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

An appropriate general citation for this material is

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.
Latent variables: General references


Some references that make key distinctions and provide diagnostic criteria.
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.
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.

![Graph showing regression analysis with x and y variables, and R^2 = 0.36]

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.

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.

<table>
<thead>
<tr>
<th>plot</th>
<th>x-trial1</th>
<th>x-trial2</th>
<th>x-trial3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.272</td>
<td>1.206</td>
<td>1.281</td>
</tr>
<tr>
<td>2</td>
<td>1.604</td>
<td>1.577</td>
<td>1.671</td>
</tr>
<tr>
<td>3</td>
<td>2.177</td>
<td>2.192</td>
<td>2.104</td>
</tr>
<tr>
<td>4</td>
<td>1.983</td>
<td>2.080</td>
<td>1.999</td>
</tr>
<tr>
<td>n</td>
<td>2.460</td>
<td>2.266</td>
<td>2.418</td>
</tr>
</tbody>
</table>

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.

Indicator reliability is a key concept.
It is useful to know 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
Ok, here is our model.

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

```
# 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)
```
The results are different now.
Here they are graphically.

Results expressed graphically

**raw units**

\[ x \overset{.597}{\rightarrow} \xi \overset{.426}{\rightarrow} \eta \overset{1.0}{\rightarrow} y \overset{1}{\rightarrow} 0 \]

**standardized units**

\[ x \overset{.20}{\rightarrow} \xi \overset{.89}{\rightarrow} \eta \overset{.673}{\rightarrow} y \overset{1}{\rightarrow} \]

R\textsuperscript{2} = .796

R\textsuperscript{2} = .453

R\textsuperscript{2} = 1.0

.55
Now, a very common application in latent variable modeling is the “multi-indicator” latent variable. Here I just show the causal situation being modeled.

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 $\xi$.  

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.80</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.60</td>
<td>0.90</td>
<td>1.0</td>
</tr>
</tbody>
</table>
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:


Now, here is a real example.
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.

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.

<table>
<thead>
<tr>
<th>Observed Correlations:</th>
<th>stems</th>
<th>infls</th>
<th>clone</th>
<th>leafh</th>
<th>leafw</th>
</tr>
</thead>
<tbody>
<tr>
<td>stems</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>infls</td>
<td>0.93</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clone</td>
<td>0.81</td>
<td>0.83</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>leafh</td>
<td>0.77</td>
<td>0.72</td>
<td>0.69</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>leafw</td>
<td>0.73</td>
<td>0.64</td>
<td>0.60</td>
<td>0.96</td>
<td>1.00</td>
</tr>
</tbody>
</table>

We ALWAYS need to look at the correlation structure of our data.
A first step is to analyze the “measurement model” using CFA.

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

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.
Illustration of some possible warning messages

```r
# 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.

Here is a common warning encountered.
Results

lavaan (0.5-12) converged normally after 72 iterations

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>23</td>
</tr>
<tr>
<td>Estimator</td>
<td>ML</td>
</tr>
<tr>
<td>Minimum Function Test Statistic</td>
<td>51.106</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>5</td>
</tr>
<tr>
<td>P-value (Chi-square)</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Model fit very poor!

Note poor fit.
Modification indices

Several ways we can ask for modification indices etc.

```r
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>"),)])
```

Here is some code for selectively extracting modification indices. Note blue part is new addition to the slide.
Here I show the whole long list of stuff spit out by lavaan. We focus in on the largest mi.

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

<table>
<thead>
<tr>
<th>lhs</th>
<th>op</th>
<th>rhs</th>
<th>mi</th>
<th>epc</th>
<th>sepc.lv</th>
<th>sepc.all</th>
<th>epc.nox</th>
</tr>
</thead>
<tbody>
<tr>
<td>stems</td>
<td>~</td>
<td>stems</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>stems</td>
<td>~</td>
<td>infls</td>
<td>10.470</td>
<td>11.784</td>
<td>11.784</td>
<td>0.341</td>
<td>0.341</td>
</tr>
<tr>
<td>stems</td>
<td>~</td>
<td>clonediam</td>
<td>17.152</td>
<td>112.521</td>
<td>112.521</td>
<td>0.392</td>
<td>0.392</td>
</tr>
<tr>
<td>stems</td>
<td>~</td>
<td>leafht</td>
<td>0.693</td>
<td>-7.889</td>
<td>-7.889</td>
<td>-0.035</td>
<td>-0.035</td>
</tr>
<tr>
<td>stems</td>
<td>~</td>
<td>leafwidth</td>
<td>2.214</td>
<td>-1.836</td>
<td>-1.836</td>
<td>-0.062</td>
<td>-0.062</td>
</tr>
<tr>
<td>infls</td>
<td>~</td>
<td>infls</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>infls</td>
<td>~</td>
<td>clonediam</td>
<td>8.773</td>
<td>11.092</td>
<td>11.092</td>
<td>0.292</td>
<td>0.292</td>
</tr>
<tr>
<td>infls</td>
<td>~</td>
<td>leafht</td>
<td>0.062</td>
<td>-0.312</td>
<td>-0.312</td>
<td>-0.010</td>
<td>-0.010</td>
</tr>
<tr>
<td>infls</td>
<td>~</td>
<td>leafwidth</td>
<td>2.906</td>
<td>-0.281</td>
<td>-0.281</td>
<td>-0.072</td>
<td>-0.072</td>
</tr>
<tr>
<td>clonediam</td>
<td>~</td>
<td>clonediam</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>clonediam</td>
<td>~</td>
<td>leafht</td>
<td>4.028</td>
<td>-21.233</td>
<td>-21.233</td>
<td>-0.085</td>
<td>-0.085</td>
</tr>
<tr>
<td>clonediam</td>
<td>~</td>
<td>leafwidth</td>
<td>0.037</td>
<td>-0.261</td>
<td>-0.261</td>
<td>-0.008</td>
<td>-0.008</td>
</tr>
<tr>
<td>leafht</td>
<td>~</td>
<td>leafht</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>leafht</td>
<td>~</td>
<td>leafwidth</td>
<td>37.863</td>
<td>One modification index is quite large.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Now we can include an error correlation/covariance as part of our model.

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

  # Error Covariances
  leafht ~~ leafwidth'
```
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.

That was the basis for our discrepancy.
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).
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


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

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