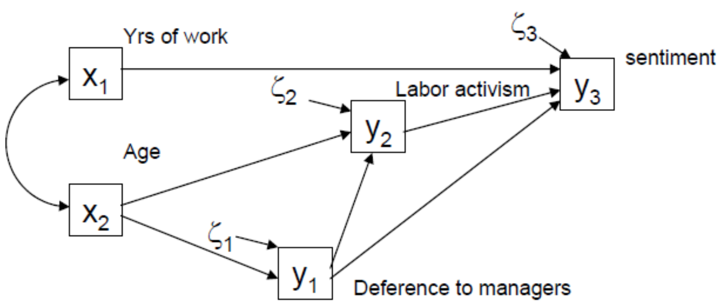
Confirmatory Factor Analysis & Structural Equation Models

Lecture 1: Overview & Path Analysis

Michael Friendly

SCS Short Course, May, 2019



Course overview

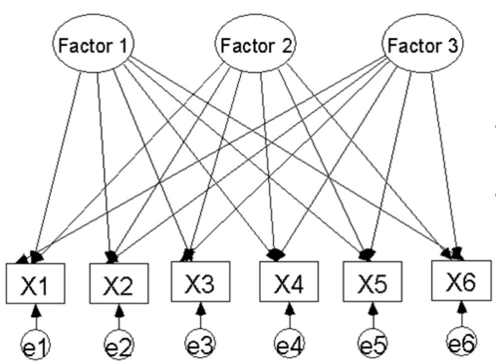
Course notes & other materials will be avaiable at:
`http://datavis.ca/courses/CFA-SEM`

- Lecture 1: Setting the stage: EFA, CFA, SEM, Path analysis
  - Goal: Understand relations among a large number of observed variables
  - Goal: Extend regression methods to (a) multiple outcomes, (b) latent variables, (c) accounting for measurement error or unreliability
  - Thinking: Equations → Path diagram → estimate, test, visualize
- Lecture 2: Measurement models & CFA
  - Effects of measurement error
  - Testing equivalence of measures with CFA
  - Multi-factor, higher-order models
- Lecture 3: SEM with latent variables

EFA, CFA, SEM?

Exploratory Factor Analysis (EFA)

- Method for “explaining” correlations of observed variables in terms of a small number of “common factors”
- Primary Q: How many factors are needed?
- Secondary Q: How to interpret the factors?

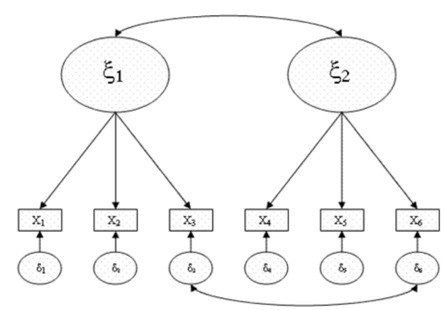


Three-factor EFA model. Each variable loads on all factors.
The factors are assumed to be uncorrelated

EFA, CFA, SEM?

Confirmatory Factor Analysis (CFA)

- Method for testing hypotheses about relationships among observed variables
- Does this by imposing restrictions on an EFA model
- Q: Do the variables have a given factor structure?
- Q: How to compare competing models?



Two-factor CFA model with non-overlapping factors
The factors are allowed to be correlated, as are two unique factors

EFA, CFA, SEM?

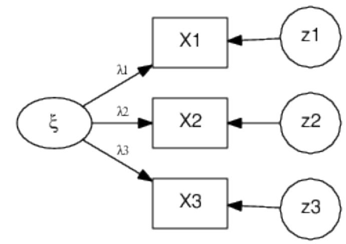
Structural Equation Models (SEM)

- Generalizes EFA, CFA to include
  - Simple and multiple regression
  - General linear model (Anova, multivariate regression, ...)
  - Path analysis — several simultaneous regression models
  - Higher-order CFA models
  - Multi-sample CFA models ("factorial invariance")
  - Latent growth/trajectory models
  - Many more ...
- A general framework for describing, estimating and testing linear statistical models

Recall basic EFA ideas

- Observed variables, $x_1, x_2, \ldots, x_p$ is considered to arise as a set of regressions on some *unobserved, latent variables* called ***common factors***, $\xi_1, \xi_2, \ldots, \xi_k$.
- That is, each variable can be expressed as a regression on the common factors. For three variables and one common factor, ξ, the model is:

$$\begin{aligned} x_1 &= \lambda_1 \xi + z_1 \\ x_2 &= \lambda_2 \xi + z_2 \\ x_3 &= \lambda_3 \xi + z_3 \end{aligned}$$



- The common factors account for correlations among the xs.
- The z_i are error terms, or unique factors

The EFA model

- For k common factors, the common factor model is

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \cdots + \lambda_{1k} + z_1 \\ x_2 &= \lambda_{21}\xi_1 + \cdots + \lambda_{2k} + z_2 \\ &\vdots \\ x_p &= \lambda_{p1}\xi_1 + \cdots + \lambda_{pk} + z_p \end{aligned}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \lambda_{21} & \cdots & \lambda_{2k} \\ \vdots & \vdots & \vdots \\ \lambda_{p1} & \vdots & \lambda_{pk} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_k \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix}$$

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \mathbf{z}$$

- This looks like a set of multiple regression models for the xs, but it is not testable, because the factors, ξ, are unobserved

The EFA model

- However, the EFA model implies a particular form for the variance-covariance matrix, $\mathbf{\Sigma}$, which is testable

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \mathbf{z} \quad \Longrightarrow \quad \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^T + \mathbf{\Psi}$$

where:

- $\mathbf{\Lambda}_{p \times k}$ = factor pattern ("loadings")
- $\mathbf{\Phi}_{k \times k}$ = matrix of correlations among factors.
- $\mathbf{\Psi}$ = diagonal matrix of unique variances of observed variables.

- Typically, it is initially assumed that factors are uncorrelated ($\Phi = I$, the identity matrix)
- Can use an oblique rotation to allow correlated factors

Limitations of EFA

- The only true statistical tests in EFA are tests for the number of common factors (when estimated by ML)

$$H_0 : k = k_0 \quad k_0 \text{ factors are sufficient}$$
$$H_a : k > k_0 \quad > k_0 \text{ factors are necessary}$$

- Substantive questions about the nature of factors can only be addressed approximately through factor rotation methods
  - Varimax & friends attempt rotation to simple structure
  - Oblique rotation methods allow factors to be correlated
  - Procrustes rotation allows rotation to a “target” (hypothesized) loading matrix

Historical development: EFA $\rightarrow$ CFA

- ML estimation for the EFA model finds estimates that minimize the difference between the observed covariance matrix, $\boldsymbol{S}$, and that reproduced by the model, $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Phi}}\widehat{\boldsymbol{\Lambda}}^{\mathsf{T}} + \widehat{\boldsymbol{\Psi}}$
  - Requires imposing k^2 restrictions for a unique solution
  - Gives a χ^2 test for goodness of fit

$$(N-1)F_{min}(\boldsymbol{S}, \widehat{\boldsymbol{\Sigma}}) \sim \chi^2 \quad \text{with } df = [(p-k)^2 - p - k]/2$$

- Joreskog (1969) proposed that a factor hypothesis could be tested by imposing restrictions on the EFA model— fixed elements in $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$, usually 0
  - Needs more than k^2 restrictions
  - The ML solution is then found for the remaining free parameters
  - The χ^2 for the restricted solution gives a test for how well the hypothesized factor structure fits.

CFA: Restricted EFA

The pattern below specifies two non-overlapping oblique factors. The x’s are the only free parameters.

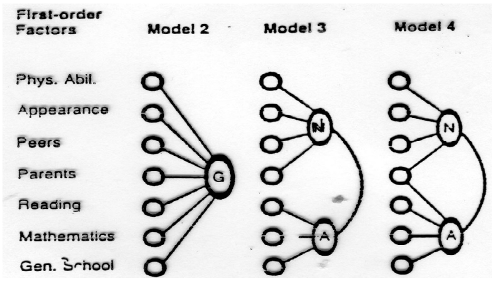
$$\Lambda = \begin{bmatrix} x & 0 \\ x & 0 \\ x & 0 \\ 0 & x \\ 0 & x \\ 0 & x \end{bmatrix} \qquad \Phi = \begin{bmatrix} 1 & \\ x & 1 \end{bmatrix}$$

- This CFA model has only 7 free parameters and $df = 15 - 7 = 8$.
- A $k = 2$-factor EFA model would have all parameters free and $df = 15 - 11 = 4$ degrees of freedom.
- If this restricted model fits (has a small χ^2/df), it is strong evidence for two non-overlapping oblique factors.
- That hypothesis cannot be tested by EFA + rotation.

Historical development: CFA $\rightarrow$ SEM

Higher-order factor analysis: The ACOVS model

- With more than a few factors, allowed to be correlated ($\boldsymbol{\Phi} \neq \boldsymbol{I}$), can we factor the factor correlations?
- In EFA, this was done by another EFA of the estimated factor correlations from an oblique rotation
- The second stage of development of CFA/SEM models combined these steps into a single model, and allowed different hypotheses to be compared



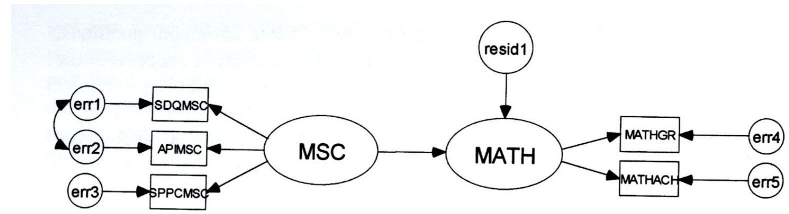
LISREL/SEM Model

- Jöreskog (1973) further generalized the ACOVS model to include structural equation models along with CFA.
- Two parts:
  - Measurement model — How the latent variables are measured in terms of the observed variables; measurement properties (reliability, validity) of observed variables. [Traditional factor analysis models]
  - Structural equation model — Specifies causal relations among observed and latent variables.
    - Endogenous variables - determined within the model (***y***)
    - Exogenous variables - determined outside the model (***x***)

Measurement models for observed variables	$\boldsymbol{x} = \Lambda_x \xi + \delta$ $\boldsymbol{y} = \Lambda_y \eta + \epsilon$
Structural eqn. for latent variables	$\eta = \boldsymbol{B}\eta + \Gamma\xi + \zeta$

LISREL/SEM Model

SEM model for measures of Math Self-Concept and MATH achievement:

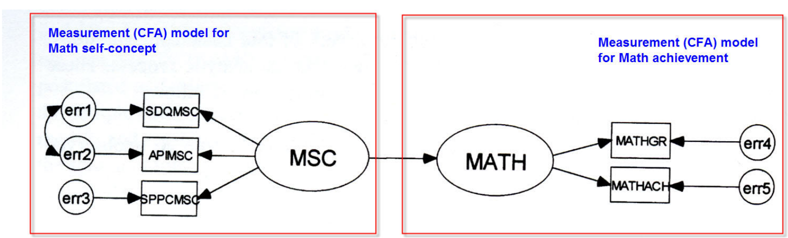


This model has:

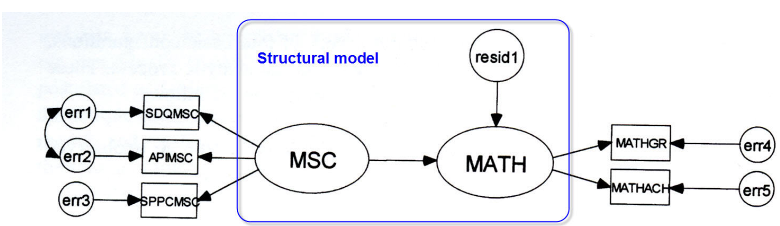
- 3 observed indicators in a measurement model for MSC (***x***)
- 2 observed indicators in a measurement model for MATH achievement (***y***)
- A structural equation predicting MATH achievement from MSC

LISREL/SEM Model

Measurement sub-models for ***x*** and ***y***



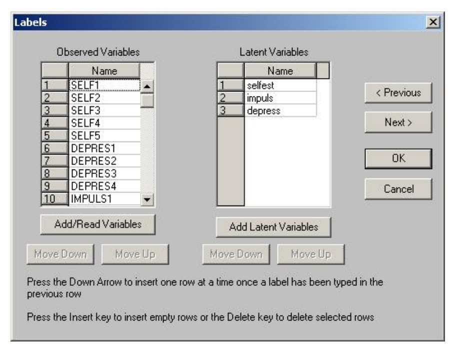
Structural model, relating ξ to η

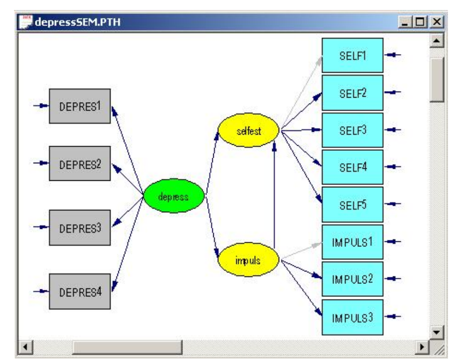


CFA/SEM software: LISREL

LISREL (http://www.ssicentral.com/) [student edition available]

- Originally designed as stand-alone program with matrix syntax
- LISREL 8.5+ for Windows/Mac: includes
  - interactive, menu-driven version;
  - PRELIS (pre-processing, correlations and models for categorical variables);
  - SIMPLIS (simplified, linear equation syntax)
  - path diagrams from the fitted model

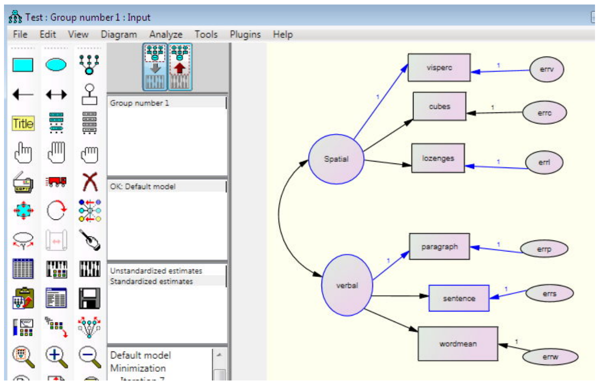




CFA/SEM software: Amos

Amos (www.ibm.com/software/products/en/spss-amos): Linear equation syntax + path diagram model description

- import data from SPSS, Excel, etc; works well with SPSS
- Create the model by drawing a path diagram
- simple facilities for multi-sample analyses
- nice comparative displays of multiple models



SAS: PROC CALIS

- SAS 9.3+: PROC CALIS
  - MATRIX (à la LISREL), LINEQS (à la EQS), RAM, ... syntax
  - Now handles multi-sample analyses
  - Multiple-model analysis syntax, e.g., Model 2 is like Model 1 except ...
  - Enhanced output controls
  - customizable fit summary table
- SAS macros `http://datavis.ca/sasmac/`:
  - `caliscmp` macro: compare model fits from PROC CALIS à la Amos
  - `csmpower` macro: power estimation for covariance structure models

R: sem, lavaan and others

- sem package (John Fox)
  - flexible ways to specify models: `cfa()`, `linearEquations()`, and `multigroupModel()`
  - `bootSem()` provides bootstrap analysis of SEM models
  - `miSem()` provides multiple imputation
  - path diagrams using `pathDiagram()` → `graphviz`
  - `polychor` package for polychoric correlations
- lavaan package (Yves Rossell)
  - Functions `lavaan()`, `cfa()`, `sem()`, `growth()` (growth curve models)
  - Handles multiple groups models
  - semTools provides tests of measurement invariance, multiple imputation, bootstrap analysis, power analysis for RMSEA, ...
- semPlot package — path diagrams for sem, lavaan, Mplus, ... models

Mplus

Mplus `https://www.statmodel.com/` [$$$, but cheaper student price]

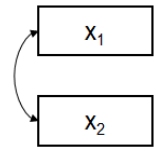
- Handles the widest range of models: CFA, SEM, multi-group, multi-level, latent group
- Variables: continuous, censored, binary, ordered categorical (ordinal), unordered categorical (nominal), counts, or combinations of these variable types
- For binary and categorical outcomes: probit, logistic regression, or multinomial logistic regression models.
- For count outcomes: Poisson and negative binomial regression models.
- Extensive facilities for simulation studies.

Caveats

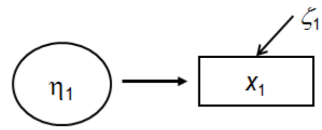
- CFA and SEM models are fit using the covariance matrix (***S***)
  - The raw data is often not analyzed
  - Graphs that can reveal potential problems often not made
- Typically, this assumes all variables are complete, continuous, multivariate normal. Implies:
  - ***S*** is a sufficient statistical summary
  - Relations assumed to be linear are in fact linear
  - Goodness-of-fit (χ^2) and other tests based on asymptotic theory ($N \rightarrow \infty$)
  - Missing data, skewed or long-tailed variables must be handled first
- Topics not covered here:
  - Using polychoric correlations for categorical indicators
  - Distribution-free estimation methods (still asymptotic)
  - Bootstrap methods to correct for some of the above
  - Multiple imputation to handle missing data

Path diagrams: Symbols

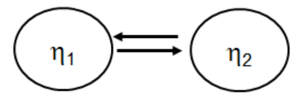
Visual representation of a set of simultaneous equations for EFA, CFA, SEM models (idea from Sewell Wright, 1920s)



- Rectangles and square boxes represent observed (manifest) variables
- Double-headed curved arrows represent unanalyzed association (correlation)



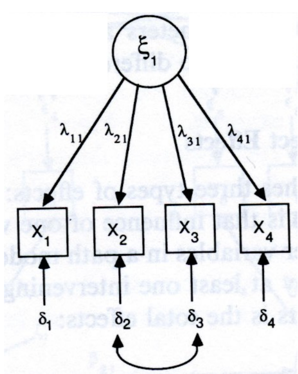
- Ellipses and circles represent unobserved (latent) variables
- Straight, single-headed arrows indicate causal relations from base to head
- Unenclosed symbol is an error term (in equation or measurement error)



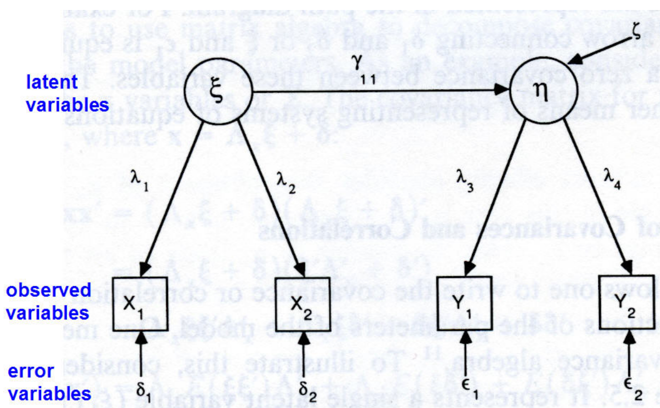
- Double single-headed arrows indicate reciprocal causation
- Allows different weights for each path

Path diagrams

Schematic Examples:



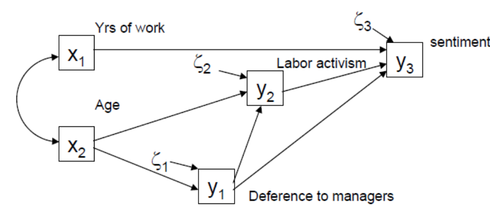
CFA, 1-factor model (correlated errors)



SEM, two latent variables, each with two indicators
Causal relation betweeen ξ (Xs) and η (Ys)

Path diagrams

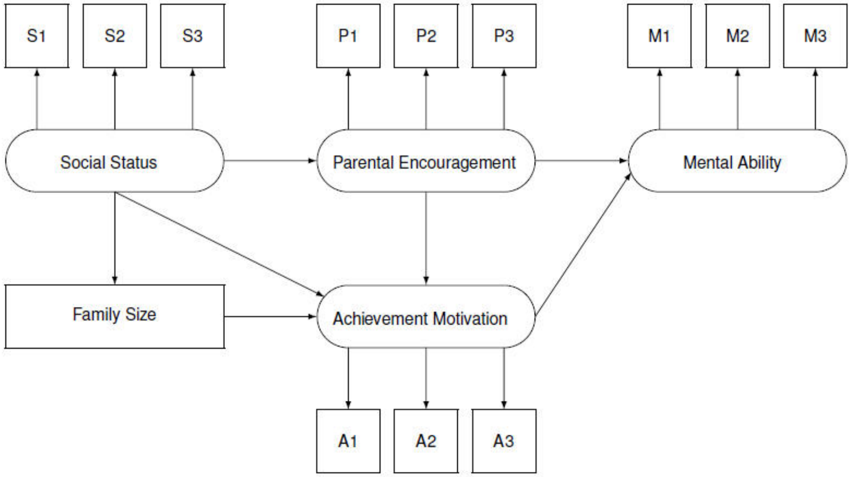
Substantive example: Path analysis model for union sentiment (McDonald & Clelland, 1984)



- No latent variables— all variables are observed indicators
- x_1, x_2 are exogenous variables— they are not explained within the model
- Correlation between x_1, x_2 is shown as a double-headed arrow
- y_1, y_2, y_3 are endogenous variables— they are explained within the model
- Causal relations are shown among the variables by single-headed arrows
- Residual (error) terms, $\zeta_1, \zeta_2, \zeta_3$ are shown as single-headed arrows to the y variables

Path diagrams

Substantive example: SEM with multiple indicators, path model for latent variables (error terms not shown)



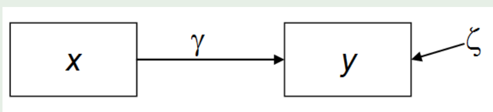
Path Analysis

- Path analysis is a simple special case of SEM
  - These models contain only observed (manifest) variables,
  - No latent variables
  - Assumes that all variables are measured without error
  - The only error terms are residuals for y (endogenous) variables
- They are comprised of a set of linear regression models, estimated simultaneously
  - Traditional approaches using MRA fit a collection of separate models
  - Multivariate MRA (MMRA) usually has all y variables predicted by all x variables
  - In contrast, SEM path models allow a more general approach, in a single model

Path Analysis: Simple examples

Simple linear regression

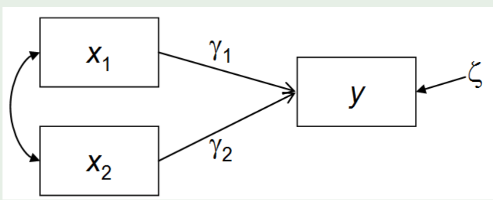
$$y_i = \gamma x_i + \zeta_i$$



- γ is the slope coefficient; ζ is the residual (error term)
- Means and regression intercepts usually not of interest, and suppressed

Multiple regression

$$y_i = \gamma_1 x_{1i} + \gamma_2 x_{2i} + \zeta_i$$

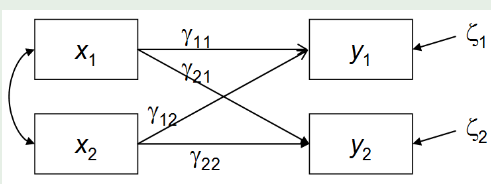


- Double-headed arrow signifies the assumed correlation between x_1 & x_2
- In univariate MRA ($y \sim x_1 + \ldots$), there can be any number of xs

Path Analysis: Simple examples

Multivariate multiple regression

$$\begin{aligned} y_{1i} &= \gamma_{11} x_{1i} + \gamma_{12} x_{2i} + \zeta_{1i} \\ y_{2i} &= \gamma_{21} x_{2i} + \gamma_{22} x_{2i} + \zeta_{2i} \end{aligned}$$



- Now need two equations to specify the model
- Note subscripts: γ_{12} is coeff of y_1 on x_2; γ_{21} is coeff of y_2 on x_1

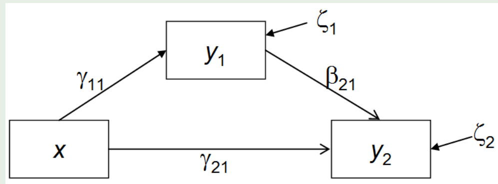
With more equations and more variables, easier with vectors/matrices

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{\Gamma} \boldsymbol{x} + \boldsymbol{\zeta}$$

Path Analysis: Simple examples

Simple mediation model

$$y_{1i} = \gamma_{11} x_i + \zeta_{1i}$$
$$y_{2i} = \gamma_{21} x_i + \beta_{21} y_{1i} + \zeta_{2i}$$



- Something new: y_1 is a dependent variable in the first equation, but a predictor in the second
- This cannot be done simultaneously via standard MRA or MMRA models

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} \gamma_{11} \\ \gamma_{21} \end{pmatrix} x + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{B}\boldsymbol{y} + \boldsymbol{\Gamma}\boldsymbol{x} + \boldsymbol{\zeta}$$

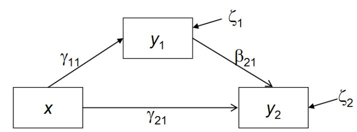
Exogenous and Endogenous Variables

Exogenous variables

- Are only independent (x) variables in the linear equations
- Never have arrows pointing at them from other variables
- They are determined outside ("ex") the model
- In path analysis models they are considered measured w/o error

Endogenous variables

- Serves as a dependent variable (outcome) in at least one equation
- If a variable has at least one arrow pointing to it, it is endogenous
- They are determined inside ("en") the model
- In path analysis models they always have error terms



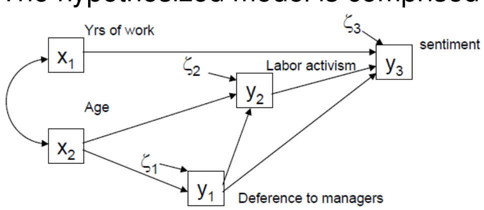
In the simple mediation model, x is exogenous, and y_1, y_2 are endogenous

Example: Union sentiment

Norma Rae example— Union sentiment among non-union Southern textile workers (McDonald & Clelland (1984); Bollen (1986))

- Exogenous variables: x_1 (years of work); x_2 (age)
- Endogenous variables: y_1 (deference to managers); y_2 (support for labor activism); y_3 (support for unions)

The hypothesized model is comprised of three linear regressions



$$y_1 = \gamma_{12} x_2 + \zeta_1$$
$$y_2 = \beta_{21} y_1 + \gamma_{22} x_2 + \zeta_2$$
$$y_3 = \beta_{31} y_1 + \beta_{32} y_2 + \gamma_{31} x_1 + \zeta_3$$

These can be expressed as a single matrix equation for the $\boldsymbol{y}$ variables:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} + \begin{bmatrix} 0 & \gamma_{12} \\ 0 & \gamma_{22} \\ \gamma_{31} & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{pmatrix}$$

The general path analysis model

The general form of a SEM path analysis model is expressed in the matrix equation

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{y} + \boldsymbol{\Gamma}\boldsymbol{x} + \boldsymbol{\zeta}$$

where:

- $\boldsymbol{y}$ is a $p \times 1$ vector of endogenous variables
- $\boldsymbol{x}$ is a $q \times 1$ vector of exogenous variables
- $\boldsymbol{B}_{p \times p}$ ("Beta") gives the regression coefficients of endogenous ($\boldsymbol{y}$) variables on other endogenous variables
- $\boldsymbol{\Gamma}_{p \times q}$ ("Gamma") gives the regression coefficients of endogenous variables on the exogenous variables ($\boldsymbol{x}$)
- $\boldsymbol{\zeta}_{p \times 1}$ is the vector of errors in the equations (i.e., regression residuals)

However, some parameters in $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ are typically fixed to 0

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{bmatrix} \quad \boldsymbol{\Gamma} = \begin{bmatrix} 0 & \gamma_{12} \\ 0 & \gamma_{22} \\ \gamma_{31} & 0 \end{bmatrix}$$

The general path analysis model

Other parameters pertain to variances and covariances of the exogenous variables and the error terms

- $\Phi_{q \times q}$ ("Phi")— variance-covariance matrix of the exogenous variables. Typically, these are all free parameters.
  For the union sentiment example, Φ is a 2×2 matrix:

$$\Phi = \begin{bmatrix} \text{var}(x_1) & \\ \text{cov}(x_1, x_2) & \text{var}(x_2) \end{bmatrix}$$

- $\Psi_{p \times p}$ ("Psi")— variance-covariance matrix of the error terms (ζ).
  Typically, the error variances are free parameters, but their covariances are fixed to 0 (models can allow correlated errors)
  For the union sentiment example, Ψ is a 3×3 diagonal matrix:

$$\Psi = \begin{bmatrix} \text{var}(\zeta_1) & & \\ 0 & \text{var}(\zeta_2) & \\ 0 & 0 & \text{var}(\zeta_2) \end{bmatrix}$$

Union sentiment: using the sem package

Read the variance-covariance matrix of the variables using **readMoments()**

```
library(sem)
union <- readMoments(diag=TRUE,
           names=c('y1', 'y2', 'y3', 'x1', 'x2'),
           text="
    14.610
    -5.250  11.017
    -8.057  11.087   31.971
    -0.482   0.677    1.559   1.021
   -18.857  17.861   28.250   7.139  215.662
")
```

The model can be specified in different, equivalent notations, but the simplest is often linear equations format, with **specifyEquations()**

```
union.mod <- specifyEquations(covs="x1, x2", text="
   y1 = gam12*x2
   y2 = beta21*y1 + gam22*x2
   y3 = beta31*y1 + beta32*y2 + gam31*x1
")
```


Union sentiment: using the sem package

Internally, sem expresses the model using "RAM" path notation (same as used by **specifyModel()**):

```
union.mod

##    Path       Parameter
## 1  x2 -> y1  gam12
## 2  y1 -> y2  beta21
## 3  x2 -> y2  gam22
## 4  y1 -> y3  beta31
## 5  y2 -> y3  beta32
## 6  x1 -> y3  gam31
## 7  x1 <-> x1 V[x1]
## 8  x1 <-> x2 C[x1,x2]
## 9  x2 <-> x2 V[x2]
## 10 y1 <-> y1 V[y1]
## 11 y2 <-> y2 V[y2]
## 12 y3 <-> y3 V[y3]
```

Fit the model using **sem()**:

```
union.sem <- sem(union.mod, union, N=173)
```


Union sentiment: Goodness-of-fit statistics

The **summary()** method prints a collection of goodness-of-fit statistics:

```
opt <- options(fit.indices = c("GFI", "AGFI", "RMSEA",  "NNFI",
         "CFI", "AIC", "BIC"))
summary(union.sem)
```

```
##
##  Model Chisquare =  1.25   Df =  3 Pr(>Chisq) = 0.741
##  Goodness-of-fit index =  0.997
##  Adjusted goodness-of-fit index =  0.986
##  RMSEA index =  0   90% CI: (NA, 0.0904)
##  Tucker-Lewis NNFI =  1.0311
##  Bentler CFI =  1
##  AIC =  25.3
##  BIC =  -14.2
##
##  ...
##
##  R-square for Endogenous Variables
##    y1    y2    y3
## 0.113 0.230 0.390
##
## ...
```

Union sentiment: Parameter estimates

```
##  Parameter Estimates
##         Estimate Std Error z value Pr(>|z|)
## gam12     -0.0874  0.0187   -4.68   2.90e-06 y1 <--- x2
## beta21    -0.2846  0.0617   -4.61   3.99e-06 y2 <--- y1
## gam22      0.0579  0.0161    3.61   3.09e-04 y2 <--- x2
## beta31    -0.2177  0.0971   -2.24   2.50e-02 y3 <--- y1
## beta32     0.8497  0.1121    7.58   3.52e-14 y3 <--- y2
## gam31      0.8607  0.3398    2.53   1.13e-02 y3 <--- x1
## V[x1]      1.0210  0.1101    9.27   1.80e-20 x1 <--> x1
## C[x1,x2]   7.1390  1.2556    5.69   1.30e-08 x2 <--> x1
## V[x2]    215.6620 23.2554    9.27   1.80e-20 x2 <--> x2
## V[y1]     12.9612  1.3976    9.27   1.80e-20 y1 <--> y1
## V[y2]      8.4882  0.9153    9.27   1.80e-20 y2 <--> y2
## V[y3]     19.4542  2.0978    9.27   1.80e-20 y3 <--> y3
```

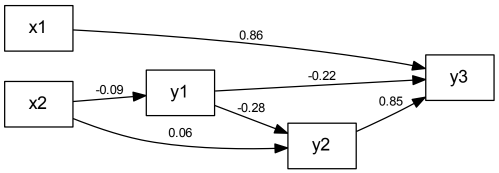
The fitted model is:

$$\begin{aligned} \hat{y}_1 &= -0.087 x_2 \\ \hat{y}_2 &= -0.285 y_1 + 0.058 x_2 \\ \hat{y}_3 &= -0.218 y_1 + 0.850 y_2 + 0.861 x_1 \end{aligned} \qquad \hat{\boldsymbol{\Psi}} = \begin{bmatrix} 12.96 & & \\ 0 & 8.49 & \\ 0 & 0 & 19.45 \end{bmatrix}$$

Union sentiment: Path diagrams

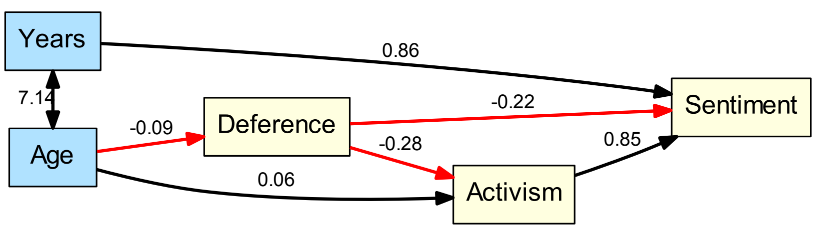
- Path diagrams for a `sem()` model can be produced using `pathDiagram(model)`
- This uses the `graphvis` program (`dot`), that must be installed first (`http://www.graphviz.org/`)
- The latest version (sem 3.1-6) uses the DiagrammeR package instead
- Edges can be labeled with parameter names, values, or both

```
pathDiagram(union.sem,
   edge.labels="values",
   file="union-sem1",
   min.rank=c("x1", "x2"))
```



Union sentiment: Path diagrams

- `dot` produces a text file describing the path diagram
- This can easily be (hand) edited to produce a nicer diagram
- Using color or linestyle for + vs. − edges facilitates interpretation



- The coefficients shown are unstandardized— on the scale of the variables
- Can also display standardized coefficients, easier to compare

Fundamental hypothesis of CFA & SEM

- The covariance matrix ($\boldsymbol{\Sigma}$) of the observed variables is a function of the parameters ($\boldsymbol{\theta}$) of the model

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

- That is, if
  - $\boldsymbol{\Sigma}$ is the population covariance matrix of the observed variables, and
  - $\boldsymbol{\theta}$ is a vector of all unique free parameters to be estimated,
  - then, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the model implied or predicted covariance matrix, expressed in terms of the parameters.
- If the model is correct, and we knew the values of the parameters, then

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

says that the population covariance matrix would be *exactly* reproduced by the model parameters

Fundamental hypothesis of CFA & SEM

Example: Consider the simple linear regression model,

$$y_i = \gamma x_i + \zeta_i$$

If this model is true, then the variance and covariance of (y, x) are

$$\begin{aligned} \mathrm{var}(y_i) &= \mathrm{var}(\gamma x_i + \zeta_i) \\ &= \gamma^2 \mathrm{var}(x_i) + \mathrm{var}(\zeta_i) \\ \mathrm{cov}(y_i, x_i) &= \gamma \mathrm{var}(x_i) \end{aligned}$$

The hypothesis $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ means that $\boldsymbol{\Sigma}$ can be expressed in terms of the model-implied parameters, γ (regression slope), $\mathrm{var}(\zeta)$ (error variance) and $\mathrm{var}(x)$:

$$\boldsymbol{\Sigma}\begin{pmatrix} y \\ x \end{pmatrix} = \begin{bmatrix} \mathrm{var}(y) & \\ \mathrm{cov}(y,x) & \mathrm{var}(y) \end{bmatrix} = \begin{bmatrix} \gamma^2 \mathrm{var}(x) + \mathrm{var}(\zeta) & \\ \gamma \mathrm{var}(x) & \mathrm{var}(x) \end{bmatrix} = \boldsymbol{\Sigma}\begin{pmatrix} \gamma \\ \mathrm{var}(\zeta) \\ \mathrm{var}(x) \end{pmatrix}$$

Fundamental hypothesis of CFA & SEM

This general hypothesis forms the basis for several important ideas in CFA and SEM

- Model ***identification***: How to know if you can find a unique solution?
- Model ***estimation***: How to fit a model to an observed covariance matrix ($\boldsymbol{S}$)?
- ***Goodness-of-fit*** statistics: How to assess the discrepancy between $\boldsymbol{S}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$?

Model identification

- A model is identified if it is possible to find a *unique* estimate for each parameter
- A non-identified model has an *infinite* number of solutions— not too useful
- Such models may be made identified by:
  - Setting some parameters to fixed constants (like $\beta_{12} = 0$ or $\mathrm{var}(\zeta_1) = 1$)
  - Constraining some parameters to be equal (like $\beta_{12} = \beta_{13}$)
- Identification can be stated as follows:
  - An unknown parameter θ is identified if it can be expressed as a function of one or more element of $\boldsymbol{\Sigma}$
  - The whole model is identified if all parameters in $\boldsymbol{\theta}$ are identified
- Complex models can often lead to identification problems, but there are a few simple helpul rules

Model identification: *t*-rule and degrees of freedom

The simplest rule, the *t*-rule says:

- The number of unknown parameters to be estimated (t) cannot exceed the number of non-redundant variances and covariances of the observed variables
- This is a necessary condition for identification, but it is not sufficient

For path analysis models, let $P = p + q$ be the total numbr of endogenous (y) and exogenous (x) variables in $\boldsymbol{\Sigma}$, and let t be the number of free parameters in $\boldsymbol{\theta}$. The t-rule is

$$P(P+1)/2 \geq t$$

The difference gives the number of degrees of freedom for the model:

$$df = P(P+1)/2 - t$$

- If $df < 0$, the model is under-identified (no unique solution)
- If $df = 0$, the model is just-identified (can't calculate goodness-of-fit)
- If $df > 0$, the model is over-identified (can calculate goodness-of-fit)

$\Longrightarrow$ Useful SEM models should be over-identified!!

Example: Union sentiment

For the Union sentiment model, the model parameters were:

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{bmatrix} \quad \boldsymbol{\Gamma} = \begin{bmatrix} 0 & \gamma_{12} \\ 0 & \gamma_{22} \\ \gamma_{31} & 0 \end{bmatrix}$$

and

$$\boldsymbol{\Phi} = \begin{bmatrix} \text{var}(x_1) & \\ \text{cov}(x_1, x_2) & \text{var}(x_2) \end{bmatrix} \quad \boldsymbol{\Psi} = \begin{bmatrix} \text{var}(\zeta_1) & & \\ 0 & \text{var}(\zeta_2) & \\ 0 & 0 & \text{var}(\zeta_2) \end{bmatrix}$$

Observed covariance matrix: $p = 3$ endogenous ys $+q = 2$ exogenous xs
$\Longrightarrow \boldsymbol{\Sigma}_{5\times5}$ has $5 \times 6/2 = 15$ variances and covariances.
12 free parameters in the model:

- 6 regression coefficients (3 non-zero in $\boldsymbol{B}$, 3 non-zero in $\boldsymbol{\Gamma}$)
- 3 variances/covariances in $\boldsymbol{\Phi}$
- 3 residual variances in diagonal of $\boldsymbol{\Psi}$

The model $df = 15 - 12 = 3 > 0$, so this model is over-identified

$\boldsymbol{B}$ rules: $\boldsymbol{B} = 0$

Another simple rule applies if no endogenous y variable affects any other endogenous variable, so $\boldsymbol{B} = 0$
For example:

$$\begin{aligned} y_1 &= \gamma_{11} x_1 + \gamma_{12} x_2 + \zeta_1 \\ y_2 &= \gamma_{21} x_1 + \gamma_{23} x_3 + \zeta_2 \\ y_3 &= \gamma_{31} x_1 + \gamma_{33} x_3 + \gamma_{34} x_4 + \zeta_3 \end{aligned}$$

- $\boldsymbol{B} = 0$ because no y appears on the RHS of an equation
- Such models are *always* identified
- This is a sufficient, but not a necessary condition
- Residuals ζ_i in such models need not be uncorrelated, i.e., $\boldsymbol{\Psi}$ can be non-diagonal ("seemingly unrelated regressions")

$\boldsymbol{B}$ rules: recursive rule

The recursive rule applies if

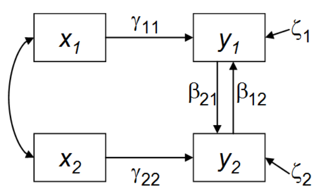
- the only free elements in $\boldsymbol{B}$ are on its lower (or upper) triangle, and
- $\boldsymbol{\Psi}$ is diagonal (no correlations amongst residuals)
- This basically means that there are no reciprocal relations among the ys and no feedback loops
- This also is a sufficient condition for model identification.

The union sentiment mode is recursive because $\boldsymbol{B}$ is lower-triangular and $\boldsymbol{\Psi}$ is diagonal

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{bmatrix} \quad \boldsymbol{\Psi} = \begin{bmatrix} \text{var}(\zeta_1) & & \\ 0 & \text{var}(\zeta_2) & \\ 0 & 0 & \text{var}(\zeta_2) \end{bmatrix}$$

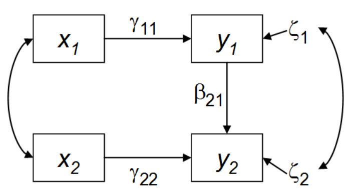
$\boldsymbol{B}$ rules: recursive rule

Non-recursive because $\boldsymbol{B}$ is not lower-triangular:



$$\boldsymbol{B} = \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix} \quad \boldsymbol{\Psi} = \begin{bmatrix} \text{var}(\zeta_1) & \\ 0 & \text{var}(\zeta_2) \end{bmatrix}$$

Non-recursive because $\boldsymbol{\Gamma}$ is not diagonal:



$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \quad \boldsymbol{\Psi} = \begin{bmatrix} \text{var}(\zeta_1) & \\ \text{cov}(\zeta_1, \zeta_2) & \text{var}(\zeta_2) \end{bmatrix}$$

Model estimation

How to fit the model to your data?

- In ordinary regression analysis, the method of least squares is used to find values of the parameters (regression slopes) that minimize the sum of squared residuals, $\sum(y_i - \widehat{y}_i)^2$.
  - This is fitting the model to the individual observations
- In constrast, SEM methods find parameter estimates that fit the model to the observed covariance matrix, $\boldsymbol{S}$.
- They are designed to minimize a function of the residual covariances, $\boldsymbol{S} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$
  - If the model is correct, then $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}$ and as $N \rightarrow \infty$, $\boldsymbol{S} = \boldsymbol{\Sigma}$.
  - There is a variety of estimation methods for SEM, but all attempt to choose the values of parameters in $\boldsymbol{\theta}$ to minimize a function $F(\bullet)$ of the difference between $\boldsymbol{S}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$

Model estimation: Maximum likelihood

- Maximum likelihood estimation is designed to maximize the likelihood ("probability") of obtaining the observed data ($\boldsymbol{\Sigma}$) over all choices of parameters ($\boldsymbol{\theta}$) in the model

$$\mathcal{L} = \Pr(\text{data} \mid \text{model}) = \Pr(\boldsymbol{S} \mid \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$$

- This assumes that the observed data are multivariate normally distributed
- ML estimation is equivalent to minimizing the following function:

$$F_{ML} = \log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| - \log|\boldsymbol{S}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) - p$$

- All SEM software obtains some initial estimates ("start values") and uses an iterative algorithm to minimize F_{ML}

Model estimation: Maximum likelihood

- ML estimates have optimal properties
  - Unbiased: $\mathcal{E}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$
  - Asymptotically consistent: as $N \rightarrow \infty$, $\widehat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}$
  - Maximally efficient: smallest standard errors
- As $N \rightarrow \infty$, parameter estimates $\widehat{\theta}_i$ are normally distributed, $\mathcal{N}(\widehat{\theta}_i, \text{var}(\theta_i))$, providing z (Wald) tests and confidence intervals

$$z = \frac{\widehat{\theta}}{s.e.(\widehat{\theta})} \qquad CI_{1-\alpha} : \widehat{\theta} \pm z_{1-\alpha/2}\, se(\widehat{\theta})$$

- As $N \rightarrow \infty$, the value $(N-1)F_{ML}$ has a χ^2 distribution with $df = P(P+1)/2 - t$ degrees of freedom, giving an overall test of model fit.

Model fit

- SEM provides R^2 values for each endogenous variable — the same as in separate regressions for each equation

```
##  R-square for Endogenous Variables
##    y1    y2    y3
## 0.113 0.230 0.390
```

- More importantly, it provides *overall measures* of fit for the entire model.
- The model for union sentiment fits very well, even though the R^2s are rather modest

```
##  Model Chisquare =  1.25   Df =  3 Pr(>Chisq) = 0.741
##  Goodness-of-fit index =  0.997
##  Adjusted goodness-of-fit index =  0.986
##  RMSEA index =  0   90% CI: (NA, 0.0904)
##  Bentler CFI =  1
##  AIC =  25.3
##  BIC =  -14.2
```

- A just-identified model will always fit perfectly— but that doesn't mean it is a good model: there might be unnecessary or trivial parameters.
- An over-identified model that fits badly might have too many fixed or constrained parameters

Model fit: χ^2 test

- The fitting function $F(\boldsymbol{S}, \widehat{\Sigma})$ used to minimize the discrepancy between $\boldsymbol{S}$ and the model estimate $\widehat{\Sigma} = \boldsymbol{\Sigma}(\widehat{\theta})$ gives a chi-square test of model fit
- If the model is correct, then the minimized value, F_{min}, has an asympotic chi-square distribution,

$$X^2 = (N-1)F_{min} \quad \sim \quad \chi^2_{df}$$

  with $df = P(P+1)/2 - t$ degrees of freedom
- This gives a test of the hypothesis that the model fits the data

$$H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

- a large (significant) X^2 indicates that the model does not fit the data.

Model fit: χ^2 test— problems

- The test statistic, $X^2 = (N-1)F_{min}$ is a function of sample size.
- With large N, trivial discrepancies will give a significant chi-square
- Worse, it tests an unrealistic hypothesis that the model fits perfectly
  - the specified model is exactly correct in all details
  - any lack-of-fit is due only to sampling error
  - it relies on asymptotic theory ($X^2 \sim \chi^2$ as $N \to \infty$) and an assumption of multivariate normality
- Another problem is parsimony— a model with additional free parameters will always fit better, but smaller models are simpler to interpret
- If you fit several nested models, $M_1 \supset M_2 \supset M_3 \ldots$, chi-square tests for the difference between models are less affected by these problems

$$\Delta X^2 = X^2(M_1) - X^2(M_2) \sim \chi^2 \text{ with } df = df_1 - df_2$$

Model fit: RMSEA

The measure of root mean square error of approximation (RMSEA) attempts to solve these problems (Browne & Cudeck, 1993)

$$\text{RMSEA} = \sqrt{\frac{X^2 - df}{(N-1)df}}$$

- Relatively insensitive to sample size
- Parsimony adjusted— denominator adjusts for greater *df*
- Common labels for RMSEA values:

RMSEA	interpretation
0	perfect fit
$\leq .05$	close fit
$.05 - .08$	acceptable fit
$.08 - .10$	mediocre fit
$> .10$	poor fit

Model fit: RMSEA

In addition, the RMSEA statistic has known sampling distribution properties (McCallum et al., 1996). This means that:

- You can calculate confidence intervals for RMSEA
- It allows to test a null hypothesis of "close fit" or "poor fit", rather than "perfect fit"

$$H_0 : RMSEA < 0.05$$
$$H_0 : RMSEA > 0.10$$

- It allows for power analysis to find the sample size (N) required to reject a hypothesis of "close fit" (RMSEA ≤ 0.05)

Incremental fit indices

- Creating new indices of goodness-of-fit for CFA/SEM models was a "growth industry" for many years— there are many possibilities
- Incremental fit indices compare the existing model with a null or baseline model
  - The null model, M_0 assumes all variables are uncorrelated— the worst possible model.
  - Incremental fit indices compare the X^2_M for model M with X^2_0 for the null model
  - All of these are designed to range from 0 to 1, with larger values (e.g., > 0.95) indicating better fit.
  - The generic idea is to calculate an R^2-like statistic, of the form

$$\frac{f(\text{null model}) - f(\text{my model})}{f(\text{null model}) - f(\text{best model})}$$

for some function $f(\bullet)$ of X^2 and df, and where the "best" model fits perfectly.

Incremental fit indices

Parsimony-adjusted indices also adjust for model *df*

- Bentler's comparative fit index (CFI) is often widely used

$$CFI = 1 - \frac{X^2_M - df_M}{X^2_0 - df_0}$$

- Tucker-Lewis Index (TLI), also called "non-normed fit index" (NNFI) are also popularly reported

$$TLI \equiv NNFI = \frac{X^2_0/df_0 - X^2_M/df_M}{X^2_0/df_0 - 1}$$

Information criteria: AIC, BIC

- Other widely used criteria, particularly when you have fit a collection of potential models are the "information criteria", ***AIC*** and ***BIC***
- Unlike the likelihood ratio tests these can be used to compare non-nested models
- Each of these uses a penalty for model complexity; BIC expresses a greater preference for simpler models as the sample size increases.

$$\begin{aligned} AIC &= X^2 - 2df \\ BIC &= X^2 - \log(N)df \end{aligned}$$

- Smaller is better

Model modification

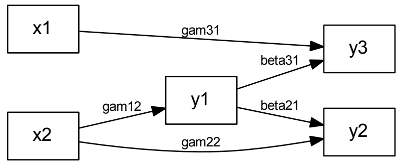
What to do when your model fits badly?

- First, note that a model might fit badly due to data problems:
  - outliers, missing data problems
  - non-normality (highly skewed, excessive kurtosis)
  - non-linearity, omitted interactions, ...
- Otherwise, bad model fit usually indicates that some important paths have been omitted, so some variances or covariances in $\boldsymbol{S}$ are poorly reproduced by the model
  - Some regression effects among $(\boldsymbol{x}, \boldsymbol{y})$ omitted (fixed to 0)?
  - Covariances among exogenous variables omitted? (all should be included)
  - Covariances among residuals might need to be included as free parameters
- Actions:
  - Examine residuals, $\boldsymbol{S} - \Sigma(\widehat{\theta})$ to see which variances/covariances are badly fit
  - Modification indices provide a way to test the impact of freeing each fixed parameter

Example: Union sentiment

To illusrate, consider what would have happened if we omitted the important path of y_3 (sentiment) on y_2 (activism) in the Union sentiment example

```
mod.bad <- specifyEquations(covs="x1, x2", text='
y1 = gam12*x2
y2 = beta21*y1 + gam22*x2
y3 = beta31*y1 +             gam31*x1
')
```



Fit the model:

```
union.sem.bad <- sem(mod.bad, union, N=173)
union.sem.bad

##
##  Model Chisquare =  50.235   Df =  4
##
##       gam12     beta21      gam22     beta31      gam31      V[x1]
##   -0.087438  -0.284563   0.057938  -0.509024   1.286631   1.021000
##     C[x1,x2]      V[x2]      V[y1]      V[y2]      V[y3]
##    7.139000 215.662000  12.961186   8.488216  25.863934
##
##  Iterations =  0
```


As expected, this model fits very badly

```
summary(union.sem.bad, fit.indices=c("RMSEA", "NNFI", "CFI"))

##
##  Model Chisquare =  50.235   Df =  4 Pr(>Chisq) = 3.2251e-10
##  RMSEA index =  0.25923   90% CI: (0.19808, 0.32556)
##  Tucker-Lewis NNFI =  0.38328
##  Bentler CFI =  0.75331
##
##  Normalized Residuals
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -0.159   0.000   0.000   0.594   0.330   5.247
##
##  R-square for Endogenous Variables
##      y1     y2     y3
## 0.1129 0.2295 0.1957
##
...
```


Normalized residuals show the differences $\boldsymbol{S} - \Sigma(\widehat{\theta})$ as approximate z-scores, so values outside of ± 2 can be considered significantly large.

```
round(normalizedResiduals(union.sem.bad), 3)

##       y1    y2     y3     x1    x2
## y1 0.000 0.000  0.103  0.477 0.000
## y2 0.000 0.000  5.246  0.330 0.000
## y3 0.103 5.246 -0.054 -0.159 1.454
## x1 0.477 0.330 -0.159  0.000 0.000
## x2 0.000 0.000  1.454  0.000 0.000
```

- This points to the one very large residual for the `y2 -> y3` (or `y3 -> y2`) path
- In this example Union sentiment ($y3$) is the main outcome, so it would make sense here to free the `y2 -> y3` path

Modification indices

- Modification indices provide test statistics for fixed parameters
- The statistics estimate the decrease in X^2 if each fixed parameter was allowed to be freely estimated
- These are $\chi^2(1)$ values, so values > 4 can be considered “significantly” large.

```
modIndices(union.sem.bad)

##
##  5 largest modification indices, A matrix (regression coefficients):
## y3<-y2 y2<-y3 x2<-y3 y3<-x2 y1<-y3
## 42.071 38.217  4.240  3.947  3.763
##
##   5 largest modification indices, P matrix (variances/covariances):
## y3<->y2 y3<->y1 x2<->y3 x1<->y3 x1<->y2
## 38.3362  3.9468  3.9468  3.9468  0.4114
```

Once again, we see large values associated with the `y2 -> y3` path

Modification indices: Caveats

- Using modification indices to improve model fit is called ***specification search***
- This is often deprecated, unless there are good substantive reasons for introducing new free parameters
  - New paths or covariances in the model should make sense theoretically
  - Large modification indices could just reflect sample-specific effects

Summary I

- ***Structural equation models*** are an historical development of EFA and CFA methods and path analysis
  - EFA and CFA attempt to explain correlations among observed variables in terms of latent variables ("factors")
  - EFA used factor rotation to obtain an interpretable solution
  - CFA imposes restrictions on a solution, and allows specific hypothesis tests
  - Higher-order CFA further generalized CFA to the ACOVS model
  - Meanwhile, path analysis developed methods for analyzing systems of equations together
  - The result, was SEM, in the form of the LISREL model

Summary II

- ***Path diagrams*** provide a convenient way to portray or visualize a SEM
  - Direct translation from/to a system of linear equations
  - Some software (AMOS graphics) allows construction of the model via a path diagram
  - Most SEM software provides for output of models and results as path diagrams
- ***Path analysis models*** provide a basic introduction to SEM
  - No latent variables— only observed ("manifest") ones
  - Does not allow for errors of measurement in observed variables
  - exogenous variables (xs)— only predictors in the linear equations
  - endogenous variables (ys)— a dependent variable in one or more equations
  - Error terms reflect errors-in-equations— unmodeled predictors, wrong functional form, etc.
- An important question in SEM models is model identification— can the parameters be uniquely estimated?
- Another important question is how to evaluate model fit?