Introduction to structural equation modelling - basic modelling

Frank Pennekamp

11/11/2020

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.

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 occurr 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 occurr for a reason versus randomness.
- We use statistics to understand, when connections are non-random.

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 occurr for a reason versus randomness.
- We use statistics to understand, when connections are non-random.
- Research questions are often about understanding cause and effect.

Why do we need to use SEM in Ecology

- System thinking

Why do we need to use SEM in Ecology

- System thinking
  - "Understanding the whole rather than the parts in isolation"

Why do we need to use SEM in Ecology

- 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.

Why do we need to use SEM in Ecology

- 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 implicitly assumes that the relationships among variables represent causal links.

Causality

- "Correlation does not imply causation"

Causality

- "Correlation does not imply causation"
- Causation indicates a relation between two variables in which one variable is affected by another.

Causality

- "Correlation does not imply causation"
- Causation indicates a relation between two variables in which one variable is affected by another.
- The arrow above indicates a causal relationship.

Regression models vs SEM

- We often have multiple observed variables.

Regression models vs SEM

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

Regression models vs SEM

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

Regression models vs SEM

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

Two goals of SEM:

1) Understand the patterns of correlation/covariance among a set of variables.

Two goals of SEM:

1) Understand the patterns of 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.

Good practice: Make a table / graph of putative causal relationships before analysis.

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.

Good practice: Make a table / graph of putative causal relationships before analysis.

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.
- Garbage in - garbage out (both data quality and ecologically meaningful).

Good practice: Make a table / graph of putative causal relationships before analysis.

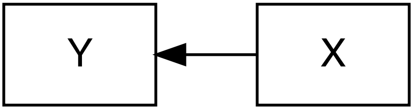
SEM modelling philosophy

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

SEM modelling philosophy

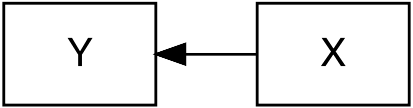
- Continuum between theory (hypothesis-driven) to exploratory (data-driven) modelling.
- Important to be explicit about the approach taken.

Graphical models:



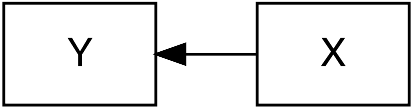
- Directed acyclic graphs (no loops).

Graphical models:



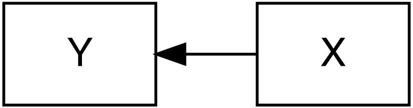
- Directed acyclic graphs (no loops).
- Variables are nodes (boxes).

Graphical models:



- 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.

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

Introduction to the dataset

Understanding the recovery of a grassland for two decades following an intensive agricultural disturbance under ambient and elevated nutrient conditions.

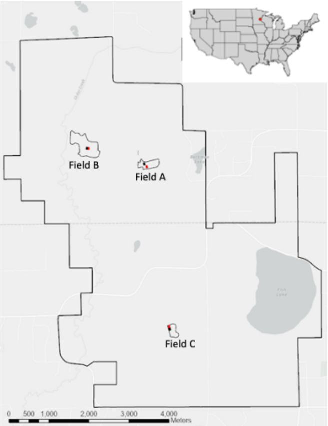
1) How has aboveground biomass changed as a function of disturbance (disking) and nutrient addition?

Introduction to the dataset

Understanding the recovery of a grassland for two decades following an intensive agricultural disturbance under ambient and elevated nutrient conditions.

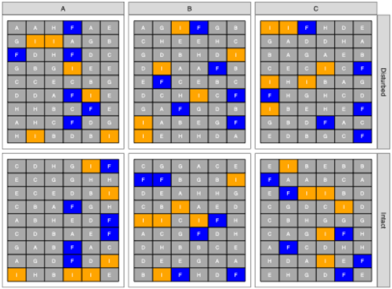
1) How has aboveground biomass changed as a function of disturbance (disking) and nutrient addition?
2) How are these effects mediated by diversity?

Introduction to the dataset



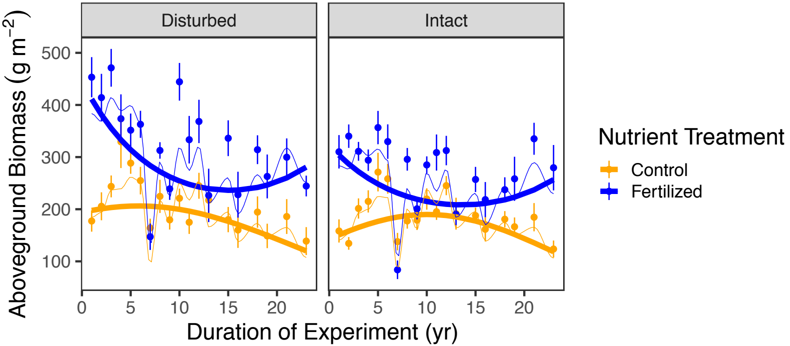
Map showing location of the study site (Cedar Creek Ecosystem Science Reserve), the location of the three study fields within the reserve, and location of the 35 x 55 m intact (black) and disturbed (red) plots within each field.

Introduction to the dataset



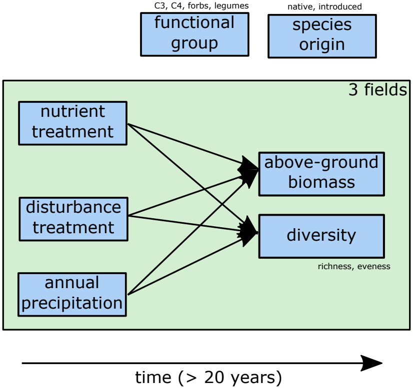
Location of the 4 x 4 m nutrient treatment plots within each 35 x 55 m Intact or Disturbed plot within each of three fields (A, B, and C). Letters indicate the nutrient treatments, and the colored plots are treatments that are the focus of the analyses presented here: Control (orange) and 9.5 g N m-2 yr-1 (blue).

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).

Question of interest

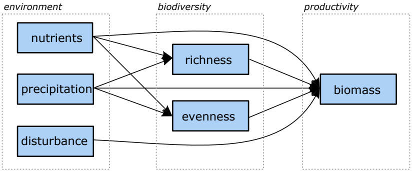


- What is the effect of richness on biomass?

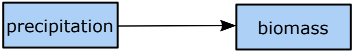
Meta-model

A metamodel summarizes the concept behind a model and links it to theory.

Productivity (biomass) is directly influence on the one hand by the environment (nutrients, disturbance and precipitation) and on the other hand by biodiversity (richness and evenness). Also some elements of the environment influence biodiversity and thus, have an additional indirect effect on productivity via biodiversity.



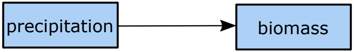
A simple bivariate model.




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

- Linear regression

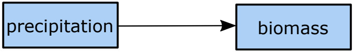
A simple bivariate model.




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

- Linear regression
- Regression coefficient quantifies the strength of relationship

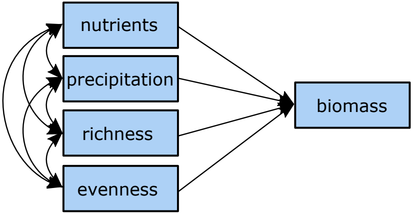
A simple bivariate model.




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

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

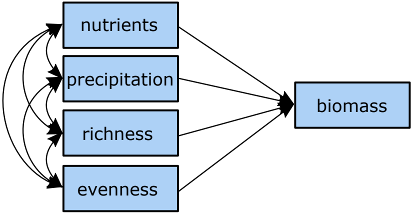
Multiple independent variables




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

- Often more than one independent variable important = multiple regression

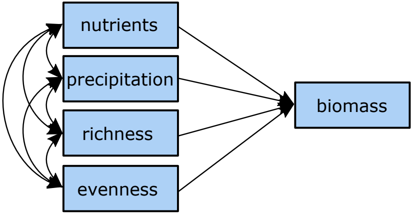
Multiple independent variables




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

- Often more than one independent variable important = multiple regression
- Estimates partial regression coefficients (i.e. effect of x1 on y when x2 is fixed)

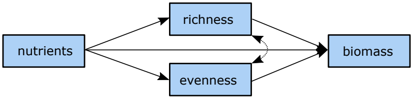
Multiple independent variables




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

- Often more than one independent variable important = multiple regression
- Estimates partial regression coefficients (i.e. effect of x1 on y when x2 is fixed)
- All effects are direct.

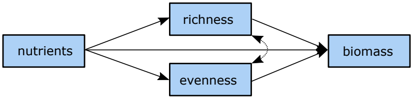
Multiple independent variables




```
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.

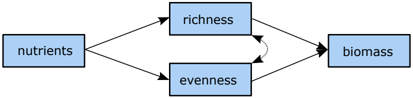
Multiple independent variables



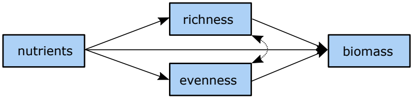

```
lm(richness ~ nutrients)
lm(evenness ~ nutrients)
lm(biomass ~ richness + evenness)
```

- 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.

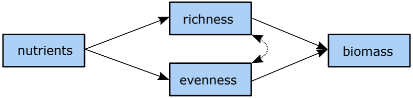
Mediation



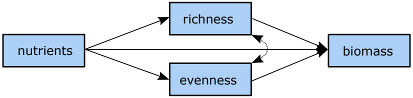
- Tests whether a particular variable has a mediating effect.



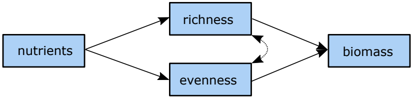
Mediation



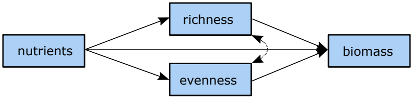
- Tests whether a particular variable has a mediating effect.
- Often used to test underlying mechanisms.



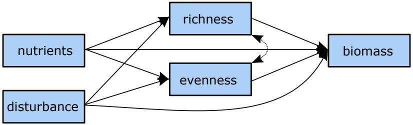
Mediation



- Tests whether a particular variable has a mediating effect.
- Often used to test underlying mechanisms.
- Full mediation versus partial mediation.

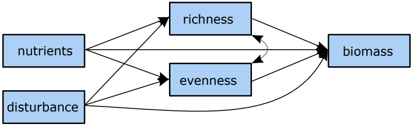


System level approach



- Variables that only have paths emanating from them (i.e., do not have arrows going into them) are called *exogenous variables*.

System level approach



- Variables that only have paths emanating from them (i.e., do not have arrows going into them) are called *exogenous variables*.
- Variables for which arrows are also directed into are called endogenous. 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 simply their bivariate correlations.

Grace's 8 rules of path coefficients

1) Unspecified relationships among exogenous variables are simply 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 simply 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 simply 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

1) Unspecified relationships among exogenous variables are simply 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.
5) Errors on endogenous variables relate the unexplained correlations or variances arising from unmeasured variables.

Grace's 8 rules of path coefficients

1) Unspecified relationships among exogenous variables are simply 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.
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.

Grace's 8 rules of path coefficients

1) Unspecified relationships among exogenous variables are simply 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.
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

1) Unspecified relationships among exogenous variables are simply 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.
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: failing to reject the null hypothesis that the χ^2 statistic is different from 0 (the ideal fit) implies a generally good representation of the data (P $>$ 0.05).

Goodness-of-fit

- χ^2 statistic: failing to reject the null hypothesis that the χ^2 statistic is different from 0 (the ideal fit) implies a generally good representation of the data (P $>$ 0.05).
- Root-mean squared error of approximation (RMSEA): this statistic penalizes models based on sample size. An 'acceptable' value is generally $<$0.10 and a 'good' value is anything $<$0.08.

Goodness-of-fit

- χ^2 statistic: failing to reject the null hypothesis that the χ^2 statistic is different from 0 (the ideal fit) implies a generally good representation of the data ($P > 0.05$).
- Root-mean squared error of approximation (RMSEA): this statistic penalizes models based on sample size. An 'acceptable' value is generally $<$0.10 and a 'good' value is anything $<$0.08.
- 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: failing to reject the null hypothesis that the χ^2 statistic is different from 0 (the ideal fit) implies a generally good representation of the data ($P > 0.05$).
- Root-mean squared error of approximation (RMSEA): this statistic penalizes models based on sample size. An 'acceptable' value is generally $<$0.10 and a 'good' value is anything $<$0.08.
- 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.
- Standardized root-mean squared residual (SRMR): the standardized difference between the observed and predicted correlations. A value $<$0.08 is considered good.

Assumptions of SEM with lavaan

- Normality

Assumptions of SEM with lavaan

- Normality
- Global estimation[1]

[1] Local estimation possible.

Assumptions of SEM with lavaan

- Normality
- Global estimation[1]
- Directed acyclic relationships[2]

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

Assumptions of SEM with lavaan

- Normality
- Global estimation[1]
- Directed acyclic relationships[2]
- Linear relationships[3]

[1] Local estimation possible.

[2] Causal loops possible.

[3] Nonlinear relationships possible.

Assumptions of SEM with lavaan

- Normality
- Global estimation[1]
- Directed acyclic relationships[2]
- Linear relationships[3]
- Backdoor criterion

[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)

Data requirements

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

Data requirements

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

Data requirements

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

Data requirements

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

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, equations)

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram, equations)
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, equations)
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, equations)
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, equations)
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, equations)
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, equations)
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 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, equations)
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 fit
9) Re-specify the model if meaningful

SEM workflow in a nutshell

1) Review the relevant theory and research literature to support model specification
2) Specify a model (e.g., diagram, equations)
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 fit
9) Re-specify the model if meaningful
10) Interpret and present results visually