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.


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 20141216. Questions about this material can be sent to sem@usgs.gov.
1. Two elements of all probabilistic equations are:
   (a) response forms and (b) linkage types.

   - What is the form of the effect of $x_j$ on $y_j$?
     (basic assumption is “simple and linear”)

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

Together, the response forms and linkage types make up the “functional form” of the relationships in the model.
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 consider assuming normal errors gives us probability statements that are not too far from valid estimates.
Variations in variables in the real world are often not “normal”.
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.
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.

\[ y_{ig} = \beta_{0g} + \beta_{1g} x_{ig} + \epsilon_{ig} \]
\[ \beta_{0g} = \gamma_{00} + \gamma_{01} W_g + \gamma_{02} Z_g + u_{0g} \]
\[ \beta_{1g} = \gamma_{10} + \gamma_{11} W_g + \gamma_{12} Z_g + u_{1g} \]

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

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”).
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.
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:

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.
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 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 using any conventional statistical package.

The software choices and features are constantly evolving. All of these are used by some ecologists.