SEM versus Multiple Regression

Jim Grace

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

This module illustrates SEM via a contrast with multiple regression.

The module on Mediation describes a study of post-fire vegetation recovery in southern California woodlands. Here I borrow that study to first consider what could be obtained from a regression study of that problem. I follow that by illustrating SEM in comparison.

An appropriate general citation for this material is

Grace, J.B. (2006) *Structural Equation Modeling and Natural Systems*. Cambridge University Press.

The specific example is drawn from results in

Grace, J.B. and Keeley, J.E. (2006) A Structural Equation Model Analysis Of Postfire Plant Diversity In California Shrublands. *Ecological Applications* 16:503–514.

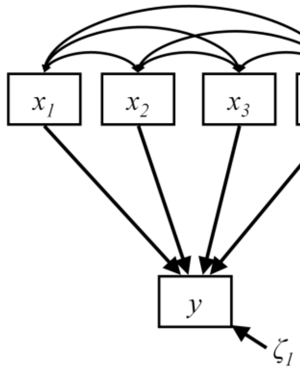
Use R&D and Ecosystems Programs. I would like to acknowledge formal review of this material by Jesse Miller and Phil Hahn, University of Wisconsin. Many helpful informal comments have contributed to the final version of this presentation. The use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Last revised 20141216. Questions about this material can be sent to sem@usgs.gov.

What if we used a multiple regression approach to the problem of understanding vegetative recovery following wildfires?

equational view of multiple regression

$$y = \alpha + \mathbf{Bx} + \varepsilon$$

graphical view of multiple regression



USGS

We are all familiar with the equation for a multiple regression. In the simple case, variations in some y variable are understood in terms of their relations with a vector of x variables. Note here the bold **x** signifies a set of predictors.

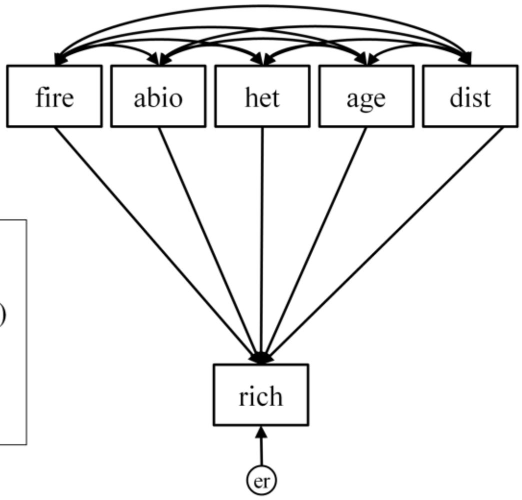
It becomes quite revealing if we borrow from the SEM toolbox the causal analysis principle of graphing the relations implied by the equation(s).

What emerges from the graphical representation is that there is a permitted but unanalyzed set of correlations among the predictors. Students of statistics know that those correlations have a huge determining influence on the coefficients that link xs to the y.

What is scientifically most important is that we scientists are not permitted to incorporate any knowledge about WHY the xs are correlated, despite the importance of those correlations. This, as we shall see, is a major loss of opportunity.

Further, the unanalyzed correlations among predictors make it darn near impossible to create a proper causal model, since there are many “unanalyzed associations” that get in the way of interpretations.

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)

USGS

3

Here we show a multiple regression designed to determine what predictors are required to predict values of richness. It might seem to students of statistics that they are seeking causal models, but stats professors will usually make it quite clear that only a parsimonious set of predictors should be expected from such a model.

Multiple regression results for initial model.

```
# multiple regression model using lm
mr.lm1 <- lm(rich ~ firesev + abiotic + hetero
      + age + distance, data=k6.dat)
```

	Est.	Std.err	Z-value	P(>|z|)	
(Intercept)	-1.44009	1.09307	-1.317	0.19126	
firesev	-0.16670	0.07976	-2.090	0.03965	*
abiotic	0.48065	0.17096	2.811	0.00614	**
hetero	0.34980	0.10778	3.245	0.00169	**
age	-0.09051	0.10418	-0.869	0.38741	
distance	0.52786	0.15712	3.360	0.00118	**

Indications are age does not contribute to the model.
Evaluations confirm.

USGS

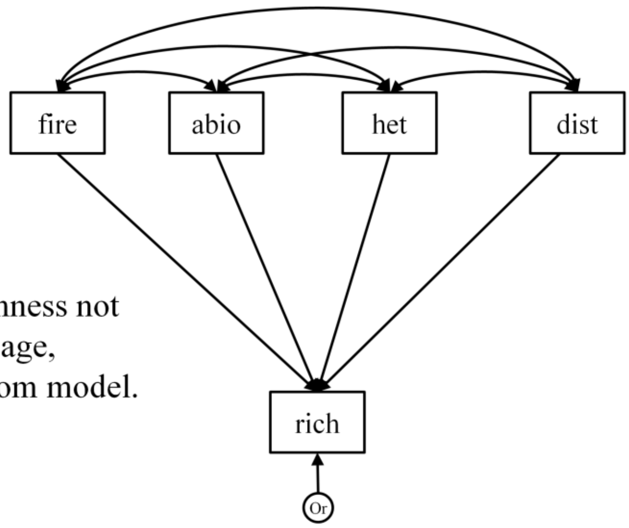
4

If we run a multiple regression model we obtain a set of parameter estimates and some assessment of whether the included predictors are needed to explain the observed variations.

Results give an indication that age is not needed in the prediction equation.

To save time, I simply mention that model comparisons confirm age can be dropped from the model.

Pruned model.



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

USGS

5

We might conclude from multiple regression findings that age of the stand that burns is not an important influence on post-fire richness. Such a conclusion, as I shall show, is not at all a proper conclusion.

An SEM Approach to the Same Problem

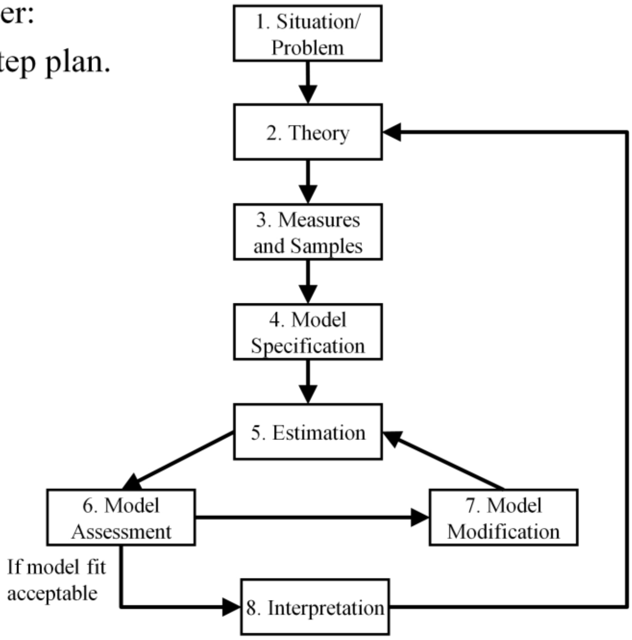
USGS

6

How might we approach the same scientific objective using SEM?

Reminder:

The 8-step plan.



USGS

7

Here is a reminder of a slide in one of our SEM Essentials modules showing 8 major steps in SEM. We will use this numbering to walk through the process.

Step 1: Situation – heterogeneous fire in heterogeneous landscape

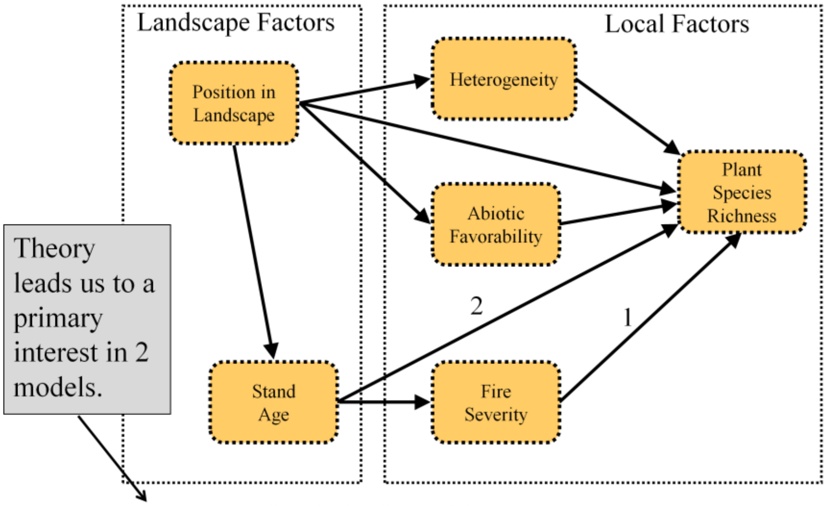
USGS

8

What is our situation?

We want to understand what controls recovery from wildfire in a heterogeneous landscape.

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

9

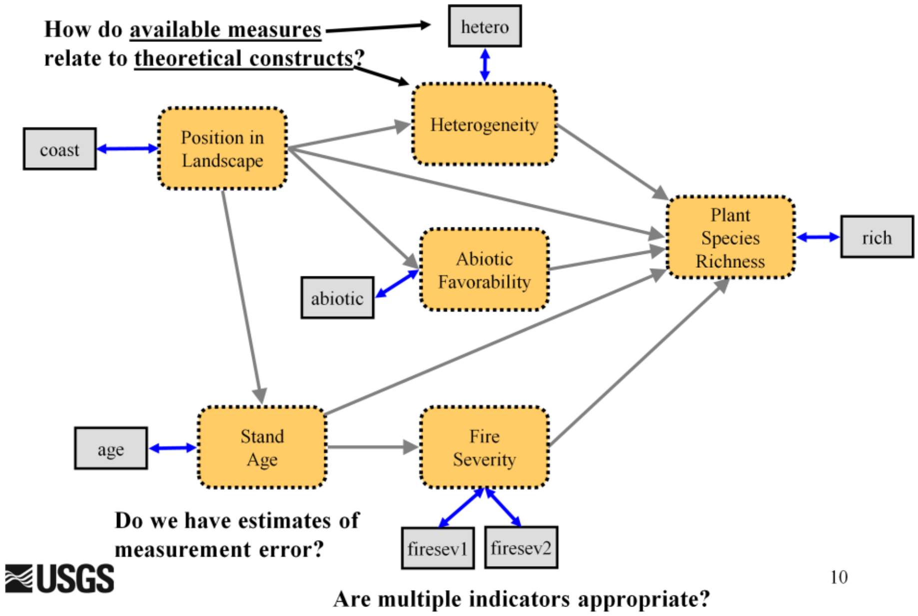
Grace and Keeley develops a kind of theory for how to think about the possible controls.

For the sake of this illustration, there were two major models of competing interest.

Model 1: The age of a stand only influences post-fire richness through its fuel-related impacts on fire severity.

Model 2: Older stands will have a reduced seed bank due to steady mortality of seeds in the seed bank. This is based on the idea that seed replenishment in the seed bank for many of the species only takes place after a fire.

Step 3: Consider measures and samples.



A key part of SEM, which is only alluded to here, is the evaluation of construct measurement. In disciplines like psychology and sociology, this is often the dominant issue to be addressed and the literature on SEM is heavily oriented to a multi-indicator factor model perspective focused on measurement issues.

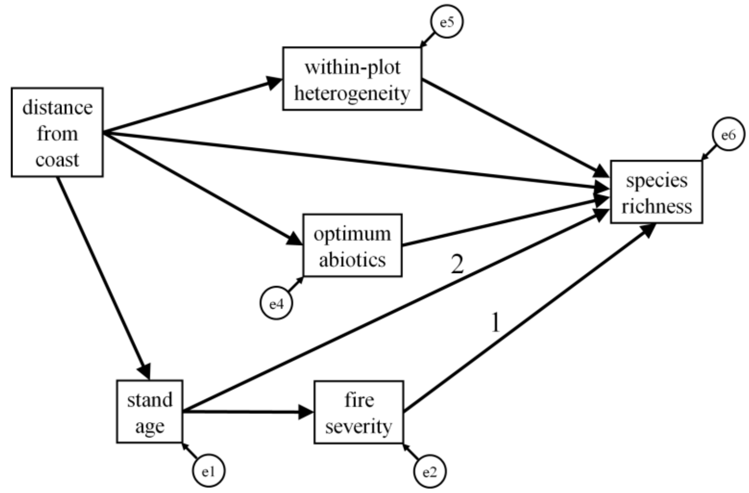
There are actually two issues here.

(1) indicator validity – do measures actually represent the theoretical entities of interest?

(2) indicator reliability – is there measurement error in estimating the true quantities of causal interest?

More about all this is presented in the module on Latent Variables in Models.

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

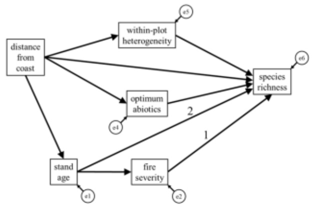
11

Here we adopt the “biometric” tradition (also the “econometric” tradition) and simply choose what we believe to be our best measures for each theoretical construct and assume no measurement error.

There are actually a number of other assumptions, some of which will be discussed in the module “Causal Modeling Revisited”.

Step 5: Estimation.

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

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



USGS

12

Here is lavaan code for the more complete model, model 2. Our purpose of running this first is to determine whether any of the models are sufficient before comparing the two models of prime theoretical interest.

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

```
# Estimate model "fire.2"
mod.2.fit <- sem(mod.2, data=k6.dat)
summary(mod.2.fit)
```

```
lavaan (0.5-15) converged normally after  20
iterations

  Number of observations                            90

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

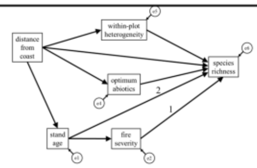
good fit means we are not missing any links.

USGS *"Minimum Function Test Statistic" = "Model Chi-Square" 13

Standard results suggest model sufficiency, i.e., no missing links.

Step 6: Model assessment – Information measures.

```
Model selection based on AIC :

> anova(mod.1.fit, mod.2.fit)
Chi Square Difference Test

           Df    AIC    BIC  Chisq Chisqdiff Df diff
Pr(>Chisq)
mod.2.fit   6 1589.0 1624.0 6.0946
mod.1.fit   7 1587.8 1620.3 6.8998    0.8052       1 0.3695
```



Delta_AIC lower for mod.1, suggests no direct path from age.

```
Model selection based on AICc :

       K    AICc   Delta_AICc AICcWt Cum.Wt      LL
mod.1 13 1589.56         0.00    0.7    0.7 -780.89
mod.2 14 1591.22         1.67    0.3    1.0 -780.49
```

USGS

Delta_AICc lower for mod.1 also,

14

We can now compare our two models of interest.

Here I show to approaches. The first is a chi-square difference test, which shows the model with direct path from age to rich (mod.2) is not significantly better that the model without that link.

AIC and AICc both favor model 1 and support conclusion direct link from age to richness is not needed.

Step 6: Model assessment - Are all paths supported?

	Est	Std.err	Z-value	P(>|z|)	Std.all
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
age	0.000				0.000
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

USGS

Indications and tests say yes.

Indications (p-values) and tests (not shown) indicate all paths in model are supported.

Step 8: Interpretation – variance explained.

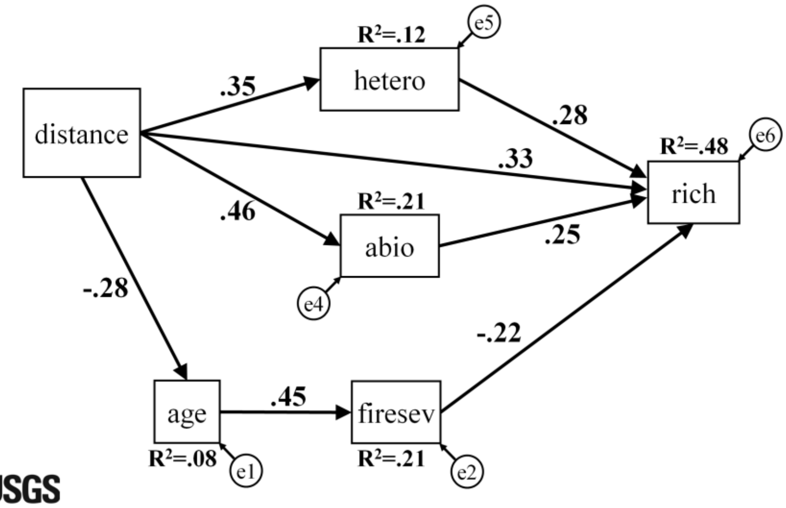
```
R-Square:

    rich                0.484
    abiotic             0.211
    hetero              0.120
    age                 0.077
    firesev             0.206
```

USGS

16

Here I extract the R-square computed for the model, which are at the bottom of the output.

Step 8: Interpretation – visualization of results.



USGS

17

And here is a visual presentation of key results.

Step 8: Interpretation – extrapolations.

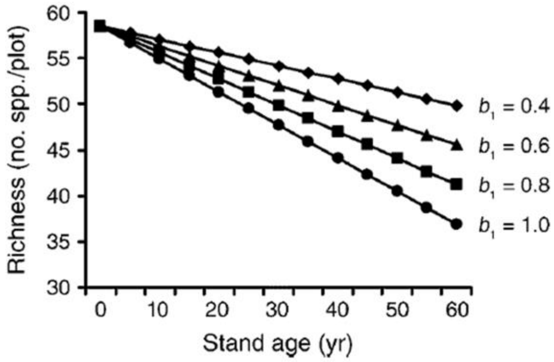


FIG. 5. Predicted sensitivity of richness to stand age at various levels of fire intensity (as a proportion of natural strength): $b_1 = 1.0$ represents the average fire severity observed in these wildfires, while values <1.0 represent expectations if fire intensity were lower, for example, through the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

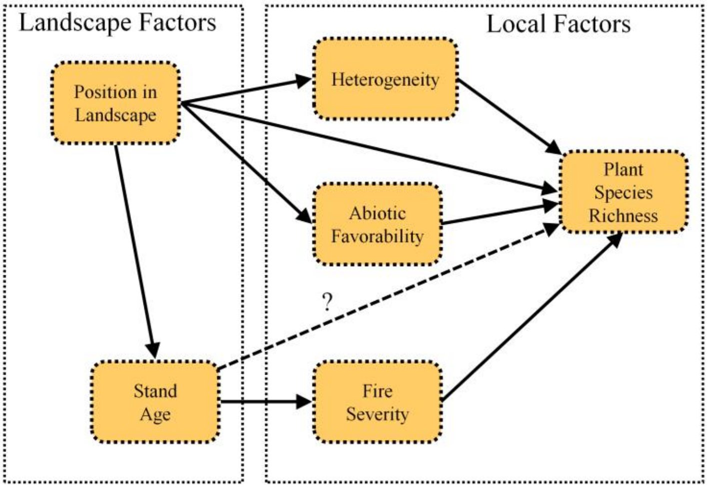
USGS

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

In this study we went on to ask what would potentially happen if prescribed fire was used to reduce fire severity.

Interestingly, the results depend on stand age and suggest that prescribed fire in older stands might enhance post-fire richness quite a lot.

Step 8: Interpretation – revising theory.



USGS

19

We then go back to our conceptual ideas. Keeley followed up this study with additional studies where various ideas and uncertainties were explored further. That is the way SEM is supposed to be used.