Structural Equation Modeling in Open-Source Software

Justin Wild
Indiana University

December 1, 2017

About the presenter

- Education Policy Studies & Inquiry Methodology doctoral candidate
  - a researcher
  - focus on methods & methodologies
  - not a programming expert
- Contributors
  - IU
    - Scott Michael
    - Michael Frisby
- Inspirational
  - Yves Rosseel's keynote speech at useR! conference, 2017

Talk Overview

- **What is SEM?**
- SEM software packages
- SEM examples: running the software
- semPlot package
- Multilevel SEM
- Wrapping up
- Slides there is no time for

What is SEM? (Yves Rosseel, 2017)

- SEM = structural equation modeling
  - Multivariate statistical modeling technique
  - Test a hypothesis about the data
  - Test a model of the data
    - We propose a data-generating model
    - The model may or may not fit the data

- What is unique about SEM?
  - May contain latent (unobserved) variables
    - Can be hypothetical 'constructs' (e.g. reading ability, depression, etc.)
    - Can be random effects (e.g. intercepts, slopes)
    - …
  - Allows for indirect effects (mediation), reciprocal effects,…
  - Drawings!!!

SEM: regression example

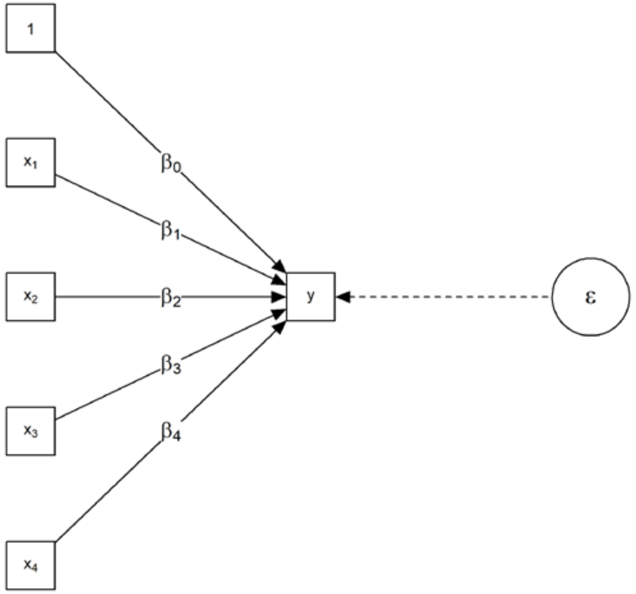
- Univariate linear regression

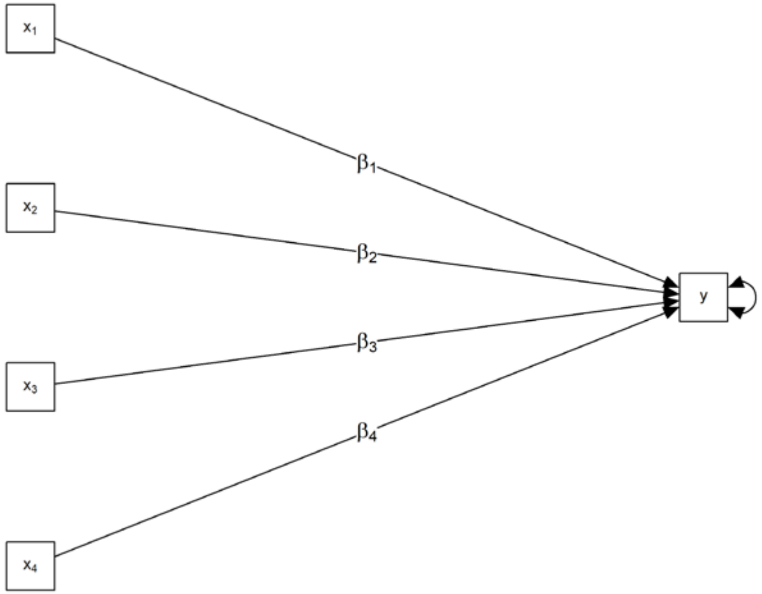
Equation: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$

Matrices: $\underset{n \times 1}{\boldsymbol{Y}} = \underset{n \times p'}{\boldsymbol{X}} \underset{p' \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\boldsymbol{e}}$

Matrix notation: $[y_1] = [1 \quad x_{11} \quad x_{12} \quad x_{13} \quad x_{14}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + [e_1]$

SEM matrix notation: $[y_1] = [\beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + [e_1]$





SEM: regression example

- Multiple (multivariate) regression

Equations:

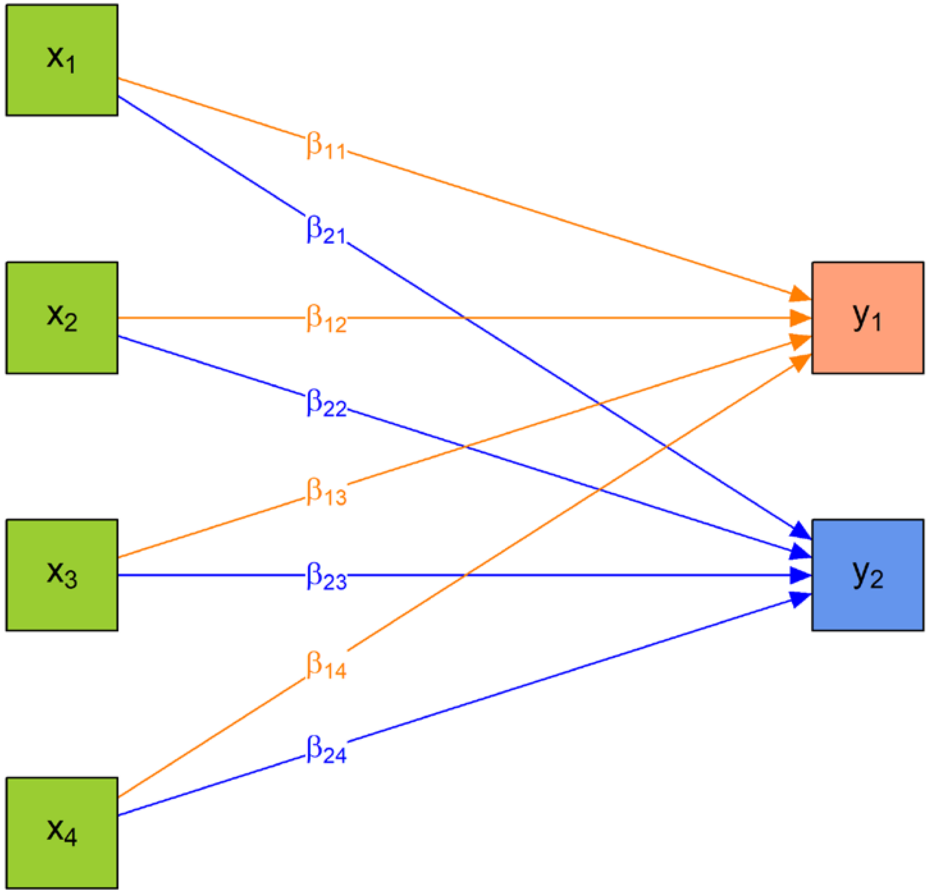
$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \beta_4 x_{14} + \epsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \beta_4 x_{24} + \epsilon_2$$

Matrix notation:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

SEM matrix notation:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

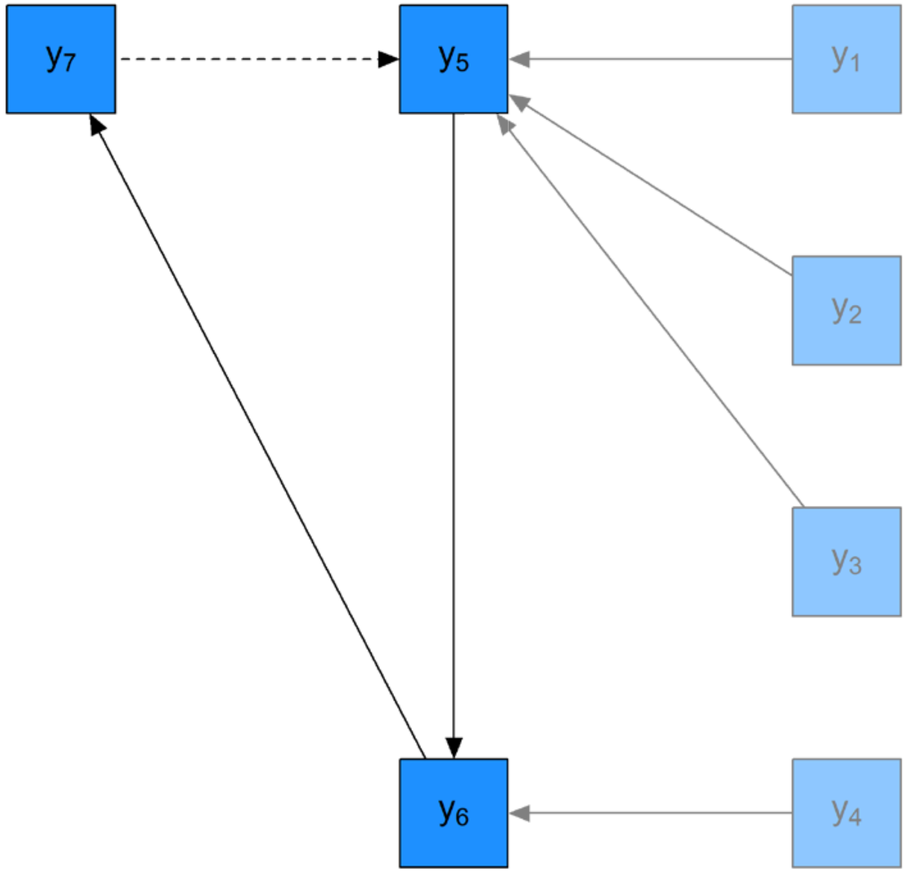


SEM: indirect effects & path analysis

- All variables are observed (manifest)
- Notice indirect effects: y_5 via y_6 on y_7
- Notice cycles: y_7 may influence y_5

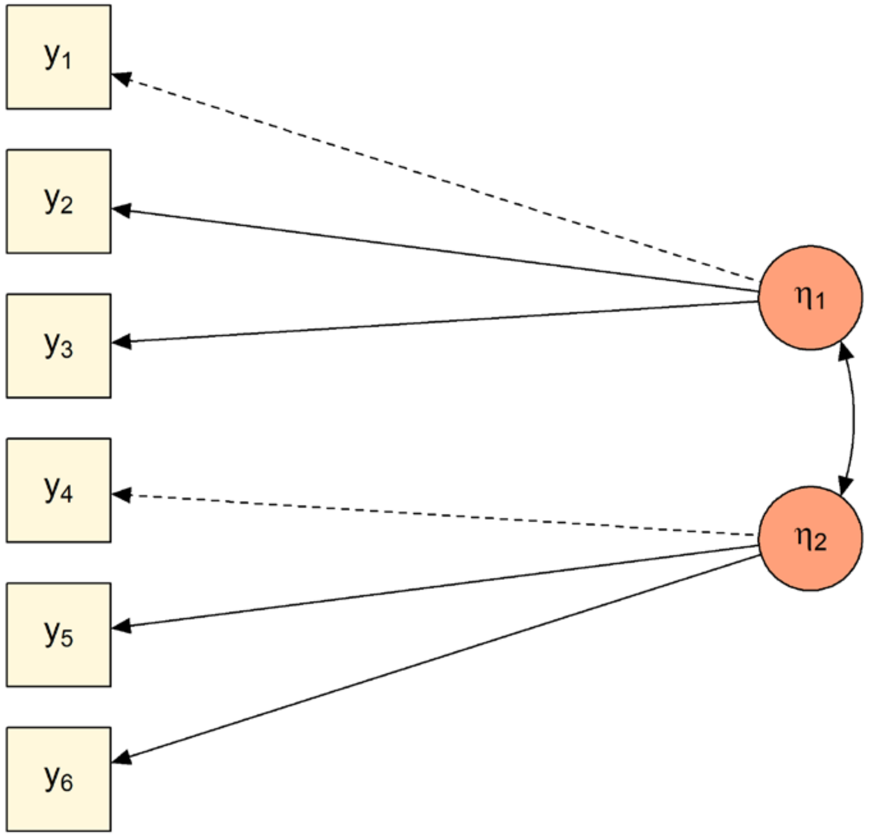
y_1 = read for school
y_2 = read with parent
y_3 = access to other media
y_4 = access to books

y_5 = reading motivation
y_6 = reading frequency
y_7 = reading ability



SEM: latent variables & path analysis

- Specify a measurement model of latent variables (η) of interest and their observed (manifest) indicators (*y*)
- Confirmatory Factor Analysis
  - Fit indices
  - Modification indices



Uses of SEM

- Widely used in social sciences: e.g. education, psychology
- Now being used in:
  - Medical sciences
  - Nueroimaging
  - Biology, ecology (e.g. climate change)
  - Operation research
  - …and growing
- SEM software is used for ‘standard’ analyses (e.g. regression) when there is need for:
  - Robust standard errors
  - Diagnostics
  - (in)equality restraints: needing a parameter to be larger
  - …

Talk Overview

- What is SEM?
- **SEM software packages**
- SEM examples: running the software
- semPlot package
- Multilevel SEM
- Wrapping up
- Slides there is no time for

SEM software packages

- Closed-source: main four
  - LISREL ('70s, Karl Jöreskog)
  - EQS ('80s, Peter Bentler)
  - AMOS ('90s, James Arbuckle)
  - Mplus (Bengt Muthén, 1998-now)
- SAS through proc CALIS, proc TCALIS
- SPSS: Bought AMOS and sells as a separate product (have not changed much since)
- Stata module (free): 'gllamm' (Sophia Rabe-Hesketh, Anders Skrondal, Andrew Pickles, since 2002)
- Statistica (SEPATH), Systat (RAMONA), Mx (Michael Neale, free, closed-source, '90s)

- Open-source
  - Interfaces between R and commercial packages:
    - REQS (Patrick Mair, Eric Wu, since 2008)
    - MplusAutomation (Michael Hallquist, since 2010)
- R packages
  - sem (John Fox, since 2001)
  - OpenMX (Steven Boker, Michael Neale, . . . since 2009)
  - lavaan (Yves Rosseel, since 2010)
  - lava (Klaus Holst, since 2012)
  - strum (Nathan Morris, since 2015; for latent variables of genetic analysis)

- Advantages
  - Single package that can do:
    - SEM
    - Multilevel SEM
    - Bootstrapping
    - Multiple imputations
    - Add weights
    - Latent categorical variables (latent classes)
    - Path diagrams
- Disadvantages
  - Black box: limited to options the developers give you
    - Example: Does not calculate average variance extracted (AVE) and composite reliability (CR), sometimes used as reliability measures
  - Not free
  - No multilevel path diagrams

- Advantages
  - Open-source; many options for the user
  - Users/coders can (re)write software for perceived needs and omissions in current packages offered
  - Not just statistics
    - Survey api to point directly to surveys
    - Security: use in conjunction with stats
    - Display results in html (e.g. Shiny, Markdown)
    - …
- Disadvantages (pertaining to SEM)
  - While R can do SEM, multilevel SEM, bootstrapping, multiple imputations, weights, path diagrams, etc. no single PACKAGE can do all (yet)
  - Packages:
    - R is a base code for which coders write packages that are then uploaded by users who need to use specific features; therefore some comfort with basic coding is advised

lavaan package in R

- Can do:
  - SEM
  - Bootstrapping
- Cannot (yet) do:
  - hierarchical/multilevel datasets (multilevel cfa, multilevel sem)
  - Latent categorical variables (latent classes)
  - Path diagrams
- More on website: http://lavaan.ugent.be/

lavaan ecosystem (non-random sample)

Path diagram plotting packages

- semPlot
- Onyx
- RAMpath

Extensions

- lavaan.survey: survey ***weights***, clustering, strata, and finite sampling corrections in SEM
- blavaan: Bayesian SEM with a lavaan interface
- semTools: collection of useful functions for SEM including ***multiple imputations***
- simsem: simulation of SEM models

Talk Overview

- What is SEM?
- SEM software packages
- **SEM examples: running the software**
- semPlot package
- Multilevel SEM
- Wrapping up
- Slides there is no time for

Example: Industrialization and Political Democracy (Bollen, 1989)

Expert ratings of the freedom of the press in 1960

The freedom of political opposition in 1960

The fairness of elections in 1960

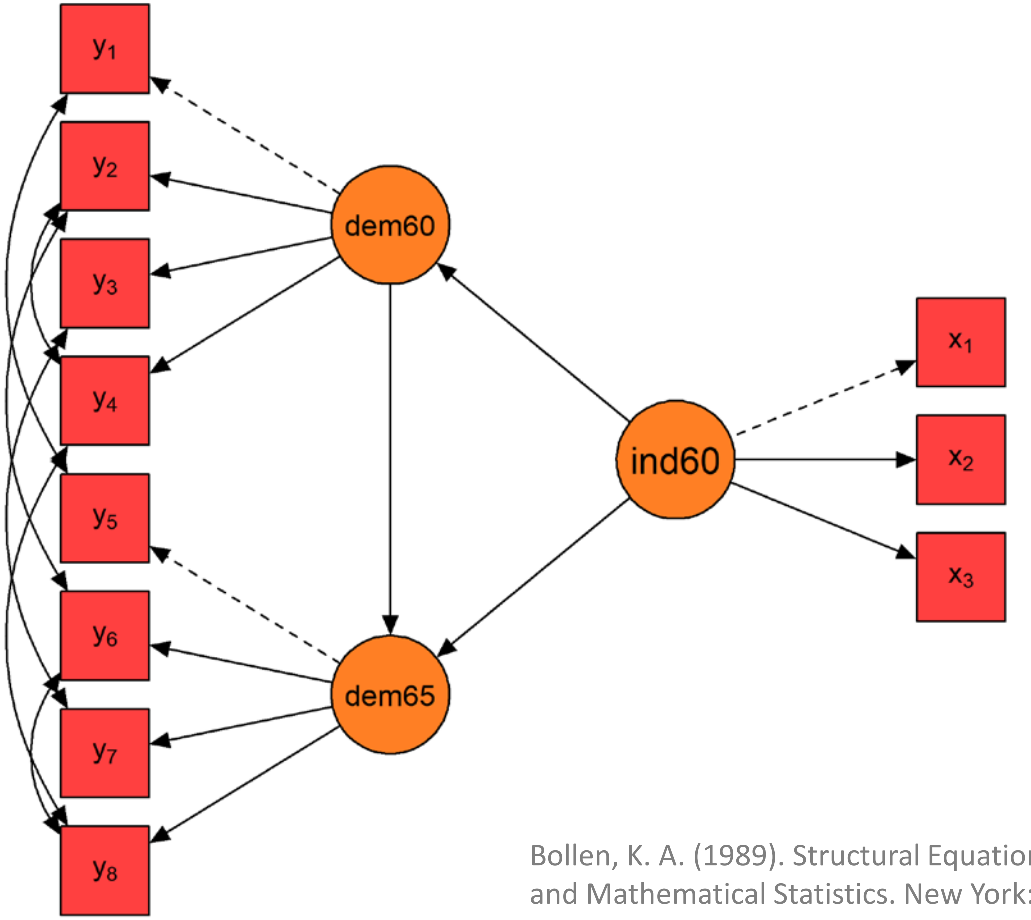
The effectiveness of the elected legislature in 1960

Expert ratings of the freedom of the press in 1965

The freedom of political opposition in 1965

The fairness of elections in 1965

The effectiveness of the elected legislature in 1965



The gross national product (GNP) per capita in 1960

The inanimate energy consumption per capita in 1960

The percentage of the labor force in industry in 1960

Bollen, K. A. (1989). Structural Equations with Latent Variables. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

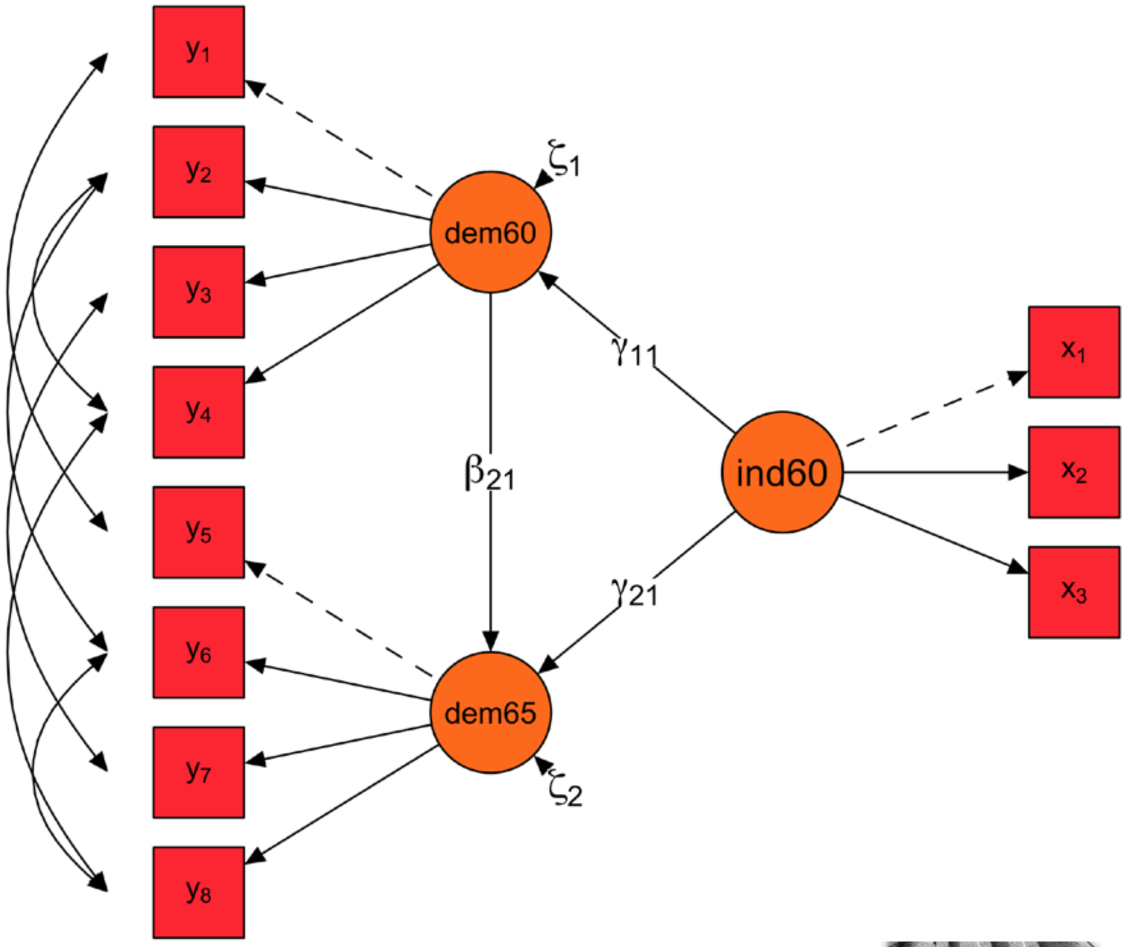
Example: Industrialization And Political Democracy

Latent Variable Model

Vector of endogenous latent variables	eta	$\boldsymbol{\eta}$
Matrix of coefficients	beta	$\boldsymbol{\beta}$
Matrix of coefficients of effect of endogenous latent variables on exogenous latent variables	gamma	$\boldsymbol{\Gamma}$
Vector of exogenous latent variables	xi	$\boldsymbol{\xi}$
Vector of errors	zeta	$\boldsymbol{\zeta}$
Covariance Matrix of latent exogenous variables	phi	$\boldsymbol{\Phi}$
Covariance Matrix of equation errors	psi	$\boldsymbol{\Psi}$

$$\boldsymbol{\eta} = \boldsymbol{\beta}\,\boldsymbol{\eta} + \boldsymbol{\Gamma}\,\boldsymbol{\xi} + \boldsymbol{\zeta}$$

$$\begin{bmatrix} dem60 \\ dem65 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} dem60 \\ dem65 \end{bmatrix} + \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix} [ind60] + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$



Example: Industrialization And Political Democracy

<u>Measurement Model</u>

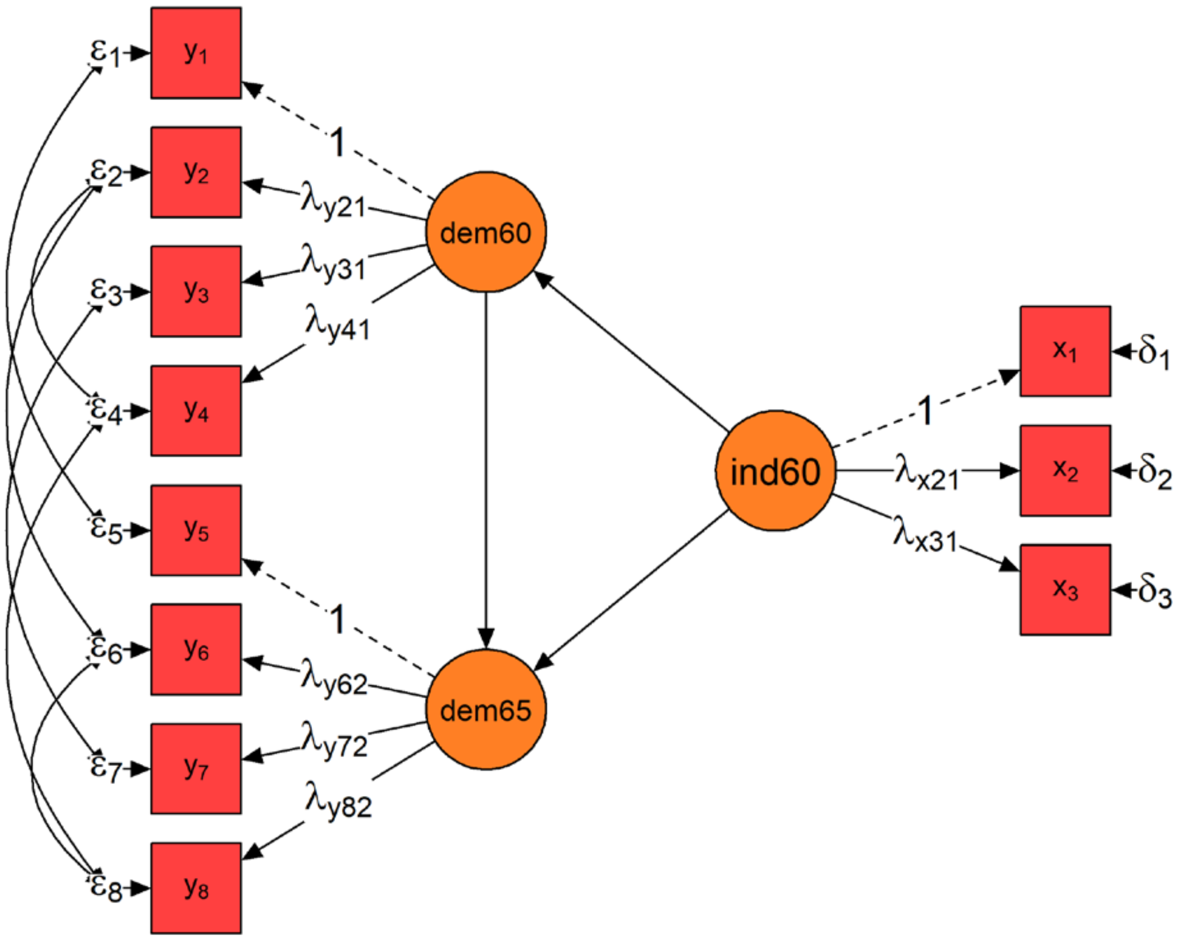
Vector of observed variables for exogenous latent variables	x	**x**
Matrix of coefficients of effect of exogenous latent variables on **x**	lambda	$\boldsymbol{\Lambda_x}$
Vector of error between exogenous latent variables and **x**	delta	$\boldsymbol{\delta}$
Vector of observed variables for endogenous latent variables	y	**y**
Matrix of coefficients of effect of endogenous latent variables on **y**	lambda	$\boldsymbol{\Lambda_y}$
Vector of error between endogenous latent variables and **y**	epsilon	$\boldsymbol{\epsilon}$
Covariance Matrix of errors of observed variables for exogenous latent variables	theta	$\boldsymbol{\Theta_\delta}$
Covariance Matrix of errors of observed variables for endogenous latent variables	theta	$\boldsymbol{\Theta_\epsilon}$

$$\mathbf{x} = \boldsymbol{\Lambda}_\mathbf{x}\, \boldsymbol{\xi} + \boldsymbol{\delta}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_{x21} \\ \lambda_{x31} \end{bmatrix} [ind60] + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

$$\mathbf{y} = \boldsymbol{\Lambda}_\mathbf{y}\, \boldsymbol{\xi} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_{y21} & 0 \\ \lambda_{y31} & 0 \\ \lambda_{y41} & 0 \\ 0 & 1 \\ 0 & \lambda_{y62} \\ 0 & \lambda_{y72} \\ 0 & \lambda_{y82} \end{bmatrix} \begin{bmatrix} dem60 \\ dem65 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \end{bmatrix}$$



R and Mplus: code side by side

```
# Political Democracy example
PolDem<-'# measurement model
    ind60 =~ x1 + x2 + x3
    dem60 =~ y1 + y2 + y3 + y4
    dem65 =~ y5 + y6 + y7 + y8

  # regressions / latent variable model
    dem60 ~ g11*ind60
    dem65 ~ g21*ind60 + b21*dem60

  # residual covariances
    y1 ~~ y5
    y2 ~~ y4 + y6
    y3 ~~ y7
    y4 ~~ y8
    y6 ~~ y8

  # Indirect
    indirect := g11*b21

  # Contrasts
    contrast := g21 - g11*b21

  # Total effect
    total    := g21 + g11*b21
  '
fit_PD<-sem(PolDem, data=PoliticalDemocracy, mimic="Mplus", estimator="MLR")

# Summary of results
summary(fit_PD, fit.measures=TRUE, rsquare=TRUE, ci=TRUE)
```

Due to concerns about *lavaan*'s capabilities, Yves has created 'mimic' arguments in the code that will calculate the answers to mirror Mplus, LISREL, and EQS

```
TITLE:     Political Democracy Example;

DATA:      FILE IS "C:/Users/statmath/Downloads/PD.dat";

VARIABLE:  NAMES ARE y1 y2 y3 y4 y5 y6 y7 y8 x1 x2 x3;
           USEVARIABLES ARE y1 y2 y3 y4 y5 y6 y7 y8 x1 x2 x3;

ANALYSIS:  ESTIMATOR IS MLR;

MODEL:     ind60 BY x1 x2 x3;
           dem60 BY y1-y4;
           dem65 BY y5 y6 y7 y8;

           dem60 ON ind60;
           dem65 ON ind60 dem60;

           y1 WITH y5;
           y2 WITH y4 y6;
           y3 WITH y7;
           y4 WITH y8;
           y6 WITH y8;

MODEL INDIRECT:
           dem65 IND dem60 ind60;


OUTPUT:    STDYX TECH1 TECH4 MODINDICES CINTERVAL;
```

R only

- Since the base R code includes complex mathematical calculations, AVE and CR can be easily computed
  - Where *i* is the number of indicators (observed variables)
  - lambda is the coefficient of effect for the indicator

$$AVE = \frac{\sum_{i=1}^{i} \lambda_i^2}{i}$$

$$CR = \frac{\left(\sum_{i=1}^{i} \lambda_i\right)^2}{\left(\sum_{i=1}^{i} \lambda_i\right)^2 + \left(\sum_{i=1}^{i} 1 - \lambda_i^2\right)}$$

R and Mplus

Let's see this example in each software package

Talk Overview

- What is SEM?
- SEM software packages
- SEM examples: running the software
- **semPlot package**
- Multilevel SEM
- Wrapping up
- Slides there is no time for

semPlot

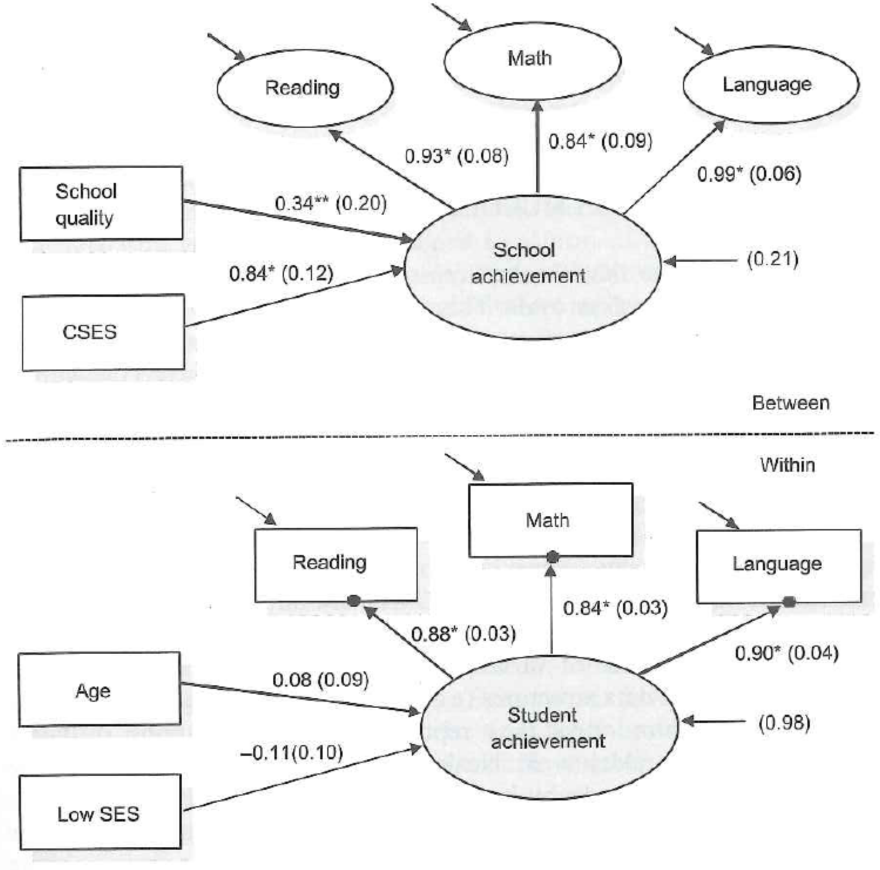
Back to the software !!!

Talk Overview

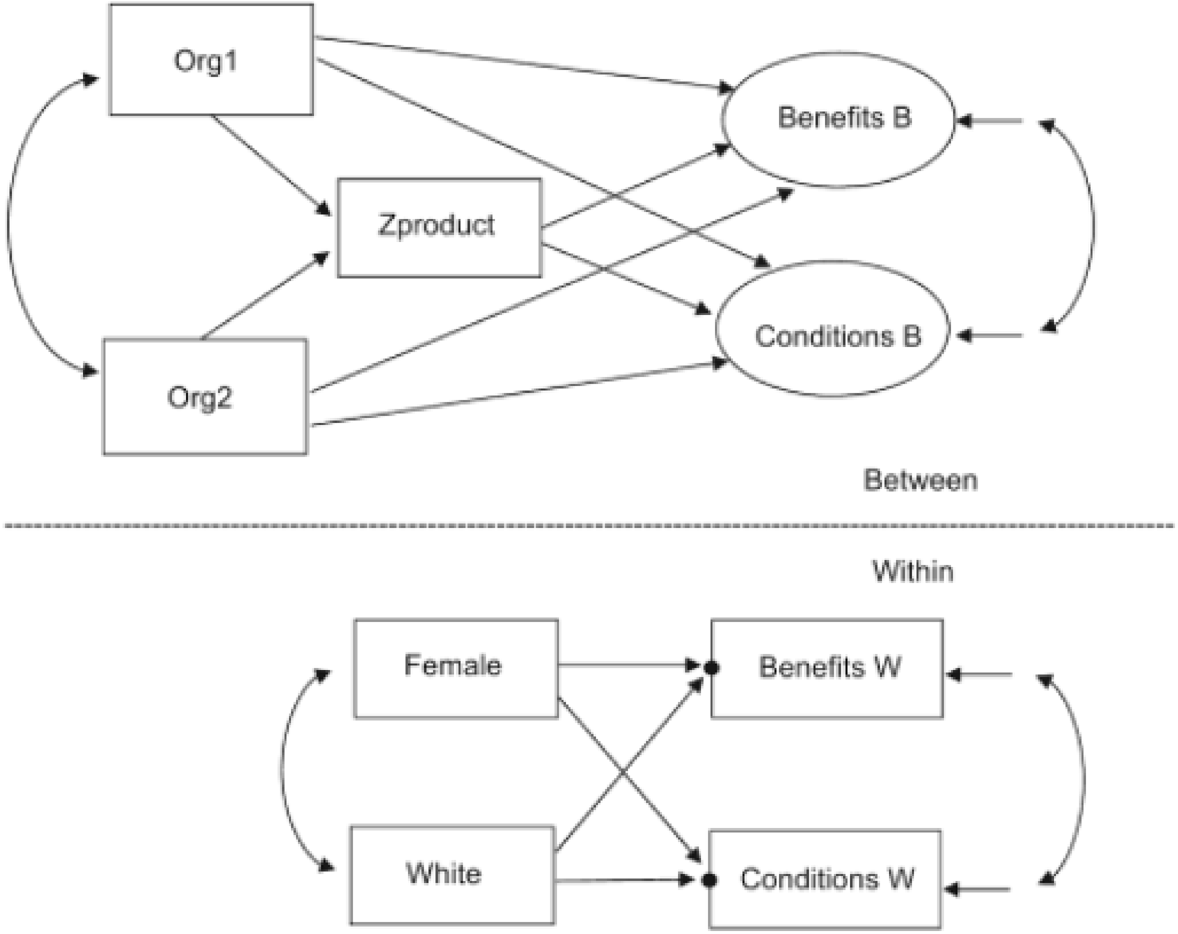
- What is SEM?
- SEM software packages
- SEM examples: running the software
- semPlot package
- **Multilevel SEM**
- Wrapping up
- Slides there is no time for

Multilevel SEM

- New R package: OpenMX
  - web page is hosted and maintained by the Social Science Research Institute and Quantitative Developmental Systems group at Penn State
  - Pre-built package binaries are hosted by the Human Dynamics Lab in the Department of Psychology at the University of Virginia
  - Specification using syntax of paths **or** matrices



Multilevel SEM: example



R and Mplus: code side by side

```r
library(OpenMx)

flatData<-read.table(file.choose())
colnames(flatData)<-c('orgid','female','white','satpay','morale','org1',
                      'org2','benefit','cond','resour','zproduct','lev1wt','lev2wt')

level1<-flatData[,c('benefit','cond','female','white','orgid')]
level2<-flatData[!duplicated(flatData$orgid),c('orgid', 'org1','org2','zproduct')]

orgModel<-mxModel("org", type="RAM",
                  mxData(level2, 'raw', primaryKey = "orgid"),
                  manifestVars=c('zproduct'),
                  latentVars  =c('Benefits','Conditions','org1','org2'),

                  # Residuals
                  mxPath(c('zproduct','Benefits','Conditions'), arrows=2, values=1),

                  # Latent covaraince
                  mxPath('Benefits','Conditions', arrows=2),

                  # Latent paths
                  mxPath('org1'    , c('Benefits', 'Conditions')),
                  mxPath('org2'    , c('Benefits', 'Conditions')),
                  mxPath('org2'    , 'zproduct'),

                  # Manifest paths
                  mxPath('zproduct', c('Benefits','Conditions'))
                 )
```

```r
empModel<-mxModel("emp", type="RAM", orgModel,
                  mxData(level1, 'raw'),
                  manifestVars=c('benefit','cond'),
                  latentVars  =c('female','white'),

                  # Latent paths
                  mxPath(c('female','white'), c('benefit','cond'), connect = "all.bivariate"),

                  # Residuals
                  mxPath(c('benefit','cond'), arrows=2, connect = "unique.pairs", values=c(1,0,1)),

                  # Connecting to BETWEEN model
                  mxPath('org.Benefits'     , 'benefit', values=1, free=FALSE, joinKey = "orgid"),
                  mxPath('org.Conditions'   , 'cond'   , values=1, free=FALSE, joinKey = "orgid")
                 )

empModel<-mxRun(empModel)
summary(empModel)
```

```
TITLE:     Multilevel example;

DATA:      FILE IS "C:/Users/statmath/Downloads/ch4mv.dat";
           FORMAT IS 7f8.0,6f8.2;

VARIABLE:  NAMES ARE orgid female white satpay morale org1
                     org2 benefit cond resour zproduct lev1wt lev2wt;
           USEVARIABLES ARE benefit cond female white org1 org2 zproduct;
           within  = female white;
           between = org1 org2 zproduct;
           CLUSTER IS orgid;

ANALYSIS:  TYPE = twolevel; ESTIMATOR IS MLR;

MODEL:     %Between%
           benefit cond ON   org1;
           zproduct     ON   org2(a);
           benefit cond ON   zproduct(b1 b2);
           benefit      WITH cond;
           %Within%
           benefit cond ON   female white;
           benefit      WITH cond;
           Model constraint: New(ind1 ind2);
           ind1=a*b1; ind2=a*b2;

OUTPUT:    STDYX TECH1 TECH4 TECH8 MODINDICES CINTERVAL;
```

Talk Overview

- What is SEM?
- SEM software packages
- SEM examples: running the software
- semPlot package
- Multilevel SEM
- **Wrapping up**
- Slides there is no time for

Advantages of SEM

- Confirmatory approach: test your model/theory
- Provides goodness-of-fit measures
- Flexible statistical modeling approach
- SEM can handle:
  - Missing data (fiml, multiple imputations)
  - (in)equality constraints
  - Mixed categorical data (binary, ordinal, count,...)
  - Mixed discrete and continuous latent variables
  - Clustered (multilevel) data
- Many other approaches are special cases (e.g. generalized linear mixed models)

Disadvantages of SEM

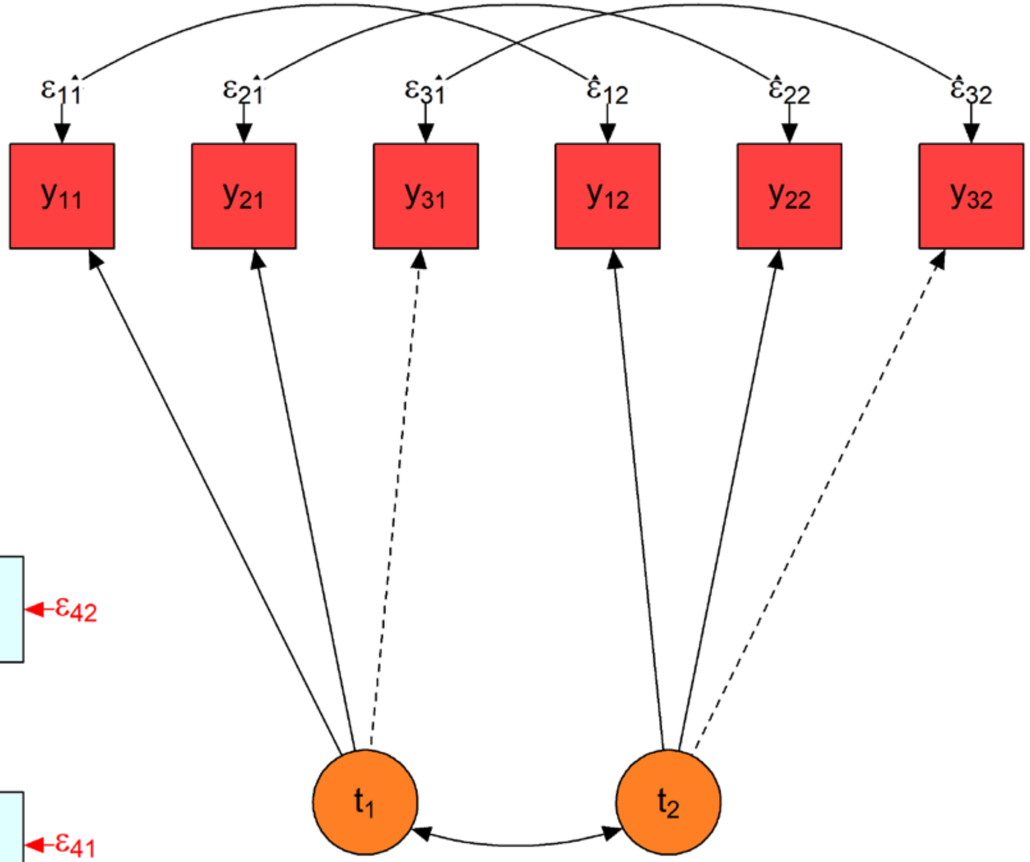
- Modeling flexibility can be overwhelming (e.g. there are connections everywhere!!!)
- Dedicated software needed (not available in SPSS)
- 'Specifying' model in software can be challenging
- As a statistical field
  - Better (?) inference for small samples
  - Understanding output can be a challenge without 'complete' knowledge of the model and techniques
- Not well connected to other branches of statistics (mixed methods,...)

Talk Overview

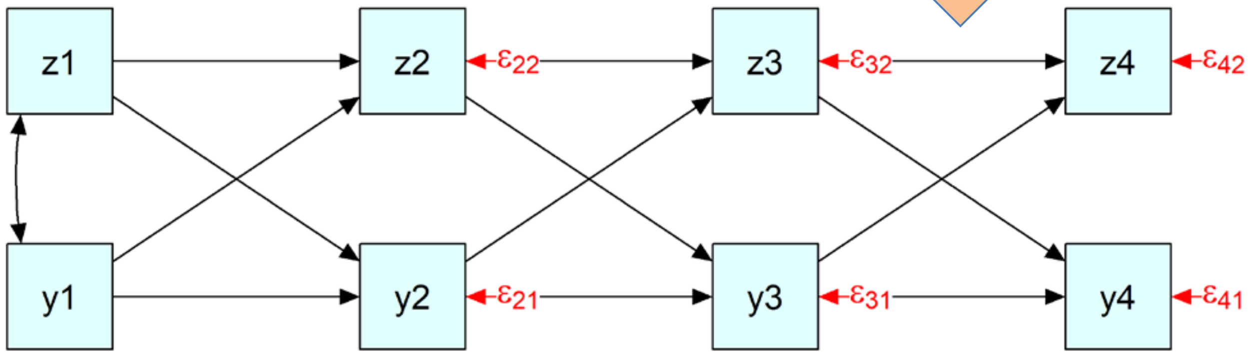
- What is SEM?
- SEM software packages
- SEM examples: running the software
- semPlot package
- Multilevel SEM
- Wrapping up
- **Slides there is no time for**

More SEM examples

- Longitudinal: paired t-test using latent variables



- Panel models with crosslagged effects



Other interesting things that we can't get into

- Lavaan ecosystem (extended)
  - nlsem (Nora Umbach and many others, since 2014): estimation of structural equation models with nonlinear effects and underlying nonnormal distributions
  - EffectLiteR (Axel Mayer, Lisa Dietzfelbinger, since 2015): using SEM to estimate average and conditional effects
  - Lavaan.shiny (W. Kyle Hamilton, since 2016): Interactive shiny application for working with different kinds of latent variable analysis, with the 'lavaan' package. Graphical output for models are provided and different estimators are supported.
  - bmem, coefficientalpha, eqs2lavaan, fSRM, influence.SEM, MIIVsem, profileR, RAMpath, regsem, RMediation, RSA, rsem, stremo, faoutlier, gimme, matrixpls, MBESS, NlsyLinks, nonnest2, piecewiseSEM, pscore, psytabs, qgraph, sesem, sirt, TAM, userfriendlyscience, . . .

More (interesting) detail about lavaan ecosystem packages

- Onyx (Timo von Oertzen, Andreas M. Brandmaier, Siny Tsang, since 2009): interactive graphical interface for SEM (written in Java)
- lavaan.survey (Daniel Oberski, since 2012): survey weights, clustering, strata, and finite sampling corrections in SEM
- semTools (Sunthud Pornprasertmanit and many others, since 2012): collection of useful functions for SEM including multiple imputations
- simsem (Sunthud Pornprasertmanit and many others, since 2012): simulation of SEM models
- RAMpath (Zhiyong Zhang, Jack McArdle, Aki Hamagami, Kevin Grimm, since 2012) : visualizations of SEM models
- semPlot (Sacha Epskamp, since 2013): visualizations of SEM models
- blavaan (Ed Merkle, Yves Rosseel, since 2015): Bayesian SEM (currently using jags) with a lavaan interface

Other interesting things that we can’t get into

- SEM and regression used to analyze the same data:
  - https://pdfs.semanticscholar.org/5280/080a464f14ffe4fdf7ef501a6ff581cb34e8.pdf
  - https://ac.els-cdn.com/S0261517709000521/1-s2.0-S0261517709000521-main.pdf?_tid=8c07ce88-d5d8-11e7-b8dd-00000aab0f27&acdnat=1512051358_f1bf721894435f4c1ff7e404b914085b
  - https://repositories.lib.utexas.edu/bitstream/handle/2152/22516/XIAO-MASTERSREPORT-2013.pdf?sequence=1&isAllowed=y

- SEM and ordinal data:
  - https://www.academia.edu/7605083/Evaluating_estimation_methods_for_ordinal_data_in_structural_equation_modeling?auto=download
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3103762/
  - http://cmss.univnt.ro/wp-content/uploads/vol/split/vol_IV_issue_1/CMSS_vol_IV_issue_1_art.002.pdf

- Latent Categorical Variables discussion & examples:
  - https://stats.idre.ucla.edu/mplus/seminars/intromplus-part2/mplus-class-notesanalyzing-data-latent-class-and-other-mixture-models/
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3823694/

- Yves Rosseel keynote:
  - https://channel9.msdn.com/Events/useR-international-R-User-conferences/useR-International-R-User-2017-Conference/KEYNOTE-Structural-Equation-Modeling-models-software-and-stories

Software websites of interest

- Stata's gllamm:
  - http://www.gllamm.org/
- REQS:
  - https://cran.r-project.org/web/packages/REQS/REQS.pdf
  - https://escholarship.org/uc/item/2pk051gv
- Mplus Automation:
  - https://cran.r-project.org/web/packages/MplusAutomation/MplusAutomation.pdf
- SEM Stat Wiki using AMOS:
  - http://statwiki.kolobkreations.com/index.php?title=Main_Page