Basic Fundamentals of Structural Equation Modeling

Jim Grace
U.S. Geological Survey
Lafayette, Louisiana USA

USGS
science for a changing world

Outline for First Section of Materials

I. Introduction and Background:
- A. SEM Essentials
- B. Doing SEM in R
- C. Model Critiquing/Evaluation

II. Basic Elements of Modeling
- D. Overview of the Modeling Process
- E. Direct and Indirect Effects
- F. SEM versus Multiple Regression
- G. Causal Modeling Principles Revisited
- H. SEM versus ANOVA and ANCOVA

I. Introduction and Background

A. SEM Essentials

Part 1: Summary Points
Part 2: Anatomy of SE Models
Part 3: Model Specifications
Part 4: Estimation

3

A. SEM Essentials

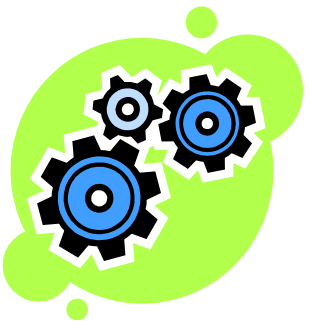
Part 1: Summary Points

4

1. SEM is a scientific framework for building and evaluating hypotheses about cause-effect connections in systems.

we use statistical and mathematical tools

within the SEM framework



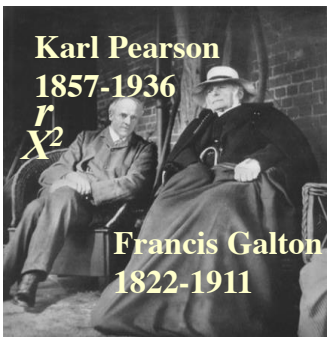
to build causal scientific understanding about the multiple processes operating in systems

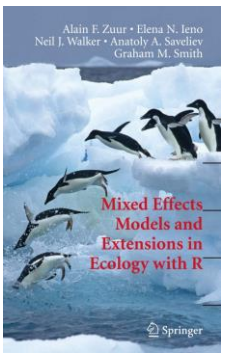
5

2. SEM is based on the fundamental notion that abstracting systems as causal, probabilistic networks is an efficient and effective way to understand the interrelations among their properties.

6

3. It is not generally appreciated that classical statistical analysis is not designed for the study of causal relations in systems.





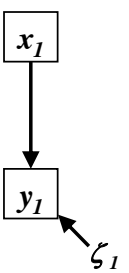
7

4. SEM is a form of graphical modeling.

equational form

$$y_1 = f(x_1)$$

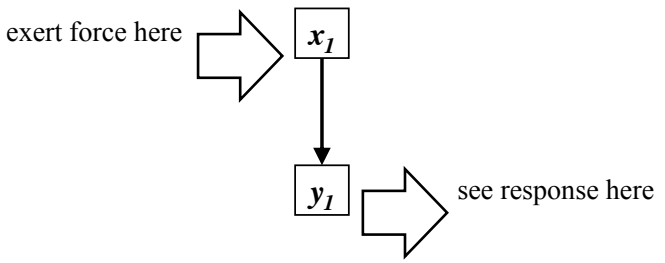
graphical form



8

5. A structural equation is one that attempts to estimate a causal effect.

Our concept of causation is that *A* is a cause of *B* if manipulation of *A* leads to a response in *B*.



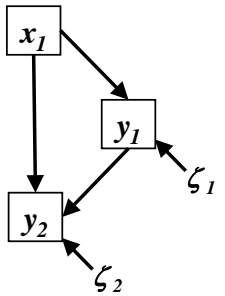
9

6. In practice, we describe SEM as a method for studying causal hypotheses.

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

10

7. Structural equation models typically involve multiple equations and represent network hypotheses.

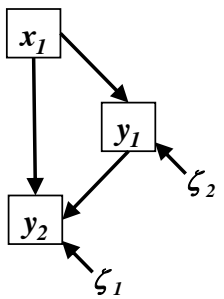


Network models require that *y*s can depend on *y*s.

$$y = f(x, y).$$

11

8. There is at least one underlying equation for each dependent variable in the network.

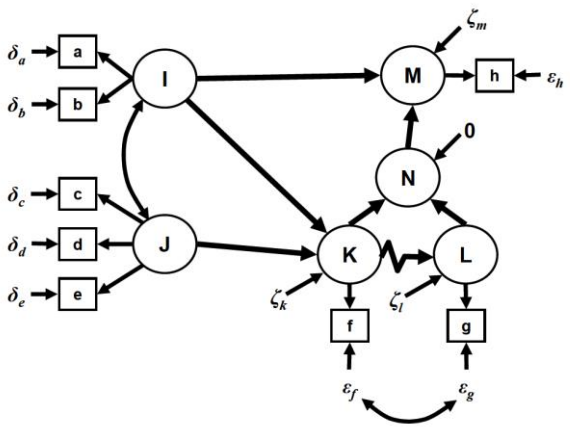


Corresponding Equations:

$$y_1 = f(x_1)$$
$$y_2 = f(x_1, y_1)$$

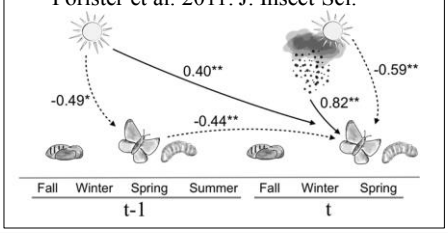
12

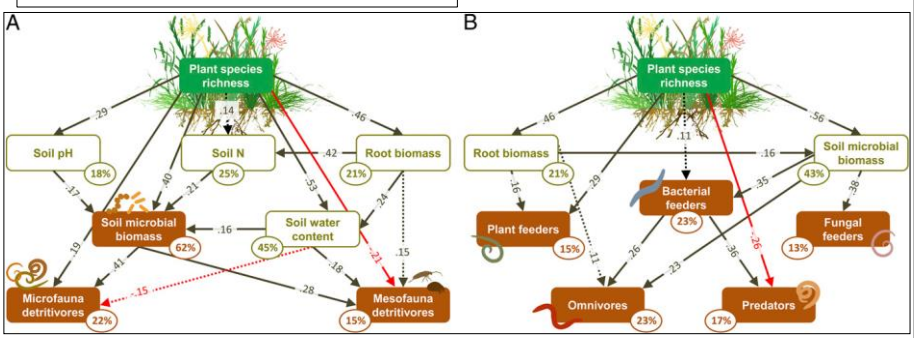
9. SEM is meant to be very flexible and capable of allowing the specification of complex hypotheses.



10. SEM presentations can also be flexible.

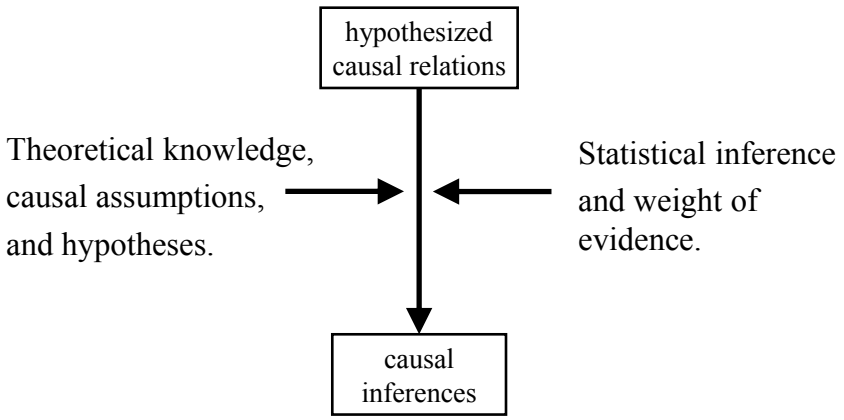
Forister et al. 2011. J. Insect Sci.





Eisenhauer et al. 2013. Proc. Nat. Acad. Sci.

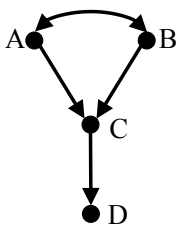
11. Investigation of causal relations requires theoretical knowledge.



No causes in, no causal estimates out."

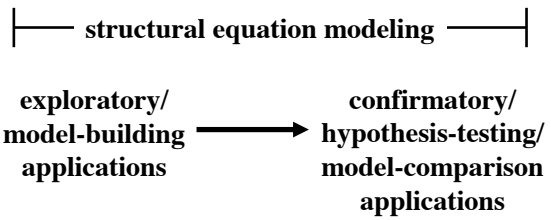
15

12. SE models include some causal assumptions, but usually also imply some <u>testable implications</u>.



All correlations must be considered for their causal implications.

13. SEM seeks to progress knowledge through sequential learning.



17

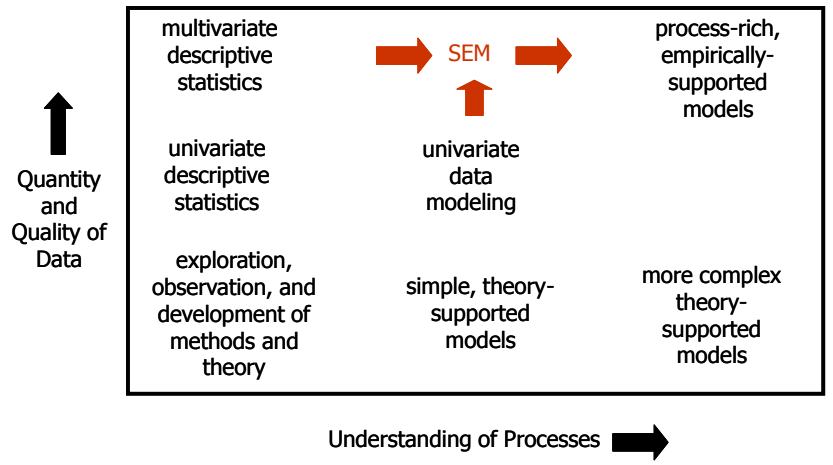
14. SEM is one of the few applications of statistical inference where the results of estimation are frequently "you have the wrong model!"

= learning!

This feedback comes from the unique feature that in SEM we compare patterns in the data to those implied by the model

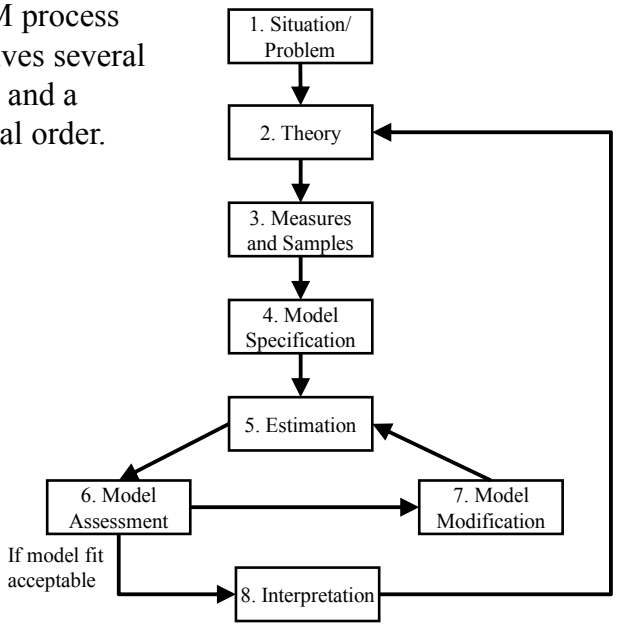
18

15. SEM fills a particular role in the scientific enterprise.



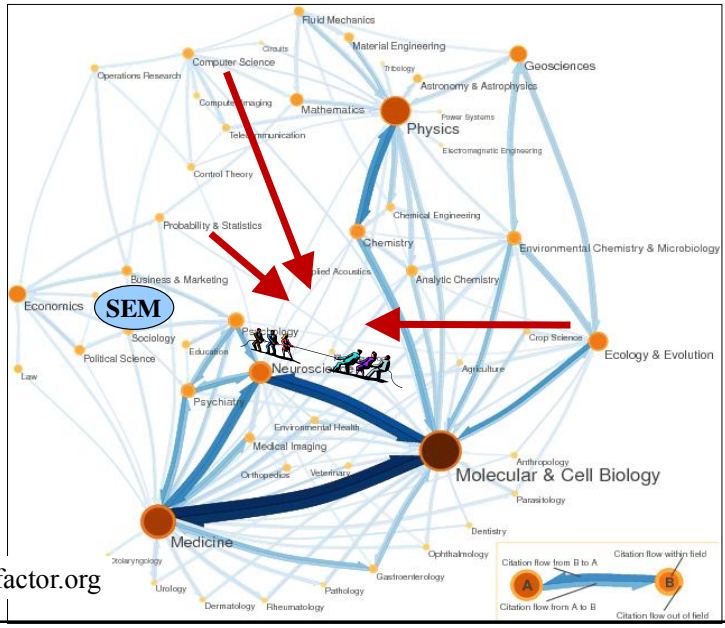
modified from Grace et al. 2009. Chapter 12, In: Real World Ecology. Springer Verlag 19

16. SEM process involves several steps and a logical order.



20

Where is the Methodology from and Where is it Headed?



eigenfactor.org

Textbooks dealing with SEM

Blunch (2013) Introduction to Structural Equation Modeling Using IBM SPSS Statistics and Amos. Sage Press.

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

Shipley (2000) Cause and Correlation in Biology. Cambridge Univ. Press.

Kline (2014) Principles and Practice of Structural Equation Modeling. (3nd Edition) Guilford Press.

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

Lee (2007) Structural Equation Modeling: A Bayesian Approach. John Wiley and Sons.

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

22

Key Papers Relevant to Ecological Applications

Grace, J.B., Anderson, T.M., Olff, H., and Scheiner, S.M. 2010. On the specification of structural equation models for ecological systems. Ecological Monographs 80:67-87. (http://www.esajournals.org/doi/abs/10.1890/09-0464.1)

Grace, J.B., Schoolmaster, D.R. Jr., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M., and Schweiger, E.W. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3(8): article 73 (http://www.esajournals.org/doi/abs/10.1890/ES12-00048.1)

also, see www.structuralequations.org

23

Discussion

24

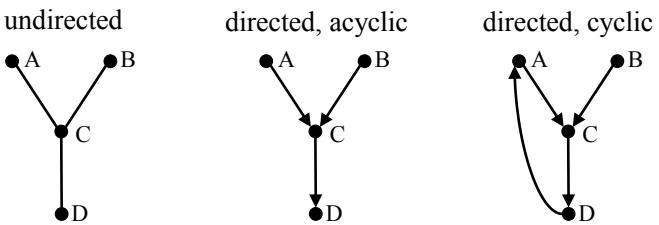
A. SEM Essentials

Part 2: Anatomy of SE Models

25

1. SEM falls within the broad field of graphical modeling, which is a fusion of probability theory and graph theory.

Types of graphs:



Graphical modeling (derived from graph theory), refers to
"nodes",
"edges", and
"pathways" along the edges.

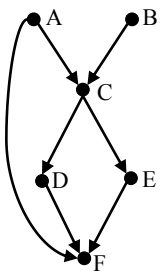
We can analyze graphs for their causal logic.
We replace the nodes in graphs with variables when we develop SE models.

26

2. Graphical modeling uses a different but overlapping terminology compared to traditional SEM literature.

Familial Relations:
parents vs. ***children***
ancestors vs. ***descendents***

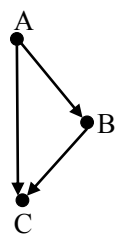
Topological Relations:
collider nodes (C and F)
forks (C, D, E)
root nodes (A, B)
terminal nodes (F)



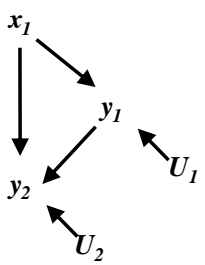
27

3. We try to follow certain nomenclatural distinctions when representing graphs, causal diagrams, and models.

graph

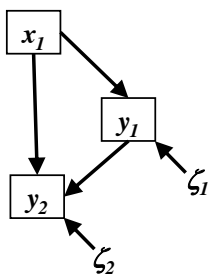


causal diagram*



*U*s are unspecified causal forces

causal statistical model



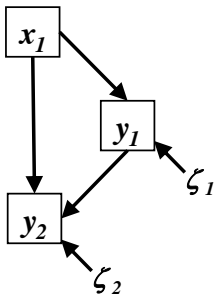
ζ (zetas) are residuals

*A **causal diagram** is a graphical tool that enables the exploration of possible causal relationships between variables in a causal model.

28

4. Models are conceptualized in nonparametric, nonlinear terms (left), but commonly represented in simple linear form (right).

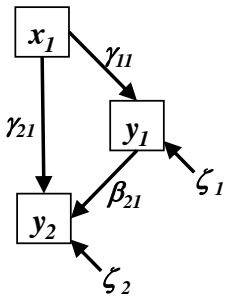
General (nonparametric) model representation



$y_1 = f(x_1)$
$y_2 = f(x_1, y_1)$

$\mathbf{Y} = f(\mathbf{X}, \mathbf{Y})$

Classical, simple linear model representation

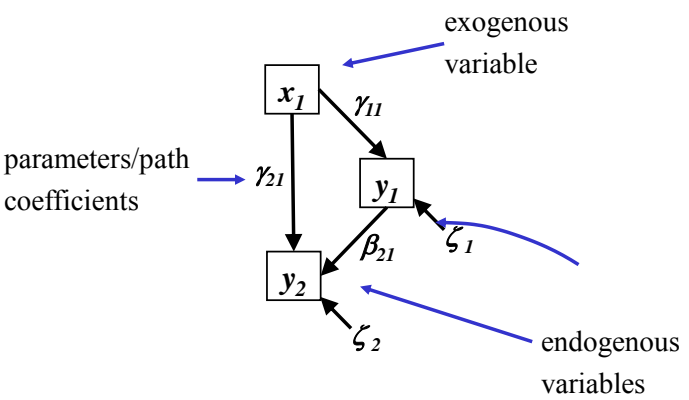


$y_1 = \alpha_1 + \gamma_{11}x_1 + \zeta_1$
$y_2 = \alpha_2 + \gamma_{21}x_1 + \beta_{21}y_1 + \zeta_2$

$\mathbf{Y} = \alpha + \Gamma\mathbf{X} + \mathbf{B}\mathbf{Y} + \zeta$

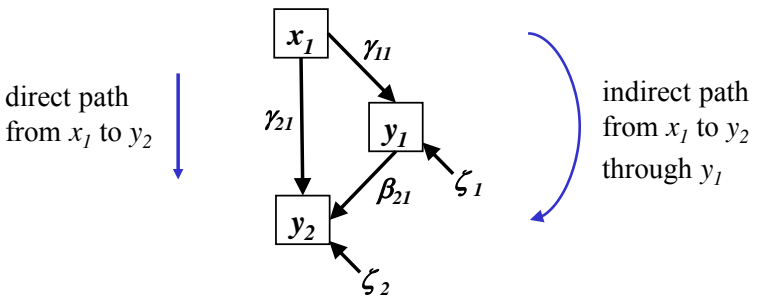
29

5. Variables are classified as exogenous or endogenous for convenience.



30

6. A key interest is partitioning relationships among pathways.



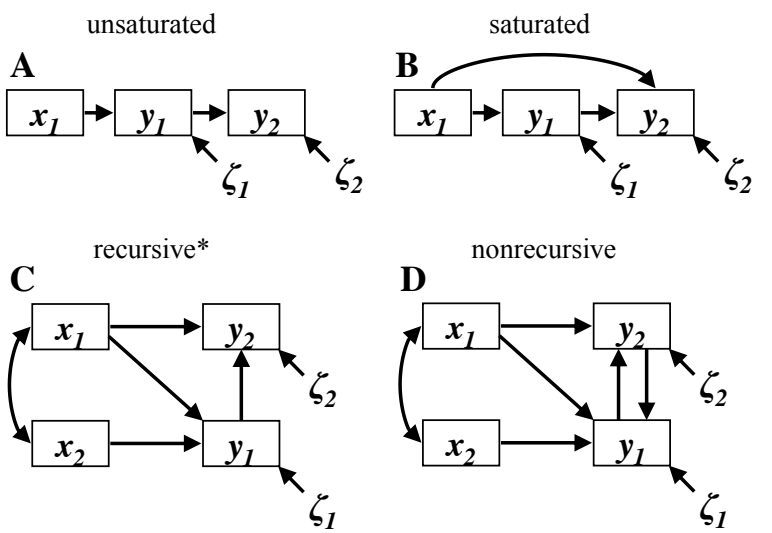
direct path strength is γ_{21} in this case.

indirect path strength is γ_{11} times β_{21} in this case.

total effect is *combination of direct and indirect effects*.

31

7. Several model architectures are possible.



32

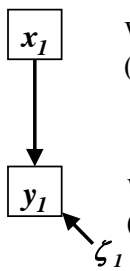
*A **recursive** sequence is where each term is defined from earlier terms in the sequence.

A. SEM Essentials

Part 3: Model Specifications

33

1. Two elements of all probabilistic equations are:
   (a) response forms and (b) linkage types.

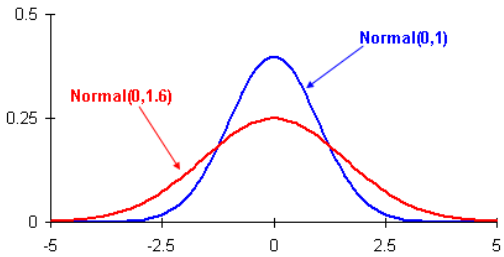


What is the form of the effect of x_1 on y_1?
(basic assumption is "simple and linear")

What is the form of the responses we expect?
(basic assumption is "Gaussian" or "normal")

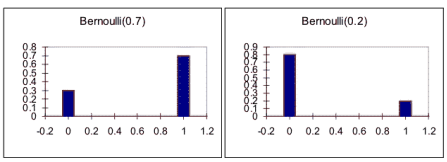
34

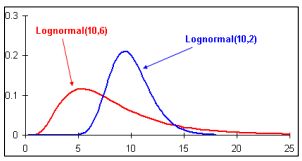
2. It is often assumed for convenience that responses follow a normal distribution.

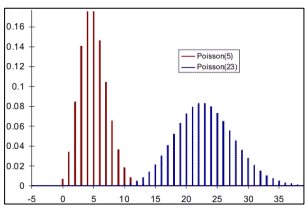


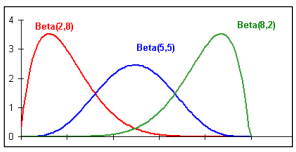
35

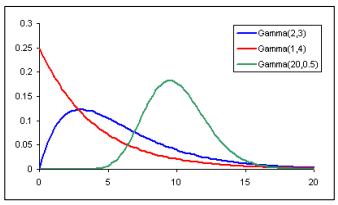
3. Real-world responses can be of various types.

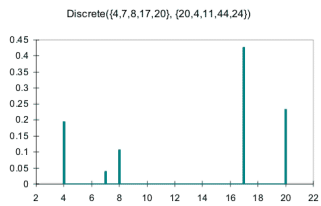












36

4. We also often assume linkages forms are linear.

Simple, linear relationship with linear equation

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

Polynomial, curvilinear equation with linear terms

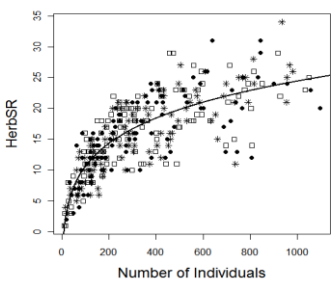
$$y_1 = \alpha_1 + \gamma_{11}x_1 + \gamma_{21}x_1^2 + \zeta_1$$

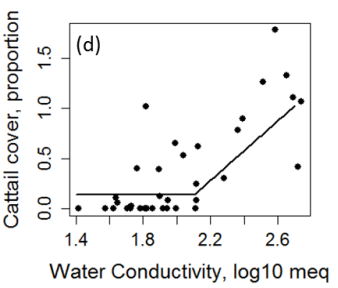
37

5. Some situations may require more complex linkage equations.

$y = a + bx^c$

$y = b_1 * x + b_2 * \text{step}(x - \psi) * (x - \psi)$





38

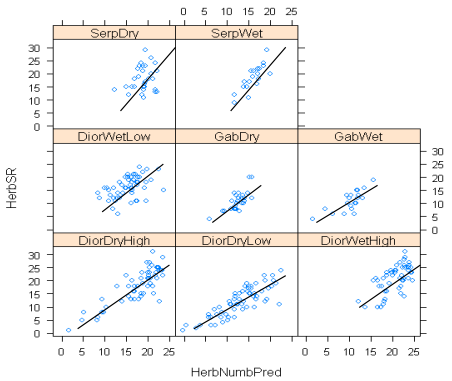
6. Equations may be hierarchical/multi-level.

$$y_{ig} = \beta_{0g} + \beta_{1g} x_{ig} + \varepsilon_{ig}$$

$$\beta_{0g} = \gamma_{00} + \gamma_{01} W_g + \gamma_{02} Z_g + u_{0g}$$

$$\beta_{1g} = \gamma_{10} + \gamma_{11} W_g + \gamma_{12} Z_g + u_{1g}$$

Parameters (e.g. β_{0g}, β_{1g}) in top equation are themselves functions of other variables.




Ecology, 92(1), 2011, pp. 108–120
© 2011 by the Ecological Society of America

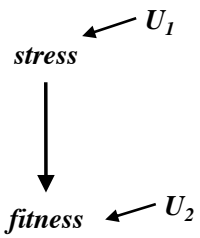
Local richness along gradients in the Siskiyou herb flora: R. H. Whittaker revisited

JAMES B. GRACE,[1] SUSAN HARRISON,[2,4] AND ELLEN I. DAMSCHEN[3]

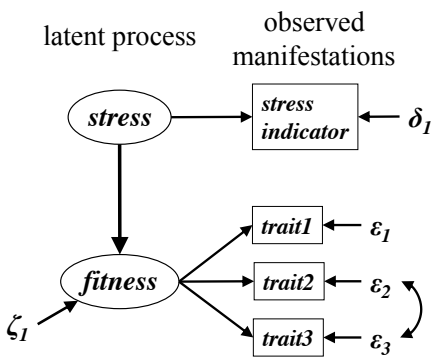
[1]U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Boulevard, Lafayette, Louisiana 70506 USA
[2]Department of Environmental Science and Policy, University of California, Davis, California 95616 USA
[3]Department of Zoology, Birge Hall, University of Wisconsin, Madison, Wisconsin 53706 USA

7. We may be missing some key variables of interest and wish to include latent variables in our models.

Causal Diagram

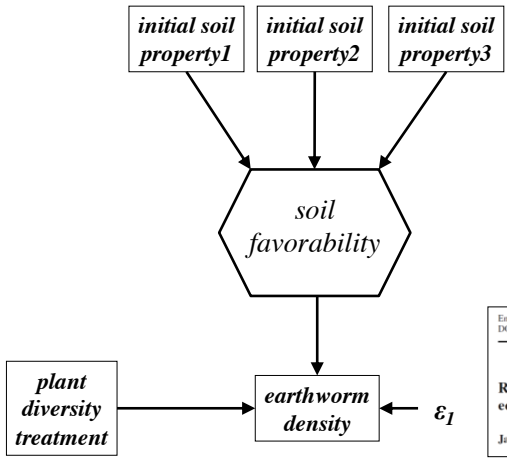


SE model



40

8. We may be interested in summarizing collective effects of groups of variables using composites.



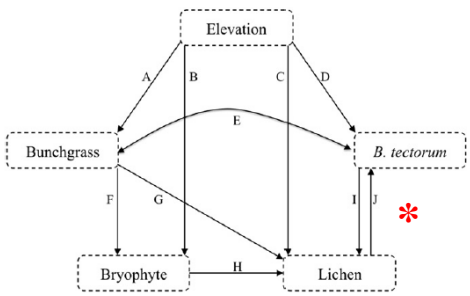
Environ Ecol Stat
DOI 10.1007/s10651-007-0047-7

Representing general theoretical concepts in structural equation models: the role of composite variables

James B. Grace · Kenneth A. Bollen

41

9. Our model may include reciprocal interactions or loops.



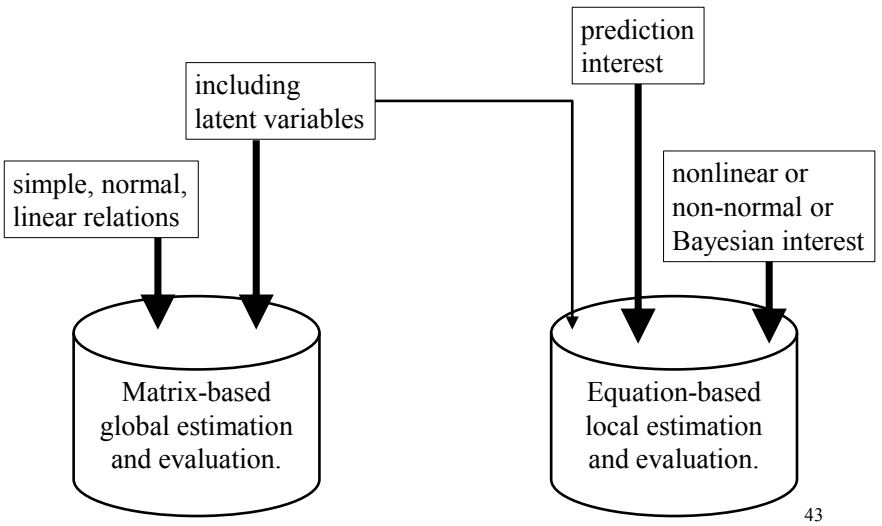
Journal of Arid Environments 84 (2012) 96–104
Contents lists available at SciVerse ScienceDirect
Journal of Arid Environments
journal homepage: www.elsevier.com/locate/jaridenv

Controls of biological soil crust cover and composition shift with succession in sagebrush shrub-steppe

E. Dettweiler-Robinson[a,*], J.D. Bakker[a], J.B. Grace[b]

42

10. The estimation approach and even software needed depends on model specification.



43

11. What software should I use?

Commercial SEM Packages (partial list)

Amos - most user-friendly current software and manual.

Mplus - favorite with advanced users.

LISREL - original software. Still being constantly updated. Lots of advanced features.

Free Packages

Lavaan SEM packages in R.

R base and various packages (for local estimation);

WinBUGS; Bayesian packages in R (local estimation and LVs).

Note: Local estimation of observed-variable models can be performed in any conventional statistical package (discussed later).

44

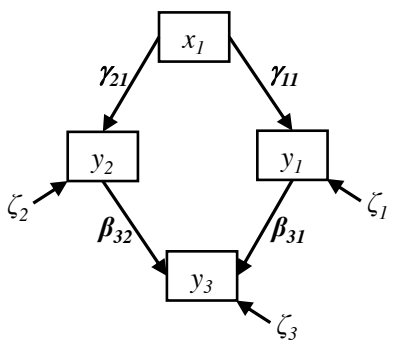
A. SEM Essentials

Part 4: Estimation

45

1. We can look at the estimation problem in a general way.

Consider a simple model where x_1 affects y_3 through two routes (via y_1 and y_2.



This model can be represented by three equations, one for each endogenous variable.

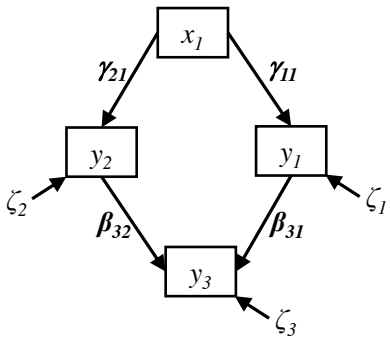
$$y_1 = \alpha_1 + \gamma_{11}x_1 + \zeta_1$$

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

$$y_3 = \alpha_3 + \beta_{31}y_1 + \beta_{32}y_2 + \zeta_3$$

46

2. A key test to be performed that involves the parameter estimates is the test of conditional independence.



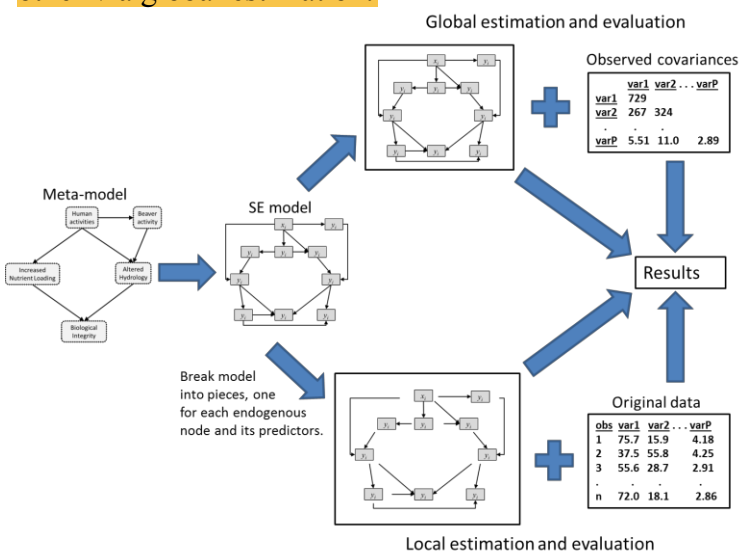
The model architecture implies;

$COV(x_1 , y_3) = \gamma_{11}*\beta_{31} + \gamma_{21}*\beta_{32}$

If this is not true, there is some additional process connecting x_1 and y_3.

47

3. There are two basic approaches to estimating the parameters in a model. One is via local "piecewise" estimation, the other via global estimation.



48

4. Local estimation involves estimating parameter values for each equation separately, then assembling the model as a collection of prediction equations.

$$y_1 = \alpha_1 + \gamma_{11}x_1 + \zeta_1$$
$$y_2 = \alpha_2 + \gamma_{21}x_1 + \zeta_2$$
$$y_3 = \alpha_3 + \beta_{31}y_1 + \beta_{32}y_2 + \zeta_3$$

49

5. Second approach to estimating the parameters is through "global" estimation methods.

- Conceptualize model as collection of vectors (of variables, intercepts, and errors) and matrices (regression parameters, error correlations).

$$y_1 = \alpha_1 + \gamma_{11}x_1 + \zeta_1$$
$$y_2 = \alpha_2 + \gamma_{21}x_1 + 0*y_1 + \zeta_2$$
$$y_3 = \alpha_3 + 0*x_1 + \beta_{31}y_1 + \beta_{32}y_2 + \zeta_3$$

$$Y = \alpha + \Gamma X + \beta Y + \zeta$$

Y = p x 1 vector of responses

α = p x 1 vector of intercepts

B = p x p coefficient matrix of ys on ys

Γ = p x q coefficient matrix of ys on xs

X = q x 1 vector of exogenous predictors

ζ = p x 1 vector of errors for the elements of y

Φ = cov (x) = q x q matrix of covariances among xs

Ψ = cov (ζ) = q x q matrix of covariances among errors

50

6. Global estimation uses matrix methods.

Pre-analysis step: Summarize raw data in variance-covariance matrices.

data

Row	x	y_1	y_2	y_3
1	40	3.5	1.04	51
2	25	4.0	0.48	31
3	15	2.6	0.95	71
4	23	4.3	1.19	64
5	24	4.0	1.30	68
.	.	.	.	.
n	15	3.8	0.69	40

variance/covariance matrix*

	x	y_1	y_2	y_3
x	1			
y_1	-0.35	1		
y_2	0.45	-0.44	1	
y_3	-0.30	0.33	-0.37	1
std dev	12.6	0.32	1.65	15.1
mean	25.6	0.69	4.56	49.2

*showing standardized matrix

plus other summary information

51

7. The basic problem in global estimation is to estimate parameters by comparing observed covariances to model-implied covariances.

compare

Observed Covariances

$$S = \begin{Bmatrix} 159 & & & \\ -1.4 & 0.10 & & \\ 9.36 & 0.23 & 2.72 & \\ -57.1 & 0.11 & -0.93 & 228 \end{Bmatrix}$$

Model-Implied Covariances

$$\Sigma = \begin{Bmatrix} \sigma_{11} & & & \\ \sigma_{12} & \sigma_{22} & & \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{13} \end{Bmatrix}$$

52

8. There is a "fundamental equation" for global estimation.

The fundamental hypothesis behind covariance-based SEM is

$$\Sigma = \Sigma(\Theta)$$

where:
Σ = population covariance matrix of observed variables,
Θ = vector of population parameter values for the model, and
$\Sigma(\Theta)$ = the covariance matrix written as a function of Θ

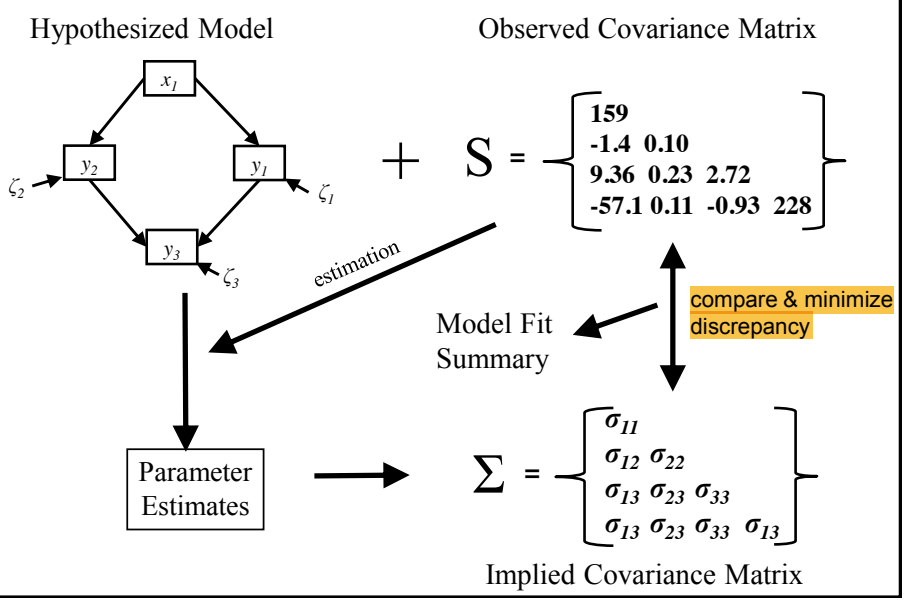
In practice, we are dealing with estimates.

$$\hat{\Sigma} = \Sigma(\hat{\Theta})$$

we aspire to $\hat{\Theta}$ such that $S = \hat{\Sigma}$

53

9. We can think of global estimation as a matching process.



10. The core of global estimation is the fitting function.

Fitting functions are designed to minimize model-data discrepancies.

Most common fitting function is based on the log likelihood ratio, which compares the likelihood for a given model to the likelihood of a model with perfect fit.

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

Note that when sample matrix and implied matrix are equal, terms 1 and 3 = 0 and terms 2 and 4 = 0. Thus,

perfect model fit yields a value of F_{ML} of 0.

55

11. Maximum likelihood estimators, such as F_{ML}, possess several important properties:

(1) asymptotically unbiased,
(2) scale invariant, and
(3) best estimators.

Assumptions:

(1) Σ-hat and S matrices are **positive definite** (i.e., that they do not have a singular determinant such as might arise from a negative variance estimate, an implied correlation greater than 1.0, or from one row of a matrix being a linear function of another), and

(2) data follow a multinormal distribution.

56

12. Parameter "identification" is a key, fundamental topic.

1. For model parameters to be estimated with unique values, they must be identified. As in linear algebra, we have a requirement that we need as many known pieces of information as we do unknown parameters.

2. Several factors can prevent identification, including:
   a. too many paths specified in model
   b. certain kinds of model specifications can make parameters unidentified
   c. multicollinearity
   d. combination of a complex model and a small sample

3. Good news is that most software checks for identification (in something called the information matrix) and lets you know when parameters are not identified (and which ones).

57

Discussion

58

B. Doing SEM in R

Part 1: Introduction to lavaan
Part 2: Local estimation of equations

59

The R environment permits several different ways to implement SEM..

Three primary implementations within the R environment:

(1) Global estimation using lavaan or sem,

(2) Local estimation using classical regression methods augumented by graph-theoretic analyses,

(3) Local estimation using Markov chain Monte Carlo methods associated with Bayesian implemention.

60

B. Doing SEM in R

Part 1: Introduction to lavaan

61

This tutorial briefly introduces the SEM R package known as **lavaan** ("latent variable analysis").

Url for the home page: http://lavaan.ugent.be/?q=node/2

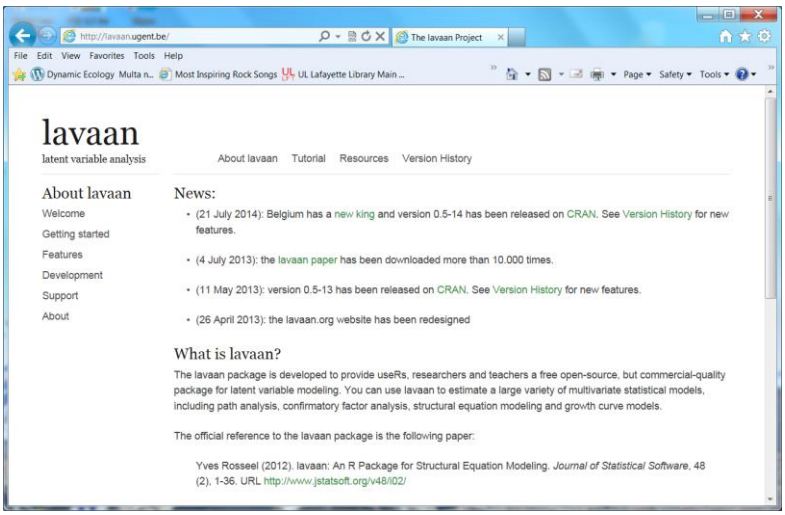
My very introductory tutorials etc. are at:
http://www.structuralequations.org

Yves Rosseel's latest (authoratative) tutorial is at:
http://lavaan.ugent.be/tutorial/tutorial.pdf

Jarrett Byrnes has some good material at:
http://jarrettbyrnes.info/ubc_sem/

62

Lavaan Home Page: http://lavaan.ugent.be/



63

1. Getting started in R: First read in data and load library.

R code: (we will **bold** command lines)

```
# Set working directory and load data

setwd("C:/Documents/LavaanTutorial")
```

```
# Read in data

k.dat<-read.csv("Keeley_rawdata_select4.csv")
```

```
# Load lavaan library

library(lavaan)
```

64

2. Getting started in R: Next, prepare data

```
# Create 4-variable example
# (pick 1st, 3rd,7th,8th vars)
x1 <- k.dat[,1]
y1 <- k.dat[,3]
y2 <- k.dat[,7]
y3 <- k.dat[,8]

# Create data set with those 4 variables
fourvars.dat <- data.frame(x1, y1, y2, y3)
```

65

3. Getting started in R: Then, examine data

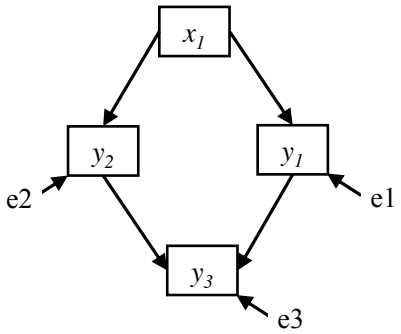
```
### Examine data and recode vars to same scale
summary(fourvars.dat)
```

```
> summary(fourvars.dat)
       x1              y1              y2               y3
 Min.   :37.04   Min.   :32.59   Min.   :0.05558   Min.   :15.00
 1st Qu.:39.46   1st Qu.:43.81   1st Qu.:0.48769   1st Qu.:37.00
 Median :51.77   Median :48.04   Median :0.63712   Median :50.00
 Mean   :49.23   Mean   :49.24   Mean   :0.69123   Mean   :49.23
 3rd Qu.:58.40   3rd Qu.:54.90   3rd Qu.:0.91468   3rd Qu.:62.00
 Max.   :60.72   Max.   :70.46   Max.   :1.53541   Max.   :85.00
>
```

```
## Recode vars to roughly same scale
x1 <- x1/100
y1 <- y1/100
y2 <- y2
y3 <- y3/100
fourvars.dat <- data.frame(x1, y1, y2, y3)
```

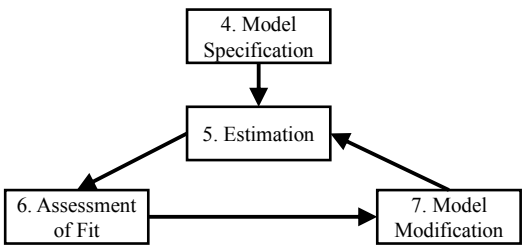
66

4. Choose a model to code.



67

5. Here we focus on the mechanics of steps 4 – 7.



68

6. Three steps for working in lavaan:

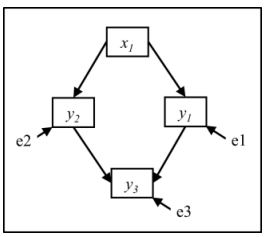
Step 1: Specify Model.

Step 2: Estimate aka "fit" Model.

Step 3: Extract Results (both estimates and assessment of fit).

69

7. Three steps for working in lavaan (continued):




```
# Step 1: Specify model

mod.1 <- 'y1 ~ x1
          y2 ~ x1
          y3 ~ y1 + y2'
```

```
# Step 2: Estimate model

mod.1.fit <- sem(mod.1, data=fourvars.dat)
```

```
# Step 3: Extract results

summary(mod1.fit)
```

70

Results Summary.

```
lavaan (0.5-12) converged normally after  31 iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic               17.729
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.000

                   Estimate  Std.err  Z-value  P(>|z|)
Regressions:
  y1 ~
    x1                0.400    0.081    4.911    0.000
  y2 ~
    x1                0.875    0.367    2.381    0.017
  y3 ~
    y1                0.935    0.171    5.475    0.000
    y2                0.129    0.041    3.121    0.002

Variances:
    y1                0.005    0.001
    y2                0.094    0.014
    y3                0.015    0.002
>
```

71

Results Summary: Closer Look.

convergence was normal

number of rows in data set

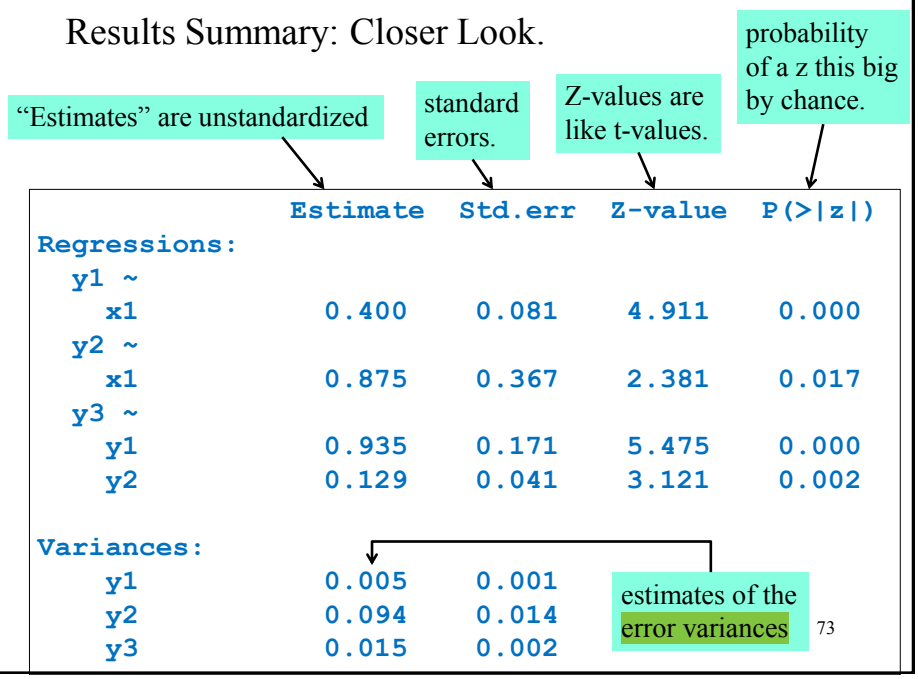
```
lavaan (0.5-12) converged normally after  31
iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic               17.729
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.000
```

default estimator is maximum likelihood

Chi-square
model df
p-value
(will discuss later)

72

Results Summary: Closer Look.



73

Additional Options for using lavaan will be introduced as we go along.

A brief overview can be found here:

www.structuralequations.com/resources/Lavaan_Syntax_Grace-BasicOnly.pdf

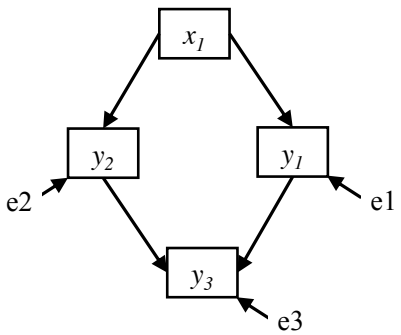
74

B. Doing SEM in R

Part 2: Local estimation of equations

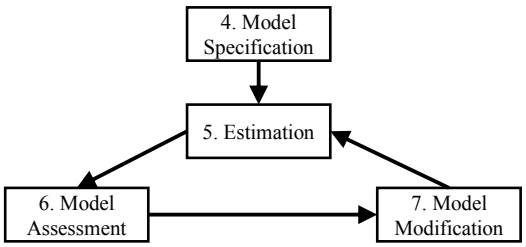
75

1. How would be evaluate this model using local estimation methods?



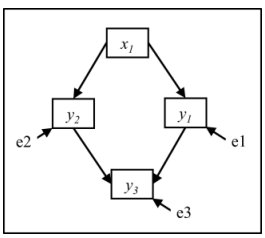
76

2. Again, we focus on the mechanics of steps 4 – 7.



77

3. Identify conditional independences as a way of thinking about specification alternatives.



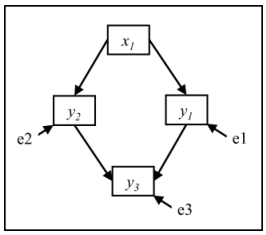
What are the conditional independence claims?

(1) $y_2 \perp y_1 \mid x_1$

(2) $y_3 \perp x_1 \mid y_1, y_2$

78

4. Specification and estimation of equations/submodels in R:

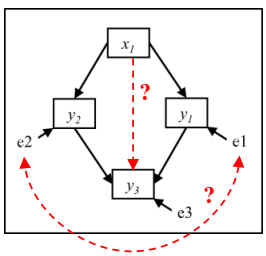


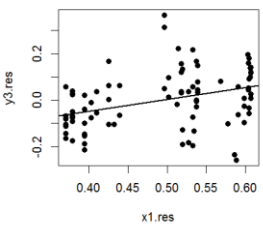

```
# Specification and estimation of submodels

y1.mod <- lm(y1 ~ x1, data=fourvars.dat)
y2.mod <- lm(y2 ~ x1, data=fourvars.dat)
y3.mod <- lm(y3 ~ y1 + y2, data=fourvars.dat)
```

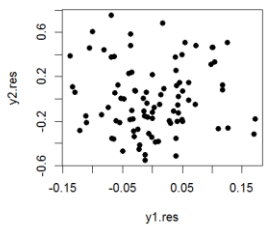
79

5. Model assessment.

```
# Capture residuals
x1.res <- x1
y1.res <- resid(y1.mod)
y2.res <- resid(y2.mod)
y3.res <- resid(y3.mod)
```

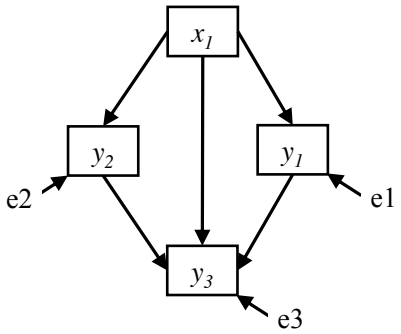






80

6. Model respecification.



81

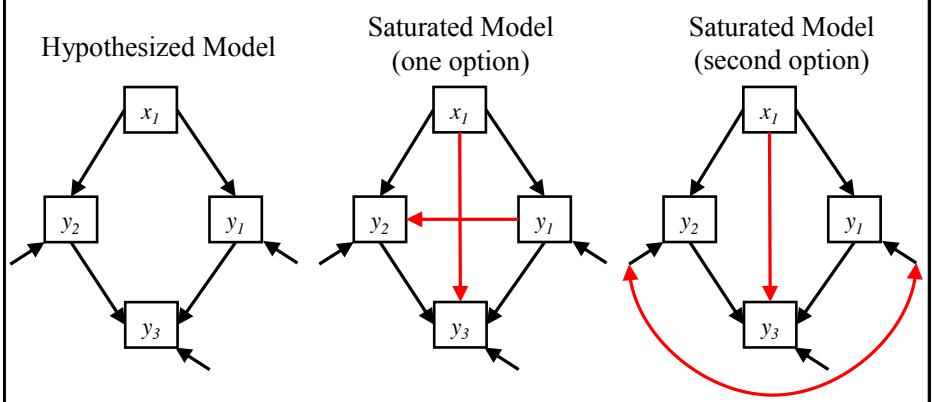
Discussion

82

C. Model Critiquing/Evaluation

83

1. The concept of Goodness of Fit involves comparing a model to one with perfect fit (a saturated model in the case of SEM).



- From an "analysis of covariance relationships" viewpoint, the question is how well our estimates reconstruct the observed covariances.
- From a scientific standpoint, the question is whether we are omitting any important processes (linkages) from our model.

1. Goodness of Fit (cont.).

- Note that in a saturated model, we will perfectly reconstruct our observed covariance matrix.
- Also note that while model fit is, in a certain sense, a comparison between models, we need to have adequate GOF in order to trust the estimates.
- For a non-network model, "fit" is evaluated simply in terms of data prediction and you rarely conclude **"wrong model"!**

85

2. The model chi-square is the most basic test statistic in global estimation.

The Model Chi-square Test.

Since the log likelihood ratio, F_{ML} follows a (chi-square) distribution, it is often used to calculate a model chi-square (X^2), as follows:

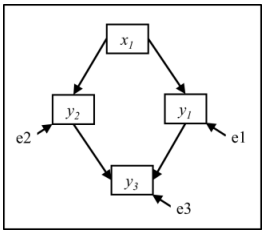
$$X^2 = n\text{-}1(F_{ML})$$

Here, n refers to the sample size, thus X^2 is a direct function of sample size.

The "single-degree-of-freedom chi-square criterion" = 3.84

86

Recall our lavaan example




```
# Step 1: Specify model

mod.1 <- 'y1 ~ x1
          y2 ~ x1
          y3 ~ y1 + y2'
```

```
# Step 2: Estimate model

mod.1.fit <- sem(mod.1, data=fourvars.dat)
```

```
# Step 3: Extract results

summary(mod1.fit)
```

87

Recall lavaan Results.

```
lavaan (0.5-12) converged normally after  31
iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic               17.729
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.000
```

```
# Request additional information

summary(mod1.fit, fit.measures=TRUE)
```

88

Other available model fit output: part 1

```
> summary(mod1.fit, fit.measures=TRUE)

lavaan (0.5-12) converged normally after  31 iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic               17.729
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.000

Model test baseline model:

  Minimum Function Test Statistic               80.731
  Degrees of freedom                                 6
  P-value                                        0.000

Full model versus baseline model:

  Comparative Fit Index (CFI)                    0.790
  Tucker-Lewis Index (TLI)                       0.369
```

89

Other available model fit output: part 2

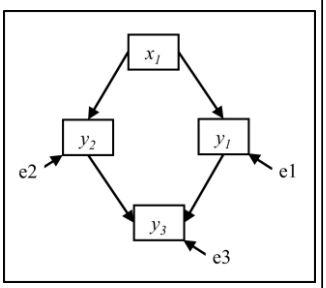
```
Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)                245.570
  Loglikelihood unrestricted model (H1)        254.435

  Number of free parameters                          7
  Akaike (AIC)                                -477.140
  Bayesian (BIC)                              -459.642
  Sample-size adjusted Bayesian (BIC)         -481.734

Root Mean Square Error of Approximation:

  RMSEA                                          0.296
  90 Percent Confidence Interval          0.180  0.429
  P-value RMSEA <= 0.05                          0.001

Standardized Root Mean Square Residual:

  SRMR                                           0.095
```

90

Modification Indices:

```
### Request Modification Indices

summary(mod1.fit, modindices=TRUE)
```

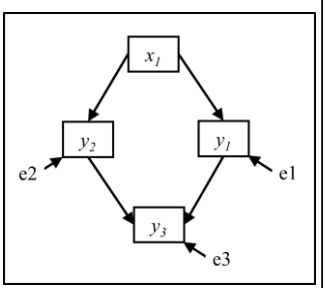
```
Modification Indices: (correlations)

   lhs op rhs      mi          epc
1   y1 ~~  y1   0.000        0.000
2   y1 ~~  y2   0.014        0.000
3   y1 ~~  y3  16.119       -0.008
4   y1 ~~  x1      NA           NA
5   y2 ~~  y2   0.000        0.000
6   y2 ~~  y3  16.119       -0.073
7   y2 ~~  x1  33.921  1085699.0
8   y3 ~~  y3   0.000        0.000
9   y3 ~~  x1  16.119        0.005
10  x1 ~~  x1   0.000        0.000
```



Here I have crossed-out non-sensical linkages (most of them!).
Looking for "mi" values larger than 3.84 (approximately).

Modification Indices (cont.):

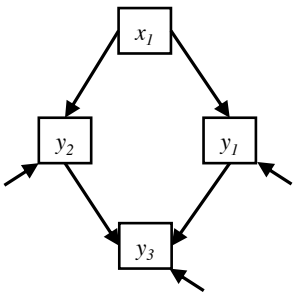
```
Modification Indices: (directed suggestions)

   lhs op rhs      mi          epc
11  y1  ~  y2   0.014        0.003
12  y1  ~  y3  10.215       -0.337
13  y1  ~  x1   0.000        0.000
14  y2  ~  y1   0.014        0.056
15  y2  ~  y3   2.107       -0.681
16  y2  ~  x1   0.000        0.000
17  y3  ~  y1   0.000        0.000
18  y3  ~  y2   0.000        0.000
19  y3  ~  x1  16.119        0.683
20  x1  ~  y1   0.000        0.000
21  x1  ~  y2   0.000        0.000
22  x1  ~  y3  12.197        0.264
```



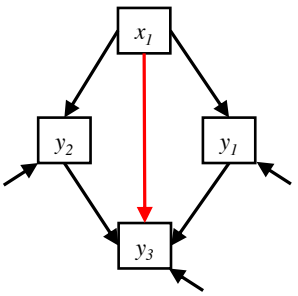
3. The way we test a model is by adding linkages.

Adding a single link.

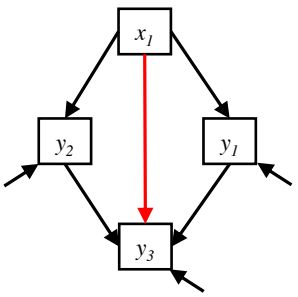
Model found to be inadequate



Suggested Improvement



93



Model #2

```
# Step 1: Specify model

mod2 <- 'y1 ~ x1
         y2 ~ x1
         y3 ~ y1 + y2 + x1'
```

```
# Step 2: Estimate model

mod.2.fit <- sem(mod.2, data=fourvars.dat)
```

```
# Step 3: Extract results

summary(mod2.fit, fit.measures=T, modindices=T)
```

94

Results (select)

```
lavaan (0.5-12) converged normally after  37 iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic                0.014
  Degrees of freedom                                 1
  P-value (Chi-square)                           0.906

Full model versus baseline model:

  Comparative Fit Index (CFI)                    1.000

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent Confidence Interval          0.000  0.119
  P-value RMSEA <= 0.05                          0.916
```

very close fit!

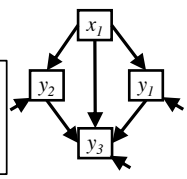
very close fit!

very close fit!

95

Full parameter output

```
# Full parameter output

summary(mod2.fit, rsq=T, standardized=T)
```

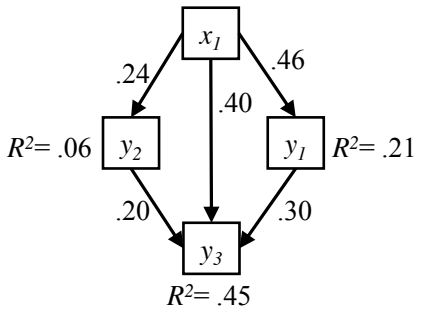



```
                   Estimate  Std.err  Z-value  P(>|z|)   Std.all
Regressions:
  y1 ~
    x1                0.400    0.081    4.911    0.000     0.460
  y2 ~
    x1                0.875    0.367    2.381    0.017     0.243
  y3 ~
    y1                0.593    0.173    3.423    0.001     0.301
    y2                0.093    0.038    2.422    0.015     0.195
    x1                0.682    0.154    4.419    0.000     0.399

Variances:
    y1                0.005    0.001                       0.789
    y2                0.094    0.014                       0.941
    y3                0.012    0.002                       0.550

R-Square:
    y1                0.211
    y2                0.059
    y3                0.450
```

additional chi-square tests could be done to validate that all included links are supported by the data.

96

Summary of selected model



Numerous possible options here. These are standardized coefficients.

97

Discussion

98

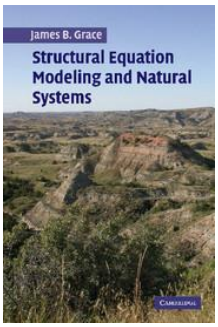
II. Basic Elements of Modeling

99

D. Overview of the Modeling Process

100

1. A workflow progression supports the SEM process.



Grace (2006)
Chapter 10

CONCEPTS & SYNTHESIS

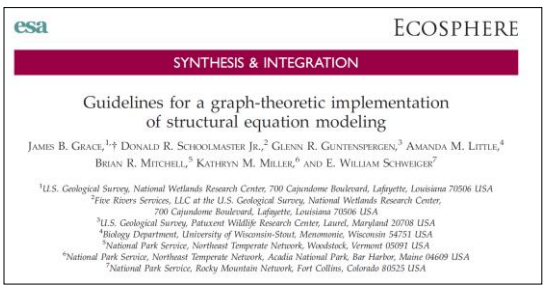
EMPHASIZING NEW IDEAS TO STIMULATE RESEARCH IN ECOLOGY

Ecological Monographs, 80(1), 2010, pp. 67–87
© 2010 by the Ecological Society of America

On the specification of structural equation models for ecological systems

JAMES B. GRACE,[1,4] T. MICHAEL ANDERSON,[2,5] HAN OLFF,[2] AND SAMUEL M. SCHEINER[3]

[1]*U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Boulevard, Lafayette, Louisiana 70506 USA*
[2]*Community and Conservation Ecology Group, Centre for Ecological and Evolutionary Studies, University of Groningen, P.O. Box 14, 9750 AA Haren, The Netherlands*
[3]*Division of Environmental Biology, National Science Foundation, Arlington, Virginia 22230 USA*



esa ECOSPHERE

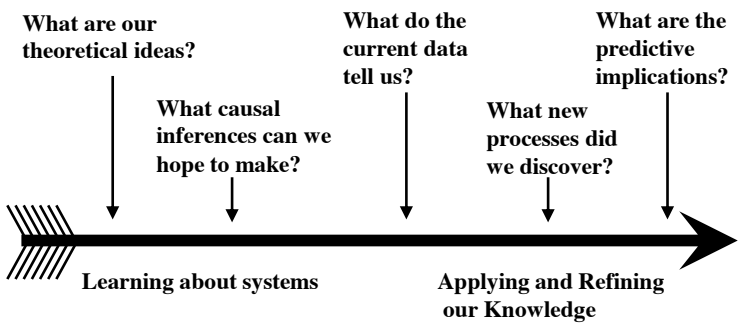
SYNTHESIS & INTEGRATION

Guidelines for a graph-theoretic implementation of structural equation modeling

JAMES B. GRACE,[1,†] DONALD R. SCHOOLMASTER JR.,[2] GLENN R. GUNTENSPERGEN,[3] AMANDA M. LITTLE,[4] BRIAN R. MITCHELL,[5] KATHRYN M. MILLER,[6] AND E. WILLIAM SCHWEIGER[7]

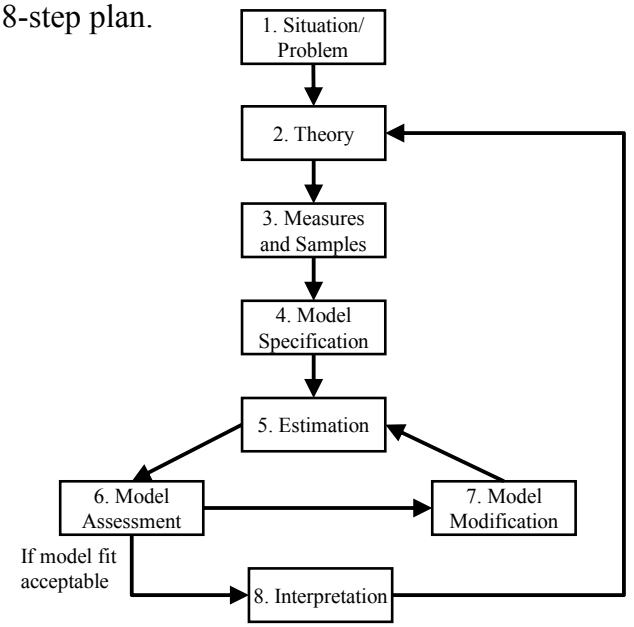
[1]*U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Boulevard, Lafayette, Louisiana 70506 USA*
[2]*Five Rivers Services, LLC at the U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Boulevard, Lafayette, Louisiana 70506 USA*
[3]*U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, Maryland 20708 USA*
[4]*Biology Department, University of Wisconsin-Stout, Menomonie, Wisconsin 54751 USA*
[5]*National Park Service, Northeast Temperate Network, Woodstock, Vermont 05091 USA*
[6]*National Park Service, Northeast Temperate Network, Acadia National Park, Bar Harbor, Maine 04609 USA*
[7]*National Park Service, Rocky Mountain Network, Fort Collins, Colorado 80525 USA*

2. We have been striving to expand the guidance.



102

3. The 8-step plan.



103

E. Direct and Indirect Effects

104

Situation: Post-Fire Recovery of Plant Communities in California Shrublands*

Analysis focus: understand post-fire recovery of plant species richness

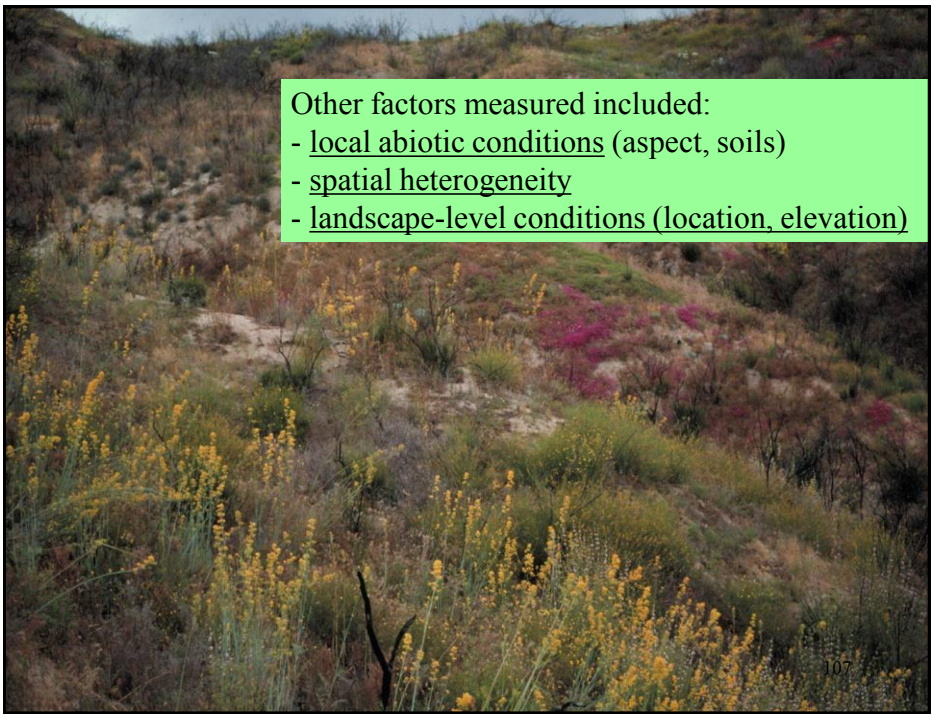
*Five year study of wildfires in Southern California in 1993. 90 plots (20 x 50m), (data from Jon Keeley et al.)

Measures and Samples

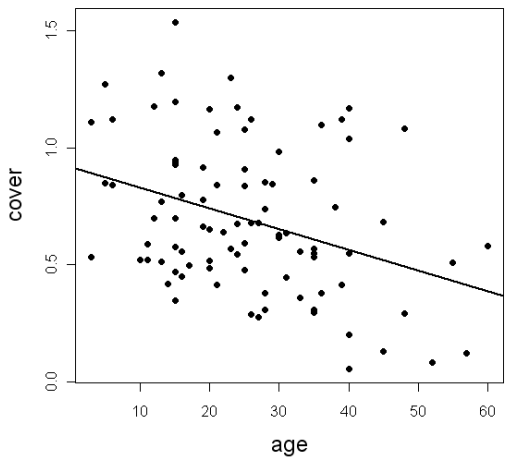
measured:
-vegetation cover
-species richness
-age of stand that burned
-fire severity

Examination of woody remains allowed for estimate of age of stand that burned as well as severity of the fires.

106

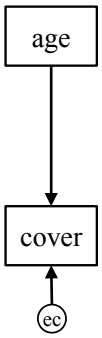


Observation: Post-fire Cover Declines with Age of Stand that Burned



108

Lavaan code for evaluating total (net) effect.




```
######### TEST OF MEDIATION ##########
# Net (total) effect of age on cover

mod.2 <- 'cover ~ age'

# Fit the model

mod.2.fit <- sem(mod.2, data=k.dat)

# Extract results

summary(mod.2.fit, stand=T, rsq=T)
```

requesting standardized estimates

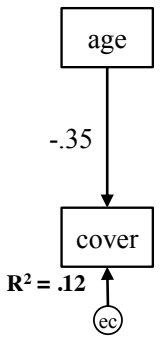
requesting r-square outputted

Lavaan results.

```
  Minimum Function Chi-square                   0.000
  Degrees of freedom                                0
  P-value                                       1.000
```

```
                   Est  Std.err  Z-value  P(>|z|)   Std.all
Regressions:
  cover ~
    age         -0.009    0.002   -3.549    0.000   -0.350

Variances:
    cover        0.087    0.013                      0.877

R-Square:

    cover        0.123
```

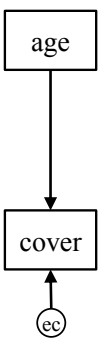
Graphical summary of net relationship.



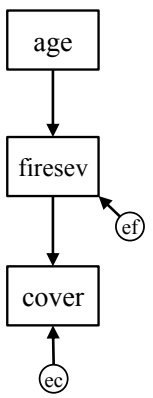
111

The Test of Mediation.

What mediates the causal effect of age on cover?



Could it be that older stands have more severe fires?

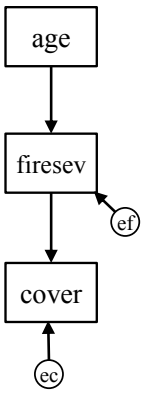


112

There are different degrees of mediation.

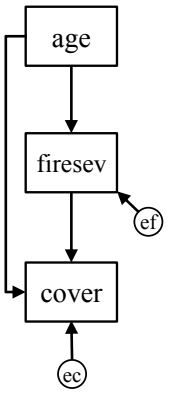
Som Possible Outcomes.

complete mediation



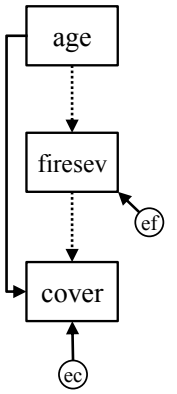

```
'cover ~ firesev
 firesev ~ age'
```

partial mediation




```
'cover ~ firesev + age
 firesev ~ age'
```

no mediation*




```
'cover ~ age'
```

Various ways of comparing models.

Option 1: Comparing model chi-squares

```
# Complete mediation

  Minimum Function Chi-square          3.297
  Degrees of freedom                       1
  P-value                              0.069


# Partial mediation

  Minimum Function Chi-square          0.000
  Degrees of freedom                       0
  P-value                              1.000
```

P-value greater than 0.05 provides classical frequentist support for the complete mediation hypothesis.

114

Various ways of comparing models (cont.).

Option 2: Using ANOVA function to compare models.

```
> anova(mod.4.fit, mod.3.fit)

Chi Square Difference Test

          Df  AIC    BIC    Chisq  Chidif Dfdif Pr>Chi
mod.4.fit  0 1069.4 1081.9  0.0000
mod.3.fit  1 1070.7 1080.7  3.2974 3.2974    1    0.069
```

Gives us same result as our direct comparison.

Calculating the magnitude of the standardized indirect effect.

Standardized total effect of age on cover:
= 0.45 x -0.44 = -.20

You can get the intercepts using the "meanstructure" option.

```
# a small digression: asking for the intercepts

mod.3a.fit <- sem(mod.3, meanstructure=T,
    data=k.dat)

summary(mod.3a.fit)
```

requesting intercepts

```
                 Est.  Std.err  Z-value  P(>|z|)
Regressions:
  cover ~
    firesev    -0.084    0.018   -4.611    0.000
  firesev ~
    age         0.060    0.012    4.832    0.000

Intercepts:
    cover       1.074    0.088   12.166    0.000
    firesev     3.039    0.351    8.647    0.000
```

Discussion

Model Comparison using Akaike Information

(similar guidance applies to Bayesian Information Criterion)

119

We can also use information theory to compare the models in a chosen set.

Another thing we can derive from the model log likelihood is the AIC (Akaike Information Criterion)

$$\text{AIC} = X^2 + 2q$$

where q = number of estimated parameters in model

note: we can view the first term in AIC as being a discrepancy and the second term a parsimony adjustment for model complexity.

120

The classic AIC is an asymptotic property.

AIC corrected for sample size is the AICc.

Because AIC is based on large sample theory, we are inclined to use the "corrected AIC", AICc, when the ratio of parameters to samples is large (i.e., information is low).

$$AICc = AIC + \left(\frac{2q(q+1)}{n-q-1} \right)$$

where q = number of estimated parameters in the model and
n = the number of samples

121

AIC comparison is approximate.

The AIC difference criteria

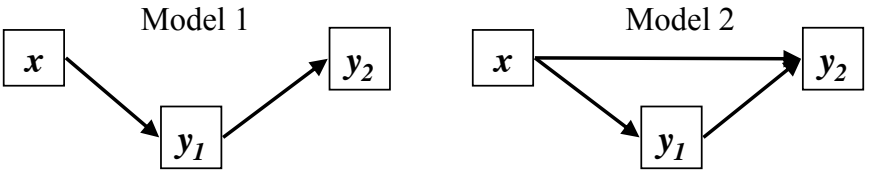
AIC diff	support for equivalency of models
0-2	substantial
4-7	weak
> 10	none

These apply to AICc as well.

Burnham, K.P. and Anderson, D.R. 2002. Model Selection and Multimodel Inference. Springer Verlag. (second edition), p 70.

122

An Illustration



remember,

$$AIC = \mathrm{X}^2 + 2q$$

$$AICc = AIC + R$$

$$R = \left(\frac{2q(q+1)}{n-q-1} \right)$$

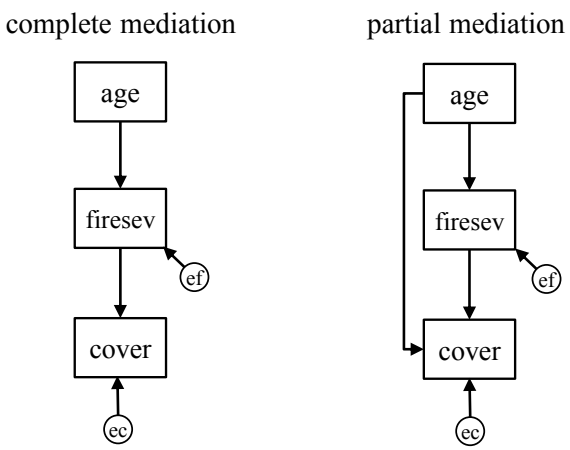
for n = 50 samples,

	X^2	q	AIC	R	AICc
Model 1	1.78	2	5.78	60/44=1.36	13.14
Model 2	0.00	3	6.00	84/43=1.95	13.98
difference	1.78	1	0.78		0.84

None of this suggests that Model 2 is better than Model 1 (we are actually looking for drop in AICc with added parameter).

123

Let's illustrate this further by revisiting our mediation model.



124

We can use AICc to compare the two models.

```
# need lavaan.modavg.R package
library(AICcmodavg)
source("lavaan.modavg.R")

aictab.lavaan(list(mod.3.fit, mod.4.fit),
    c("Full", "Partial"))
```

```
Model selection based on AICc :

        K    AICc  Delta_AICc AICcWt Cum.Wt
Partial 5  1069.66    0.00     0.64   0.64
Full    4  1070.82    1.16     0.36   1.00
```

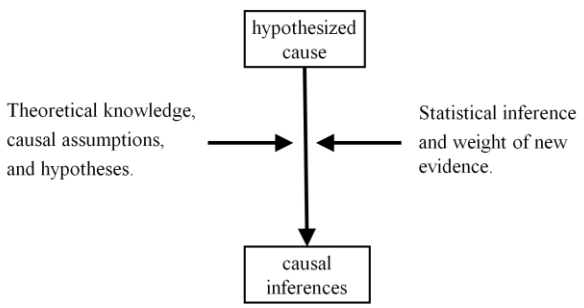
Delta_AICc < 2, leads to same conclusion, that Full Mediation Hypothesis is supported.

125

Final thoughts on "The Decision Problem".

1. There are a variety of approaches to weighing the statistical evidence regarding a link.

2. Remember, in SEM we also bring in a priori causal knowledge.



3. It is possible to formalize the combining of evidence using Bayes theorem, though often our combining is based on the scientist's expert judgement.

126

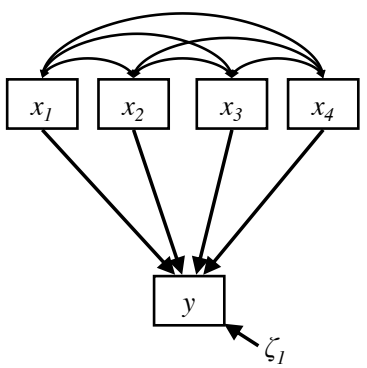
F. SEM versus Multiple Regression

127

What if we used a multiple regression approach?

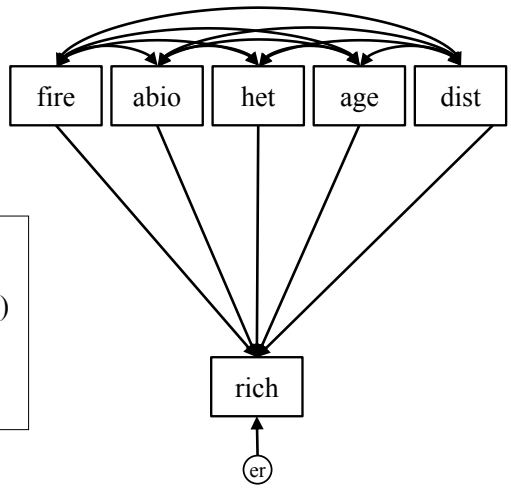
$$y = \alpha + \mathbf{B}\mathbf{x} + \varepsilon$$

graphical view of multiple regression



128

Multiple regression for post-fire species richness variations.



Possible predictors:
- fire severity (fire)
- abiotic favorability (abio)
- heterogeneity (het)
- age of burned stand (age)
- distance from coast (dist)

129

Multiple regression results for initial model.

```
# multiple regression model in lavaan
'rich ~ firesev + abiotic + hetero + age
        + distance'
```

```
                   Est.    Std.err  Z-value  P(>|z|)
Regressions:
  rich ~
    firesev      -0.167      0.077   -2.163    0.031
    abiotic       0.481      0.165    2.910    0.004
    hetero        0.350      0.104    3.359    0.001
    age          -0.091      0.101   -0.899    0.368
    distance      0.528      0.152    3.478    0.001
```

Pruning approach: remove least significant term (age) and rerun model.

130

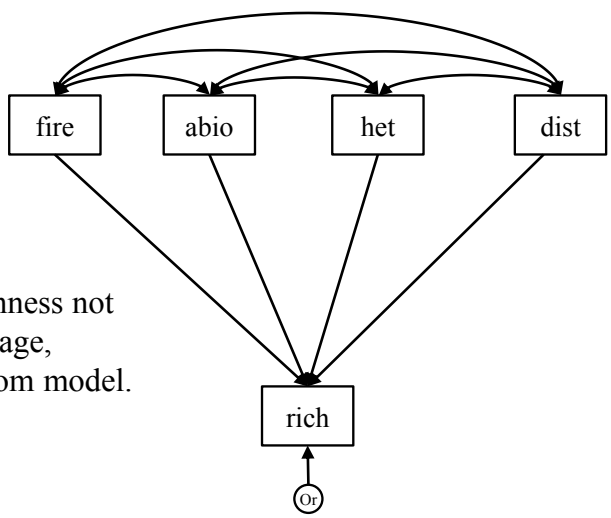
Prune #1 results.

	Est.	Std.err	Z-value	P(>|z|)
Regressions:				
rich ~				
firesev	-0.195	0.071	-2.764	0.006
abiotic	0.475	0.166	2.864	0.004
hetero	0.352	0.105	3.370	0.001
distance	0.550	0.150	3.657	0.000

All remaining pathways highly significant.

131

Pruned model.



Results imply richness not affected by stand age, which dropped from model.

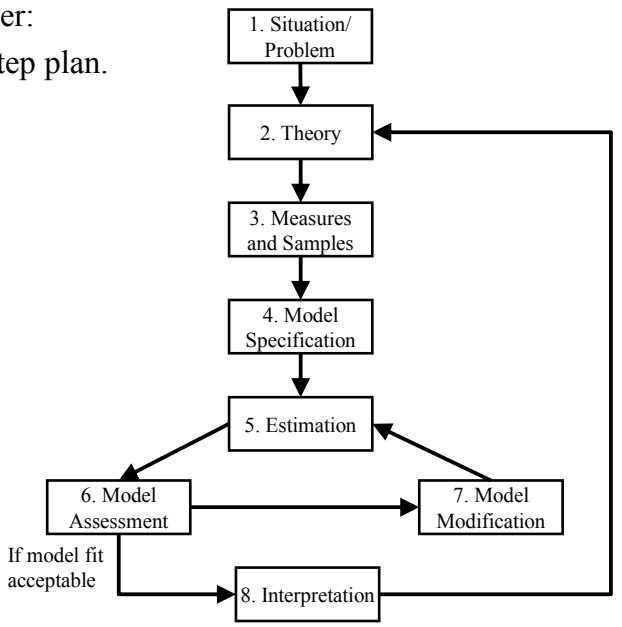
132

An SEM Approach to the Same Problem

133

Reminder:

The 8-step plan.

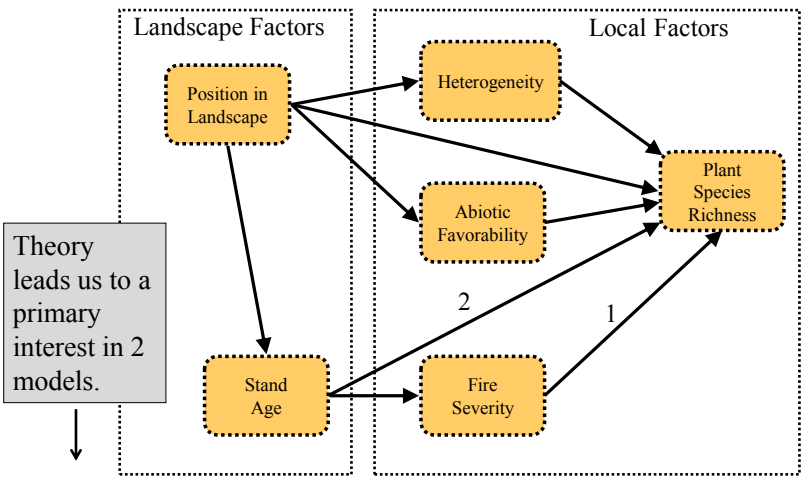


134

Step 1: Situation – heterogeneous fire in heterogeneous landscape

135

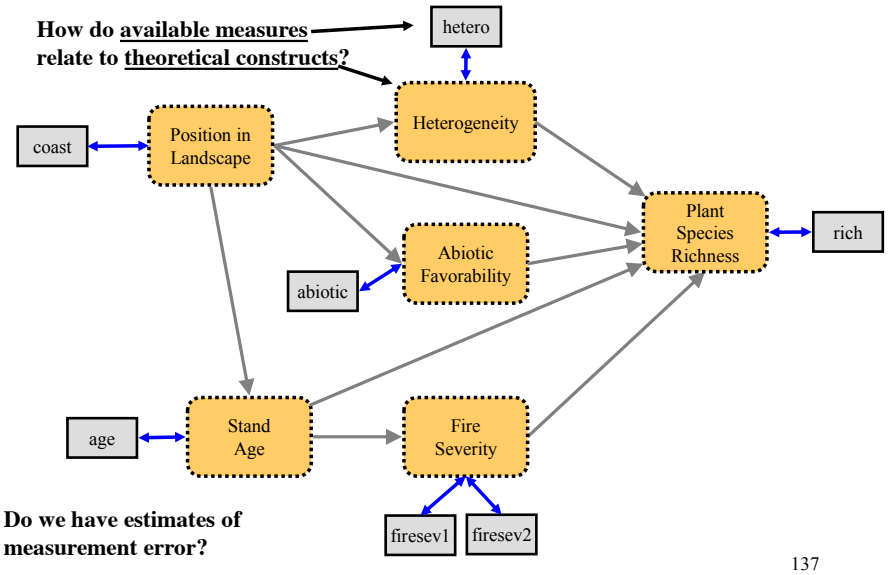
Step 2: Develop theory (see Grace and Keeley 2006)



Does seedbank decline with age?
- mod.1 only has path from fire severity to richness
- mod.2 also has path from stand age to richness

136

Step 3: Consider measures and samples.

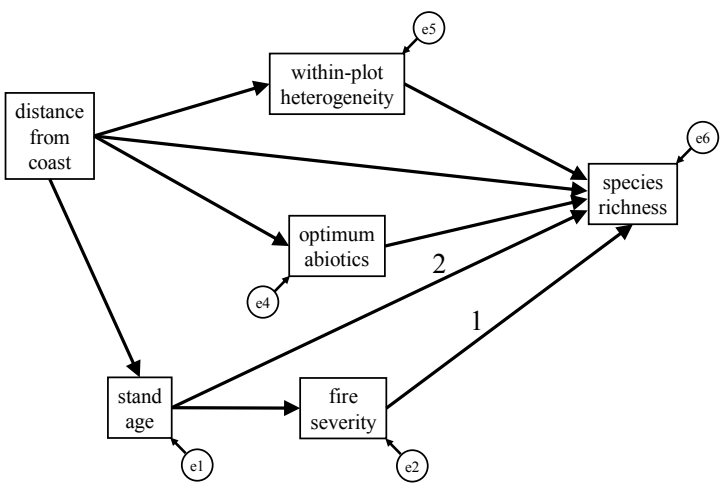


Do we have estimates of measurement error?

Are multiple indicators appropriate?

137

Step 4: Model specifications.



Does seedbank decline with age?

- mod.1 only has path from fire severity to richness
- mod.2 also has path from stand age to richness

138

Step 5: Estimation.

```
########### SEM FOR FIRE RECOVERY STUDY #########
# First run most comprehensive model "fire.2"
# and check for missing paths

# specify "fire.2"
fire.2 <- 'rich ~ abiotic + hetero + distance
                + firesev + age
           abiotic ~ distance
           hetero ~ distance
           age ~ distance
           firesev ~ age'
```

139

Step 6: Model assessment – the model chi-square test.

```
# Estimate model "fire.2"
fire.2.fit <- sem(fire.2, data=k.dat)

summary(fire.2.fit, fit.measures=T)
```

request wide array of fit measures

```
lavaan (0.5-15) converged normally after  22
iterations

  Number of observations                    90

  Estimator                                 ML
  Minimum Function Test Statistic        6.095
  Degrees of freedom                         6
  P-value (Chi-square)                   0.413
```

good fit means we are not missing any links.

Note: "Minimum Function Test Statistic" = "Model Chi-Square"

140

Step 6: Model assessment – comparative fit index.

```
Fitted model:
  Minimum Function Test Statistic          6.095
  Degrees of freedom                           6
  P-value (Chi-square)                     0.413

Baseline (null) model:
  Minimum Function Test Statistic        132.230
  Degrees of freedom                          15
  P-value                                  0.000

User model versus baseline model:

  Comparative Fit Index (CFI)              0.999
```

141

Step 6: Model assessment – RMSEA.

```
Root Mean Square Error of Approximation:

  RMSEA                                    0.013
  90 Percent Confidence Interval    0.000  0.138
  P-value RMSEA <= 0.05                    0.551
```

142

Step 6: Model assessment – Information measures.

```
Number of free parameters                        14
 Akaike (AIC)                                4075.767
 Bayesian (BIC)                              4110.764
 Sample-size adjusted Bayesian (BIC)         4066.579
```

143

Step 6: Model assessment – Modification indices.

```
# getting modification indices
modindices(fire.2.fit)
```

← requesting modification indices

```
# largest modification index

17   firesev ~~ distance 2.862
```

Mod. index well below 3.84 threshold, suggesting little room for improvement.

144

Step 6: Model assessment – residuals.

asking for standardized residuals

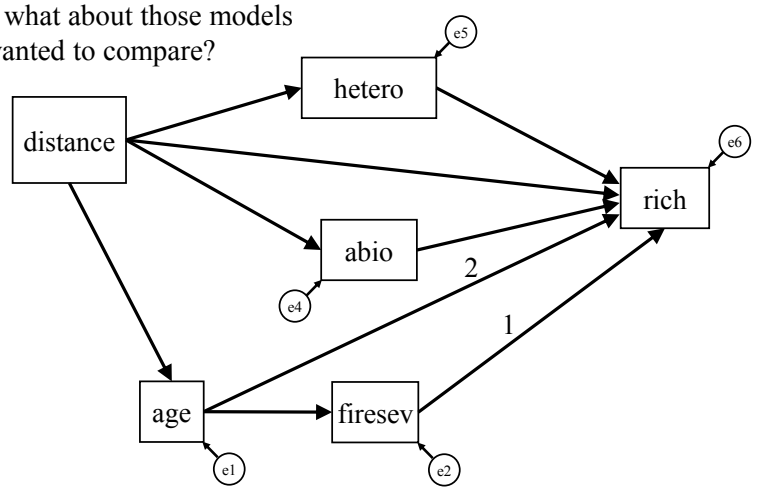
```
# getting residuals
resid(fire.2.fit, type="standardized")
```

```
          rich   abiotc hetero firesv age    distnc
rich       0.906
abiotic    1.274  0.000
hetero     0.714  1.292  0.001
firesev   -1.293 -1.475 -0.084  0.002
age       -0.033 -0.066  0.004  0.002     NA
distance   0.818  0.003  0.004 -1.615     NA  0.000
```

Largest residual is also between fire severity and distance from coast.

Step 6: Model assessment – model comparisons.

Now what about those models we wanted to compare?



- mod.1 only has path from fire severity to richness
- mod.2 also has path from stand age to richness

Step 6: Model assessment – model comparison with BICc (sample size adjusted BIC).

```
# Model fire.1
 Number of free parameters                      13
  Akaike (AIC)                                4074.572
  Bayesian (BIC)                              4107.069
  Sample-size adjusted Bayesian (BIC)         4066.040

# Model fire.2
 Number of free parameters                      14
  Akaike (AIC)                                4075.767
  Bayesian (BIC)                              4110.764
  Sample-size adjusted Bayesian (BIC)         4066.579

# Sample-size adjusted BIC difference           0.539
```

Results imply model with extra path (path 2) not justified.

147

Step 6: Model assessment – log-likelihood comparison

```
# Select results from output

> anova(fire.1.fit, fire.2.fit)

Chi Square Difference Test

           Df    AIC  BIC   Chisq
fire.2.fit  6 4075.8 4110.8 6.0946
fire.1.fit  7 4074.6 4107.1 6.8998
diff             1.2    3.7 0.8052
```

BIC and chi-square diff test both suggest model without path 2 is preferred.

148

Step 8: Interpretation – first looking at the parameter estimates.

	Est	Std.err	Z-value	P(>|z|)	Std.all
Regressions:					
rich ~					
abiotic	0.475	0.163	2.909	0.004	0.248
hetero	0.352	0.103	3.410	0.001	0.275
distance	0.550	0.150	3.663	0.000	0.330
firesev	-0.195	0.068	-2.874	0.004	-0.219
abiotic ~					
distance	0.400	0.081	4.911	0.000	0.460
hetero ~					
distance	0.450	0.129	3.498	0.000	0.346
age ~					
distance	-0.396	0.144	-2.747	0.006	-0.278
firesev ~					
age	0.597	0.124	4.832	0.000	0.454

149

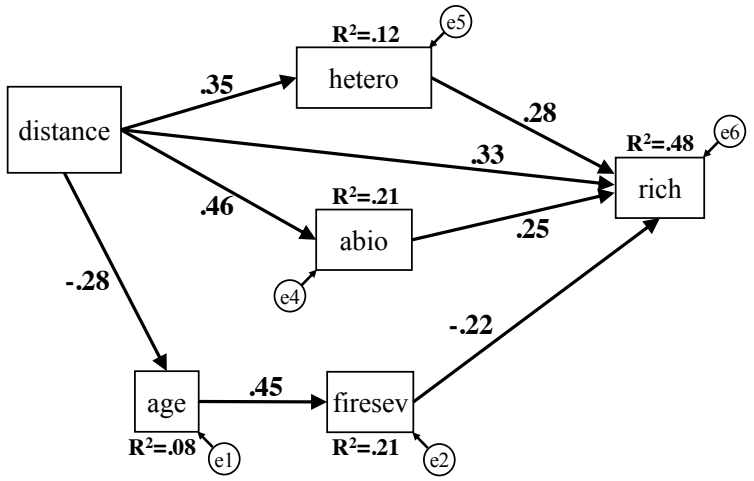
Step 8: Interpretation – variance explained.

R-Square:

rich	0.484
abiotic	0.211
hetero	0.120
age	0.077
firesev	0.206

150

Step 8: Interpretation – visualization of results.



Grace and Bollen 2006. Interpreting the results from regression and structural equation models. Bulletin Ecol. Soc. Amer. 86:283–295.

151

Step 8: Interpretation – extrapolations.

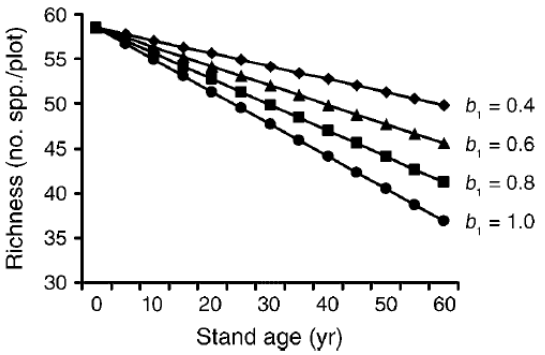
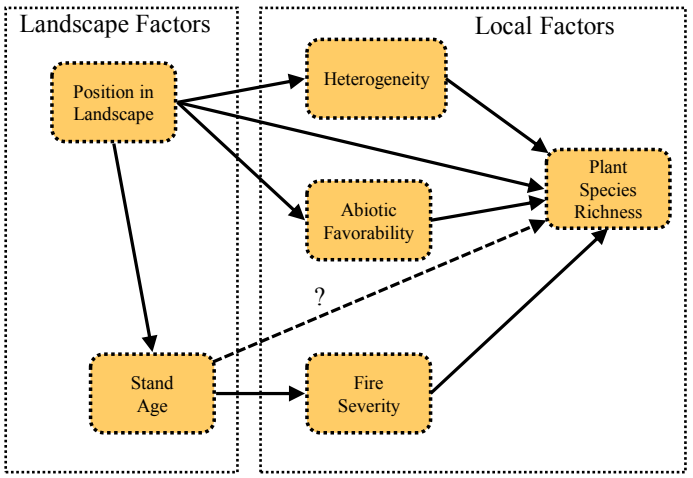


FIG. 5. Predicted sensitivity of richness to stand age at various levels of fire intensity (as a proportion of natural strength): $b_1 = 1.0$ represents the average fire severity observed in these wildfires, while values <1.0 represent expectations if fire intensity were lower, for example, through the use of prescribed burning techniques under more moderate weather and fuel conditions.

Grace and Keeley 2006 – Prescribed fire could be highly effective in protecting diversity loss.

Step 8: Interpretation – revising theory.



153

Discussion

154

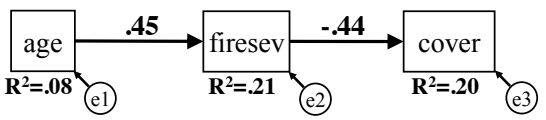
G. More Causal Modeling Principles

155

How we can throw around the word "causal"?

causal linkages versus causal estimates.

Let's consider this snippet from our fire-effects SEM.



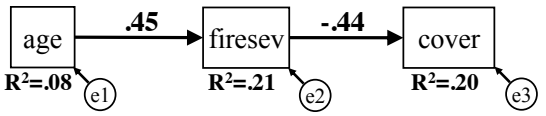
"How strongly can we defend the contention that this is a causal hypothesis (linkage are cause-effect)?"

- Time sequence is good:
  Age before fire → severity of fire → recovery from fire.

- Temporal logic is irreversible (arrows can't go the other way).

156

Causal linkages versus causal estimates (continued).



A second question, "How strongly can we defend the parameter estimates as unbiased causal predictions?"

This one is much harder.
A strong causal prediction would be that any individual plot in a hypothetical population of plots would increase in fire severity by 0.45 if we increased stand age by one unit.

157

There is skepticism about causal modeling from some quarters.

Caution is OK, *but* we should not be afraid to do science!

How far down the rabbit hole does this debate go?
Even in controlled experiments we cannot know for sure how individuals assigned to one treatment group would have behaved had they been in the other treatment group. For example, we might ask

"If Ms. X had been given the drug, would she have then survived?"

At a certain level, this counterfactual cannot be answered with certainty.

158

Bottom line: we should not be cavalier about the limits of interpretation.

Generally, with SEM we are evaluating causal hypotheses. Building confidence about causal conclusions requires persistent investigation.

Therefore, we might say,

"Our model results support the conclusion that X affects Y through Z."

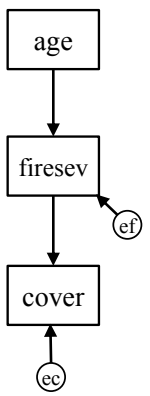
Best not to speak as if by doing SEM we automatically have causal conclusions.

159

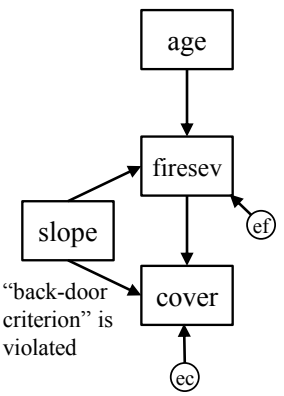
How can things go wrong?

Confounding – our ultimate concern.

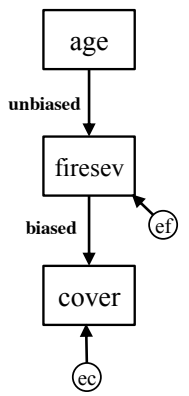
What if we hypothesized this model?



But this was the true model.



Our estimates for the initially hypothesized model would be . . .



160

Solution is to model the confounding factor.

H. SEM versus ANOVA and ANCOVA

Example where randomized experiments were devised to control for confounding and avoid biased estimates of causal effects.

Great idea, when feasible, but there is a lot more we can do with the data from experiments.

161

Example: Complex ecological forcing in eelgrass beds: *A global, comparative-experimental approach*

Parallel experiments
nutrient enrichment
grazer exclusion

Exploit natural gradients
biodiversity
abiotic env't (T, S, light)

Data from:
Field-based Experimental Study of the Importance of Small Herbivores in a Seagrass Ecosystem:

Matthew A Whalen and J Emmett Duffy

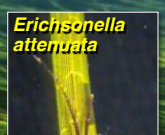
Whalen, Duffy, and Grace, 2013. *Ecology* 94:510-520.
(http://www.esajournals.org/doi/abs/10.1890/12-0156.1)




WILLIAM & MARY
VIMS
VIRGINIA INSTITUTE OF MARINE SCIENCE
SCHOOL OF MARINE SCIENCE

163

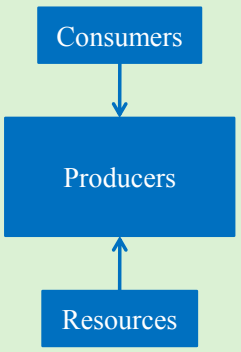
York River, Virginia:
Major herbivores are invert crustaceans - these grazers control epiphytes and promote the eelgrass

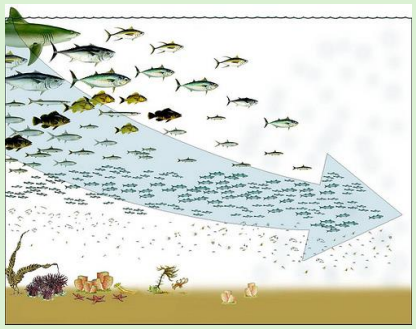


164

The Big Question

Are seagrasses controlled by bottom-up forces or trophic cascade?





Subtext: Is nutrient runoff or overfishing causing seagrass declines?

165

Preliminary Study:
Virginia site

Experimental Design:

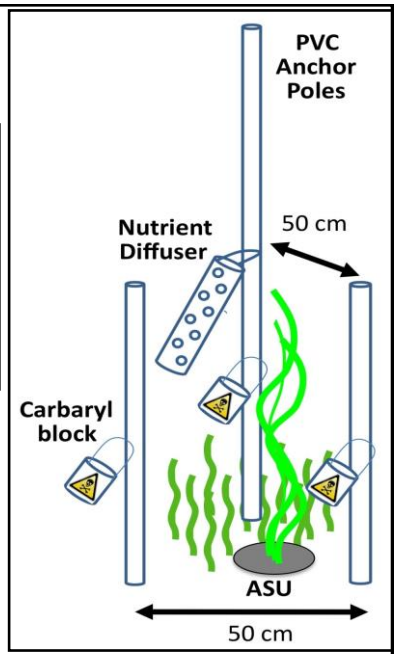
Treatments:
- pesticide
- nutrient addition
- combination
- controls

8 reps @ 5 trts = 40 plots

Pesticide effects:

Crustaceans: reduced 58-96%
Algal biomass: increased 130-748%

Nutrients: nonsignificant effects



166

A Primary ANOVA result:
Means for pesticide effect on epiphytes

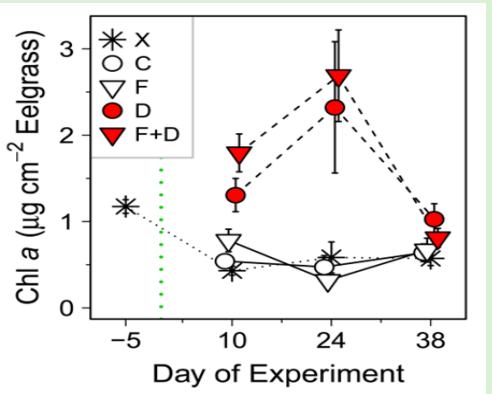
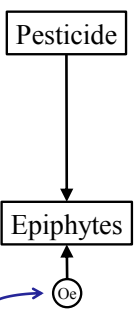


Illustration of ANOVA-type model

```
# Read Whalen Seagrass Data
w.dat<-read.csv("WhalenData.csv")
```

```
# ANOVA Model
anovaModel<-'epiphytes ~ pesticides'
```



We are using slightly different notation here.

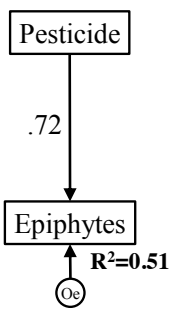
Illustration of ANOVA-type model (cont.)

```
# Fit ANOVA Model
anovaFit<-sem(anovaModel, data=w.dat)

# Get Results
summary(anovaFit, standardized=T, rsq=T)
```

```
                   Est    SE      Z      P Std.all
Regressions:
  epiphytes ~
    pesticides   0.998  0.154   6.48  0.000  0.716

Variances:
    epiphytes    0.227  0.051                0.488

R-Square:
    epiphytes    0.512
```

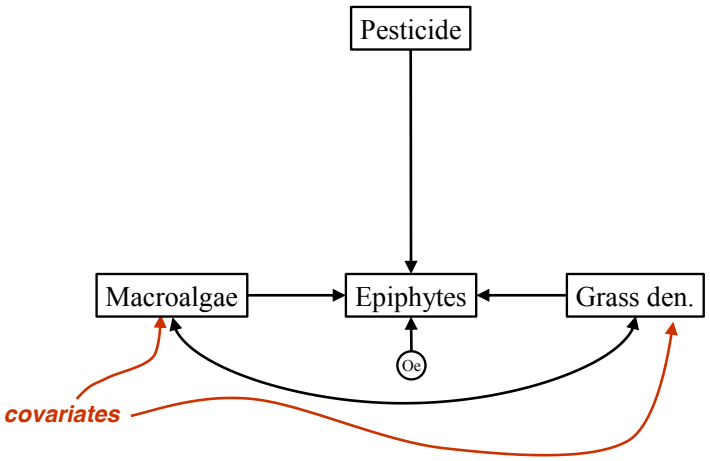
169

Results



170

Illustration of ANCOVA-type model

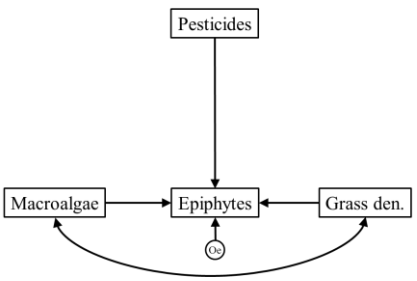


Note: in ANCOVA, covariates are not allowed to correlate with treatment variables.

171

Illustration of ANCOVA-type model

```
# Simple ANCOVA Model
ancovaModel<-'epiphytes ~ pesticides
                        + macroalgae + grass'
```



172

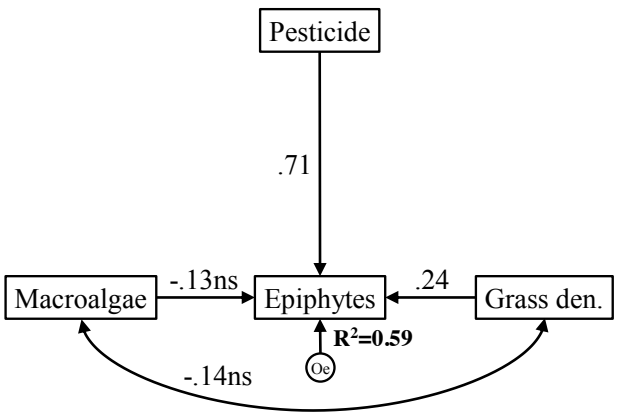
Results

```
ancova.fit <- sem(ancovaModel, data=w.dat)

summary(ancova.fit, stand=T, rsq=T)
```

We request r-squares.

```
                   Est     SE       Z      P Std.all
Regressions:
  epiphytes ~
    pesticides   0.998  0.154   6.48  0.000  0.716

Variances:
    epiphytes    0.227  0.051                0.488

R-Square:
    epiphytes    0.512
```


Results (visual)



1. Variance explanation for epiphytes improves.
2. Grass density promotes epiphyte development.
3. Macroalgae have nonsignificant negative effect on epiphytes.

The test of mediation

Does reduction of Gammarids explain promotion of epiphytes by pesticide?

note that Gammarids are class of crustations whose abundance is most strongly reduced by the pesticide.

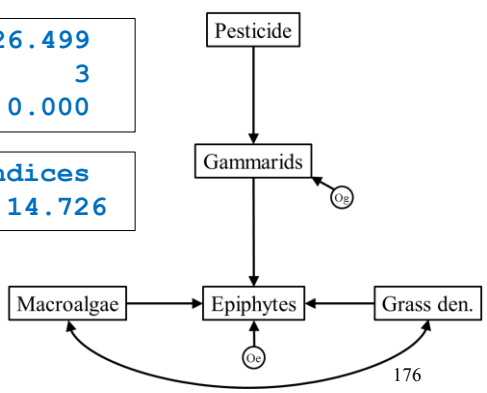
Lavaan code and results

```
# SEM Model 1 "sem1"

sem1 <- 'epiphytes ~ macroalgae + grass + gammarids
         gammarids ~ pesticide'
```

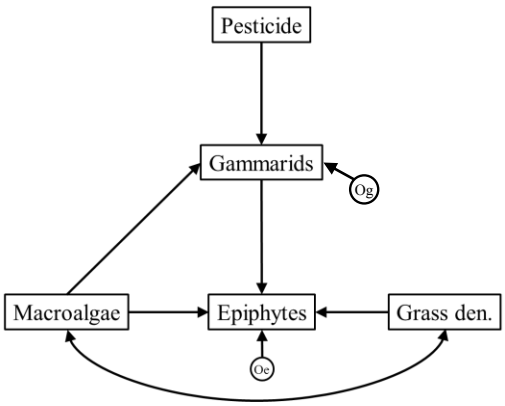
```
Chi-square                 26.499
Degrees of freedom          3
P-value                    0.000
```

```
# Select Modification Indices
gammarids ~ macroalgae  14.726
```

So, we should add path from macroalgae to gammarids.



Modifying our model: adding needed linkages




```
# New Model - SEM Model 2 "sem2"

sem2 <- 'epiphytes ~ macroalgae + grass + gammarids
         gammarids ~ pesticide + macroalgae'
```

Results

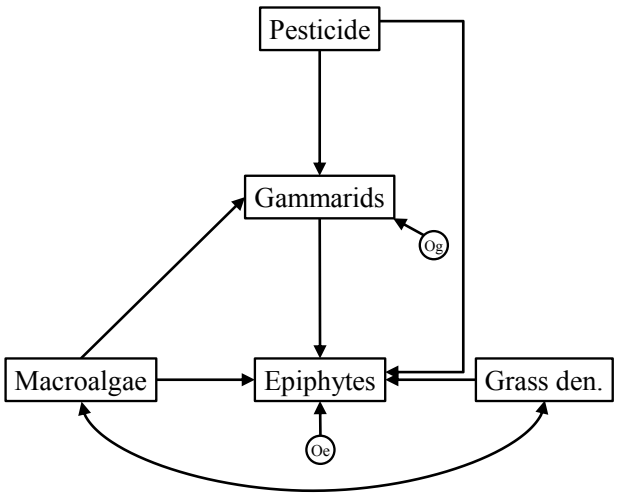
```
Chi-square                8.136
Degrees of freedom            2
P-value                   0.017
```

```
# Chi-square difference test
anova(sem1.fit, sem2.fit)

Chisq-diff = 18.363,
df-dif     =  1
p          = < 0.001
```

```
# Select Modification Indices
gammarids ~ grass     3.319
epiphytes ~ pesticide 4.205
```

178

New model to examine: “sem3”



179

sem3 model and results

```
# SEM Model 3 “sem3”

sem3 <- 'epiphytes ~ macroalgae + grass + gammarids
                   + pesticide
         gammarids ~ pesticide + macroalgae'
```

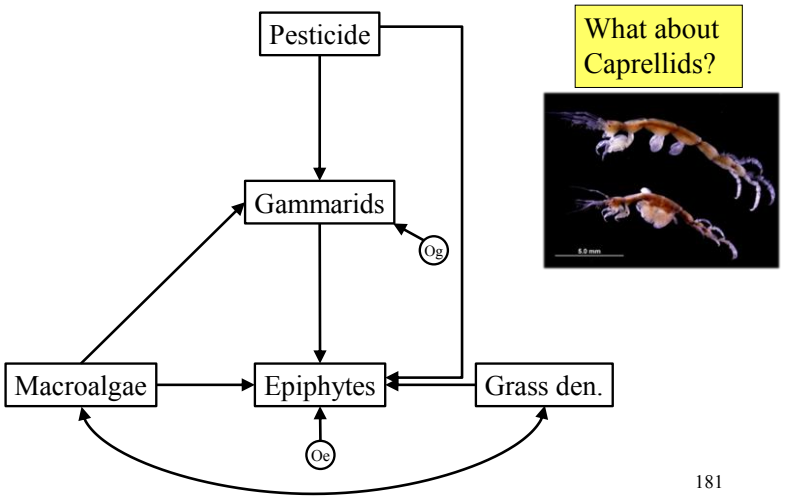
```
Chi-square                3.465
Degrees of freedom            1
P-value                   0.063
```

```
# Chi-square difference test
anova(sem1.fit, sem2.fit)

Chisq-diff =  4.671
df-dif     =  1
p          = < 0.031
```

180

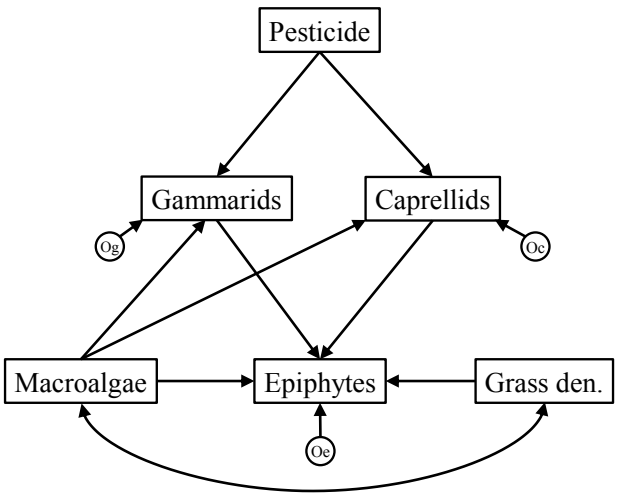
We can go further.

What is mediating the remaining effect of pesticide on epiphytes?



181

"sem4" model



182

"sem4" model and results

```
# SEM Model 4 "sem4"

sem4 <- 'epiphytes ~ macroalgae + grass + gammarids
                   + caprellids
         gammarids ~ pesticide + macroalgae
         caprellids ~ pesticide + macroalgae'
```

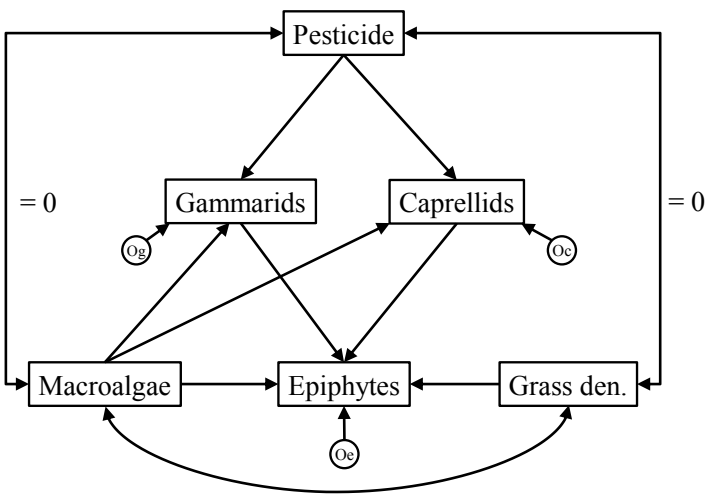
```
Chi-square                  9.182
Degrees of freedom              4
P-value                     0.057
```

```
# Select Modification Indices
nothing obvious
```

183

What if we wanted to include some constraits?

Here we force the correlations between treatment and covariates to equal 0.



184

"sem5" model and results

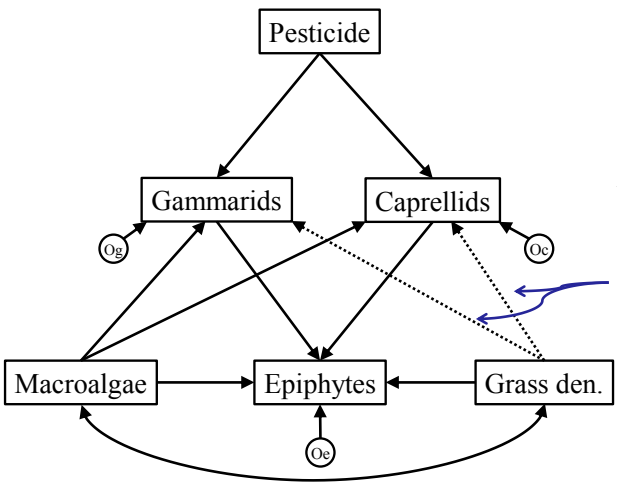
```
# SEM Model 5 "sem5"

sem5 <- 'epiphytes ~ macroalgae + grass + gammarids
                   + caprellids
         gammarids ~ pesticide + macroalgae
         caprellids ~ pesticide + macroalgae
         pesticide ~~ 0*macroalgae
         pesticide ~~ 0*grass'

sem5.fit <- sem(sem5, data=w.dat, fixed.x=F)
```

```
# Chi-square difference test
anova(sem4.fit, sem5.fit)

Chisq-diff =  1.363
df-dif     =  3
p          =  highly ns
```

```
# note we must
declare
"fixed.x=FALSE"
to work with
exogenous
correlations.
```

185

Final accepted model



Note that we show the paths from Grass den to illustrate we tested them (optional).

Chi-square = 5.432, df = 5, p = 0.366

186

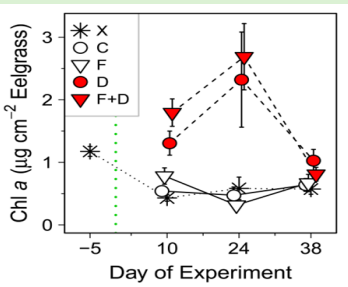
Results

	Est.	Std.err	Z-val	P(>|z|)	Std.all
Regressions:					
epiphytes ~					
macroalgae	0.105	0.040	2.612	0.009	0.290
grass	0.405	0.100	4.034	0.000	0.389
gammarids	-0.329	0.057	-5.828	0.000	-0.663
caprellids	-0.240	0.085	-2.834	0.005	-0.335
gammarids ~					
pesticide	-2.053	0.215	-9.570	0.000	-0.748
macroalgae	0.304	0.057	5.347	0.000	0.418
grass	0.315	0.164	1.922	0.055	0.150
caprellids ~					
pesticide	-0.748	0.231	-3.239	0.001	-0.393
macroalgae	0.243	0.061	3.965	0.000	0.481
grass	0.231	0.176	1.311	0.190	0.159
R-Square:					
epiphytes	0.645				
gammarids	0.756				
caprellids	0.411				

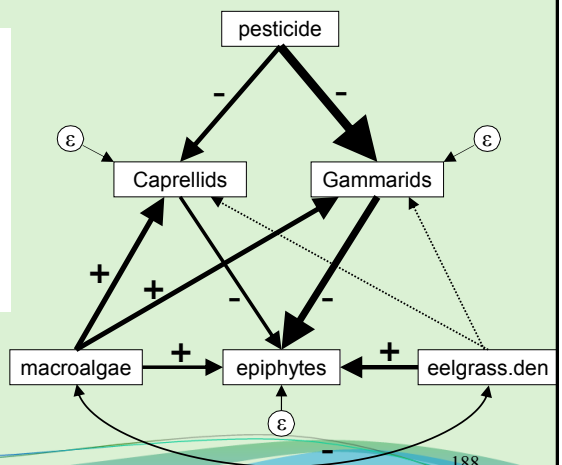
187

Our Inference

Our model results imply that behind this summary of mean responses



is a network of effects like this.



188

Lessons about using SEM with experimental data

1. Test of mediation is neglected concept in biometrics. This results directly from the limitations of classic ANOVA and ANCOVA analyses.

2. A further limitation of ANCOVA is that if we used mediating variables as covariates, the results would indicate no significant treatment effect!

3. SEM easy to implement with simple experimental designs. With blocking, nested designs, etc., more work required for SEM analyses.

4. We generally recommend performing classic analyses along with SEM analyses and reporting both. Classical analyses can more easily detect interactions and in SEM you have to work to examine them (more on that later).

189

Discussion

190