Introduction to structural equation modeling and mixed models in R

Day 5

Oksana Buzhdygan

oksana.buzh@fu-berlin.de

24.02.2023

Day 5 – Part 1

Day 5 – Part 1

Outline

- Basics of SEM

  - ✓ From regression to SEM
  - ✓ SEM history. SEM in natural sciences.
  - ✓ SEM workflow process. Where do I start?
  - ✓ First impression of ‘lavaan’

Day 5 – Part 1

Outline

- Basics of SEM

  - ✓ **From regression to SEM**
  - ✓ SEM history. SEM in natural sciences.
  - ✓ SEM workflow process. Where do I start?
  - ✓ First impression of ‘lavaan’

From regression to SEM

Aim of regression model:

- (How) does variable $\boldsymbol{x}$ impact variable $\boldsymbol{y}$?
- Can we better predict values for variable $\boldsymbol{y}$, if we account for variable $\boldsymbol{x}$?

$$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{x}$$

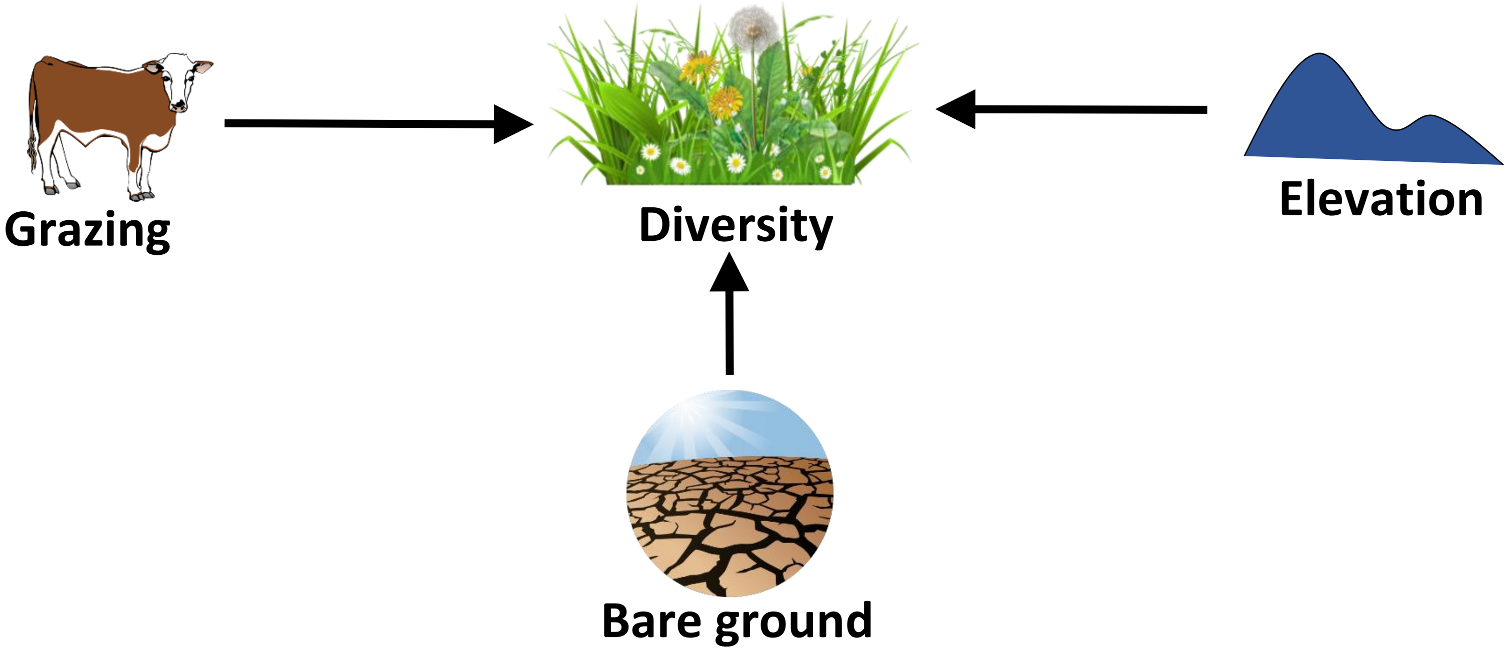
From regression to SEM

Aim of regression model:

- (How) does variable x impact variable y?
- Can we better predict values for variable y, if we account for variable x?

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Buzhdygan, *et al.* 2021 *PLosONE*



Univariate relationships

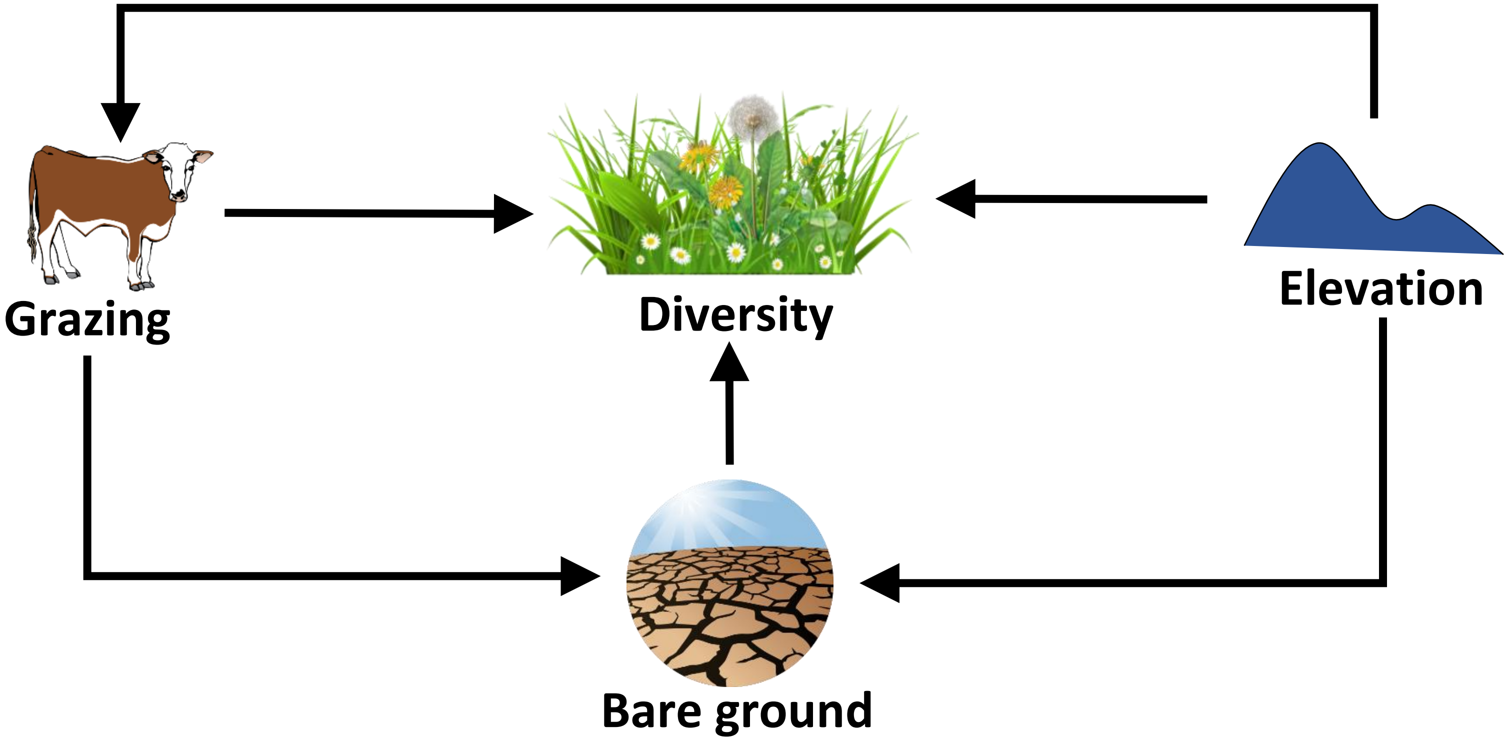
- involve response variable explained by a set of predictors

From regression to SEM

SEM:

- Tests **systems of relationships** (multivariate) rather than a dependant variable and a set of predictors (univariate relationships)

Buzhdygan, *et al.* 2021 *PLosONE*



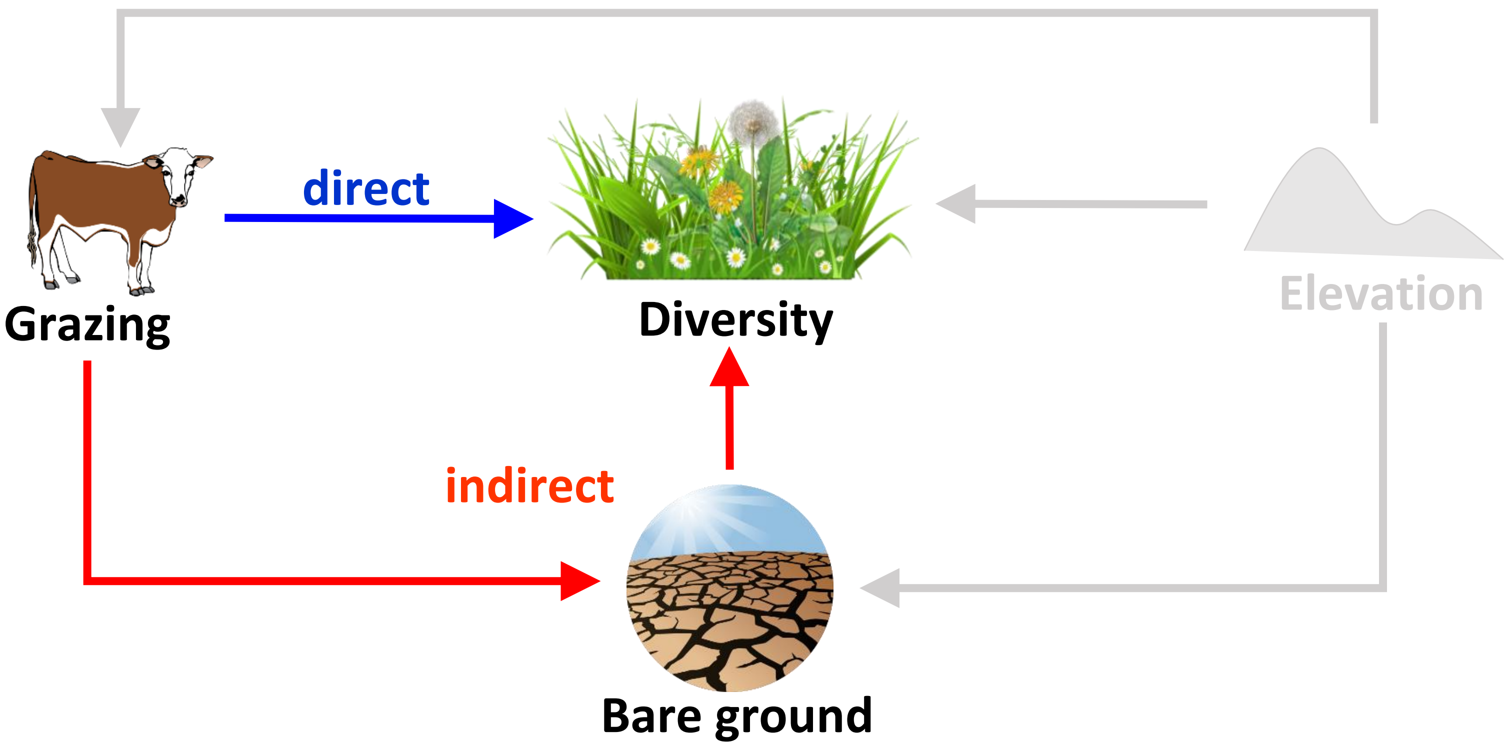
Multivariate relationships

- involve simultaneous influences and responses

From regression to SEM

SEM:

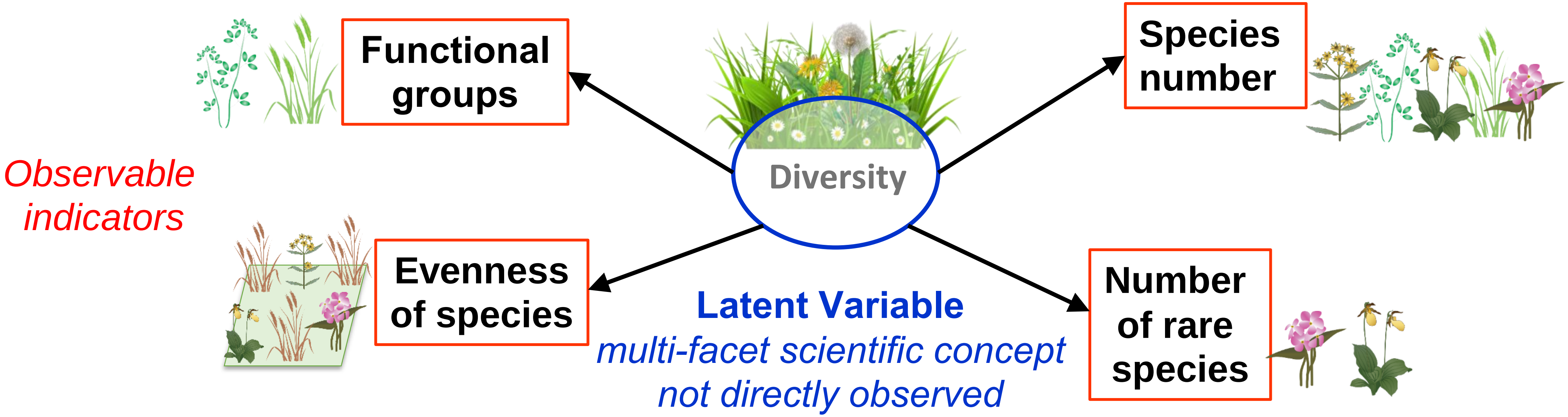
- Tests **systems of relationships** (multivariate relationships)
- Allows testing **indirect and direct effects** of variables on other variables



From regression to SEM

SEM:

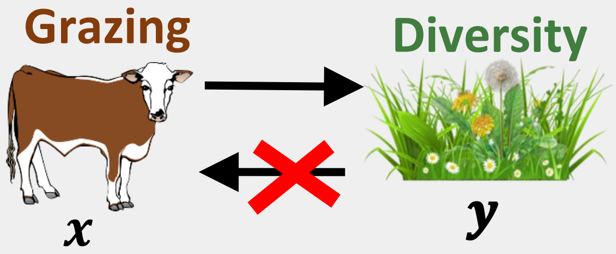
- Tests **systems of relationships** (multivariate relationships)
- Allows testing **indirect and direct effects** of variables on other variables
- Involves complex, multi-faceted **constructs,** approximated by observed indicators



What is SEM?

Structural

There is
hypothesized underlying <u>structure</u>
to study system
(a <u>cause and an effect</u>)…



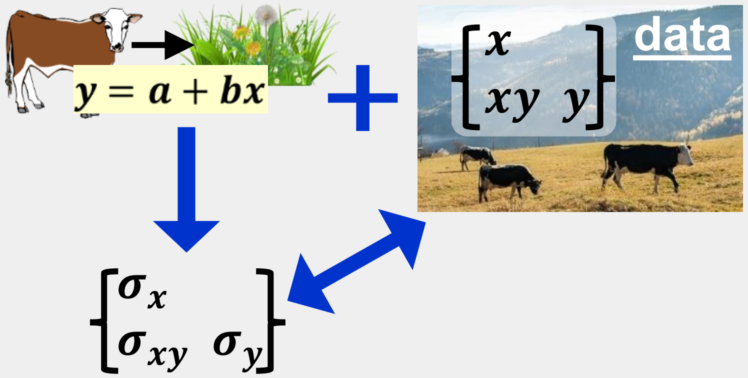
Equation

…that can be translated to
a series of
mathematical <u>equations</u>…

$$y = a + bx$$

Modelling

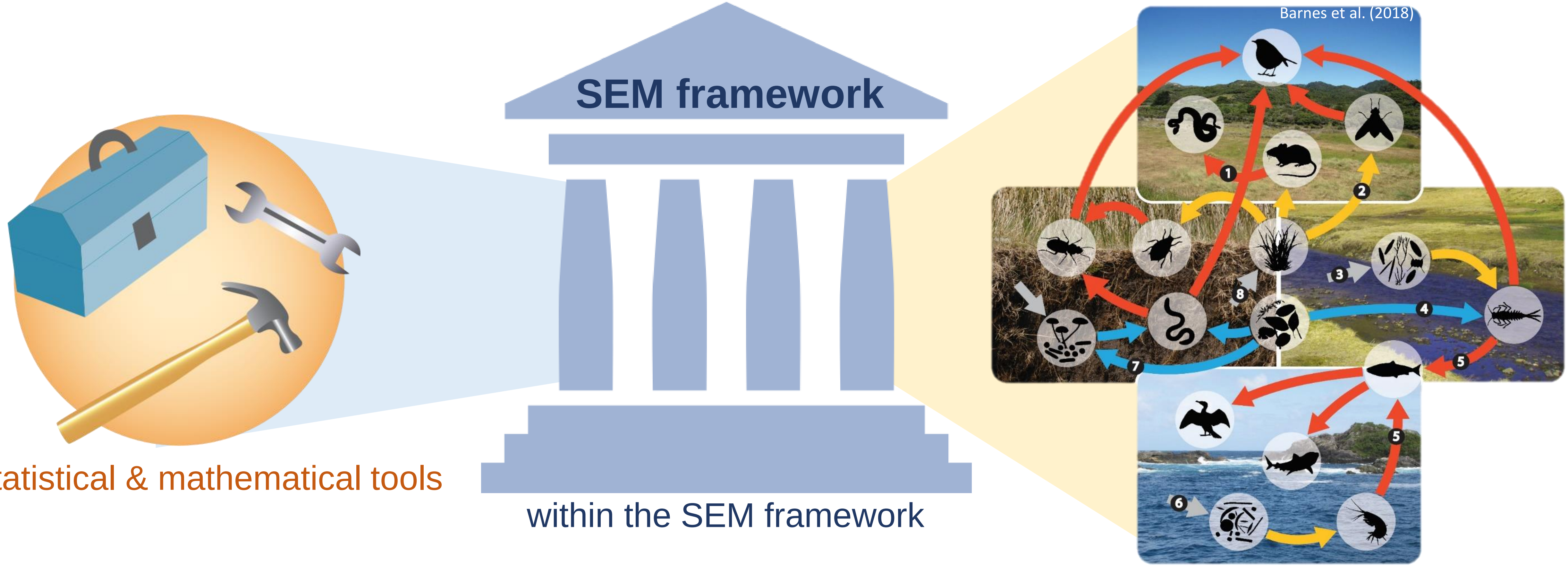
……which can be <u>modelled against data</u>
to support or refuse
the proposed <u>structure</u>



What is SEM?

SEM is a framework

- **not one** statistical method or technique



Statistical & mathematical tools

within the SEM framework

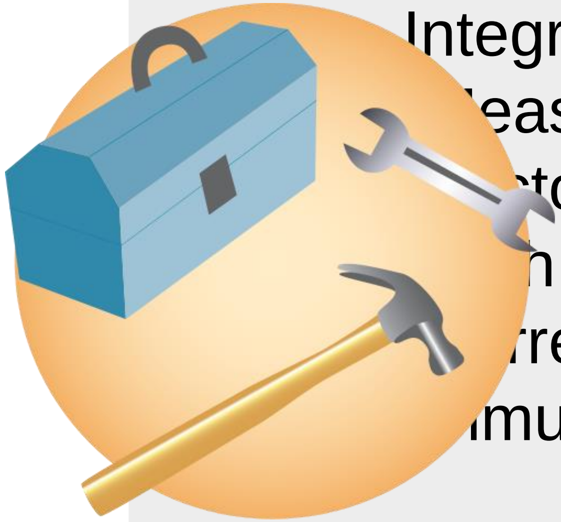
to understand multiple processes in complex systems

What is SEM?

SEM is a framework

- **not one** statistical method or technique

SEM framework



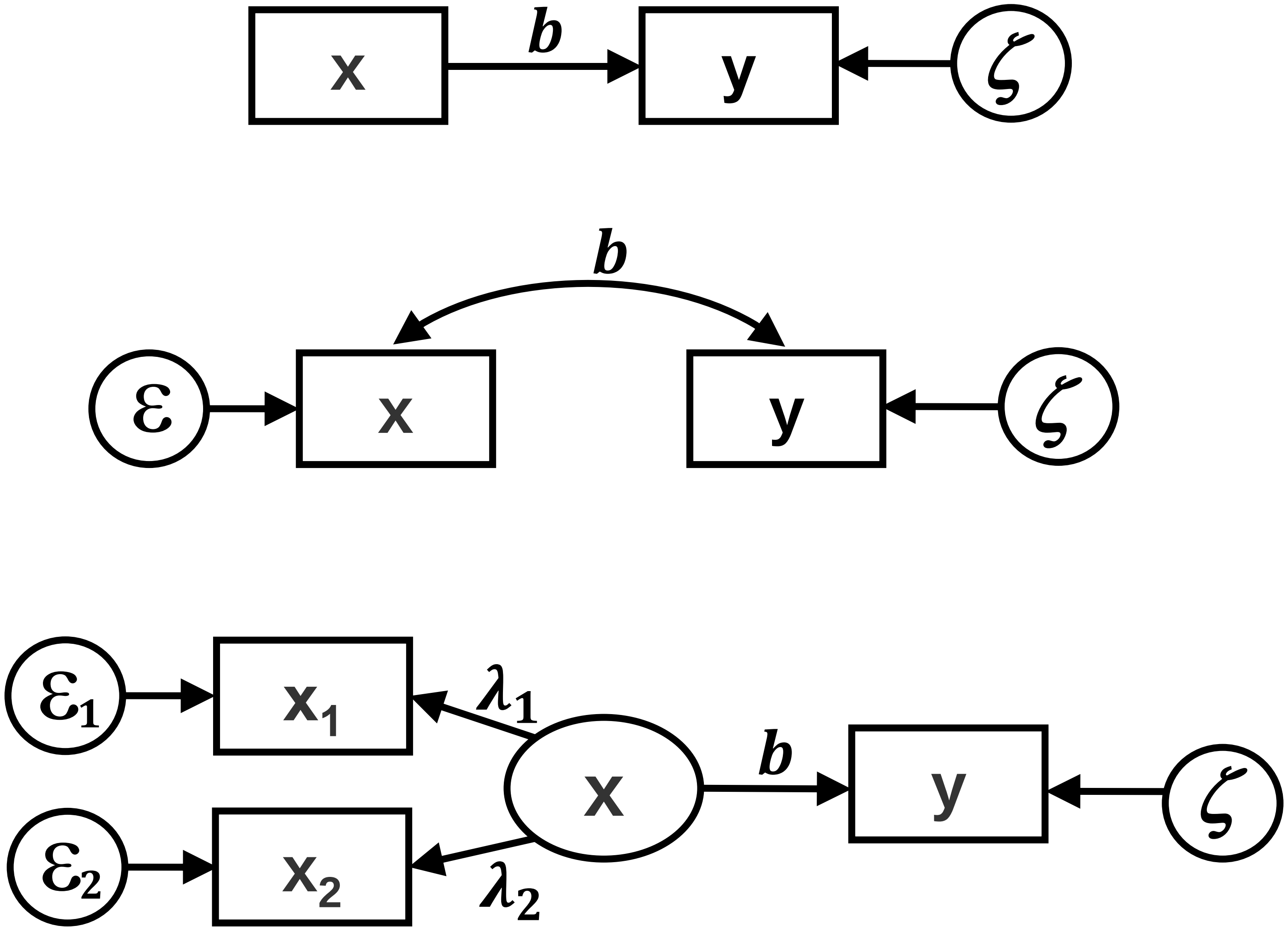
Integrates:
- Measurement theory
- Factor analysis
- Path analysis
- Correlation & Regression
- Simultaneous equations

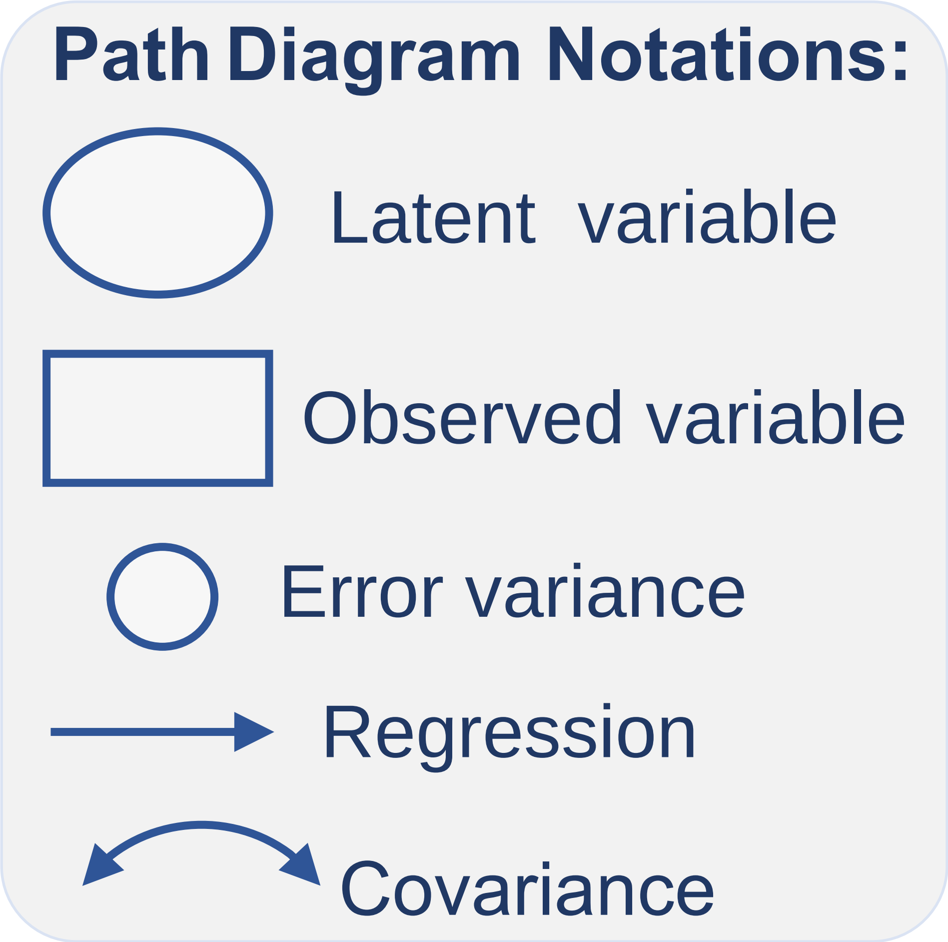
Statistic

within the SEM framework

to understand multiple processes in complex systems

SEM is Graphical Modelling





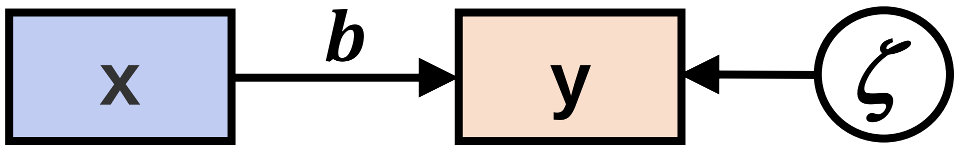
What is SEM?

SEM is Graphical Modelling

Equation form:

$$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{bx} + \zeta$$

Graphical form (Path Diagram):



Exogenous variable

- have arrows directed only ***out of it*** (i.e., no arrows going into it)

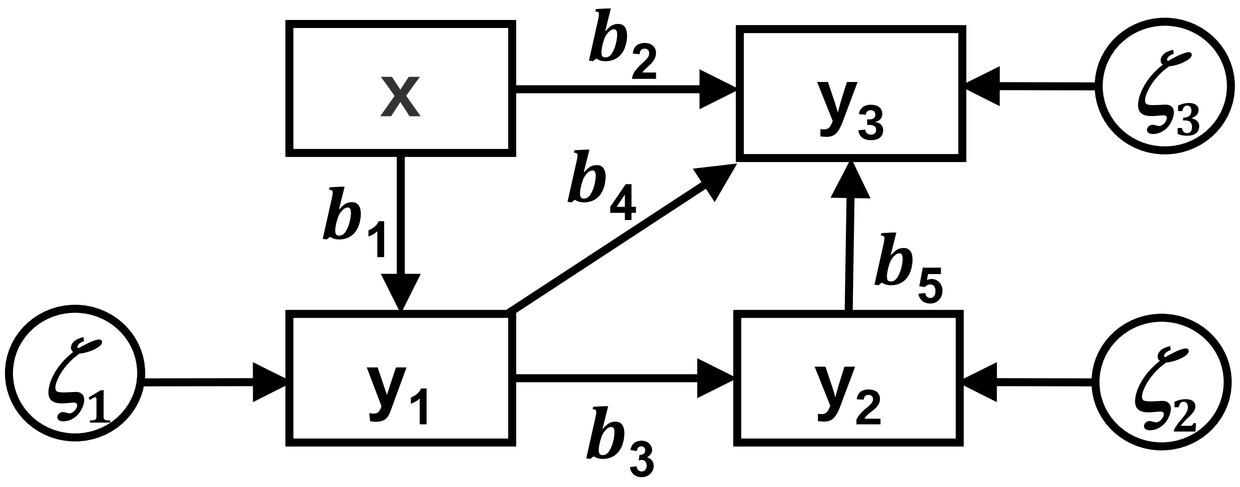
Endogenous variable

- for which arrows are also directed ***into it***
- can also have arrows directing out of it, but **must be predicted** at the same time

What is SEM?

SEM as complex networks

Path Diagram:



Corresponding equations:

$$y_1 = b_1 x + \zeta_1$$
$$y_2 = b_3 y_1 + \zeta_2$$
$$y_3 = b_2 x + b_4 y_1 + b_5 y_2 + \zeta_3$$

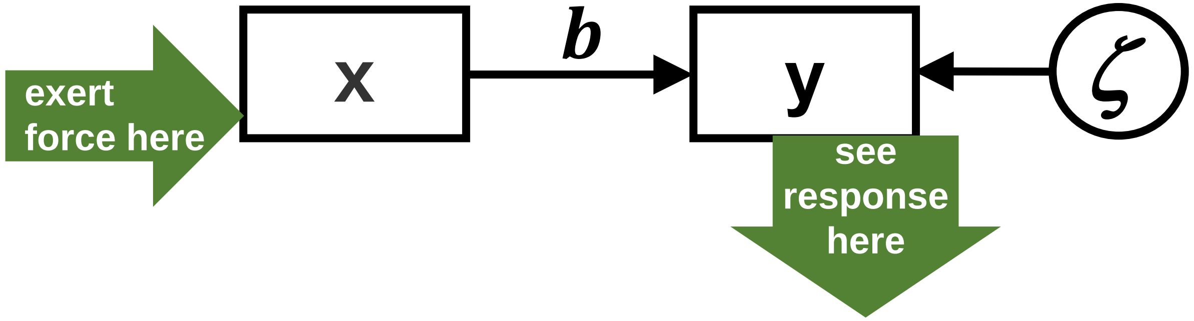
SEM addresses

- multivariate relationships (simultaneous influences and responses)
- mechanical understanding (direct & indirect effects)

What is SEM?

Implies direction of relationships

Graphical form **with causality:**



Cause-Effect Relations

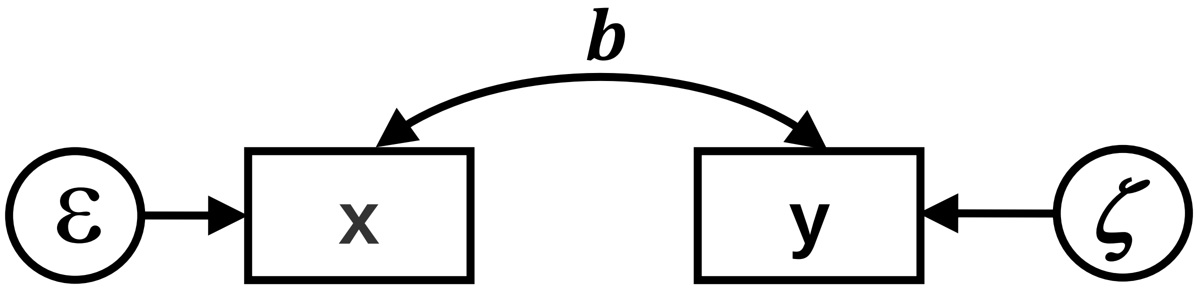
x causes **y**

(if manipulation of **x**
leads to a response in **y**)

"correlation does not imply causation"

R.A. Fisher

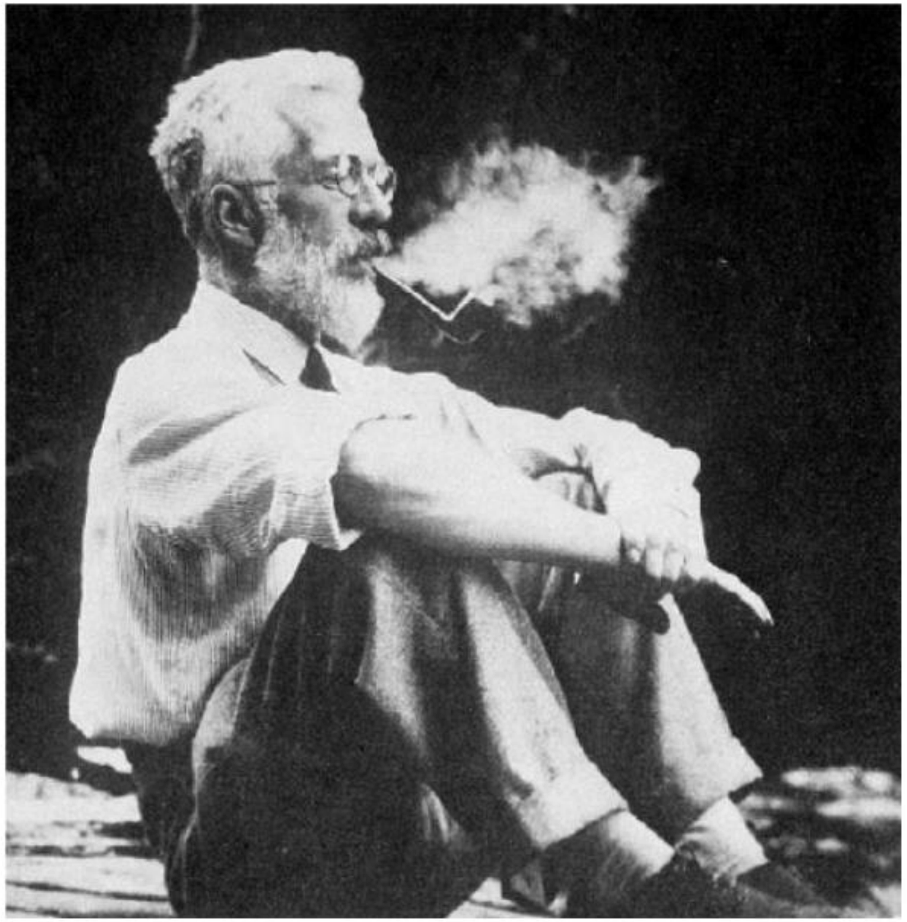
Graphical form **without causality:**



Correlation

y and **x** tend to be observed
at the same time

Causality *vs* Correlation



R.A. Fisher smoking a pipe, 1956.

Photo: [0.1016/j.endeavour.2004.02.003]

Smoking

Lung Cancer

"correlation does not imply causation"

R.A. Fisher

Smoking → Lung Cancer

Causality *vs* Correlation

Ambulance cars tend to be observed in traffic jams

Photo: https://www.aninews.in

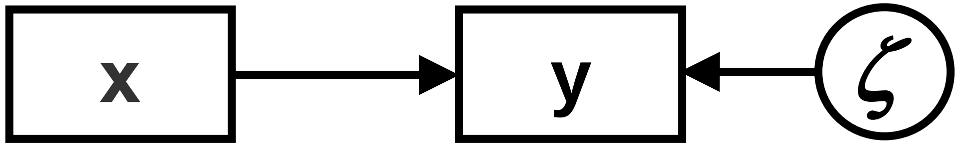
Ambulance ↔ Car accidents in traffic jams

Ambulance ← Car accidents in traffic jams

Ambulance → Car accidents in traffic jams

What is SEM?

SEM is not a method for discovering causes



- SEM results is **not a proof of causal claims**.
- SEM relies upon the **causal assumptions made by us**, when building the model.

We assume that **x** causes **y** from:

Not from SEM

!

- Research design
- Prior observation
- Prior statistical models
- Prior experimentation
- Logical arguments
- Some or all of the above

The **credibility of the SEM** depends on the **credibility of the causal assumptions** made by the **researcher**

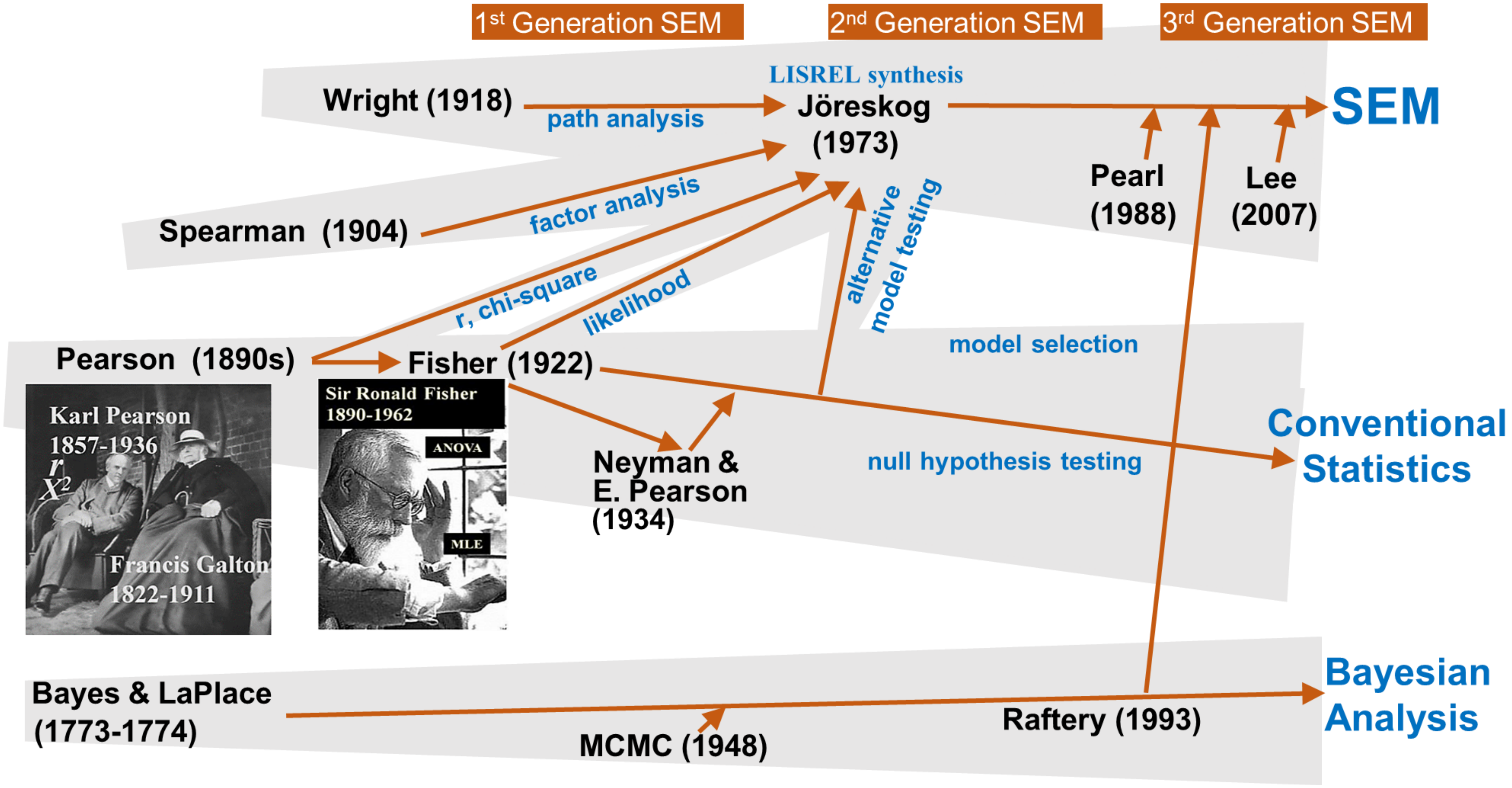
Day 5 – Part 1

Outline

- Basics of SEM

  - ✓ From regression to SEM
  - ✓ **SEM history. SEM in natural sciences.**
  - ✓ SEM workflow process. Where do I start?
  - ✓ First impression of ‘lavaan’

SEM History



adapted from Grace, 2006

SEM in Ecology and Evolution

Jim B. Grace www.structuralequations.org

SEM adaptation to the needs of ecology and evolutionary biology

- Grace (2010) Structural Equation Modeling for Observational Studies. *Journal of Wildlife Management*, 72:14-22
- Grace et al. (2010) On the specification of structural equation models for ecological systems. *Ecological Monographs*, 80, 67-87.
- Grace, Bollen (2005) Interpreting the Results from Multiple Regression and Structural Equation Models. *Bulletin of the Ecological Society of America*, 86, 283-295.
- Grace (2015) Taking a systems approach to ecological systems. *Journal of Vegetation Science* 26, 1025-1027.

Jon Lefcheck https://jslefche.github.io/sem_book

Tools for SEM in R:
piecewiseSEM

Jarrett Byrnes https://jebyrnes.github.io/semclass

Tools for SEM in R:
sem.additions
collaborate on *sem* and *lavaan*

Learn More about SEM

- Grace (2006) Structural Equation Modeling and Natural Systems. Cambridge Univ. Press.

- Shipley (2000) Cause and Correlation in Biology. Cambridge Univ. Press.
- Kline (2012) Principles and Practice of Structural Equation Modeling. (3rd Edition) Guilford Press.

- Bollen (1989) Structural Equations with Latent Variables. John Wiley and Sons.

- Hoyle (2012) Handbook of Structural Equation Modeling. Guilford Press.

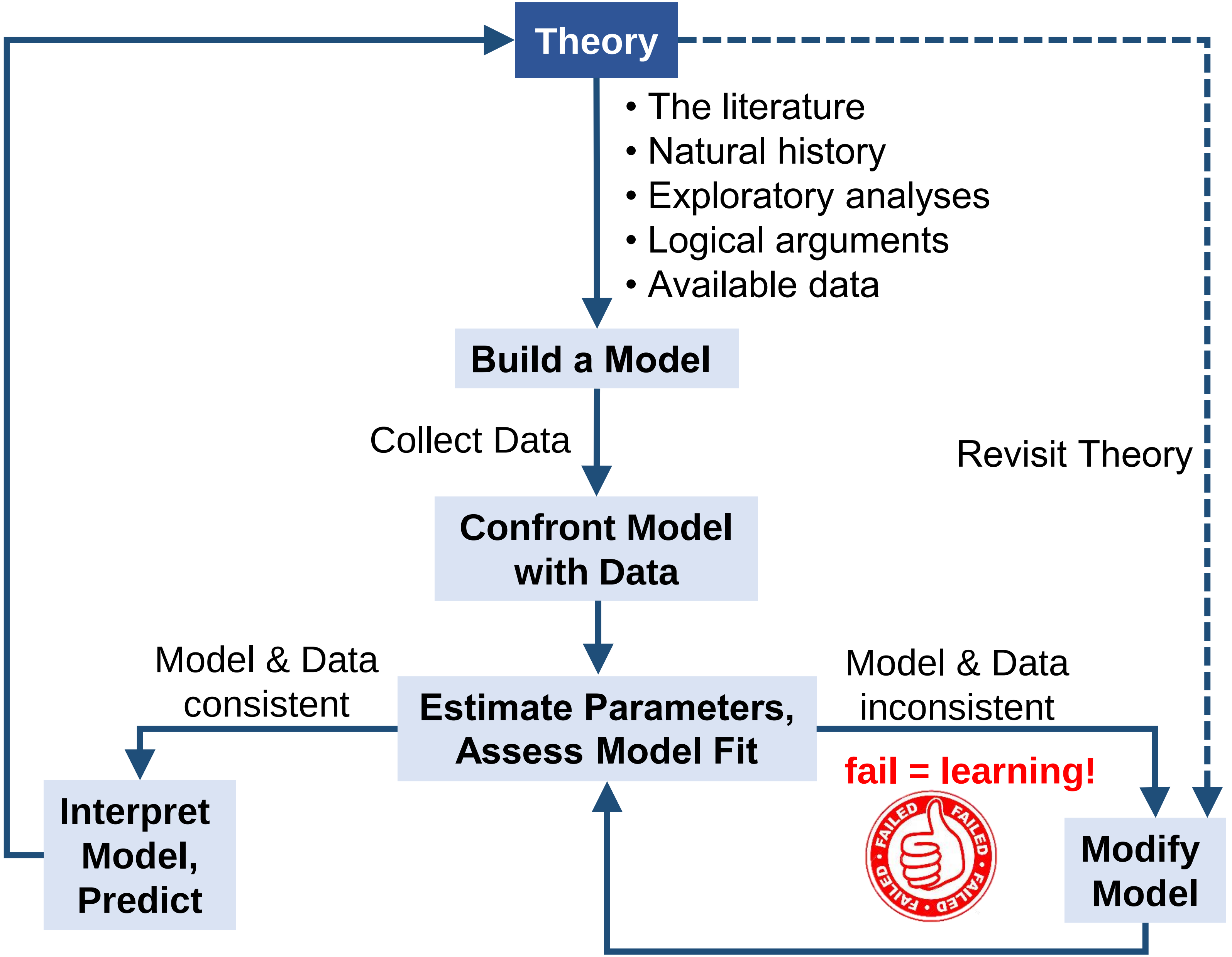
Day 5 – Part 1

Outline

- Basics of SEM
  - ✓ From regression to SEM
  - ✓ SEM history. SEM in natural sciences.
  - ✓ **SEM workflow process. Where do I start?**
  - ✓ First impression of ‘lavaan’

Where to start?

SEM workflow process



- multi-step process to build knowledge through sequential learning
- fail implies learning from your data and through revisiting theory

Where to start?

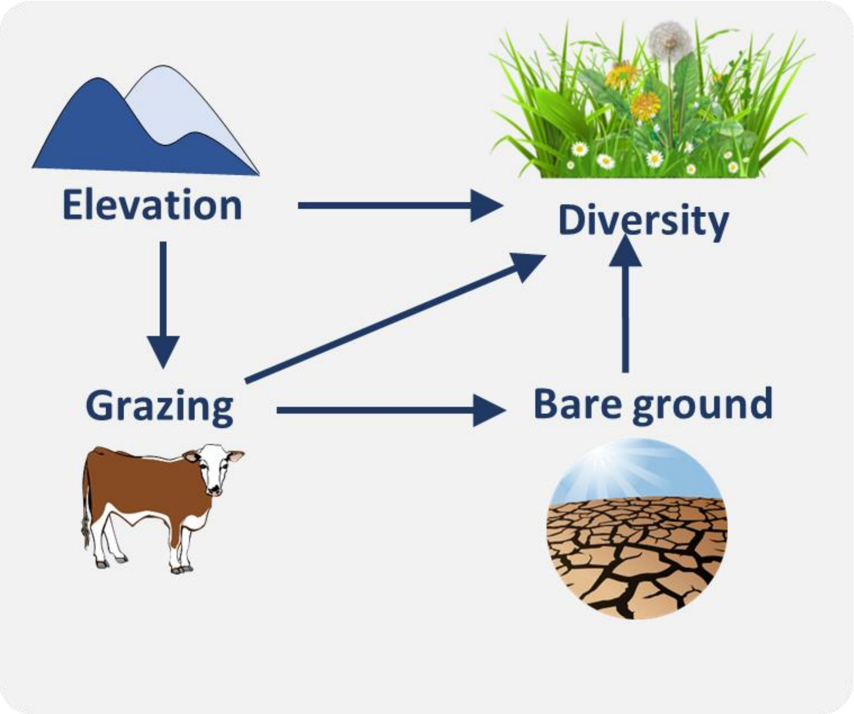
Theory

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

Build a Model

Collect Data

Confront Model with Data

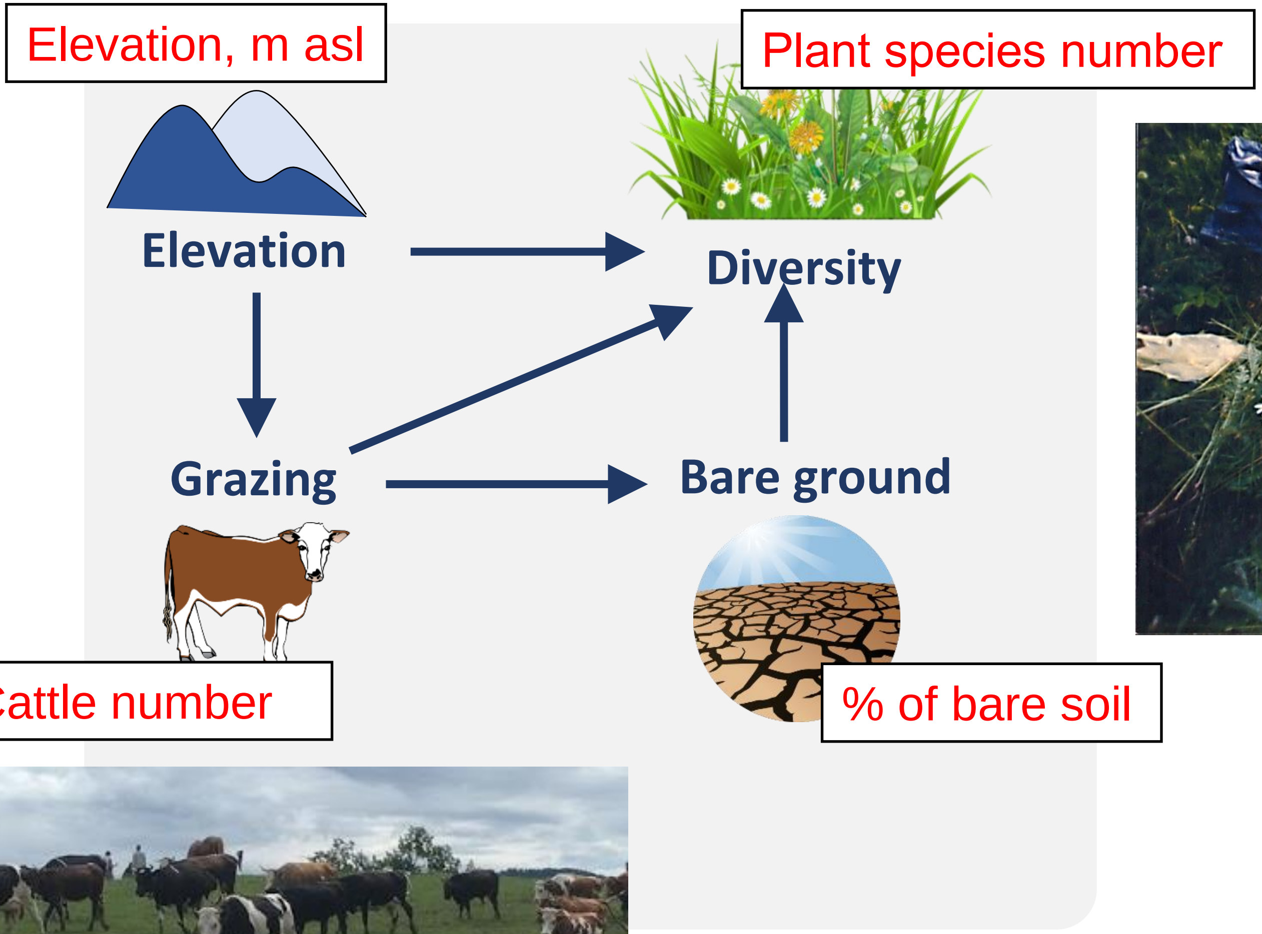


Buzhdygan, *et al.* 2020 *PLosONE*

Study area

Map: Google Earth

Collect Data and Parametrise Model



Where to start?

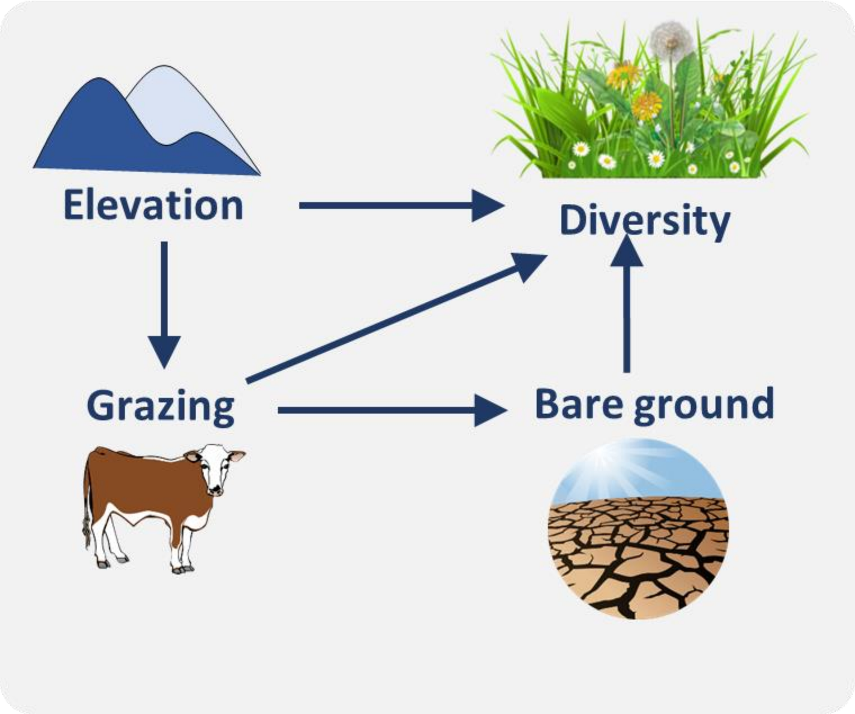
Theory

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

Build a Model

Collect Data

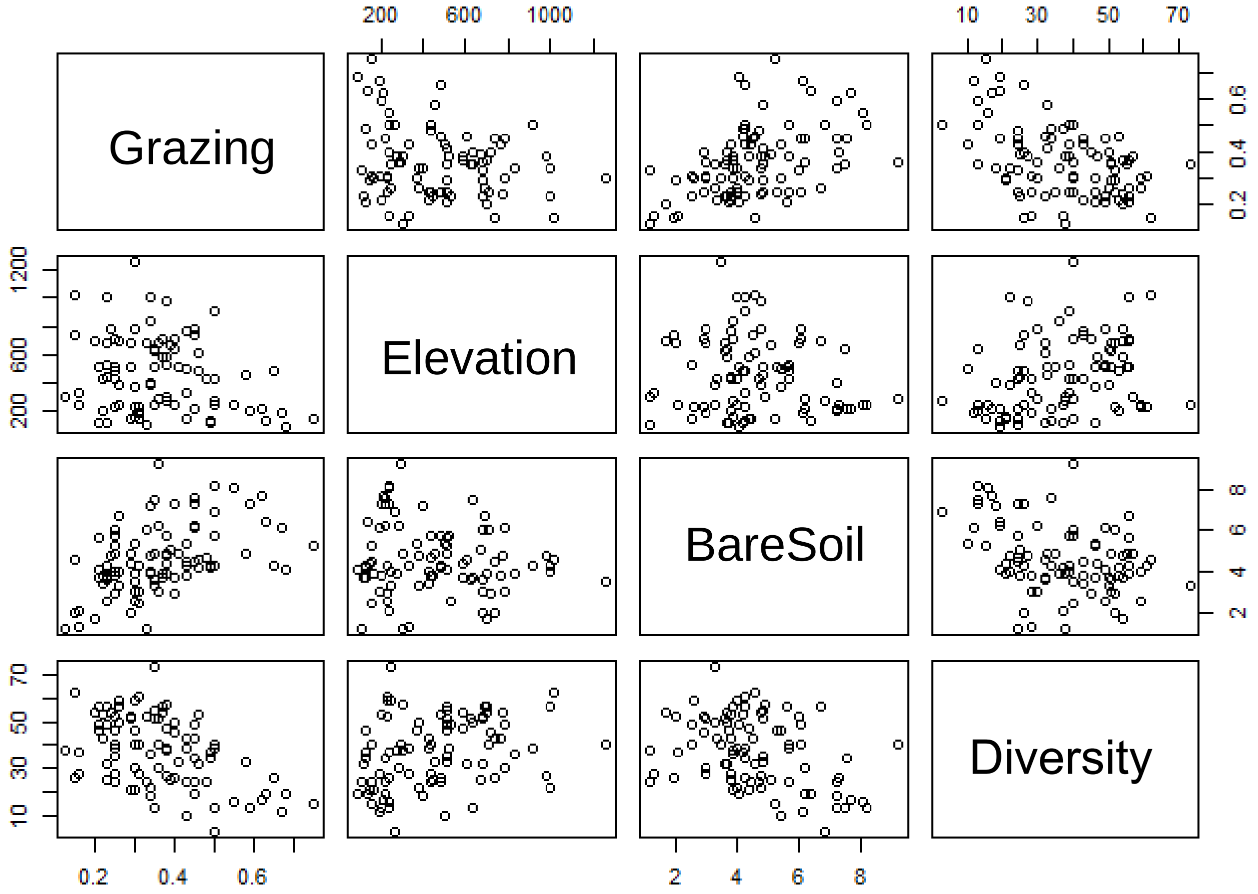
Confront Model with Data

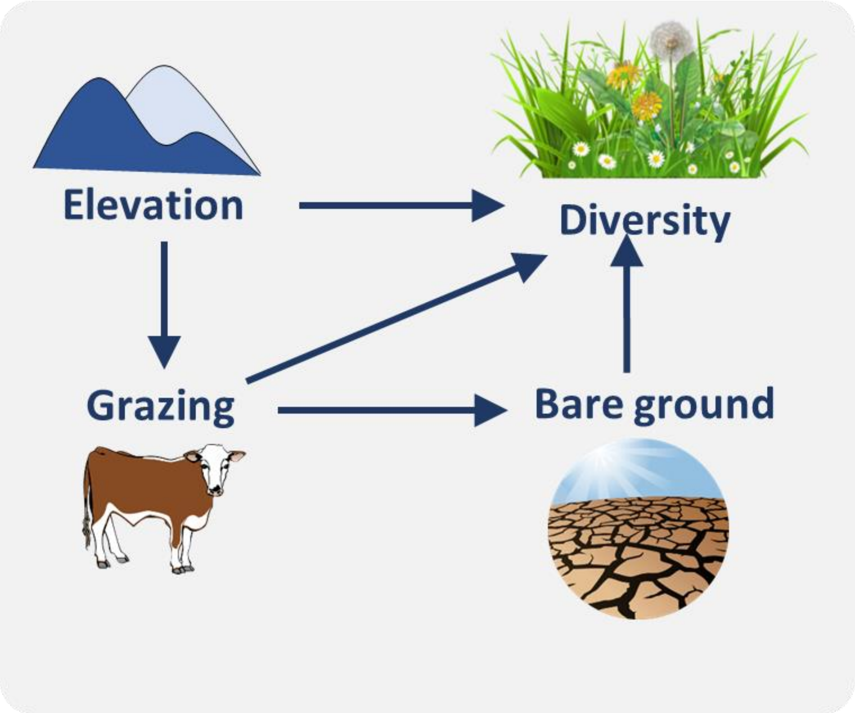


Buzhdygan, *et al.* 2020 *PLosONE*

Where to start?

```
data <- read.csv("Grassl_data.csv")
names(data)

# view the data
pairs(data)
```





Buzhdygan, *et al.* 2020 *PLosONE*

Day 5 – Part 1

Outline

- Basics of SEM

  ✓ From regression to SEM

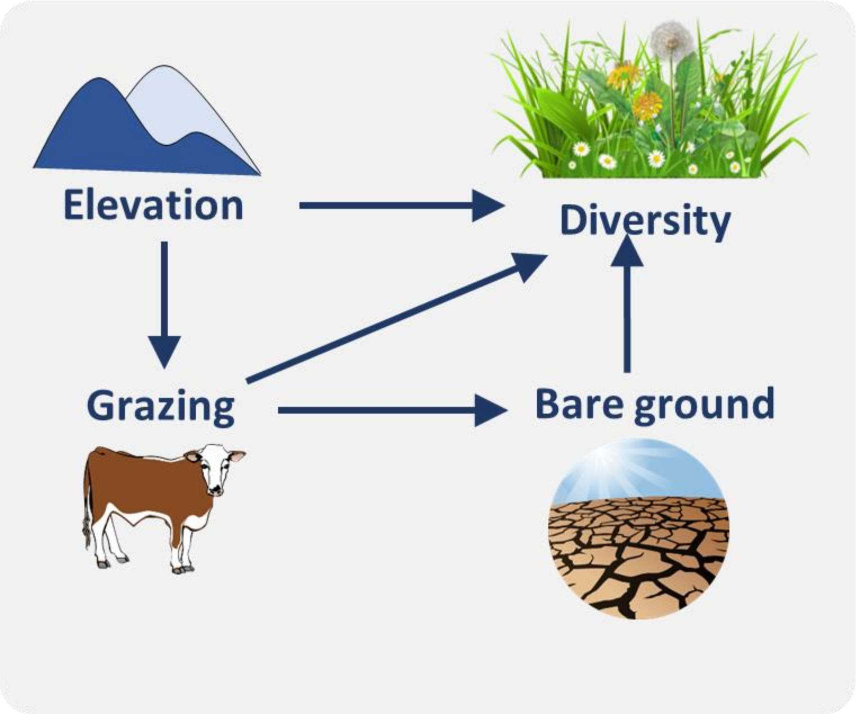
  ✓ SEM history. SEM in natural sciences.

  ✓ SEM workflow process. Where do I start?

  ✓ **First impression of ‘lavaan’**

First impression of ‘lavaan’

```
# Coding SEM
library(lavaan)

# Specify model structure

sem_mod <- '
    Grazing  ~  Elevation
    BareSoil  ~  Grazing
    Diversity ~ Elevation + Grazing + BareSoil
'
```

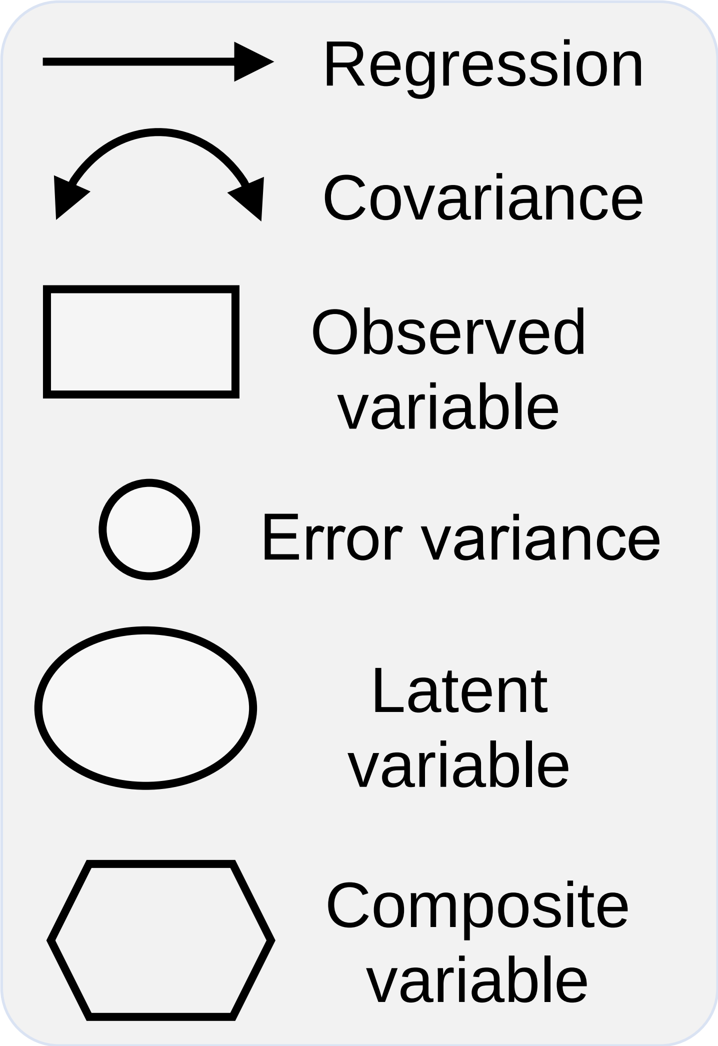


Buzhdygan, *et al.* 2020 *PLosONE*

First impression of ‘lavaan’

```
# Coding SEM
library(lavaan)

# Specify model structure

sem_mod <- '
    Grazing  ~  Elevation
    BareSoil  ~  Grazing
    Diversity ~ Elevation + Grazing + BareSoil
'
```

Path Diagram Notations:



Specification operators in ‘lavaan’

formula type	operator	meaning
Regression	~	“regressed on”
Correlation	~~	"correlated with"
Intercept	~ 1	"estimates intercept"
Latent variable	=~	"is measured by"
Composite	<~	"is caused by"

First impression of ‘lavaan’

lavaan

latent variable analysis

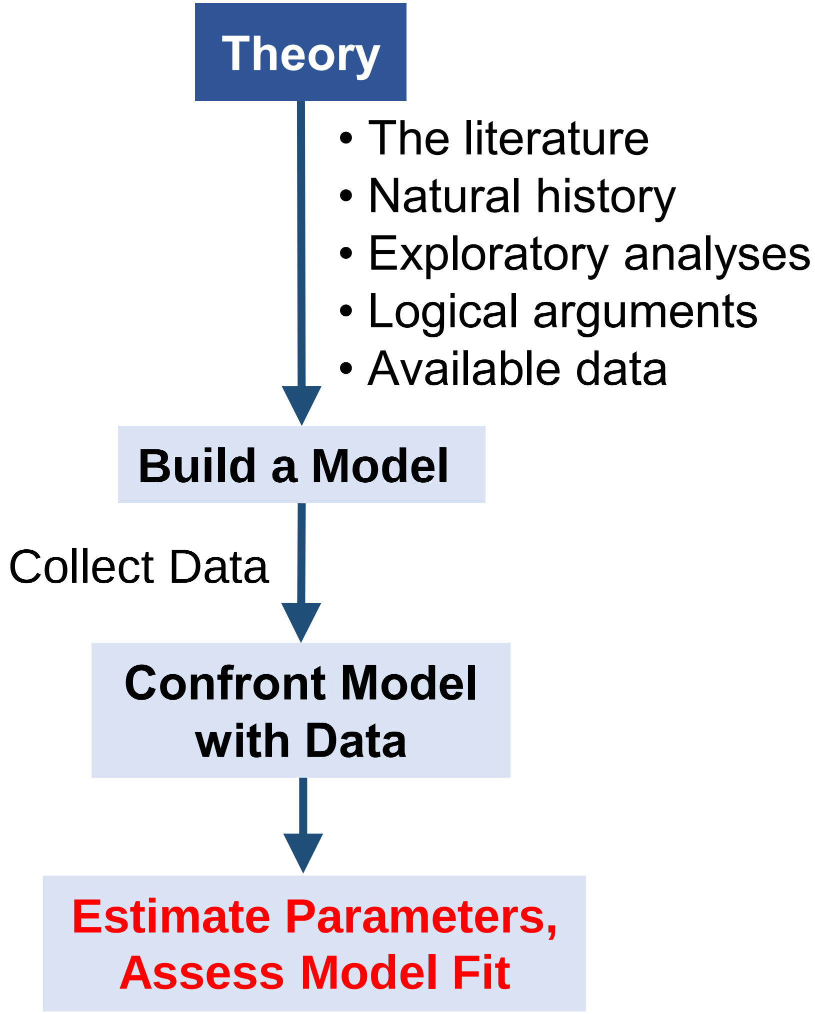
https://lavaan.ugent.be

What is lavaan?

- Stands for **LA**tent **VA**riable **AN**alaysis
- Written by Yves Roseel in 2010
- Currently in version 6
- Uses R lm syntax

First impression of ‘lavaan’

```
# Coding SEM
library(lavaan)

# Specify model structure

sem_mod <- '
    Grazing  ~  Elevation
    BareSoil  ~  Grazing
    Diversity ~ Elevation + Grazing + BareSoil
`


# Estimate parameters, assess model fit
sem.fit <- sem(sem_mod, data=data)

# extract results
summary(sem.fit)
```



First impression of ‘lavaan’

When you fit the model

```
# Error about data scales
Warning message:
In lav_data_full(data = data, group = group, cluster = cluster,   :
  lavaan WARNING: some observed variances are (at least) a factor 1000 times larger than others; use varTable(fit) to investigate
```

First impression of 'lavaan'

```
# Call the model-implied covariance matrix

lavInspect(sem.fit, "obs")$cov

>
             Grazng     BareSl      Dvrsty        Elevtn
Grazing       0.017
BareSoil      0.102      2.685
Diversity    -0.904     -8.969     217.200
Elevation    -8.439    -55.722     1125.614     65289.346
```

- The covariance matrix is Ok, - there are no data problems.
- This is a likelihood algorithm problem – we can ignore the WARNING
- If you are worried about it, rescale data and see if answers change

First impression of 'lavaan'

```
# Check the data scales

varTable(sem.fit)


       name   idx nobs   type  exo  user    mean       var
1   Grazing    1    90 numeric   0     0   0.361     0.017
2  BareSoil    3    90 numeric   0     0   4.587     2.716
3 Diversity    4    90 numeric   0     0  37.022   219.640
4 Elevation    2    90 numeric   1     0 456.856 66022.934



 # Transform the data: recode vars to roughly same scale

 data$Diversity    <- data$Diversity/10

 data$Elevation <- data$Elevation/100


 # Repeat model estimation using transformed data
```

First impression of ‘lavaan’

```
# extract results
summary(sem.fit)

lavaan 0.6-9 ended normally after 55 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         8

  Number of observations                            90

Model Test User Model:

  Test statistic                                 0.021
  Degrees of freedom                                 1
  P-value (Chi-square)                           0.886
```

continued

Assessed model fit

More soon!
(in the part 3)

First impression of 'lavaan'

```
# Results from SEM model

> summary(sem.fit , standardize = T)

Parameter Estimates:
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  Grazing ~
    Elevation        -0.000    0.000   -2.475    0.013   -0.000   -0.252
  BareSoil ~
    Grazing           5.963    1.161    5.136    0.000    5.963    0.476
  Diversity ~
    Elevation         0.011    0.005    2.062    0.039    0.011    0.190
    Grazing         -37.259   11.739   -3.174    0.002  -37.259   -0.331
    BareSoil         -1.696    0.913   -1.856    0.063   -1.696   -0.189
```

Standardised coefficients for latent variables

Standardised coefficients for all variables

Raw unstandardized coefficients

standard errors

Wald statistic

Probability of a z this big by chance

First impression of 'lavaan'

```
# Results from SEM model

> summary(sem.fit , standardize = T)

Parameter Estimates:
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  Grazing ~
    Elevation        -0.000    0.000   -2.475    0.013   -0.000   -0.252
  BareSoil ~
    Grazing           5.963    1.161    5.136    0.000    5.963    0.476
  Diversity ~
    Elevation         0.011    0.005    2.062    0.039    0.011    0.190
    Grazing         -37.259   11.739   -3.174    0.002  -37.259   -0.331
    BareSoil         -1.696    0.913   -1.856    0.063   -1.696   -0.189
```

Standardised coefficients for latent variables

Standardised coefficients for all variables

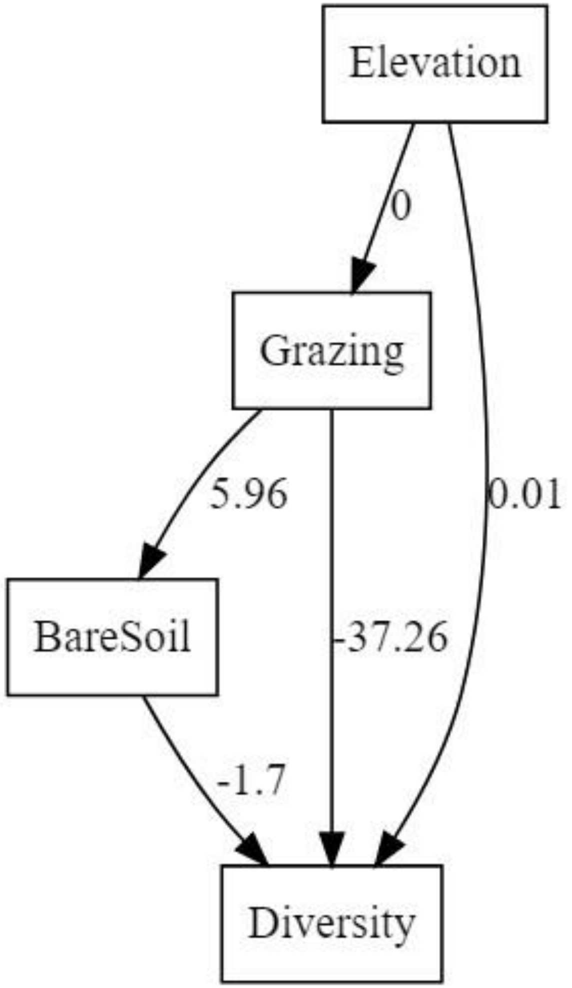
Raw unstandardized coefficients

standard errors

Wald statistic

Probability of a z this big by chance

```
library(lavaanPlot)
lavaanPlot(model = sem.fit,
coefs = TRUE)
```

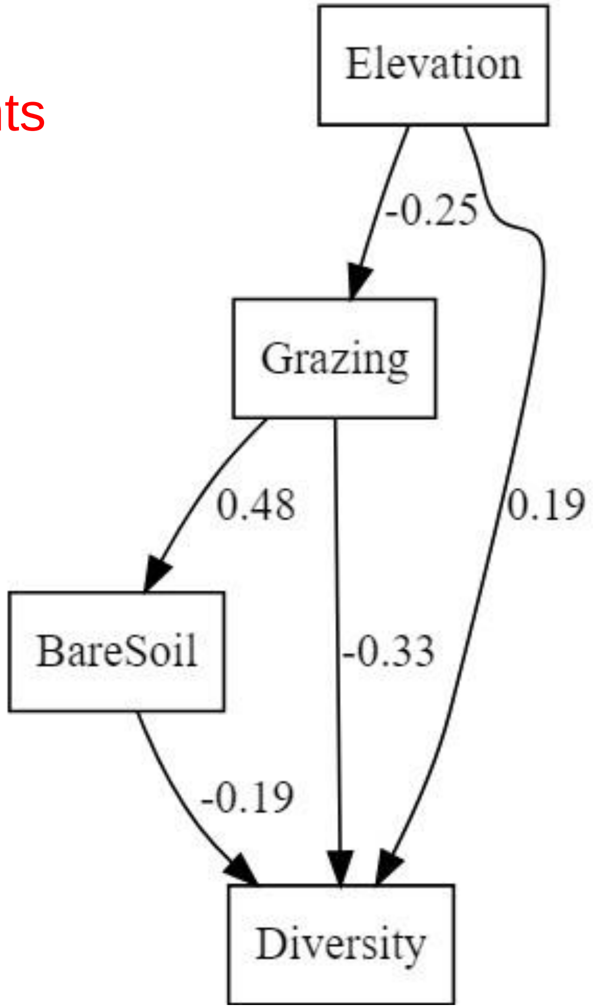


First impression of ‘lavaan’

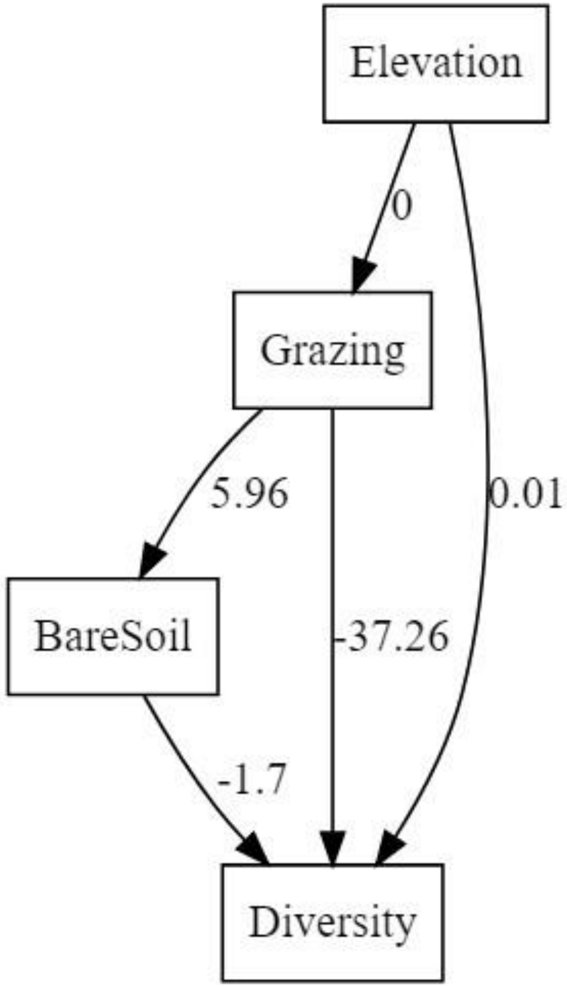
```
lavaanPlot(model = sem.fit,
coefs = TRUE, stand=TRUE)
```

Standardised coefficients

- comparable across the entire model




```
lavaanPlot(model = sem.fit,
coefs = TRUE)
```

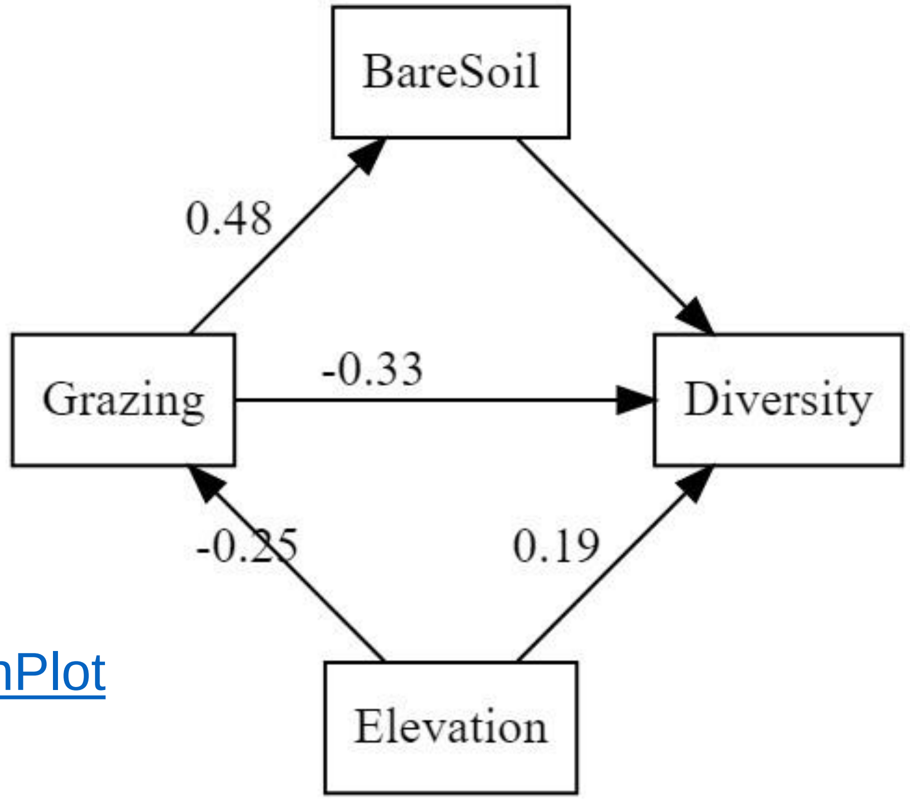


First impression of ‘lavaan’

```
lavaanPlot(model = sem.fit,
           coefs = TRUE, stand=TRUE,
graph_options = list(layout = "circo"),
                       sig = 0.05)
```

changes layout

only shows coefficients p≤0.05



See more:
https://cran.r-project.org/web/packages/lavaanPlot

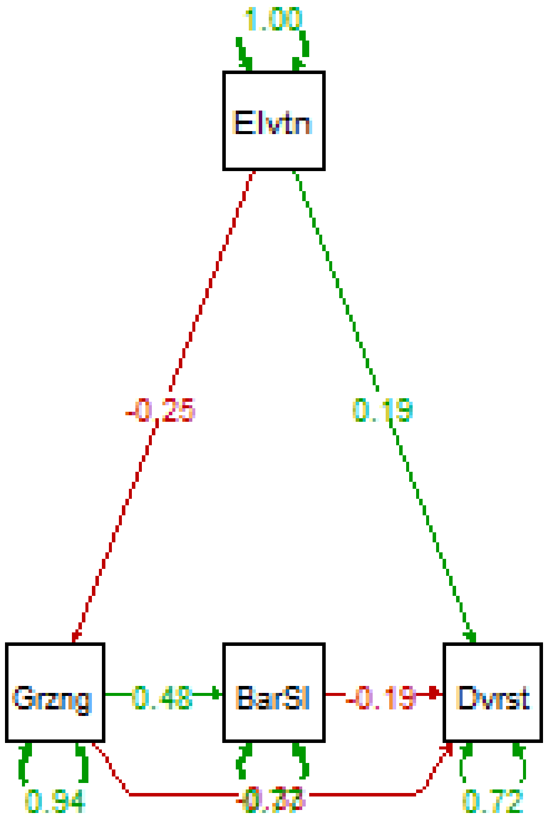
First impression of ‘lavaan’

Characters in node labels

```
library(semPlot)
semPaths(sem.fit, what='std', nCharNodes=6, sizeMan=10,
         edge.label.cex=1.25, curvePivot = TRUE,
         fade=FALSE)
```

Curved links

No transparency of links



See more:

http://sachaepskamp.com/semPlot/examples

Day 5 – Part 2

Day 5 – Part 2

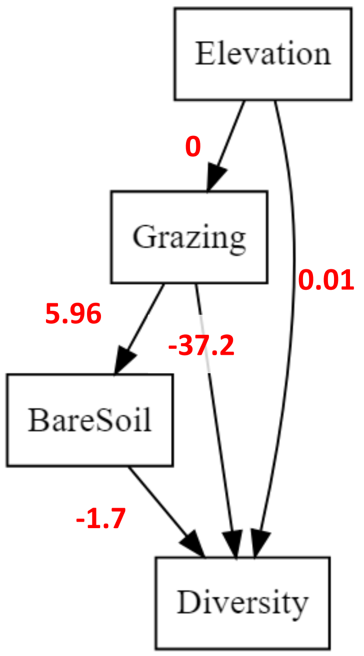
Outline

- Understanding path coefficients
  - ✓ Variance, covariance, correlation, regression coefficients
  - ✓ Indirect effects
  - ✓ Unexplained variances

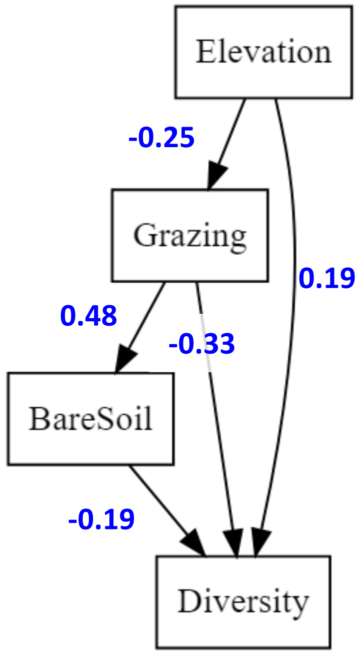
Path Coefficients

```
# Results from SEM model

> summary(sem.fit , standardize = T)

Parameter Estimates:
Regressions:
                   Estimate   Std.Err  z-value  P(>|z|)   Std.lv   Std.all
  Grazing ~
    Elevation        -0.000     0.000   -2.475    0.013   -0.000    -0.252
  BareSoil ~
    Grazing           5.963     1.161    5.136    0.000    5.963     0.476
  Diversity ~
    Elevation         0.011     0.005    2.062    0.039    0.011     0.190
    Grazing         -37.259    11.739   -3.174    0.002  -37.259    -0.331
    BareSoil         -1.696     0.913   -1.856    0.063   -1.696    -0.189
```

Unstandardized



Standardized

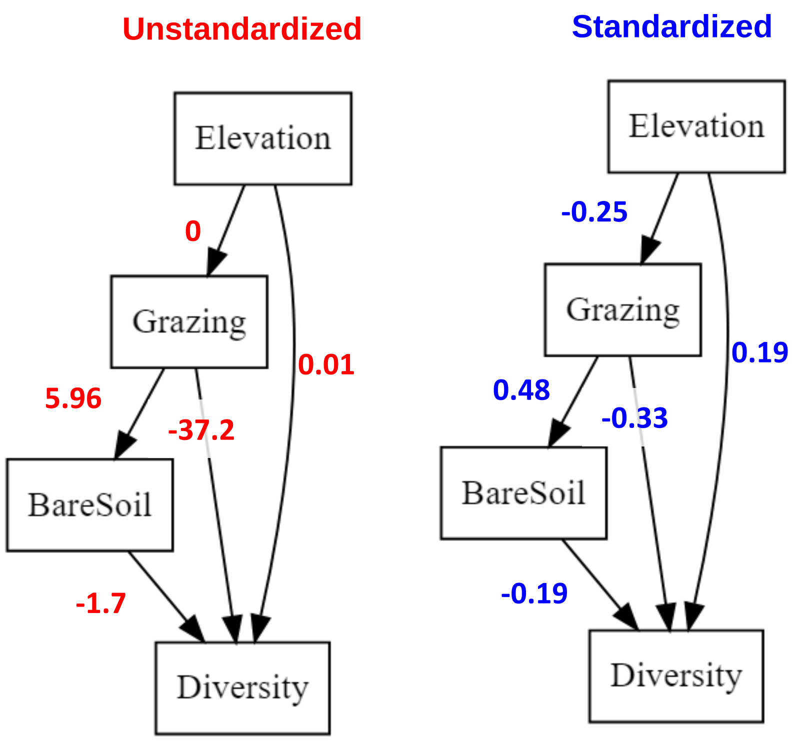


Path Coefficients

```
# Results from SEM model

> summary(sem.fit , standardize = T)

Parameter Estimates:
Regressions:
                  Estimate          Std.all
  Grazing ~
    Elevation       -0.000           -0.252
  BareSoil ~
    Grazing          5.963            0.476
  Diversity ~
    Elevation        0.011            0.190
    Grazing        -37.259           -0.331
    BareSoil        -1.696           -0.189
```

The building blocks of path coefficients

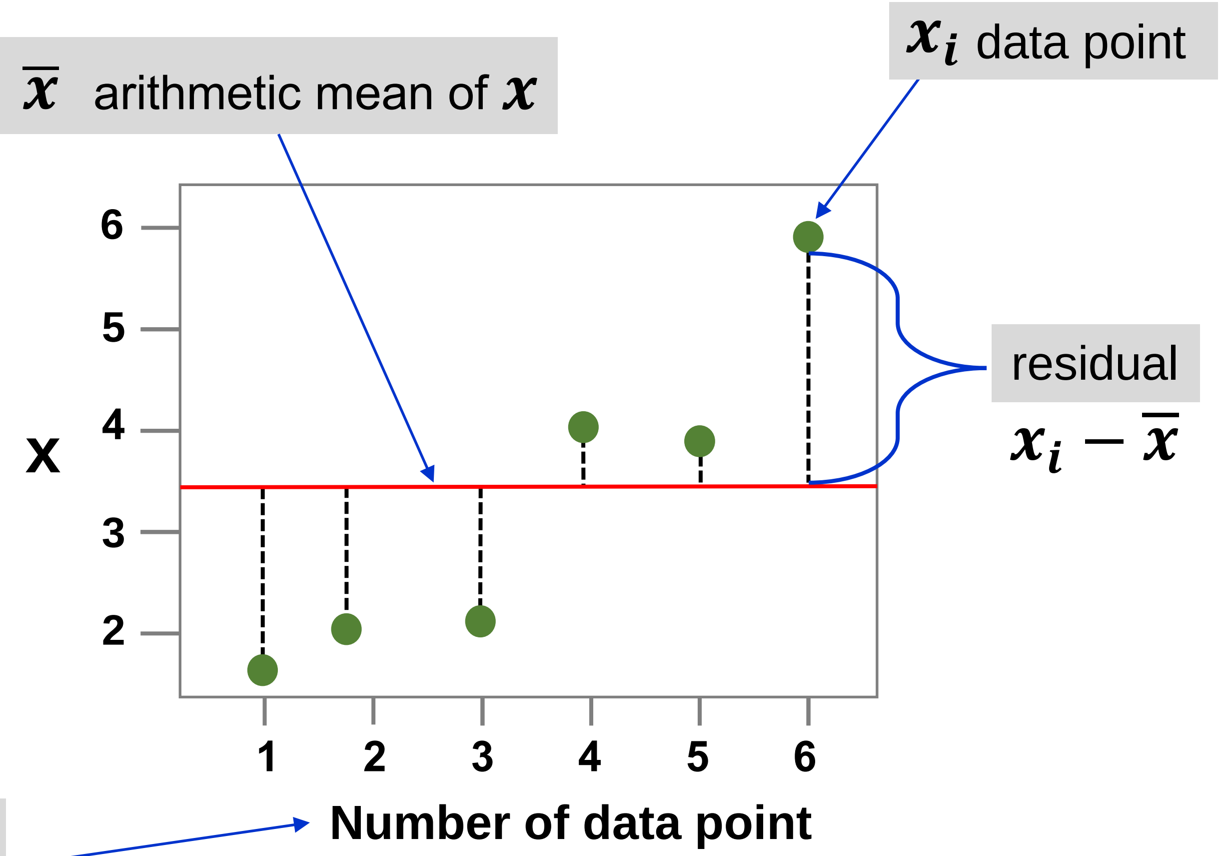
- variances,
- covariances,
- correlations,
- regression coefficients



What is Variance?

```
# in R

x <- c(1, 2, 3, 4)

var(x) # Variance

[1] 1.666667
```

$$VAR_x = \frac{\sum(x_i - \bar{x})^2}{n-1}$$



sample size
n = 6

Path Coefficients

```
# in R
x <- c(1, 2, 3, 4)
var(x) # Variance
[1] 1.666667

y <- c(70, 30, 10, 90)
var(y) # Variance
[1] 1333.333

cov(x,y) # Covariance
[1] 6.666667


> mean(x)
[1] 2.5
> mean(y)
[1] 50
```

What is Covariance?

- Dependency between two variables
- Scaled to the raw values

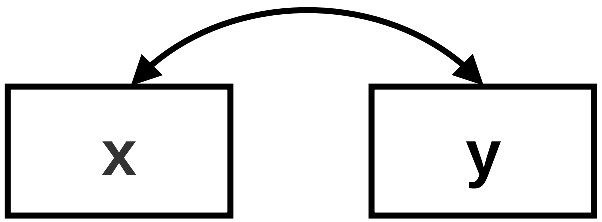
$$VAR_x = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

Covariance Matrix

	x	y
x	1.66	
y	**6.66**	1333.3

$$VAR_y = \frac{\sum(y_i - \overline{y})^2}{n-1}$$

$$COV_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$



Path Coefficients

What is Correlation?

```
# Covariance
cov(x,y)
[1] 6.666667

# Correlation
cor(x, y)
[1] 0.14
# calculate by hand
 cov(x, y)/(sd(x)*sd(y))
[1] 0.14
```

Covariance Matrix

Raw Covariance Matrix

	x	y
x	1.66	
y	6.66	1333.3

$$COV_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation Matrix

Standardised Covariance Matrix

	x	y
x	1	
y	0.14	1

$$r_{xy} = \frac{COV_{xy}}{SD_x \times SD_y}$$

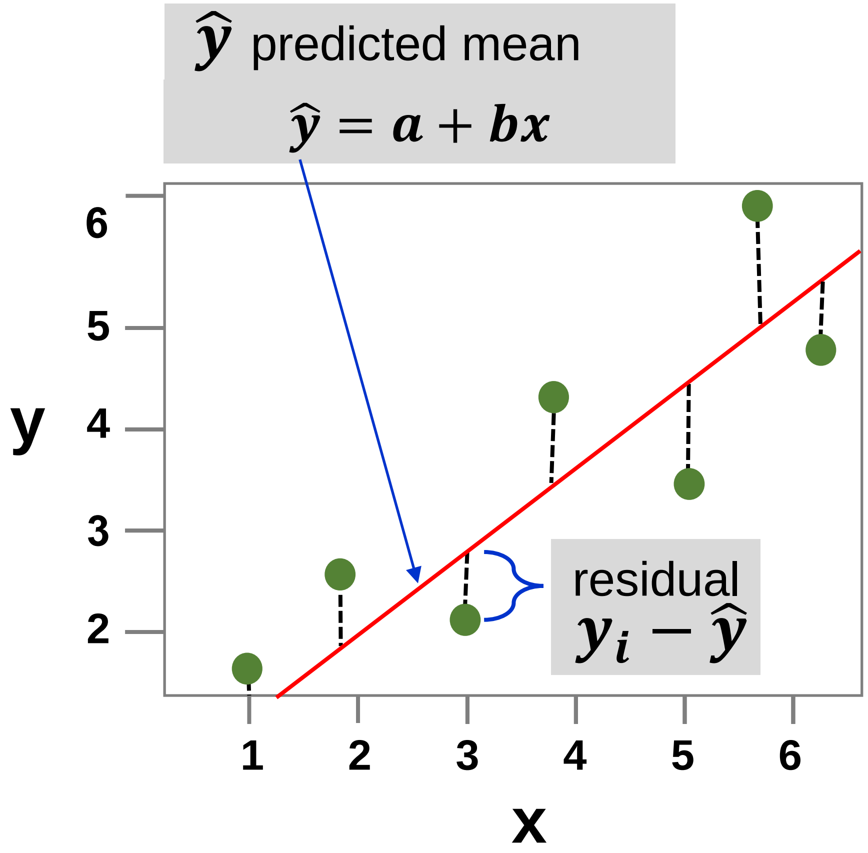
standard deviation of the mean
(the square-root of the variance)

Path Coefficients

What is Regression Coefficient?

$$y = a + \color{blue}{b}x$$

$$\boxed{x} \xrightarrow{\color{blue}{b_{xy}}} \boxed{y}$$



$$b_{xy} = \frac{COV_{xy}}{VAR_x}$$

Unstandardized regression coefficient

$$b_{xy} = r_{xy} = \frac{COV_{xy}}{SD_x \times SD_y}$$

Standardized regression coefficient

Path Coefficients

What is Regression Coefficient?

$$y = a + bx$$

$$x \xrightarrow{b_{xy}} y$$

When two variables are connected by a single path, the coefficient of that path is the correlation coefficient

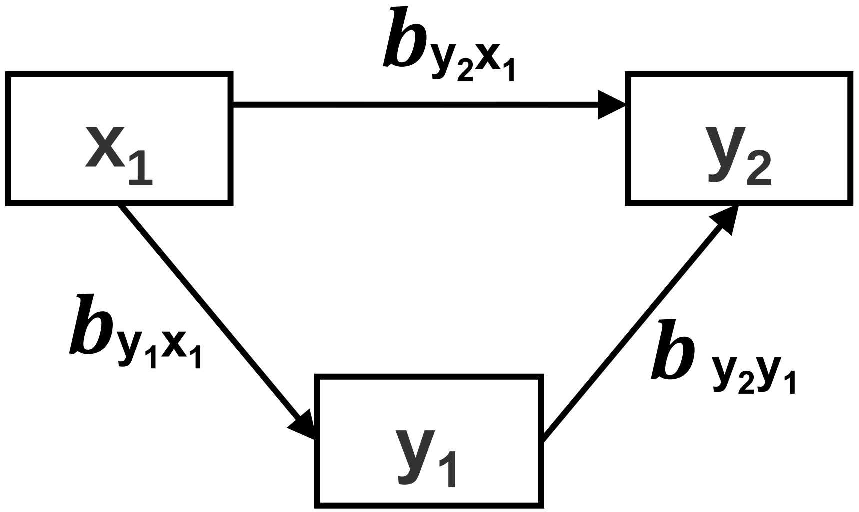
$$b_{xy} = \frac{COV_{xy}}{VAR_x}$$

Unstandardized regression coefficient

$$\boxed{b_{xy} = r_{xy}} = \frac{COV_{xy}}{SD_x \times SD_y}$$

Standardized regression coefficient

Path Coefficients



When variables are connected by more than one path, each path coefficient is the <u>'partial'</u> regression coefficient.

takes the bivariate correlation between $\mathbf{x_1}$ and $\mathbf{y_2}$

removes the joint influence of $\mathbf{x_1}$ and $\mathbf{y_1}$ on $\mathbf{y_2}$

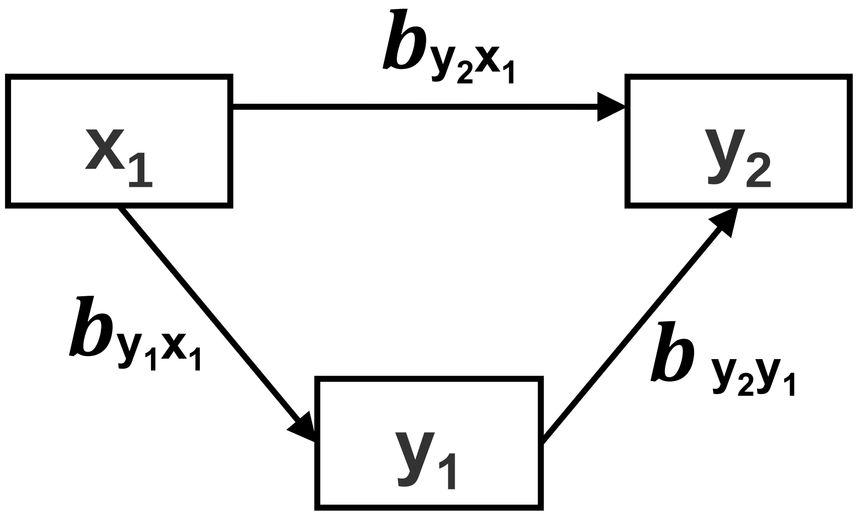
$$b_{y2x1} = \frac{r_{x1y2} - (r_{x1y1} \times r_{y1y2})}{1 - r_{x1y1}^2}$$

scales this effect by the shared variance between $\mathbf{x_1}$ and $\mathbf{y_1}$

Corresponding equations:

$$y_1 = b_1x_1$$
$$y_2 = b_2x_1 + b_3y_1$$

Path Coefficients

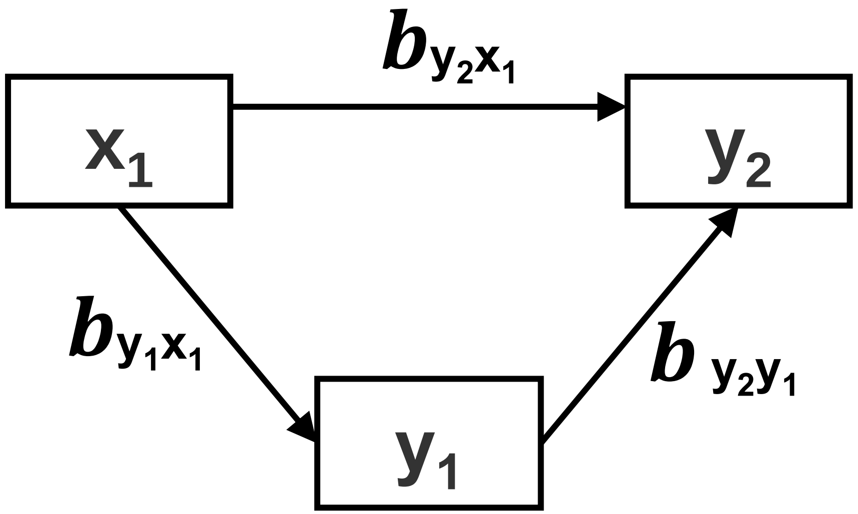


When variables are connected by more than one path, each path coefficient is the <u>'partial'</u> regression coefficient.

Standardized $$b_{y2x1} = \frac{r_{x1y2} - (r_{x1y1} \times r_{y1y2})}{1 - r_{x1y1}^2}$$

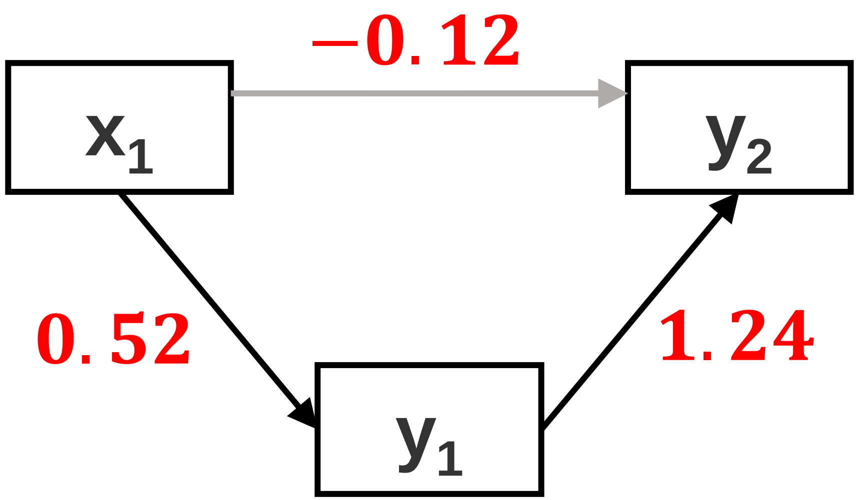
Unstandardized $$b_{y2x1} = \frac{SD_{y2}}{SD_{x1}} \times \frac{r_{x1y2} - (r_{x1y1} \times r_{y1y2})}{1 - r_{x1y1}^2}$$

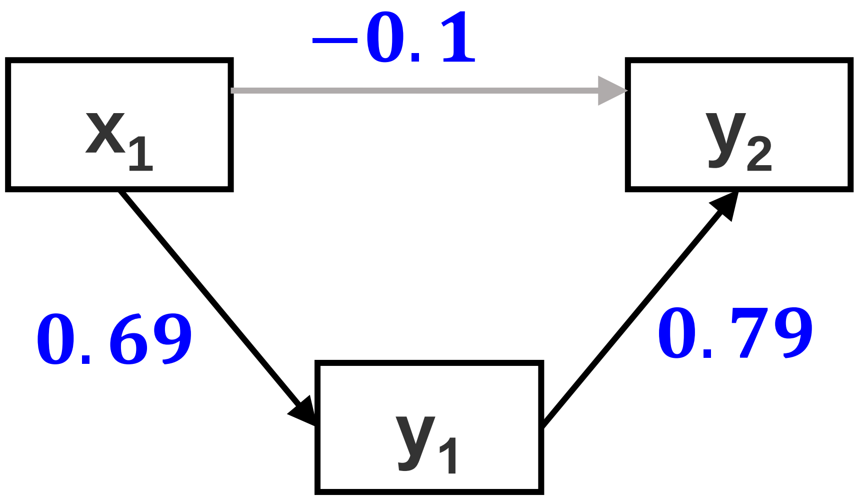
Path Coefficients




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

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

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

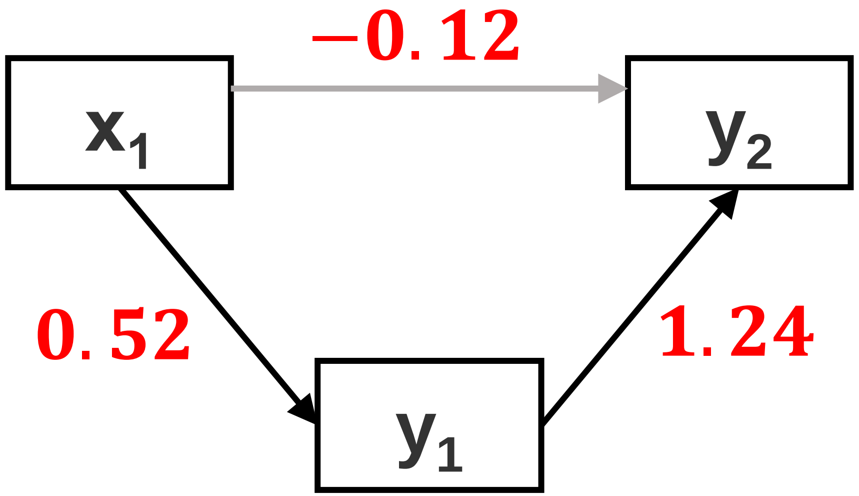
Path Coefficients

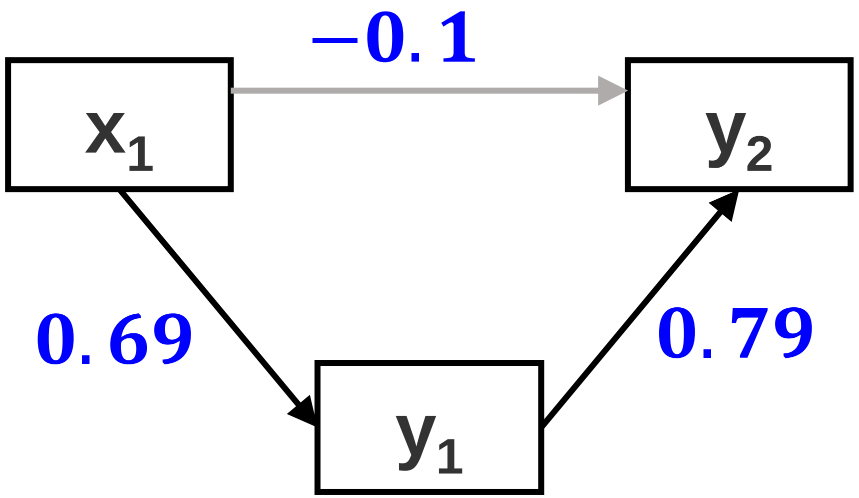





```
# Results
...
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  y1 ~
    x1                0.517    0.054    9.525    0.000    0.517    0.690
  y2 ~
    x1               -0.116    0.113   -1.034    0.301   -0.116   -0.099
    y1                1.239    0.150    8.248    0.000    1.239    0.787
```

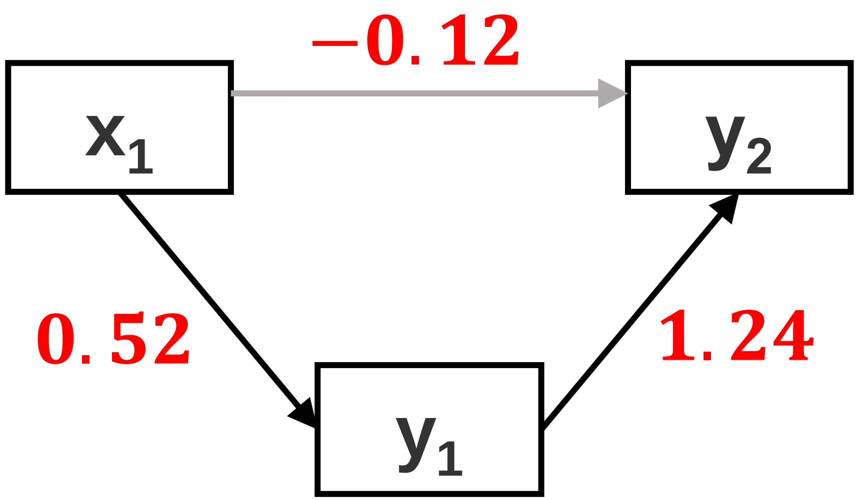
Path Coefficients

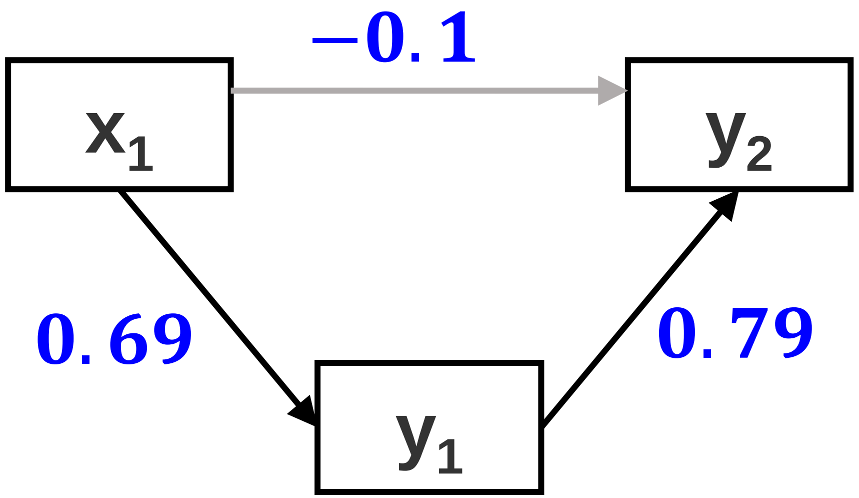





```
# to call the unstandardised estimates for each predictor:

coef(sem.fit1)
>
 y1~x1  y2~x1  y2~y1 y1~~y1 y2~~y2
 0.517 -0.116  1.239  0.036  0.081
```

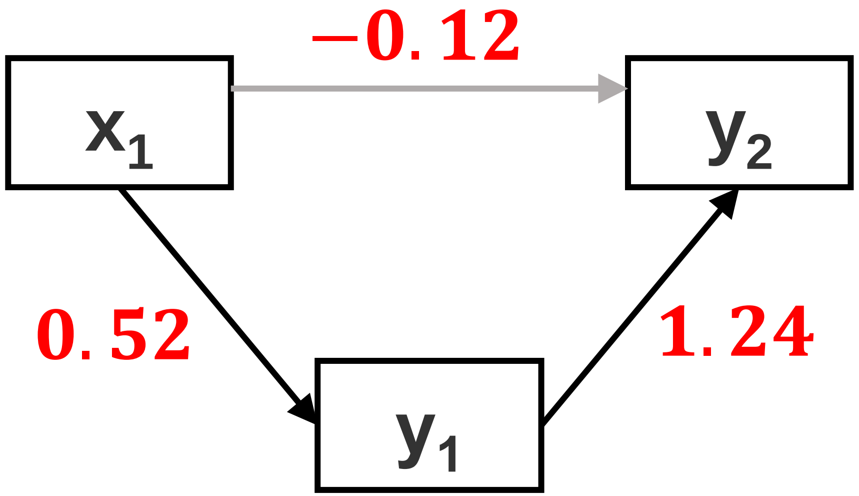
Path Coefficients

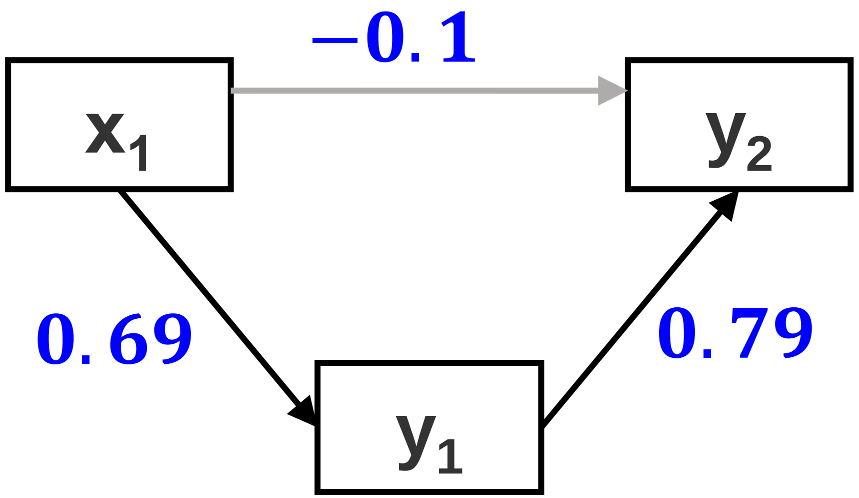





```
# or (also Unstandardized coeficientes)
> parameterEstimates(sem.fit1)
  lhs op rhs    est    se      z pvalue ci.lower ci.upper
1  y1  ~  x1  0.517 0.054  9.525  0.000    0.411    0.623
2  y2  ~  x1 -0.116 0.113 -1.034  0.301   -0.337    0.104
3  y2  ~  y1  1.239 0.150  8.248  0.000    0.944    1.533
4  y1 ~~  y1  0.036 0.005  7.071  0.000    0.026    0.046
5  y2 ~~  y2  0.081 0.011  7.071  0.000    0.059    0.104
6  x1 ~~  x1  0.122 0.000     NA     NA    0.122    0.122
```

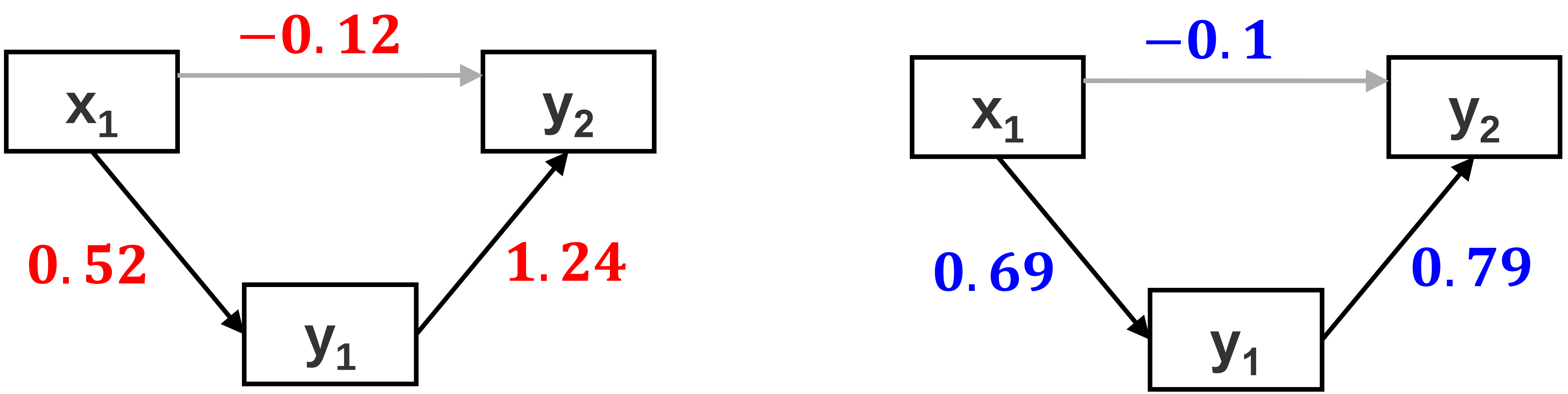
Path Coefficients





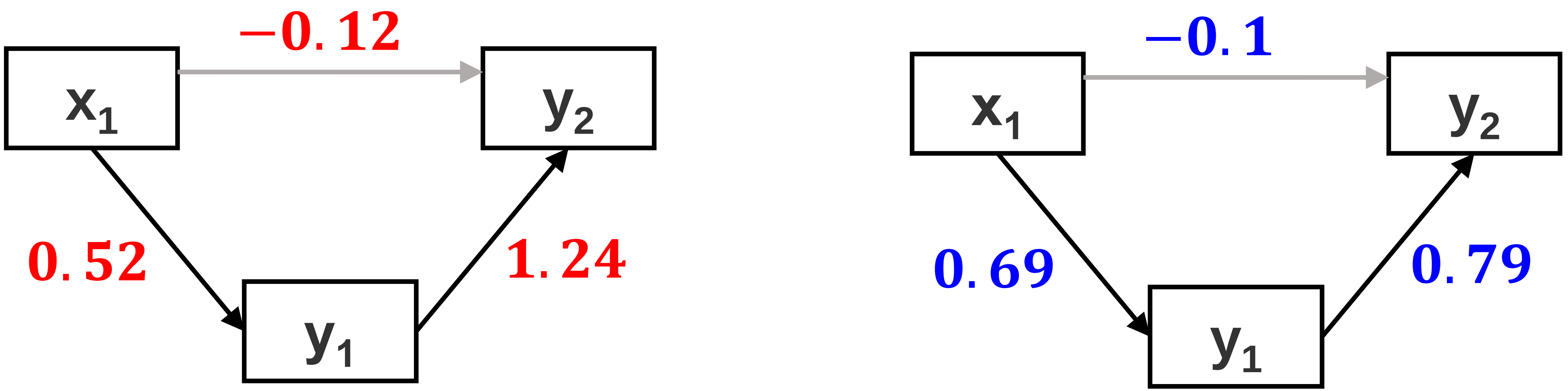

```
# Standardised coeficientes
> standardizedSolution(sem.fit1)
  lhs op rhs est.std    se      z pvalue ci.lower ci.upper
1  y1  ~  x1   0.690 0.046 15.067    0.0    0.600    0.779
2  y2  ~  x1  -0.099 0.095 -1.037    0.3   -0.285    0.088
3  y2  ~  y1   0.787 0.077 10.214    0.0    0.636    0.938
4  y1 ~~  y1   0.524 0.063  8.304    0.0    0.401    0.648
5  y2 ~~  y2   0.478 0.068  7.051    0.0    0.345    0.610
6  x1 ~~  x1   1.000 0.000     NA     NA    1.000    1.000
```

Path Coefficients




```
# To standardize the effect:
coef(sem.fit1)[1] *sd(data1$x1)/sd(data1$y1)
>
y1~x1
0.6896943

# check with the result table from lavaan
standardizedSolution(sem.fit1)[1 ,  "est.std"]
>
[1] 0.6896943
```

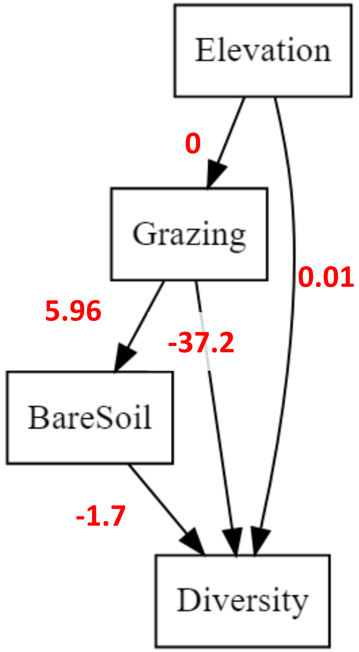
Path Coefficients

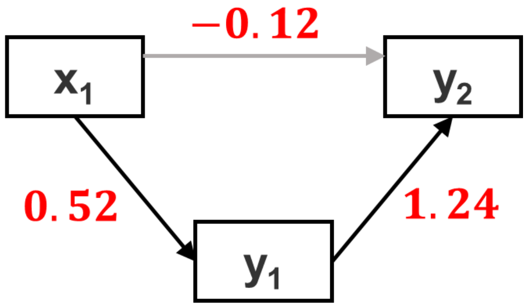


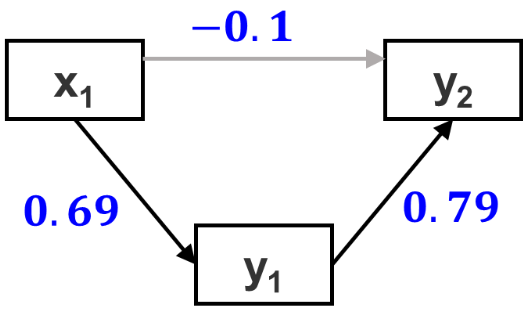

```
# To unstandardize the effect:
standardizedSolution(sem.fit1)[1 , "est.std"] * sd(data1$y1)/sd(data1$x1)
>
[1] 0.51705

# check with the result table from lavaan
coef(sem.fit1)[1]
>
 y1~x1
0.51705
```

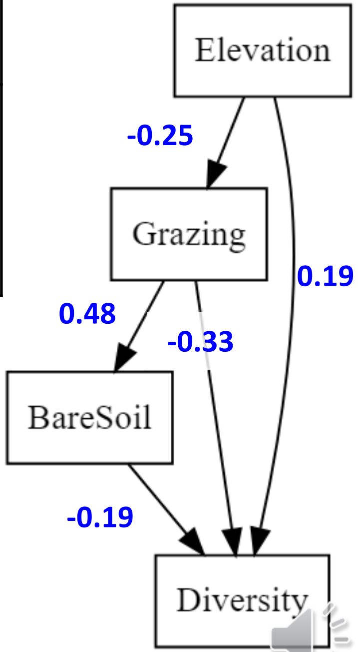
Path Coefficients

Unstandardized path coefficients	Standardized path coefficients
• Good for prediction: coefficients are in raw units	• Good for ranking: coefficients are in equivalent units
• Has direct real world meaning	• Less clear real world meaning
• Can be compared across pathways or models that have identical units	• Can be compared across all pathways in all models









Day 5 – Part 2

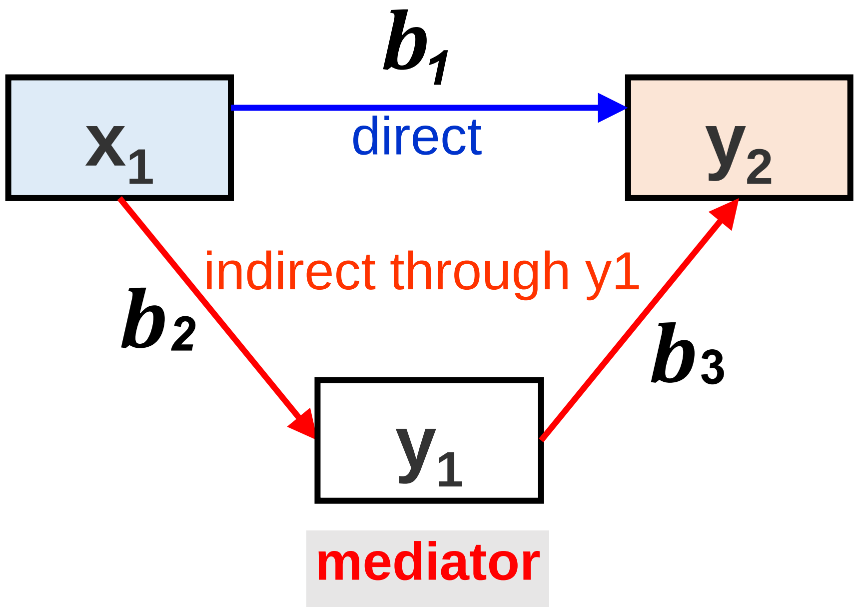
Outline

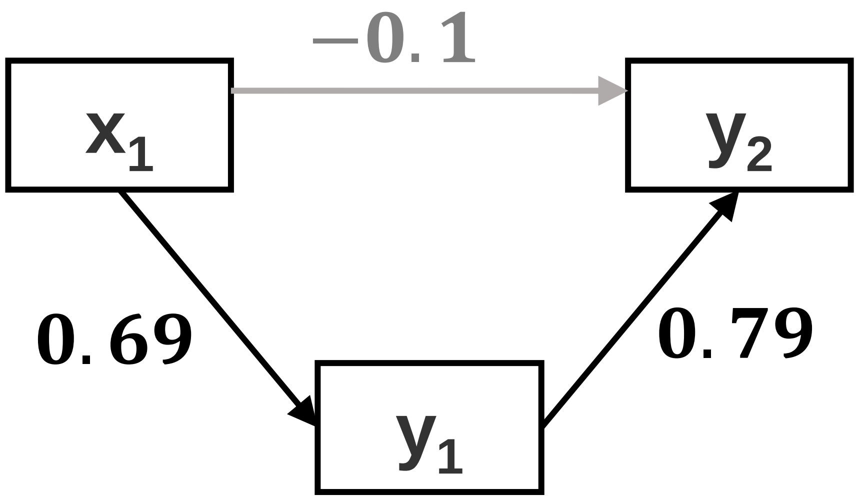
- Understanding path coefficients

  - ✓ Variance, covariance, correlation, regression coefficients
  - ✓ **Indirect effects**
  - ✓ Unexplained variances

Indirect effects

Effects of x_1 on y_2





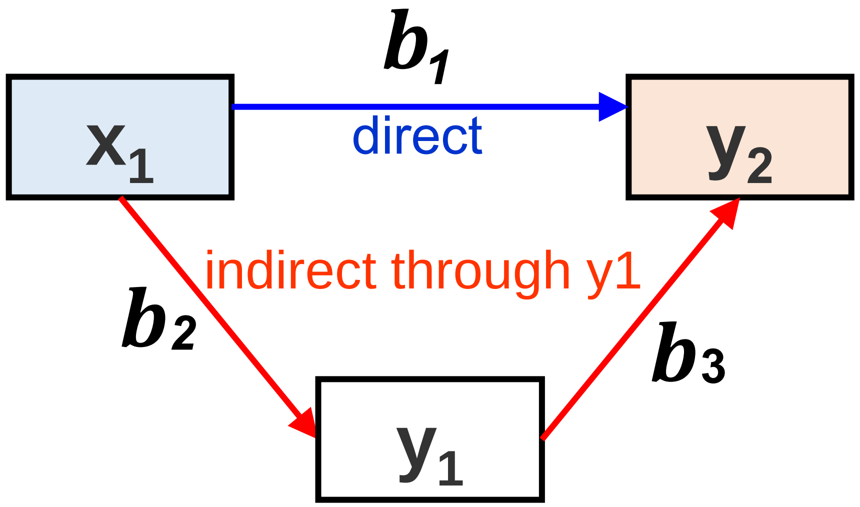
direct b_1

indirect $b_2 \times b_3$

Total effect = direct + indirect

direct	**-0.1**
indirect	$0.69 \times 0.79 =$ **0.54**
total	$-0.1 + 0.55 =$ **0.44**

Indirect effects



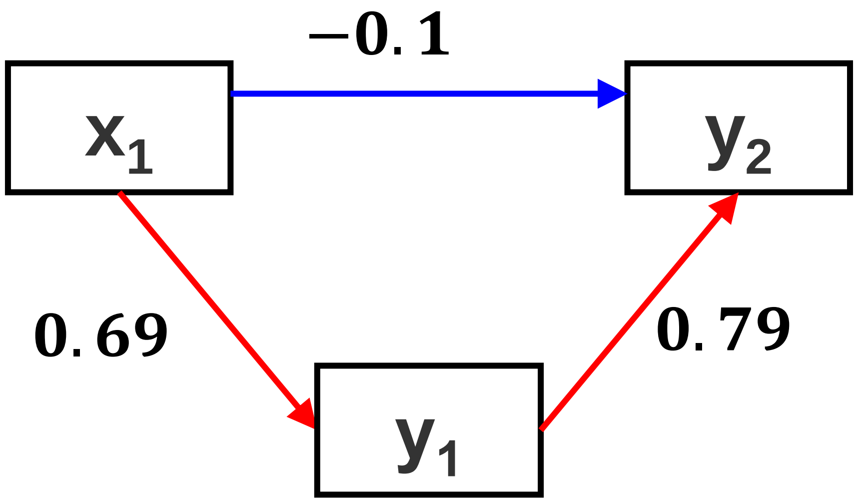

```
# Naming the coefficients in lavaan
sem_mod1 <- '
     y2 ~ b1*x1 + b3*y1
     y1 ~ b2*x1
     # define direct, indirect and total effects
     direct   := b1
     indirect := b2*b3
     total    := b1 + (b2*b3)
     # or
     # total   := direct + indirect

'
sem.fit1 <- sem(sem_mod1, data=data1)
summary(sem.fit1, standardize = T)
```

direct b_1

indirect $b_2 \times b_3$

Total effect = direct + indirect

Indirect effects



direct	**-0.1**
indirect	$0.69 \times 0.79 =$ **0.54**
total	$-0.1 + 0.55 =$ **0.44**

The total effect is equivalent to the total correlation

```
> summary(sem.fit1, standardize = T)
...
Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    direct           -0.116    0.113   -1.034    0.301   -0.116   -0.099
    indirect          0.640    0.103    6.235    0.000    0.640    0.543
    total             0.524    0.106    4.959    0.000    0.524    0.444
```

Day 5 – Part 2

Outline

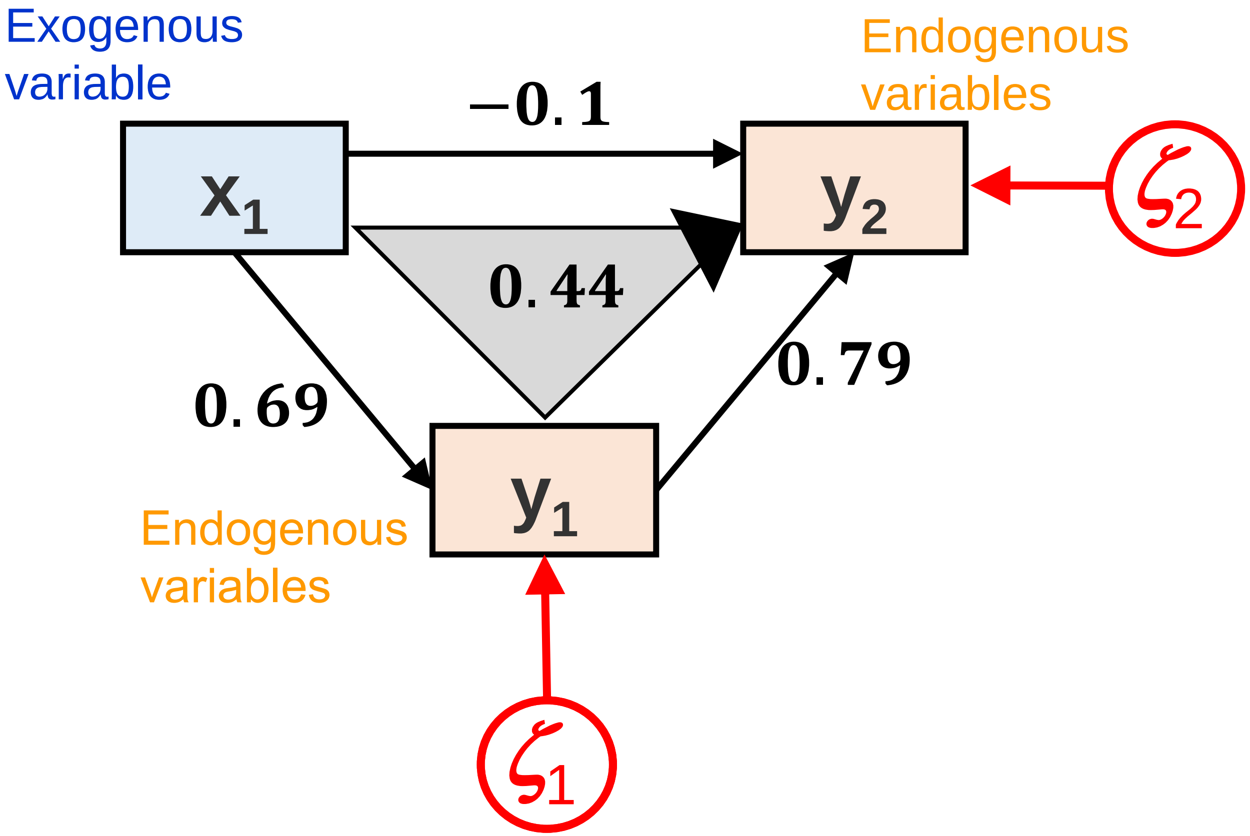
- Understanding path coefficients

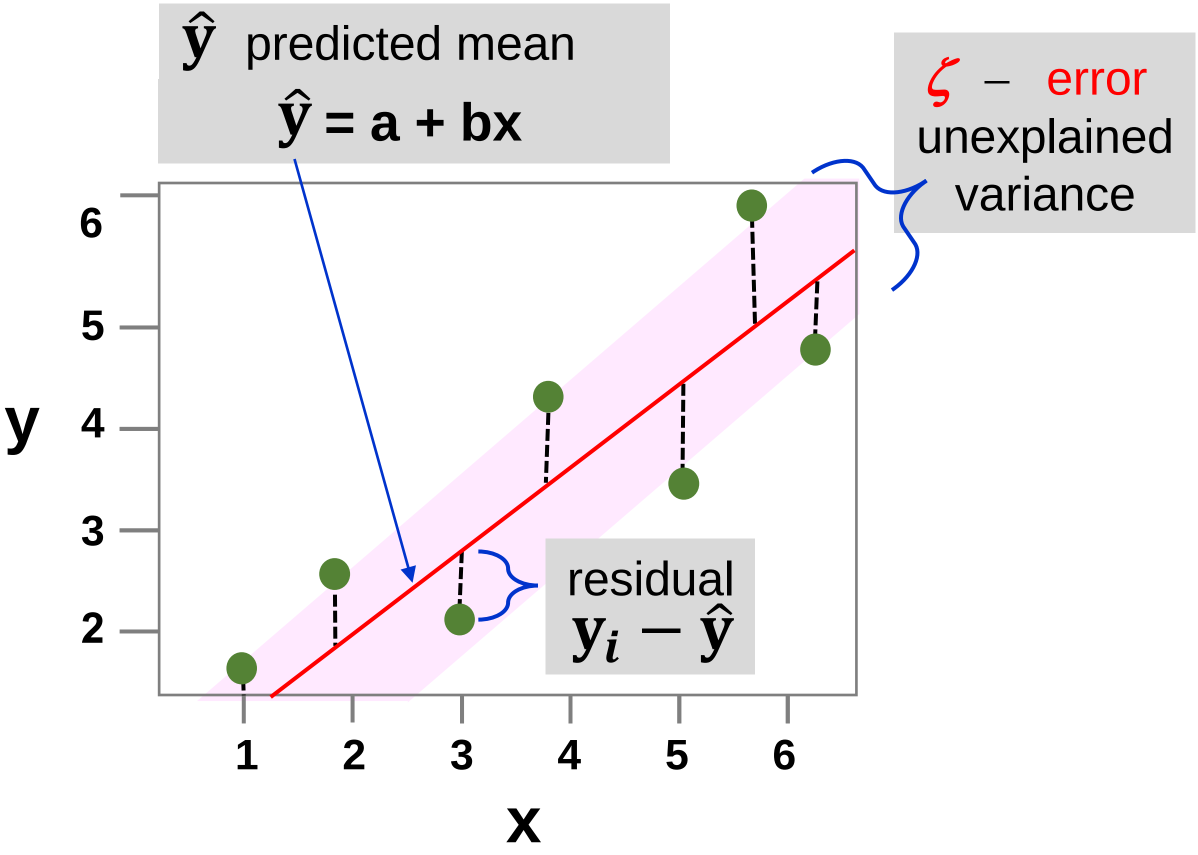
  ✓ Variance, covariance, correlation, regression coefficients

  ✓ Indirect effects

  ✓ **Unexplained variances**

Unexplained variances



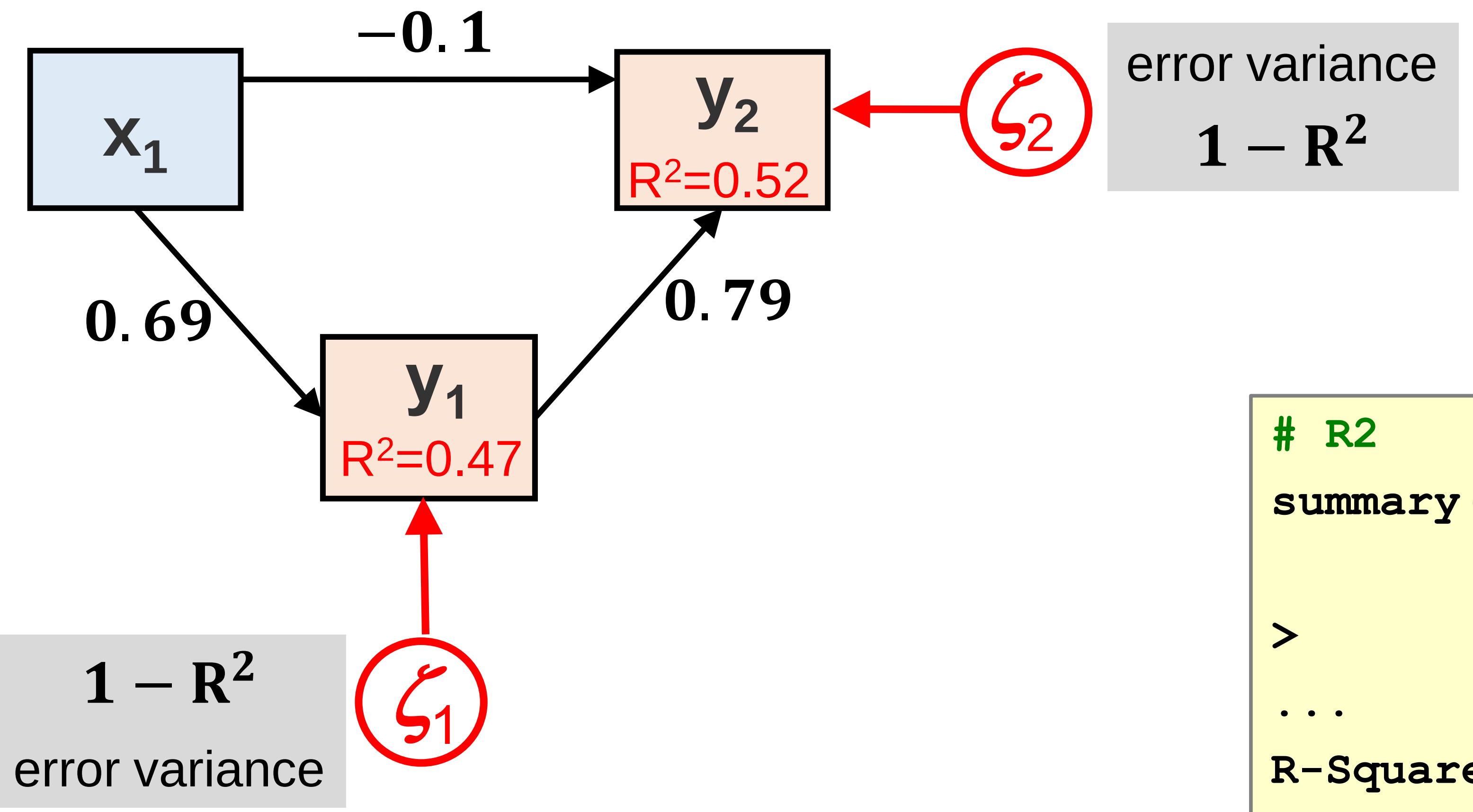


Equation form:

$$y_1 = a_1 + b_1 x_1 + \zeta_1$$

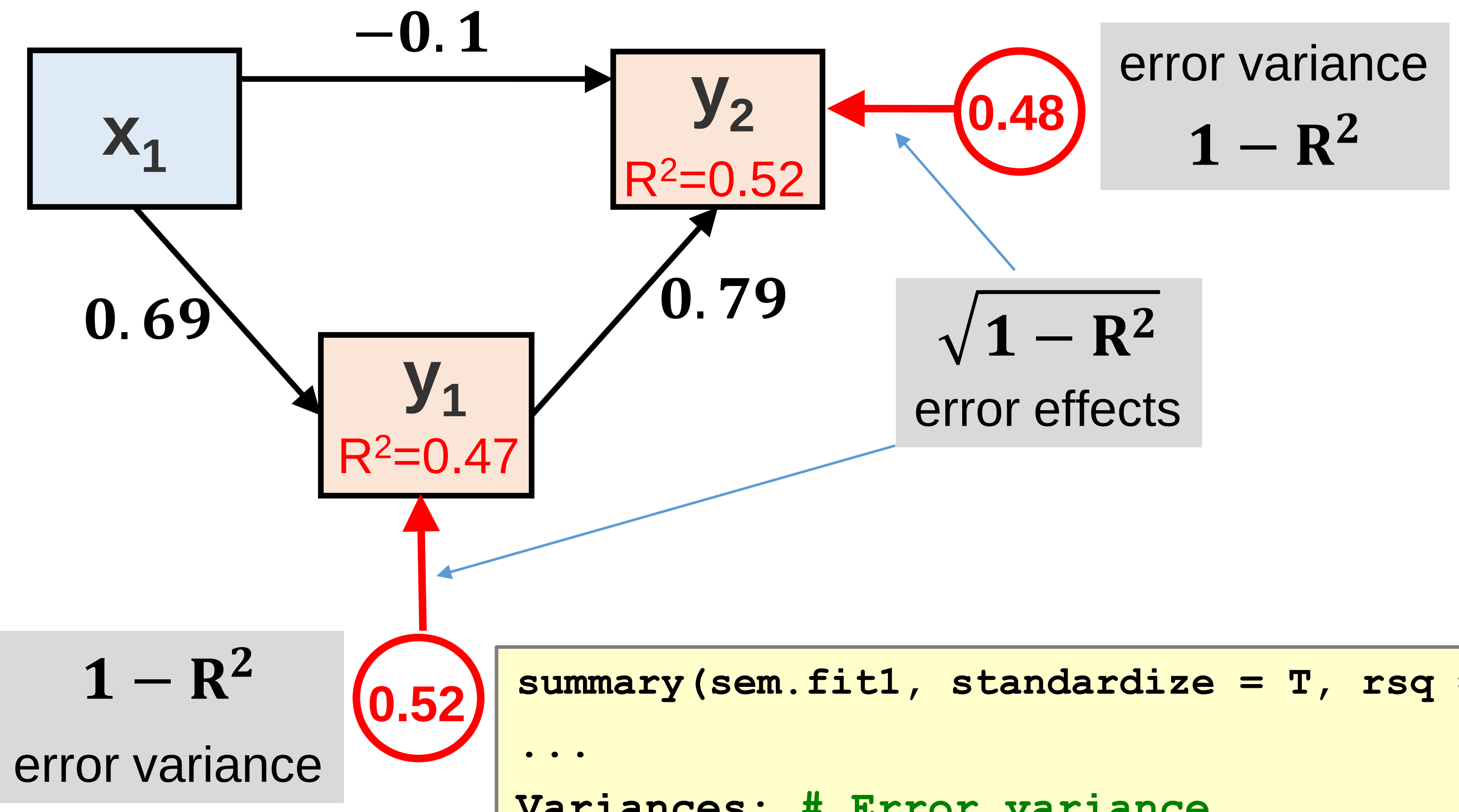
$$y_2 = a_2 + b_2 x_1 + b_3 y_1 + \zeta_2$$

Unexplained variances



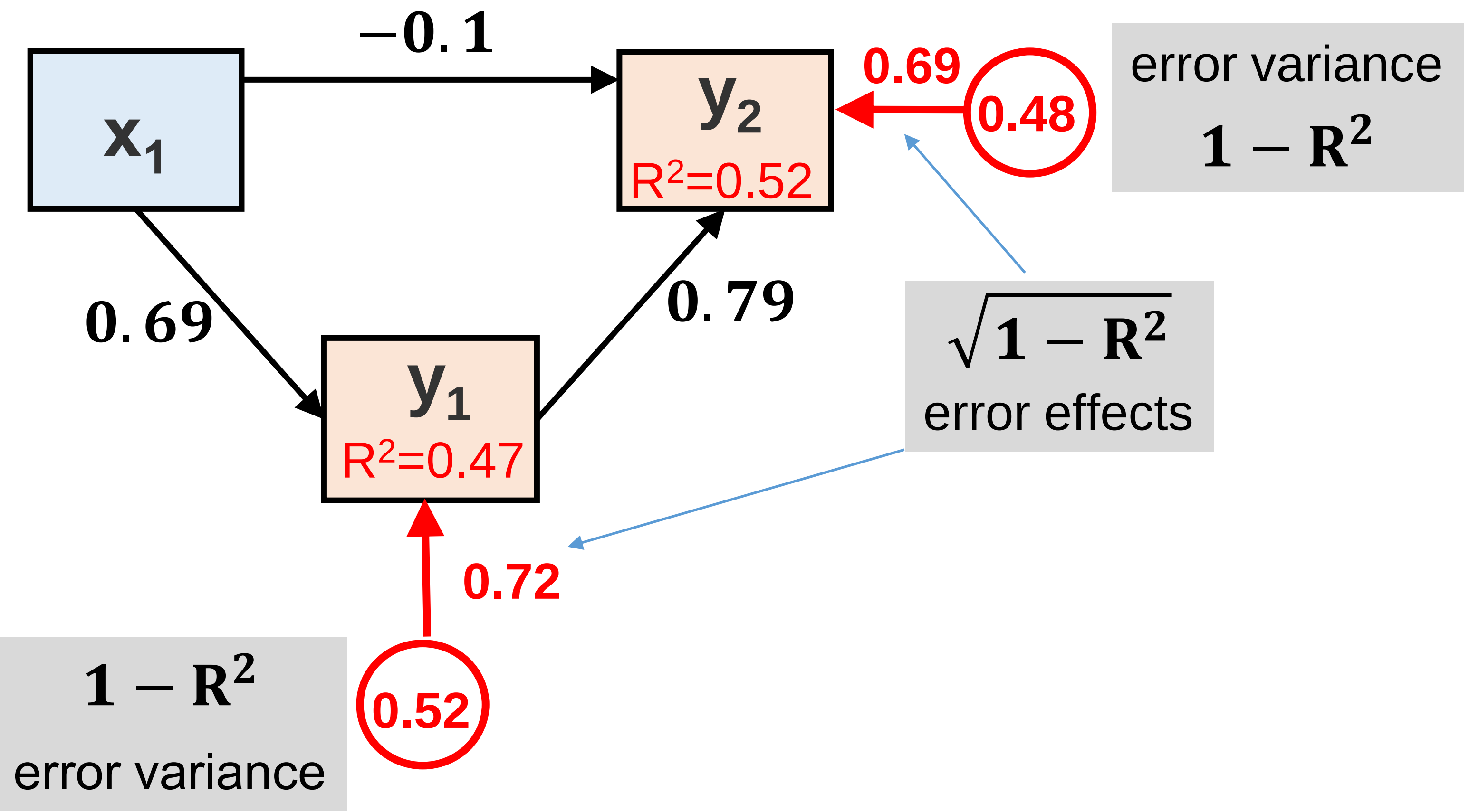

```
# R2
summary(sem.fit1, standardize = T,
                                      rsq = T)
>
...
R-Square:
                          Estimate
    y1                       0.476
    y2                       0.517
```

Unexplained variances

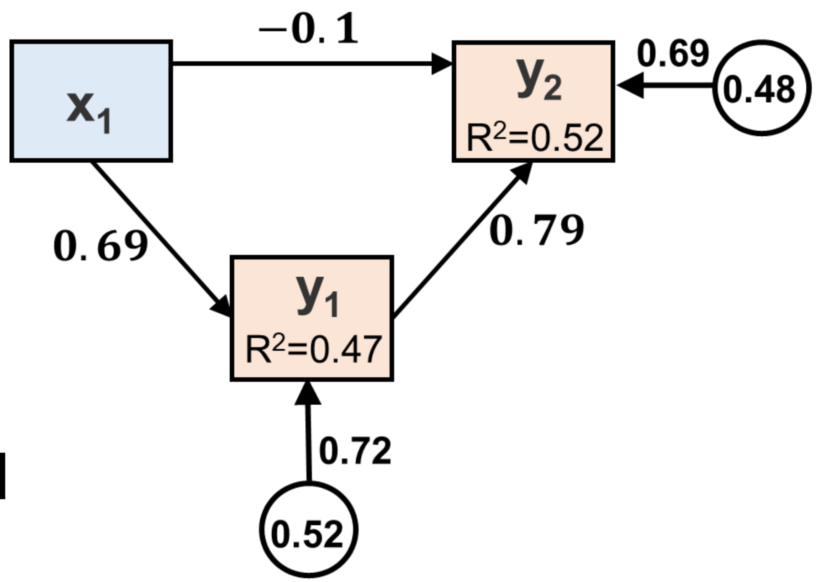



```
summary(sem.fit1, standardize = T, rsq = T)
...
Variances: # Error variance
                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .y1                 0.036    0.005    7.071    0.000    0.036    0.524
   .y2                 0.082    0.012    7.071    0.000    0.082    0.483
```

Unexplained variances



Unexplained variances

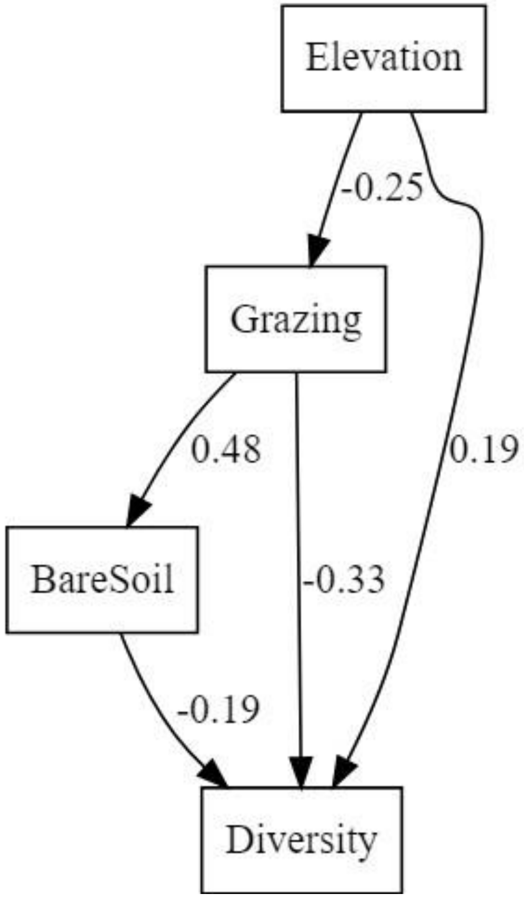


The major points to remember are:

- standardized coefficients reflect (partial) correlations;
- the indirect effect of one variable on another is obtained by multiplying the individual path coefficients (standardized or unstandardized);
- the total effect is the sum of direct and indirect paths;
- the bivariate correlation is the sum of the total effect plus any undirected paths.

Day 5 Task 1

Effects of grazing on plant diversity along elevation gradient




```
# data
data <- read.csv("Grassl_data.csv")
```

Day 5 Task 1

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

1. Calculate the standardised direct, indirect and total effects of ***grazing*** on ***diversity*** (do this in lavaan in R)
2. Define the exogenous and endogenous variables in the model
3. For each endogenous variable get the following:
   - the variance explained by the model
   - the error variance
   - the effect of the error (path coefficient with the error variance).

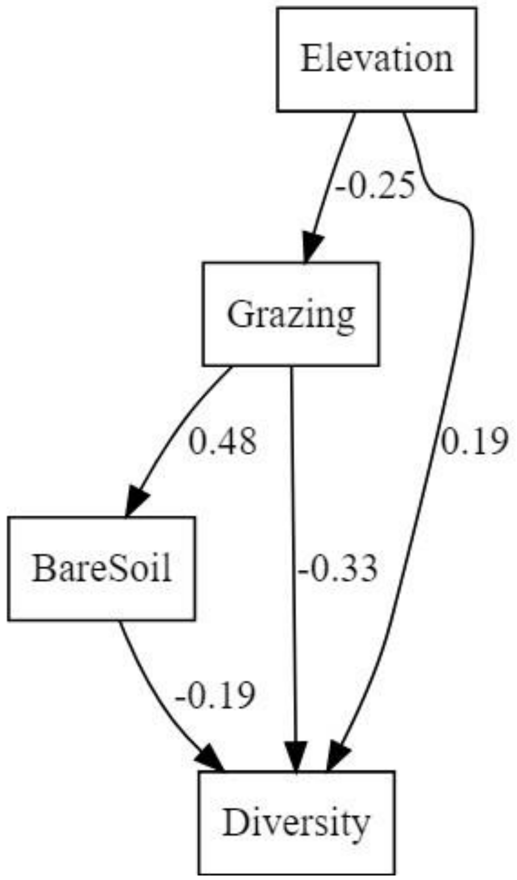


Fig. 1

Day 5 – Part 3

Day 5 – 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)
>
```

	x1	y1	y2
x1	0.12	0.06	0.06
y1	0.06	0.07	0.08
y2	0.06	0.08	0.17

= **S**

Observed
variance-covariance matrix

Maximum-Likelihood Estimation

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

Implied
(model-estimated)
variance-covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \begin{Bmatrix} \sigma_x & & \\ \sigma_{xy_1} & \sigma_{y_1} & \\ \sigma_{xy_2} & \sigma_{y_1y_2} & \sigma_{y_2} \end{Bmatrix}$$

Likelihood Function:

tr trace of the matrix

p number of endogenous variables

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

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

$\boldsymbol{S}$ observed covariance matrix

q number of exogenous variables

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

Perfect model fit

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

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

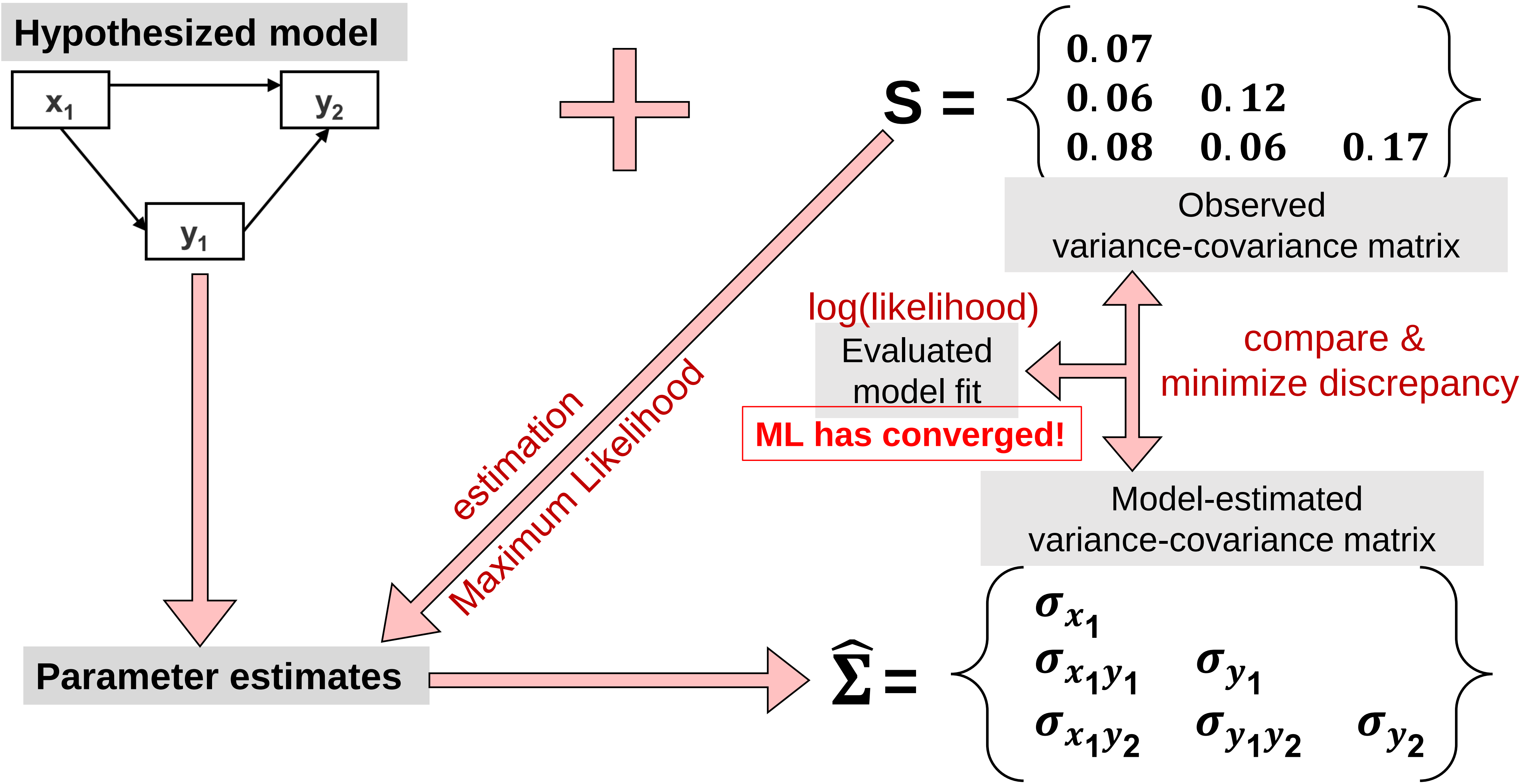
$\boldsymbol{S}$ observed covariance matrix

q number of exogenous variables

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

- scale invariant
- asymptotically unbiased
- efficient

Covariance-based SEM



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

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

- Unknowns are the model parameters to be estimated

Model Identifiability

- 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**

$a+b=8$

$a=3b$

$(3b)+b=8$

$4b=8$

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

$a+2=8$

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

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?

- 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

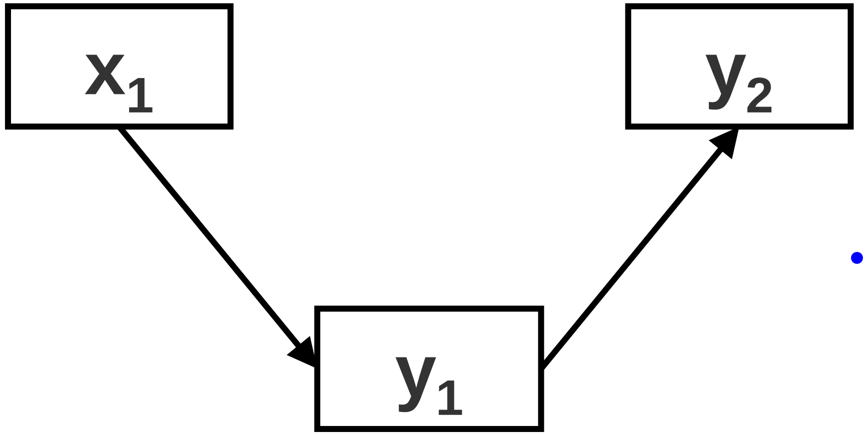
We can fit model !

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

- Unknowns are the model parameters to be estimated

Model Identifiability

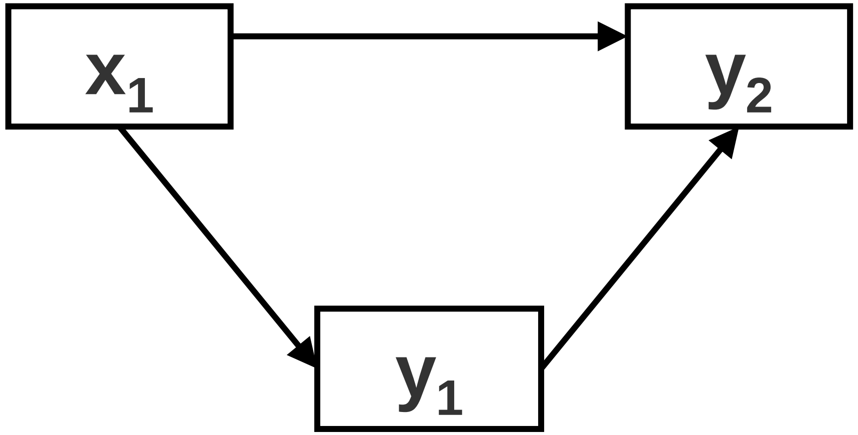
Can I fit my model?



Overidentified (Unsaturated)

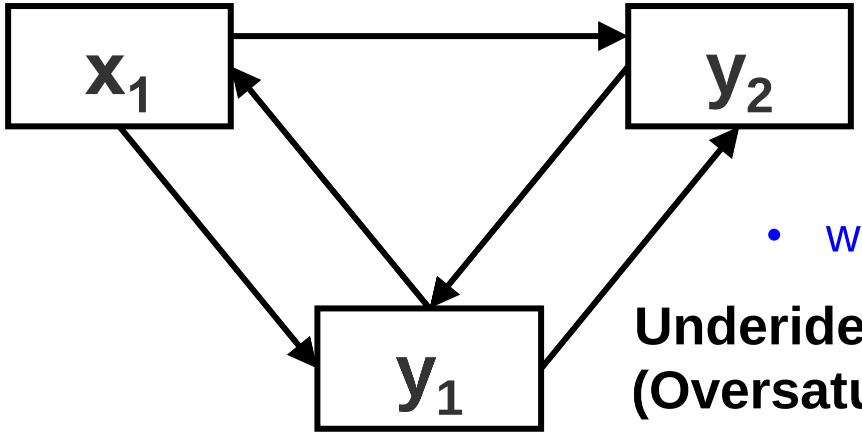
Recursive models

- 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.
```



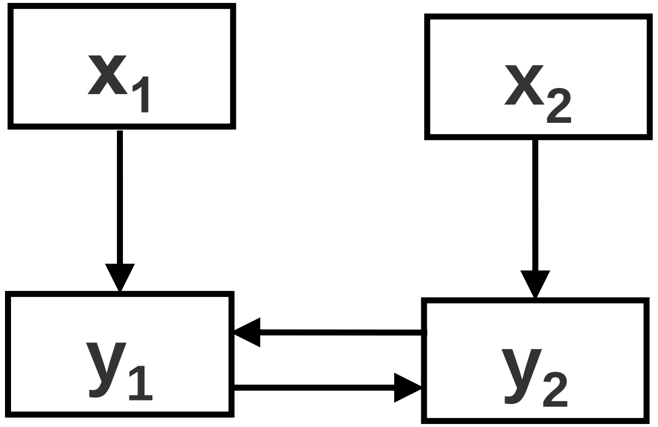
Non-recursive models

- with bidirectional feedbacks

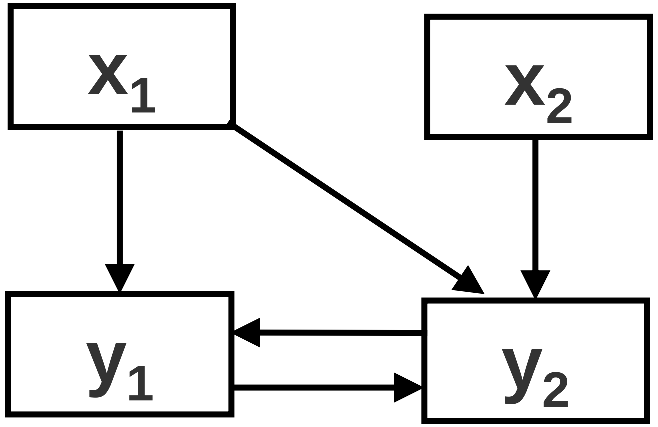
Underidentified (Oversaturated)

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

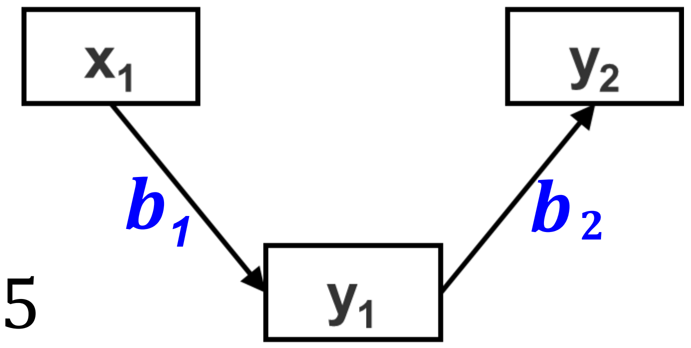
Observed variance-covariance matrix

$s = 3$

$t_{max} = 6$

$t = 2 + 3 = 5$

$5 < 6$



Overidentified (Unsaturated)

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

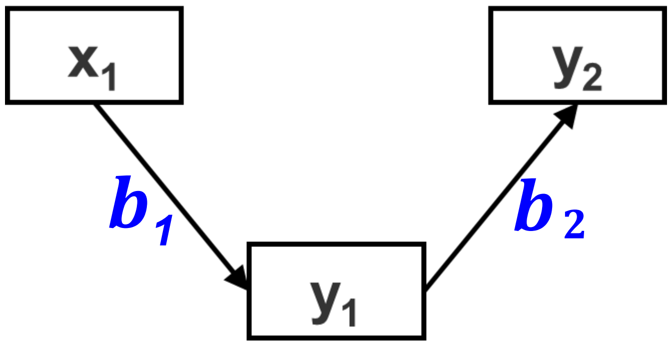
Minimum requirement $\boldsymbol{n} = \boldsymbol{p} \times 5$

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 5 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 5 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?

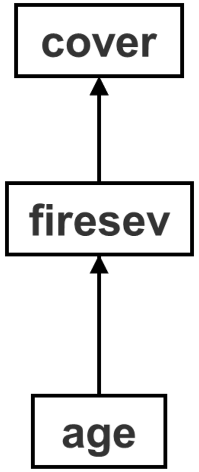


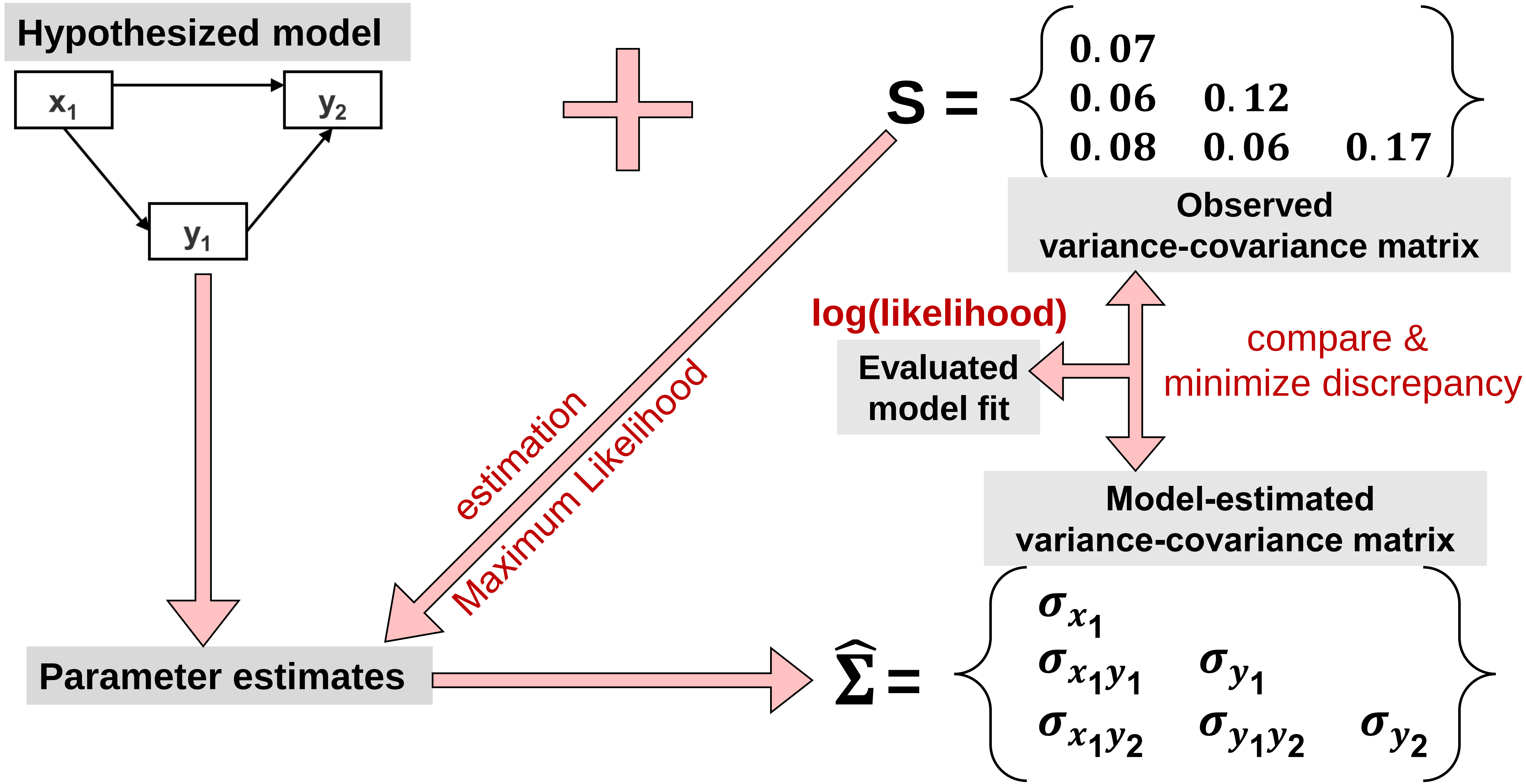
Fig. 1

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**

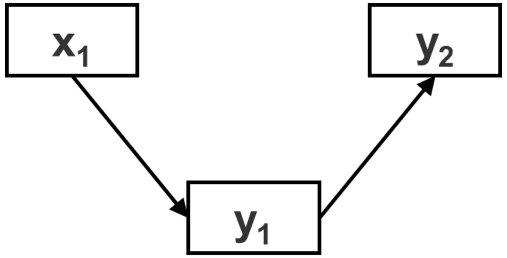
Covariance-based SEM



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

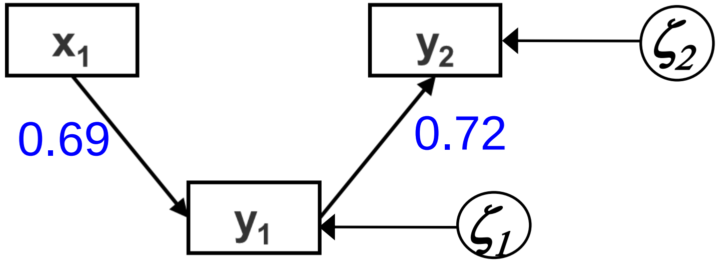
$$\hat{\boldsymbol{\Sigma}} = \left\{ \begin{matrix} \sigma_{x_1} & & \\ \sigma_{x_1 y_1} & \sigma_{y_1} & \\ \sigma_{x_1 y_2} & \sigma_{y_1 y_2} & \sigma_{y_2} \end{matrix} \right\}$$

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)
```

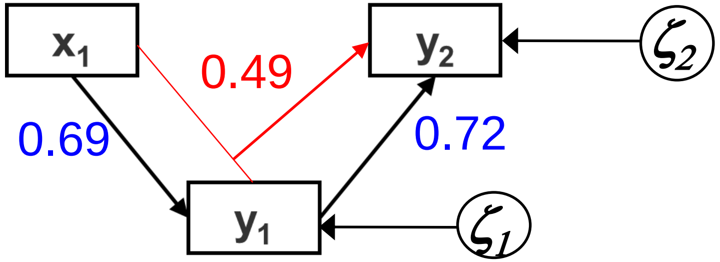
Goodness of fit




```
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

$r(x_1 y_2)$ 0.49 = 0.69 x 0.72



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)

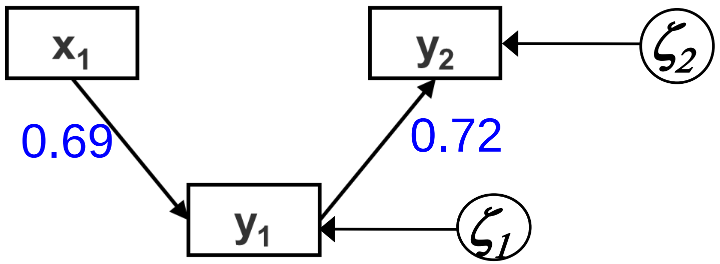
	x_1	y_1	y_2
x_1	1.00		
y_1	0.69	1.00	
y_2	0.49	0.72	1.00

residual

0.444-0.496=**-0.052**

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

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

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

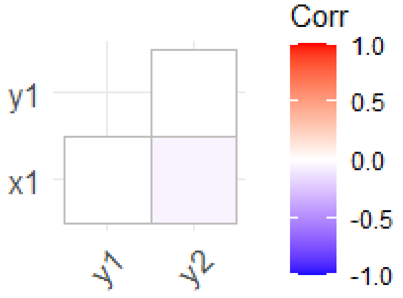
Model implied matrix (scaled)

	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**

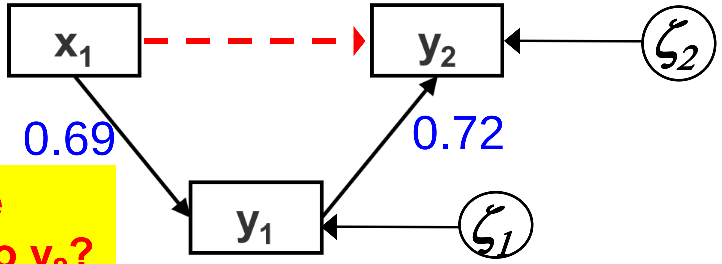
Residuals r (scaled)

	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

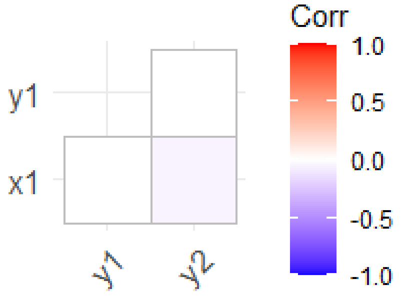
Is it good enough?

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

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

Perfect model fit

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

$\hat{\mathbf{\Sigma}}$ 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

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

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**

DF degrees of freedom

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

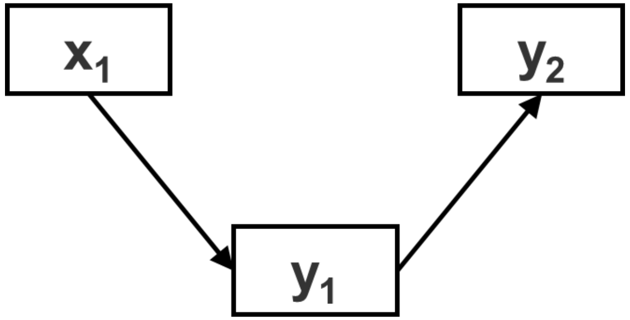
from the t-rule

s number of observed variables

t number of parameters to be estimated by the model

Goodness of fit

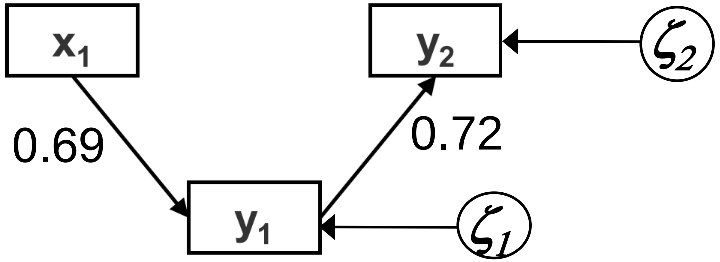
```
data1 <- read.table("Data/SEMdata1.txt", header = T)

# 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

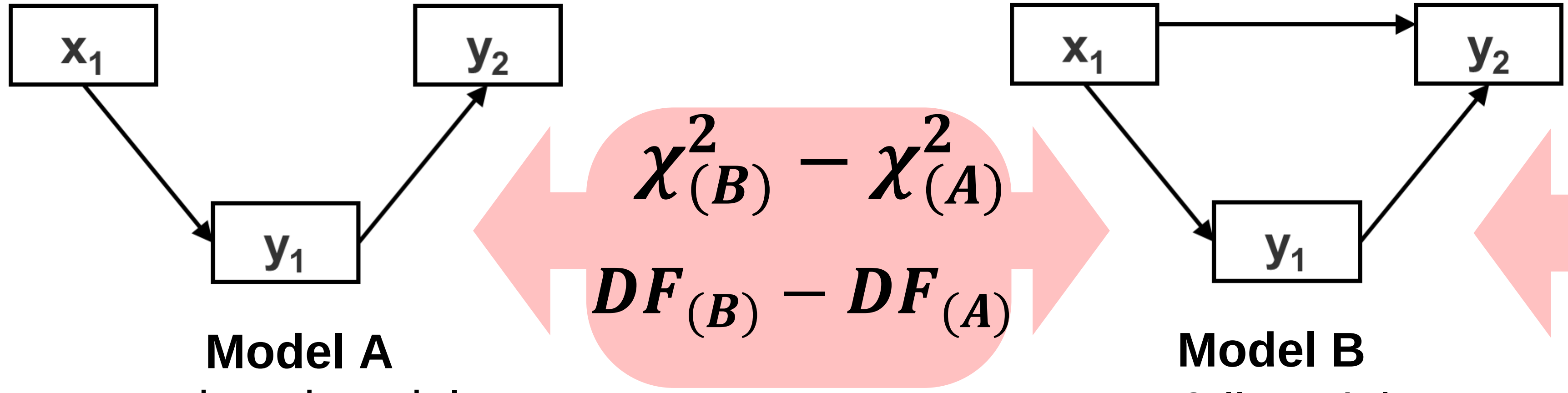
p

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

Goodness of fit

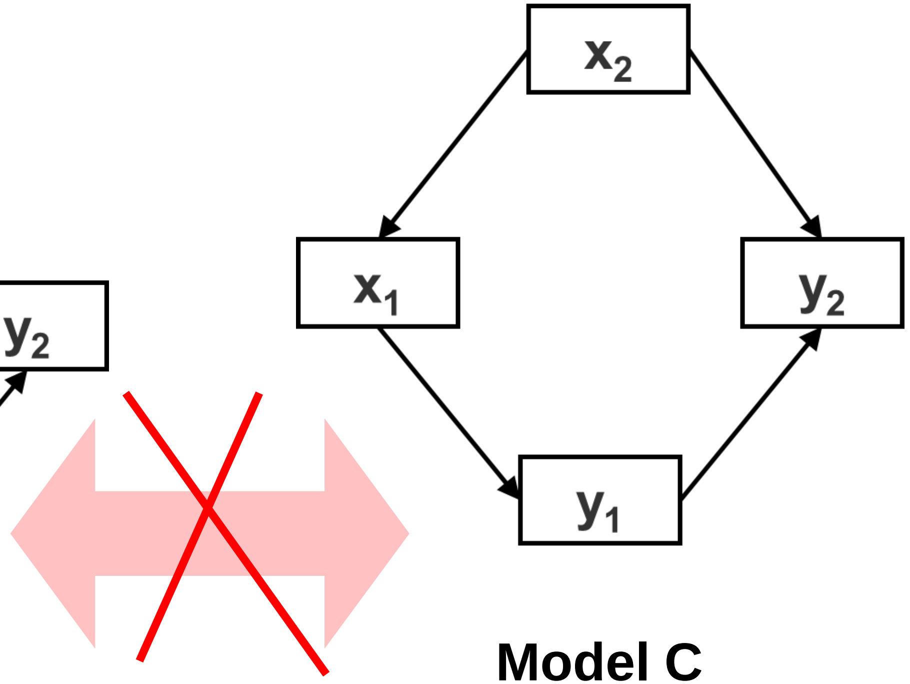
χ^2 – difference test:

- 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

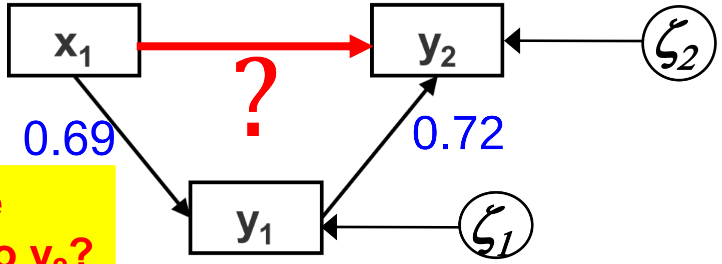
A is nested in B

- the same variables but less 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

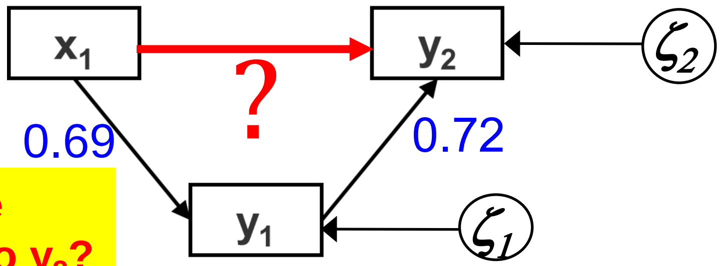
residual

0.444-0.496=**-0.052**

χ^2 statistics:

χ^2 = 1.06, DF=1, n=100, p = 0.3

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:

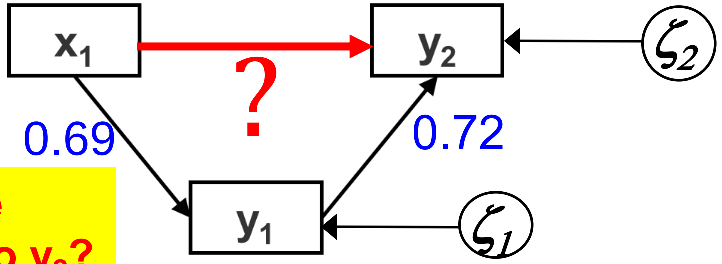
χ^2 = 1.06, DF=1, n=100, p = 0.3

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?

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

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.

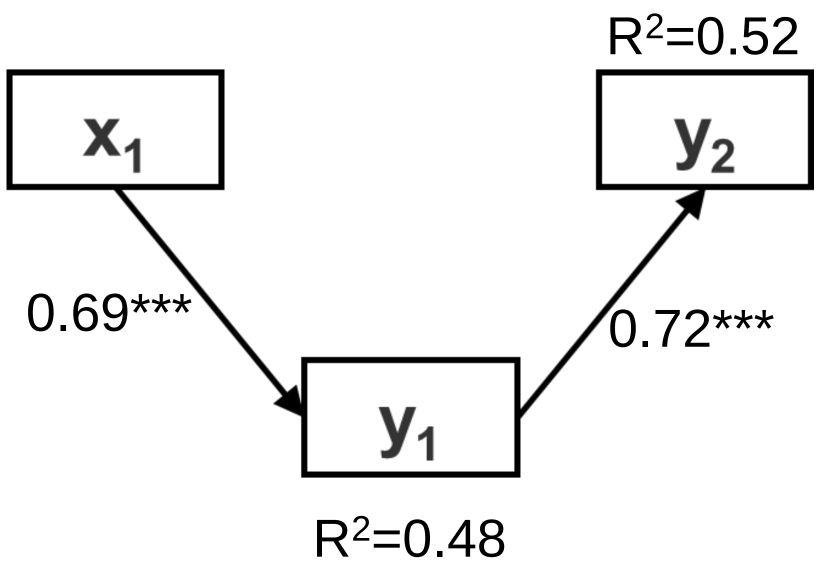
Goodness of fit

… 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 5 Task 3

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.

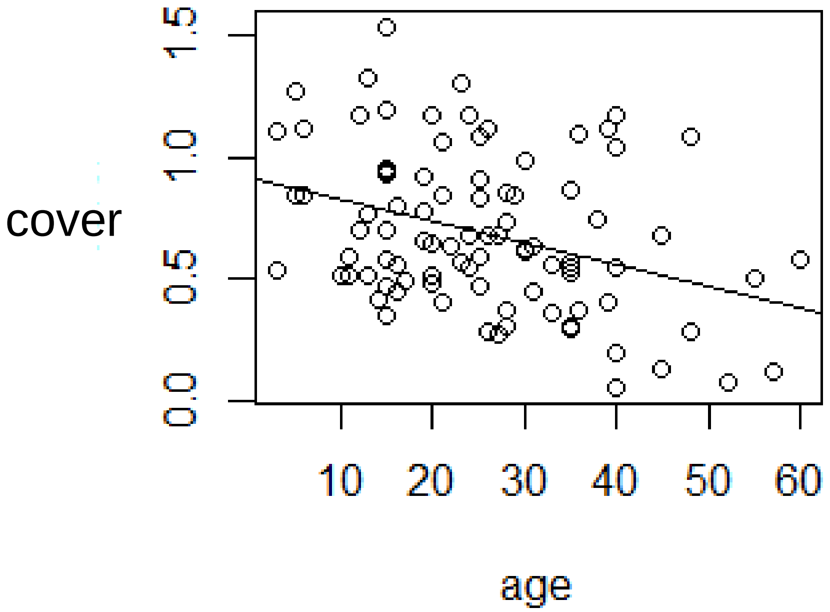
A number of measures were taken, including:

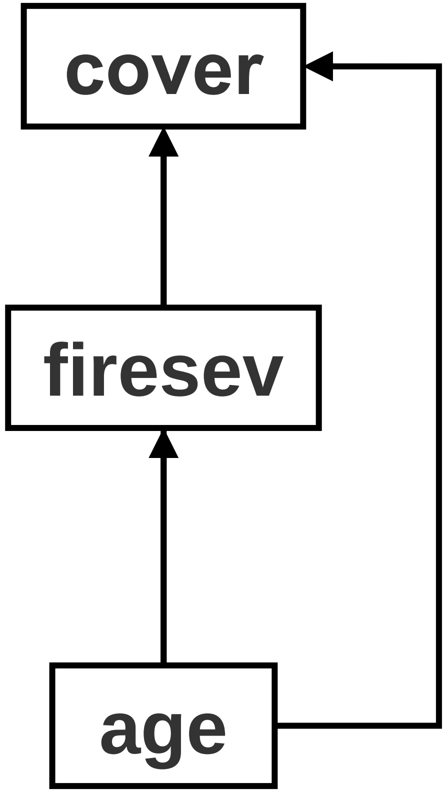
- Vegetation cover **"cover"**
- Age of stands that burned **"age"**
- Fire severity **"firesev"**

```
# 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 5 Task 3





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 5 Task 3

4. Fit the model on Fig. 1 in lavaan and get the path coefficients.

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

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

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

7. For the final model calculate direct, indirect and total effects of age on cover.

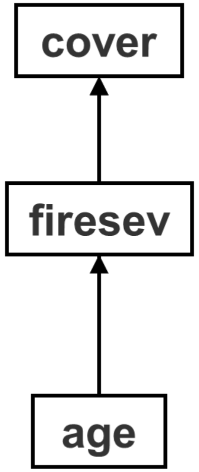


Fig. 1

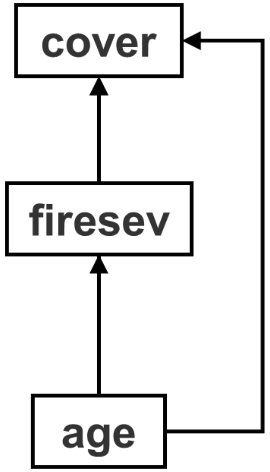


Fig. 2