Introduction to Structural Equation Modelling

About Moritz Link

- PhD-Student at University Koblenz-Landau, Germany (link@uni-landau.de)

- Working on:
  - Monitoring program for small streams (partners: UFZ, UBA)
  - Fungicide effects on aquatic fungi
  - Data analysis with SEM

About Moritz Link

- PhD-Student at University Koblenz-Landau, Germany (link@uni-landau.de)

- Working on:
  - Monitoring program for small streams (partners: UFZ, UBA)
  - Fungicide effects on aquatic fungi
  - Data analysis with SEM

About Moritz Link

- PhD-Student at University Koblenz-Landau, Germany (link@uni-landau.de)

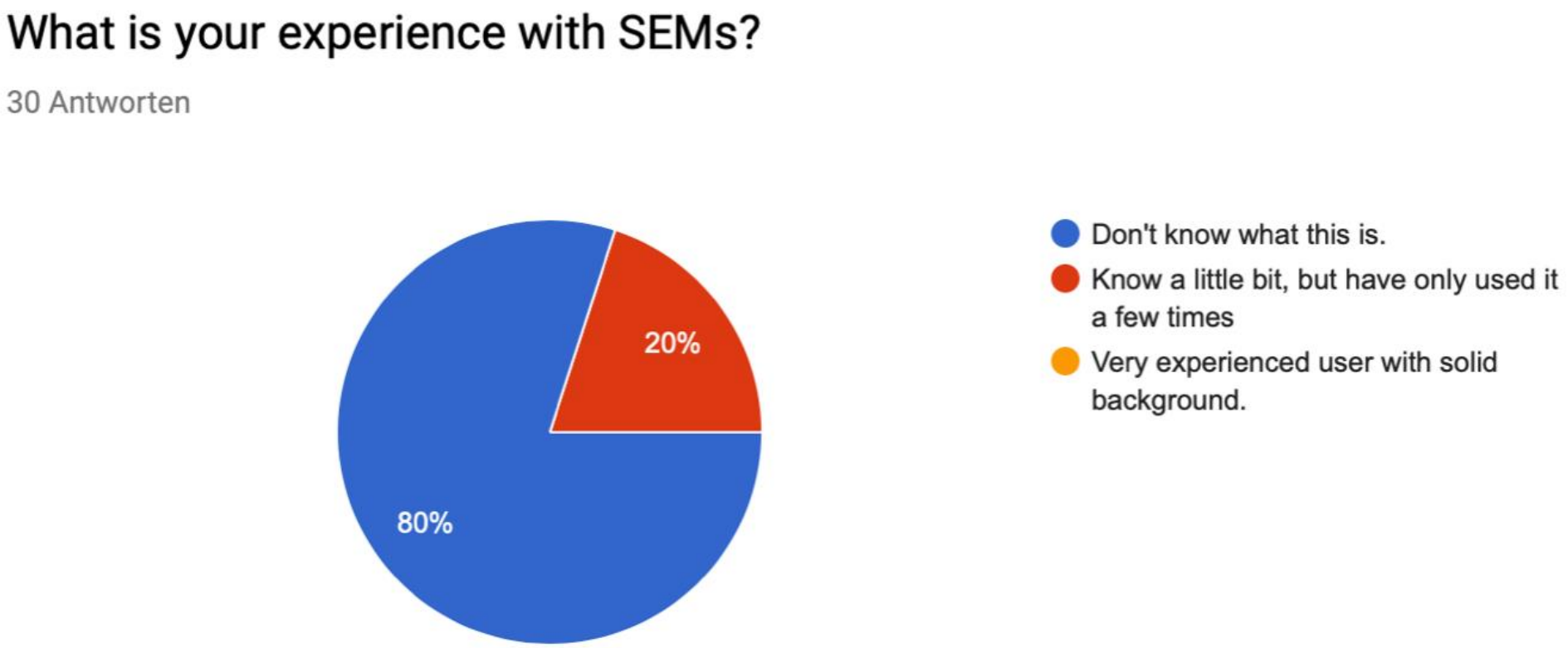
- Working on:
  - Monitoring program for small streams (partners UFZ, UBA)
  - Fungicide effects on aquatic fungi
  - Data analysis with SEM
- NOT an expert in SEM!
- Experts are these guys:
  - **Jarrett Byrnes, http://byrneslab.net/project/byrnes/**
  - **Jon Lefcheck, https://jonlefcheck.net/**

About Moritz Link

- My first thoughts on this workshop:

About Moritz Link

- Results of the poll:



About this workshop

- 1 Background
- 2 Definitions and rules
- 3 Two SEM packages for R
- 4 Let’s fool around with R
- 5 What we did not cover

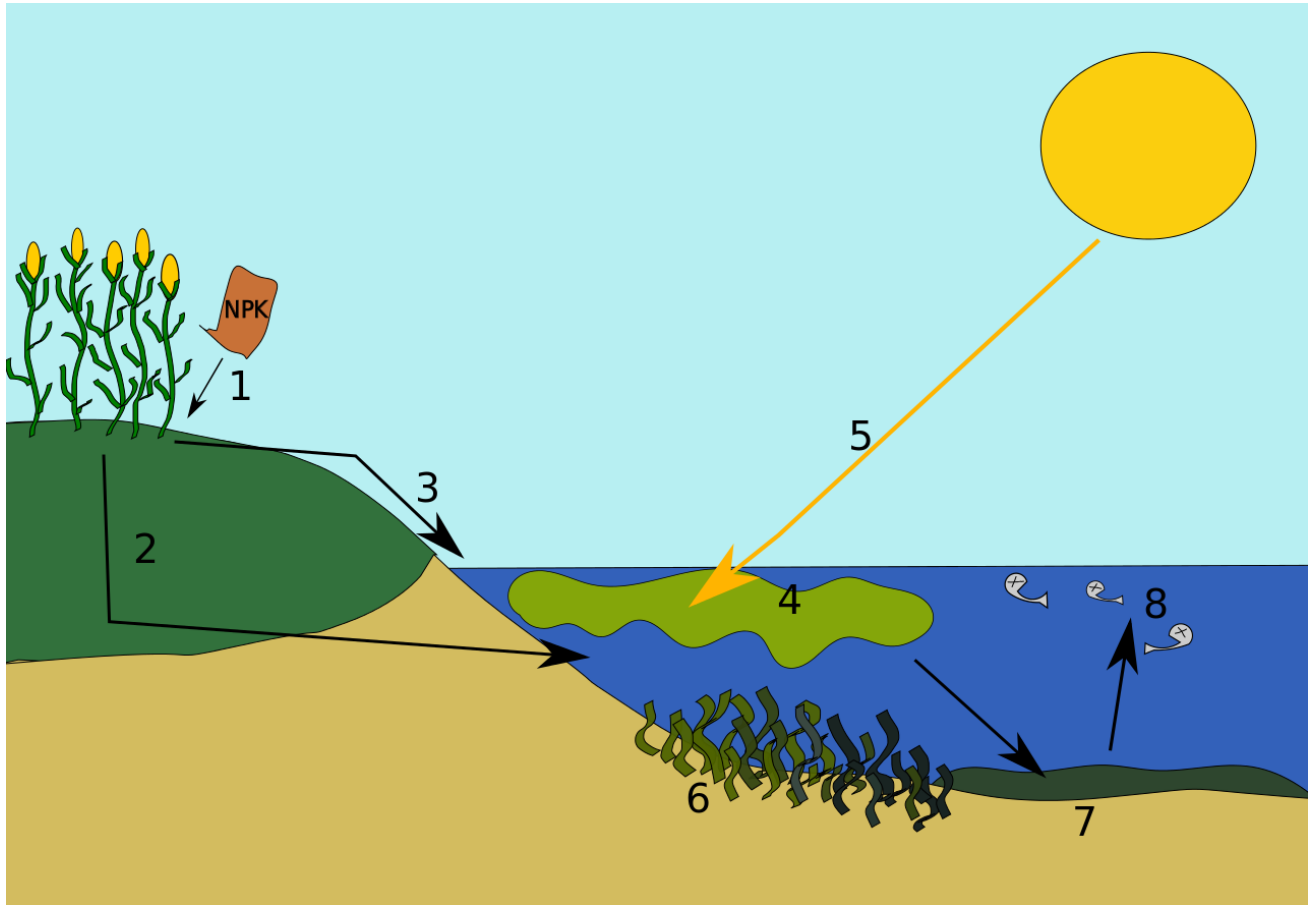
About this workshop

- 1 Background
- 2 Definitions and rules
- 3 Two SEM packages for R
- 4 Let’s fool around with R
- 5 What we did not cover

Background of SEM

Background of SEM

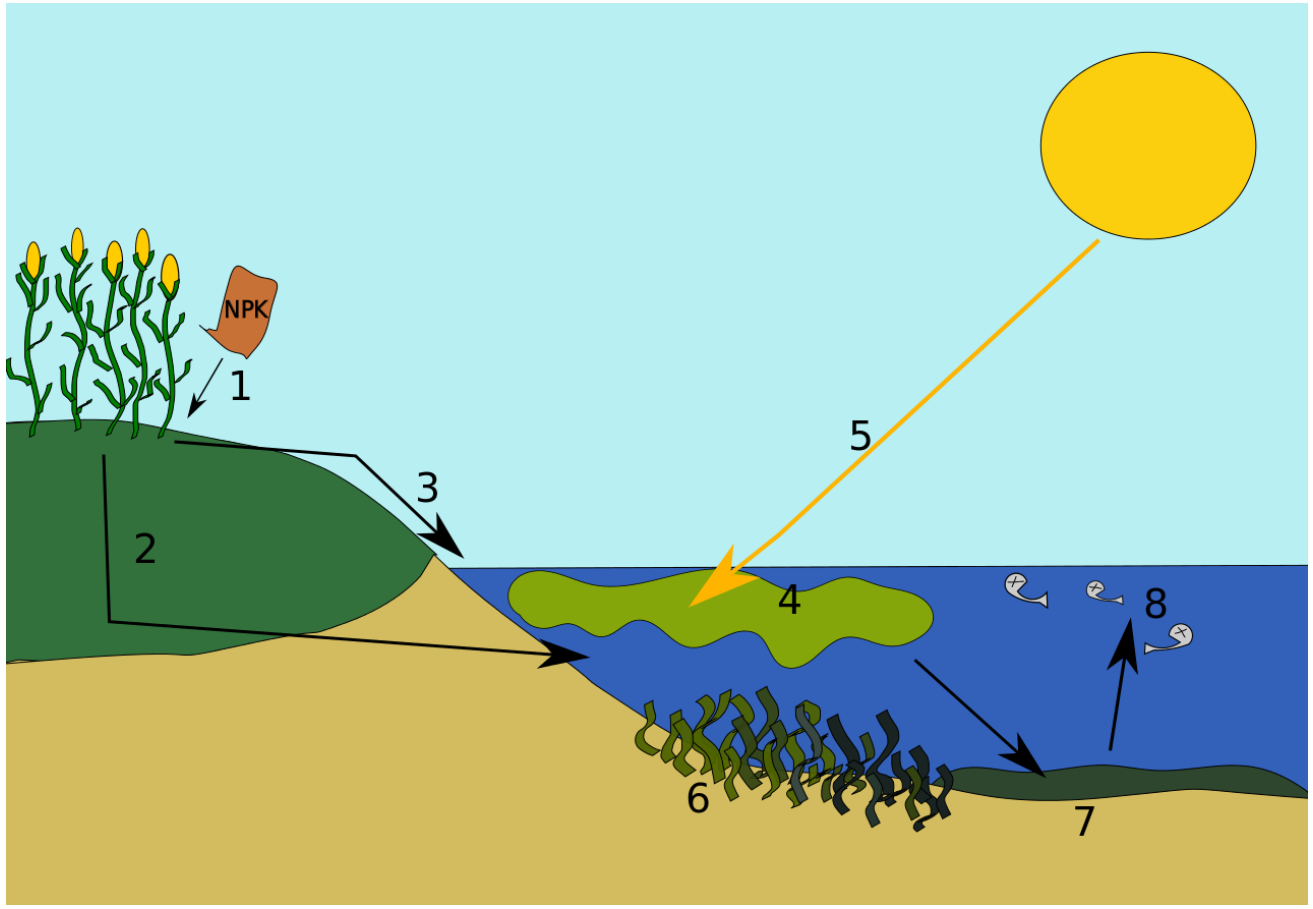
- What is Structural Equation Modelling?



By Kungfucrab - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=49234478

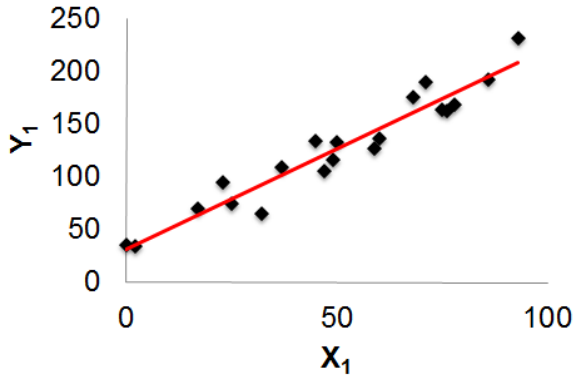
Background of SEM

- What is Structural Equation Modelling?
  - Observing a process (**Structure**)
  - Expressing it as mathematical **equations**
  - Using collected data to **model**/test your proposed process
- Other names:
  - Confirmatory path analysis
  - Directed acyclic graph models



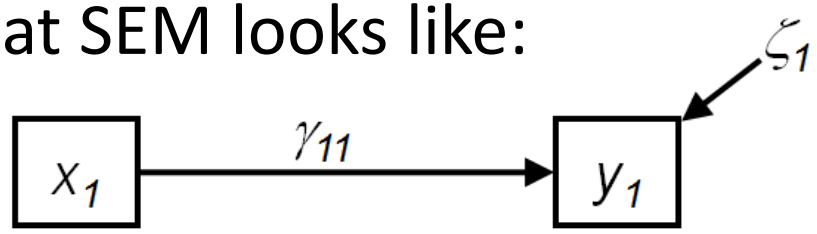
Background of SEM

- That's like regression, right!?

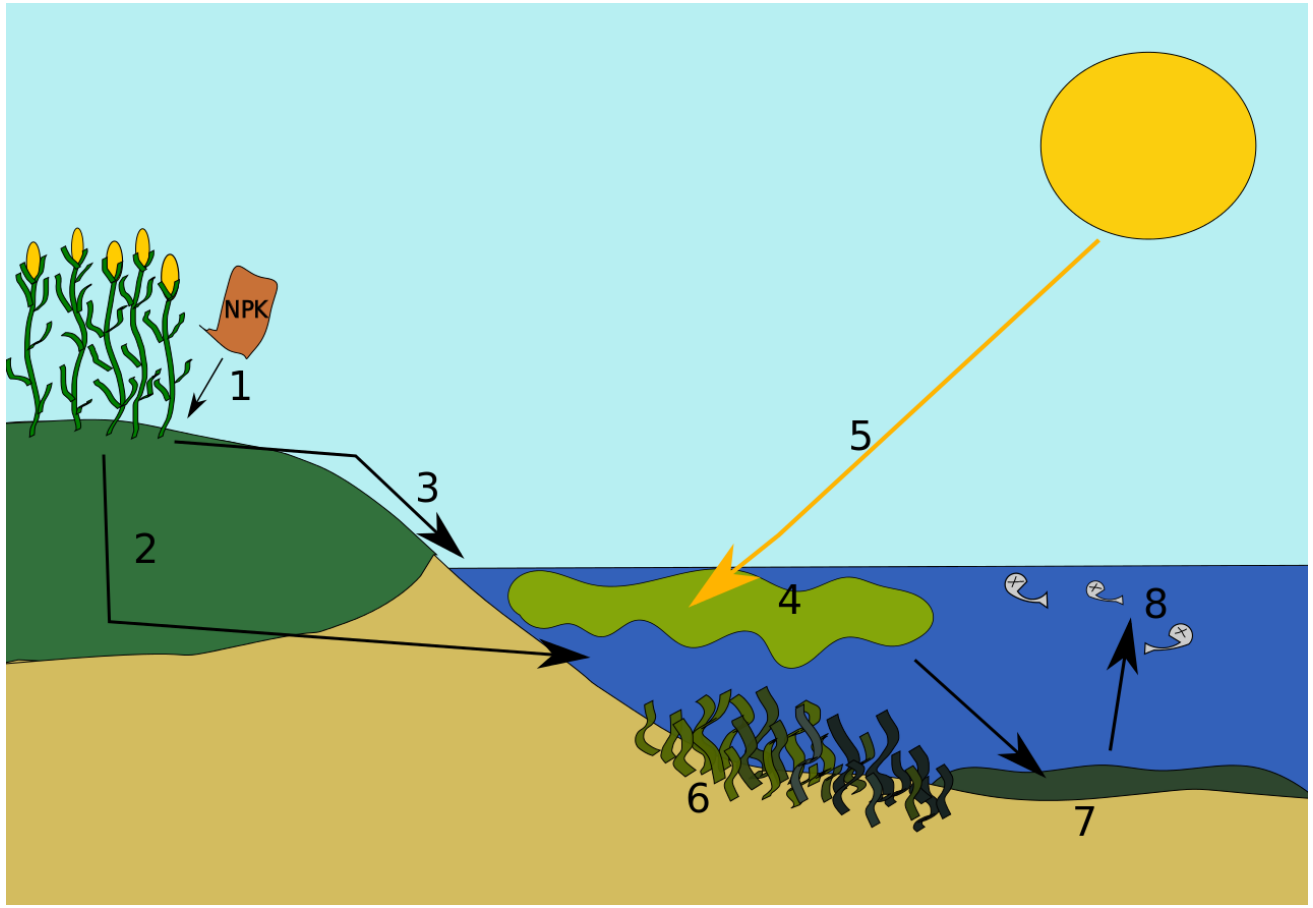


$$y_1 \sim \alpha_1 + \beta_1 X_1 + \varepsilon$$

- What SEM looks like:



$$y_1 \sim \alpha_1 + \gamma_{11} x_1 + \zeta_1$$



By Kungfucrab - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=49234478

Background of SEM

- SEM is a tool and can provide a:
  - Conceptual framework
  - Workflow process
  - Means of modeling systems or networks
  - Means of testing hypotheses
  - Method of learning

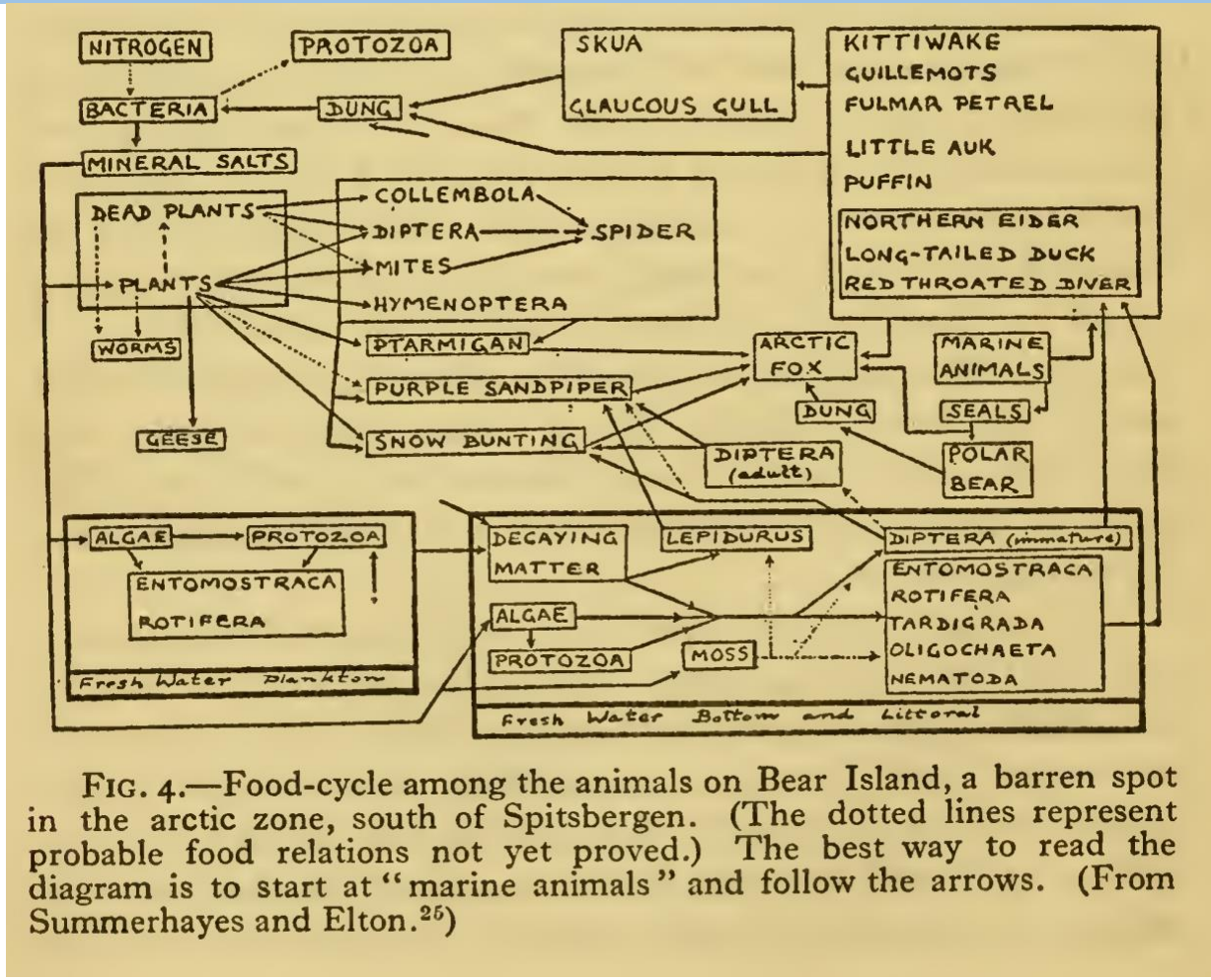


FIG. 4.—Food-cycle among the animals on Bear Island, a barren spot in the arctic zone, south of Spitsbergen. (The dotted lines represent probable food relations not yet proved.) The best way to read the diagram is to start at "marine animals" and follow the arrows. (From Summerhayes and Elton.[25])

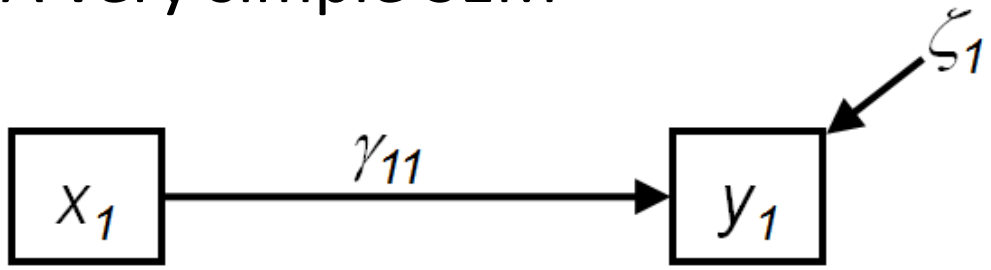
Charles Elton, 'Animal Ecology' (1927, p.58)

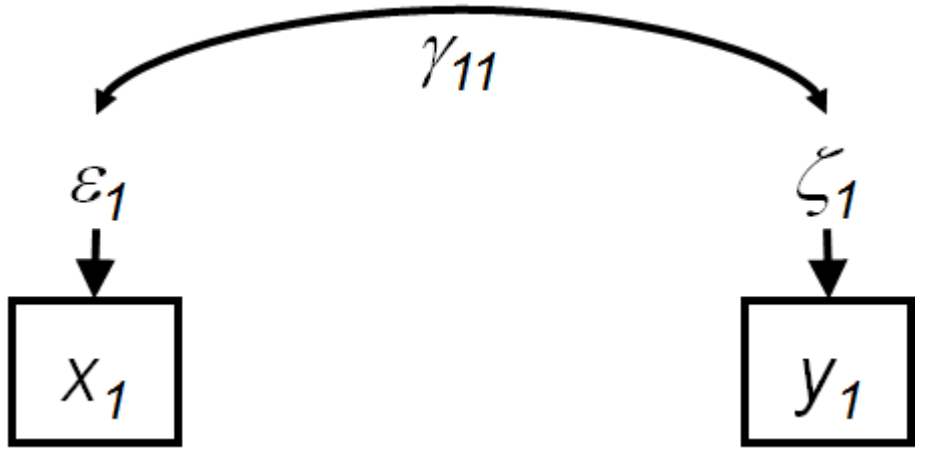
Background of SEM

- **Important**:
  - SEM assumes a causal effect
  - Causality **needs sufficient evidence** (prior observations, experiments, e.g.)

As always: Correlation does not mean causation!

A very simple SEM





Background of SEM

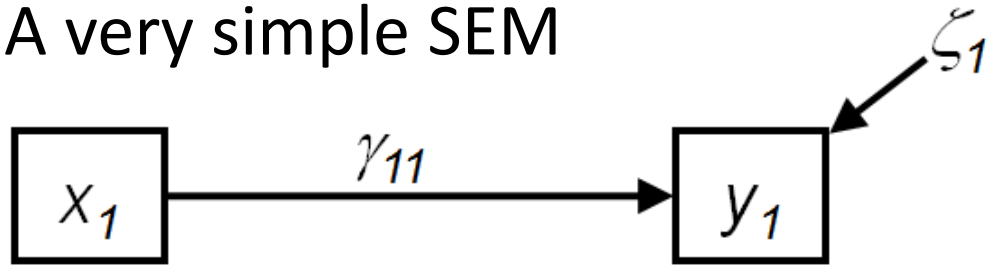
- **Important**:
  - SEM assumes a causal effect
  - Causality **needs sufficient evidence** (prior observations, experiments, e.g.)
  - Combining inferences across the model, SEM analysis direct and indirect effects in a system

$$y_1 = \alpha_1 + \gamma_{11} x_1 + \zeta_1$$

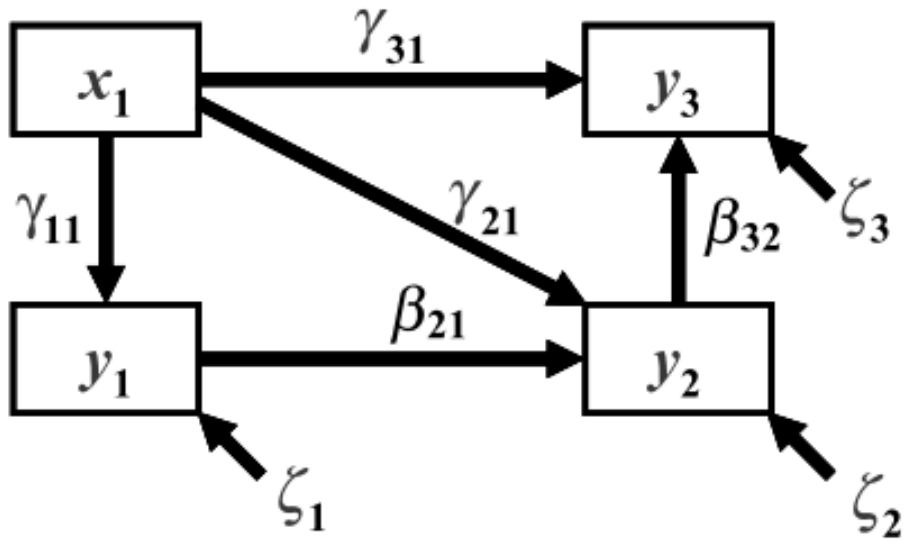
$$y_2 = \alpha_2 + \beta_{21} y_1 + \gamma_{21} x_1 + \zeta_2$$

$$y_3 = \alpha_3 + \beta_{32} y_2 + \gamma_{31} x_1 + \zeta_3$$

A very simple SEM



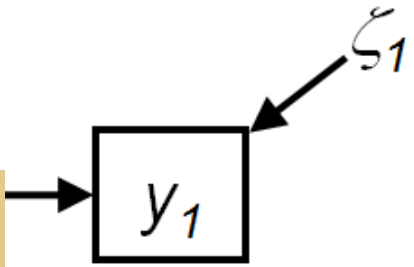
Nature brings complexity



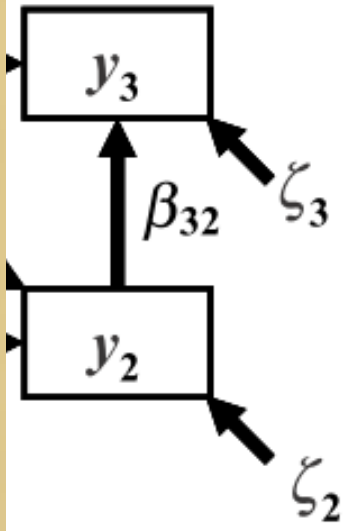
Background of SEM

If your sample size is large enough…ALL THE INFERENCE!

A very simple SEM



olexity



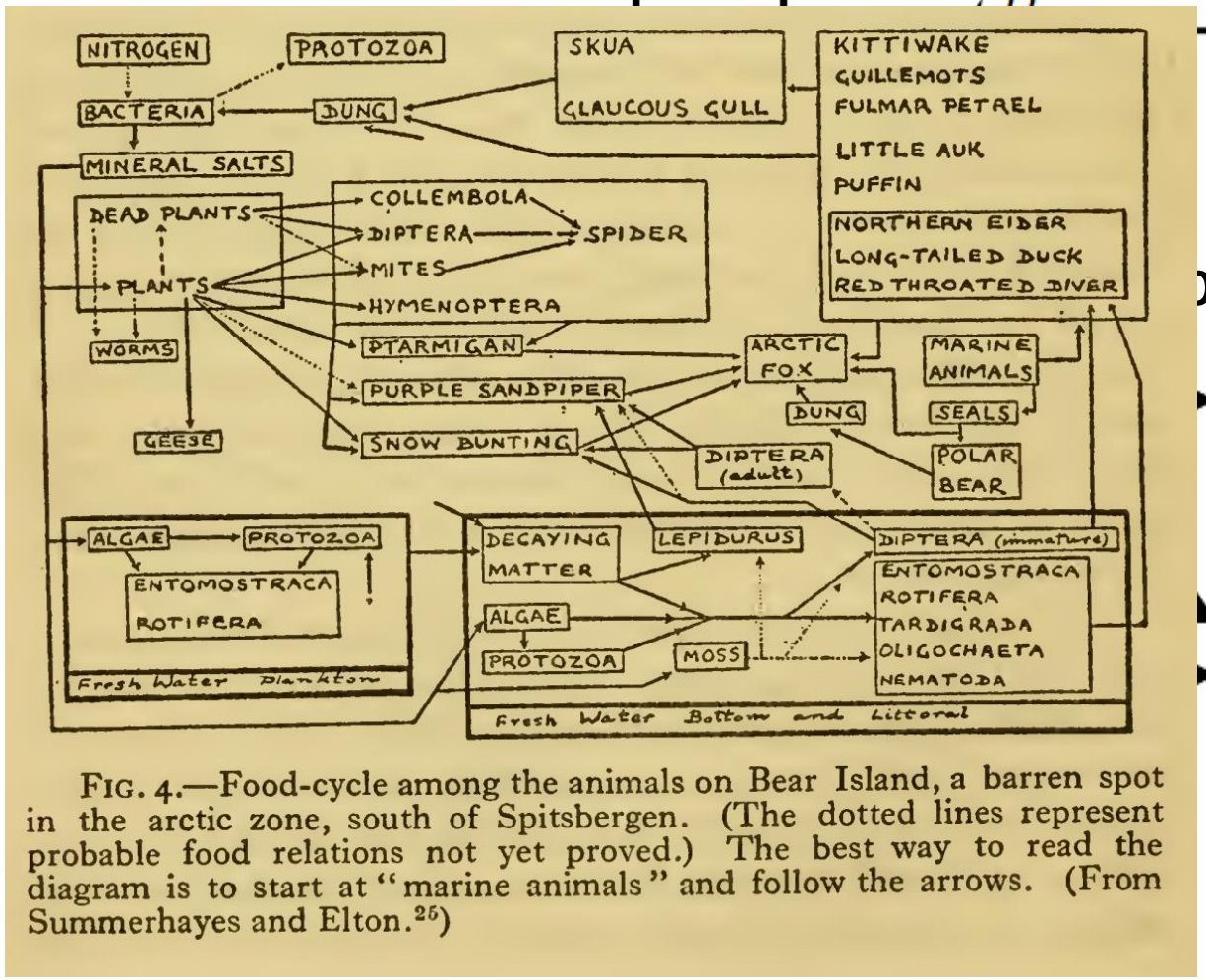
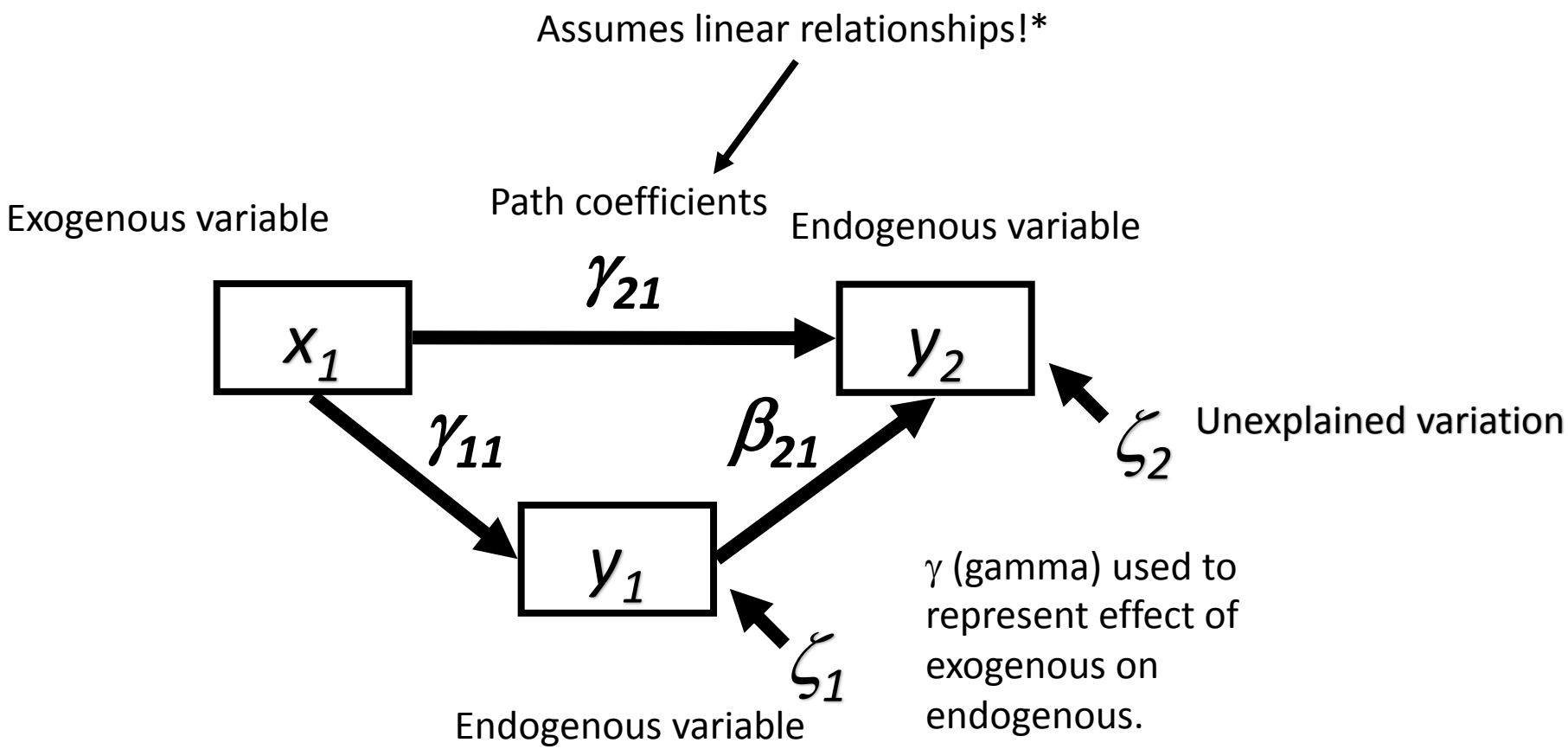


FIG. 4.—Food-cycle among the animals on Bear Island, a barren spot in the arctic zone, south of Spitsbergen. (The dotted lines represent probable food relations not yet proved.) The best way to read the diagram is to start at "marine animals" and follow the arrows. (From Summerhayes and Elton.[25])

Charles Elton, 'Animal Ecology' (1927, p.58)

Definitions and rules

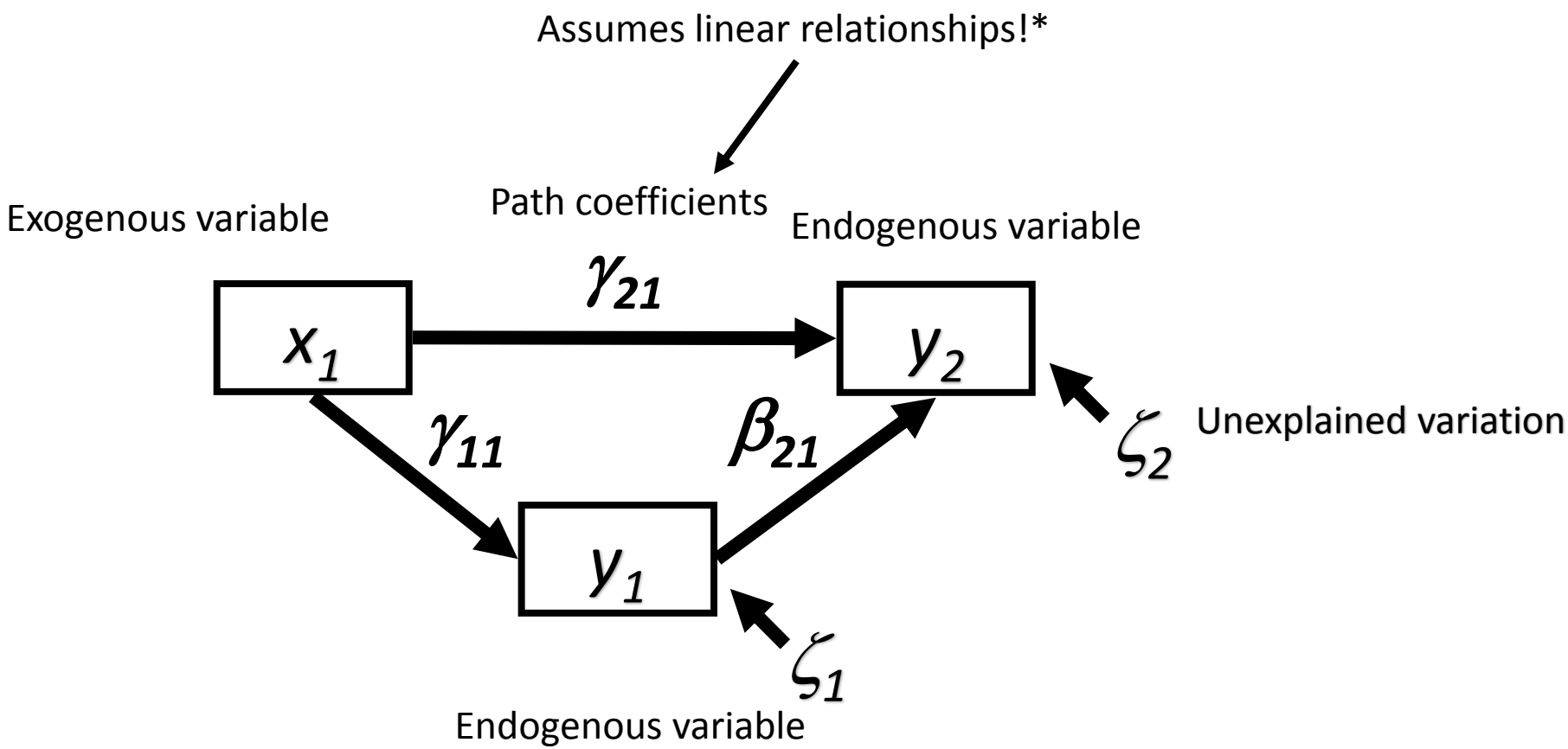
Definitions and rules



γ (gamma) used to represent effect of exogenous on endogenous.

β (beta) used to represent effect of endogenous on endogenous.

Definitions and rules

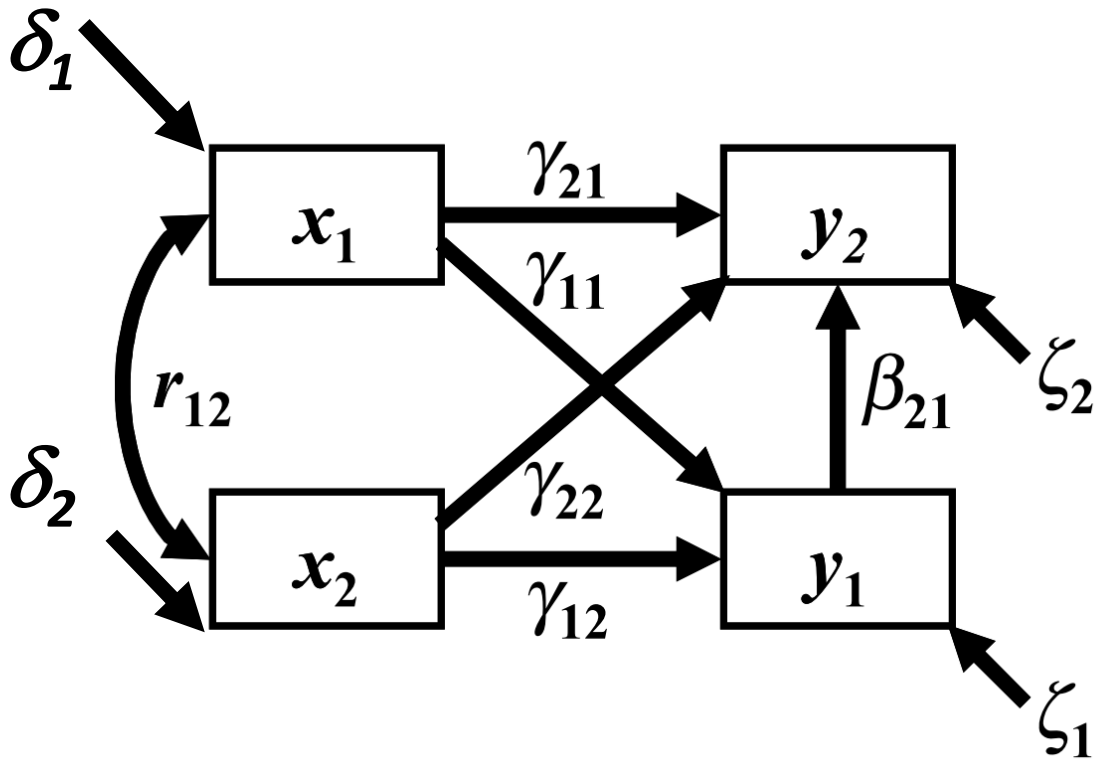


Direct effect: $x_1 \rightarrow y_2$

Indirect effect: $x_1 \rightarrow y_1 \rightarrow y_2$

Definitions and rules

- You can't just come up with any model!
- Beware of the **t-rule**, check if your model can be estimated



$$\underline{t\text{-rule}} = t \leq n(n+1)/2$$

t = # of unknowns (parameters to be estimated)

n = # of knowns (observed variables)

Definitions and rules

- Path coefficients:
  - Covariances for model fitting (depend on measuring unit)
  - Standardized covariances (correlations) for interpretation

Raw Covariance Matrix

	x_1	x_2	y_1
x_1	0.81		
x_2	0.87	1.63	
y_1	0.88	1.80	4.98

variance

covariance

Standardized Covariance Matrix

	x_1	x_2	y_1
x_1	1.0		
x_2	0.76	1.0	
y_1	0.44	0.63	1.0

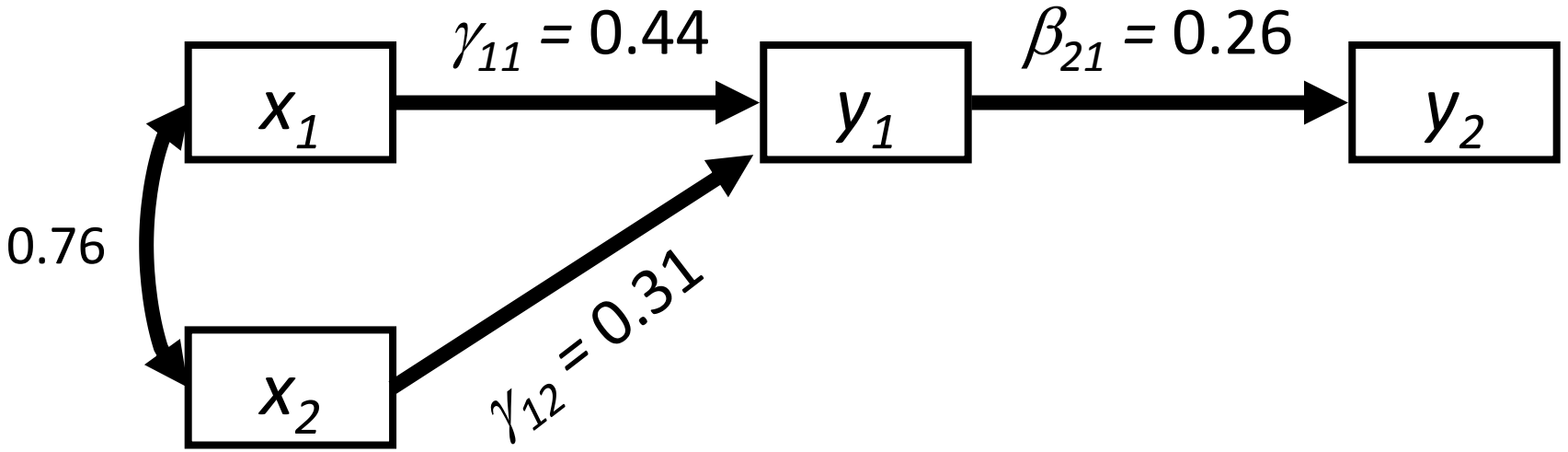
correlation

Definitions and rules

- Path coefficients: **8 rules**

1: path coefficients for unanalyzed relationships between exogenous variables are simply the correlations or covariances

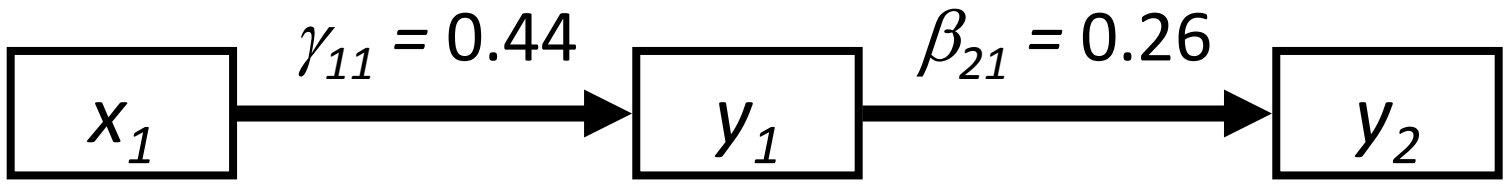
2: For variables connected by a single causal path, the coefficient equals the correlation coefficient



Definitions and rules

- Path coefficients: **8 rules**

3: strength of a compound along a path (x_1 -> y_1 -> y_2) is the product of the coefficients along that path

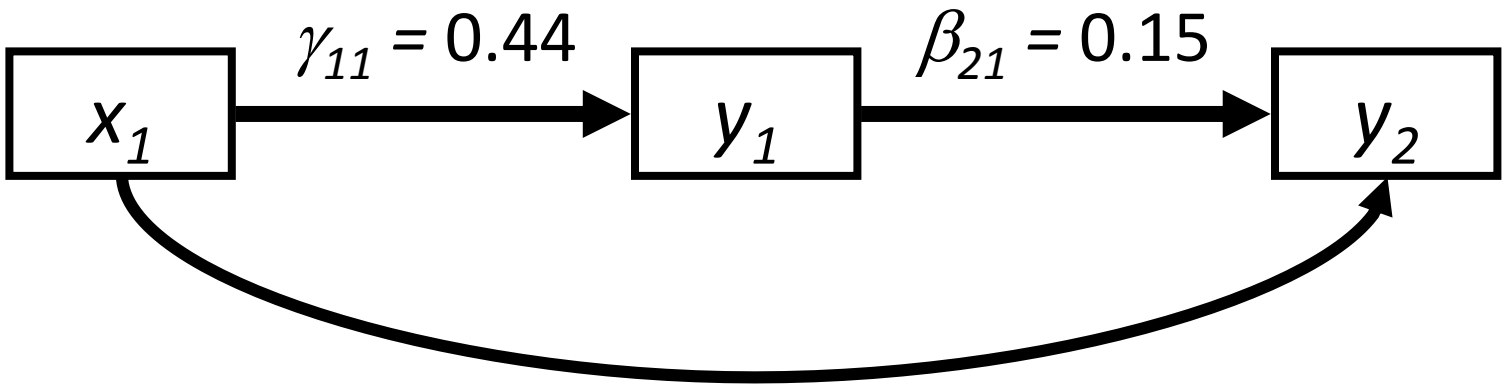


x_1 and y_2 are conditionally independent!

Definitions and rules

- Path coefficients: **8 rules**

4: When variables are connected by more than one pathway, the coefficients are ‘partial’ regression coefficients

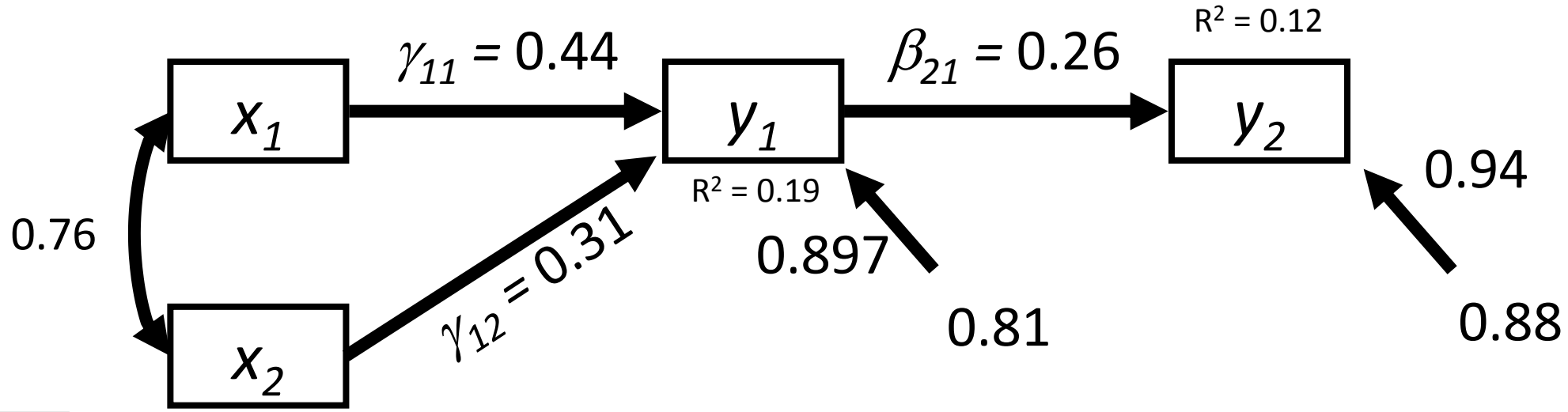


x_1 and y_2 are NOT conditionally independent!

Definitions and rules

- Path coefficients: **8 rules**

5: path from error variables represent prediction error



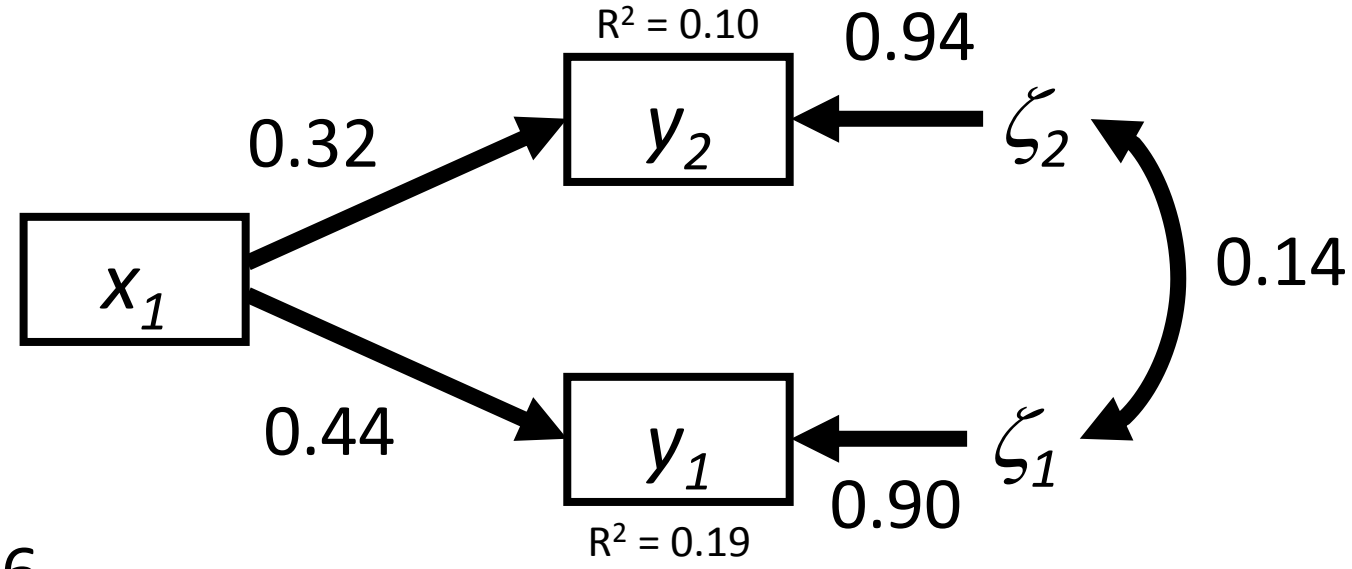
path $= \sqrt{1 - R^2_{y_i}}$

zeta $= 1 - R^2$

Definitions and rules

- Path coefficients: **8 rules**

6: unanalyzed residual correlations between endogenous variables are partial correlations or covariances



Total correlation between y_1 and y_2:

0.32 * 0.44 + 0.94 * 0.14 * 0.90 = 0.26

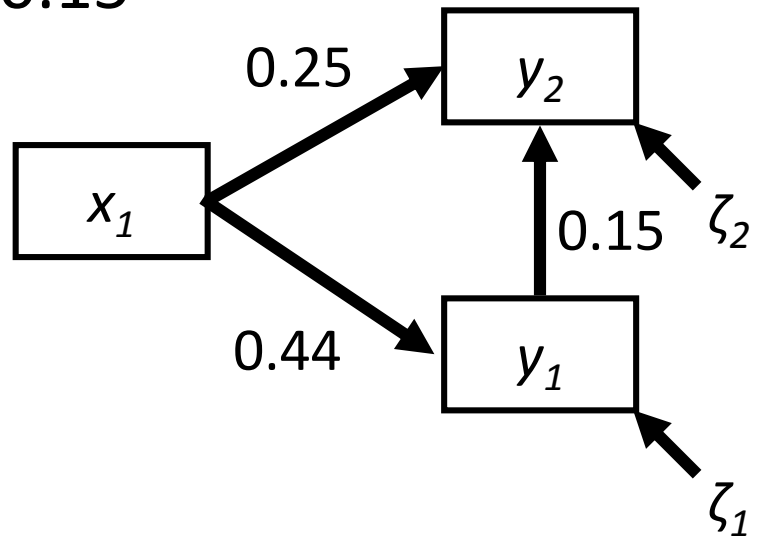
Definitions and rules

- Path coefficients: **8 rules**

7: total effect one variable has on another equals the sum of its direct and indirect effects

Total effect = 0.25 + 0.44 * 0.15

8: sum of all pathways between two variables equals the correlation



Definitions and rules

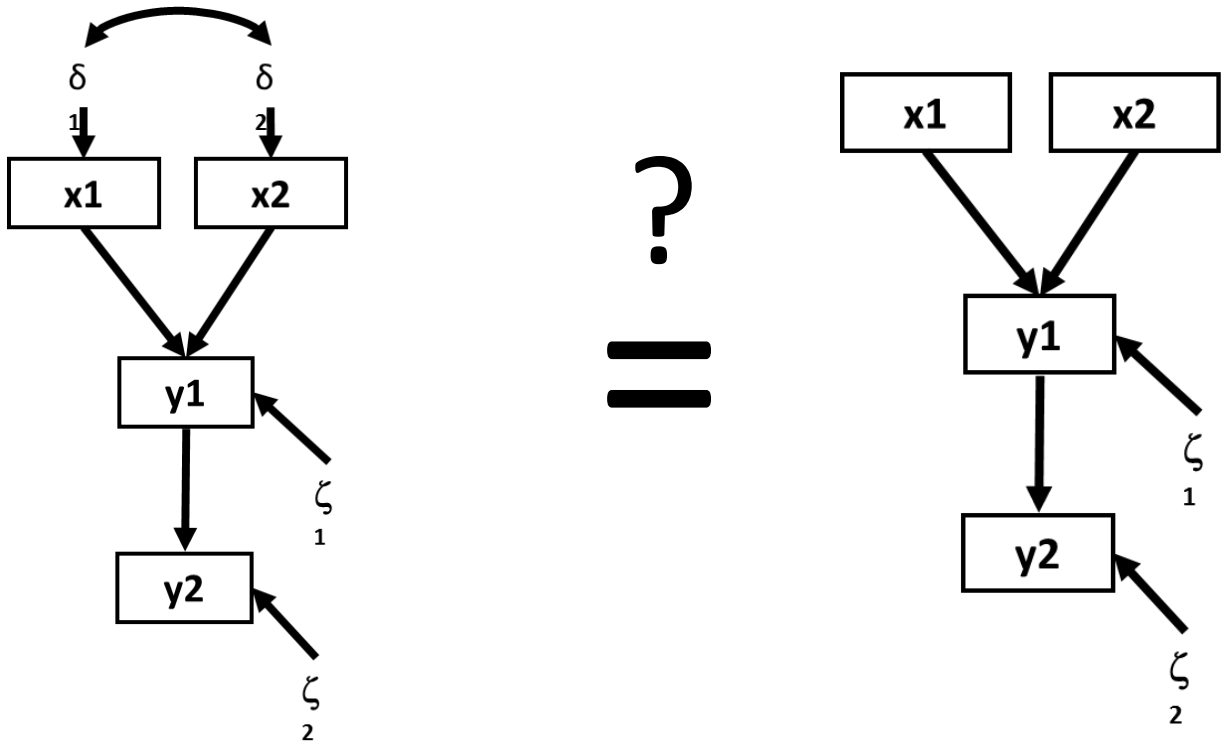
- Identifiability of your model:

- What to avoid:
  - Too many paths relative to number of variables (t-rule)
  - High multi-collinearity (r>0.9)
  - Complex models with small sample sizes
  - Certain structures

Definitions and rules

- Identifiability of your model:

- t-rule again: **$t \leq (p+q)(p+q+1)/2$**
  - t = parameters, p = end. variables, q = exog. variables

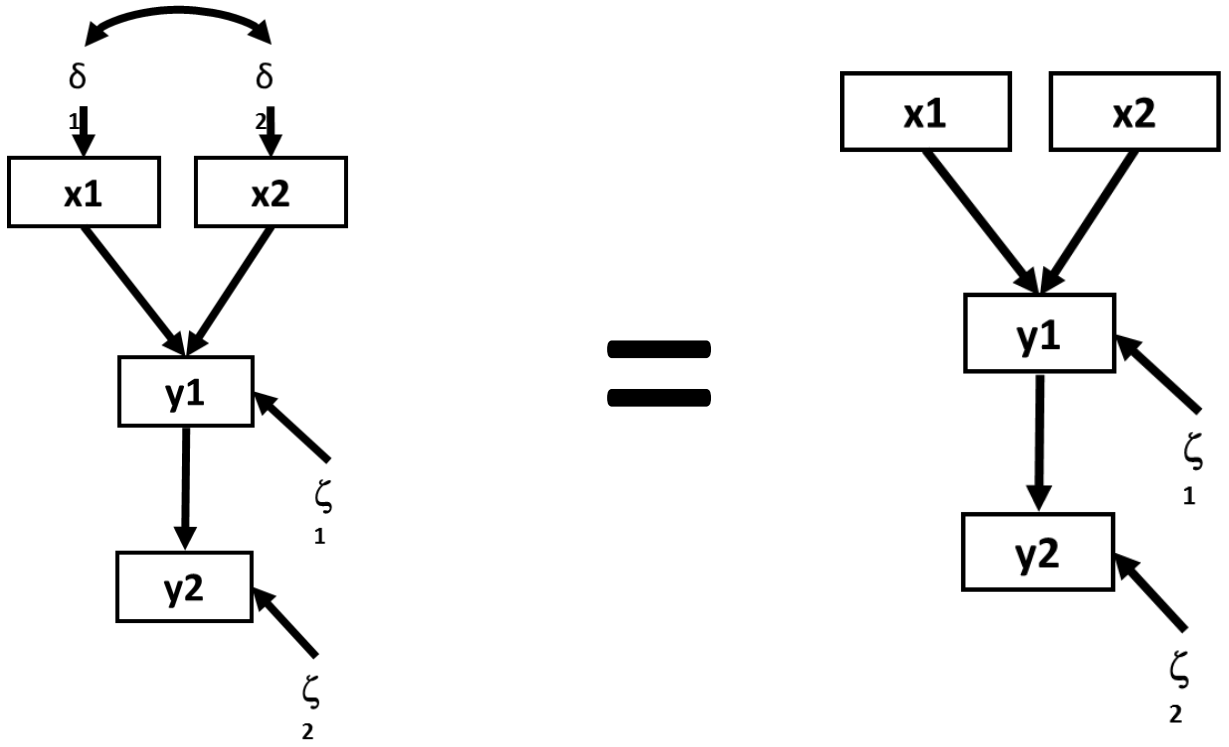


Definitions and rules

- Identifiability of your model:

- t-rule again: **$t \leq (p+q)(p+q+1)/2$**
  - t = parameters, p = end. variables, q = exog. variables

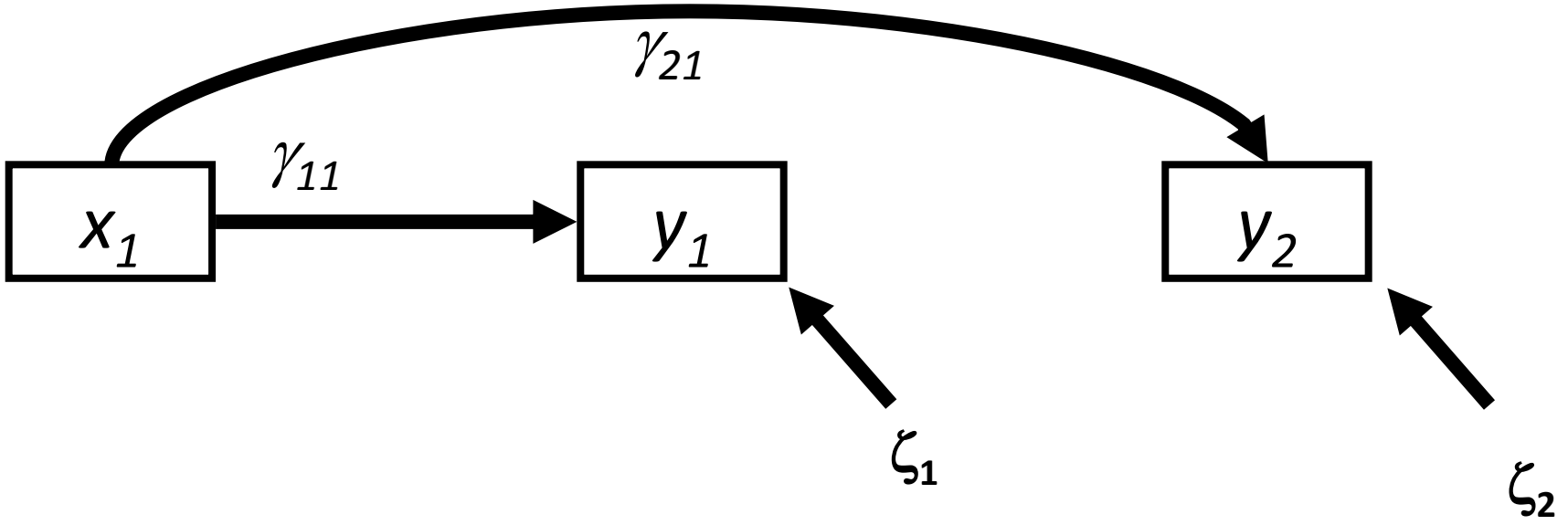
YES, DF for both models = 2



Definitions and rules

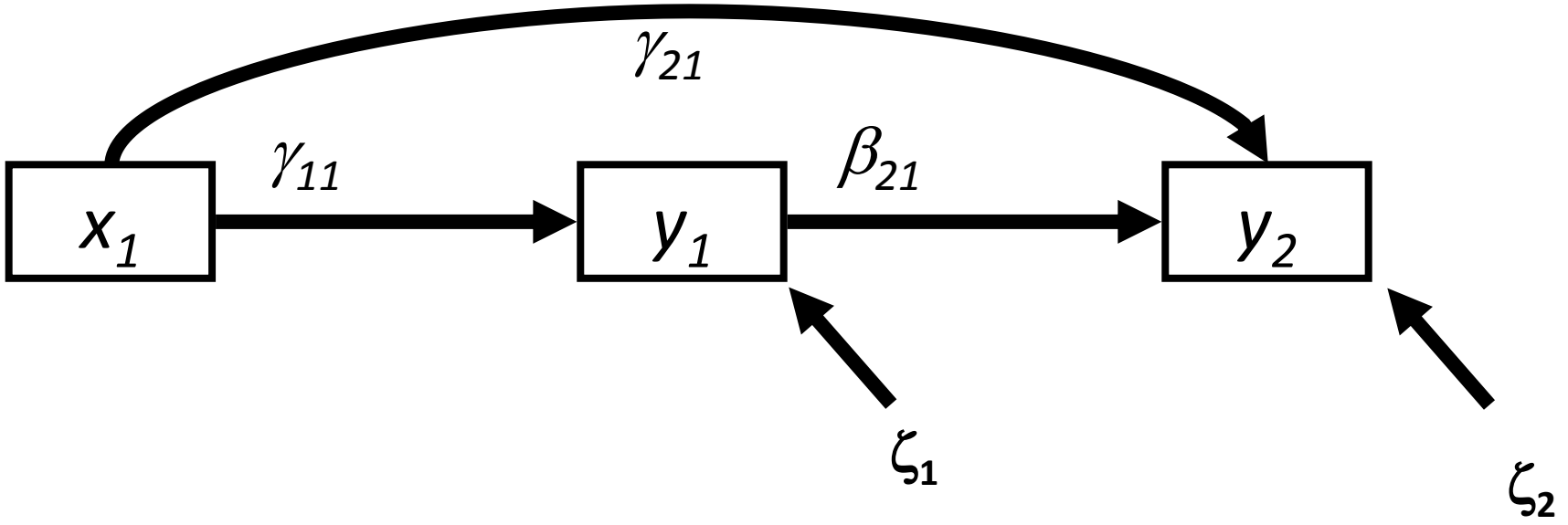
- Identifiability of your model:

- **Many regressions,** no relationships between endogenous variables, sufficient condition



Definitions and rules

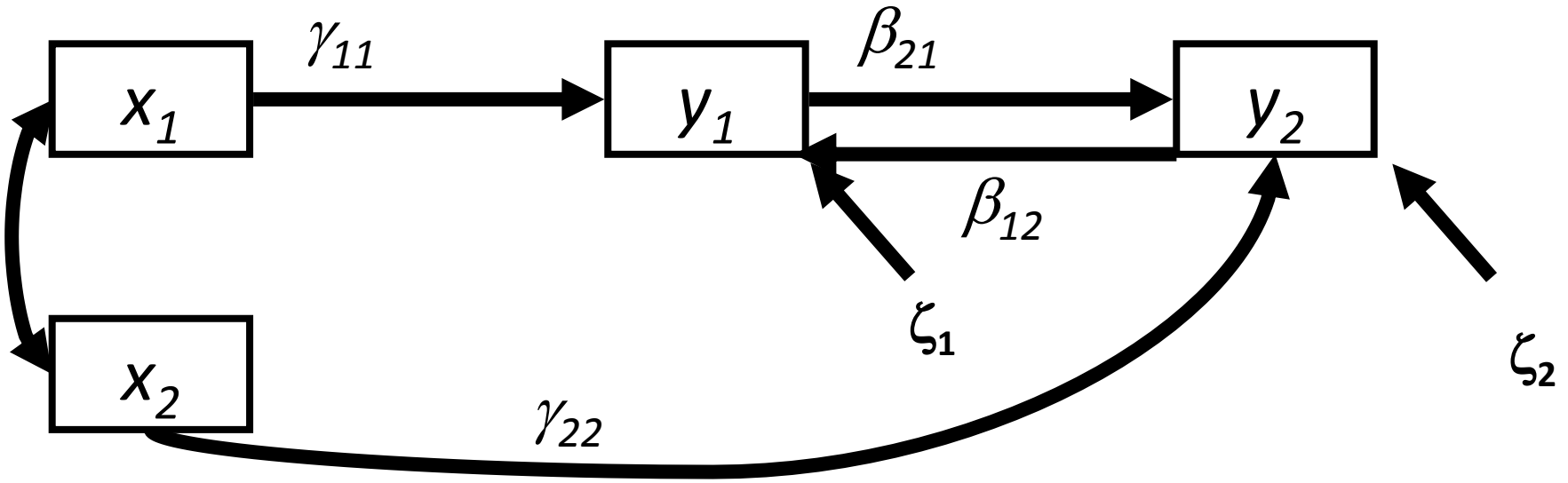
- Identifiability of your model:
- **If no feedbacks,** model is recursive**:** sufficient condition



Definitions and rules

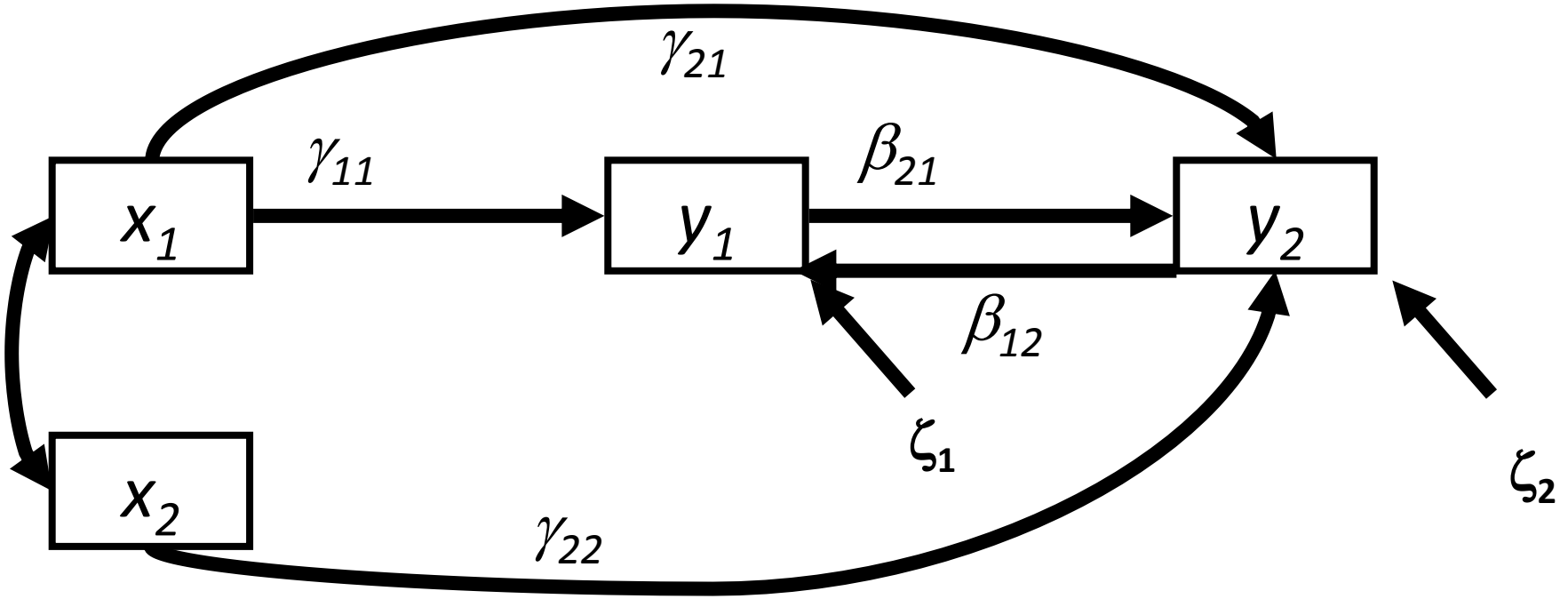
- Identifiability of your model:

- **With feedbacks,** model is non-recursive, but y's have unique information: sufficient condition



Definitions and rules

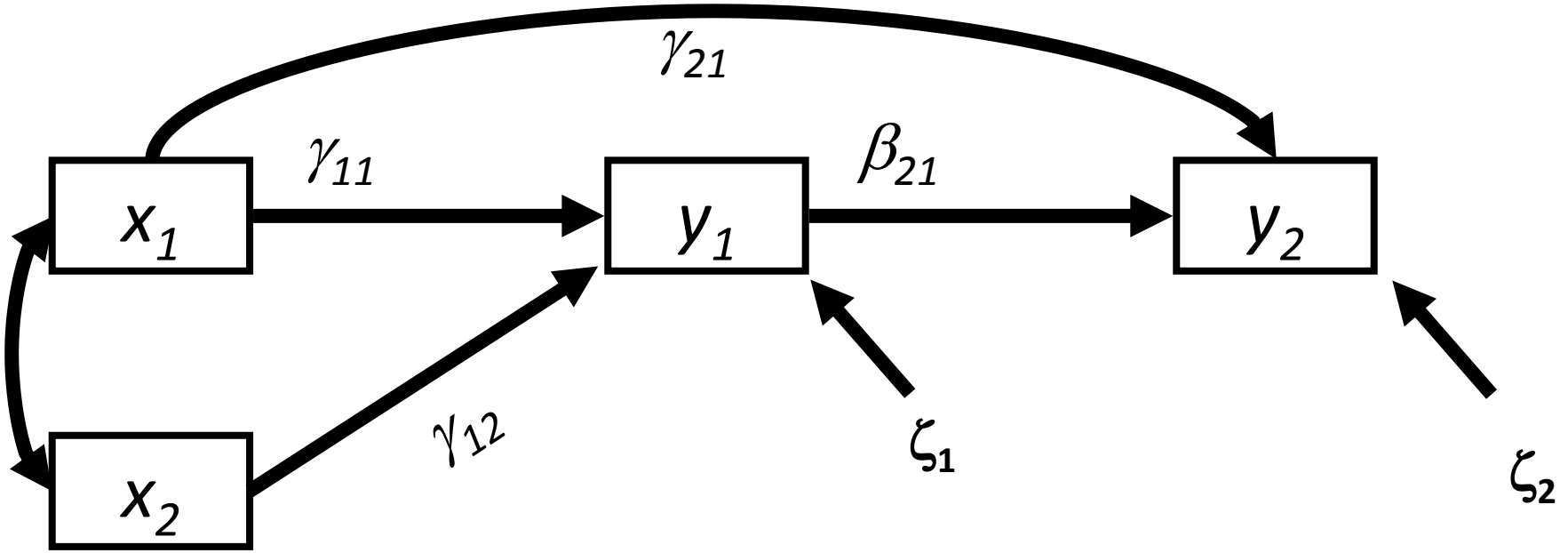
- Identifiability of your model:
- **Non-recursive model with feedbacks: Not identified**



Definitions and rules

- Identifiability of your model:

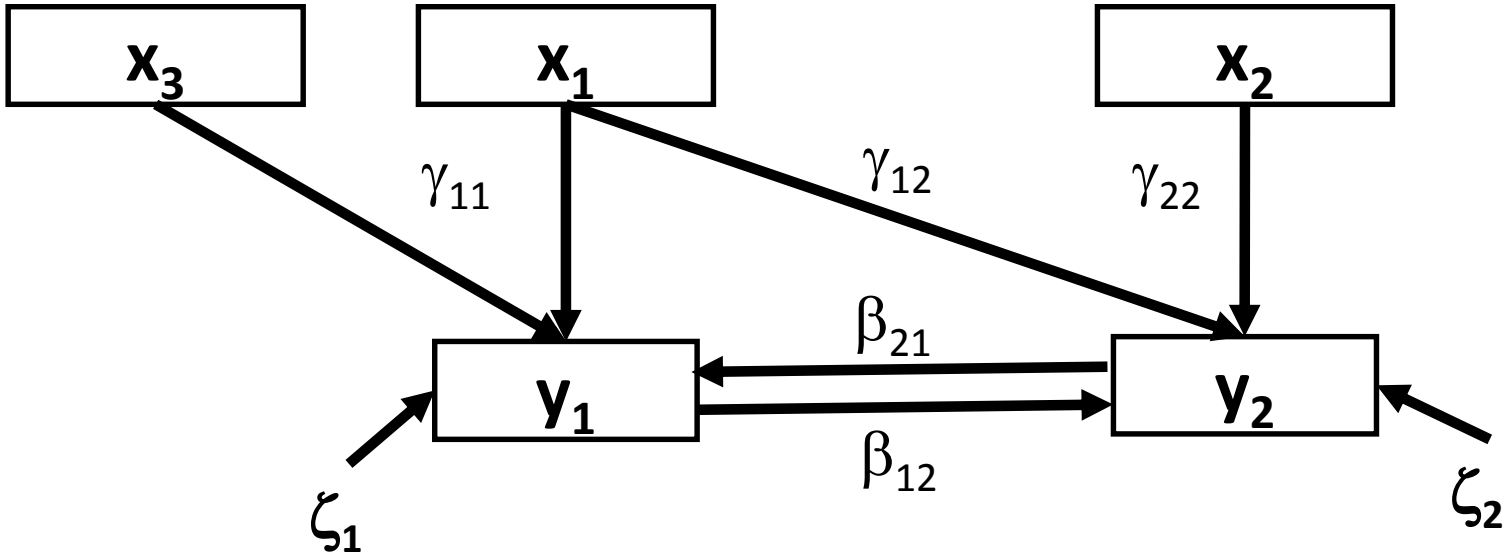
- The **Order condition: G ≤ H, necessary condition**
  - G = incoming paths
  - H = number of exogenous variables + indirectly connected endogenous vars



Definitions and rules

- Identifiability of your model:

- The **Rank condition: Sufficient condition**
- Slightly more complicated, not relevant for this course



Definitions and rules

- Identifiability of your model:

- The **Rank condition: Sufficient condition**
- Slightly more complicated, not relevant for this course

- Check here for detailed explanations:

  https://www.bauer.uh.edu/rsusmel/phd/ec1-16.pdf

  https://ebrary.net/1028/economics/identification

  https://stat.utexas.edu/software-faqs/lisrel/146-training/software/655-lisrel-assessing-model-identification

Definitions and rules

- What to keep in mind about sample size:
  - An exogenous variable's effect gets weaker the further you get into the model
  - Sample size and the number of estimated parameters determine our ability to detect inferences
  - Sample size sets an upper limit to the complexity of your model
  - Sample size influences your ability to detect lack of model fit
- Rule of thumb
  - At least five samples per estimated parameter, better 20 samples
  - Estimated parameters are your path coefficients, but not your variances

Two SEM packages for R

Two SEM packages for R

- lavaan (Yves Rosseel, Daniel Oberski, Jarrett Byrnes, Leonard Vanbrabant, Victoria Savalei, Ed Merkle, Michael Hallquist, Mijke Rhemtulla, Myrsini Katsikatsou , Mariska Barendse, Michael Chow, Terrence D. Jorgensen)
  - http://www.lavaan.org
  - Based on maximum likelihood estimation
- Assumptions behind F_{ml}
  - Multivariate normality
  - No missing da in calculations of the observed covariance matix
  - No redundant variables
  - ‘large’ sample size

Two SEM packages for R

- lavaan syntax:
  - Operators

formula type	operator	mnemonic
latent variable definition	=~	is measured by
regression	~	is regressed on
(residual) (co)variance	~~	is correlated with
intercept	~ 1	intercept

Two SEM packages for R

- piecewiseSEM (by John Lefcheck)
  - http://github.comjslefche/

- What's different compared to lavaan
  - Variance-covariance matrices estimated **for each endogenous variable**
  - Incorporates various solutions (Poission, Gamma, etc.)
  - Translates the complete SEM into a list of structural equations
  - No non-recursive models, no latent or composite variables

- Depending on your data, piecewiseSEM might be the better solution

Two SEM packages for R

- piecewiseSEM (by John Lefcheck)
  - http://github.comjslefche/

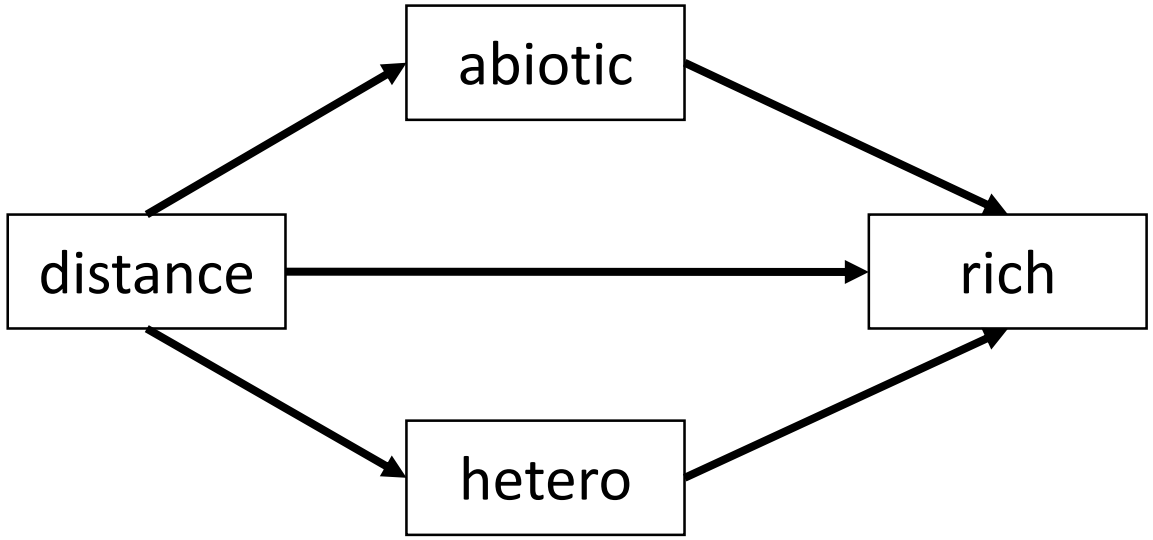
- What's different compared to lavaan
  - Variance-covariance matrices estimated **for each endogenous variable**
  - Incorporates various solutions (Poission, Gamma, etc.)
  - Translates the complete SEM into a list of structural equations
  - No non-recursive models, no latent or composite variables

- Depending on your data, piecewiseSEM might be the better solution

Let’s fool around in R

Fit this model

1. Fit this model!
2. Fill in Standardized Coefficients and R^2 for this model
3. Calculate summed direct and indirect effects of distance on richness



What we did not cover

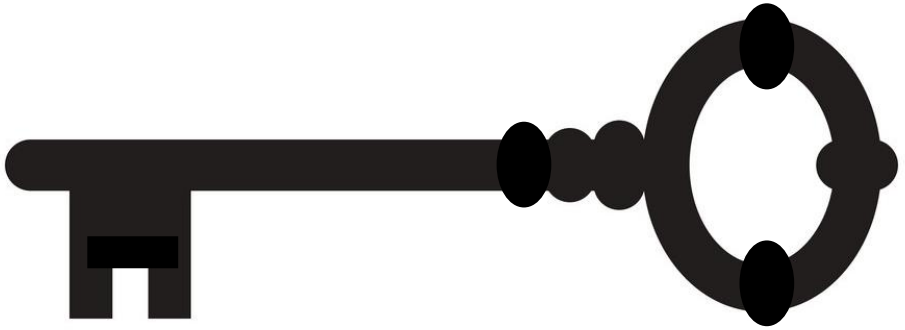
What we did not cover

… but what is also crucial for SEM

- Latent variables
- Composite variables
- Multigroup models
- piecewiseSEM

- A bunch of other stuff!!

…but our time is up and you are cramming your head since 10 a.m.!



Thank you…

Questions?
Suggestions?

Did this meet your expectations?