Introduction to structural equation modelling

Basic modelling

Frank Pennekamp

08.11.2021

Department of Evolutionary Biology and Environmental Sciences

University of Zurich

Quick introduction of participants

- Who are you?

Quick introduction of participants

- Who are you?
- Why do you want to learn about SEM?

Quick introduction of participants

- Who are you?
- Why do you want to learn about SEM?
- What is your research question for day 3?

General information

Introduction of the Swiss SEM team

- Dr. Frank Pennekamp (main instructor)

Introduction of the Swiss SEM team

- Dr. Frank Pennekamp (main instructor)
- Dr. James Grace (advanced topics and model clinic)

Introduction of the Swiss SEM team

- Dr. Frank Pennekamp (main instructor)
- Dr. James Grace (advanced topics and model clinic)
- Dr. Rachel Korn (course development)

Introduction of the Swiss SEM team

- Dr. Frank Pennekamp (main instructor)
- Dr. James Grace (advanced topics and model clinic)
- Dr. Rachel Korn (course development)
- Dr. Noémie Pichon, Dr. Fletcher Halliday, Dr. Eliane Meier, Dr. Hugo Saiz, Dr. Debra Zuppinger-Dingley, Rebecca Oester, Annabelle Constance, Fabienne Wiederkehr (course development)

Schedule & content

- *Day 1*:

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems
    - Fitting SEMs to data (live demo)

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems
    - Fitting SEMs to data (live demo)
    - Model pruning, visualization and reporting

Schedule & content

- *Day 1*:
  - General introduction to SEM to model ecological systems
  - Fitting SEMs to data (live demo)
  - Model pruning, visualization and reporting
  - Discussion with James Grace

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems
    - Fitting SEMs to data (live demo)
    - Model pruning, visualization and reporting
    - Discussion with James Grace
- *Day 2*:

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems
    - Fitting SEMs to data (live demo)
    - Model pruning, visualization and reporting
    - Discussion with James Grace
- *Day 2*:
    - Latent and composite variables

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems
    - Fitting SEMs to data (live demo)
    - Model pruning, visualization and reporting
    - Discussion with James Grace
- *Day 2*:
    - Latent and composite variables
    - Interactions

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems
    - Fitting SEMs to data (live demo)
    - Model pruning, visualization and reporting
    - Discussion with James Grace
- *Day 2*:
    - Latent and composite variables
    - Interactions
    - Complex sampling designs

Schedule & content

- *Day 1*:
  - General introduction to SEM to model ecological systems
  - Fitting SEMs to data (live demo)
  - Model pruning, visualization and reporting
  - Discussion with James Grace
- *Day 2*:
  - Latent and composite variables
  - Interactions
  - Complex sampling designs
  - Discussion with James Grace

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems
    - Fitting SEMs to data (live demo)
    - Model pruning, visualization and reporting
    - Discussion with James Grace
- *Day 2*:
    - Latent and composite variables
    - Interactions
    - Complex sampling designs
    - Discussion with James Grace
- *Day 3*:

Schedule & content

- *Day 1*:
    - General introduction to SEM to model ecological systems
    - Fitting SEMs to data (live demo)
    - Model pruning, visualization and reporting
    - Discussion with James Grace
- *Day 2*:
    - Latent and composite variables
    - Interactions
    - Complex sampling designs
    - Discussion with James Grace
- *Day 3*:
    - Self-study with possibility to meet with instructor(s)

- What the course is about:

- What the course is about:
    - Global estimation with R package lavaan

Overview

- What the course is about:
    - Global estimation with R package lavaan
    - Hands on exercises and live coding

- What the course is about:
  - Global estimation with R package lavaan
  - Hands on exercises and live coding
  - We will work with a single, ecological dataset (Seabloom et al. 2020)

Overview

- What the course is about:
  - Global estimation with R package lavaan
  - Hands on exercises and live coding
  - We will work with a single, ecological dataset (Seabloom et al. 2020)
- What will not be covered

Overview

- What the course is about:
    - Global estimation with R package lavaan
    - Hands on exercises and live coding
    - We will work with a single, ecological dataset (Seabloom et al. 2020)
- What will not be covered
    - Local estimation of SEMs (with piecewiseSEM)

Overview

- What the course is about:
  - Global estimation with R package lavaan
  - Hands on exercises and live coding
  - We will work with a single, ecological dataset (Seabloom et al. 2020)
- What will not be covered
  - Local estimation of SEMs (with piecewiseSEM)
  - Advanced topics like incorporating random effects, feedbacks, temporal autocorrelation

Learning objectives

- Participants understand the advantages and limits of SEMs to draw inferences from data

Learning objectives

- Participants understand the advantages and limits of SEMs to draw inferences from data
- Participants are able to fit, interpret and visualize a SEM with `lavaan`.

Learning objectives

- Participants understand the advantages and limits of SEMs to draw inferences from data
- Participants are able to fit, interpret and visualize a SEM with `lavaan`.
- Participants are able to apply SEM to their own dataset

Getting started with Structural Equation Modeling

Research questions

- Ecology is about the interactions between organisms and their environment.

Research questions

- Ecology is about the interactions between organisms and their environment.
- We often have ideas how things could be connected in ecological systems.

- Ecology is about the interactions between organisms and their environment.
- We often have ideas how things could be connected in ecological systems.
- To test hypotheses, we need a way to dissect when they occur for a reason versus randomness.

Research questions

- Ecology is about the interactions between organisms and their environment.
- We often have ideas how things could be connected in ecological systems.
- To test hypotheses, we need a way to dissect when they occur for a reason versus randomness.
- We use statistics to understand what is signal and what is noise.

Research questions

- Ecology is about the interactions between organisms and their environment.
- We often have ideas how things could be connected in ecological systems.
- To test hypotheses, we need a way to dissect when they occur for a reason versus randomness.
- We use statistics to understand what is signal and what is noise.
- Research questions are often about cause and effect.

- Promotes system thinking:

- Promotes system thinking:
    - “Understanding the whole rather than the parts in isolation”

Why do we need to use SEM in Ecology

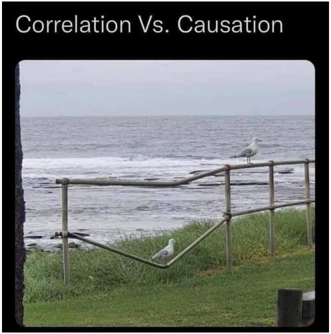
- Promotes system thinking:
    - "Understanding the whole rather than the parts in isolation"
- SEM unites multiple variables in a single causal network: simultaneous tests of multiple hypotheses.

- Promotes system thinking:
  - “Understanding the whole rather than the parts in isolation”
- SEM unites multiple variables in a single causal network: simultaneous tests of multiple hypotheses.
- Causality is central:

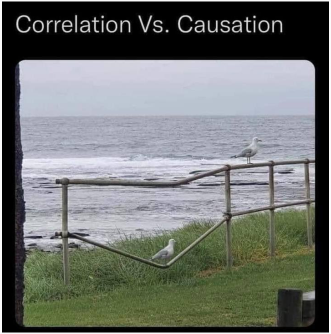
Why do we need to use SEM in Ecology

- Promotes system thinking:
    - “Understanding the whole rather than the parts in isolation”
- SEM unites multiple variables in a single causal network: simultaneous tests of multiple hypotheses.
- Causality is central:
    - SEM assumes that the specified relationships among variables represent causal links.

Causality

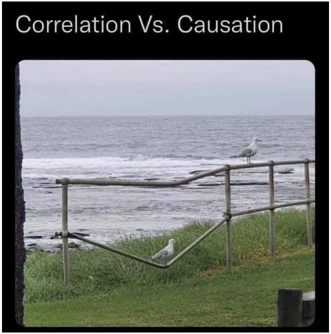


- "Correlation does not imply causation".

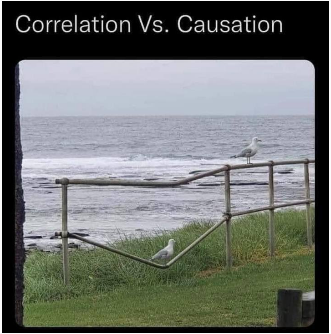


- “Correlation does not imply causation”.
- Everything else being equal, seeing variation in X leading to variation in Y.

Causality



- "Correlation does not imply causation".
- Everything else being equal, seeing variation in X leading to variation in Y.
- Experiments to isolate effect of X on Y.



- "Correlation does not imply causation".
- Everything else being equal, seeing variation in X leading to variation in Y.
- Experiments to isolate effect of X on Y.
- Experiments not always feasible, hence development of SEM.

Differences and similarities between SEM and regression models

Regression models vs SEM

- We often have multiple observed variables.

Regression models vs SEM

- We often have multiple observed variables.
- Include terms to test for interactions.

Regression models vs SEM

- We often have multiple observed variables.
- Include terms to test for interactions.
- We want to test and evaluate multivariate causal relationship.

Regression models vs SEM

- We often have multiple observed variables.
- Include terms to test for interactions.
- We want to test and evaluate multivariate causal relationship.
- Test direct and indirect effects on assumed causal relationships.

Regression models vs SEM

- We often have multiple observed variables.
- Include terms to test for interactions.
- We want to test and evaluate multivariate causal relationship.
- Test direct and indirect effects on assumed causal relationships.
- Incorporate observed and latent variables.

Two goals of SEM:

1) Understand the underlying causal network driving the correlation/covariance among a set of variables.

Two goals of SEM:

1) Understand the underlying causal network driving the correlation/covariance among a set of variables.
2) Explain as much of their variance as possible with the model specified.

Thinking about the model

A SEM is usually specified based on theory to determine and validify a proposed causal process and/or model.

Which variables to include?

- Supported by theory.

Thinking about the model

A SEM is usually specified based on theory to determine and validify a proposed causal process and/or model.

Which variables to include?

- Supported by theory.
- Ecologically meaningful.

A SEM is usually specified based on theory to determine and validify a proposed causal process and/or model.

Which variables to include?

- Supported by theory.
- Ecologically meaningful.
- Garbage in - garbage out (both data quality and ecologically meaningful).

SEM modelling philosophy

- Continuum between theory (hypothesis-driven) to exploratory (data-driven) modelling.

SEM modelling philosophy

- Continuum between theory (hypothesis-driven) to exploratory (data-driven) modelling.
- Important to be explicit about the approach taken.

Introduction to the dataset

Introduction to the dataset

We will use an experimental dataset collected at the Cedar Creek Ecosystem Science Reserve to examine long-term consequences of human-driven environmental changes ecosystem responses to:

- Disturbance

Introduction to the dataset

We will use an experimental dataset collected at the Cedar Creek Ecosystem Science Reserve to examine long-term consequences of human-driven environmental changes ecosystem responses to:

- Disturbance
- Nitrogen deposition

Introduction to the dataset

We will use an experimental dataset collected at the Cedar Creek Ecosystem Science Reserve to examine long-term consequences of human-driven environmental changes ecosystem responses to:

- Disturbance
- Nitrogen deposition
- Changes in precipitation

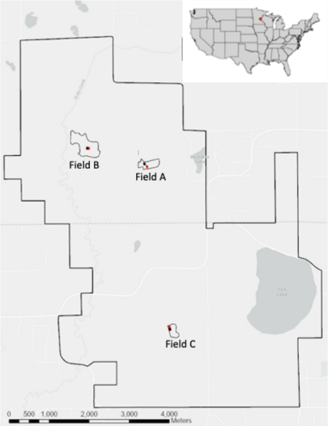
Research questions:

1) How has aboveground biomass changed as a function of disturbance (disking) and nutrient addition?

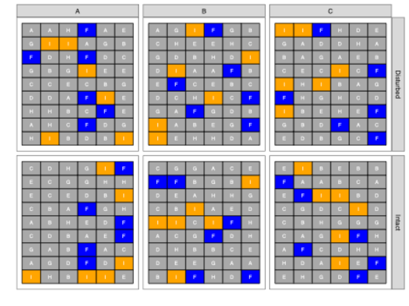
Research questions:

1) How has aboveground biomass changed as a function of disturbance (disking) and nutrient addition?
2) How are these effects mediated by diversity?

Introduction to the dataset

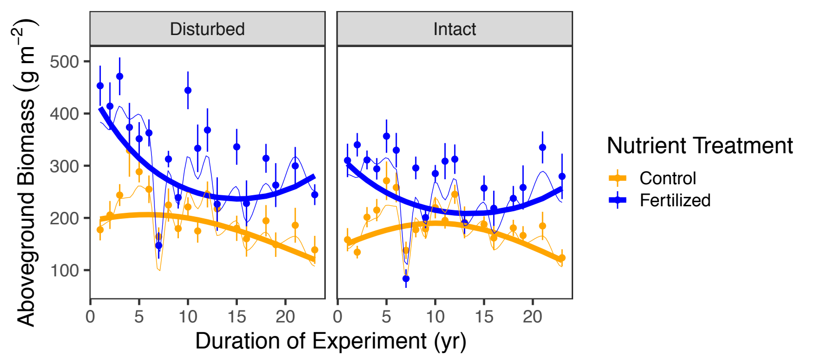


Location of the study site (Cedar Creek Ecosystem Science Reserve), the location of the three study fields within the reserve, and location of the 35 x 55 m intact (black) and disturbed (red) plots within each field.



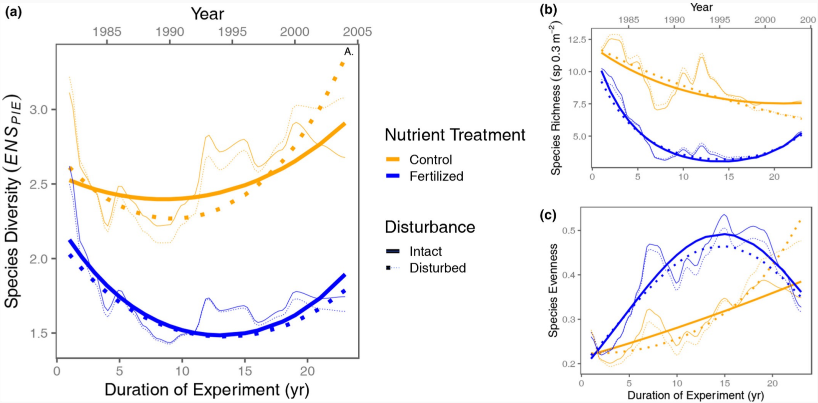
Location of the 4 x 4 m nutrient treatment plots within each 35 x 55 m Intact or Disturbed plot within each of three fields (A, B, and C). Letters indicate the nutrient treatments.

Introduction to the dataset



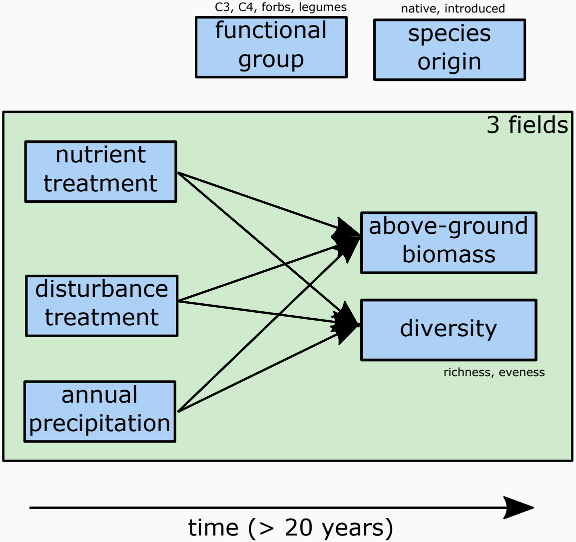
Effect of soil disturbance (disking) and nutrient enrichment on live, aboveground plant biomass. Colors indicate nutrient addition treatment: Control and NPK+ (all nutrients plus 9.5 g N m-2 yr-1).

Introduction to the dataset

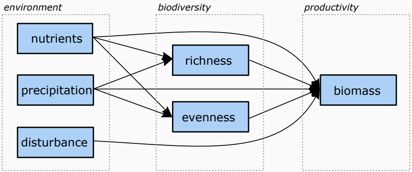


Effect of soil disturbance (disking) and nutrient enrichment on (a) diversity (ENSPIE), (b) richness (S, species 0.3 m−2), and (c) evenness (ENSPIE S−1).

Question of interest: what is the effect of richness on biomass?



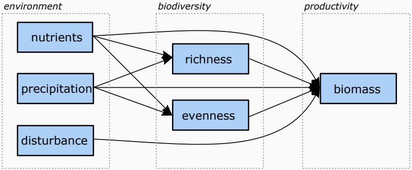
Meta-model



Meta-model are conceptual models that allow to link data with theory.

1) Productivity (biomass) is directly influenced by the environment (nutrients, disturbance and precipitation)

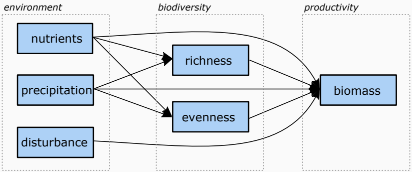
Meta-model



Meta-model are conceptual models that allow to link data with theory.

1) Productivity (biomass) is directly influenced by the environment (nutrients, disturbance and precipitation)
2) Productivity (biomass) is directly influenced by biodiversity (richness and evenness).

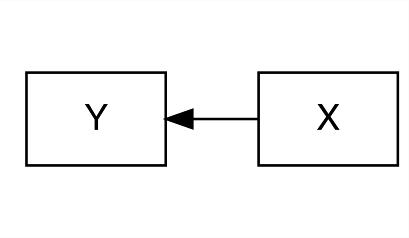
Meta-model



Meta-model are conceptual models that allow to link data with theory.

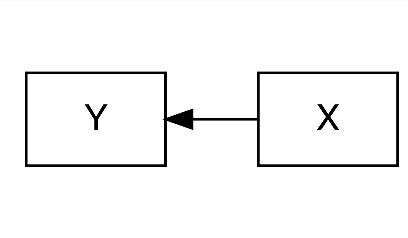
1) Productivity (biomass) is directly influenced by the environment (nutrients, disturbance and precipitation)
2) Productivity (biomass) is directly influenced by biodiversity (richness and evenness).
3) The environment also influences biodiversity and thus, have an indirect effect on productivity via biodiversity.

Graphical models:

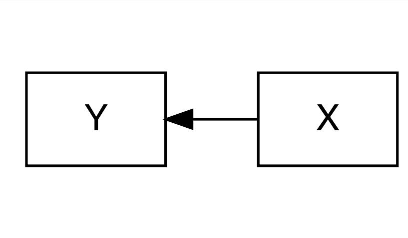


- Directed acyclic graphs (no loops).

Graphical models:

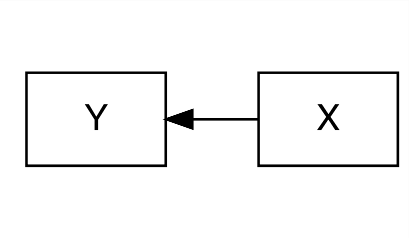


- Directed acyclic graphs (no loops).
- Variables are nodes (boxes).



- Directed acyclic graphs (no loops).
- Variables are nodes (boxes).
- Edges (one-headed arrows) are causal relationships such as X affects Y.

Graphical models:



- Directed acyclic graphs (no loops).
- Variables are nodes (boxes).
- Edges (one-headed arrows) are causal relationships such as X affects Y.
- Edges (two-headed arrows) are non-causal relationships such as X and Y are correlated.

Exercise:

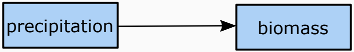
Exercise:

- Draw a meta-model of the dataset you will work with on Thursday.

Exercise:

- Draw a meta-model of the dataset you will work with on Thursday.
- Make a table with putative causal relationships.

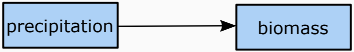
A simple bivariate model.




```
lm(biomass ~ precipitation)
```

- Linear regression

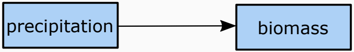
A simple bivariate model.




```
lm(biomass ~ precipitation)
```

- Linear regression
- Regression coefficient quantifies the strength of relationship

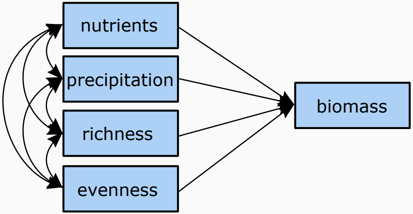
A simple bivariate model.




```
lm(biomass ~ precipitation)
```

- Linear regression
- Regression coefficient quantifies the strength of relationship
- Change in Y for one unit change in X

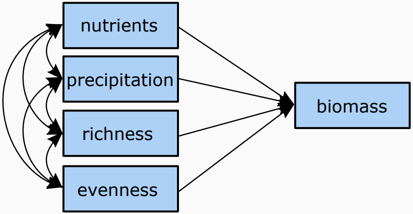
Multiple independent variables




```
lm(biomass ~ precipitation + nutrients + ...)
```

- Often more than one independent variable important → multiple regression

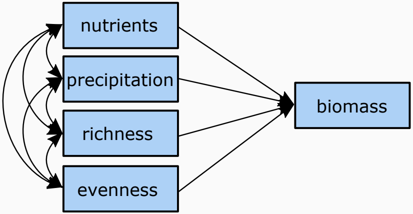
Multiple independent variables




```
lm(biomass ~ precipitation + nutrients + ...)
```

- Often more than one independent variable important $\longrightarrow$ multiple regression
- Estimates partial regression coefficients (i.e. effect of precipitation on biomass when nutrient addition is fixed)

Multiple independent variables




```
lm(biomass ~ precipitation + nutrients + ...)
```

- Often more than one independent variable important → multiple regression
- Estimates partial regression coefficients (i.e. effect of precipitation on biomass when nutrient addition is fixed)
- Only direct effects.

From regression models to SEM

SEM: variance-covariance matrix

```
##             [,1]      [,2]      [,3]
## [1,] 1875.3209  429.8712  462.4775
## [2,]  429.8712 1306.9817 -262.8231
## [3,]  462.4775 -262.8231  755.5193
```

- Comparison of specified SEM to observed variance-covariance matrix

SEM: variance-covariance matrix

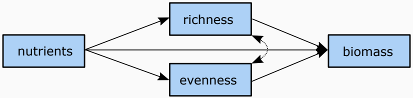
```
##              [,1]      [,2]      [,3]
## [1,] 1875.3209  429.8712  462.4775
## [2,]  429.8712 1306.9817 -262.8231
## [3,]  462.4775 -262.8231  755.5193
```

- Comparison of specified SEM to observed variance-covariance matrix
- The variances appear along the diagonal and covariances appear in the off-diagonal elements

```
##           [,1]       [,2]       [,3]
## [1,] 1.0000000  0.2745780  0.3885349
## [2,] 0.2745780  1.0000000 -0.2644881
## [3,] 0.3885349 -0.2644881  1.0000000
```

- Correlation matrix is standardized variance-covariance matrix

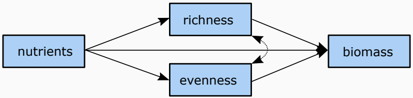
Multiple independent variables




```
lm(richness ~ nutrients)
lm(evenness ~ nutrients)
lm(biomass ~ richness + evenness)
```

- Indirect effect is the effect of an independent variable on a dependent variable through one or more intervening or mediating variables.

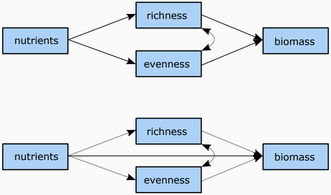
Multiple independent variables




```
lm(richness ~ nutrients)
lm(evenness ~ nutrients)
lm(biomass ~ richness + evenness)
```

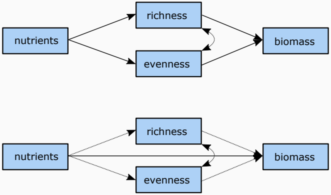
- Indirect effect is the effect of an independent variable on a dependent variable through one or more intervening or mediating variables.
- Indirect effects can be quantified by the product of the compound path.

Mediation



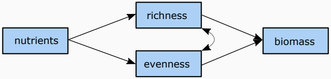
- Tests whether a particular variable has a mediating effect on a path.

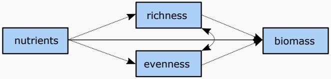
Mediation



- Tests whether a particular variable has a mediating effect on a path.
- Often used to test underlying mechanisms.

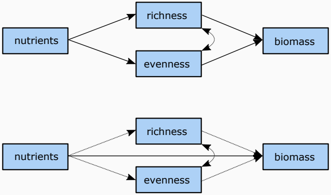
Mediation





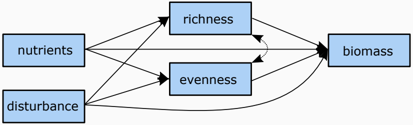
- Tests whether a particular variable has a mediating effect on a path.
- Often used to test underlying mechanisms.
- In our example, we could ask whether the effect of nutrients on biomass is mediated through biodiversity.

Mediation



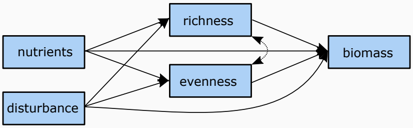
- Tests whether a particular variable has a mediating effect on a path.
- Often used to test underlying mechanisms.
- In our example, we could ask whether the effect of nutrients on biomass is mediated through biodiversity.
- Possibilities: complete mediation, partial mediation, no mediation.

System level approach

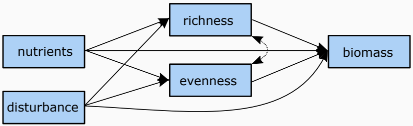


- Exogenous variables only have paths emanating from them (i.e., do not have arrows going into them).

System level approach



- Exogenous variables only have paths emanating from them (i.e., do not have arrows going into them).
- Endogenous variables have paths directed into them.



- Exogenous variables only have paths emanating from them (i.e., do not have arrows going into them).
- Endogenous variables have paths directed into them.
- An endogenous variable can also have arrows directing out of it, but the sole condition is that they must be predicted.

Grace's 8 rules of path coefficients

1) Unspecified relationships among exogenous variables are their bivariate correlations.

Grace's 8 rules of path coefficients

1) Unspecified relationships among exogenous variables are their bivariate correlations.
2) When two variables are connected by a single path, the coefficient of that path is the regression coefficient.

Grace's 8 rules of path coefficients

1) Unspecified relationships among exogenous variables are their bivariate correlations.
2) When two variables are connected by a single path, the coefficient of that path is the regression coefficient.
3) The strength of a compound path (one that includes multiple links) is the product of the individual coefficients.

Grace's 8 rules of path coefficients

1) Unspecified relationships among exogenous variables are their bivariate correlations.
2) When two variables are connected by a single path, the coefficient of that path is the regression coefficient.
3) The strength of a compound path (one that includes multiple links) is the product of the individual coefficients.
4) When variables are connected by more than one pathway, each pathway is the 'partial' regression coefficient.

Grace's 8 rules of path coefficients

5) Errors on endogenous variables relate the unexplained correlations or variances arising from unmeasured variables.

5) Errors on endogenous variables relate the unexplained correlations or variances arising from unmeasured variables.
6) Unanalyzed (residual) correlations among two endogenous variables are their partial correlations.

5) Errors on endogenous variables relate the unexplained correlations or variances arising from unmeasured variables.
6) Unanalyzed (residual) correlations among two endogenous variables are their partial correlations.
7) The total effect one variable has on another is the sum of its direct and indirect effects.

Grace’s 8 rules of path coefficients

5) Errors on endogenous variables relate the unexplained correlations or variances arising from unmeasured variables.
6) Unanalyzed (residual) correlations among two endogenous variables are their partial correlations.
7) The total effect one variable has on another is the sum of its direct and indirect effects.
8) The total effect (including undirected paths) is equivalent to the total correlation.

Goodness-of-fit

- χ^2 *statistic*: good fit when failing to reject the null hypothesis that the χ^2 statistic is different from 0 ($P > 0.05$).

- χ^2 *statistic*: good fit when failing to reject the null hypothesis that the χ^2 statistic is different from 0 ($P > 0.05$).
- *Comparative fit index (CFI)*: this statistic considers the deviation from a 'null' model. In most cases, the null estimates all variances but sets the covariances to 0. A value > 0.9 is considered good.

Goodness-of-fit

- χ^2 *statistic*: good fit when failing to reject the null hypothesis that the χ^2 statistic is different from 0 (P > 0.05).
- *Comparative fit index (CFI)*: this statistic considers the deviation from a 'null' model. In most cases, the null estimates all variances but sets the covariances to 0. A value > 0.9 is considered good.
- *Root-mean squared error of approximation (RMSEA)*: statistic penalizes models based on sample size. A value < 0.10 is acceptable, and anything < 0.08 is good.

Goodness-of-fit

- χ^2 *statistic*: good fit when failing to reject the null hypothesis that the χ^2 statistic is different from 0 ($P > 0.05$).
- *Comparative fit index (CFI)*: this statistic considers the deviation from a 'null' model. In most cases, the null estimates all variances but sets the covariances to 0. A value > 0.9 is considered good.
- *Root-mean squared error of approximation (RMSEA)*: statistic penalizes models based on sample size. A value < 0.10 is acceptable, and anything < 0.08 is good.
- *Standardized root-mean squared residual (SRMR)*: the standardized difference between the observed and predicted correlations. A value < 0.08 is considered good.

- (Multivariate) normality of endogenous variables

Assumptions of SEM with lavaan

- (Multivariate) normality of endogenous variables
- Global estimation based on variance-covariance matrix[1]

[1] Local estimation possible.

Assumptions of SEM with lavaan

- (Multivariate) normality of endogenous variables
- Global estimation based on variance-covariance matrix[1]
- Directed (acyclic) relationships[2]

[1] Local estimation possible.
[2] Causal loops possible.

Assumptions of SEM with lavaan

- (Multivariate) normality of endogenous variables
- Global estimation based on variance-covariance matrix[1]
- Directed (acyclic) relationships[2]
- Linear relationships[3]

---

[1] Local estimation possible.

[2] Causal loops possible.

[3] Nonlinear relationships possible.

Model identifiability

- Underidentified: not enough pieces of information to identify parameters uniquely (df < 0).

Model identifiability

- Underidentified: not enough pieces of information to identify parameters uniquely (df < 0).
- Saturated: Just enough information to uniquely identify parameters, but no df to check model fit (df = 0).

Model identifiability

- Underidentified: not enough pieces of information to identify parameters uniquely (df < 0).
- Saturated: Just enough information to uniquely identify parameters, but no df to check model fit (df $= 0$).
- Over-identified: parameters can be uniquely identified and positive dfs to test model goodness-of-fit (df > 0).

"t-rule" to quickly gauge whether a model is under-, just, or overidentified:

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

t = number of unknowns (parameters to be estimated, i.e. variances & covariances)
n = number of knowns (observed variables).

The LHS is how many pieces of information we want to know.
RHS: information we have (number of unique cells in the observed variance-covariance matrix).

Data requirements

- Replication should be at least 5x the number of estimated coefficients (not error variances or other correlations).

Data requirements

- Replication should be at least 5x the number of estimated coefficients (not error variances or other correlations).
- To estimate two relationships, at least n = 10 required to fit model.

Data requirements

- Replication should be at least 5x the number of estimated coefficients (not error variances or other correlations).
- To estimate two relationships, at least n = 10 required to fit model.
- Ideally, replication is 5-20x the number of estimated parameters.

Data requirements

- Replication should be at least 5x the number of estimated coefficients (not error variances or other correlations).
- To estimate two relationships, at least n = 10 required to fit model.
- Ideally, replication is 5-20x the number of estimated parameters.
- The larger the sample size, the more precise (unbiased) the estimates.

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)
3) Determine model identification

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)
3) Determine model identification
4) Select measures for the variables represented in the model

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)
3) Determine model identification
4) Select measures for the variables represented in the model
5) Collect data

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)
3) Determine model identification
4) Select measures for the variables represented in the model
5) Collect data
6) Conduct preliminary descriptive statistical analysis (e.g., scaling, missing data, collinearity issues, outlier detection)

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)
3) Determine model identification
4) Select measures for the variables represented in the model
5) Collect data
6) Conduct preliminary descriptive statistical analysis (e.g., scaling, missing data, collinearity issues, outlier detection)
7) Estimate parameters in the model

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)
3) Determine model identification
4) Select measures for the variables represented in the model
5) Collect data
6) Conduct preliminary descriptive statistical analysis (e.g., scaling, missing data, collinearity issues, outlier detection)
7) Estimate parameters in the model
8) Assess model goodness-of-fit

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)
3) Determine model identification
4) Select measures for the variables represented in the model
5) Collect data
6) Conduct preliminary descriptive statistical analysis (e.g., scaling, missing data, collinearity issues, outlier detection)
7) Estimate parameters in the model
8) Assess model goodness-of-fit
9) Check for missing or unnecessary links

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram)
3) Determine model identification
4) Select measures for the variables represented in the model
5) Collect data
6) Conduct preliminary descriptive statistical analysis (e.g., scaling, missing data, collinearity issues, outlier detection)
7) Estimate parameters in the model
8) Assess model goodness-of-fit
9) Check for missing or unnecessary links
10) Interpret and present results visually

Lavaan syntax

Define model:

```
simple <-
"mass.above ~ nadd + rich + even + precip.mm + disk
rich ~ nadd + precip.mm
even ~ nadd + precip.mm"
```

Fit model:

```
fit.simple <- sem(simple, data = seabloom)
```

Formula type	R	Meaning	Example
regression	`~`	is regressed on	`y ~ x`
correlation	`~~`	correlate errors for	`y1 ~~ y2`
latent	`=~`	set reflective indicators	`Height =~ y1 + y2 + y3`
composite	`<~`	set formative indicators	`Comp1 <~ 1*x1 + x2 + x3`
intercept	`~ 1`	estimate mean for `y`	`y ~ 1`
labelling	`*`	name coefficients	`y ~ b1*x1 + b2*x2`
defining	`:=`	define quantity	`Total := b1*b3 + b2`

Questions?

Live coding session

Your turn: working with the Seabloom dataset

Exercise 1

- Exploration of dataset (variables and treatments)

Exercise 1

- Exploration of dataset (variables and treatments)
- Check collinearity and normality

Exercise 1

- Exploration of dataset (variables and treatments)
- Check collinearity and normality
- Fitting linear models to estimate coefficients

Exercise 1

- Exploration of dataset (variables and treatments)
- Check collinearity and normality
- Fitting linear models to estimate coefficients
    - Multiple regression (direct effects of predictors on AGB)

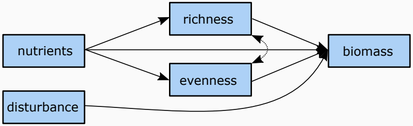
Exercise 1

- Exploration of dataset (variables and treatments)
- Check collinearity and normality
- Fitting linear models to estimate coefficients
    - Multiple regression (direct effects of predictors on AGB)
    - Multiple regression (indirect effects on richness and evenness)

Exercise 1

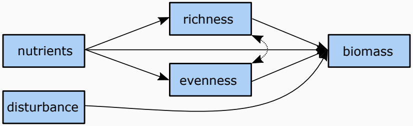
- Exploration of dataset (variables and treatments)
- Check collinearity and normality
- Fitting linear models to estimate coefficients
    - Multiple regression (direct effects of predictors on AGB)
    - Multiple regression (indirect effects on richness and evenness)
    - What can you conclude?

Exercise 2



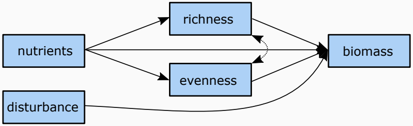
- Fitting of above SEM to Seabloom data:

Exercise 2



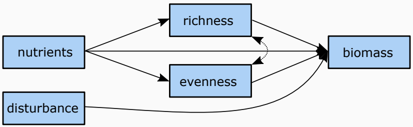
- Fitting of above SEM to Seabloom data:
    - Assess model goodness of fit

Exercise 2



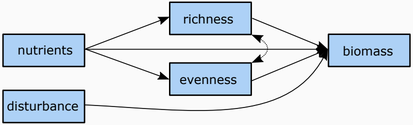
- Fitting of above SEM to Seabloom data:
    - Assess model goodness of fit
    - Investigate the modification indices. Are there paths to add that are reasonable?

Exercise 2



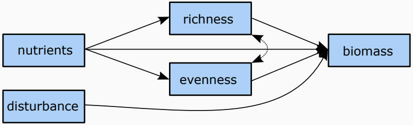
- Fitting of above SEM to Seabloom data:
    - Assess model goodness of fit
    - Investigate the modification indices. Are there paths to add that are reasonable?
    - Check model summary.

Exercise 3



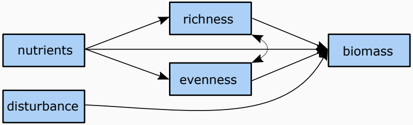
- Model analysis:

Exercise 3



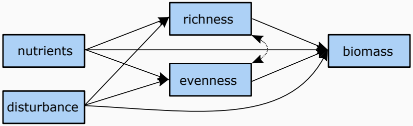
- Model analysis:
    - calculate standardized coefficients

Exercise 3



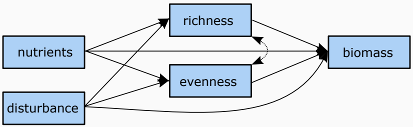
- Model analysis:
    - calculate standardized coefficients
    - add derived quantities

Exercise 4



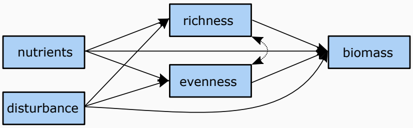
- Perform mediation analysis:

Exercise 4



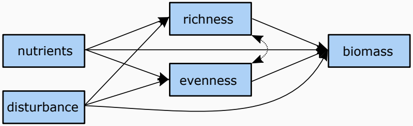
- Perform mediation analysis:
    - Is the effect of disturbance is mediated via its effect on richness and eveness, rather than directly on biomass.

Exercise 4



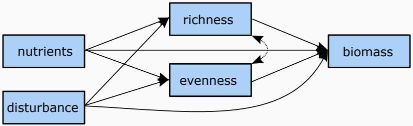
- Perform mediation analysis:
    - Is the effect of disturbance is mediated via its effect on richness and eveness, rather than directly on biomass.
    - Add paths from disk to rich and even, remove the path to mass.above

Exercise 4



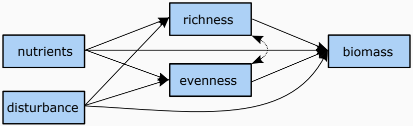
- Perform mediation analysis:
  - Is the effect of disturbance is mediated via its effect on richness and eveness, rather than directly on biomass.
  - Add paths from disk to rich and even, remove the path to mass.above
  - Compare model fit to simple model. What do you conclude?

Exercise 5



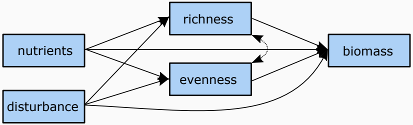
- Saturated model:

Exercise 5



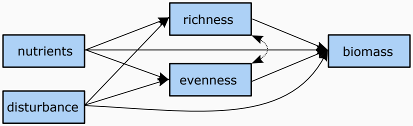
- Saturated model:
  - Model comparison with simpler models used previously.

Exercise 5



- Saturated model:
  - Model comparison with simpler models used previously.
  - Perform model pruning.

Exercise 5



- Saturated model:
    - Model comparison with simpler models used previously.
    - Perform model pruning.
    - Decide on most parsimonious model and summarize model.