Introduction to structural equation modeling and mixed models in R

Day 3 – Part 3: SEM

Oksana Buzhdygan

oksana.buzh@fu-berlin.de

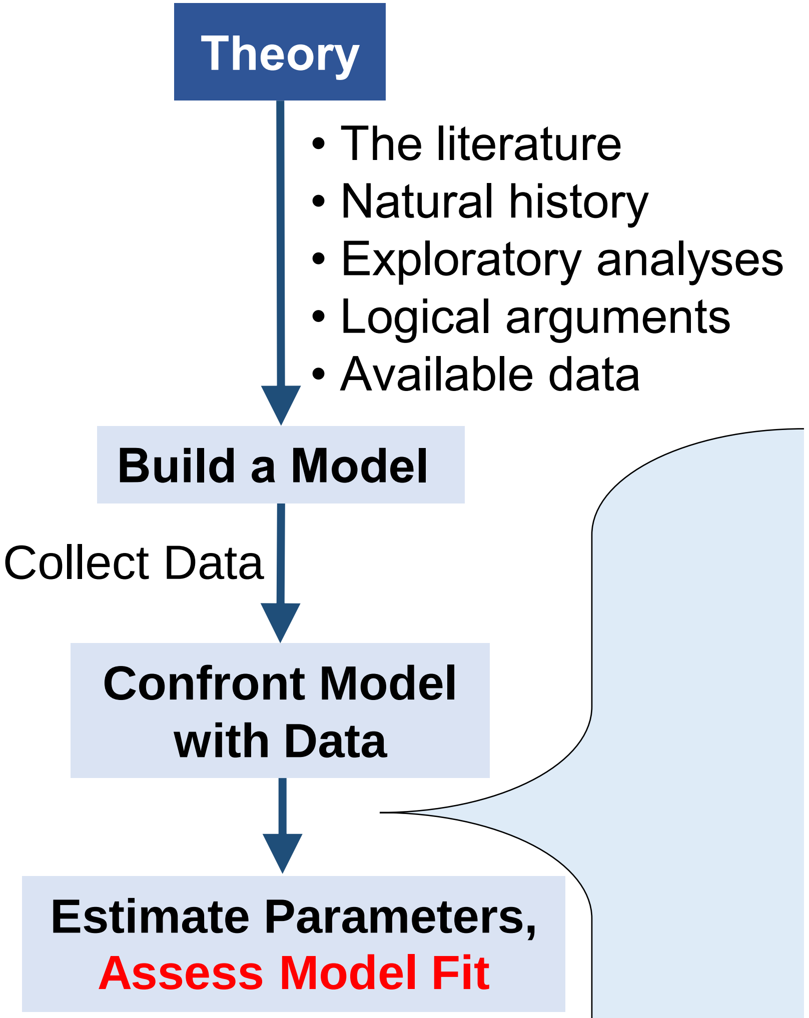
Day 3 – Part 3

Outline

- Introduction to Covariance-based SEM

  - ✓ SEM using likelihood and covariance matrices
  - ✓ Model Identifiability
  - ✓ Sample Size for SEM
  - ✓ Assessing model fit: χ^2, related indices

SEM workflow process



Theory

- The literature
- Natural history
- Exploratory analyses
- Logical arguments
- Available data

Build a Model

Collect Data

Confront Model with Data

Estimate Parameters, Assess Model Fit

How well our data correspond to our model?

Two Paradigms for model estimation

Covariance-Based Estimation

(*lavaan*)

Global estimation:

- reproduce a single variance-covariance matrix

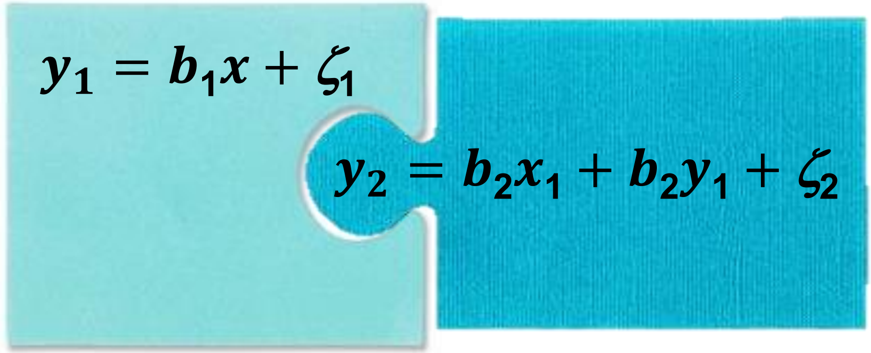
$$\left\{\begin{matrix} \sigma_x & & \\ \sigma_{xy_1} & \sigma_{y_1} & \\ \sigma_{xy_2} & \sigma_{y_1y_2} & \sigma_{y_2} \end{matrix}\right\}$$

Local Equation Estimation

(*piecewiseSEM*)

Local estimation:

- fit a model for each response
- strings together the inferences



$$y_1 = b_1x + \zeta_1$$

$$y_2 = b_2x_1 + b_2y_1 + \zeta_2$$

Covariance-based SEM

```
cov(data1)
>
           y1         x1         y2
y1 0.06939250 0.06384005 0.07851289
x1 0.06384005 0.12346977 0.06469415
y2 0.07851289 0.06469415 0.17174160
```

$= \mathbf{S}$

Observed
variance-covariance matrix

Maximum-Likelihood Estimation

$\mathbf{S} = \hat{\boldsymbol{\Sigma}}$

Implied
(model-estimated)
variance-covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \left\{\begin{array}{ccc} \sigma_x & & \\ \sigma_{xy_1} & \sigma_{y_1} & \\ \sigma_{xy_2} & \sigma_{y_1y_2} & \sigma_{y_2} \end{array}\right\}$$

Covariance-based SEM

Maximum Likelihood ML

Likelihood Function:

tr trace of the matrix

p number of endogenous variables

$$\boldsymbol{F_{ML}} = log\left|\widehat{\boldsymbol{\Sigma}}\right| + tr\left(\boldsymbol{S}\widehat{\boldsymbol{\Sigma}}^{-1}\right) - log|\boldsymbol{S}| - (p + q)$$

$\widehat{\boldsymbol{\Sigma}}$ modeled covariance matrix

$\boldsymbol{S}$ observed covariance matrix

q number of exogenous variables

Covariance-based SEM

Maximum Likelihood ML

Likelihood Function:

tr trace of the matrix

p number of endogenous variables

$$F_{ML} = log|\widehat{\boldsymbol{\Sigma}}| + tr(\mathbf{S}\widehat{\boldsymbol{\Sigma}}^{-1}) - log|\mathbf{S}| - (p + q)$$

$\widehat{\boldsymbol{\Sigma}}$ modeled covariance matrix

$\mathbf{S}$ observed covariance matrix

q number of exogenous variables

Perfect model fit

$$F_{ML} = 0$$

Covariance-based SEM

Maximum Likelihood ML

Likelihood Function:

tr trace of the matrix

p number of endogenous variables

$$\boldsymbol{F_{ML}} = log\left|\widehat{\boldsymbol{\Sigma}}\right| + tr\left(\mathbf{S}\widehat{\boldsymbol{\Sigma}}^{-1}\right) - log|\mathbf{S}| - (p + q)$$

$\widehat{\boldsymbol{\Sigma}}$ modeled covariance matrix

$\boldsymbol{S}$ observed covariance matrix

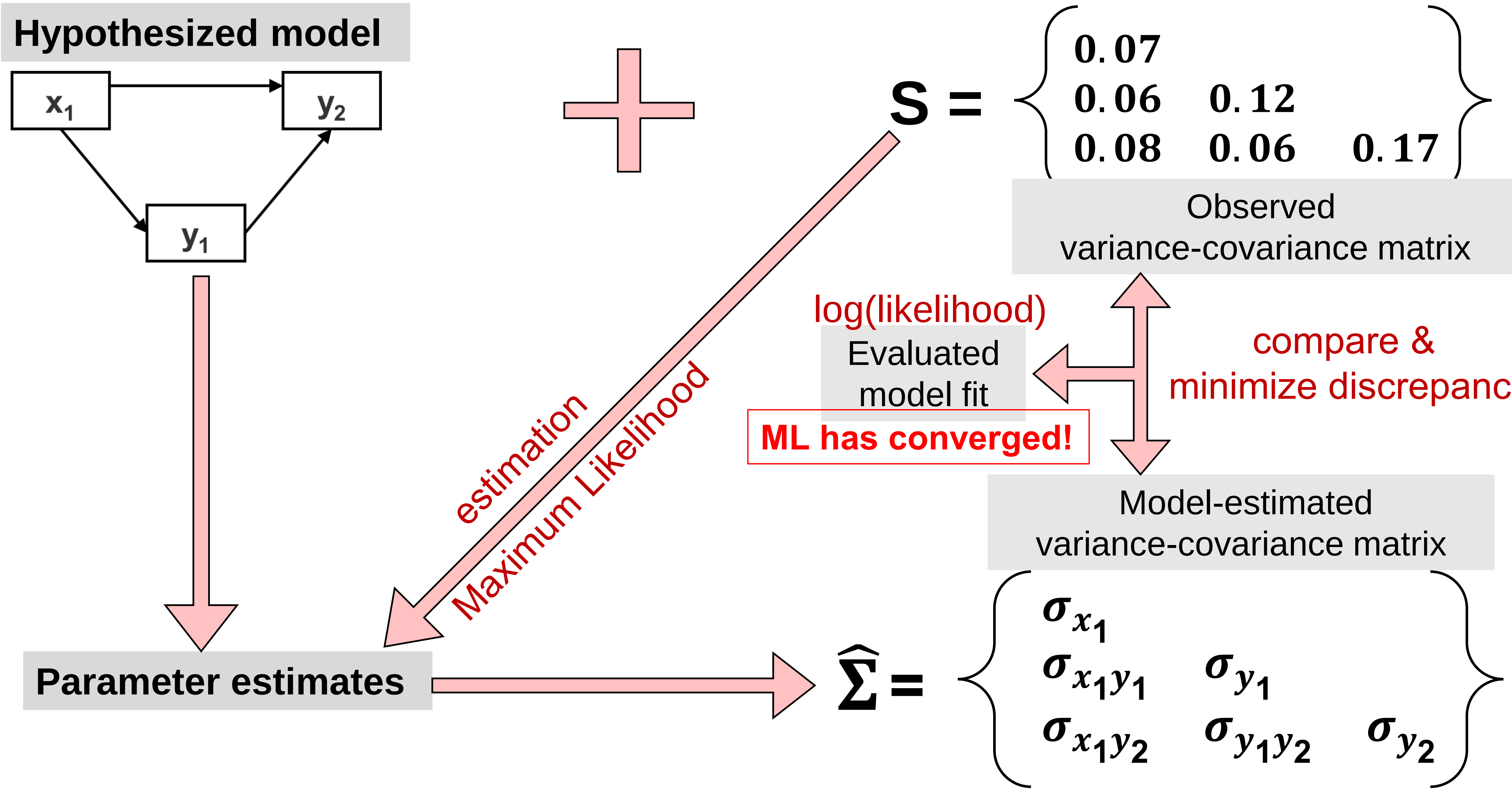
q number of exogenous variables

Desirable properties of $\boldsymbol{F_{ML}}$:

- scale invariant
- asymptotically unbiased
- efficient

Covariance-based SEM

Global Estimation



Hypothesized model

$$\mathbf{S} = \left\{\begin{matrix} 0.07 & & \\ 0.06 & 0.12 & \\ 0.08 & 0.06 & 0.17 \end{matrix}\right\}$$

Observed variance-covariance matrix

log(likelihood)

Evaluated model fit

ML has converged!

compare & minimize discrepancy

estimation

Maximum Likelihood

Model-estimated variance-covariance matrix

Parameter estimates

$$\widehat{\boldsymbol{\Sigma}} = \left\{\begin{matrix} \sigma_{x_1} & & \\ \sigma_{x_1 y_1} & \sigma_{y_1} & \\ \sigma_{x_1 y_2} & \sigma_{y_1 y_2} & \sigma_{y_2} \end{matrix}\right\}$$

Day 3 – Part 3

Outline

- Introduction to Covariance-based SEM

✓ SEM using likelihood and covariance matrices

✓ **Model Identifiability**

✓ Sample Size for SEM

✓ Assessing model fit: χ^2, related indices

Model Identifiability

Can I fit my model?

- To fit a model we need enough ‘known’ pieces of information to produce unique estimates of ‘unknown’ parameters

We can not fit the model !

$a+b=8$

Unidentified

- **no unique estimates**

$a+b=8$

$a=3b$

$2a-4=4b$

Overidentified

- **more ‘known’ than ‘unknown’**

We can evaluate model fit !

$a+b=8$

$a=3b$

$(3b)+b=8$

$4b=8$

$\mathbf{b}=8/4=\mathbf{2}$

$a+2=8$

$\mathbf{a}=8-2=\mathbf{6}$

Just Identified

- **unique estimates**

b=2

a=6

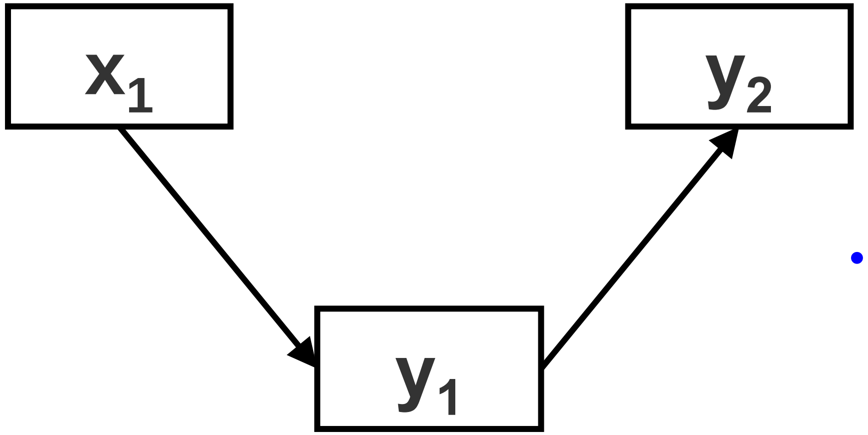
We can fit model !

- In SEM ‘knowns’ are the variances & covariances of observed variables

- Unknowns are the model parameters to be estimated

Model Identifiability

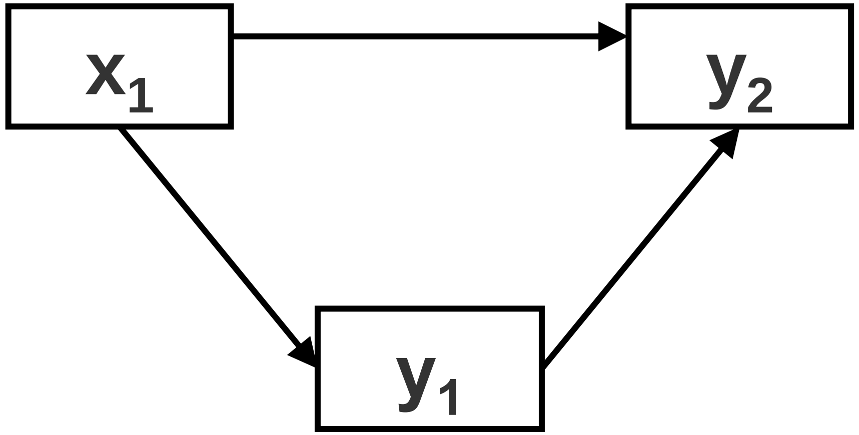
Can I fit my model?



Overidentified (Unsaturated)

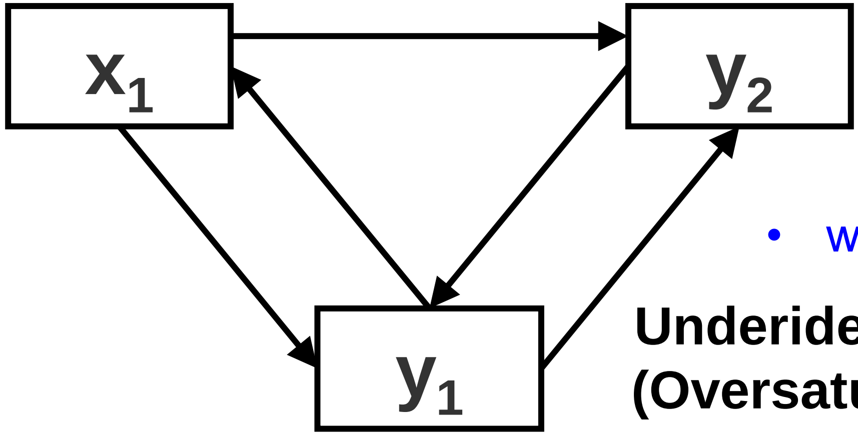
Recursive models

- all causal effects are unidirectional



Just Identified (Saturated)

```
lavaan WARNING:
Could not compute standard errors!
...This may be a symptom that the
model is not identified.
```



Non-recursive models

- with bidirectional feedbacks

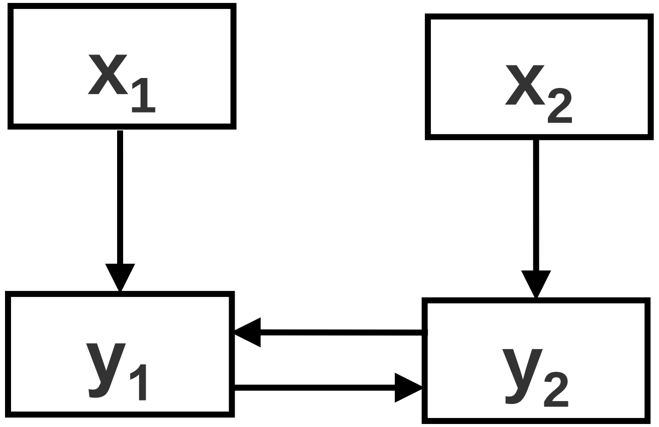
Underidentified (Oversaturated)

Model Identifiability

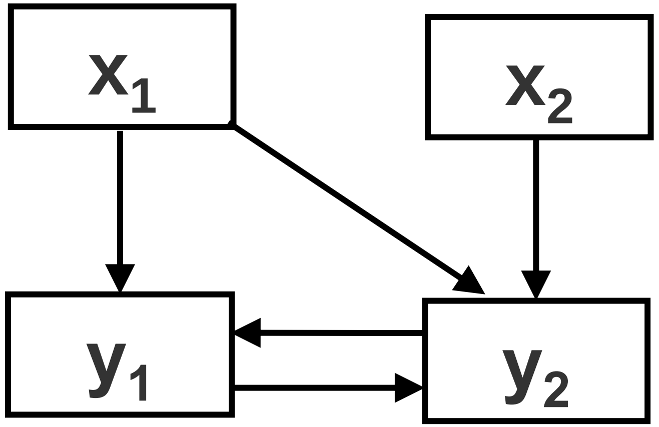
Can I fit my model?

Can non-recursive models be identified?

YES:
if responses have unique information



NO:
if not enough information for unique solution



Model Identifiability

Can I fit my model?

Assessing identification status: t-rule

$$\mathbf{DF} = t_{max} - t$$

maximum number of parameters that can be estimated, given s

s number of observed variables

$$t_{max} = \frac{s(s+1)}{2}$$

$t = t_{max}$ Just identified

$t > t_{max}$ Unidentified

$t < t_{max}$ Overidentified

t number of parameters to be estimated by the model

	x$_1$	y$_1$	y$_2$
x$_1$	0.07		
y$_1$	0.06	0.12	
y$_2$	0.08	0.06	0.17

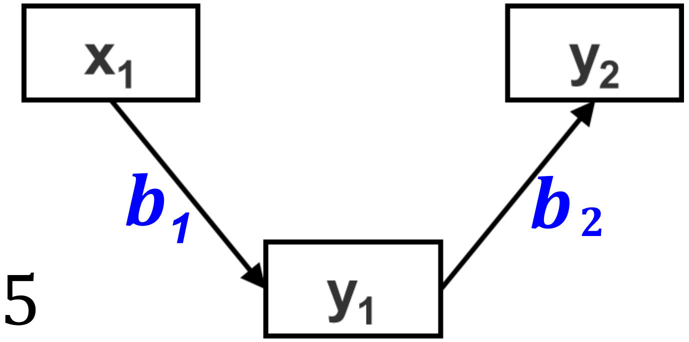
Observed variance-covariance matrix

$s = 3$

$t_{max} = 6$

$t = 2 + 3 = 5$

$5 < 6$



Overidentified
(Unsaturated)

Day 3 – Part 3

Outline

- Introduction to Covariance-based SEM

✓ SEM using likelihood and covariance matrices

✓ Model Identifiability

✓ **Sample Size for SEM**

✓ Assessing model fit: χ^2, related indices

Sample Size

Is my sample size enough?

The basic rule-of-thumb:

$\boldsymbol{n}$ sample size

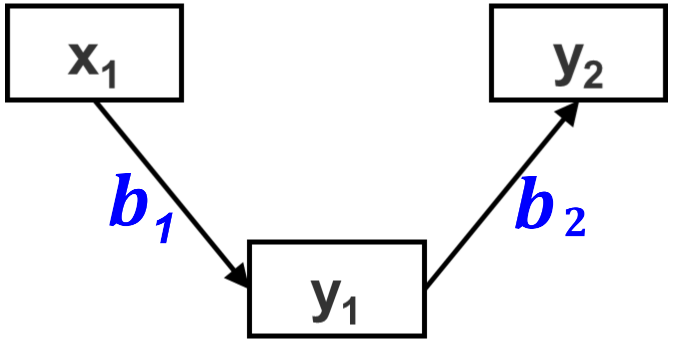
Minimum requirement $\boldsymbol{n} = \boldsymbol{p} \times 5$

Ideally $\boldsymbol{n} = \boldsymbol{p} \times 20$

$\boldsymbol{p}$ number of path coefficients

$$\boldsymbol{k} = \frac{\boldsymbol{p}^{\frac{3}{2}}}{\boldsymbol{n}} \approx \mathbf{0}$$

The larger the sample size, the more precise (unbiased) the estimates will be.



$\boldsymbol{p} = 2$

$\boldsymbol{n} = 2 \times 5 = 10 \quad \boldsymbol{k} = \mathbf{0.16}$

$\boldsymbol{n} = 2 \times 20 = 40 \quad \boldsymbol{k} = \mathbf{0.03}$

Day 3 – Part 3

Outline

- Introduction to Covariance-based SEM

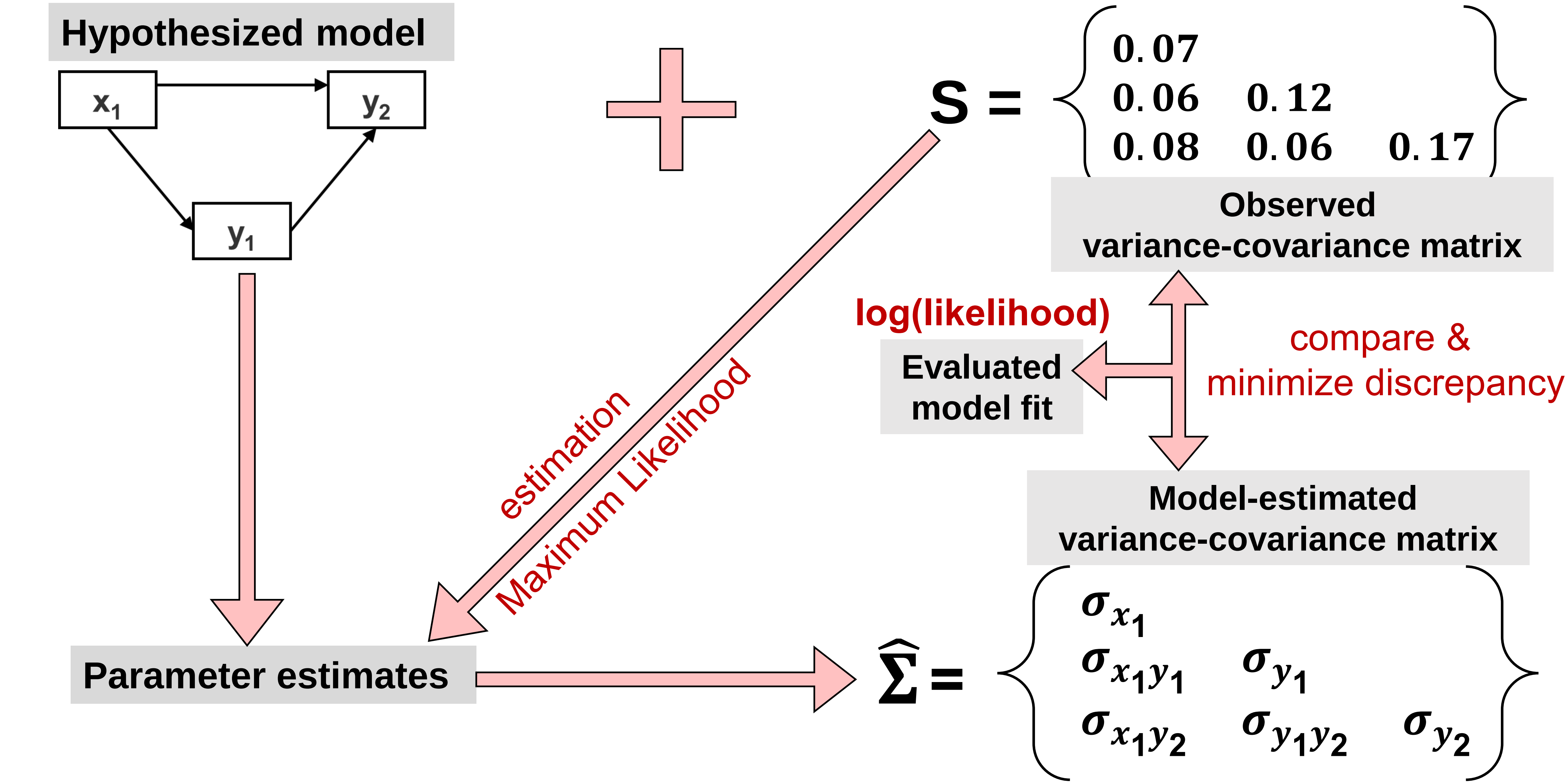
  ✓ SEM using likelihood and covariance matrices

  ✓ Model Identifiability

  ✓ Sample Size for SEM

  ✓ **Assessing model fit: χ^2, related indices**

Covariance-based SEM



Hypothesized model

$$S = \begin{Bmatrix} 0.07 & & \\ 0.06 & 0.12 & \\ 0.08 & 0.06 & 0.17 \end{Bmatrix}$$

Observed variance-covariance matrix

log(likelihood)

Evaluated model fit

compare & minimize discrepancy

estimation

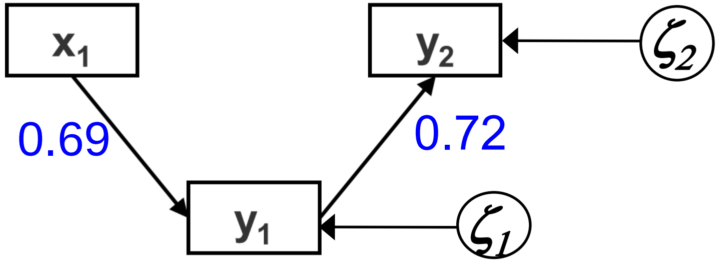
Maximum Likelihood

Model-estimated variance-covariance matrix

Parameter estimates

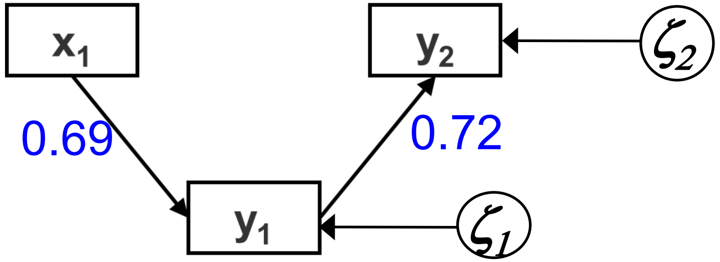
$$\hat{\Sigma} = \begin{Bmatrix} \sigma_{x_1} & & \\ \sigma_{x_1 y_1} & \sigma_{y_1} & \\ \sigma_{x_1 y_2} & \sigma_{y_1 y_2} & \sigma_{y_2} \end{Bmatrix}$$

Goodness of fit



```
data1 <- read.table("Data/SEMdata1.txt", header = T)

# Specify the model in lavaan
sem_mod1 <- ` y1 ~ x1
              y2 ~  y1
`
# Fit the model
sem.fit1 <- sem(sem_mod1, data=data1)

# Extract results
summary(sem.fit1, standardize = T)
```

Goodness of fit



Observed covariance matrix (scaled)

	x_1	y_1	y_2
x_1	1.00		
y_1	0.69	1.00	
y_2	0.44	0.72	1.00

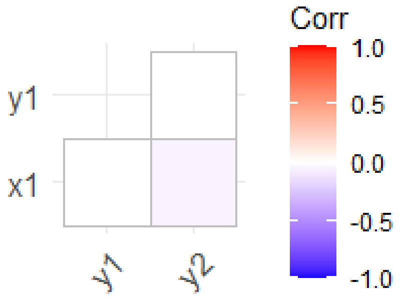
Model implied matrix (scaled)

	x_1	y_1	y_2
x_1	1.00		
y_1	0.69	1.00	
y_2	0.49	0.72	1.00

residual
0.444-0.496=**-0.052**

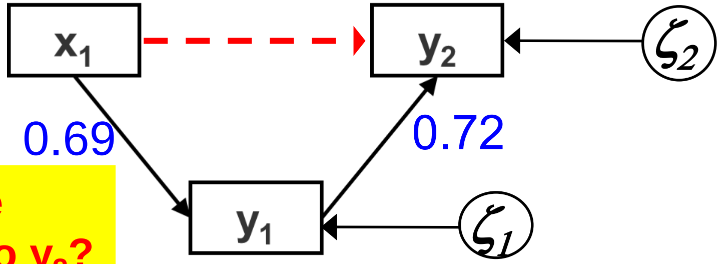
Residuals r (scaled)

	x_1	y_1	y_2
x_1	0		
y_1	0	0	
y_2	-0.052	0	0



```
# Model implied covariance matrix (standardised)
lavInspect(sem.fit1, what="cor.all")

# Observed covariance matrix (standardised)
lavCor(sem.fit1)

# Residuals (standardised)
resid(sem.fit1, "cor")


 library(ggcorrplot)
 ggcorrplot(resid(sem.fit1,type="cor")$cov,
                                    type="lower")
```

Goodness of fit



Should there be a path from x_1 to y_2?

Observed covariance ma

	x_1
x_1	1.0
y_1	0.6
y_2	0.4

Model implied
x (scaled)

y_1	y_2
1.00	
0.72	1.00

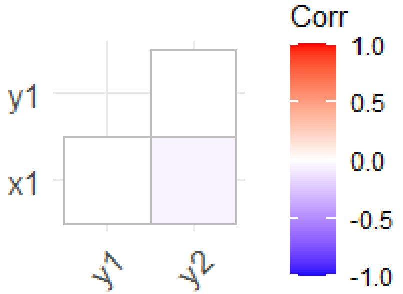
Is it good enough?

0.444-0.496=**-0.052**

Residuals r (scaled)

	x_1	y_1	y_2
x_1	0		
y_1	0	0	
y_2	-0.052	0	0

Look for r>0.1



```
# Model implied covariance matrix (standardised)
lavInspect(sem.fit1, what="cor.all")

# Observed covariance matrix (standardised)
lavCor(sem.fit1)

# Residuals (standardised)
resid(sem.fit1, "cor")


 library(ggcorrplot)
 ggcorrplot(resid (sem.fit1,type="cor")$cov,
                                    type="lower")
```

Goodness of fit

Likelihood Function:

p number of endogenous variables

$$\boldsymbol{F_{ML}} = log\left|\widehat{\boldsymbol{\Sigma}}\right| + tr\left(\mathbf{S}\widehat{\boldsymbol{\Sigma}}^{-1}\right) - log|\mathbf{S}| - (p + q)$$

Perfect model fit

$\boldsymbol{F_{ML}} = 0$

$\widehat{\boldsymbol{\Sigma}}$ modeled covariance matrix

$\boldsymbol{S}$ observed covariance matrix

q number of exogenous variables

$$\boldsymbol{\chi^2 = (n - 1)F_{ML}}$$

χ^2 model fit

$\boldsymbol{n}$ sample size

Goodness of fit

$$\chi^2 = (n-1)F_{ML}$$

n sample size

DF degrees of freedom

$$DF = \frac{s(s+1)}{2} - t$$

from the t-rule

s number of observed variables

t number of parameters to be estimated by the model

Goodness of fit

$$\chi^2 = (n-1)F_{ML}$$

$\boldsymbol{n}$ sample size

H0: no difference between model-implied and observed covariance matrices χ^2 =0 (the model fits perfectly)

Good fit: P > 0.05 failing to reject **H0**

- **Large χ^2 implies LACK of fit**
- **Scaling by sample size**

$\boldsymbol{DF}$ degrees of freedom

$$DF = \frac{s(s+1)}{2} - t$$

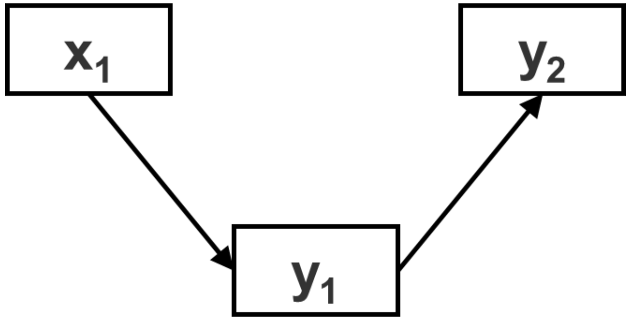
$\boldsymbol{s}$ number of observed variables

$\boldsymbol{t}$ number of parameters to be estimated by the model

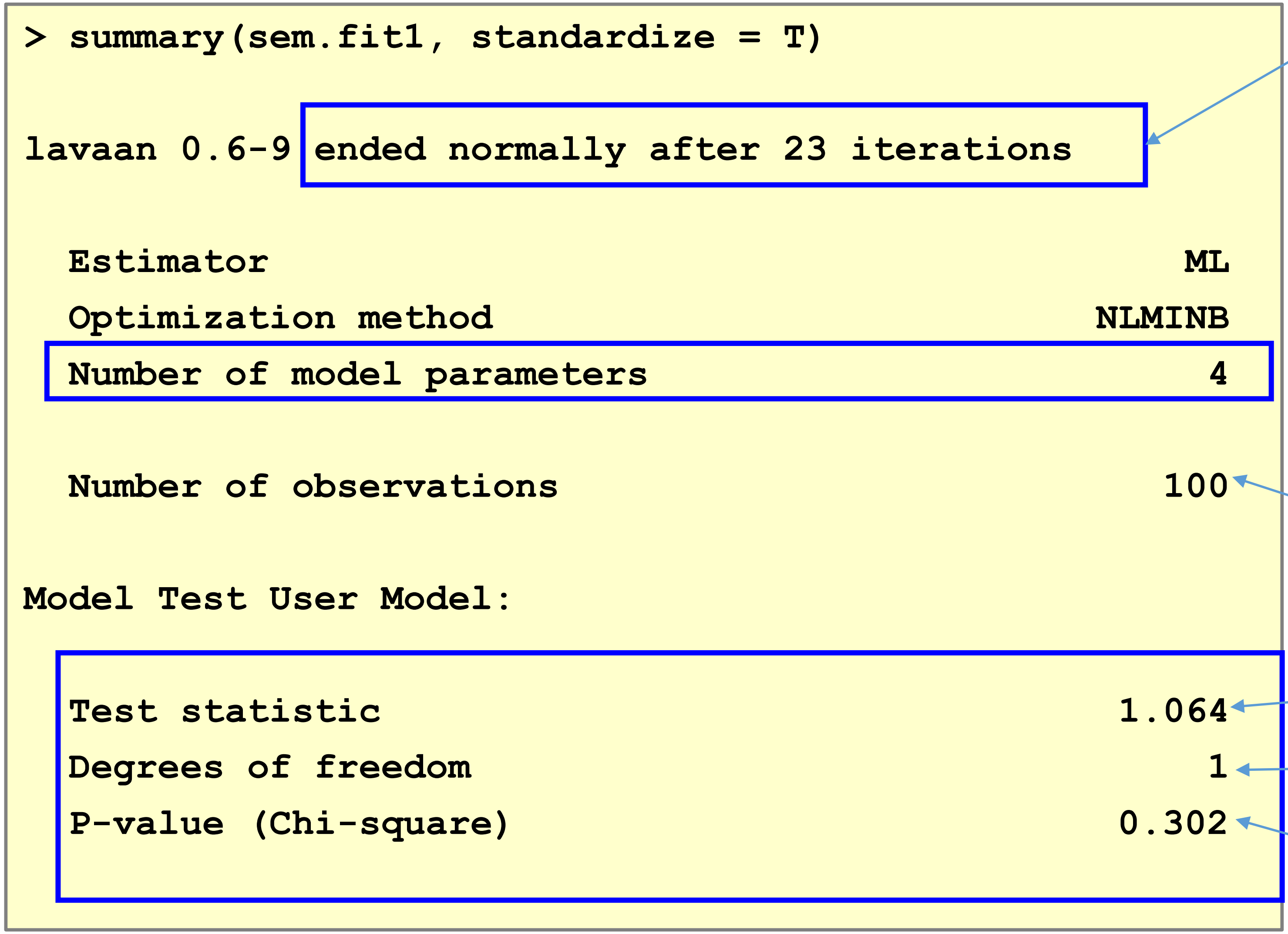
from the t-rule

Goodness of fit

```
data1 <- read.table("Data/SEMdata1.txt", header = T)

# SEM model in lavaan
sem_mod1 <- ` y1 ~ x1
              y2 ~  y1
`
sem.fit1 <- sem(sem_mod1, data=data1)

summary(sem.fit1, standardize = T)
```

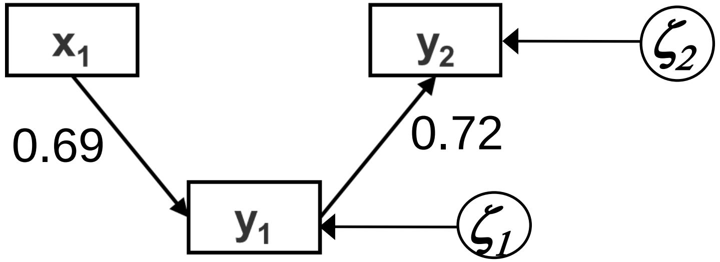


Goodness of fit



```
> summary(sem.fit1, standardize = T)

lavaan 0.6-9 ended normally after 23 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         4

  Number of observations                           100

Model Test User Model:

  Test statistic                                 1.064
  Degrees of freedom                                 1
  P-value (Chi-square)                           0.302
```

ML converged normally



Number of rows in dataset

χ^2

DF

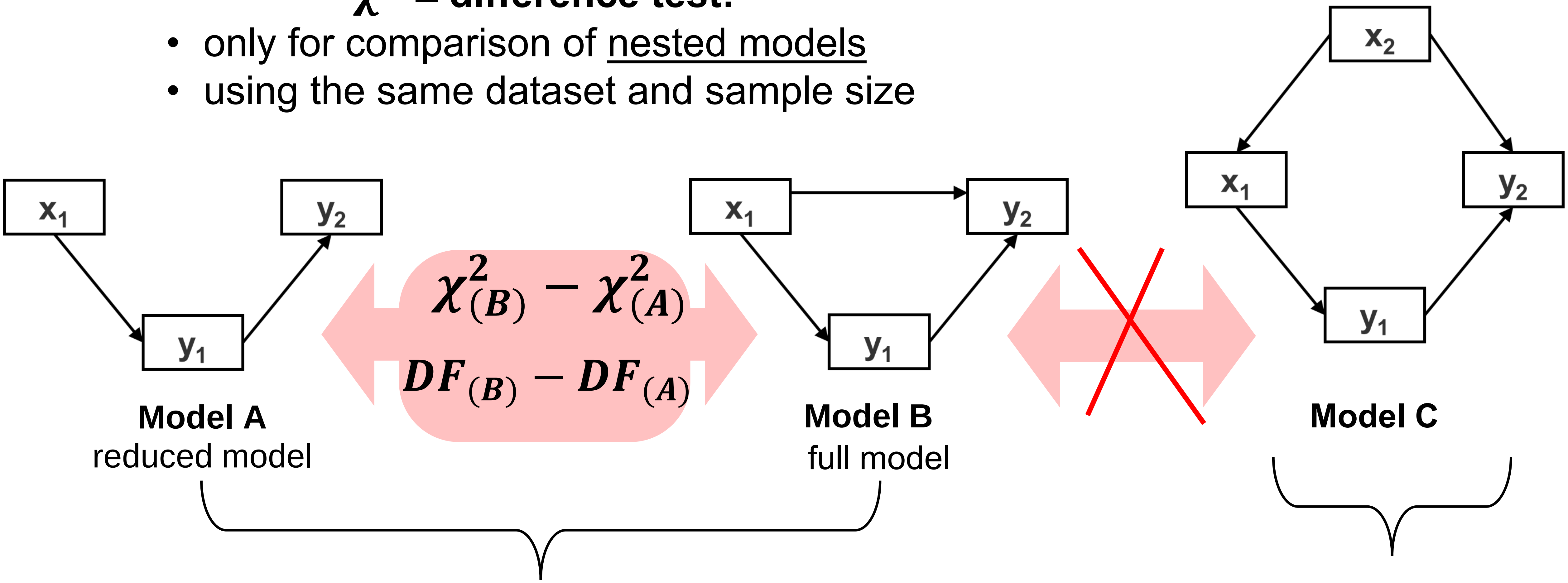
p

p>0.05 means no discrepancy between sample and observed covariance matrix (GOOD FIT)

Goodness of fit

χ^2 – difference test:

- only for comparison of <u>nested models</u>
- using the same dataset and sample size



$$\chi^2_{(B)} - \chi^2_{(A)}$$

$$DF_{(B)} - DF_{(A)}$$

Model A
reduced model

Model B
full model

Model C

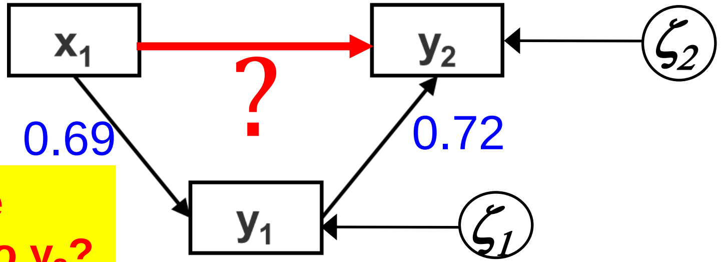
A is nested in B

- the same variables but additional parameter to be estimated

C is not nested (in A or B)

- has additional variable x_2

Goodness of fit



Should there be a path from x_1 to y_2?

Observed covariance matrix (scaled)

	x_1	y_1	y_2
x_1	1.00		
y_1	0.69	1.00	
y_2	0.44	0.72	1.00

Model implied matrix (scaled)

	x_1	y_1	y_2
x_1	1.00		
y_1	0.69	1.00	
y_2	0.49	0.72	1.00

residual
0.444-0.496=**-0.052**

χ^2 statistics:

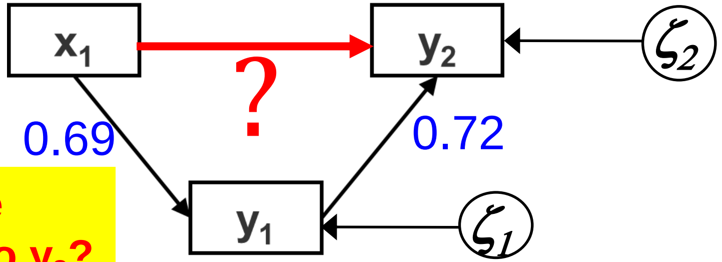
χ^2 = 1.06, DF=1, n=100, p = 0.3

χ^2 – difference test:

- only for comparison of nested models
- using the same dataset and sample size

```
# SEM model 1
sem_mod1 <- ` y1 ~ x1
              y2 ~  y1
`
sem.fit1 <- sem(sem_mod1, data=data1)

# SEM model 2
sem_mod2 <- ` y1 ~ x1
              y2 ~  y1 + x1
`
sem.fit2 <- sem(sem_mod2, data=data1)

# Chi-Squared Difference Test
anova(sem.fit1, sem.fit2)
```

Goodness of fit



Should there be a path from x_1 to y_2?

χ^2 – difference test:

- only for comparison of nested models
- using the same dataset and sample size

χ^2 statistics:

χ^2 = 1.06, DF=1, n=100, p = 0.3

- **Our model is good enough**
- **No modifications needed**

```
# results

> anova(sem.fit1, sem.fit2)

Chi-Squared Difference Test

         Df     AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
sem.fit2  0 -5.8616 7.1643 0.0000
sem.fit1  1 -6.7977 3.6230 1.0639     1.0639       1     0.3023
```

Goodness of fit

But, Sample Size dependency?

$$\boldsymbol{\chi^2 = (n - 1)F_{ML}}$$

$\boldsymbol{n}$ sample size

50 samples: χ^2 = 1.78, DF=1, p = 0.182

p>0.05 good fit

100 samples: χ^2 = 3.60, DF=1, p = 0.058

p decrease with higher n

200 samples: χ^2 = 7.24, DF=1, p = 0.007

Goodness of fit

```
# SEM model in lavaan
sem_mod1 <- ` y1 ~ x1
              y2 ~  y1
`
sem.fit1 <- sem(sem_mod1, data=data1)

summary(sem.fit1, standardize = T,
       fit.measures=T) # fit measures
```

```
# results (fit.measures=T)
lavaan 0.6-9 ended normally after 23 iterations
...
 Model Test Baseline Model:

  Test statistic                              138.453
  Degrees of freedom                                3
  P-value                                       0.000

User Model versus Baseline Model:
  Comparative Fit Index (CFI)                   1.000
  Tucker-Lewis Index (TLI)                      0.999

Loglikelihood and Information Criteria:
  Loglikelihood user model (H0)                 7.399
  Loglikelihood unrestricted model (H1)         7.931
# continued on the next page
```

Goodness of fit

```
# SEM model in lavaan
sem_mod1 <- ‘ y1 ~ x1
              y2 ~  y1
‘
sem.fit1 <- sem(sem_mod1, data=data1)

summary(sem.fit1, standardize = T,
       fit.measures=T) # fit measures
```

```
# continued
...
 Akaike (AIC)                                  -6.798
  Bayesian (BIC)                                3.623
  Sample-size adjusted Bayesian (BIC)          -9.010

Root Mean Square Error of Approximation:

  RMSEA                                         0.025
  90 Percent confidence interval - lower        0.000
  90 Percent confidence interval - upper        0.268
  P-value RMSEA <= 0.05                         0.360

Standardized Root Mean Square Residual:

  SRMR                                          0.021
```

Goodness of fit

```
# call the fit measures in lavaan
fitMeasures(sem.fit1)
```

```
> fitMeasures(sem.fit1)
               npar                fmin               chisq                  df              pvalue 
              4.000               0.005               1.064               1.000               0.302 
     baseline.chisq         baseline.df     baseline.pvalue                 cfi                 tli 
            138.453               3.000               0.000               1.000               0.999 
               nnfi                 rfi                 nfi                pnfi                 ifi 
              0.999               0.977               0.992               0.331               1.000 
                rni                logl   unrestricted.logl                 aic                 bic 
              1.000               7.399               7.931              -6.798               3.623 
             ntotal                bic2               rmsea      rmsea.ci.lower      rmsea.ci.upper 
            100.000              -9.010               0.025               0.000               0.268 
       rmsea.pvalue                 rmr          rmr_nomean                srmr        srmr_bentler 
              0.360               0.003               0.003               0.021               0.021 
srmr_bentler_nomean                crmr         crmr_nomean          srmr_mplus   srmr_mplus_nomean 
              0.021               0.030               0.030               0.021               0.021 
              cn_05               cn_01                 gfi                agfi                pgfi 
            362.085             624.659               0.993               0.955               0.165 
```

Goodness of fit

Recommended
minimum of fit measures:

Measure	Name	Description	Cut-off for ‘good’ fit
χ^2	Model Chi-Square	Assess overall fit and the discrepancy between the observed and model-implied covariance matrices. Sensitive to sample size. H0: The model fits perfectly. (Present: χ^2, DF, p)	p-value > 0.05
RMSEA	Root Mean Square Error of Approximation	The square-root of the difference between the observed and model-implied covariance matrices. A parsimony-adjusted index. Values closer to 0 represent a good fit. RMSEA < 0.10 is generally ‘acceptable’ value. (Present: RMSEA, 90%CI, p_{RMSEA})	RMSEA < 0.08
CFI	Comparative Fit Index	Compares the fit of a model to the fit of a ‘null’ model (which estimates all variances but sets the covariances to 0). Low sensitivity to sample size.	CFI ≥ 0.90
SRMR	Standardized Root Mean Square Residual	The standardized difference between the observed and model-implied covariance matrices.	SRMR < 0.08

Principles and Practice of Structural Equation Modeling. Rex B. Kline. 2005.

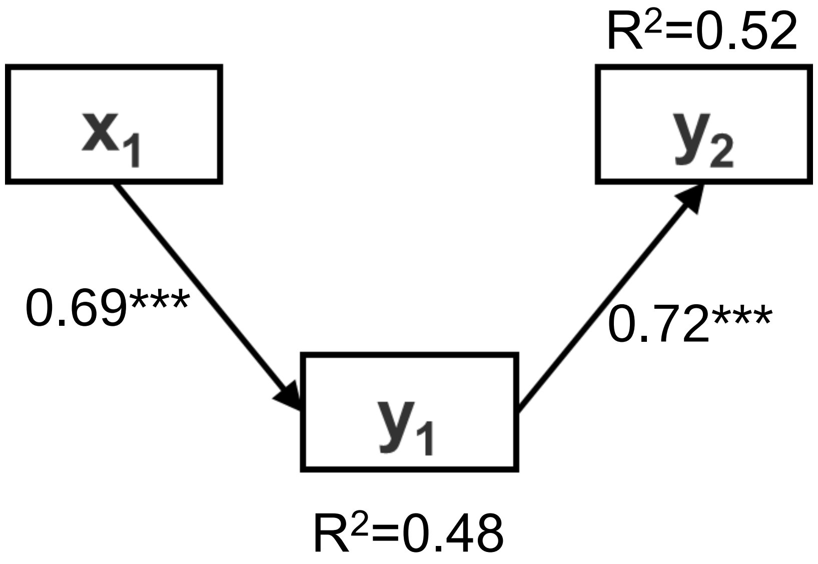
Goodness of fit

… and more:

Measure	Name	Description	Cut-off for ‘good’ fit
GFI	Goodness of Fit	GFI is the proportion of variance accounted for by the estimated population covariance. Analogous to R^2.	GFI ≥ 0.95
AGFI	Adjusted Goodness of Fit	AGFI favours parsimony.	AGFI ≥0.90
NFI	Normed-Fit Index	An NFI of 0.95, indicates that the model of interest improves the fit by 95% relative to the null model.	NFI ≥ 0.95
NNFI	Non-Normed-Fit Index	NNFI is preferable for smaller samples.	NNFI ≥ 0.95
TLI	Tucker Lewis index	Sometimes the NNFI is called the Tucker Lewis index (TLI)	

More comprehensive overview: http://davidakenny.net/cm/fit.htm

Goodness of fit



Indirect Effect of x1 on y2 = 0.496

Example of how to present the fit statistics:

χ^2 = 1.06, DF=1, n=100, p = 0.3

RMSEA=0.025, (CI = 0, 0.27) , p_{RMSEA}=0.36

CFI=1.00

SRMR=0.021

```
# plot the model
library(lavaanPlot)
lavaanPlot(model = sem.fit1,
                   coefs = TRUE, stand=TRUE,
                   stars = 'regress', # shows stars for regr coef
                   digits = 2) # limit the digits
```

Goodness of fit

Important points:

In SEM we assess overall model fit:

- Is your model adequate?
- Are you missing any paths?

When you are missing important paths:

- your parameter estimates may be incorrect
- your model is misspecified

Day 3 Task 2

California, USA.

Photos credit: USFS, and Jon Keeley, USGS

doi.org/10.1186/s42408-019-0041-0

doi.org/10.1071/WF07049

Postfire recovery of plant communities in California shrublands

Following fires, 90 plots were established 20x50m.

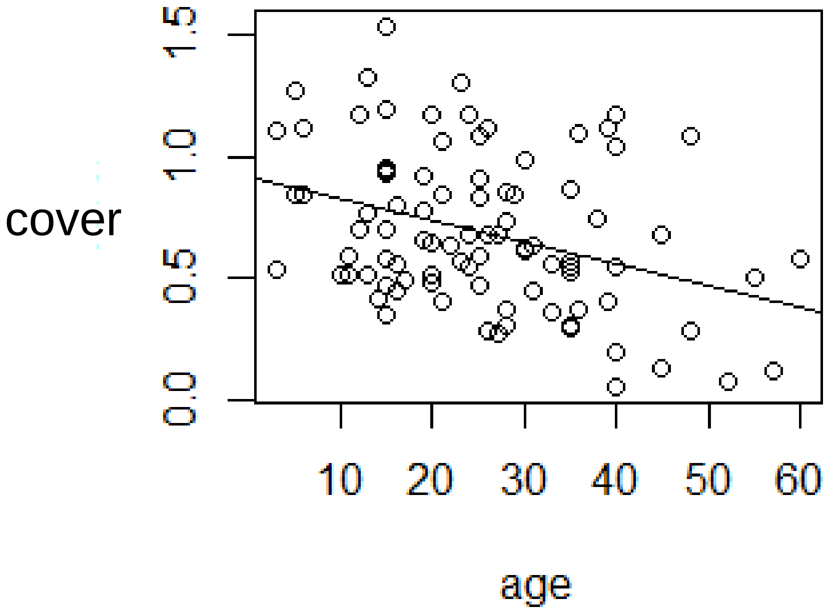
A number of measures were taken, including:

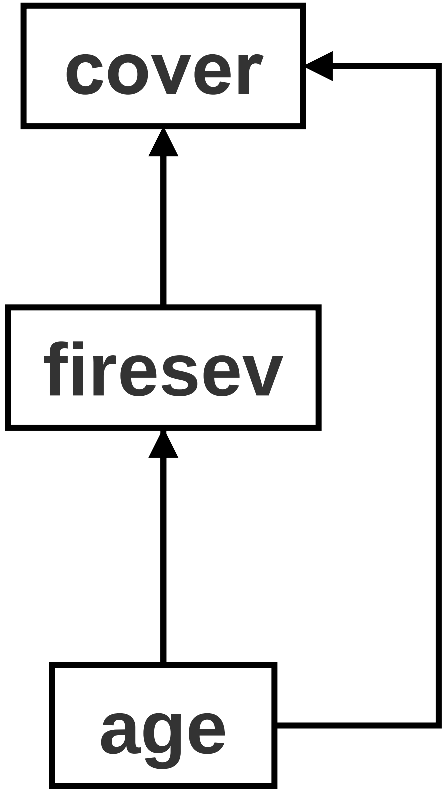
- Vegetation cover "**cover**"
- Age of stands that burned "**age**"
- Fire severity "**firesev**"

```
# Keeley data
library(piecewiseSEM)
data(keeley)
```

Data: Grace, J.B. and Keeley, J.E. 2006. A structural equation model analysis of postfire plant diversity in California shrublands. Ecological Applications 16:503-514

Day 3 Task 2





Data: Grace, J.B. and Keeley, J.E. 2006. A structural equation model analysis of postfire plant diversity in California shrublands. Ecological Applications 16:503-514

Day 3 Task 2

For the model on **Fig. 1**:

1. Check what is the model identifability status:

- identified, underidentified, or overidentified model?
- saturated or unsaturated model?
- recursive or non-recursive?

2. Assess if the sample size is enough to fit this model?

3. Fit the model in lavaan and get the path coefficients.

4. Get the fit indices and assess goodness of fit.

5. Test if link from “age” to “cover” is missing (see **Fig 2**)

For this use a Likelihood Ratio Test (χ^2 – difference test)

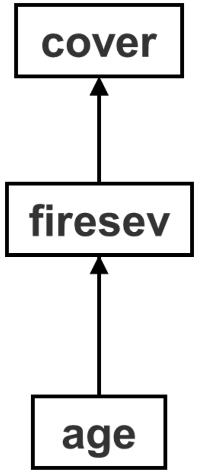


Fig. 1

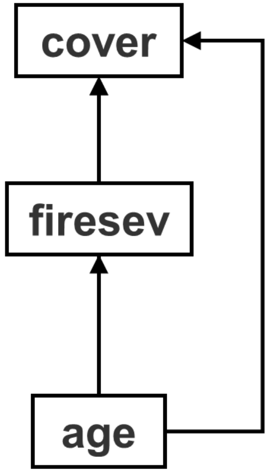


Fig. 2