SEM Essentials: Model Specifications

U.S. Department of the Interior
U.S. Geological Survey

1

In this module I talk about how responses and relationships in models are specified. This relates to more purely statistical issues and, therefore, to software choices for estimation.

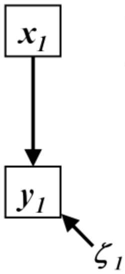
Grace, J.B., Schoolmaster, D.R. Jr., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M., and Schweiger, E.W. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3(8): article 73 (44 pages). http://www.esajournals.org/doi/pdf/10.1890/ES12-00048.1.

Notes: IP-056512; Support provided by the USGS Climate & Land Use R&D and Ecosystems Programs. I would like to acknowledge formal review of this material by Jesse Miller and Phil Hahn, University of Wisconsin. Many helpful informal comments have contributed to the final version of this presentation. The use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Last revised 17.01.29.

Source: https://www.usgs.gov/centers/wetland-and-aquatic-research-center/science/quantitative-analysis-using-structural-equation

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



What is the <u>form of the effect</u> of x_1 on y_1?
(basic assumption is “simple and linear”)

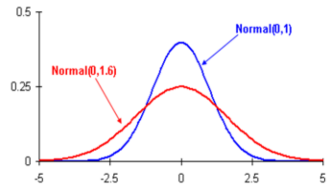
What is the <u>form of the responses</u> we expect?
(basic assumption is “Gaussian” or “normal”)

USGS

2

Together, the response forms and linkage types make up the “functional form” of the relationships in the model.

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



Distribution has a <u>mean</u> and a <u>precision</u> (= 1/variance).

<u>Gaussian (normal) responses for y1 ~ x1</u>:
each link involves two relationships:

(for [i] samples)

Each observation of **y1** is drawn from a normal distribution with predicted value **y1.hat** and precision **y1.precision**.

y1[i] ~ dnorm(y1.hat[i], y1.precision) # eqn1

y1.hat[i] ← $b_{1.0} + b_{1.1}$*x1[i] # eqn2

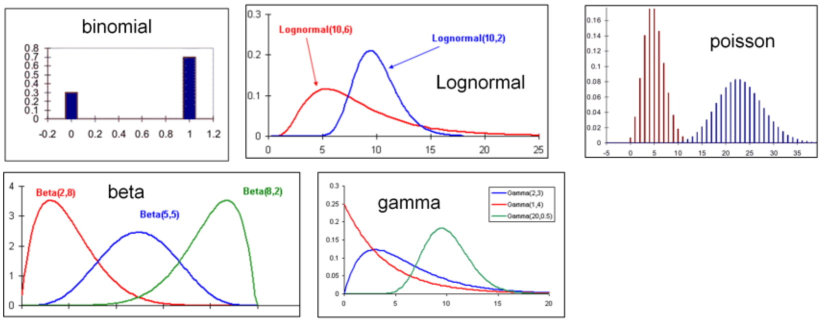
USGS

Predicted value **y1.hat** is linear function of the predictors.

3

We understand that most of the time responses are not strictly Gaussian or “normal”. Rather, we use a Gaussian approximation to the problem, which means we by assuming normal errors, our probability statements that are not too far from valid estimates.

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



poisson responses for y1 ~ x1:
(for [i] samples)

y1[i] ~ dpoisson(y1.hat[i]) # eqn1 (note poisson only has 1 parameter)

y1.hat[i] $\leftarrow b_{1.0} + b_{1.1}$*x1[i] # eqn2

USGS

Variations in variables in the real world are often not “normal”.

Images from http://www.vosesoftware.com/.

4. We also often assume linkage forms are linear and additive.

Simple, linear relationship with linear equation

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

Polynomial, curvilinear equation with linear terms

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

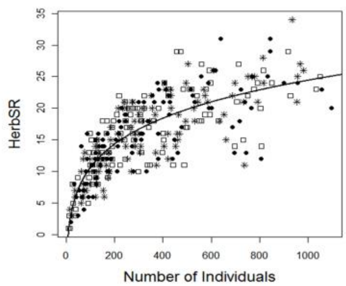
USGS

5

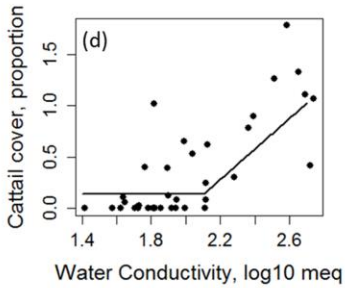
A lot can be done with linear equations. For example, the so-called Taylor expansion series of equations can be used for polynomial modeling of relationships that are highly curvilinear. Additivity means terms are summed to produce the predictions. Making such assumptions can be very handy.

5. Some situations may require more complex linkage equations.

$y = a + bx^c$



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



USGS

6

Of course in many situations more complex equations have advantages over linear additive approximations in fitting relationships. The equation on the left is a power function. The one on the right is a step or threshold function.

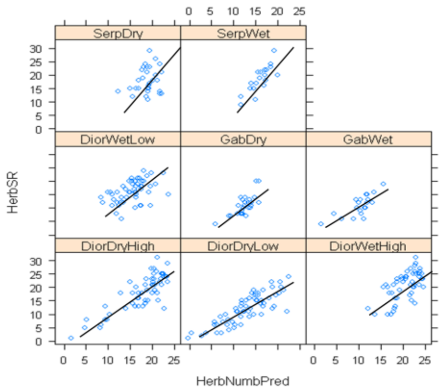
6. Equations may be hierarchical/multi-level.

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

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

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

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



Ecology, 92(1), 2011, pp. 108–120

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

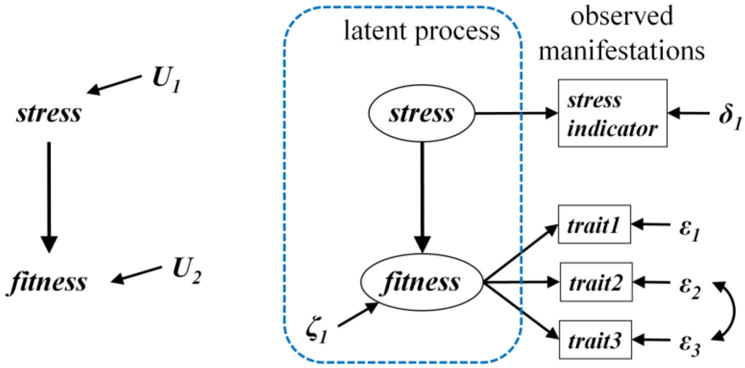
USGS

Data relationships are often hierarchical (also known more generally as “multi-level”). It is very popular in the quantitative community these days to recognize this feature of data by using hierarchical equations where the terms in one equation themselves have equations (left upper inset box). This allows us to represent situations where regression relationships for different groups have different slopes and intercepts.

In this example, herb species richness “HerbSR” is a function of the number of herbs “HerbNumbPred” but also the region sampled (e.g., “SerpDry”).

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

Causal Diagram | SE model

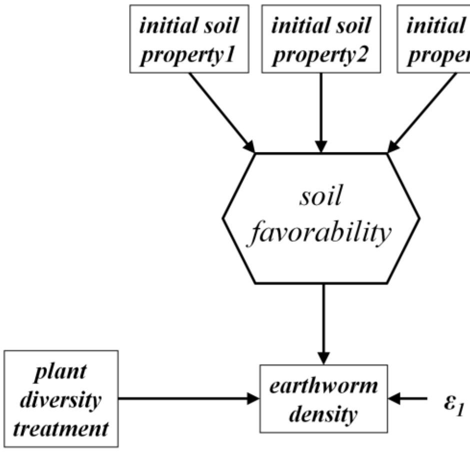


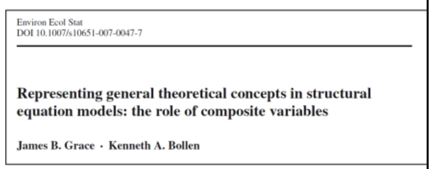
USGS

8

Latent variables are used for various modeling purposes, such as to represent a more general concept, but ultimately they behave like a variable with missing values. There are some subtle issues related to modeling with latent variables, so I generally treat this as an advanced topic and details of modeling with latent variables is covered later.

8. We may be interested in summarizing collective effects of groups of variables using composite variables (e.g., soil favorability).




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

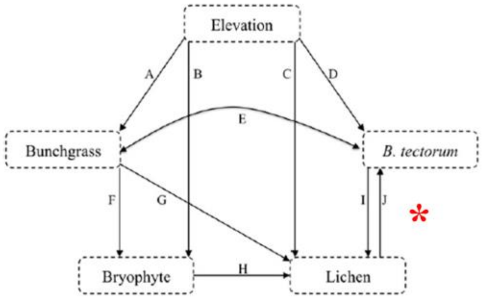
USGS

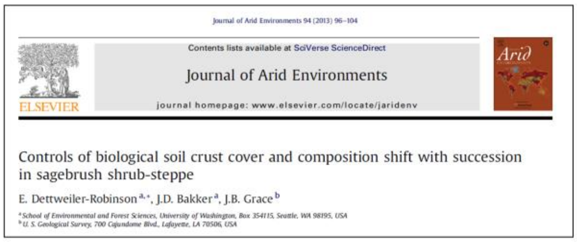
9

A “composite” variable is one that combines multiple causal influences. I often contrast composite variables with latent variables, but that is a bit of an oversimplification. This topic is covered in more detail in a separate module. It is also covered in detail in the following paper:

Grace et al. 2010. On the specification of structural equation models for ecological systems. Ecological Monographs 80:67-87.

9. Our model may include reciprocal interactions or loops.




Journal of Arid Environments 94 (2013) 96–104

Contents lists available at SciVerse ScienceDirect

Journal of Arid Environments

journal homepage: www.elsevier.com/locate/jaridenv

ELSEVIER

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

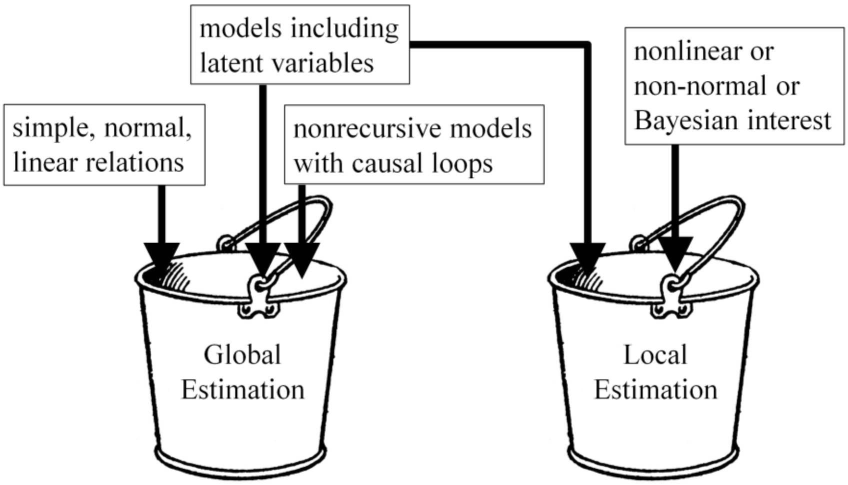
E. Dettweiler-Robinson[a,*], J.D. Bakker[a], J.B. Grace[b]

[a] School of Environmental and Forest Sciences, University of Washington, Box 354115, Seattle, WA 98195, USA
[b] U. S. Geological Survey, 700 Cajundome Blvd., Lafayette, LA 70506, USA

USGS

When models are non-recursive, such as when they contain reciprocal interactions, special requirements apply for both estimation and valid interpretation. What is important here is simply the fact that only certain estimation methods are well suited to handling such cases.

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



USGS

(More on this distinction to follow.)

11

So, what I have been building up to is this slide, which represents the idea that our choice of estimation strategy depends on model specification. As the module on "Estimation" will describe, there are two basic approaches.

In classical SEM (what is implemented in specific SEM software packages at the moment), a matrix-based approach is taken. In this approach, the data are summarized in the form of a matrix of covariances. This allows for a ready handling of latent variables and estimation in models with causal loops. Also, there are procedures for relaxing assumptions related to Gaussian relations, but these are limited.

When we want to include more complex functional forms in our models and we don't intend to work with latent variables or nonrecursive models, we can use any sort of local estimation method. Essentially there is where we estimate each equation separately.

Bayesian methods permit us to estimate latent relations using local estimation methods and also permit a great variety of functional forms. What they cannot handle, however, are non-recursive relations.

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, piecewiseSEM (for local estimation);

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

USGS

Note: Local estimation of observed-variable models can be performed using any conventional statistical package.

12

The software choices and features are constantly evolving. All of these are used by some ecologists, though there is a gradual migration to R amongst many.