Modeling with Latent Variables

Jim Grace

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

In this module I give a few basics for working with latent variable models.

An appropriate general citation for this material is

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.

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 17.02.05.

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

What is a latent variable?

"A variable for which we do not have measurements."

Q: How should we think about latent variables in models?.

A: A single latent variable acts like a single missing variable.

Levels of abstraction:
- True values for *y*.
- General properties of *y*.
- A general theoretical/hypothetical concept of interest.

USGS 2

It is useful to note that latent variables range in their level of abstraction from simply meaning "the true value" to being "deeply latent" ideas that are highly abstract and of uncertain reality.

General References.

Grace, J.B. and Bollen, KA. 2008. Representing general theoretical concepts in structural equation models: the role of composite variables. Environmental and Ecological Statistics 15:191-213. (http://link.springer.com/article/10.1007/s10651-007-0047-7)

(http://www.odum.unc.edu/content/pdf/Bollen%20Grace%20Bollen%20(preprint%202008)%20Environ%20and%20Ecol%20Stats**.pdf**)

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)

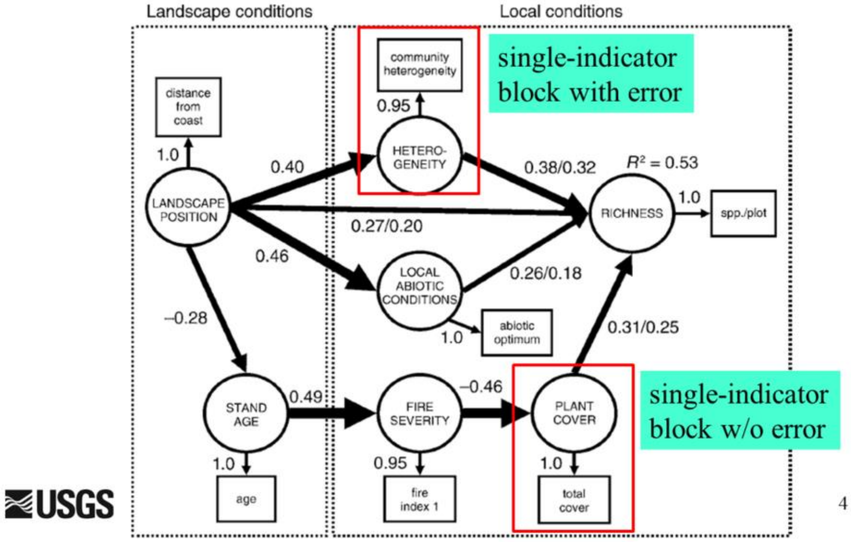
Bollen, K.A. 2012. Latent variables in structural equation modeling. Chapter 4, In: Hoyle, R.H. (ed.) Handbook of Structural Equation Modeling. Guilford Press, New York.

USGS

Some references that make key distinctions and provide diagnostic criteria.

Ecological Examples Using Latent Variables.

Grace and Keeley (2006) A structural equation model analysis of postfire plant diversity in California shrublands. *Ecol. Apps*. 16:503-514.

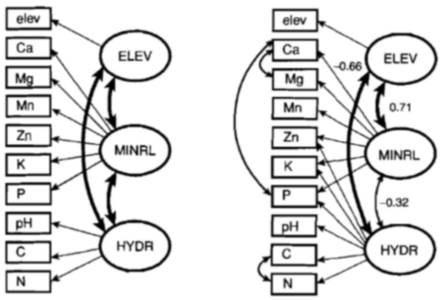


Let's look at some examples of the use of latent variables in models. This model is from the published version of the fire SEM published by Grace and Keeley. Here the investigators used latent variables, even though they did not use multiple indicators, because they wanted to distinquish the concepts of interest from the measurements available. Also, the investigators had estimates of precision (and therefore, measurement error) for a couple of the properties of interest.

The interpretation of this model, as represented, is that there are interactions amongst the latent factors and we observe the surface manifestations of those hidden processes (with some error).

Ecological Examples Using Latent Variables (cont.).

Grace, J.B. (2003) Examining the relationships between environmental variables and ordination axes using latent variables and structural equation modeling. *Ch 7 in Pugesek et al. Structural Equation Modeling in Ecological and Evolutionary Biology. Cambridge Univ. Press*



	Latent factor		
Measured variable	ELEV	MINRL	HYDR
elev	1.0	—	—
Ca	—	0.66	—
Mg	—	0.43	—
Mn	—	0.99	—
Zn	—	0.66	0.81
K	—	0.53	0.84
P	—	0.39	0.93
pH	—	—	−0.66
C	—	—	0.78
N	—	—	0.67

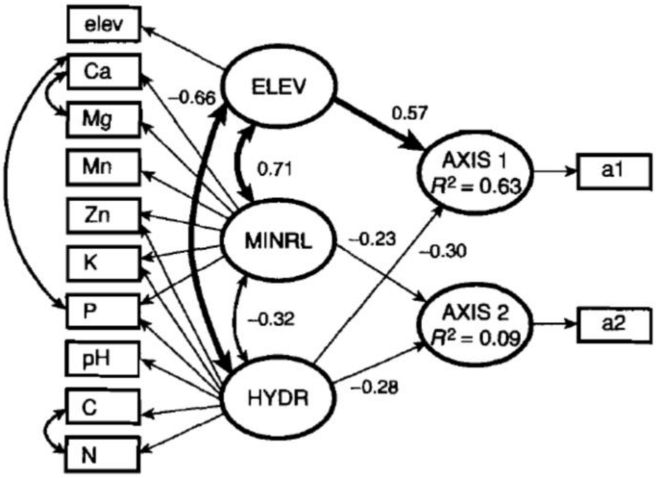
5

USGS

Another example of the use of latent variables is illustrated in this slide and the next. In the first phase of the analysis portrayed here, two latent soil properties were hypothesized to explain the intercorrelations among a measured set of soil variables. Because of the potential for elevation to have a separate influence on the plant community, it was included as a third latent variable in the model. On the left is the originally hypothesized model and to the right of it is the model selected as the best representation of the system. Factor loadings provide additional information to help interpret the system.

The model type presented here is typically referred to as a "confirmatory factor model" (CFA). This type of model is very commonly used in social sciences and psychology. A separate module showing details of the analysis is also available, along with a practice exercise.

Ecological Examples Using Latent Variables (cont.).



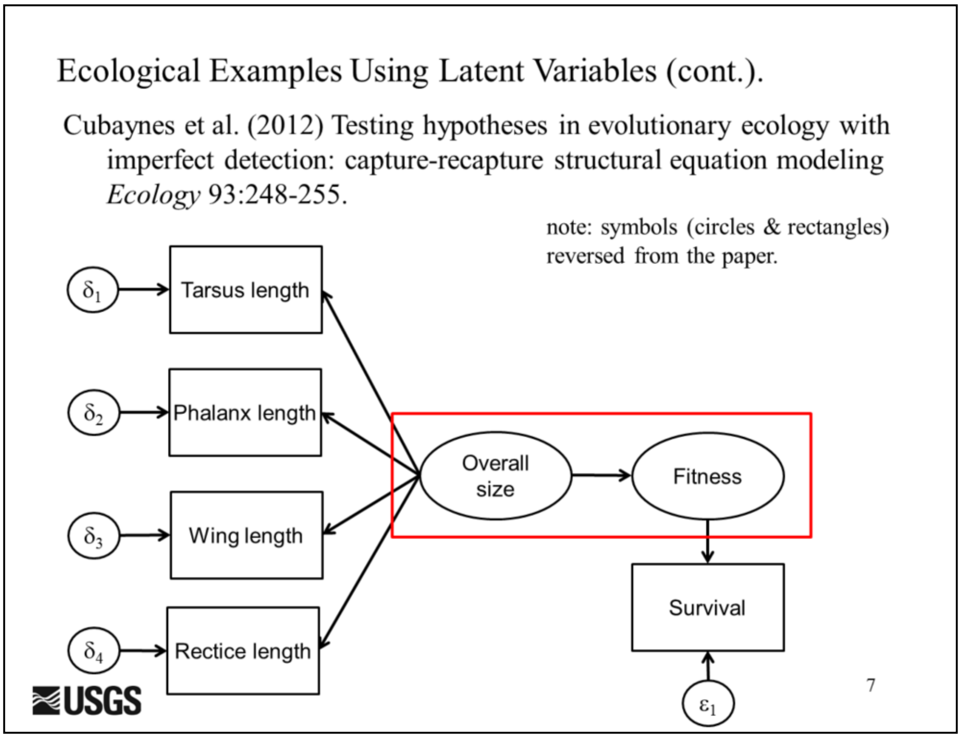
6

In the second phase of the analysis initiated on the previous slide, the two dimensions of a community ordination are related to the latent soil properties. The source for the original analysis is

Grace, J. B., Allain, L. & Allen, C. (2000b). Vegetation associations in a rare community type - coastal tallgrass prairie. ***Plant Ecology,*** **147,** 105-1-15.

The paper concluded that

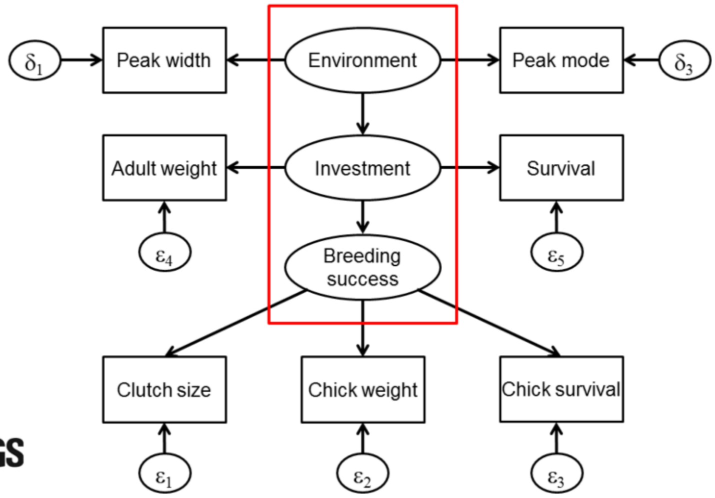
"The dominant environmental influence on species composition was found to be elevation and a host of correlated factors including those associated with soil organic content. A secondary group of factors, consisting primarily of soil cations, was found to explain additional variance among plots. Overall, this prairie was found to contain plant associations that are now rare in the surrounding landscape. Within the prairie, plant groups were largely separated by a suite of environmental conditions associated with topography. These results suggest that conservation and restoration efforts will need to carefully consider local topographic influences in order to be successful."



There have been a few studies dealing with wildlife that have incorporated SEM elements, often in combination with procedures for dealing with imperfect detection. Here is one dealing with the effect of body size on fitness in black birds.

Ecological Examples Using Latent Variables (cont.).

Cubaynes et al. (2012) Testing hypotheses in evolutionary ecology with imperfect detection: capture-recapture structural equation modeling *Ecology* 93:248-255. note: symbols (circles & rectangles) reversed from paper.



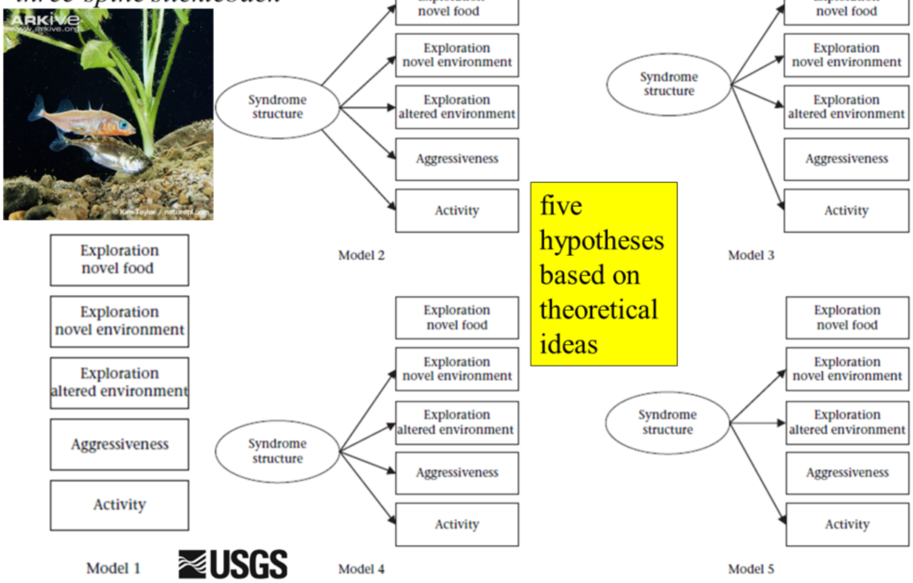
USGS

8

Cubanyes et al. also present in their paper an example of the use of SEM for the bird know as the blue tit.

Ecological Examples Using Latent Variables (cont.).

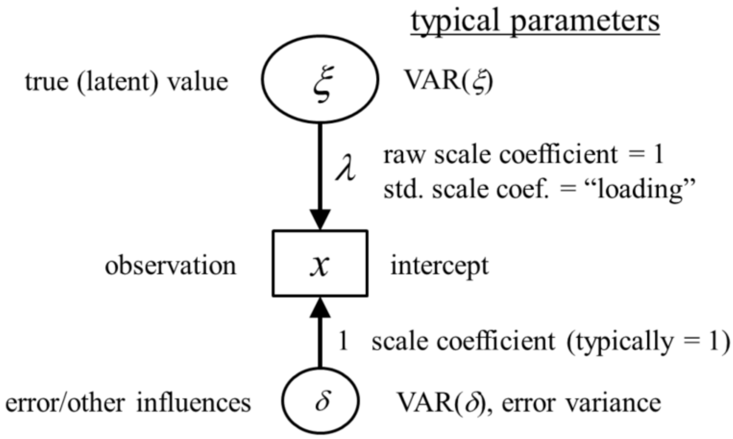
three-spine stickleback



Dinglemanse et al. (2010) A method for exploring the structure of behavioural syndromes to allow formal comparison within and between data sets. *Animal Behavior* 73:439-450.

In this example, the authors proposed a number of behavior strategies and sought to assess the empirical support for them.

LV Fundamentals: The Single-Indicator LV block.



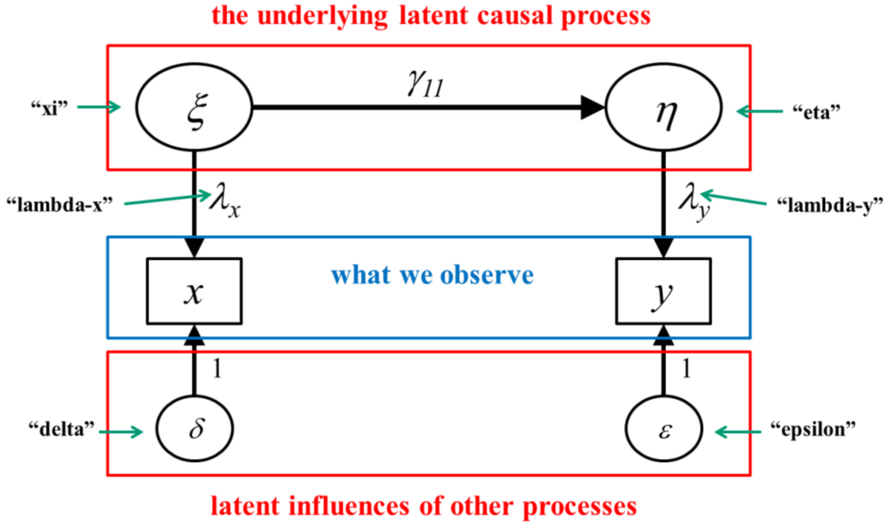
USGS

10

Now, getting to the technical bit, traditionally we use solid-line ovals for latent variables and rectangles for observed variables.

Note that technically the error term is a latent variable, though we don’t always show it that way.

A single-indicator latent regression.



USGS

11

Causation is presumed to flow from latent to observed variables (typically). Stated differently, the things we observe emanate from a latent, unseen causal world.

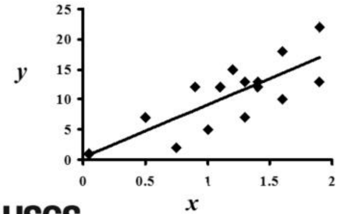
One reason for latent variables - address measurement error.

Observed variable models assume all variables are measured without error.

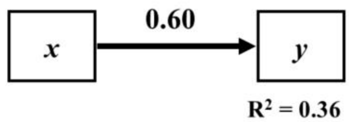
This applies to all classical statistical models, as well as to observed variable (aka “manifest”) SE models.

So, what difference does it make?

Imagine we observe this.



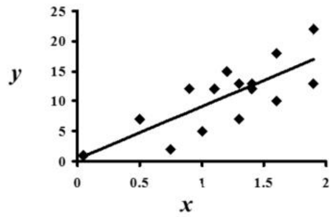
The regression / SE relationship would be.

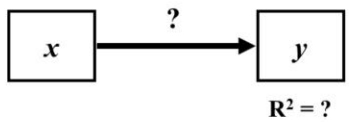


12

The issue of measurement error and its effects is virtually ignored in most statistical training, though that is starting to change. There is a very strong case for dealing explicitly with measurement error because ignoring it leads to downward bias in parameters.

Addressing measurement error (cont.).





A problem is, error in measuring *x* is assigned to the error in predicting *y*.

So, the true effect of *x* on *y* is typically underestimated to either a large or small degree.

USGS

13

Error in measuring x is interpreted as error in predicting y.

We can estimate measurement error using multiple measures.

Imagine that some of the observed variance in *x* is due to error of measurement.

Calibration data set based on repeated measurement trials.

plot	***x*-trial1**	***x*-trial2**
1	**0.556**	**0.419**
2	**-1.803**	**-1.141**
3	**0.385**	**0.497**
4	**0.616**	**-0.608**
.	**.**	**.**
n	**0.946**	**0.586**

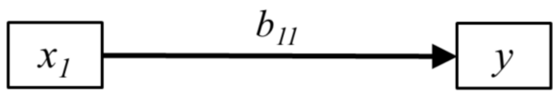
If, average correlation between trials, COR(x_1,x_2) = 0.90, then the average **reliability** of any given set of measurements is estimated at 0.90.

USGS

14

Indicator reliability is a key concept.

Illustration of empirical results ignoring measurement error.



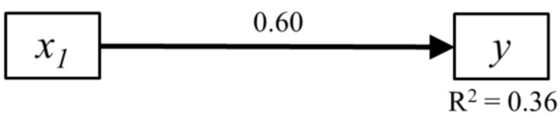

```
# specify model
mod.1 <- 'y ~ x1'

# fit model using input covariance matrix "input.cov2"
mod.1.fit <- sem(mod.1, sample.cov=input.cov2,
                 sample.nobs=100)

# request output
summary(mod.1.fit, standardized=T, rsq=T)
```

15

Here is the model and associated code.

Lavaan results ignoring measurement error.

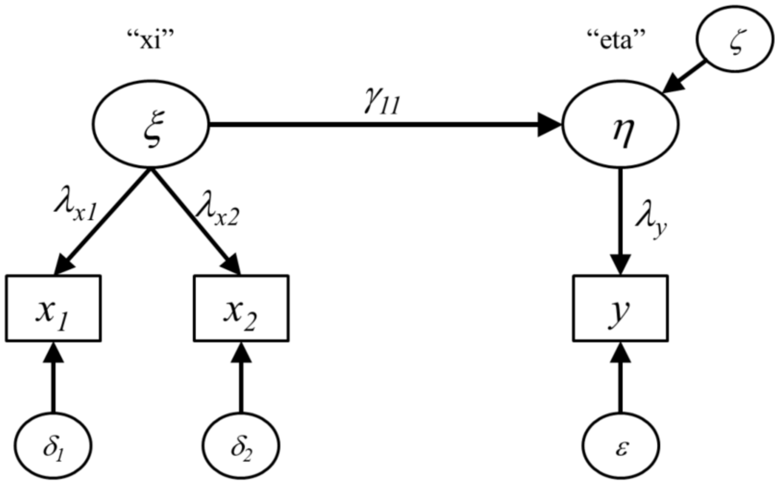
```
Parameter estimates:

                   Estimate  Std.err  Z-value  P(>|z|)   Std.all
Regressions:
  y ~
    x1                0.594    0.080    7.413    0.000     0.596

Variances:
    y                 0.642    0.091                       0.645

R-Square:

    y                 0.355
```



16

And here are the results from our regression example ignoring measuremen error.

A common practice in SEM is to use multiple measurements as "multiple indicators" in the model.



USGS 17

Here is the model we are going to code in the next slide. Using multiple indicators for xi is one way to estimate and control for measurement error.

Specifying 2-indicator latent regression model in lavaan.

```
# specify model
mod.2 <- '
   # declare latent variables using "=~" operator
      xi  =~ lambda*x1 + lambda*x2
      eta =~ y

   # declare latent regression
      eta ~ xi'

# fit model
mod.2.fit <- sem(mod.2, sample.cov=input.cov2,
                 sample.nobs=100)

# request output
summary(mod.2.fit, standardized=T, rsq=T)
```

note we estimate a single lambda for both indicators to achieve identification.

USGS

18

There are several important things to be aware of here.

(1) Latent variable models with only 2 indicators are locally non-identified. To solve this problem, we can (a) ensure x1 and x2 have equal variances, in this case by standardizing the data. (2) When latent variables are included we must specify a fixed value for some parameter associated with the LV to achieve identification. The lavaan default is to set the loading from the LV to the first-mentioned indicator to 1.0. (3) For single-indicator LVs, the default measurement error is set to 0.0.

Lavaan results from model with multiple indicators.

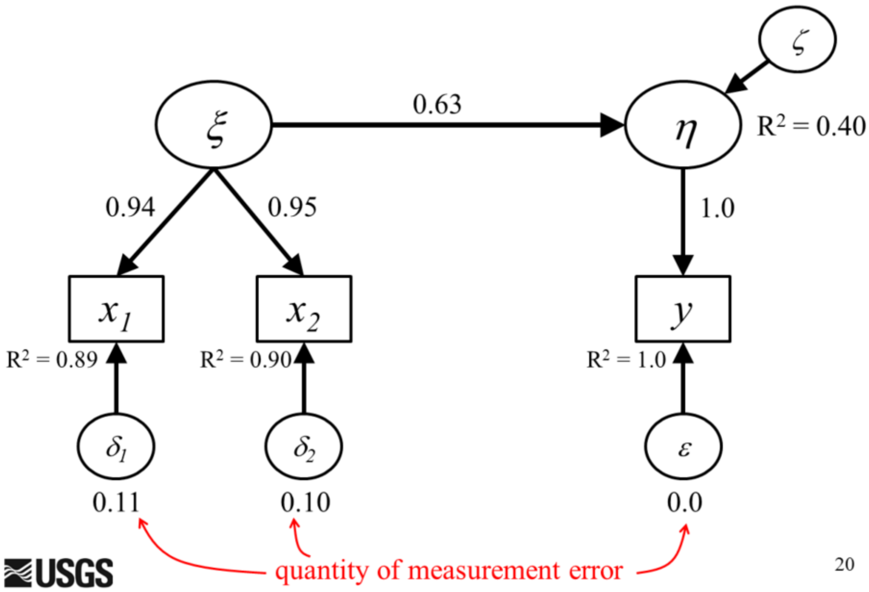
```
                    Estimate   Std.err   Z-value   P(>|z|)    Std.all
Latent variables:
  xi =~
    x1 (lmbd)         1.000                                     0.944
    x2 (lmbd)         1.000                                     0.949
  eta =~
    y                 1.000                                     1.000
Regressions:
  eta ~
    xi                0.667     0.086     7.734     0.000       0.631
Variances:
    x1                0.109     0.038                           0.109
    x2                0.098     0.038                           0.099
    y                 0.000                                     0.000
    xi                0.891     0.134                           1.000
    eta               0.599     0.088                           0.602
R-Square:
    x1                0.891
    x2                0.901
    y                 1.000
    eta               0.40
```

As expected, prediction is "better" when some of the unexplained variation is attributed to measurement error.

19

Here are the results for the latent regression, showing a greater R-square.

Results shown on graph.



20

Various results are summarized here on the graph.

How do we compute the quantity of measurement error?

In this example, reliability, = 0.90. This means COR(x_1,x_2) = λ_{x1}*λ_{x2} = 0.90.

In the previous slide we confirm that = λ_{x1}*λ_{x2} = 0.94*0.95 = 0.90.

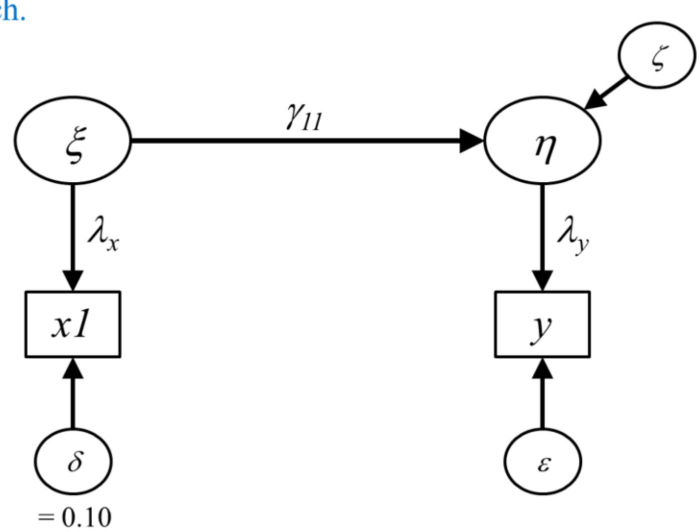
Standardized measurement error (θ) = 1- λ^2 = 0.10.
and,
Absolute Measurement Error = θ*VAR(x).

In this example we standardized x_i, so VAR(x) = 1.0 and absolute measurement error = 0.10.

USGS 21

It is useful to know how to compute measurement error and correct for it without using multiple indicators explicitly.

We can specify the measurement error using a single-indicator approach.



It is useful to be able to use general knowledge we have about measurement error or indicator reliability to correct for its effects in our models. We need to be careful with this practice, however, because specifying lots of measurement error can lead to model instability and questionable results.

The lavaan code for this model is given in the next slide.

We specify measurement error in lavaan using “~~”.

```
#lv model with error specified
mod3 <- '
   # declare latent variables
      xi =~ x1
      eta =~ y

   # declare latent regression
      eta ~ xi

   # specifying error variance for x
      x1 ~~ 0.10*x1'

mod.3.fit <- sem(mod.3, sample.cov=input.cov2,
                 sample.nobs=100)
```

variance for x is ‘x ~~ x’, we fix the value to 0.10 by premultiplying

USGS

23

In lavaan, we can tell the program how much measurement error we think we have for our x variable and it can adjust the estimates of parameters accordingly.

Here we are only specifying imperfect reliability for one indicator, x. We could also do the same for y. By not specifying measurement error for y, we are assuming perfect measurement.

Results adjusted for measurement error.

USGS

```
Latent variables:  Estimate  Std.err  z-value  P      Std.all
 xi =~
    x              1.000                              0.949
  eta =~
    y              1.000                              1.000

Regressions:
  eta ~
    xi             0.660     0.090    7.364    0.000  0.628

Variances:
    x              0.100                              0.100
    y              0.000                              0.000
    xi             0.900     0.141                    1.000
    eta            0.603     0.091                    0.606

R-Square:
    x                 0.900
    y                 1.000
    eta               0.394
```

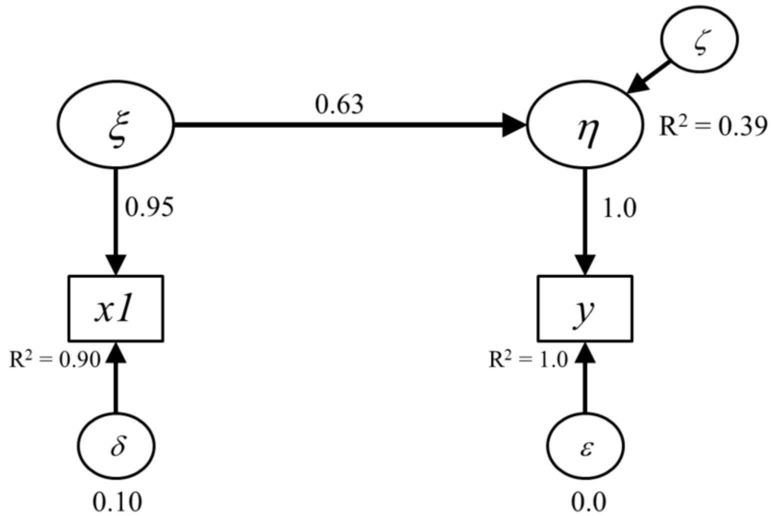
this is value we set

Minus rounding error, results same as for the 2-indicator model.

24

We get similar results as for the 2-indicator model, with the difference being attributable to rounding errors.

Single-indicator results presented graphically.



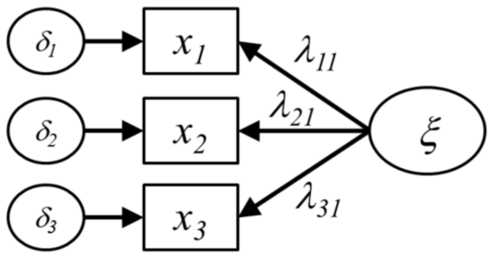
USGS

25

And here is the graphical representation for the model assuming 10% of the variance in x1 is due to measurement error.

The multi-indicator latent variable.

the hypothesis



the data

	x_1	x_2	x_3
x_1	**1.0**		
x_2	**0.80**	**1.0**	
x_3	**0.60**	**0.90**	**1.0**

This model hypothesizes that the correlations/covariances between x_1, x_2, and x_3 can all be explained by a single influence.

Lambdas will be selected that best resolve the three covariances.

There are an implied set of scores for ξ.

USGS

26

Now, a very common application in latent variable modeling involves the use of the "multi-indicator" latent variable. Here I just show the causal situation being modeled. The roots of this idea go back the early studies of human intelligence and its modern application to human studies is widespread.

Example of multi-indicator type model.

The Example: The general performance of transplanted plants as a function of their genetic dissimilarity to local populations.

from:

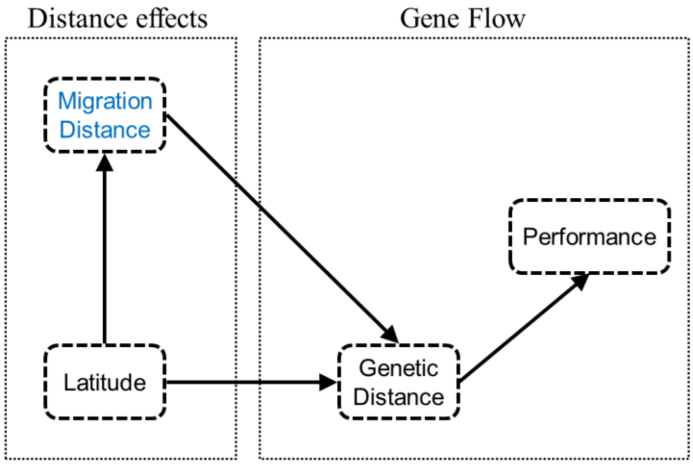
Travis, S.E. and Grace, J.B. 2010. Predicting performance for ecological restoration: a case study using *Spartina alterniflora*. *Ecological Applications* 20:192-204.

27

USGS

Now, here is a real example that employs latent variables in a restoration study.

Theory suggests the following for transplanted *Spartina*.



but, what do we mean by performance?

28

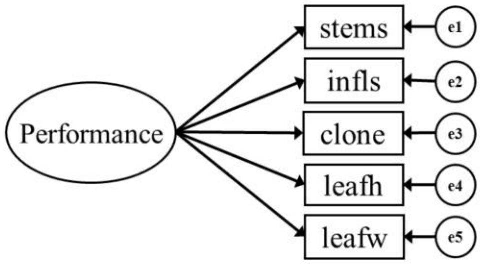
Here is our conceptual meta-model. Our example focuses on modeling "performance" as a generalize response, not one characterized by a single indicator.

“Performance” is a latent construct.

Word performance implies complex, intercorrelated response by many traits reflecting some underlying, unmeasured cause or causes.

Be aware that simply linking a bunch of measures to a latent variable does not mean you have correctly specified the model. You must justify causal assumptions.

Note this model hypothesizes we have five observed responses whose intercorrelations are consistent with a single underlying cause.



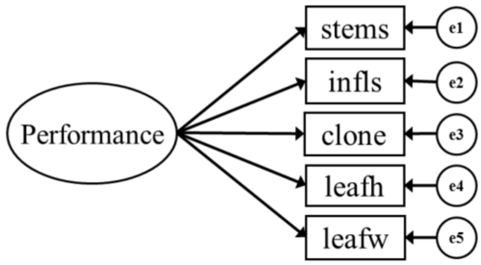
USGS

29

It is common that plant scientists will take several measures of plant properties, thinking that one may prove to be the most sensitive indicator of performance. We can use all the measures and confirmatory factor analysis to evaluate the latent relationships among variables.

“Performance” is a latent construct (cont.).

Examination of correlations among candidate indicators gives us notion of whether pattern is consistent with what is implied by our model.



Observed Correlations:

	stems	infls	clone	leafh	leafw
stems	1.00				
infls	0.93	1.00			
clone	0.81	0.83	1.00		
leafh	0.77	0.72	0.69	1.00	
leafw	0.73	0.64	0.60	0.96	1.00

USGS

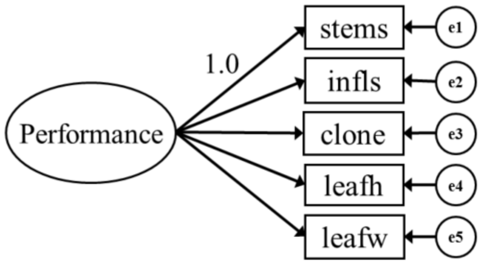
30

We ALWAYS need to look at the correlation structure of our data. If there really is a common latent factor, the observed variables should be consistently and uniformly correlated. Our data suggest the leaf height and leaf width are especially highly correlated, probably due to evolutionary constraints to morphology.

Specifying the “confirmatory factor model” (CFA).

1. Note when including a latent variable, we have increased the number of parameters to estimate and need to “fix” some parameters (specify their values).

2. Lavaan sets first loading = 1.0.




```
lvmod.1 <- '
     # Latent variable definition
       Perform=~ stems + infls + clonediam
               + leafht + leafwdth'
```

31

A first step is to analyze the “measurement model” using CFA.

Illustration of some possible warning messages.

```
# fit model

lvmod.1.fit <- sem(lvmod.1, data=perf.dat)
```

```
Warning message:
In lavaan(model = lvmod.1, data = perf.dat,
model.type = "sem",   :
  lavaan WARNING: some estimated variances are
negative
```

This may or may not be a problem for us. The question we have to consider next is, are there some estimated variances that are significantly negative.

USGS

We use the "sem" function here, but there is also a lavaan function "cfa" specifically for this type of model.

Here a common warning is encountered for this type of model.

Results.

```
lavaan (0.5-12) converged normally after  72 iterations

  Number of observations                            23

  Estimator                                         ML
  Minimum Function Test Statistic               51.106
  Degrees of freedom                                 5
  P-value (Chi-square)                           0.000
```

Model fit very poor!

USGS

33

Note poor fit.

Modification indices.

Several ways we can ask for modification indices etc.

```
modindices(lvmod.1.fit)  #this gives us everything

mi <- modindices(lvmod.1.fit) #create index object

print(mi[mi$op == "~",])   #request only ~ links

print(mi[mi$op == "~~",])  #request only ~~ links
```

USGS

34

Here is some code for selectively extracting modification indices.

Modification indices.

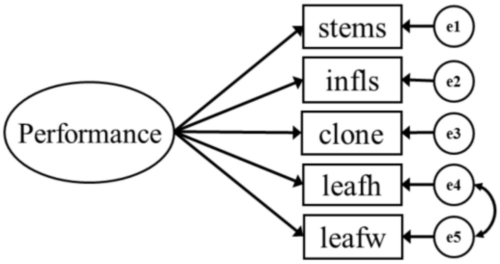
```
mi <- modindices(lvmod.1.fit) #create index object
print(mi[mi$op == "~~",])  #request only ~~ links
```

lhs	op	rhs	mi	epc	sepc.lv	sepc.all	epc.nox
stems	~~	stems	0.000	0.000	0.000	0.000	0.000
stems	~~	infls	10.470	11.784	11.784	0.341	0.341
stems	~~	clonediam	17.152	112.521	112.521	0.392	0.392
stems	~~	leafht	0.693	-7.889	-7.889	-0.035	-0.035
stems	~~	leafwdth	2.214	-1.836	-1.836	-0.062	-0.062
infls	~~	infls	0.000	0.000	0.000	0.000	0.000
infls	~~	clonediam	8.773	11.092	11.092	0.292	0.292
infls	~~	leafht	0.062	-0.312	-0.312	-0.010	-0.010
infls	~~	leafwdth	2.906	-0.281	-0.281	-0.072	-0.072
clonediam	~~	clonedia	0.000	0.000	0.000	0.000	0.000
clonediam	~~	leafht	4.028	-21.233	-21.233	-0.085	-0.085
clonediam	~~	leafwdth	0.037	-0.261	-0.261	-0.008	-0.008
leafht	~~	leafht	0.000	0.000	0.000	0.000	0.000
leafht	~~	leafwdth	37.863	[illegible]	[illegible]	[illegible]	[illegible]
leafwdth	~~	leafwdth	0.000	0.000	0.000	0.000	0.000
Perform	~~	Perform	0.000	0.000	0.000	0.000	0.000

One modification index is quite large.

Here I show the whole long list of stuff spit out by lavaan. We focus in on the largest mi (modification index value) and will incorporate a correlation between leaf height and width in our model (next slide).

Modified model with added error covariance.




```
lvmod.2 <- ' # Latent variable definition
             Perform=~ stems + infls + clonediam
                       + leafht + leafwdth

             # Error Covariances
               leafht ~~ leafwdth'
```

USGS

36

Now, we can include an error correlation/covariance as part of our model using the code shown in red.

Results for revised model.

```
lavaan (0.5-12) converged after  91 iterations

 Number of observations                          23

 Estimator                                     ML
 Minimum Function Chi-square                 7.40
 Degrees of freedom                          4
 P-value                                     0.116
```

Huge drop in discrepancy! Now model fit good (esp. for a lv model).

The significant drop in model chi-square (from 51.1 to 7.4) can serve as a formal test of the added link.
Or, you could do an AICc model comparison.

USGS

We found the basis for the observed model discrepancy.

Results for revised model (cont.).

	Estimate	Std.err	Z-value	P(>|z|)
Latent variables:				
Perform =~				
stems	1.000			
infls	0.117	0.016	7.173	0.000
clonediam	1.086	0.096	11.319	0.000
leafht	0.697	0.127	5.509	0.000
leafwdth	0.082	0.018	4.529	0.000
Covariances:				
leafht ~~				
leafwdth	10.831	3.432	3.156	0.002

USGS

38

Now here are some of the results. For more on this paper see

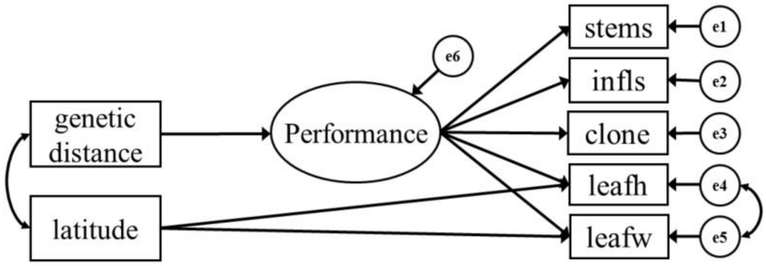
Travis, S.E. and Grace, J.B. 2010. Predicting performance for ecological restoration: a case study using *Spartina alterniflora*. *Ecological Applications* 20:192-204.

[selected as Recommended Reading by the Faculty of 1000: http://f1000biology.com/article/id/2305956/evaluation]

[featured in a Research Brief by Conservation Maven: http://www.conservationmaven.com/frontpage/predicting-the-performance-of-plant-restoration.html]

Putting performance into context in the full model.

Now we put performance into a broader context by evaluating its relationship to two driving factors, genetic distance and latitude. (simplification of full model)



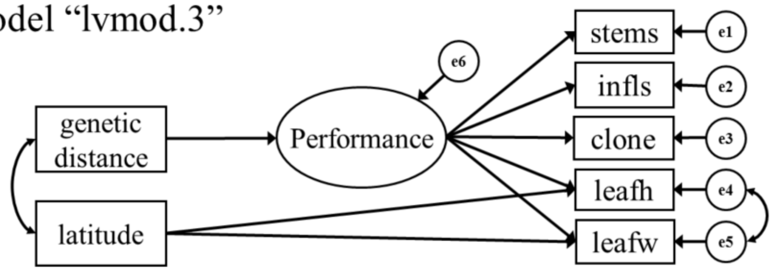
We have reason to believe based on past studies that leafht and lfwidth will respond directly to those climatic factors associated with latitude.

39

USGS

Here is a simplified version of the full model of interest in the study. The interest was in whether measures of genetic distance could be used to predict plant performance in a new location. Latitude was included as a control variable because it is known that the climatic differences found at different latitudes can also influence plant morphology.

Model “lvmod.3”

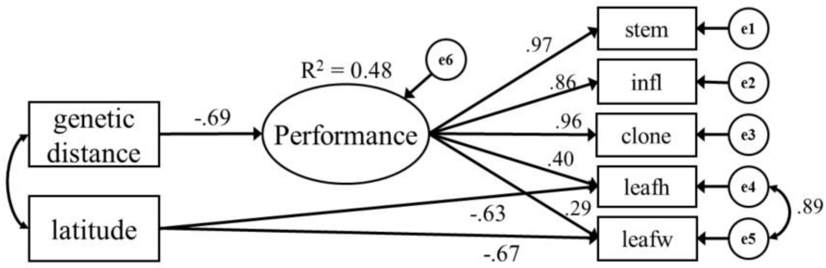



```
lvmod.3 <- ' # Latent variable definition
               Perform=~ stems + infls + clonediam
                        + leafht + leafwdth

             # Error Covariances
               leafht ~~ leafwdth

             # Regressions
               Perform ~ geneticdist
               leafht ~ latitude
               leafwdth ~ latitude'
```

40

Here the code for the responses of performance measures to genetic distance and latitude are shown in red.

Results and interpretation.



Leaf ht and width more related to latitudinal ecotype development than performance response.

chi-square = 19.523
df = 11
p = 0.052

A few results. For a more complete picture of the findings, see the Travis and Grace (2010) paper.

41

And here are key results.