USGS
science for a changing world

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. Questions about this material can be sent to sem@usgs.gov.

Last revised 15.09.11.

What is a latent variable?

“A variable for which we do not have measurements.”

How should we think about latent variables in models?

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

How should we think about latent variables?

Latent variables: General references

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)

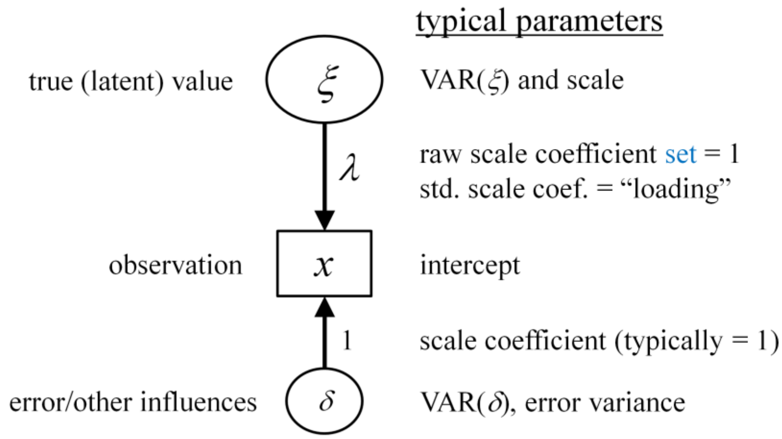
Bollen, K.A. 2012. Latent variables in structural equation modeling. Chapter 4, In: Hoyle, R.H. (ed.) Handbook of Structural Equation Modeling. Guilford Press, New York.

USGS

3

Some references that make key distinctions and provide diagnostic criteria.

The single-indicator LV block

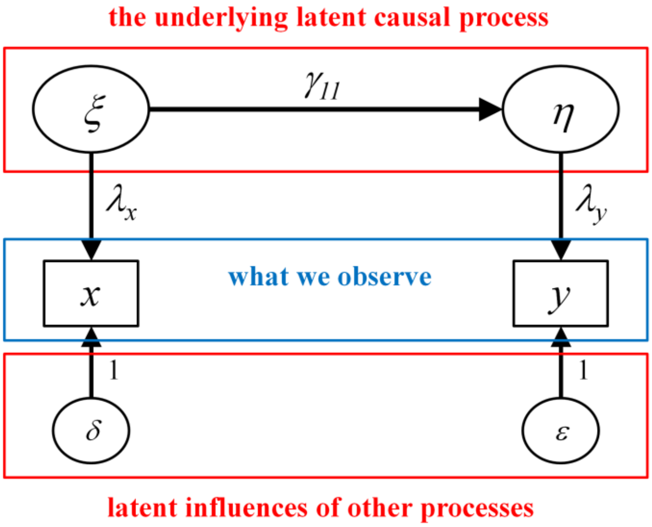


USGS

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 regression



USGS

5

Causation is from latent to observed variables (typically).

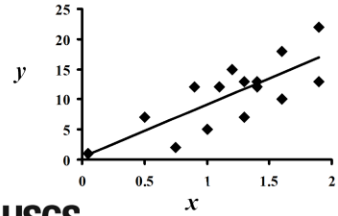
One reason to use latent variables is to address measurement error.

Observed variable models assume all variables are measured without error.*

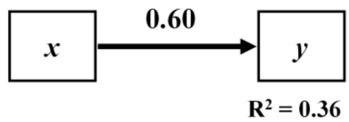
(*This applies to all classical statistical models, as well as to observed variable SE models.)

So, what difference does it make?

Imagine we observe this.



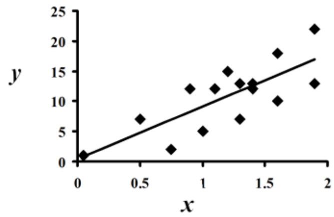
The regression / SE relationship would be.

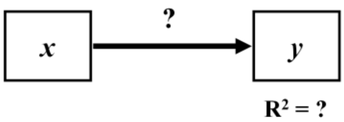


USGS

The issue of measurement error and its effects is virtually ignored in most statistical training, though that is starting to change.

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

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

We can estimate measurement error by hand.

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	x-trial3
1	1.272	1.206	1.281
2	1.604	1.577	1.671
3	2.177	2.192	2.104
4	1.983	2.080	1.999
.	.	.	.
n	2.460	2.266	2.418

If, average correlation between trials = 0.90,

then, the average **reliability** of any given set of measurements is: r = 0.90, the average correlation between any two sets of measurements across the sample.

8

Indicator reliability is a key concept.

How to compute measurement error.

Measurement Error Variance = $(1 - r^2)$ times the variance of x

So, if reliability, r, = 0.90, then

Standardized Measurement Error is $(1 - r^2) = 0.19$

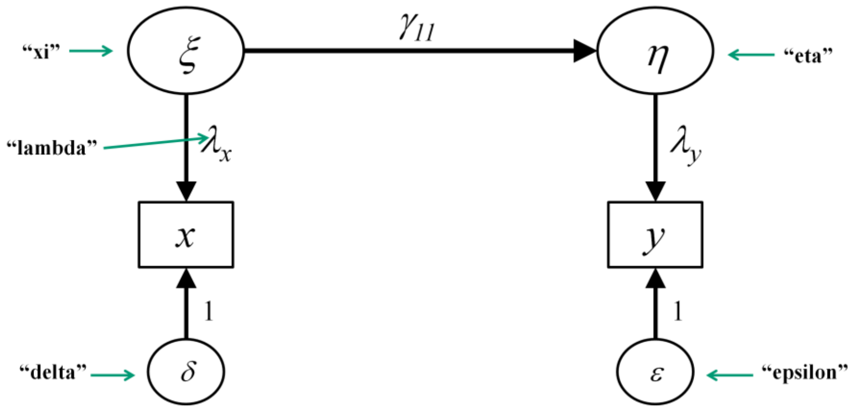
and, Absolute Measurement Error = 0.19*VAR(x)

Imagine VAR(x) = 3.14,
Absolute Measurement Error Variance = 0.19 x 3.14 = 0.597

USGS

It is useful to know how to compute measurement error.

Ok, here is our model.



USGS

10

Here is the model we are going to code in the next slide.

Specifying measurement error in lavaan

```
# lv model with error specified
lv.mod2 <- '
   # declare latent variables
      xi =~ x
      eta =~ y

   # declare latent regression
      eta ~ xi

   # specifying error variance for x
      x ~~ 0.597*x'

# fit model
lv.fit2 <- sem(lv.mod2, sample.cov= mod1.cov,
sample.nobs= 15)
```

variance for x is ‘x ~~ x’, we fix the value to 0.597 by premultiplying

USGS

11

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.

Results adjusted for measurement error

Not the same results as for observed variable model.

```
Latent variables:  Estimate Std.err  z-value  P       Std.all
 xi =~
    x              1.000                              0.892
  eta =~
    y              1.000                              1.000

Regressions:
  eta ~
    xi             0.426    0.152    2.808    0.005   0.673

Variances:
    x              0.597                              0.204
    y              0.000                              0.000
    xi             2.334    1.070                     1.000
    eta            0.510    0.226                     0.547

R-Square:
    x                 0.796
    y                 1.000
    eta               0.453
```

std beta greater than 0.60

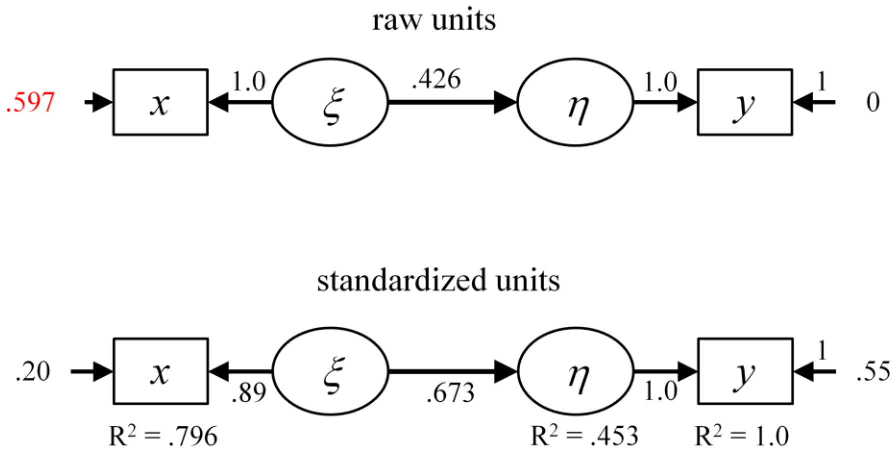
here is the error we specified

R-square est now higher

12

The results are different now.

Results expressed graphically



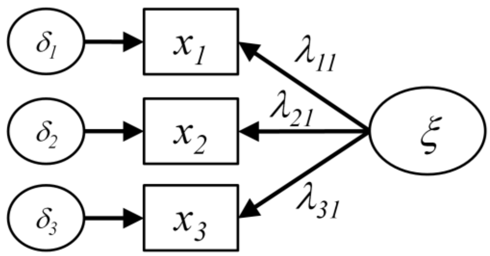
USGS

13

Here they are graphically.

The multi-indicator latent variable – Confirmatory Factor Analysis

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

14

Now, a very common application in latent variable modeling is the "multi-indicator" latent variable. Here I just show the causal situation being modeled.

USGS

Example of multi-indicator type model

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

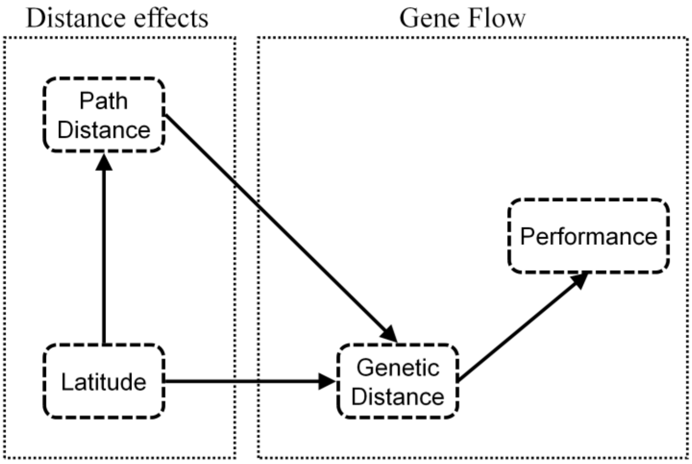
from:

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

15

Now, here is a real example.

Theory suggests the following for transplanted *Spartina*.



but, what do we mean by performance?

![USGS]

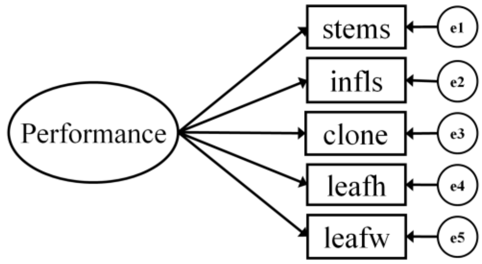
Here is our conceptual meta-model. Our example focuses on modeling "performance" as a generalize response, not one characterized by a single indicator.

“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.



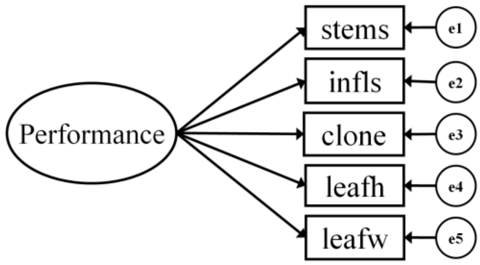
USGS

17

Again, note the direction of cause and effect being specified

“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

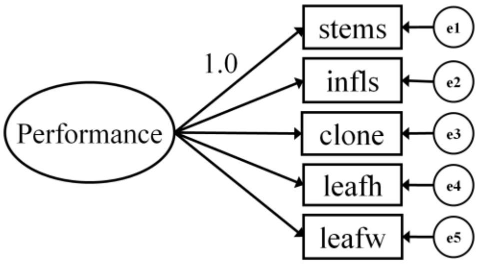
18

We ALWAYS need to look at the correlation structure of our data.

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'
```

USGS

19

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

Here is a common warning encountered.

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

21

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

# only values great than 3
print(mi1[([mi1$mi > 3.0,] & [!(mi1$mi=="<NA>"),])])
```

USGS

22

Here is some code for selectively extracting modification indices. Note blue part is new addition to the slide.

Modification indices

USGS

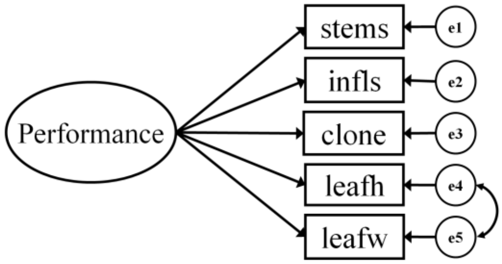
```
mi <- modindices(lvmod.1.fit) #create index object
print(mi[mi$op == "~~",])  #request only ~~ links
```

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.

Modified model with added error covariance.




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

             # Error Covariances
               leafht ~~ leafwdth'
```

USGS

24

Now we can include an error correlation/covariance as part of our model.

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

That was the basis for our discrepancy.

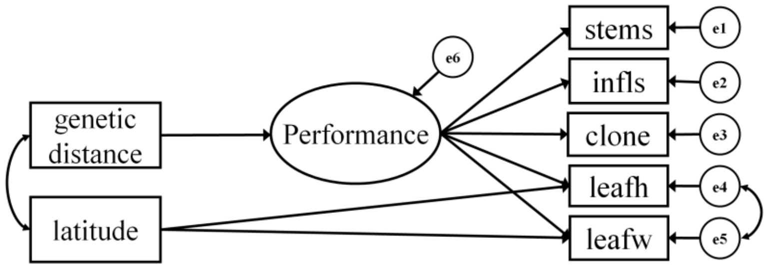
Results for revised model (cont.)

```
                       Estimate   Std.err Z-value  P(>|z|)   Std.all
Latent variables:
 Perform =~
    stems                 1.000                                0.970
    infls                 0.117     0.016    7.173    0.000    0.858
    clonediam             1.086     0.096   11.319    0.000    0.960
    leafht                0.697     0.127    5.509    0.000    0.776
    leafwdth              0.082     0.018    4.529    0.000    0.705

Covariances:
   leafht ~~
    leafwdth             10.831     3.432    3.156    0.002    0.943

R-Square:
    stems                 0.942
    infls                 0.736
    clonediam             0.921
    leafht                0.603
    leafwdth              0.497
```


Now here are some of the results for the measurement model. While not definitive, the p-values suggest all the parameters in the model are importantly different from zero. It is rare that p-values this small are associated with ignorable relationships (except at very large sample sizes).

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.

27

USGS

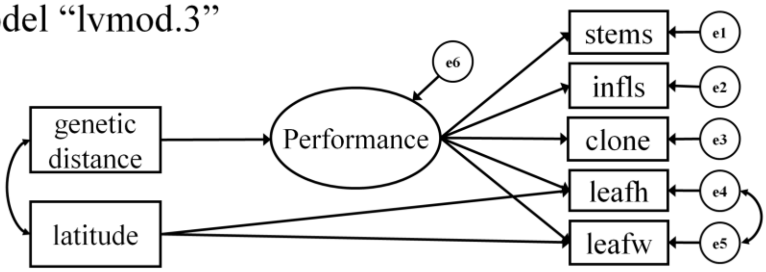
While this tutorial has focused on the modeling of performance as a general, latent factor, here I show more of the full ecological model, which includes the effects of genetic distance on performance and the effects of latitude as a predictor of specific leaf traits associated with ecotypic differentiation. For a more on this study, 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]

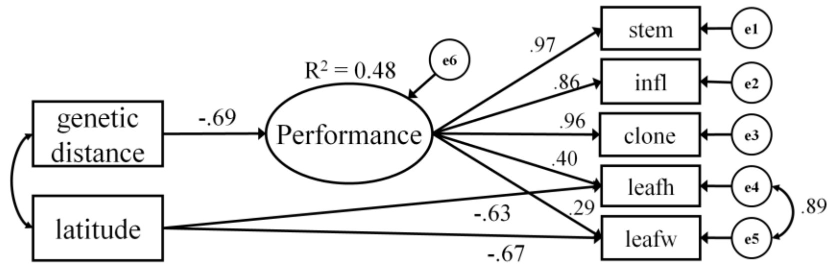
Model “lvmod.3”




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

             # Error Covariances
               leafht ~~ leafwdth

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

28

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

USGS

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

More information can be found at http://www.nwrc.usgs.gov/SEM

USGS

I hope this overview has been useful. For more information, go to our webpage or search for examples involving your subject of interest. Questions and comments can be sent to sem@usgs.gov. Please note I cannot guarantee responses to individual inquiries, but will try to incorporate suggestions in future tutorials. – Thanks!