support for categorical data!
    - Binary, Categorical, and Continuous DVs

My Favourite SEM Books







Idre: http://www.ats.ucla.edu/stat/