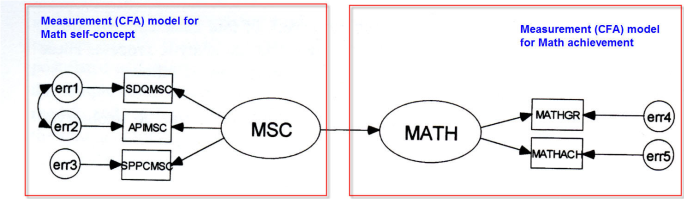
CFA & SEM

Lecture 3: Structural Equation Models with Latent Variables and Other Topics

Michael Friendly

SCS Short Course

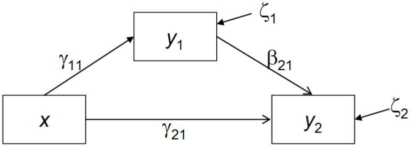


Overview: The full SEM

Path analysis models:

- We started here, with models for observed variables only
- With exogenous variables ($\boldsymbol{x}$) and endogenous variables ($\boldsymbol{y}$), these have the form
$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{y} + \Gamma\boldsymbol{x} + \zeta$$
- These models do not allow for measurement error in the $\boldsymbol{x}$ or $\boldsymbol{y}$ variables
- The only errors are the disturbance terms ζ ("errors in equations"), allowing for unmeasured or omitted predictors
- e.g., the simple mediation model:

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

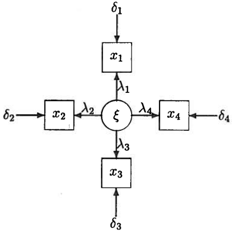


Overview: The full SEM

Confirmatory factor analysis (CFA) models:

- We next considered CFA measurement models, allowing for observed indicators to be expressed as regressions on unobserved, latent variables.
- For a set of observed variables ($\boldsymbol{x}$), there can be one or more factors, ξ and the error terms δ can reflect both specific variance and unreliability

$$\boldsymbol{x} = \boldsymbol{\Lambda}_x \xi + \boldsymbol{\delta}$$



the one-factor congeneric model:

$$\begin{aligned} x_i &= \lambda_i \xi + \delta_i \\ \boldsymbol{\Sigma} &= \boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathsf{T}} + \boldsymbol{\Theta} \end{aligned}$$

The complete SEM model: LISREL form

- Now imagine that we have q observed exogenous variables, $\boldsymbol{x}$, and p endogenous variables, $\boldsymbol{y}$
- We can allow for errors of measurement with measurement models for each:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \\ \boldsymbol{y} &= \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \end{aligned}$$

  - Measurement error is accounted for in the (co)variances of $\boldsymbol{\delta}$ ($\boldsymbol{\Theta}_\delta$) and $\boldsymbol{\epsilon}$ ($\boldsymbol{\Theta}_\epsilon$)
  - Errors of measurement can be allowed to be correlated— $\boldsymbol{\Theta}_\delta$ and $\boldsymbol{\Theta}_\epsilon$ need not be diagonal
- These are connected by the structural model,

$$\eta = \boldsymbol{B}\eta + \boldsymbol{\Gamma}\xi + \zeta$$

  - The coefficients in $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ represent the linear regressions for the true, latent constructs
  - These are not biased by measurement error
  - ζ now reflects the pure errors in equations.

The Eight LISREL matrices

The main matrices of regression coefficients in this general model are:

- $\underset{(q \times n)}{\boldsymbol{\Lambda_x}}$ ("lambda-x"): factor loadings of the observed exogenous variables $\boldsymbol{x}$ on their latent variables $\boldsymbol{\xi}$
- $\underset{(p \times m)}{\boldsymbol{\Lambda_y}}$ ("lambda-y"): factor loadings of the observed endogenous variables $\boldsymbol{y}$ on their latent variables $\boldsymbol{\eta}$
- $\underset{(m \times n)}{\boldsymbol{B}}$ ("beta"): Coefficients for the regressions of $\boldsymbol{\eta}$ on $\boldsymbol{\eta}$. ($\boldsymbol{B}$ must have zeros on the diagonal, and is usually lower (upper) triangular.)
- $\underset{(m \times n)}{\boldsymbol{\Gamma}}$ ("gamma"): Coefficients for the regressions of $\boldsymbol{\eta}$ on $\boldsymbol{\xi}$

The Eight LISREL matrices

In addition, there are four variance-covariance matrices:

- $\underset{(q\times q)}{\mathbf{\Theta}_\delta}$ ("theta-delta"): residual variances (and possibly covariances) for exogenous observed variables
- $\underset{(q\times q)}{\mathbf{\Theta}_\epsilon}$ ("theta-epsilon"): residual variances (and possibly covariances) for endogenous observed variables
- $\underset{(n\times n)}{\mathbf{\Phi}}$ ("phi"): covariance matrix for exogenous latent variables, ξ
- $\underset{(m\times m)}{\mathbf{\Psi}}$ ("psi"): covariance matrix of residual terms, ζ, from the structural regression model

The complete SEM model: Σ

In this form, the relationship between the population covariance matrix Σ and the parameters is

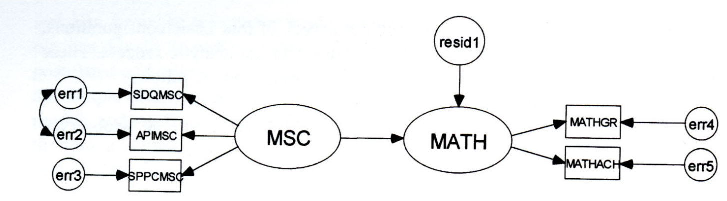
$$\underset{(q+p \times q+p)}{\Sigma} = \Sigma \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{bmatrix} \underset{(q \times q)}{\Sigma_{\boldsymbol{xx}}} & \underset{(q \times p)}{\Sigma_{\boldsymbol{xy}}} \\ \underset{(p \times q)}{\Sigma_{\boldsymbol{yx}}} & \underset{(p \times p)}{\Sigma_{\boldsymbol{py}}} \end{bmatrix}$$

where

$$\begin{aligned} \Sigma_{\boldsymbol{xx}} &= \Lambda_x \Phi \Lambda_x^{\mathsf{T}} + \Theta_\delta \\ \Sigma_{\boldsymbol{xy}} &= \Lambda_x \Phi \Gamma^{\mathsf{T}} (\boldsymbol{I} - \boldsymbol{B})^{-1\,T} \Lambda_y^{\mathsf{T}} \\ \Sigma_{\boldsymbol{yy}} &= \Lambda_y (\boldsymbol{I} - \boldsymbol{B})^{-1} [\Gamma \Phi \Gamma^{\mathsf{T}} + \Psi] (\boldsymbol{I} - \boldsymbol{B})^{-1\,T} \Lambda_y^{\mathsf{T}} + \Theta_\epsilon \end{aligned}$$

Example: General SEM Model

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

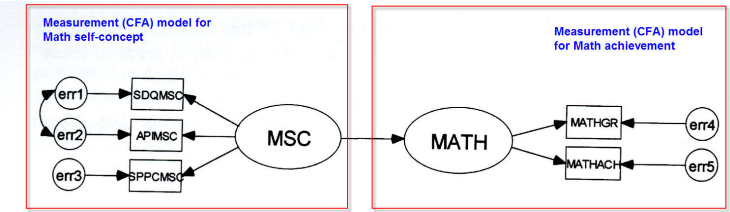


This model has:

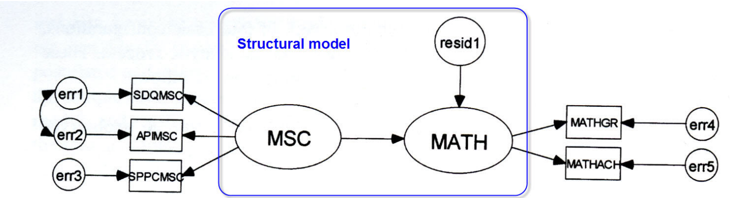
- 3 observed indicators in a measurement model for MSC (***x***)
- 2 observed indicators in a model for MATH achievement (***y***)
- A structural equation predicting MATH achievement from MSC
- Correlated errors for two MSC variables

Example: General SEM Model

Measurement sub-models for $\boldsymbol{x}$ and $\boldsymbol{y}$



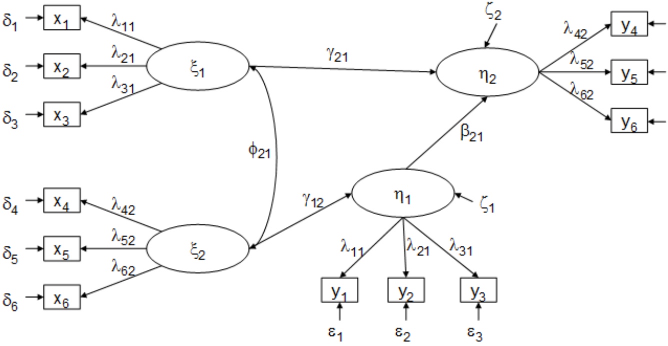
Structural model, relating ξ to η



Example: General SEM Model

Here is another example, with 6 x variables and 6 y variables

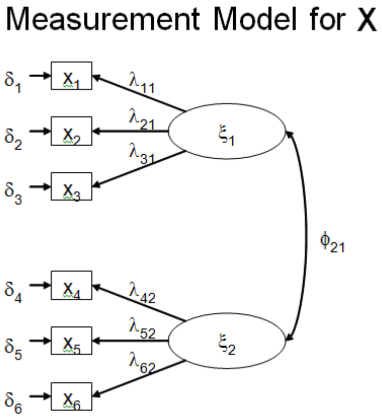
- What are the measurement models for x and y?
- What is the structural part of the model?

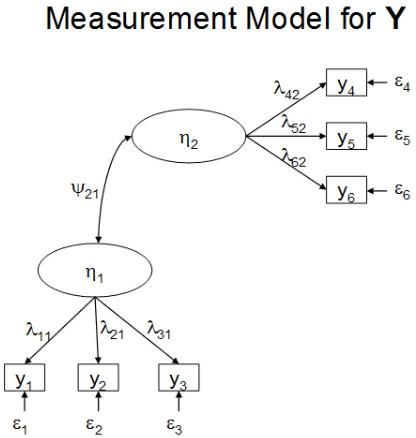


Example: General SEM Model

Measurement models for *x* and *y*:

- The *x* variables are assumed to measure correlated factors, in two congeneric sets
- Same for the *y* variables



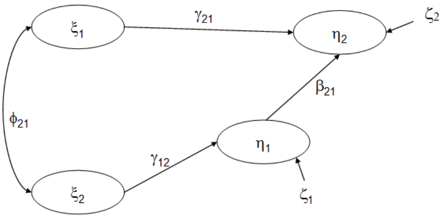


Example: General SEM Model

Structural model:

- η_2 is predicted only by ξ_1
- η_1 is predicted only by ξ_2 and η_2

Structural Model



$$
\begin{aligned}
\eta_1 &= \beta_{21}\eta_2 + \gamma_{12}\xi_2 + \zeta_1 \\
\eta_2 &= \gamma_{21}\xi_1 + \zeta_2
\end{aligned}
$$

Example: Health care utilization

A study[1] was carried out to address these questions:

- Do age, stress and poor sense of self predict perceived ill health and health care utilization?
- Does perceived ill health directly predict health care utilization?
- Does perceived ill health serve as in intervening variable between age, life stress, poor sense of self, and health care utilization?

Raw data is available for a sample of N=445. We can (and should!) also examine the following:

- Is there evidence of serious departure from univariate and multivariate normality?
- More important: Are the relationships among the variables at least approximately linear?
- Are there possible multivariate outliers that might affect the results?

The MVN package provides some useful tools for these questions.

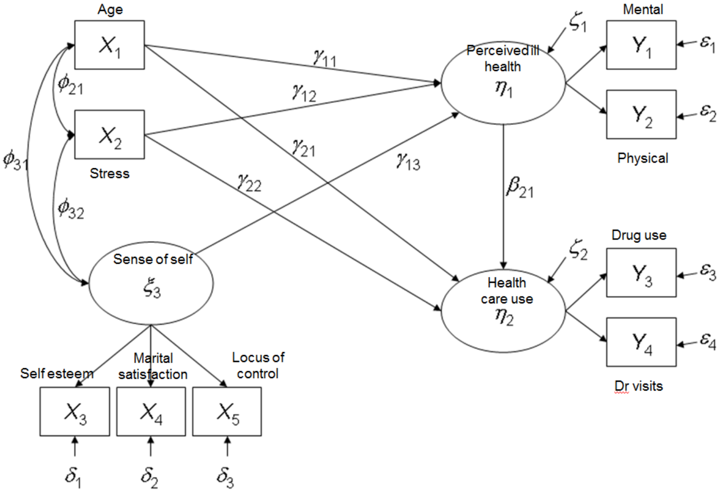
[1] This example taken from Flora (2014)

Variables

The variables are:

- Age (x_1)
- Stress (x_2)
- Sense of self: latent variable measured by three indicators
  - Self-esteem (x_3);
  - Marital satisfaction (x_4)
  - Locus of conrol (x_5)
- Perceived ill health: latent variable measured by
  - number of mental health problems (y_1)
  - number of physical health problems (y_2)
- Health care utilization: latent variable measured by
  - frequency of prescription drug use (y_3)
  - number of visits to health professionals (y_4)

The path diagram for the proposed model:



Data screening

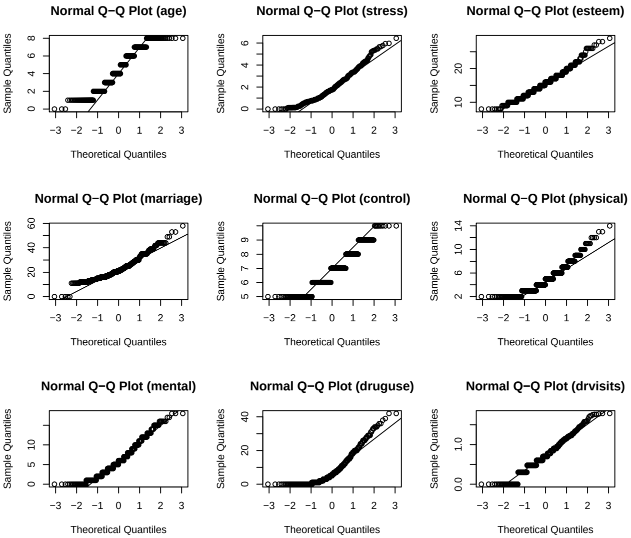
At a minimum,

- Make univariate QQ plots to assess univariate normality
- Make a χ^2 QQ plot to assess multivariate normality
- You can also use uni- (Shapiro-Wilks) and multivariate (Mardia) tests

```r
library(MVN)
#healthdat <- read.table("R/healthutil.txt")
MVN::mvn(healthdat, univariateTest = "SW")$univariateNormality

##            Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk     age       0.9376  <0.001      NO
## 2 Shapiro-Wilk  stress       0.9522  <0.001      NO
## 3 Shapiro-Wilk  esteem       0.9761  <0.001      NO
## 4 Shapiro-Wilk marriage      0.9478  <0.001      NO
## 5 Shapiro-Wilk  control      0.9120  <0.001      NO
## 6 Shapiro-Wilk physical      0.9144  <0.001      NO
## 7 Shapiro-Wilk  mental       0.9531  <0.001      NO
## 8 Shapiro-Wilk  druguse      0.8512  <0.001      NO
## 9 Shapiro-Wilk drvisits      0.9750  <0.001      NO
```

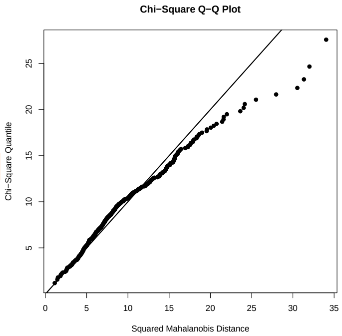
Univariate normal QQ plots:

```
uniPlot(healthdat)
```



Multivariate tests:

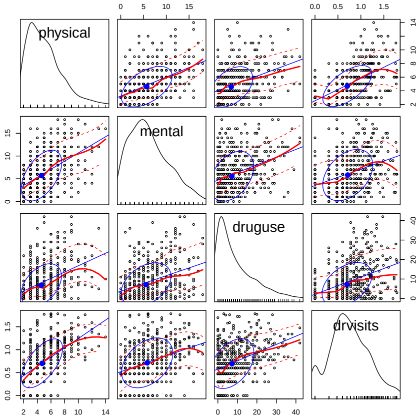
```r
mardiaTest(healthdat, qqplot=TRUE)
```



- Univariate and multivariate tests show strong evidence of non-normality
- The χ^2 QQ plot shows that there may be some multivariate outliers
- Possible actions:
  - Transform variables
  - Use robust SEM methods for tests
  - Use bootstrap methods for tests

Assessing linearity

```r
library(car)
scatterplotMatrix(healthdat[,6:9], cex=0.8,
    ellipse=TRUE, levels=0.68, col=c("blue", "red", "black"))
```



- Non-linear relationships are a more serious problem for SEM
- A simple way to assess this is a scatterplot matrix, showing non-parametric smooth curves
- These plots show some slight non-linearities, but perhaps not too serious
- The diagonal panels show generally skewed distributions

What's wrong with multivariate regression?

Rather than SEM, you might consider fitting two multivariate linear models (MLM):

$$
\begin{aligned}
(y_3, y_4) &\sim x_1 + x_2 + x_3 + x_4 + y_1 + y_2 \\
(y_1, y_2, y_3, y_4) &\sim x_1 + x_2 + x_3 + x_4
\end{aligned}
$$

For example,

```
health.mlm1 <- lm(cbind(druguse, drvisits) ~
  age + stress + esteem + marriage + control + physical + mental,
  data=healthdat)
health.mlm2 <- lm(cbind(physical, mental, druguse, drvisits) ~
  age + stress + esteem + marriage + control, data=healthdat)
```

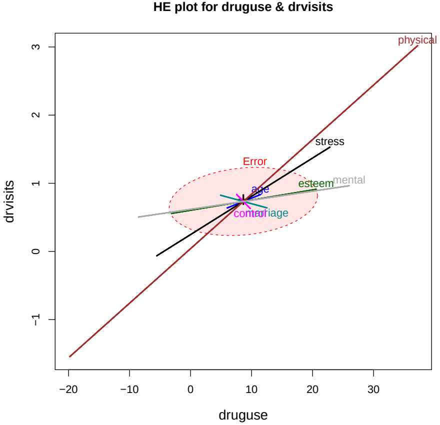
What's wrong with multivariate regression?

```
library(car)
Anova(health.mlm1)

##
## Type II MANOVA Tests: Pillai test statistic
##          Df test stat approx F num Df den Df  Pr(>F)
## age       1    0.0011      0.2      2    436   0.786
## stress    1    0.0445     10.2      2    436 4.9e-05 ***
## esteem    1    0.0137      3.0      2    436   0.050 *
## marriage  1    0.0022      0.5      2    436   0.612
## control   1    0.0008      0.2      2    436   0.837
## physical  1    0.2446     70.6      2    436 < 2e-16 ***
## mental    1    0.0280      6.3      2    436   0.002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problems with this approach:

- Doesn't provide a single, overall model
- Doesn't allow for errors of measurement in ***x*** or ***y***
- All predictors in each model are included for all responses

Nevertheless, the MLM provides some useful graphical displays not available for SEMs



- Hypothesis-Error (HE) plots show relations of *x*s to *y*s
- Significant predictors project outside the Error ellipse
- Directions show their relations to the *y*s

Fitting the SEM model

Using lavaan, the model can be specified as follows:

```
library(lavaan)
health.mod1 <- '
      # latent variables
Self =~ esteem + marriage + control
Ill  =~ physical + mental
Util =~ druguse + drvisits
    # structural regressions
Ill  ~ age + stress + Self
Util ~ age + stress + Ill
    # covariances (Phi)
Self ~~ age + stress
age ~~ stress
'
```

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

```
health.sem1 <- lavaan::sem(health.mod1, data=healthdat,
                   estimator="ML", fixed.x=F)
```

Assessing model fit

Quick look at fit indices:

```
fitMeasures(health.sem1, c("chisq", "df", "pvalue", "cfi", "rmsea"))

##   chisq      df  pvalue     cfi   rmsea
## 111.521  20.000   0.000   0.878   0.101
```

This model doesn't fit very well. Examine modification indices to see why not:

```
modindices(health.sem1, minimum.value=20)

##          lhs op      rhs     mi    epc sepc.lv sepc.all sepc.nox
## 30      Self =~   mental 38.710  0.581   1.482    0.353    0.353
## 66 physical ~~ drvisits 35.327  0.278   0.278    0.666    0.666
## 70   mental ~~ drvisits 21.326 -0.341  -0.341   -0.360   -0.360
```

- Be cautious of revising a model just based on modification indices
- Any changes should make sense substantively— it makes no sense to add `mental` as an indicator of `Self`
- The largest covariance MI is for the error covariance between `physical` (y_2) and visits to health professionals, `drvisits` (y_4)

Revised model

Add a covariance betweeen `physical` and `drvisits`:

```
health.mod2 <- paste(health.mod1,
'physical ~~ drvisits
')
```

Fit the new model:

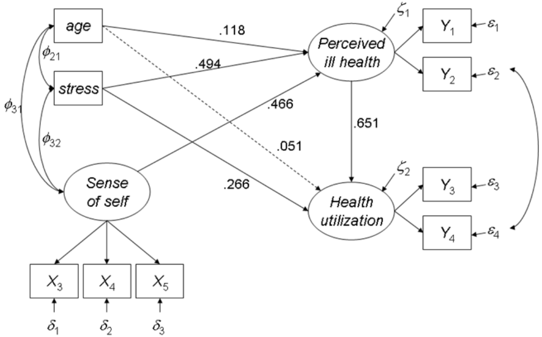
```
health.sem2 <- lavaan::sem(health.mod2, data=healthdat, estimator="ML", fi
fitMeasures(health.sem2, c("chisq", "df", "pvalue", "cfi", "rmsea"))

##  chisq     df pvalue    cfi  rmsea
## 68.568 19.000  0.000  0.934  0.077
```

Test whether this is a significant improvement:

```
anova(health.sem1, health.sem2)

## Chi Square Difference Test
##
##             Df   AIC   BIC Chisq Chisq diff Df diff Pr(>Chisq)
## health.sem2 19 18249 18356  68.6
## health.sem1 20 18290 18393 111.5         43       1    5.6e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model interpretation

Examine the standardized estimates in the path diagram (only $\widehat{\gamma}$ and $\widehat{\beta}$ shown here)



- Age has only a tiny effect on Util — remove/ignore it
- Stress and Self strongly predict (perceived) Ill health
- Stress and Ill strongly predict health utilization

Two-wave longitudinal models

- SEM is also very useful when the same variables are measured on two (or more) occasions in a longitudinal design
- In general, longitudinal studies seek to:
  - assess changes in outcomes over time
  - relate these to background variables or intervening treatment interventions
- Assume we have two observed measures, used on two occasions: y_1 and y_2 measure the latent variable η_1 on occasion 1; y_3 and y_4 measure the latent variable η_2 on occasion 2.
- The measurement model (with reference variables) is

$$\begin{aligned} y_1 &= 1\eta_1 + \epsilon_1 \\ y_2 &= \lambda_1\eta_1 + \epsilon_2 \\ y_3 &= 1\eta_2 + \epsilon_3 \\ y_4 &= \lambda_2\eta_2 + \epsilon_4 \end{aligned}$$

Two-wave longitudinal models

- Main interest is in the stability of η over time. This gives the structural equation
$$\eta_2 = \beta\eta_1 + \zeta$$
- If the same latent construct is measured on both equations, we should have $\widehat{\beta} \approx 1$ and $\mathrm{var}(\zeta)$ small
- One wrinkle is that the errors of measurement, ϵ_i are likely to be correlated for the same measure given on multiple occasions.
- This can be allowed for by allowing Θ_ϵ to be non-diagonal, e.g., $\theta_{31} \neq 0, \theta_{42} \neq 0$
- Let Ω be the covariance matrix of (η_1, η_2). Then, the correlation between η_1 and η_2 is

$$\rho = Corr(\eta_1, \eta_2) = \left[\frac{\omega_{21}}{\omega_{11}\omega_{22}}\right]^{1/2} = \left[\frac{\sigma_{32}\sigma_{41}}{\sigma_{21}\sigma_{43}}\right]^{1/2}$$

Example: Stability of Alienation

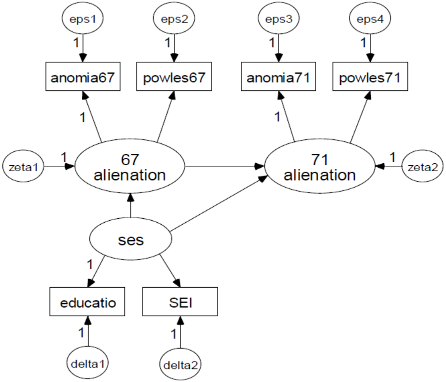
Data from Wheaton et. al (1997)

- Attitude measurements of N=932 people in rural Illinois were collected in 1967 and 1971
- Scales of anomia and powerless were both taken as indicators of a latent variable,alienation
- Background variables are
  - Respondent's education (of schooling)
  - Duncan's Socioeconomic status index (SEI)
  - These are taken as indicators of a latent SES variable

Model A

Model A from Jöreskog & Sörbom (1984)

- Endogenous latent variables: Alienation67 (η_1) and Alienation71 (η_2);
- SES (exogenous latent) influences both Alienation67 and Alienation71
- NB: $\epsilon_1, \epsilon_2 \ldots$ represent not only errors of measurement, but also specificity



Data

The data set is given as the 6×6 covariance matrix:

```
library(sem)
S.wheaton <- readMoments(
    names=c('Anomia67','Powerless67','Anomia71',
            'Powerless71','Education','SEI'),
    text="
   11.834
    6.947    9.364
    6.819    5.091   12.532
    4.783    5.028    7.495    9.986
   -3.839   -3.889   -3.841   -3.625   9.610
  -21.899  -18.831  -21.748  -18.775  35.522  450.288
")
```

There are $6 \times 7/2 = 21$ sample moments (6 variances and 15 covariances)

Specifying the model

The model in the path diagram has the following form for the measurement models:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_1 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \end{bmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ \lambda_3 \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$$

The structural model for η_1 and η_2 is:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_1 & 0 \end{bmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \xi + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$

- This model has 15 parameters (6 regression weights, 9 variances)
- This leaves $21 - 15 = 6$ df

Fitting the model

Translating this into linear equations and variances, we have:

```
wh.model.A <- specifyEquations(text="
Anomia67       = 1*Alienation67
Powerless67    = lamy1*Alienation67
Anomia71       = 1*Alienation71
Powerless71    = lamy2*Alienation71
Education      = 1*SES
SEI            = lamx*SES
Alienation67   = gam1*SES
Alienation71   = gam2*SES + beta*Alienation67
V(Anomia67)    = the1
V(Anomia71)    = the2
V(Powerless67) = the3
V(Powerless71) = the4
V(SES)         = phi
")
```

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

```
sem.wh.A <- sem(wh.model.A, S.wheaton, 932)
```

Assessing model fit

Model A does not fit particularly well by conventional criteria:

```
summary(sem.wh.A, fit.indices = c("RMSEA", "NNFI", "CFI"))

##
##  Model Chisquare =  71.47   Df =  6 Pr(>Chisq) = 2.0417e-13
##  RMSEA index =  0.10826   90% CI: (0.086585, 0.13145)
##  Tucker-Lewis NNFI =  0.92266
##  Bentler CFI =  0.96907
##
...
```

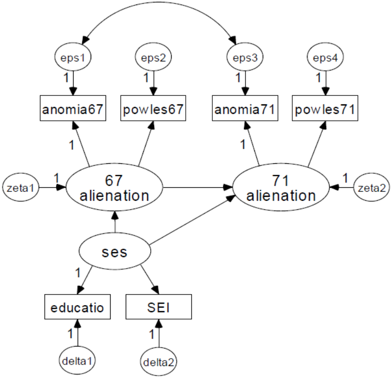
Examine modification indices (***A***: regression coef.; ***P***: covariances)

```
print(modIndices(sem.wh.A), n.largest=3)

##
##  3 largest modification indices, A matrix (regression coefficients):
##    Anomia71<-Anomia67    Anomia67<-Anomia71 Powerless71<-Anomia67
##                58.724                51.912                46.156
##
##   3 largest modification indices, P matrix (variances/covariances):
##    Anomia71<->Anomia67 Powerless71<->Anomia67 Anomia71<->Powerless67
##                 63.706                 49.829                 49.752
```

Model B

The largest MI is the covariance for `Anomia71<->Anomia67`– set it free

- Add covariance between ϵ_1 and ϵ_2
- Could also add a covariance between ϵ_2 and ϵ_4 (Model C)
- There are often equivalent models that improve fit equally, but in different ways.



Model B

The `update()` function makes it easy to add (or remove) parameters. "`<->`" specifies a covariance

```
wh.model.B <- update(wh.model.A, text="
 add, Anomia67 <-> Anomia71,         the13"
 )
```

Fit model B:

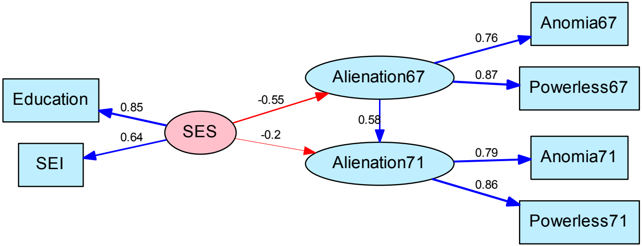
```
sem.wh.B <- sem(wh.model.B, S.wheaton, 932)
summary(sem.wh.B, fit.indices = c("RMSEA", "NNFI", "CFI"))

##
##  Model Chisquare =  6.3307   Df =  5 Pr(>Chisq) = 0.27536
##  RMSEA index =  0.016908   90% CI: (NA, 0.050905)
##  Tucker-Lewis NNFI =  0.99811
##  Bentler CFI =  0.99937
...
```

This fits very well!

Model interpretation

Path diagram with (standardized) coefficient estimates:

```
pathDiagram(sem.wh.B,
   same.rank=c("Alienation67, Alienation71"),
   min.rank=c("Education", "SEI"),
   edge.labels = "values", edge.colors = c("blue", "red"),
   node.colors = c("pink", "lightblue1"),
   edge.weight="proportional", standardize=TRUE)
```



Other models

Given a model that fits reasonably well, it is often useful to ask if we can make the model simpler in some ways

- Are there non-significant paths or latent variables that could be eliminated from the model?
- Are there free parameters that could be constrained to be equal?
  - e.g., perhaps we could set $\lambda_1 = \lambda_2$?

```
coef(sem.wh.B)[1:2]
```

```
##    lamy1   lamy2
## 1.02653 0.97092
```

  - could test whether the scales are τ-equivalent, i.e., $\lambda_1 = 1$ and/or $\lambda_2 = 1$
  - could test whether the variances of errors are equal $(\mathrm{var}(\epsilon_1) = \mathrm{var}(\epsilon_3)$; $\mathrm{var}(\epsilon_2) = \mathrm{var}(\epsilon_4))$
- What happens if we remove the effects of SES?

Power and Sample Size for CFA and SEM

Bad news Determining the required sample size, or the power of a hypothesis test are far more complex in CFA and SEM than in other statistical applications (correlation, ANOVA, etc.)

- SEM involves both measurement and structural sub-models
- There are often many parameters involved
- Hard to tell where lack of fit comes from
- Logic of hypothesis tests is reversed from usual NHST

Good news There are a few things you *can* do to choose a sample size intelligently.

- Rules of thumb for EFA models
- Using desired standard errors
- Overall approach based on RMSEA
- Some useful methods for individual parameters

Power and Sample Size for CFA and SEM

Rules of thumb for EFA

For EFA, there is little statistical basis for determining the appropriate sample size, and little basis for determining power (but the overall approach of CFA can be used).

Some traditional "rules of thumb" for EFA:

- ***The more the better!***
  - Reliability and replicability increase directly with $\sqrt{N}$.
  - More reliable factors can be extracted with larger sample sizes.
- Absolute minimum– $N = 5p$, but you should have $N > 100$ for any non-trivial factor analysis. Minimum applies only when communalities are high and p/k is high.
- Most EFA and CFA studies use $N > 200$, some as high as 500-600.
- Safer to use at least $N > 10p$.
- The lower the reliabilities, the larger N should be.

Using desired standard errors

- An alternative approach for EFA/CFA/SEM considers the standard errors of correlations, in relation to sample size.
- This usually provides more informed guidance than the rules of thumb. It can be shown that,

$$\sigma(\rho) = \frac{1-\rho^2}{\sqrt{N}} + \mathcal{O}(N^{-1})$$

so, we could determine the sample size to make the standard error of a "typical" correlation smaller than some given value.

$$\sqrt{N} > \frac{1-\rho^2}{\sigma(\rho)}$$

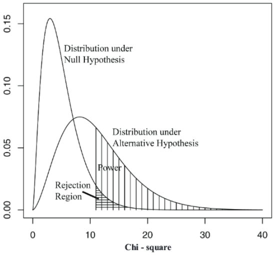
Using desired standard errors

	Sample size				
ρ	50	100	200	400	800
0.1	0.140	0.099	0.070	0.050	0.035
0.3	0.129	0.091	0.064	0.046	0.032
0.5	0.106	0.075	0.053	0.038	0.027
0.7	0.072	0.051	0.036	0.026	0.018

- Standard error decreases as $|\rho|$ increases.
- So, if you want to keep the standard error less than 0.05, you need $N = 400$ when the "typical" correlation is only 0.1, but $N = 100$ when the "typical" correlation is 0.7.
- In many behavioural and psychology studies, correlations among *different scales* are modest, at best ($0.1 \leq \rho \leq 0.3$).
- For typical scale analysis, one should expect the correlations among items on the *same scale* to be much higher ($0.7 \leq \rho \leq 0.9$), $\Rightarrow$ smaller required sample size for the same standard error.

Power analysis

Recall the basis for power analysis using χ^2 tests:



- Under H_0 (perfect fit) the test statistic $X^2 = (N-1)F_{min} \sim \chi^2(df)$
- Reject H_0 if $X^2 > \chi^2_{1-\alpha}(df)$
- Under H_1, X^2 gives larger values, a non-central $\chi^2(df, \lambda > 0)$ distribution
- Power = $\Pr(X^2 > \chi^2_{1-\alpha} \mid H_1)$

Power and Sample size for CFA and SEM

- ***Problems:*** The situation in CFA wrt power analysis is typically reversed compared with other forms of hypothesis tests—
    - $X^2 = (N-1)F_{min}$, so large $N \Rightarrow$ reject H_0.
    - With small specification errors, large sample size will magnify their effects $\Rightarrow$ reject H_0.
    - With large specification errors, small sample size will mask their effects $\Rightarrow$ accept H_0.
- ***Solutions:***
    - Use an interpretable statistic that maps directly to the χ^2 non-centrality parameter, λ
    - Turn the test around, so rather than testing $H_0 : \lambda = 0$ (perfect fit) we can test $H_0 : \lambda < \lambda_0$ (acceptable fit)

Power and Sample size for CFA and SEM

- ***Overall RMSEA approach:***
  MacCallum, Browne and Sugawara (1996) approach allows for testing a null hypothesis of 'not-good-fit', so that a significant result provides support for good fit.
  - Effect size is defined in terms of a null hypothesis and alternative hypothesis value of the root-mean-square error of approximation (RMSEA) index. Typical values for RMSEA:

$\leq .05$	close fit
$.05 - .08$	fair
$.08 - .10$	mediocre
$> .10$	poor

  - These values, together with the df for the model being fitted, sample size (*N*), and error rate (α), allow power to be calculated.

Power and Sample size for CFA and SEM

- The CSMPOWER macro
  - See: http://datavis.ca/sasmac/csmpower.html
  - Retrospective power analysis— uses the RMSEA values from the OUTRAM= data set from PROC CALIS for the model fitted.
  - Prospective power analysis— values of RMSEA, DF and N must be provided through the macro arguments.

Example: Retrospective power analysis

Here, we examine the power for the test of Lord's two-factor model for speeded and unspeeded vocabulary tests, where $N = 649$.

```
title "Power analysis: Lord's Vocabulary Data";
title2 "Lord's data: H1- X1 and X2 parallel,
        Y1 and Y2 parallel, rho=1";
proc calis data=lord cov summary outram=ram1;
   lineqs  x1  = betax F1  + e1,
           x2  = betax F1  + e2,
           y1  = betay F2  + e3,
           y2  = betay F2  + e4;
   std  F1 F2 = 1 1,
        e1 e2 e3 e4 = vex vex vey vey;
   cov  F1 F2 = 1;
run;

*-- Power analysis from RMSEA statistics in this model;
title 'Retrospective power analysis';
%csmpower(data=ram1);
```

Example: Retrospective power analysis

Results include:

Alpha	df	Name of Variable	N	H0 fit value	Ha fit value	Power
0.05	6	RMSEAEST	649	0.05	0.08977	0.75385
		RMSEALOB	649	0.05	0.06349	0.19282
		RMSEAUPB	649	0.05	0.11839	0.99202

With this sample size, we have power of 0.75 to distinguish between a fit with RMSEA=0.05 and one with RMSEA=0.09.

Example: Prospective power analysis

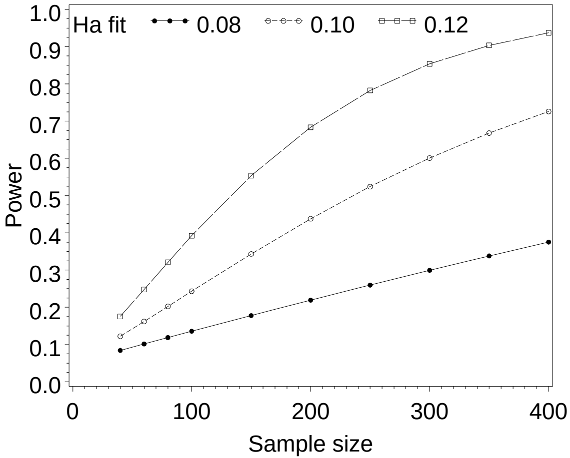
For prospective power analysis, we specify the RMSEA for alternative hypotheses of 'not good fit' with the RMSEAA= parameter (for H_a).

```
*--; title  'Prospective power analysis';
%csmpower(df=6, rmseaa=%str(.08 to .12 by .02),
          plot=%str(power*n =rmseaa));
```

Results include a listing:

Alpha	df	N	H0 fit value	Ha fit value	Power
0.05	6	40	0.05	0.08	0.08438
		40	0.05	0.10	0.12243
		40	0.05	0.12	0.17575
		60	0.05	0.08	0.10168
		60	0.05	0.10	0.16214
		60	0.05	0.12	0.24802
		80	0.05	0.08	0.11883
		80	0.05	0.10	0.20262
		80	0.05	0.12	0.32093
		100	0.05	0.08	0.13585
		100	0.05	0.10	0.24333
		100	0.05	0.12	0.39214
		...	...	...	...
		400	0.05	0.08	0.37545
		400	0.05	0.10	0.72599
		400	0.05	0.12	0.93738

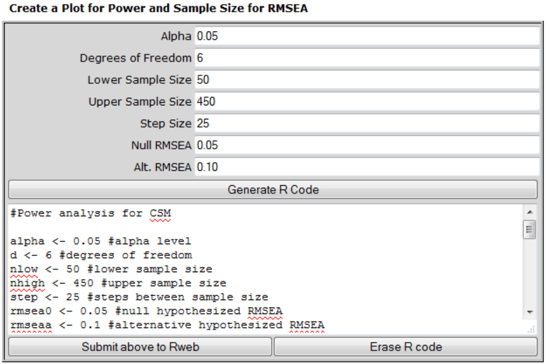
Plot of Power by N for each level of RMSEAA:

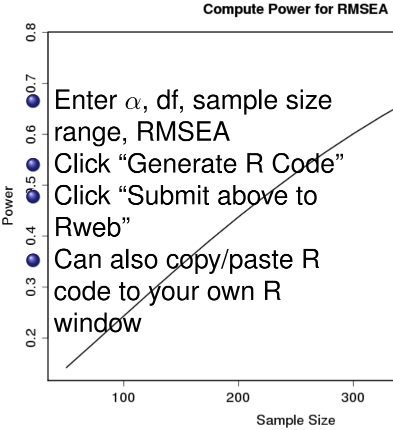


- For the most stringent test of H_0 : RMSEA $= 0.05$ vs. H_a : RMSEA $= 0.08$, the largest sample size, $N = 400$ only provides a power of 0.375.
- Good thing they used $N = 649$!

Online RMSEA power calculator

Several online web applications use R with a forms interface, e.g., Preacher & Coffman, `http://www.quantpsy.org/rmsea/rmsea.htm`





- Enter α, df, sample size range, RMSEA
- Click "Generate R Code"
- Click "Submit above to Rweb"
- Can also copy/paste R code to your own R window

semTools package

The semTools package contains functions for these purposes:

- **plotRMSEApower()**: plot power based on population RMSEA given sample size range
- **findRMSEApower()**: find power based on population RMSEA given a sample size
- **findRMSEAsamplesize()**: find minium sample size for given power based on population RMSEA

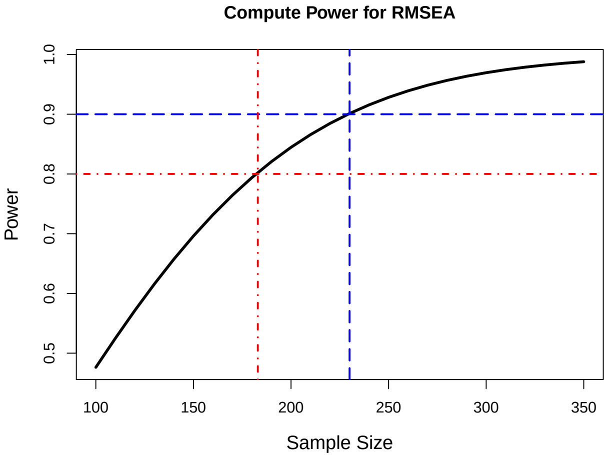
What sample size is required for power = (0.8, 0.9) to detect difference bewtween RMSEA_0=0.025 and RMSEA_A=0.08 with df=23?

```
library(semTools)
findRMSEAsamplesize(rmsea0=0.025, rmseaA=0.08, df=23, power=0.80)

## [1] 183

findRMSEAsamplesize(rmsea0=0.025, rmseaA=0.08, df=23, power=0.90)

## [1] 230
```

semTools package

Plot the power curve:

```
plotRMSEApower(rmsea0=.025, rmseaA=.08, df=23, 100, 350, 10, cex.lab=1.25,
abline(h=c(0.8, 0.9), col=c("red", "blue"), lty=4:5, lwd=2)
abline(v=c(183, 230), col=c("red", "blue"), lty=4:5, lwd=2)
```



Individual model specifications

- The overall approach evaluates power or required sample size for the whole model.
- It does not distinguish among the *a priori* specifications of free and fixed parameters implied by the model being tested.
- Things become more difficult when the focus is on power for deciding on some one or a few specifications (parameters) in a model.
  - In a mediation model, how to determine sample size to test the mediator effect?
  - In a higher-order CFA model, what sample size do I need to distinguish among competing models for 2^{nd}-order factors?
  - In a complex SEM, how to distinguish among competing models?

Individual model specifications

There are some promising results:

- Satorra (1989): modification indices— $\Delta\chi^2$ for *fixed parameters*— in a model approximate the χ^2 non-centrality parameters required to determine power for a specific fixed parameter.
- Similarly, Wald tests, $\chi_1^2 = (\text{par}/s(\text{par}))^2$ approximate the χ^2 non-centrality parameters required to determine power for *free parameters*.
- These χ^2 values should be studied in relation to the estimated change in the parameter (ECP).
  - A large $\Delta\chi^2$ with a small ECP simply reflects the high power to detect small differences which comes with large N.
  - Similarly, a small $\Delta\chi^2$ with a large ECP reflects low power for large differences with small N.

See Kaplan, "Statistical power in structural equation models", `www.gsu.edu/~mkteer/power.html` for further discussion and references on these issues.

Comparing nested models

A simpler method was suggested by McCallum, Browne & Cai (2006) to find sample size or compute power in the comparison of two nested models

- Recall that difference between two nested models, $A \subset B$, with degrees of freedom df_A and df_B can be tested with the likelihood ratio test

$$\Delta X^2 = X^2_A - X^2_B = (N-1)(F^A_{min} - F^B_{min}) \sim \chi^2 \text{ with } df = df_A - df_B$$

- Under H_0 : Models A and B do not differ in fit, $\Delta X^2 \sim \chi^2(df_A - df_B)$
- Under H_1 : Model B fits better, ΔX^2 is a non-central χ^2 with non-centrality $\lambda = (N-1)(F^A_{min} - F^B_{min})$
- This can be specified in terms of (the population) RMSEA as

$$RMSEA \equiv \epsilon = \sqrt{F_{min}/df} \quad \Longrightarrow \quad \lambda = (N-1)(df_A \epsilon^2_A - df_B \epsilon^2_B)$$

- Thus, you can find the power or sample size needed to detect a difference between two models

The semTools package contains functions for these purposes:

- **plotRMSEApowernested()**: plot power of nested model RMSEA, given sample size range
- **findRMSEApowernested()**: find the power for a given sample size in nested model comparison
- **findRMSEAsamplesizenested()**: find minium sample size for given power in nested model comparison

Examples: Model A has 22 df, model B has 20 df. What sample size do I need to detect a difference between $\text{RMSEA}_A = 0.075$ and $\text{RMSEA}_B = 0.05$ with power=0.9?

```
findRMSEAsamplesizenested(rmsea1A = 0.075, rmsea1B = 0.05,
  dfA = 22, dfB = 20, power=0.9)

## [1] 173
```

What is the power if I only have N=100?

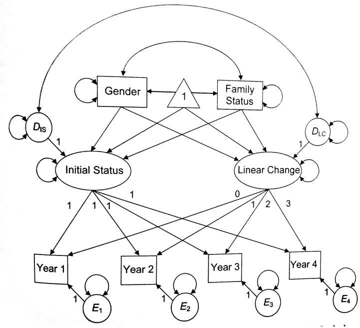
```
findRMSEApowernested(rmsea1A = 0.075, rmsea1B = 0.05,
  dfA = 22, dfB = 20, n = 100)

## [1] 0.67513
```

SEM Extensions I

- A variety of methods handle ***non-normal distributions***
  - Robust ML (Satorra-Bentler) corrects the χ^2 statistic and standard errors for excess kurtosis
  - Asymptotically distribution-free (ADF) methods do something similar
  - Bootstrap methods avoid normality assumptions by re-sampling from the data— data-based standard errors
- ***categorical variables***
  - Likert scales with > 5 ordered categories can usually be treated as continuous, applying robust ML
  - Otherwise, one can use polychoric correlations rather than Pearson correlations
  - Usually this is done via a form of weighted least squares rather than ML estimation
- ***missing data*** is readily handled using multiple imputation methods

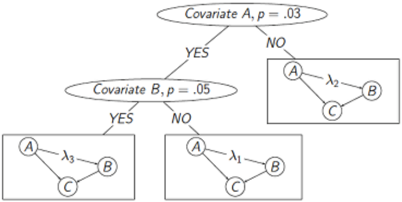
SEM Extensions II

- ***latent growth models*** extend the SEM approach to longitudinal data, allowing for measurement error
  - Can be used to investigate systematic change, or growth, and inter-individual variability in this change
  - Can incorporate time-invariant or time-varying exogenous covariates
  - (Alternatives are: repeated measure ANOVA/MANOVA, mixed-models)



SEM Extensions III

- ***structural equation mixture models*** (SEMM):
  - Supposes latent class variables that partition the data into subgroups
  - Correlations arise from a mixture of multivariate normal distributions
  - Subgroup models are linear, but overall model can allow nonlinear relations
- ***structural equation model trees*** (semtree):
  - Combine strengths of SEM and recursive-partitioning decision trees ("CART")
  - Partitions dataset recursively into subsets with significantly different parameter estimates in SEM



Summary

- The general SEM allows measurement models for exogenous ($\boldsymbol{x}$) and endogenous ($\boldsymbol{y}$) variables in terms of latent variables ξ and η

$$\boldsymbol{x} = \Lambda_x \xi + \delta \qquad \boldsymbol{y} = \Lambda_y \eta + \epsilon$$

- This allows for treating errors of measurement and reduces bias
- These are connected by a structural model

$$\eta = \boldsymbol{B}\eta + \Gamma\xi + \zeta$$

- Path analysis models are the special case of no latent variables
- CFA models are the special case of only one set of observed variables
- When raw data are available, data screening is an important prelude to SEM modeling
- SEM methods have been extended to handle a wide variety of data structures and new model types!