Introduction Structural Equation Modeling in R

11.02.2022 – Rémy Beugnon

Who am I?

Ph.D. in the TreeDì graduate school @iDiv

Working on:

Ph.D.: Tree diversity effect on forest carbon cycle in subtropical forest

PostDoc: Vegetation diversity mediation of microclimatic fluctuations.

Education:

Studies agricultural engineering (France)

Master in Ecology and Evolution (France)

Ph.D. in Ecology (Leipzig)

Find me here:

email: remy.beugnon@idiv.de

Twitter: @BeugnonRemy

Web: https://remybeugnon.netlify.app

Goals

1. Understand the concept behind Structural Equation Modeling
2. Build SEMs: protocol and rules
3. Fit SEMs in R
4. Analyze and show fit outputs
5. Read and understand SEMs in articles

SUMMARY

This morning

- Introduction: SEM?

break

- Build SEMs
- Methods to fit SEMs

lunch break

This afternoon

- Fit SEMs in R
- Read your results

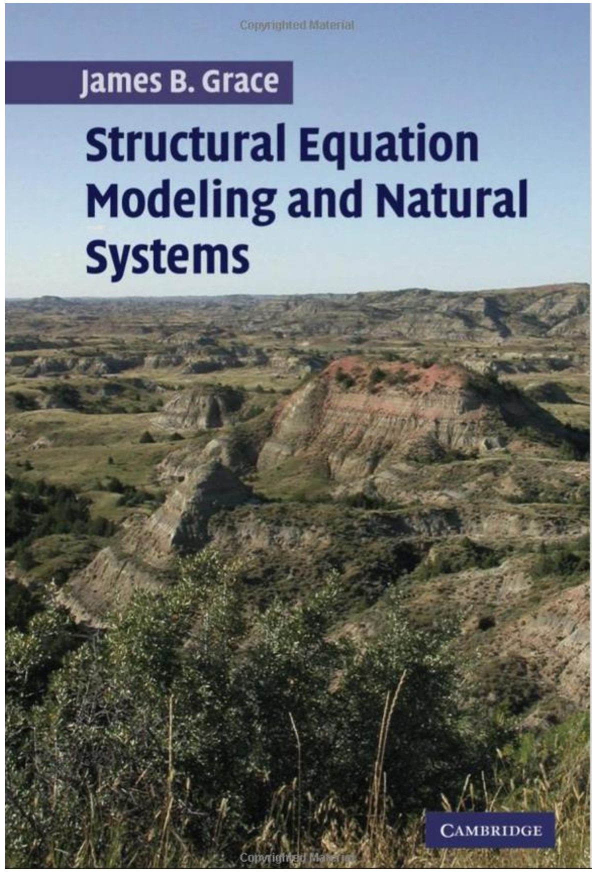
break

- Show your results

Read SEMs in articles

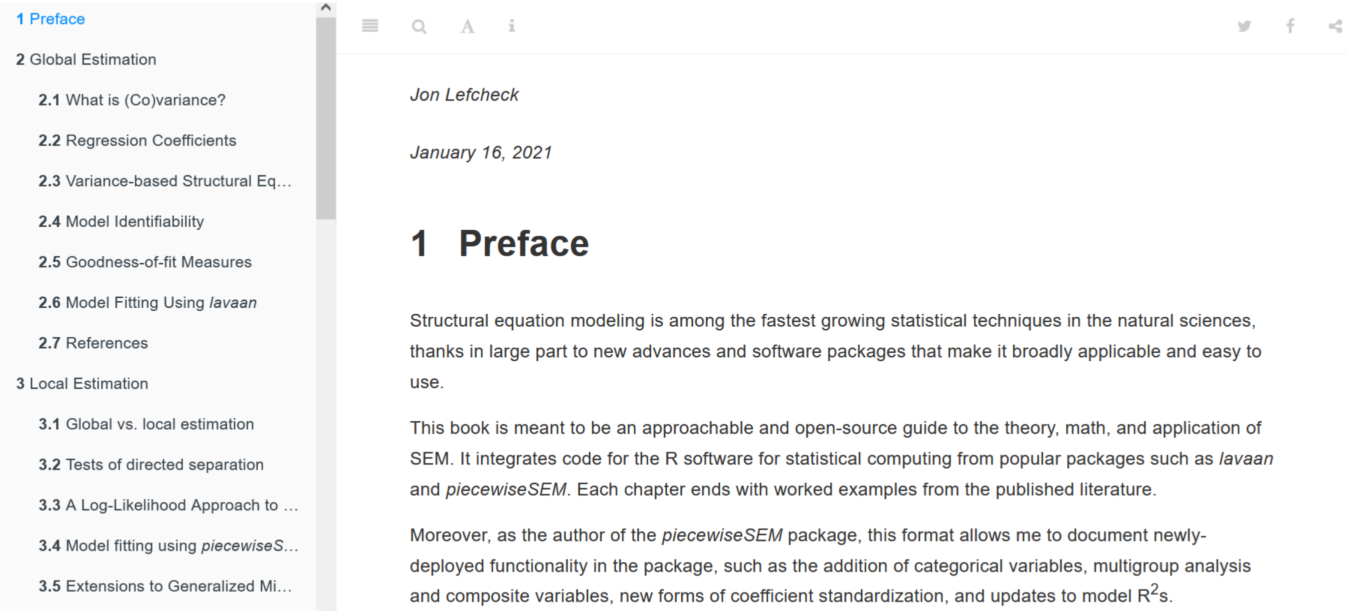
- Add-ons

Structural Equation Modeling



Global estimations (*lavaan* R package)

Jon Lefcheck online book
https://jslefche.github.io/sem_book/index.html



Jon Lefcheck

January 16, 2021

1 Preface

Structural equation modeling is among the fastest growing statistical techniques in the natural sciences, thanks in large part to new advances and software packages that make it broadly applicable and easy to use.

This book is meant to be an approachable and open-source guide to the theory, math, and application of SEM. It integrates code for the R software for statistical computing from popular packages such as *lavaan* and *piecewiseSEM*. Each chapter ends with worked examples from the published literature.

Moreover, as the author of the *piecewiseSEM* package, this format allows me to document newly-deployed functionality in the package, such as the addition of categorical variables, multigroup analysis and composite variables, new forms of coefficient standardization, and updates to model R^2s.

Global and local estimations (*piecewiseSEM* R package)

Structural Equation Modeling

Modeling our world…

Structural Equation Modeling

Modeling our world…

… by using a set of equations …

Structural Equation Modeling

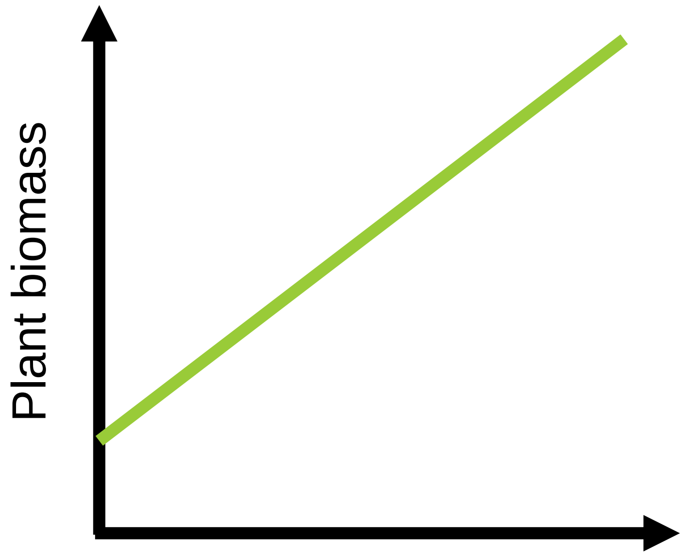
Modeling our world…

… by using a set of equations …

… in a structured order.

Structural Equation Modeling

Adding water to my plants makes them grow

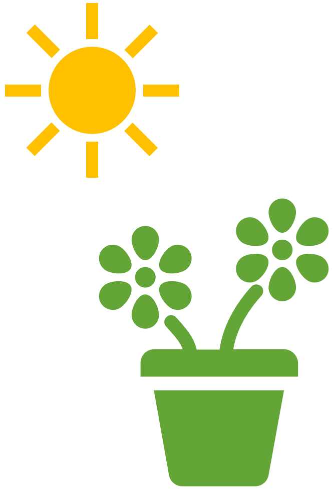


$$plant\ biomass \sim water$$

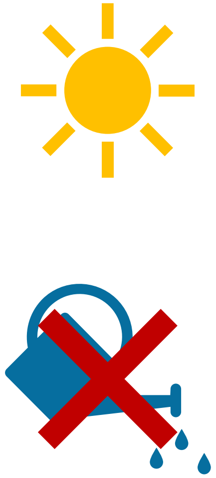
$$plant.biomass \sim \mu + \alpha \times water + \varepsilon$$

Structural Equation Modeling

Adding water to my plants makes them grow



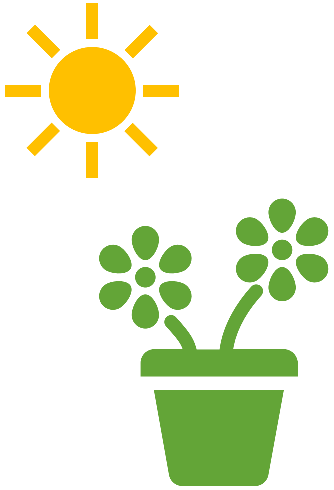
Warming my plants makes them grow



Warming reduces the water availability

Structural Equation Modeling

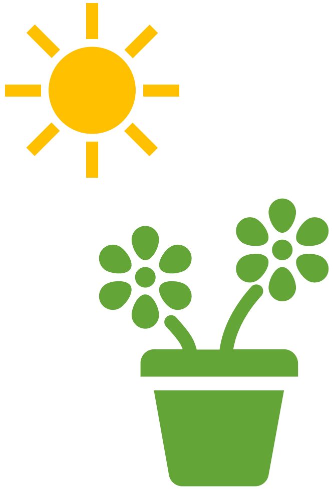
Adding water to my plants makes them grow



Warming my plants makes them grow

Structural Equation Modeling

Adding water to my plants makes them grow



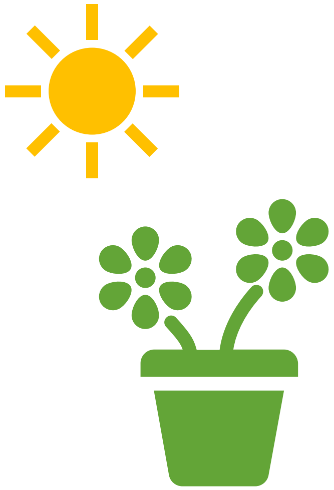
Warming my plants makes them grow

$$plant\ biomass \sim water + temperature$$

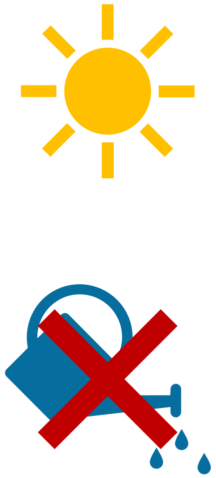
What is an SEM?

Structural Equation Modeling

Adding water to my plants makes them grow



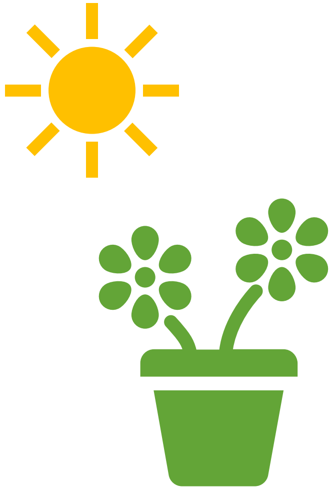
Warming my plants makes them grow

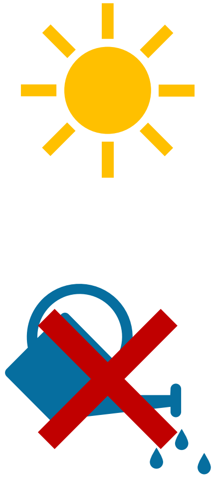


Warming reduces the water availability

$$\textit{plant biomass} \sim \textit{water} + \textit{temperature}$$

Structural Equation Modeling





Adding water to my plants makes them grow

Warming my plants makes them grow

Warming reduces the water availability

$$plant\ biomass \sim water + temperature$$

$$water \sim temperature$$

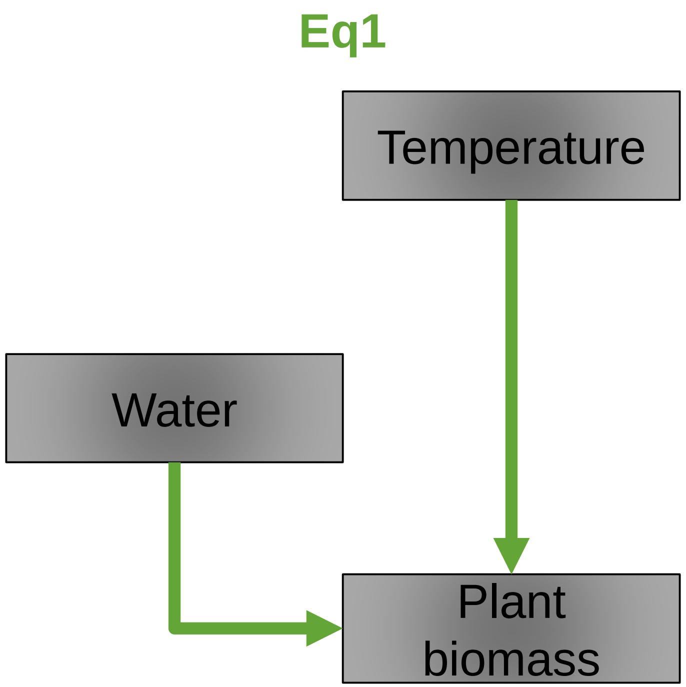
What is an SEM?

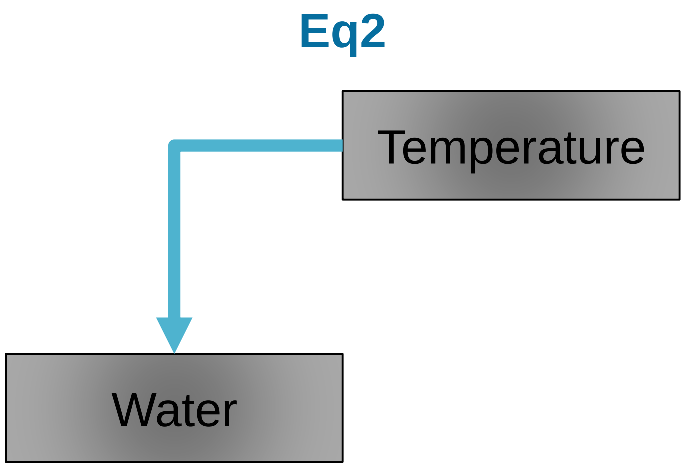
Structural Equation Modeling

Eq1: $plant\ biomass \sim water + temperature$
Eq2: $water \sim temperature$

Structural Equation Modeling

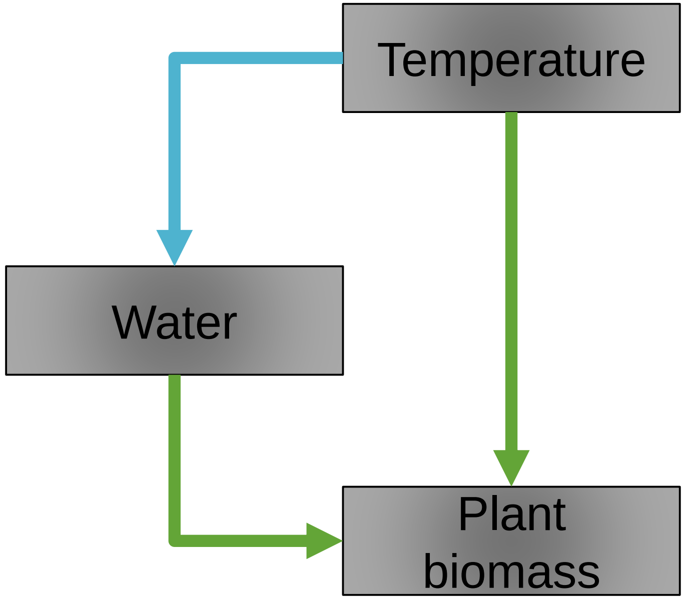
Eq1: $plant\ biomass \sim water + temperature$
Eq2: $water \sim temperature$



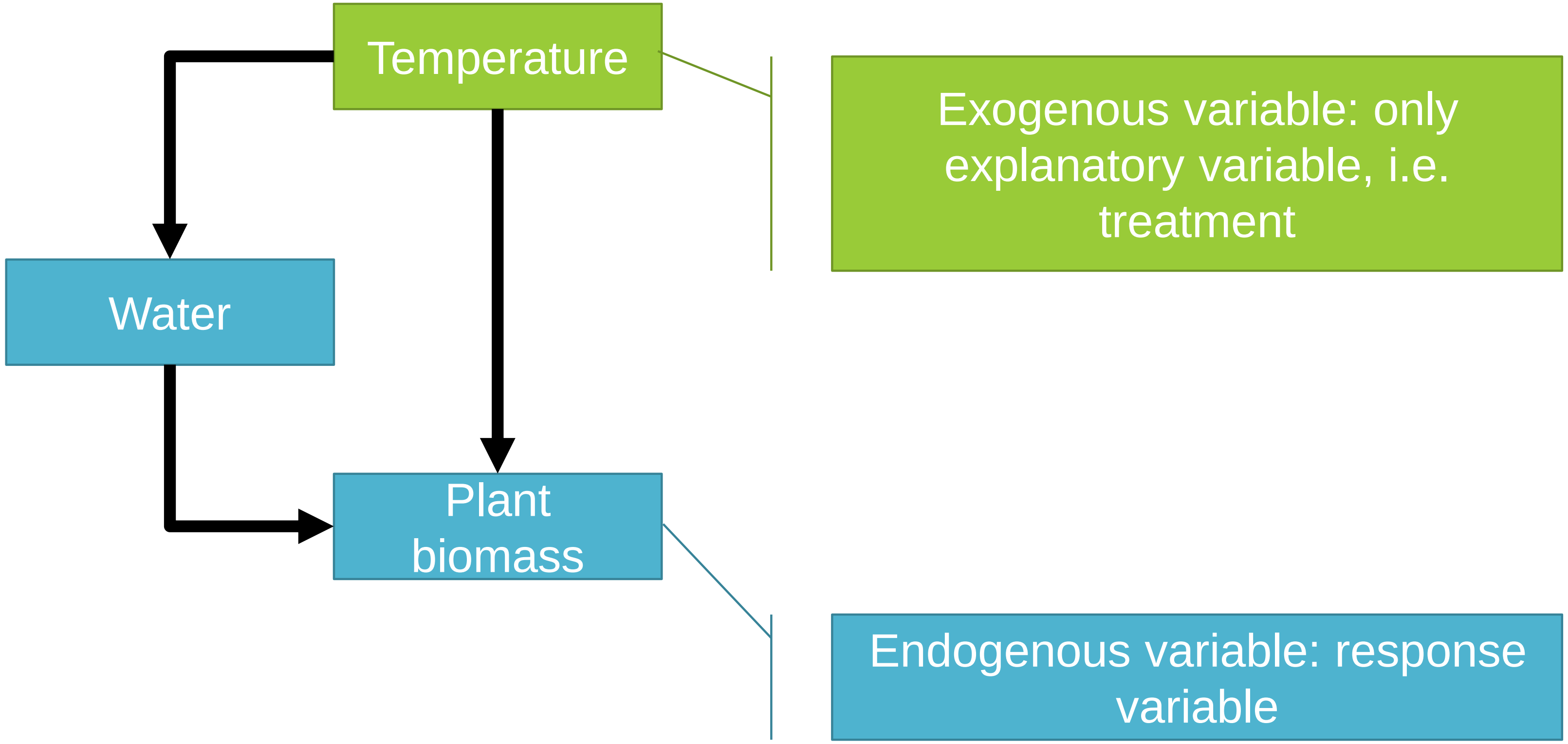


Structural Equation Modeling

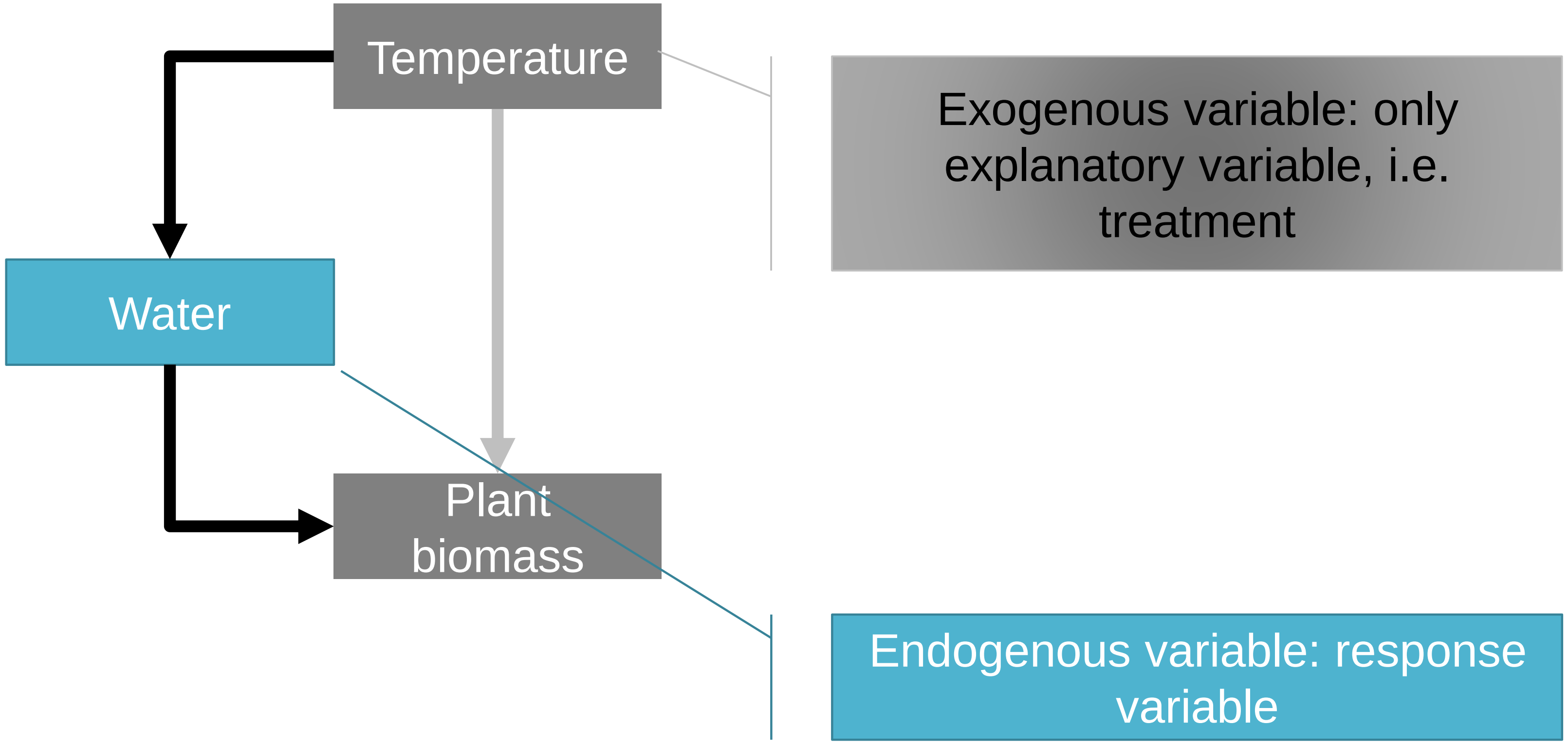
Eq1: $plant\ biomass \sim water + temperature$
Eq2: $water \sim temperature$



Exogenous *vs.* endogenous variables

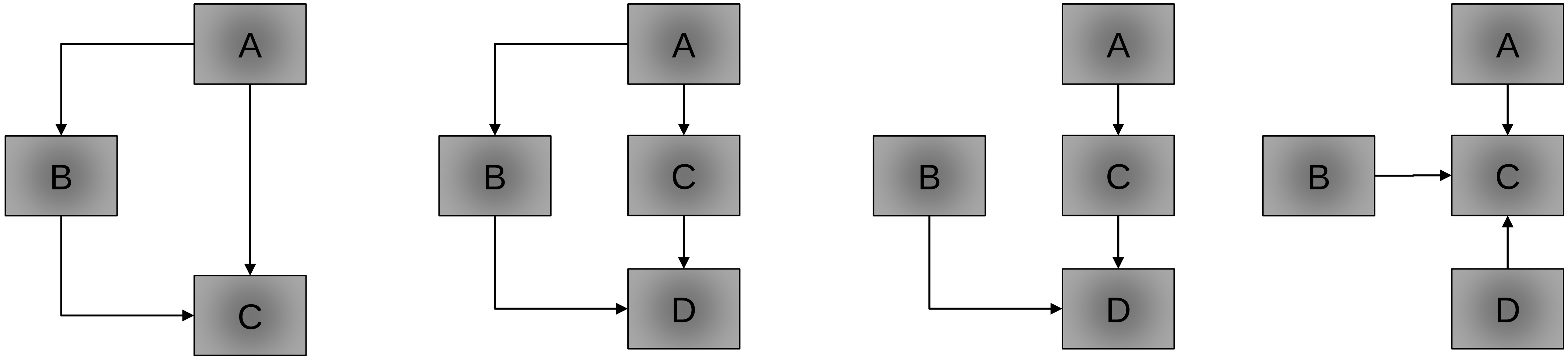


Exogenous *vs.* endogenous variables



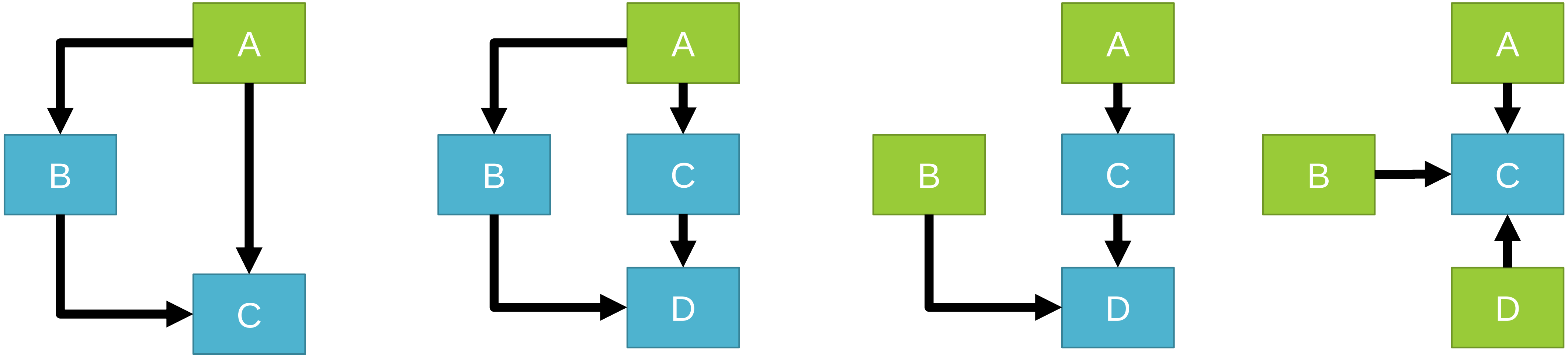
Practical # 1

1 - Identify exogenous and endogenous variables



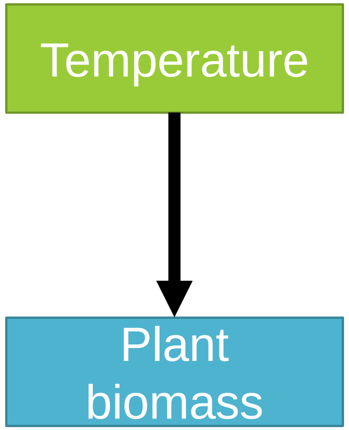
Practical # 1

1 - Identify exogenous and endogenous variables



Causation *vs.* Correlation

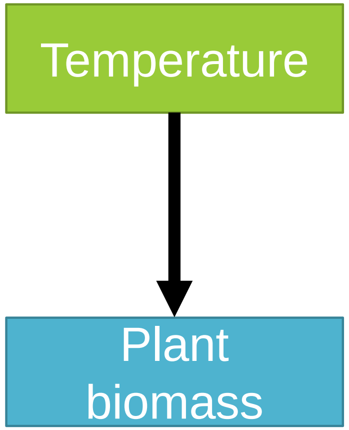
Causal relation



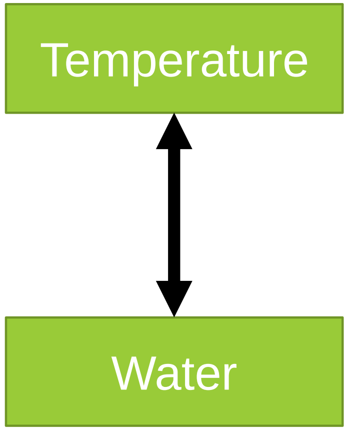
$$Plant\ biomass \sim Temperature$$

Causation *vs.* Correlation

Causal relation $\neq$ **Correlation**



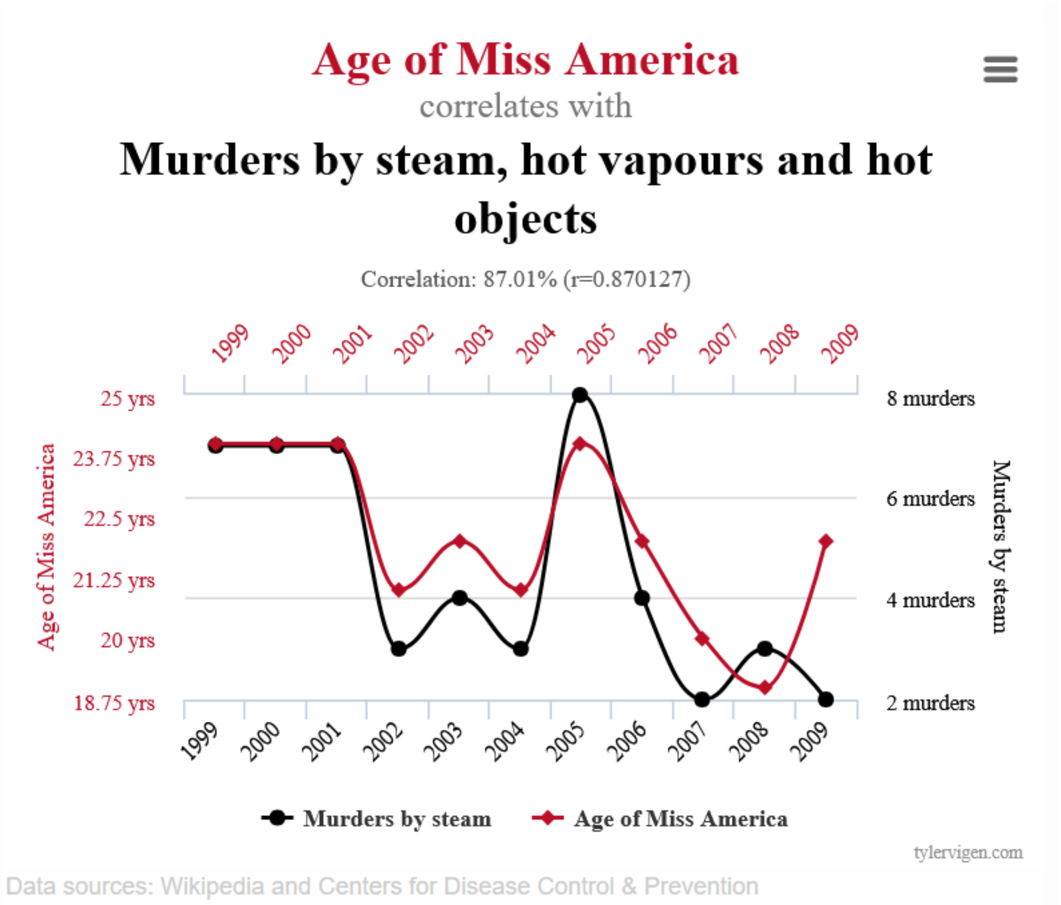
?



$Plant\ biomass \sim Temperature$

$cor(Temperature, Water)$

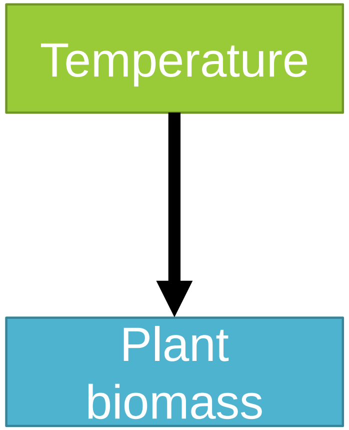
Example:



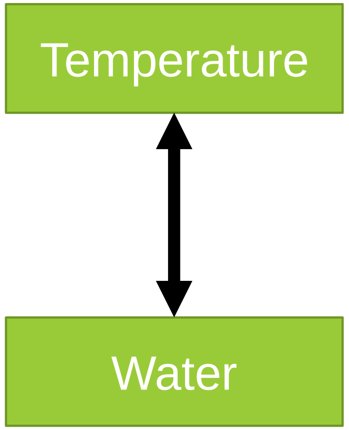
If you have ne clue about causality in your dataset check graphical lasso for instance

Causation *vs.* Correlation

Causal relation $\neq$ **Correlation**

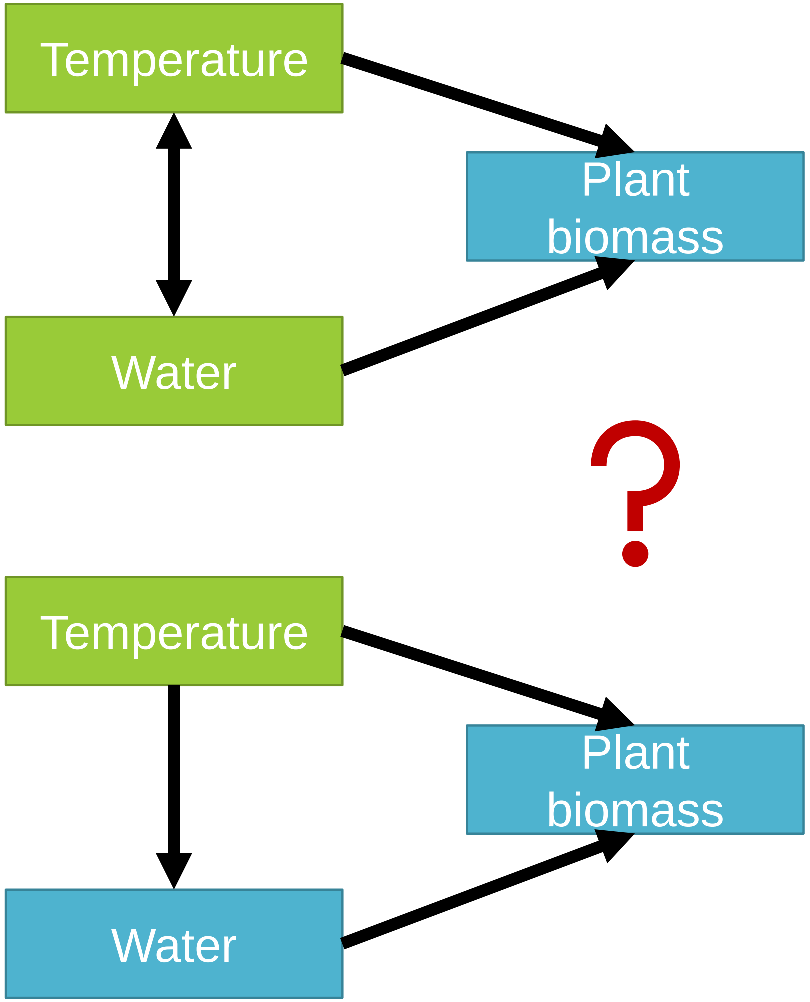


?



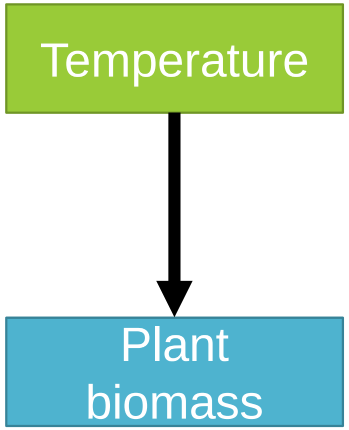
$Plant\ biomass \sim Temperature$

$cor(Temperature, Water)$

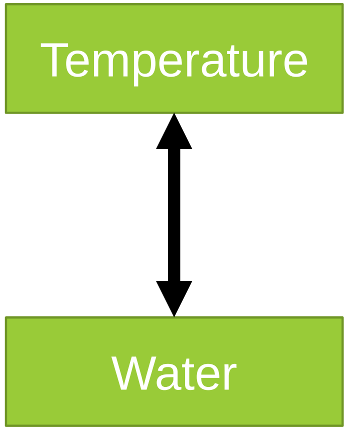


Causation *vs.* Correlation

Causal relation $\neq$ **Correlation**



?



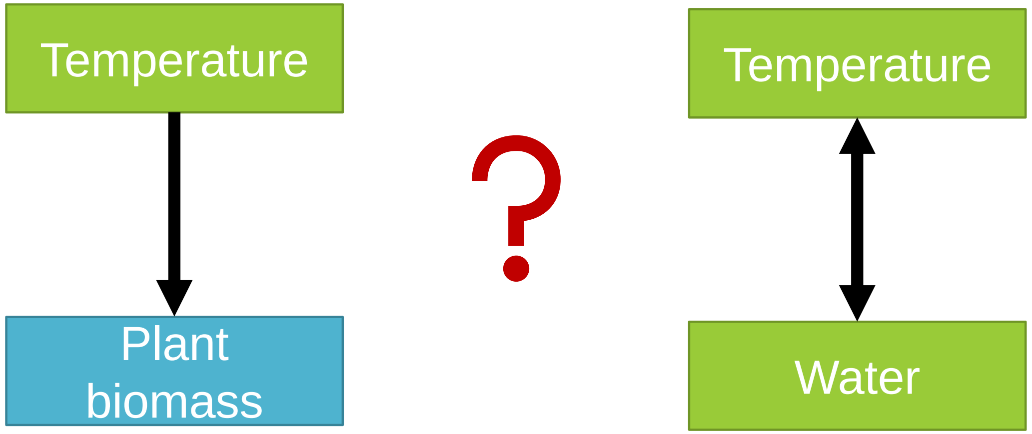
$Plant\ biomass \sim Temperature$

$cor(Temperature, Water)$

"SEM results should not be taken as proof of causal claims, but instead as evaluations or tests of models representing causal hypotheses" ---James Grace

Causation *vs.* Correlation

Causal relation $\neq$ **Correlation**

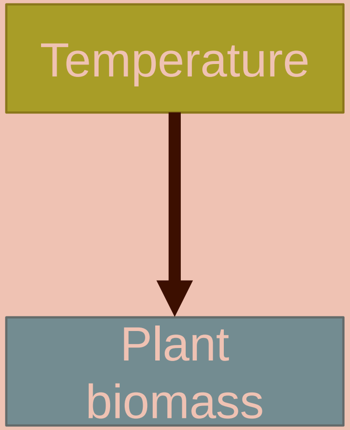


$Plant\ biomass \sim Temperature$

$cor(Temperature, Water)$

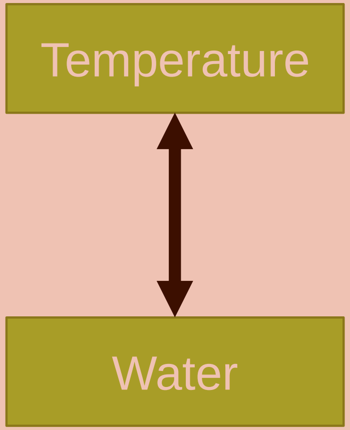
DANGER ZONE

Causal relation



$Plant\ biomass \sim Temperature$

Correlation

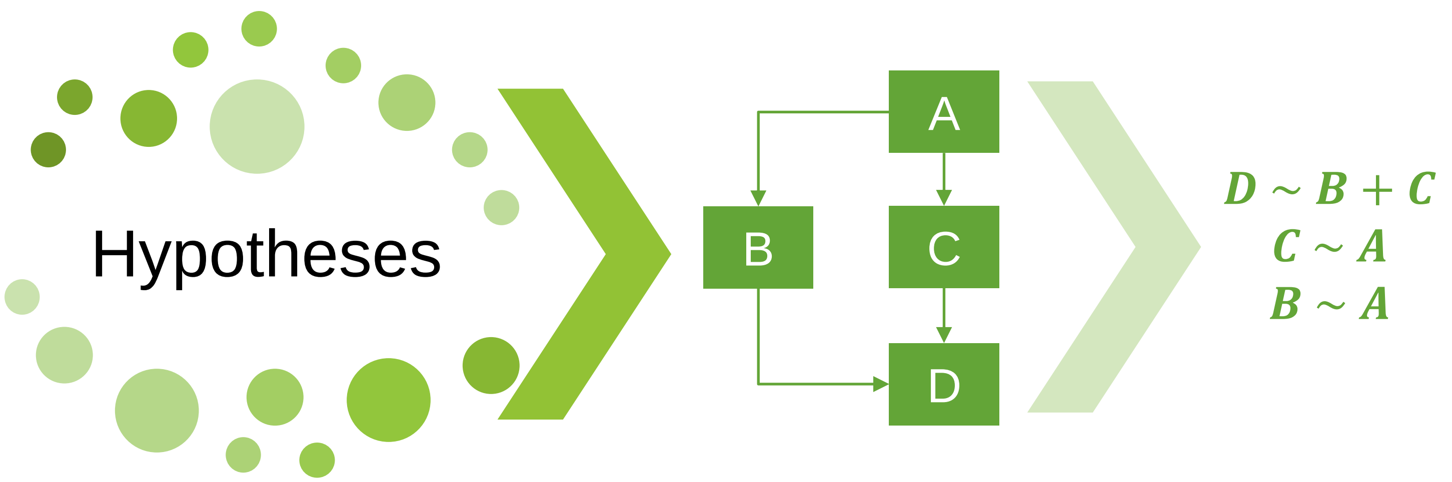


$cor(Temperature, Water)$

BACKUP ALL YOUR HYPOTHESES BY LITERATURE

If you have no clue about causality in your dataset check graphical lasso for instance

Build an SEM



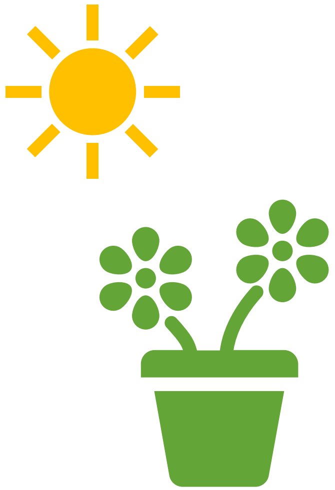
Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

H1

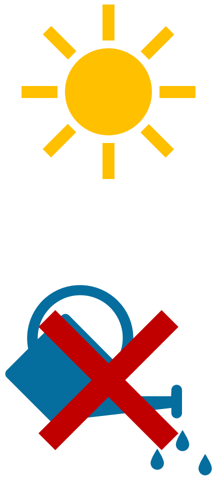
Adding water to my plants makes them grow

H2



Warming my plants makes them grow

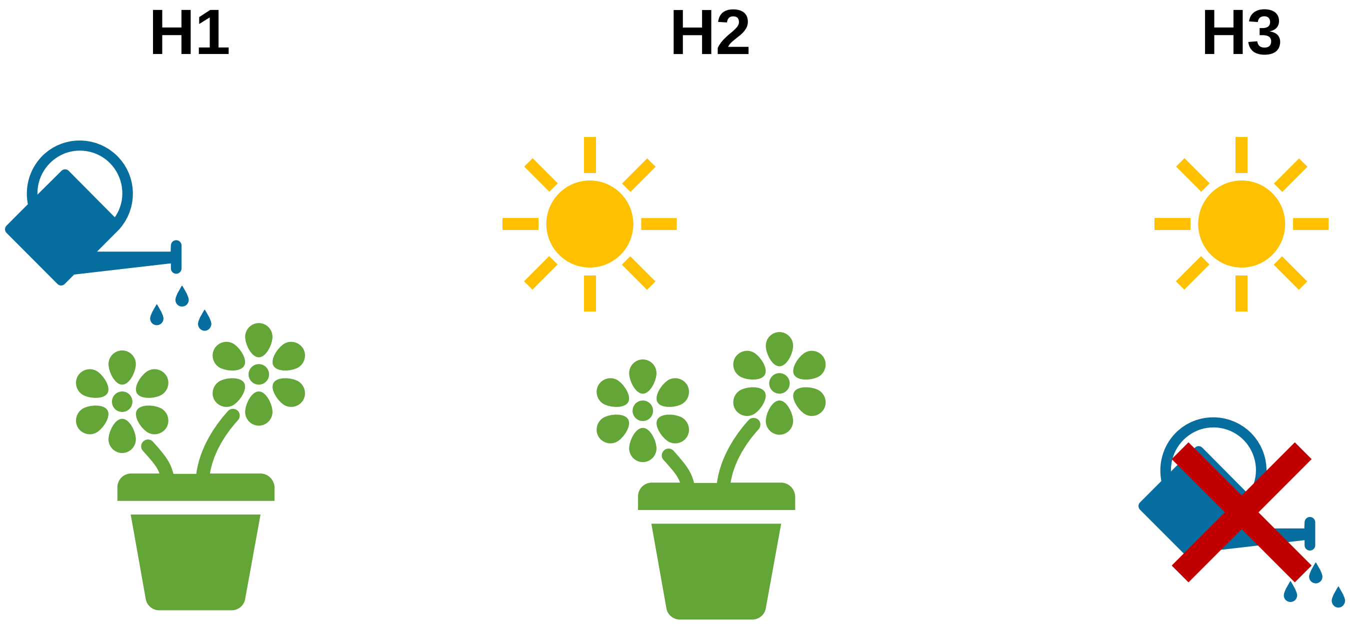
H3



Warming reduces the water availability

Protocol to build and fit SEMs in R

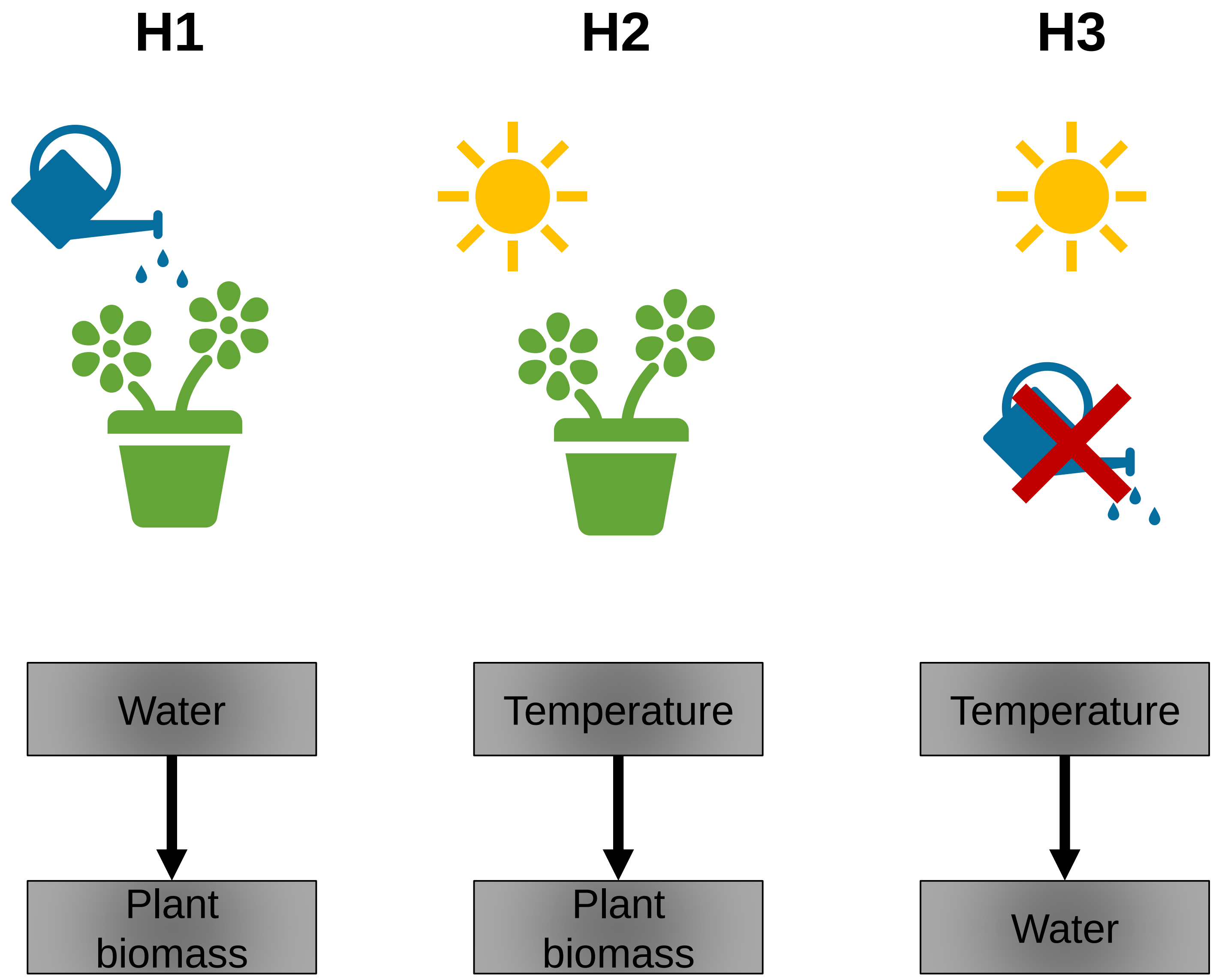
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



Hypotheses		References
H1	Adding water to my plants makes them grow	Smith *et al*. 2020
H2	Warming my plants makes them grow	Dupont *et al*. 2006
H3	Warming reduces the water availability	Doe *et al*. 1994

Protocol to build and fit SEMs in R

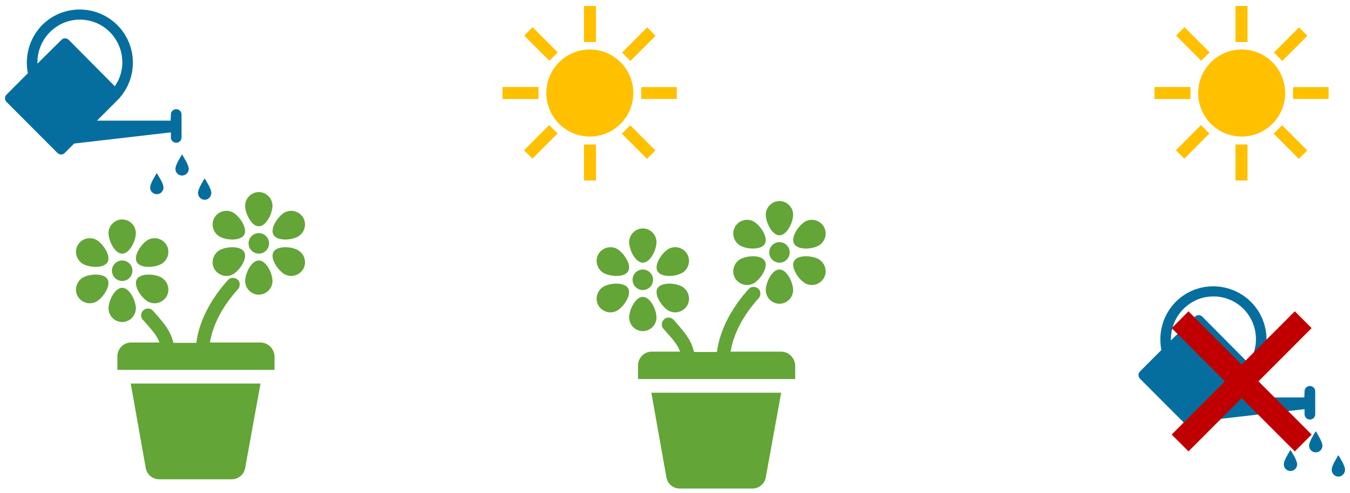
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

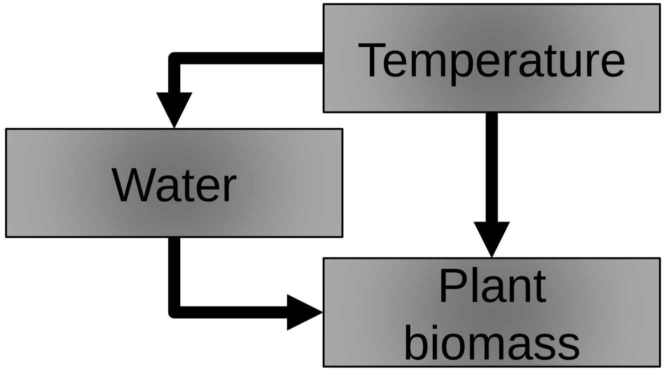


Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

H1 H2 H3



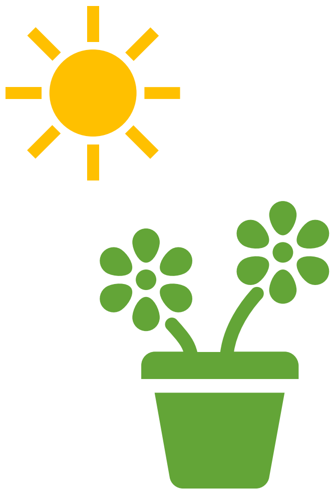


Protocol to build and fit SEMs in R

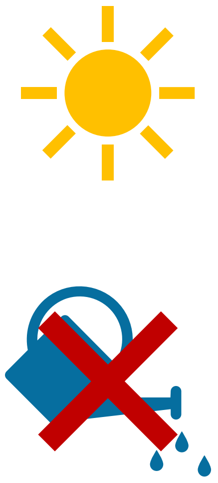
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

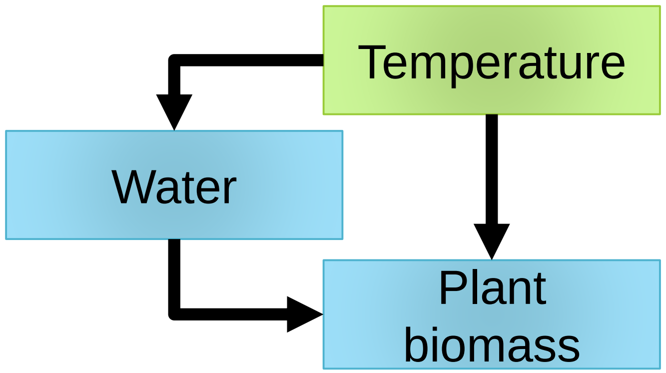
H1

H2



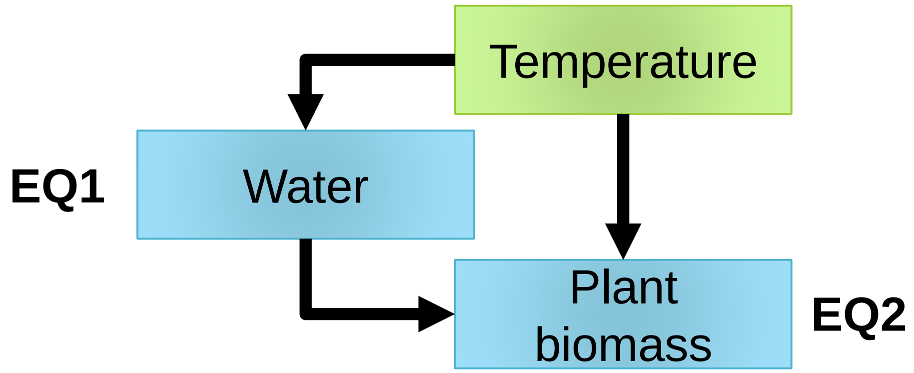
H3





Protocol to build and fit SEMs in R

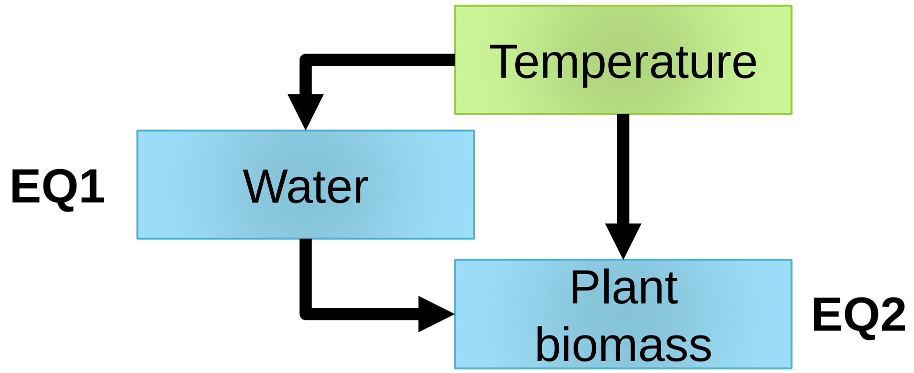
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



1 equation per endogenous variable

Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



EQ1 $plant\ biomass \sim water + temperature$

EQ2 $water \sim temperature$

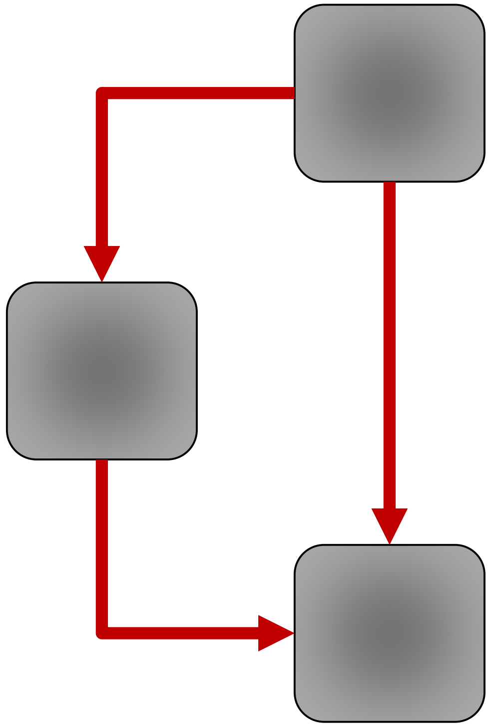
Model saturation

Can I fit my model?

Model saturation

Can I fit my model?

t-rule: $t \leq \frac{n(n+1)}{2}$



n = Known: variables (n = 3)

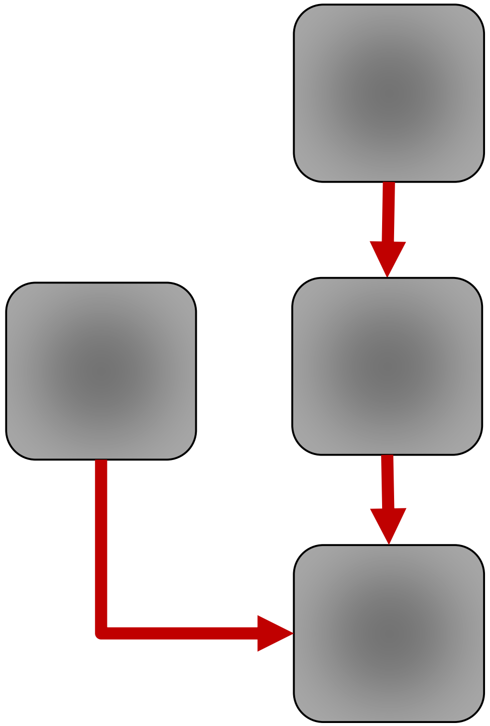
t = Unknown: relations + variances
(t = 3 + 3 = 6)

$$t = 6 \leq 6 = \frac{3 \times (3+1)}{2}$$

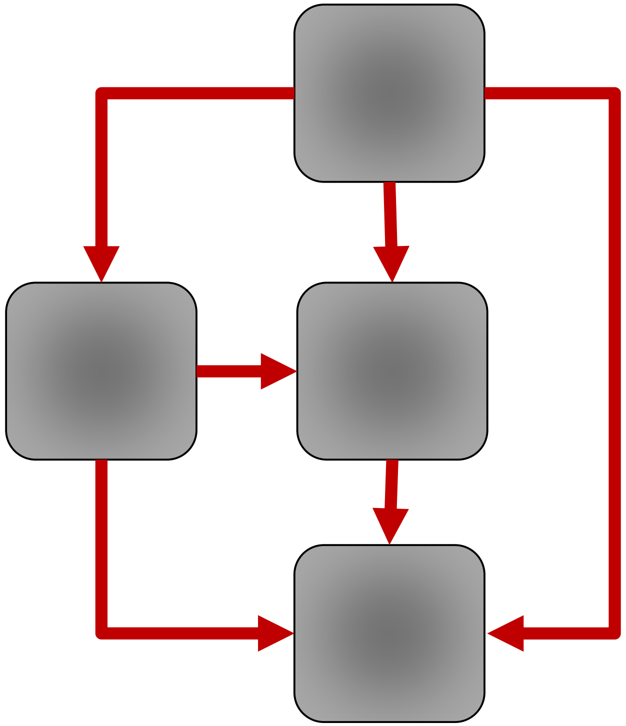
Model saturation

Can I fit my model?

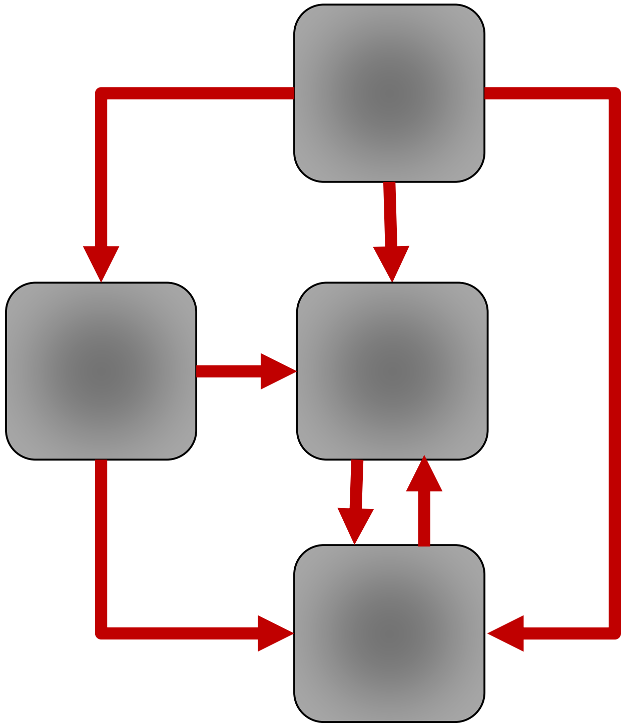
t-rule: $t \leq \frac{n(n+1)}{2}$



$$t = 7 < 10 = \frac{4 \times (4+1)}{2}$$



$$t = 10 = 10 = \frac{4 \times (4+1)}{2}$$

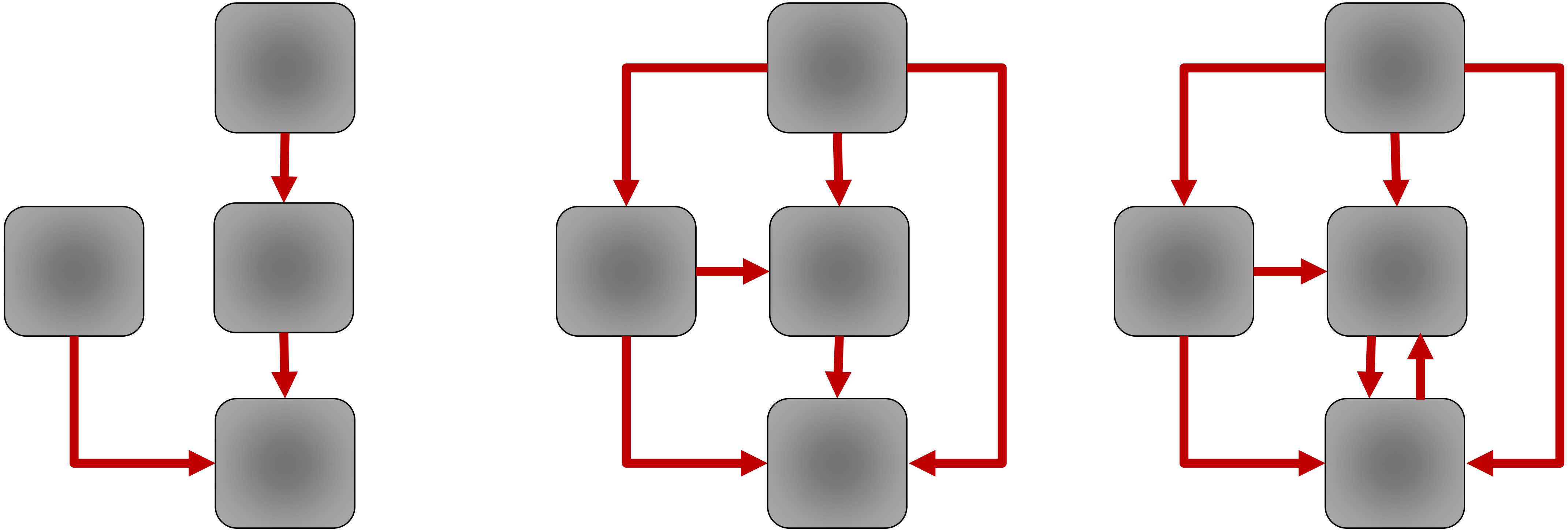


$$t = 11 > 10 = \frac{4 \times (4+1)}{2}$$

Model saturation

Can I fit my model?

t-rule: $t \leq \frac{n(n+1)}{2}$



$$t = 7 < 10 = \frac{4 \times (4+1)}{2}$$

Undersaturated

$$t = 10 = 10 = \frac{4 \times (4+1)}{2}$$

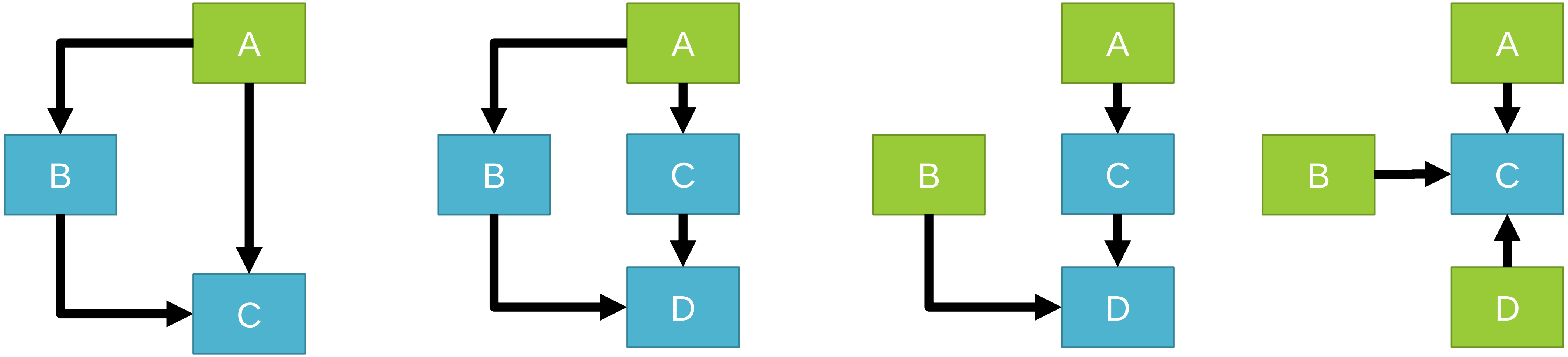
Saturated

$$t = 11 > 10 = \frac{4 \times (4+1)}{2}$$

Oversaturated

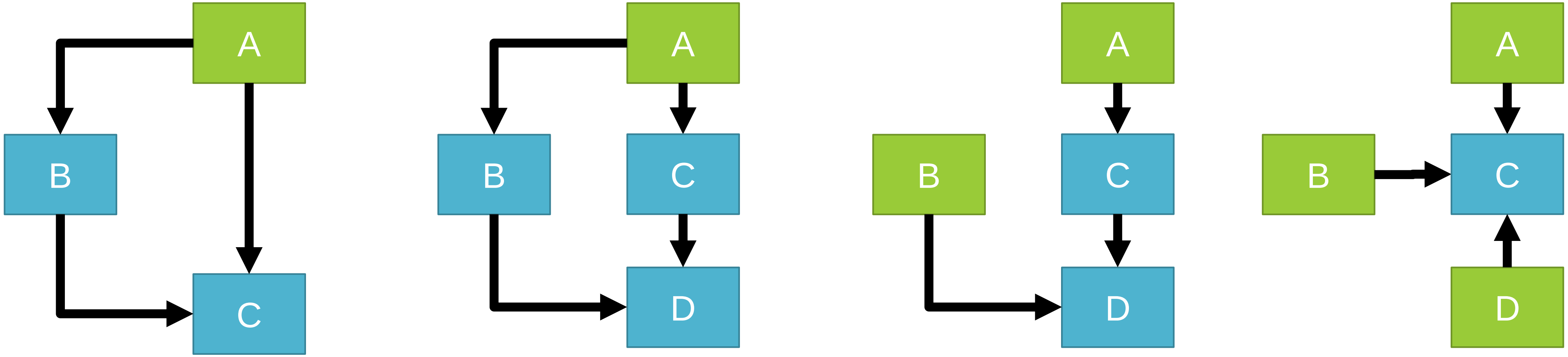
Practical # 1.2

1 - Identify exogenous and endogenous variables



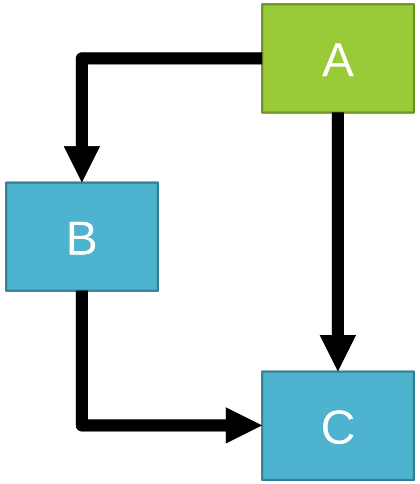
Practical # 1.2

1 - Identify exogenous and endogenous variables
2 – Write the equations

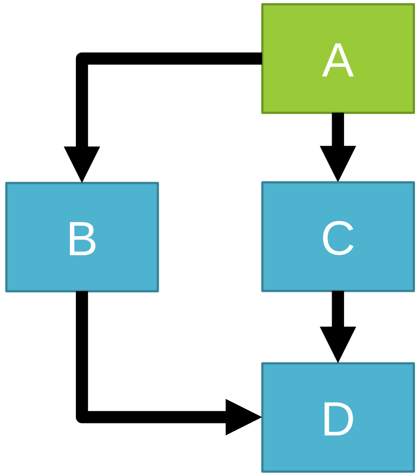


Practical # 1.2

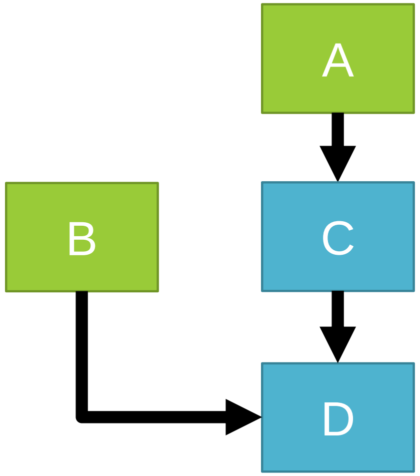
1 - Identify exogenous and endogenous variables
2 – Write the equations



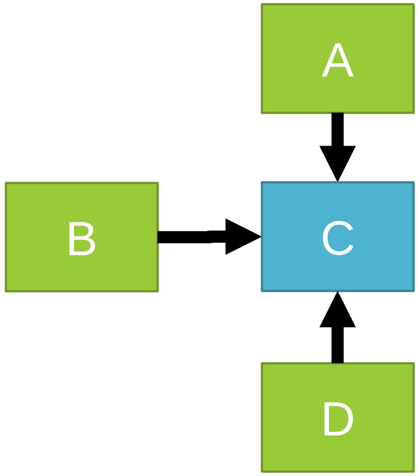
$$C \sim A + B$$
$$B \sim A$$



$$D \sim B + C$$
$$C \sim A$$
$$B \sim A$$



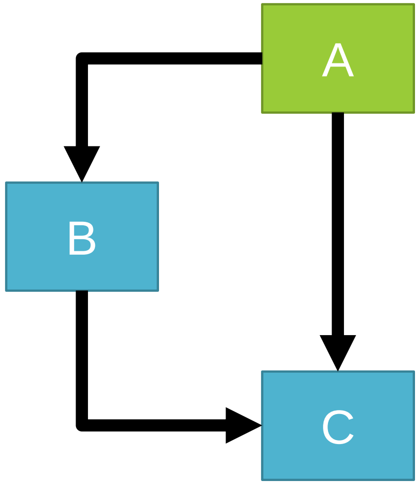
$$D \sim B + C$$
$$C \sim A$$



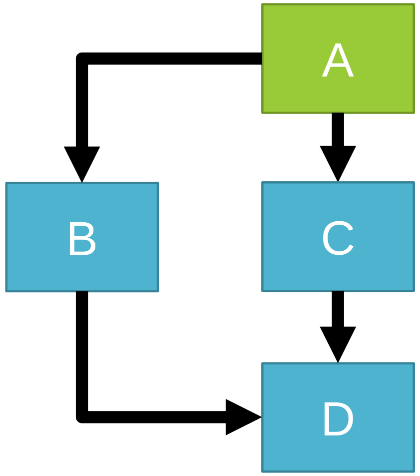
$$C \sim A + B + D$$

Practical # 1.2

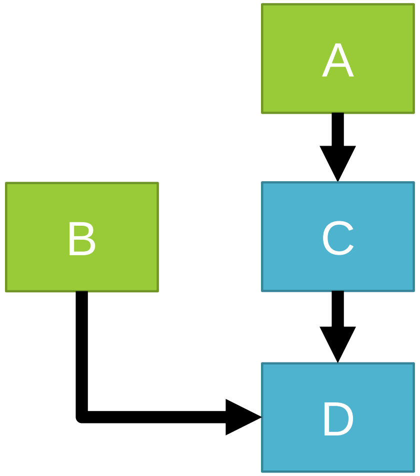
1 - Identify exogenous and endogenous variables
2 – Write the equations
3 – calculate the t-value



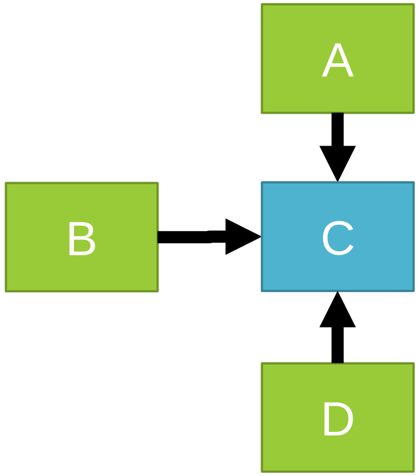
$C \sim A + B$
$B \sim A$



$D \sim B + C$
$C \sim A$
$B \sim A$



$D \sim B + C$
$C \sim A$

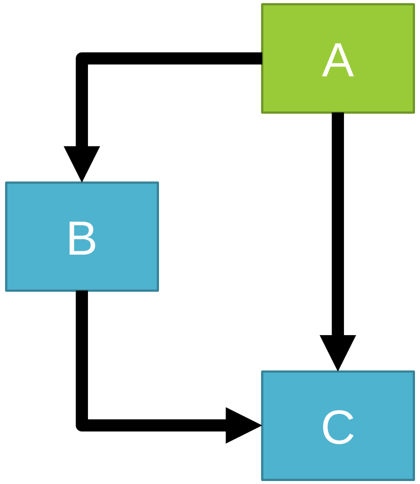


$C \sim A + B + D$

t-rule: $t \leq \frac{n(n+1)}{2}$

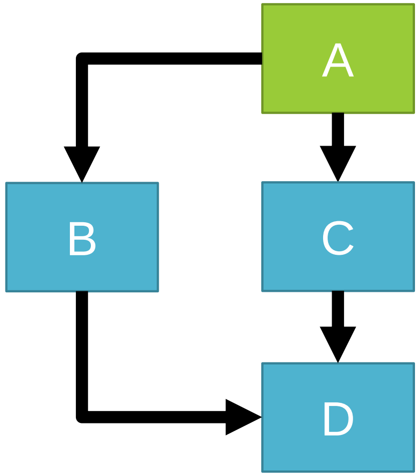
Practical # 1.2

1 - Identify exogenous and endogenous variables
2 – Write the equations
3 – calculate the t-value



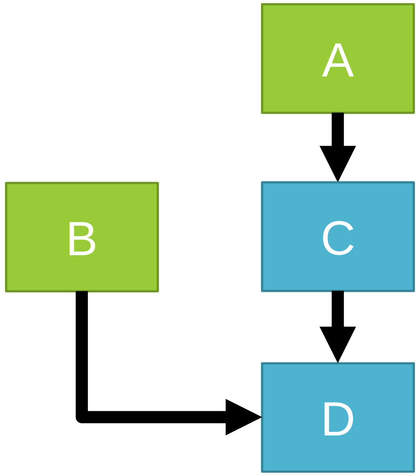
$$C \sim A + B$$
$$B \sim A$$

$$t = 6 \leq 6$$



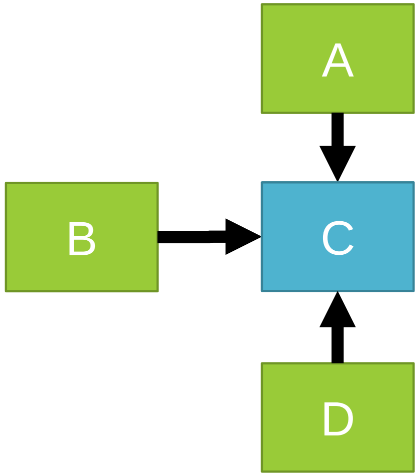
$$D \sim B + C$$
$$C \sim A$$
$$B \sim A$$

$$t = 8 \leq 10$$



$$D \sim B + C$$
$$C \sim A$$

$$t = 7 \leq 10$$



$$C \sim A + B + D$$

$$t = 7 \leq 10$$

Fit an SEM

$$D \sim B + C$$
$$C \sim A$$
$$B \sim A$$

$$D = \mu_D + \alpha_{D1} \times B + \alpha_{D2} \times C + \varepsilon_D$$
$$C = \mu_C + \alpha_C \times A + \varepsilon_C$$
$$B = \mu_B + \alpha_B \times A + \varepsilon_B$$

Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

Fit your simple models and check for anomalies:

$$plant\ biomass \sim water + temperature$$

$$water \sim temperature$$

Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

Fit your simple models and check for anomalies:

$$\boldsymbol{plant\ biomass \sim water + temperature}$$

$$\boldsymbol{water \sim temperature}$$

```
mod1 = lm('plant.biomass ~ water + temperature', data = data)

mod2 = lm('water ~ temperature', data = data)
```

Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

Fit your simple models and check for anomalies:

$$plant\ biomass \sim water + temperature$$

$$water \sim temperature$$

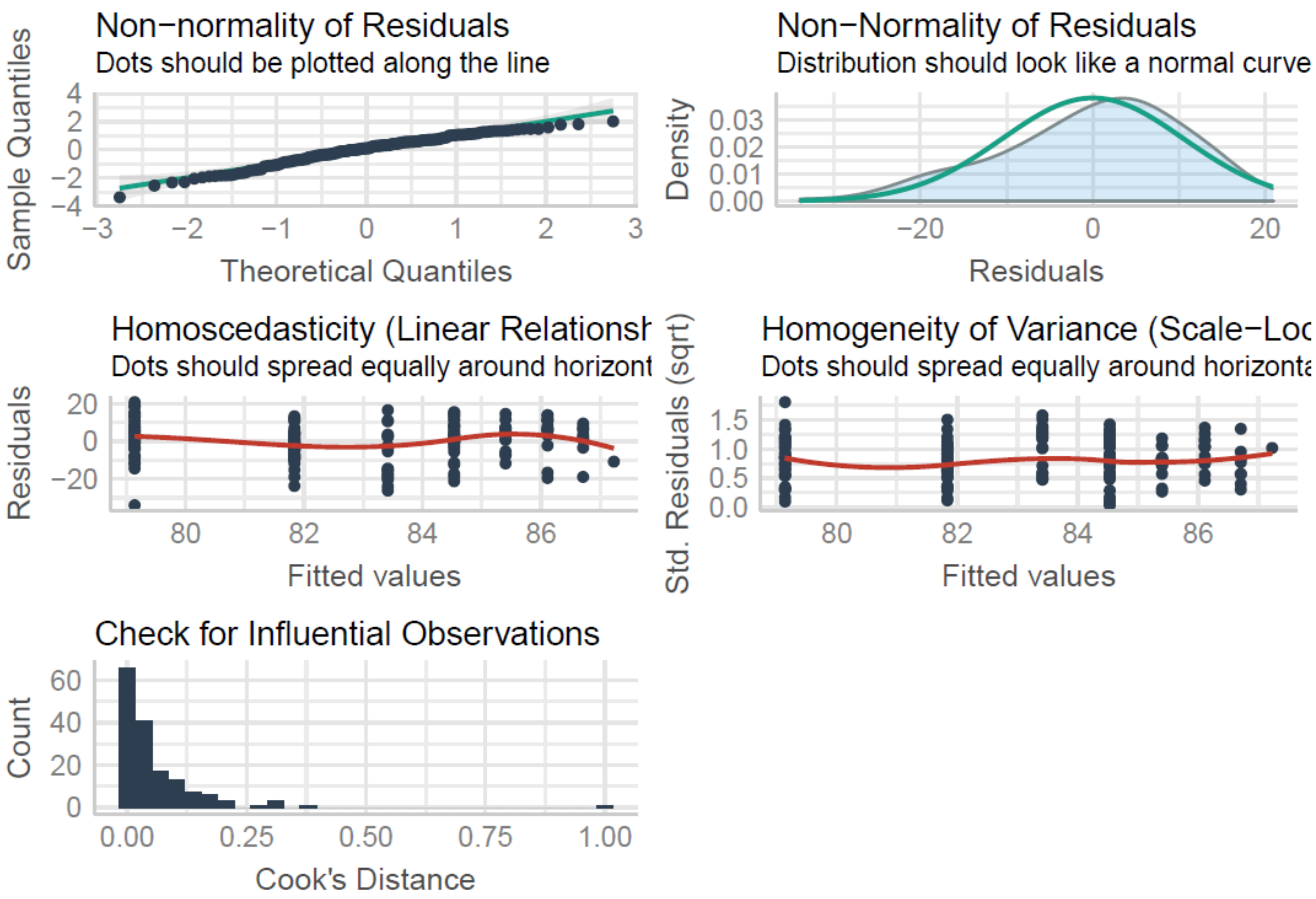
```
mod1 = lm('plant.biomass ~ water + temperature', data = data)

mod2 = lm('water ~ temperature', data = data)
```

Check model quality, e.g. *performance* package

Protocol to build and fit SEMs in R

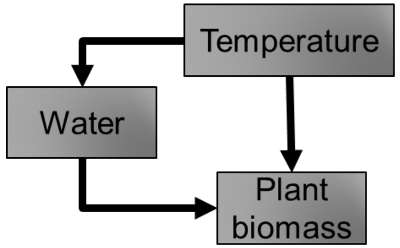
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

Model quality



Protocol to build and fit SEMs in R

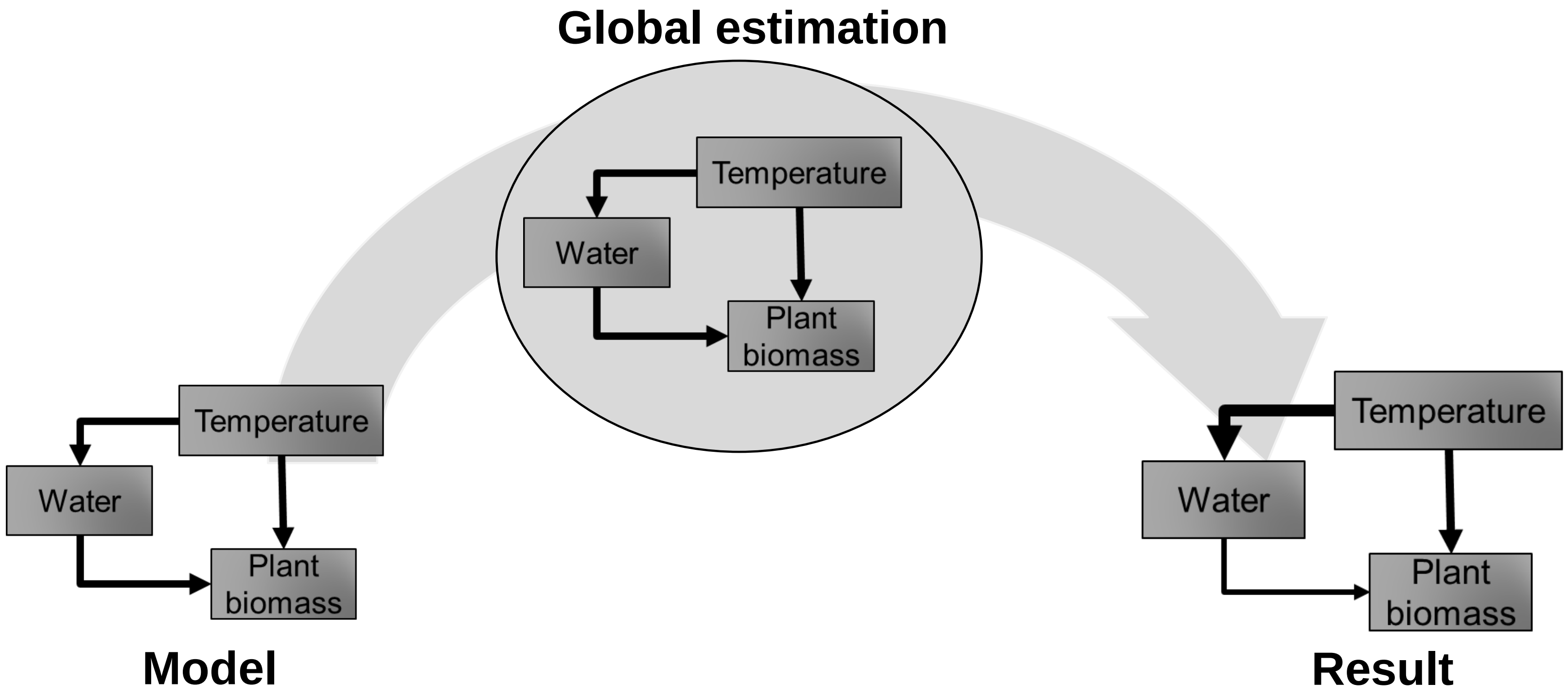
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



Model

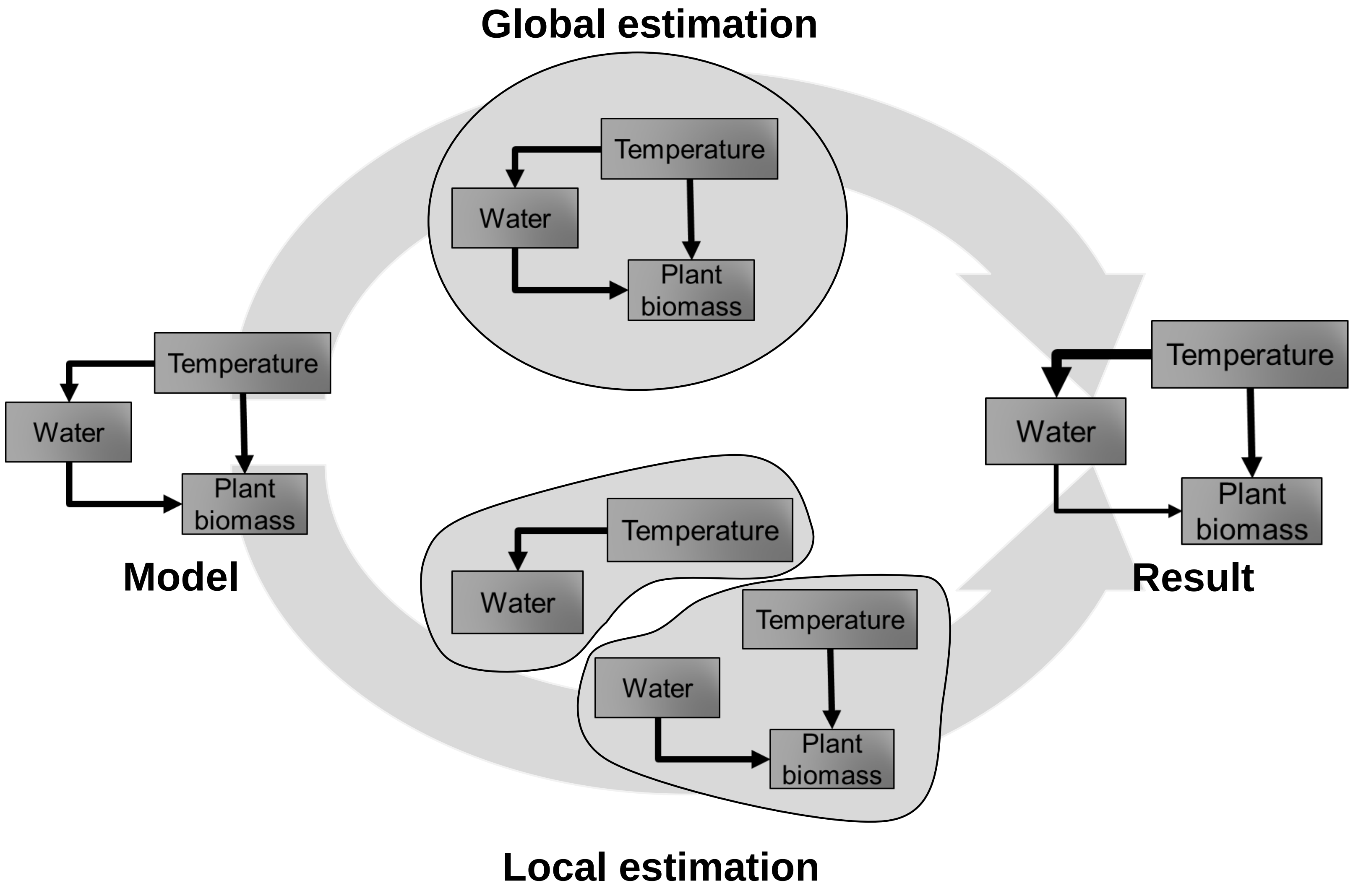
Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



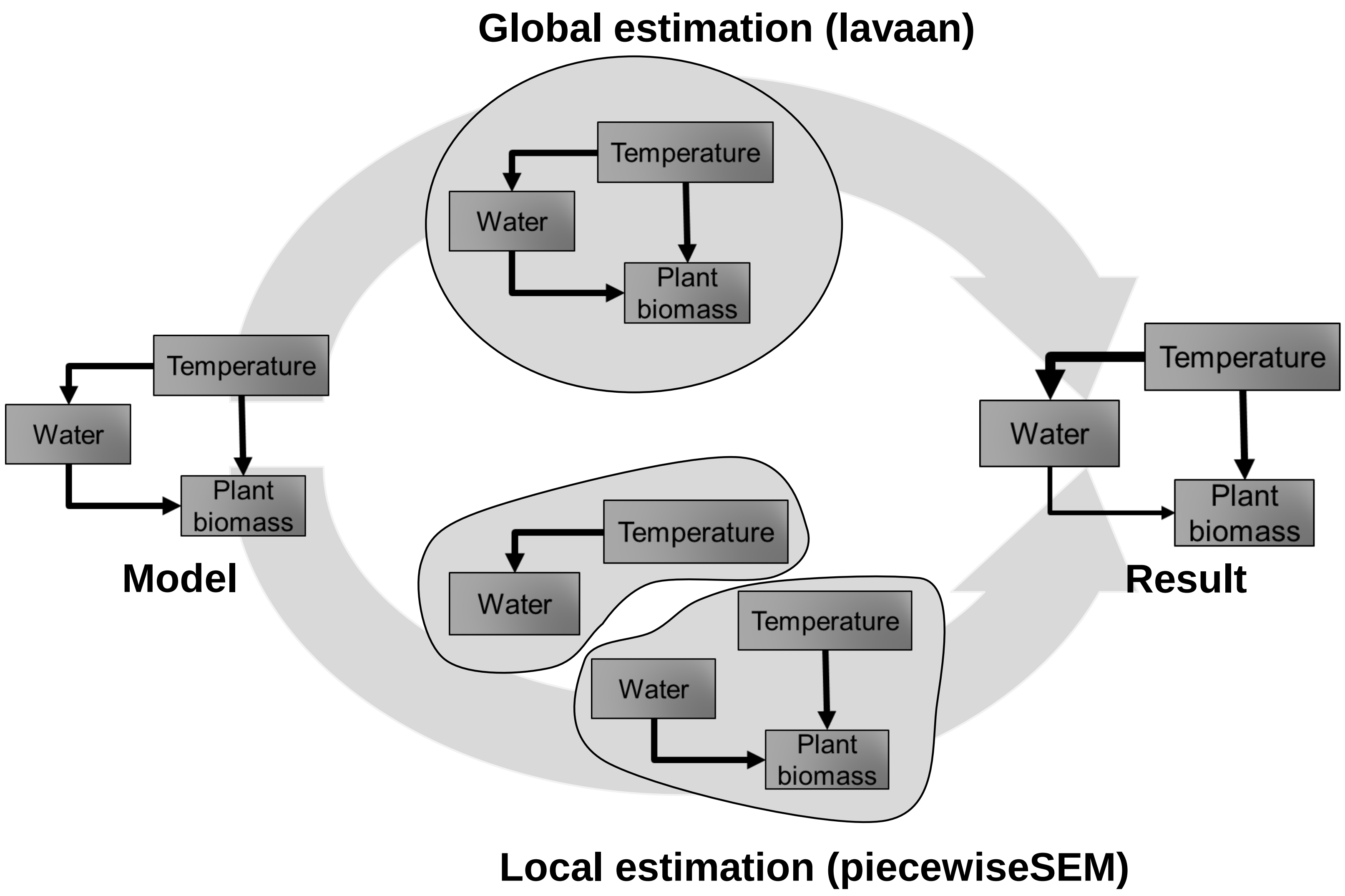
Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



Protocol to build and fit SEMs in R

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results

Global estimation
lavaan

Pros
- Old and stable method

Cons
- Only accept linear models without interactions

Local estimation
piecewiseSEM

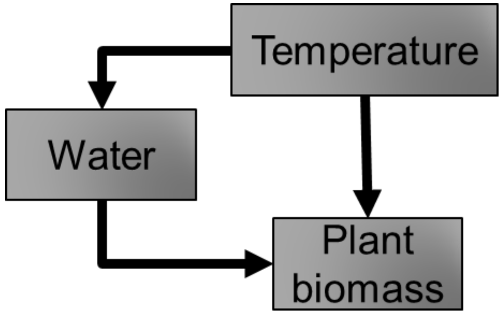
Pros
- Accept all kind of models
- Flexible
- Model can be fitted on different datasets

Cons
- Sensitive to overfit

For most models both method give the same results

Protocol to build and fit SEMs in R

Model



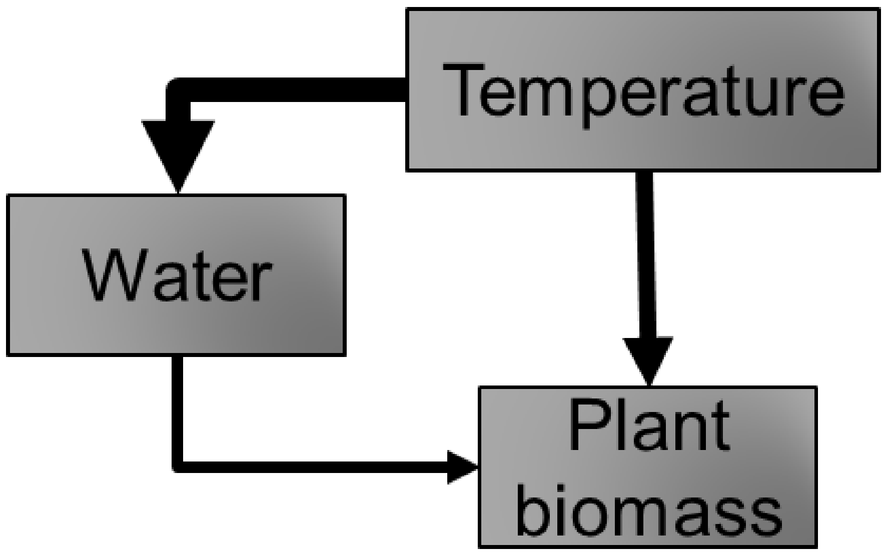
Global estimation
lavaan

```
library(lavaan)

model = ‘
plant.biomass ~ water + temperature
water ~ temperature
’

fit = sem(model, data = data)
```

Local estimation
piecewiseSEM

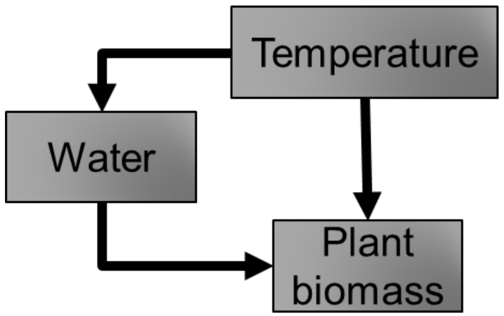
```
library(piecewiseSEM)

fit = psem(
lm(data = data, formula = ‘plant.biomass ~ water + temperature’),
lm(data = data, formula = ‘water ~ temperature’))
```

Read your SEM output



Read results

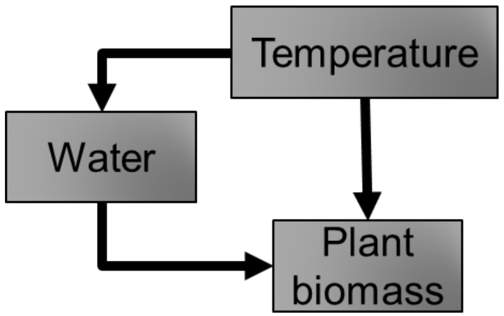
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



Did my SEM fit well the data?

Read results

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



Did my SEM fit well the data?

Forget about model *p*-value here

Useful and complementary indices:

CFI > 0.9
SRMR < 0.1 - 0.08
RMSEA < 0.08

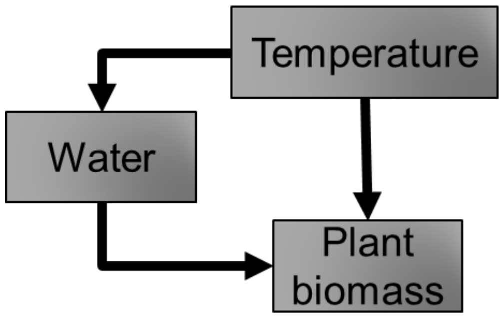
•*Comparative fit index* (CFI): this statistic considers the deviation from a ‘null’ model.

•*Standardized root-mean squared residual* (SRMR): the standardized difference between the observed and predicted correlations.

•*Root-mean squared error of approximation* (RMSEA): this statistic penalizes models based on sample size.

Read results

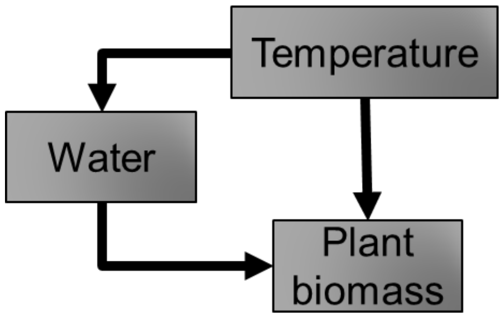
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



What can we learn from this SEM?

Read results

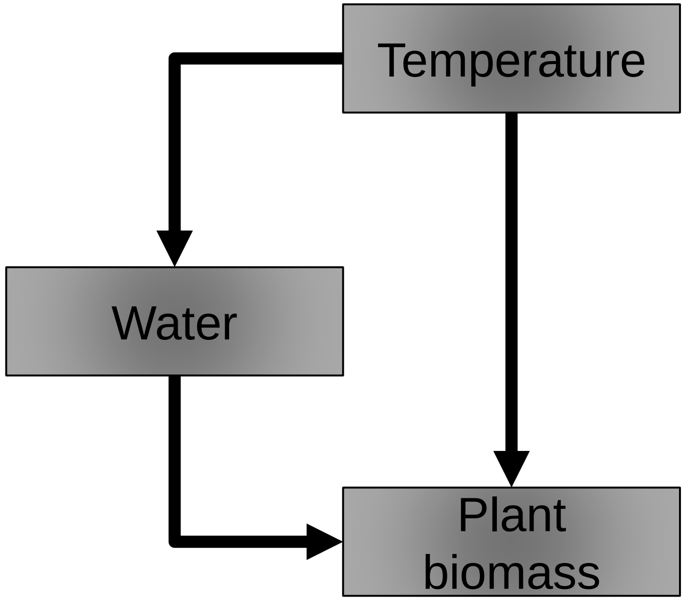
- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



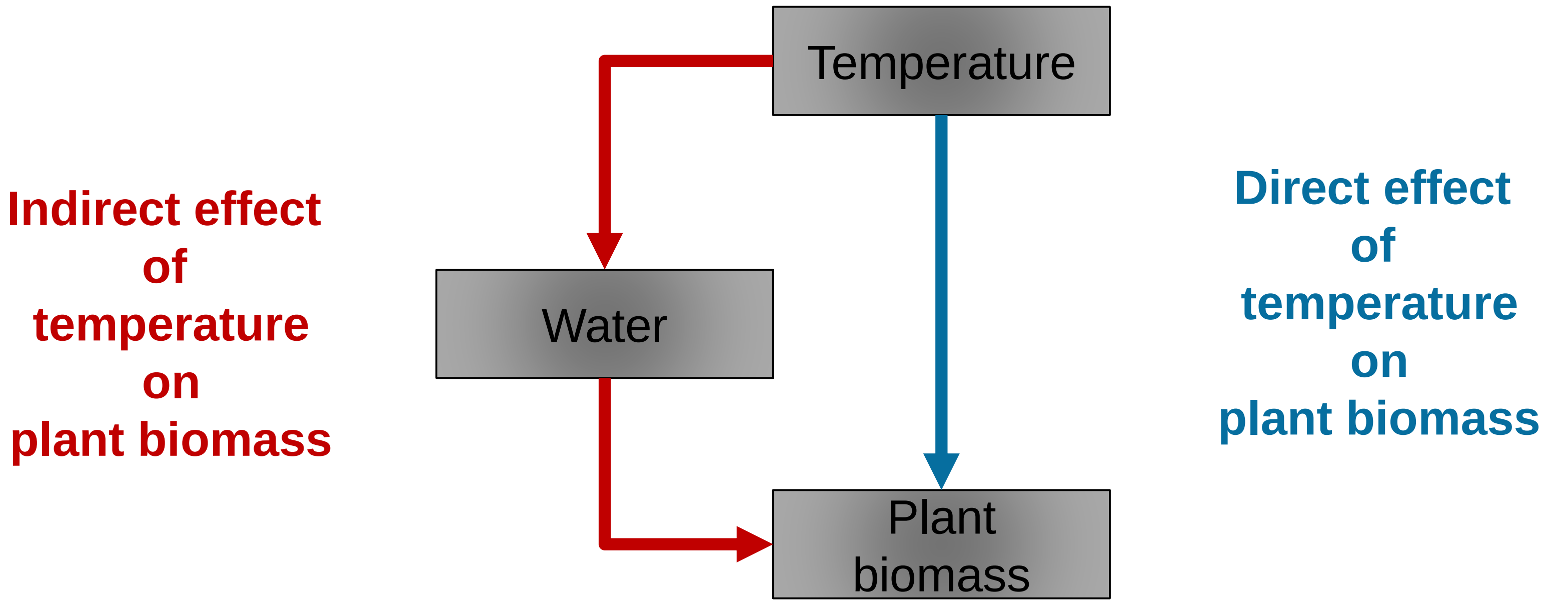
What can we learn from this SEM?

- The effect of temperature on plant biomass
- The effect of soil water content on plant biomass
- The effect of temperature on soil water content
- The mediation of this effect by soil water content

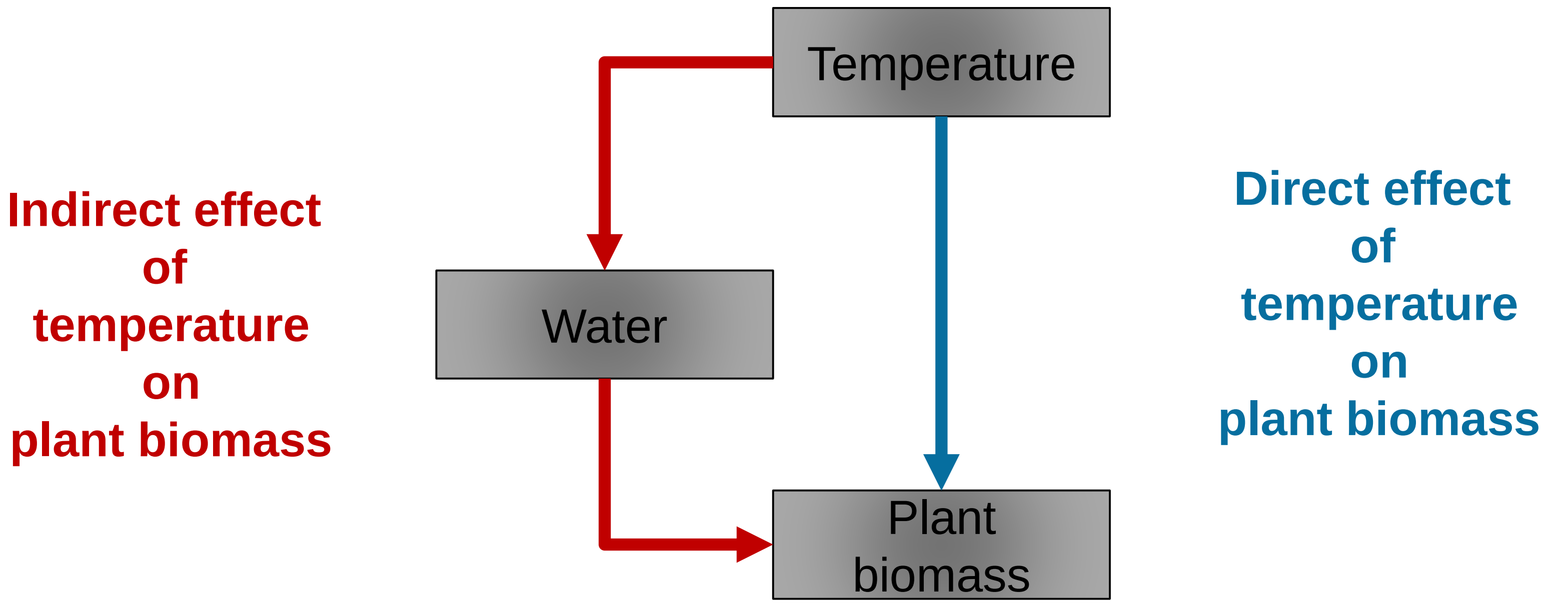
Direct *vs.* indirect effects (pathway analysis)



Direct *vs.* indirect effects (pathway analysis)

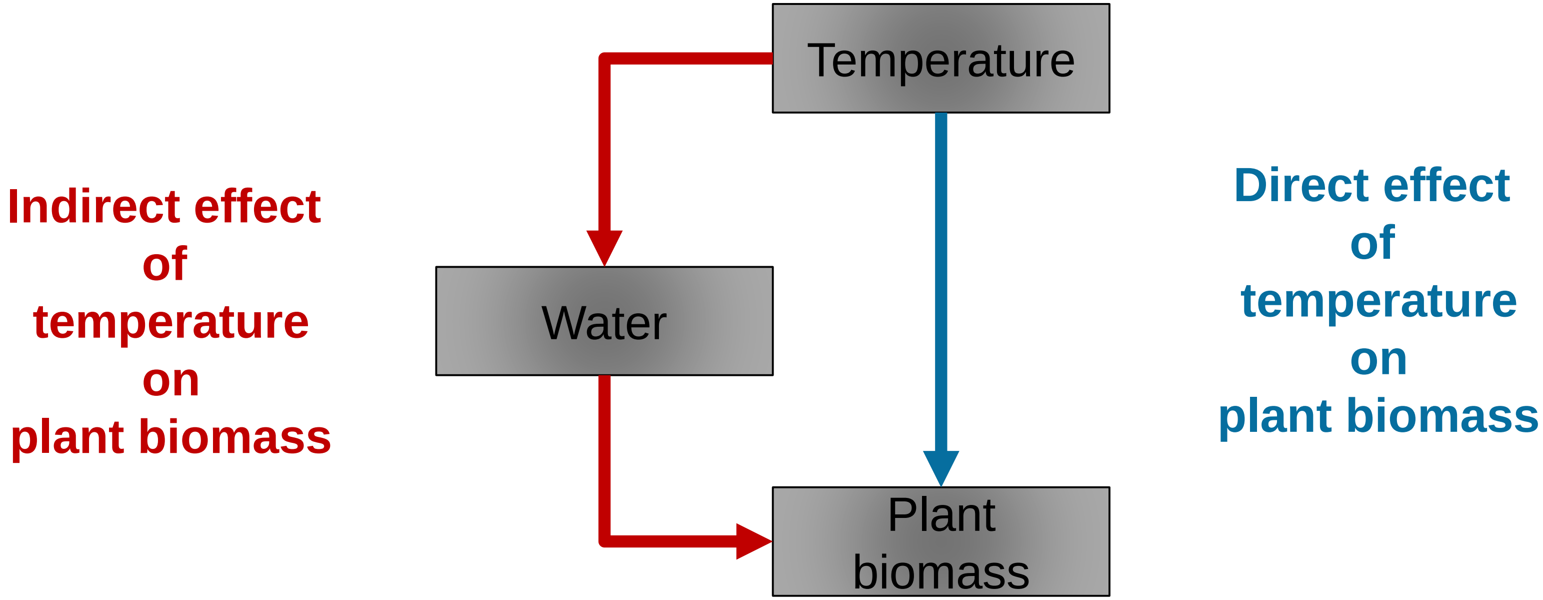


Direct *vs.* indirect effects (pathway analysis)



$$plant\ biomass \sim water + temperature$$
$$water \sim temperature$$

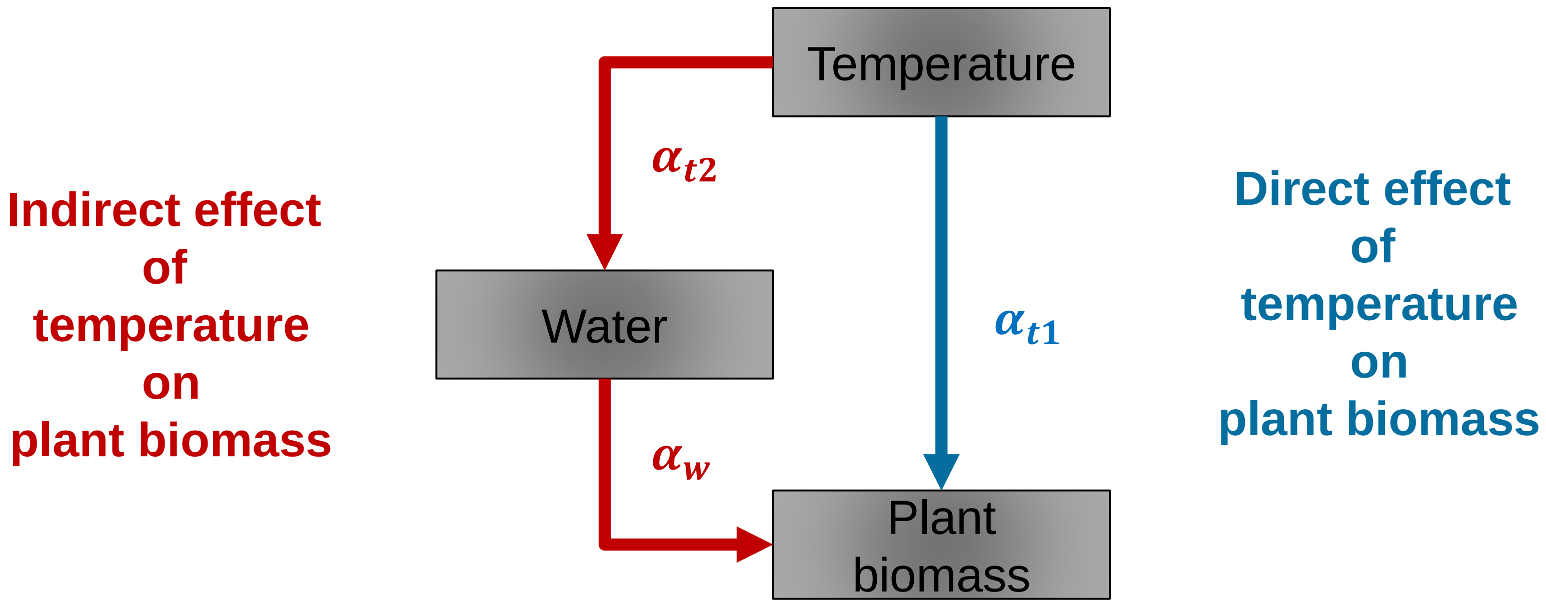
Direct *vs.* indirect effects (pathway analysis)



$$plant\ biomass \sim water + temperature$$
$$water \sim temperature$$

$$plant\ biomass = \mu_p + \alpha_w \times water + \alpha_{t1} \times temperature + \varepsilon_p$$
$$water = \mu_w + \alpha_{t2} \times temperature + \varepsilon_w$$

Direct *vs.* indirect effects (pathway analysis)

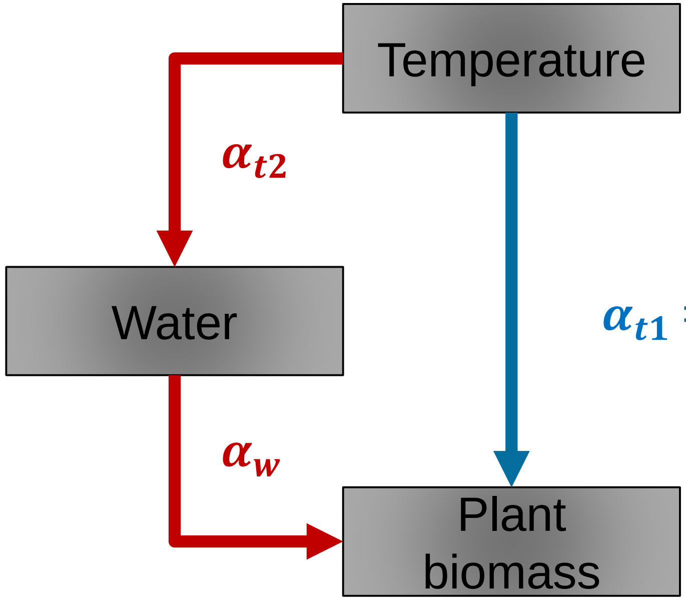


$$plant\ biomass \sim water + temperature$$
$$water \sim temperature$$

$$plant\ biomass = \mu_p + \alpha_w \times water + \alpha_{t1} \times temperature + \varepsilon_p$$
$$water = \mu_w + \alpha_{t2} \times temperature + \varepsilon_w$$

Direct *vs.* indirect effects (pathway analysis)

Indirect effect of temperature on plant biomass $= \alpha_w \times \alpha_{t2}$



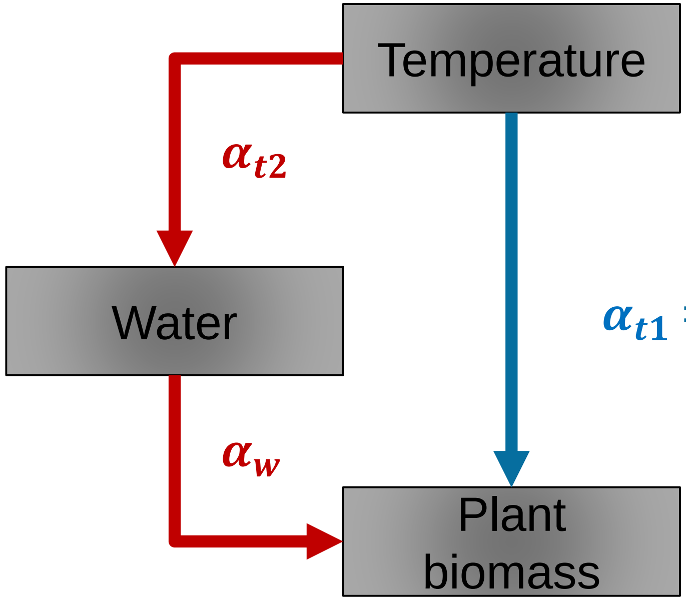
$\alpha_{t1} =$ **Direct effect of temperature on plant biomass**

$$plant\ biomass \sim water + temperature$$
$$water \sim temperature$$

$$plant\ biomass = \mu_p + \alpha_w \times water + \alpha_{t1} \times temperature + \varepsilon_p$$
$$water = \mu_w + \alpha_{t2} \times temperature + \varepsilon_w$$

Direct *vs.* indirect effects (pathway analysis)

Indirect effect of temperature on plant biomass $= \alpha_w \times \alpha_{t2}$



α_{t1} = **Direct effect of temperature on plant biomass**

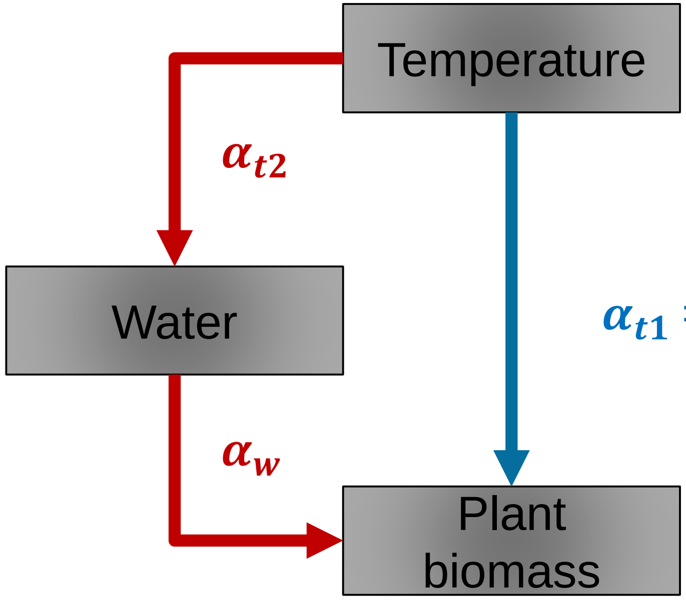
$$\boldsymbol{plant\ biomass \sim water + temperature}$$
$$\boldsymbol{water \sim temperature}$$

$$\boldsymbol{plant\ biomass = \mu_p + \alpha_w \times water + \alpha_{t1} \times temperature + \varepsilon_p}$$
$$\boldsymbol{water = \mu_w + \alpha_{t2} \times temperature + \varepsilon_w}$$

Total effect = direct + indirect effect

Direct *vs.* indirect effects (pathway analysis)

Indirect effect of temperature on plant biomass $= \alpha_w \times \alpha_{t2}$



α_{t1} = **Direct effect of temperature on plant biomass**

$$plant\ biomass \sim water + temperature$$
$$water \sim temperature$$

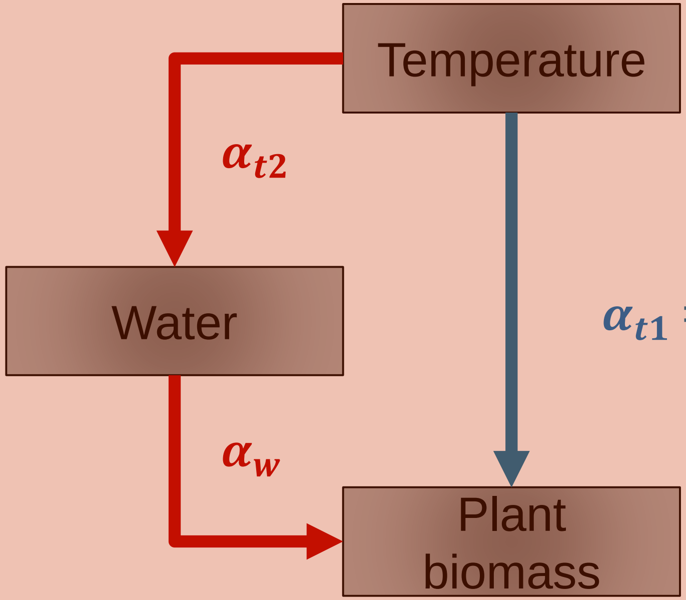
$$plant\ biomass = \mu_p + \alpha_w \times water + \alpha_{t1} \times temperature + \varepsilon_p$$
$$water = \mu_w + \alpha_{t2} \times temperature + \varepsilon_w$$

Total effect = direct + indirect effect

$$total\ effect = \alpha_{t1} + \alpha_w \times \alpha_{t2}$$

DANGER ZONE

Indirect effect of temperature on plant biomass $= \alpha_w \times \alpha_{t2}$



$\alpha_{t1} =$ **Direct effect of temperature on plant biomass**

$$plant\ biomass \sim water + temperature$$
$$water \sim temperature$$

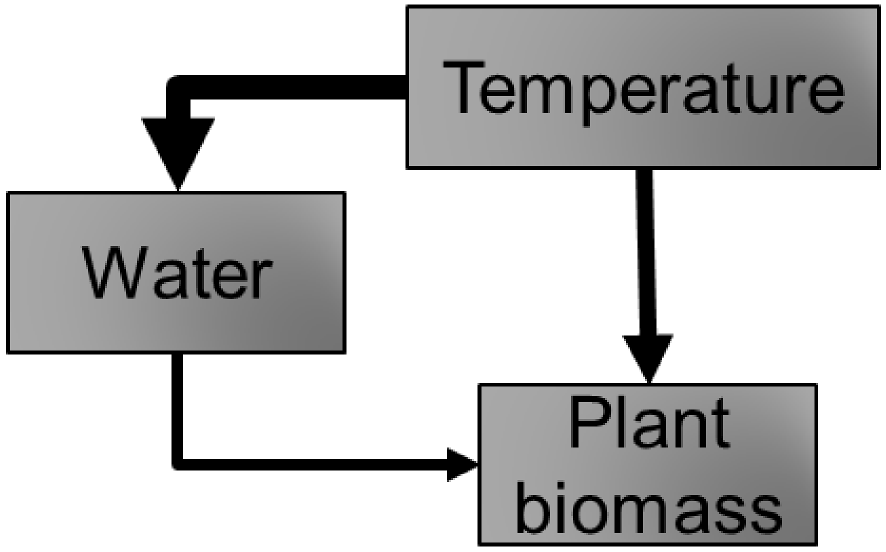
$$plant\ biomass = \mu_p + \alpha_w \times water + \alpha_{t1} \times temperature + \varepsilon_p$$
$$water = \mu_w + \alpha_{t2} \times temperature$$

Total effect = direct + indirect effect

$$total\ effect = \alpha_{t1} + \alpha_w \times \alpha_{t2}$$

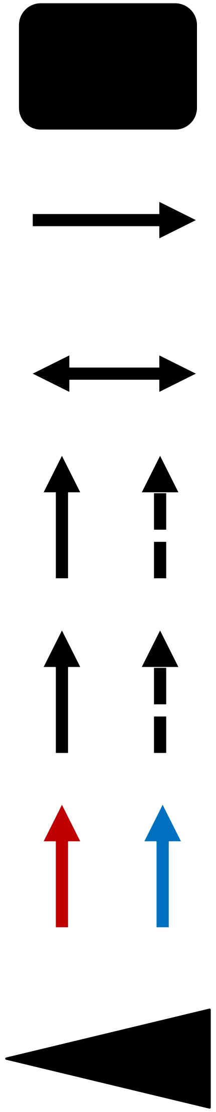
VARIABLES NEED TO BE RESCALED BEFOREHAND

Read SEMs in articles



Rules of etiquette

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



Squared boxes are variables

Single headed arrows are causal relationships

Double headed arrows are correlations

Significant *vs.* non-significant effects

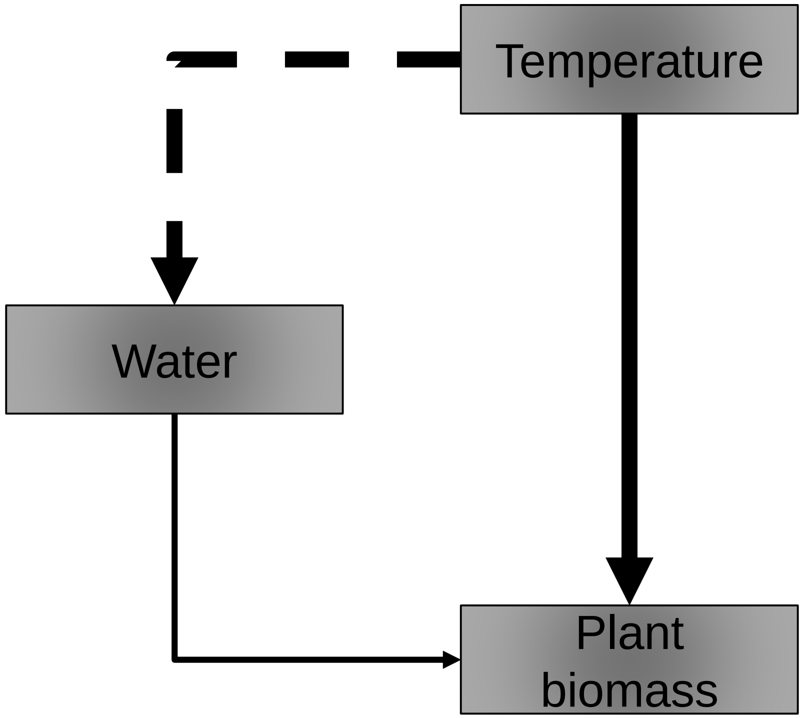
Positive vs. negative effects

Positive vs. negative effects

Arrow size is proportional to the effect size

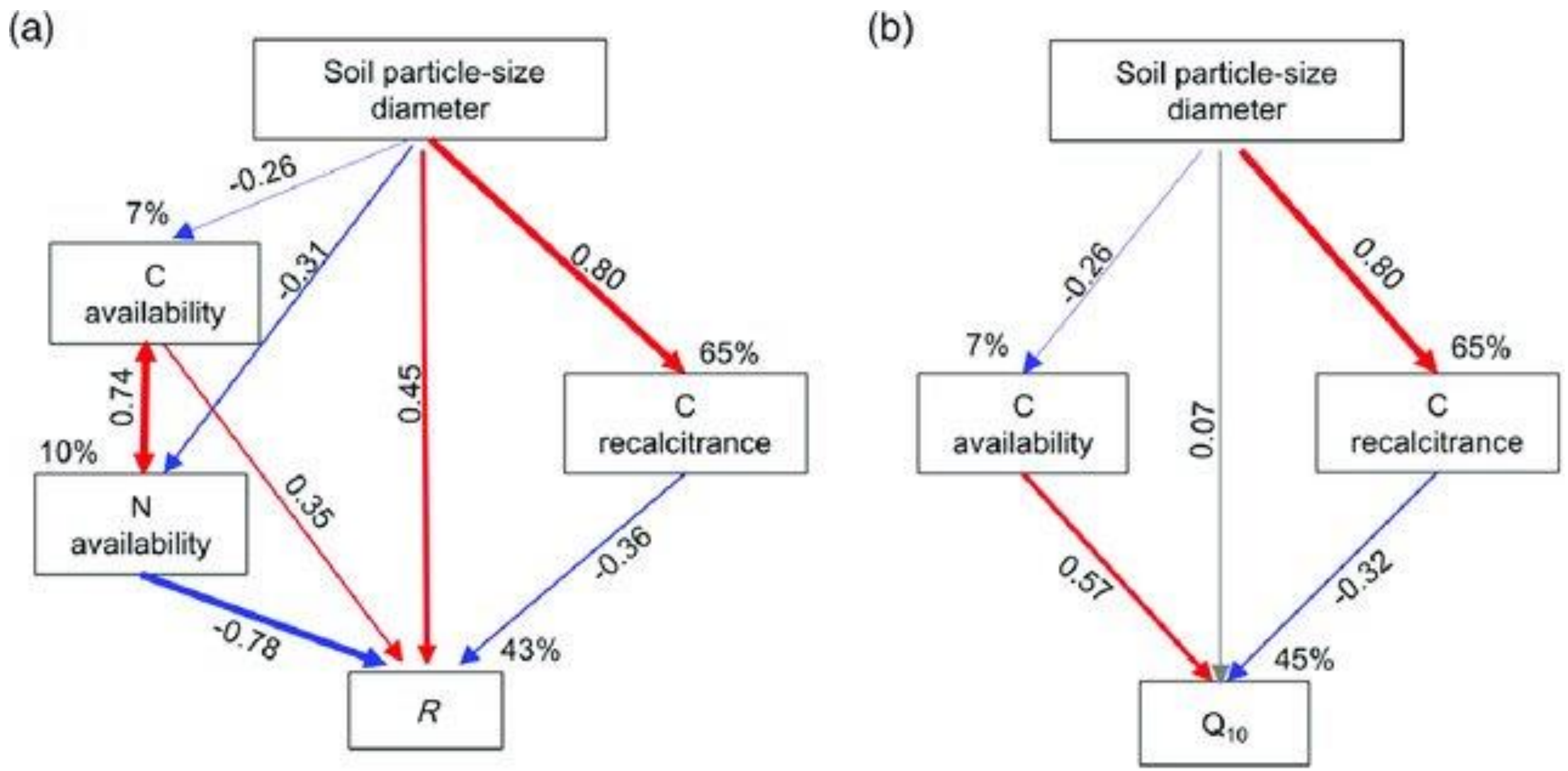
Rules of etiquette

- Write down the hypotheses (+ REF)
- Identify the model structure
- Write the equations
- Fit the SEM in R
- Read the results
- Show the results



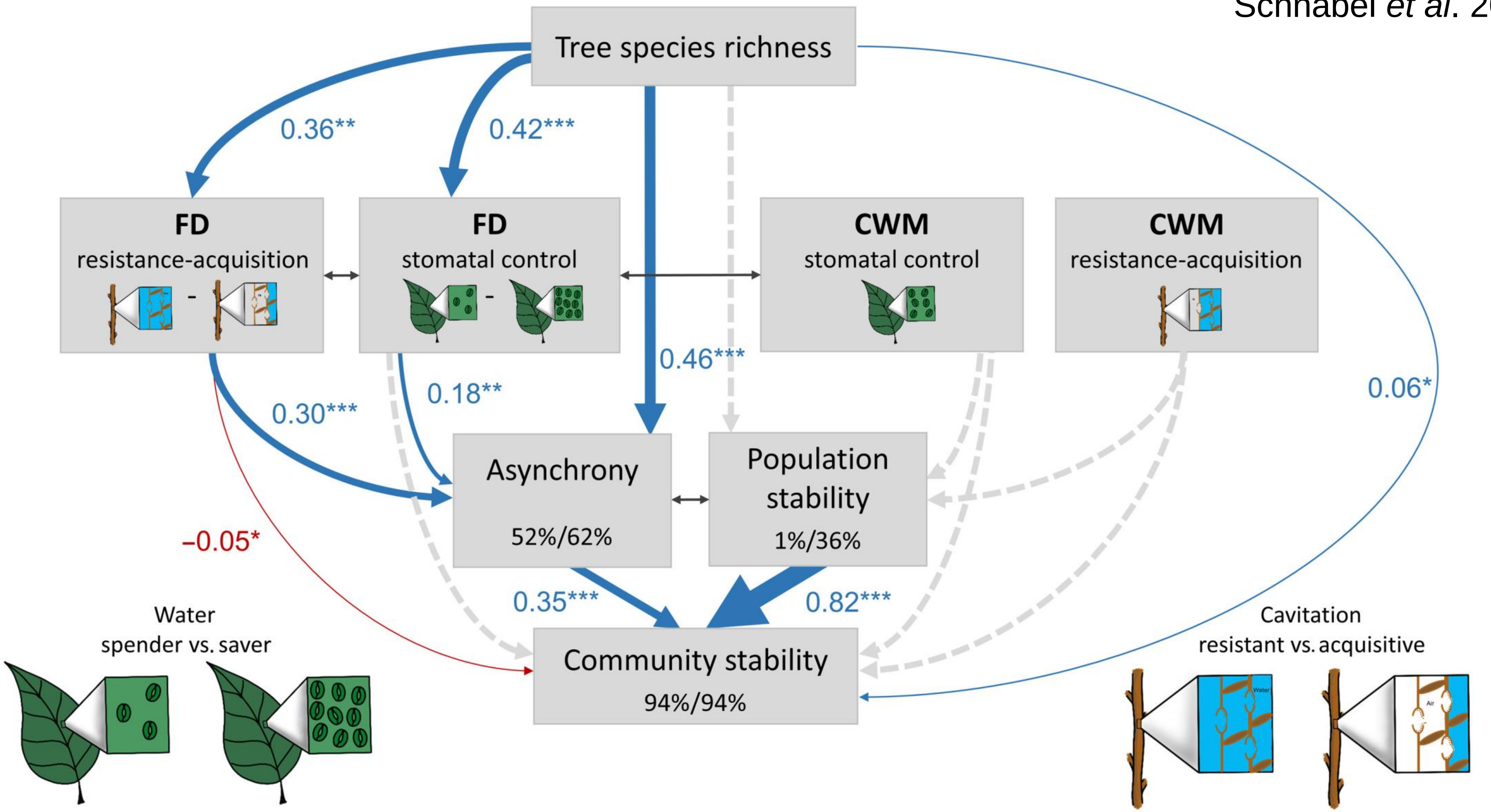
Real articles

Ding *et al*. 2017



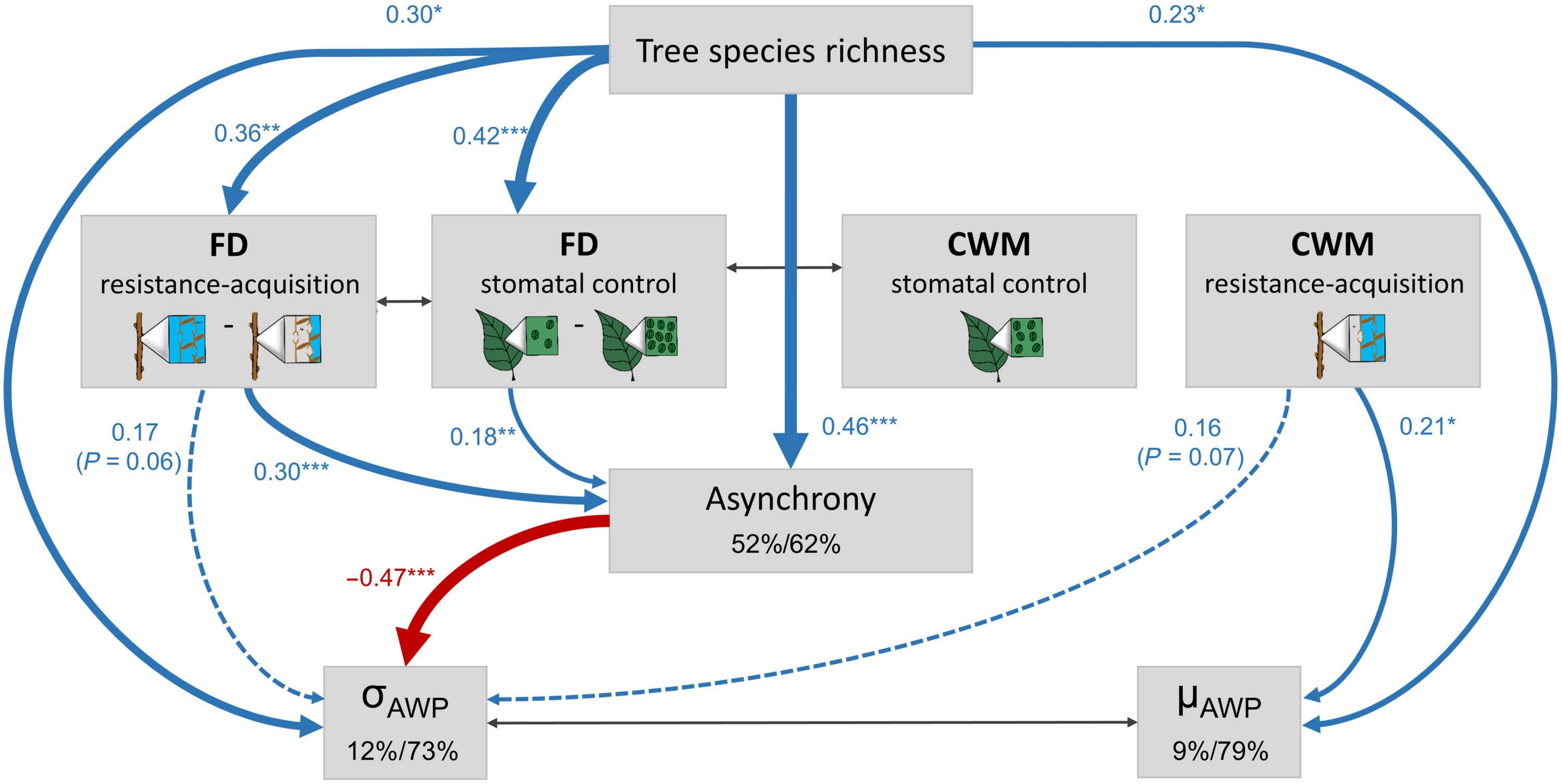
Real articles

Schnabel *et al*. 2021



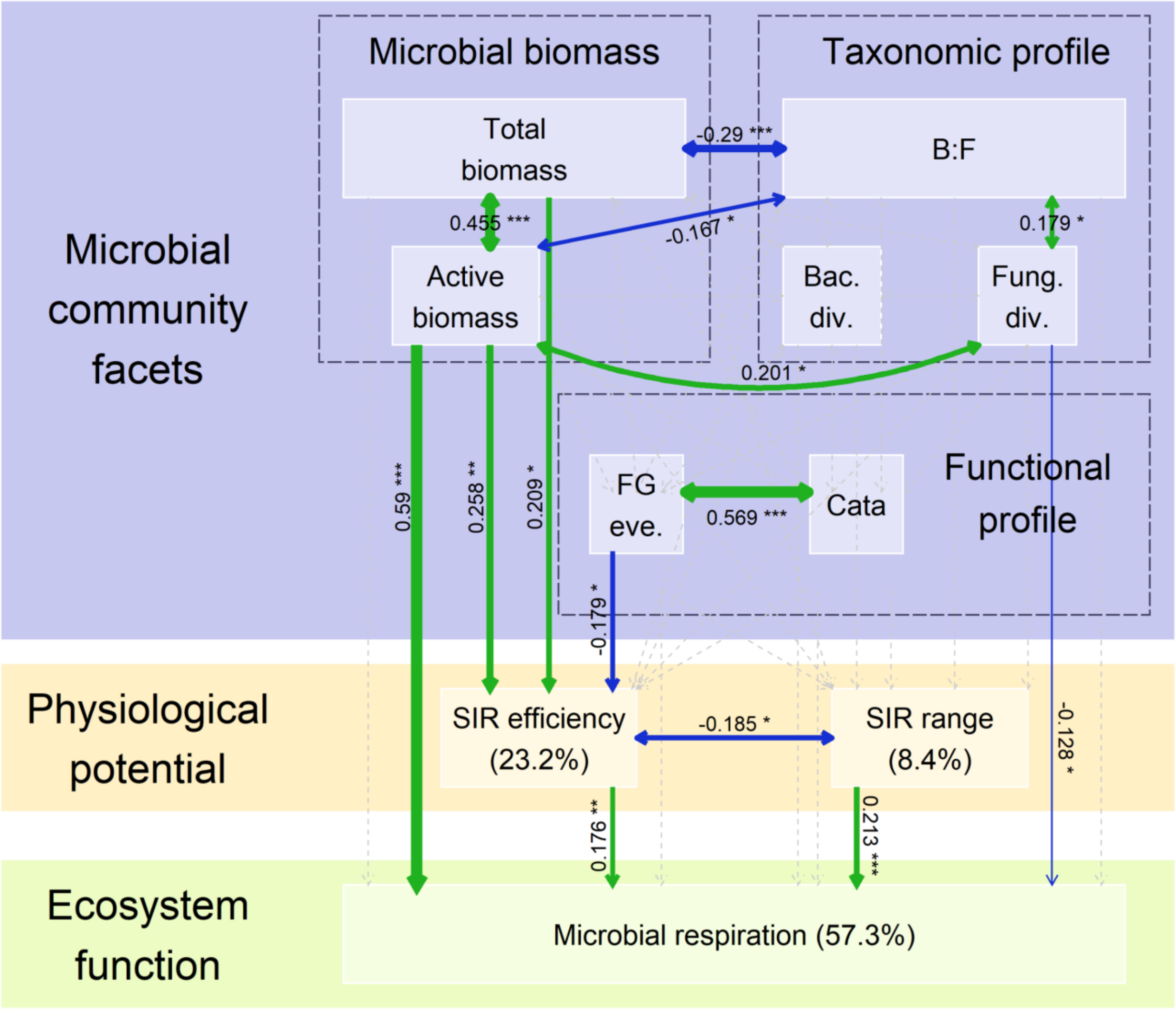
Real articles

Schnabel *et al*. 2021



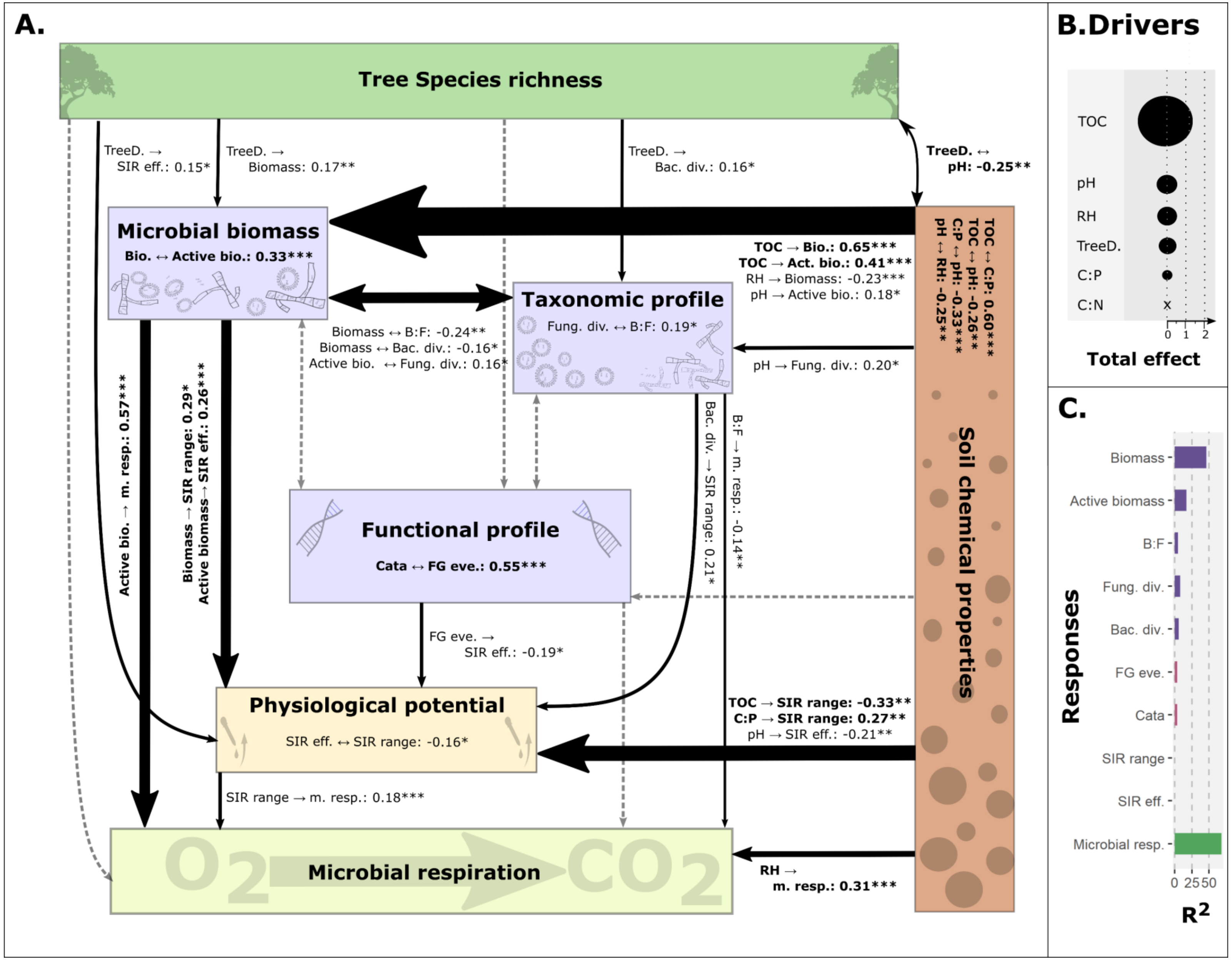
Real articles

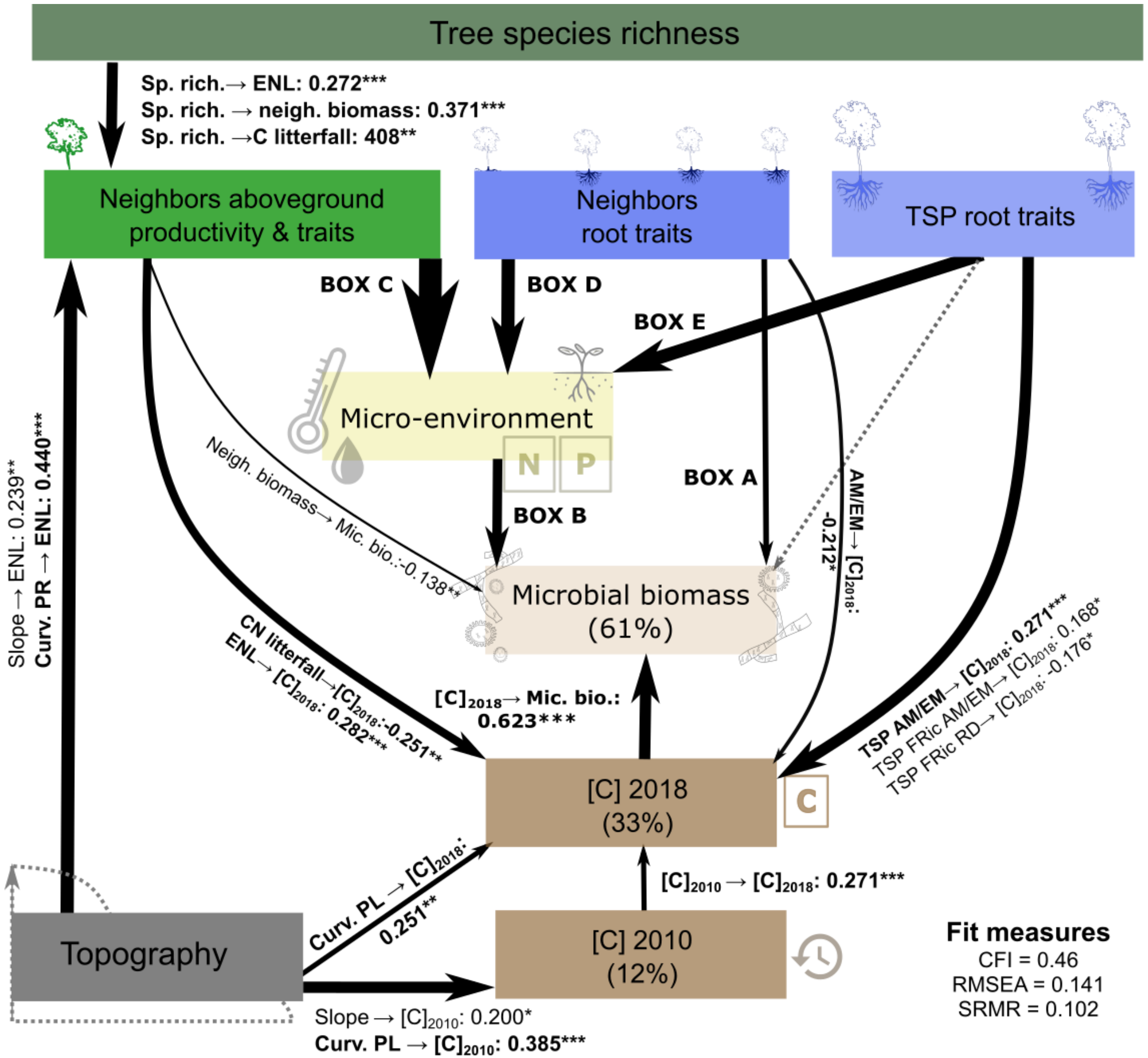
Beugnon, Du *et al.* 2021



Real articles

Beugnon, Du *et al.* 2021





BOX A
AM/EM→ Mic. bio.: -0.211***
RD → Mic. bio.: -0.124*

BOX B
Temperature → Mic. bio.: -0.270***
Litter CN → Mic. bio.: 0.242***

BOX C:
ENL → Temperature: -0.446***
ENL → Litter CN: -0.324***
CN litterfall → Litter CN : 0.239**
CN litterfall → Soil N: -.197*

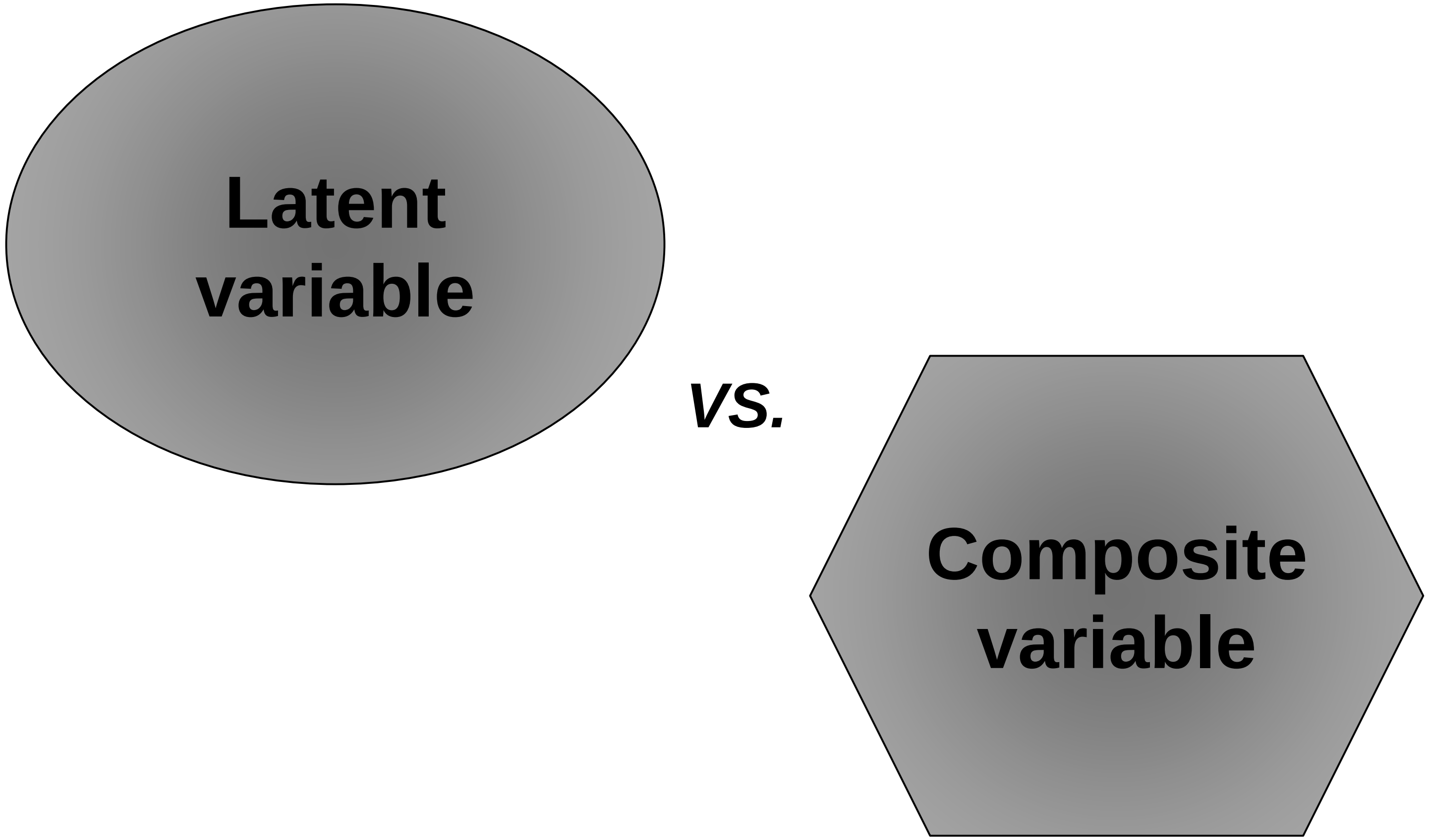
BOX D:
SRL → RH: -0.218**
FDis AM/EM → Litter CN: 0.173*
AM/EM → Litter CN: 0.315***

BOX E:
TSP RD → RH: -0.218**
TSP FRic RD → RH: 0.198*
TSP FRic AM/EM → RH: 0.173*
TSP AM/EM → Soil N: 0.246**

Useful links

- Introduction to Stat in R: https://remybeugnon.netlify.app/post/intro-to-stat-in-r/
- Introduction to SEM in R: https://remybeugnon.netlify.app/post/intro-to-sem-in-r/
- SEM book: https://jslefche.github.io/sem_book/
- Lavaan tutorials: https://lavaan.ugent.be/tutorial/sem.html

Add Ons



Thank you for your attention

Please fill the evaluation form for the future students