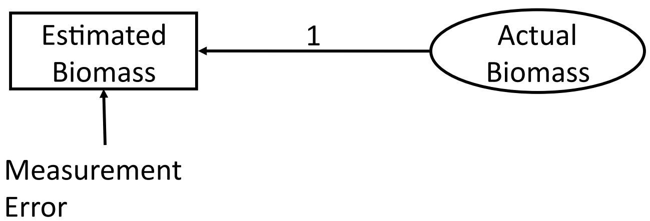
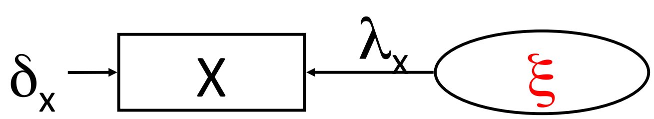
Latent Variables



Latent Variables

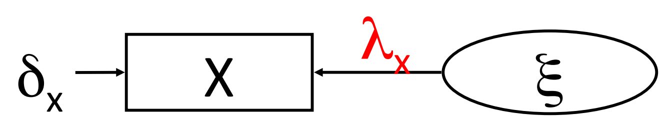
1. What is a latent variable?
2. Latent variables with multiple indicators
3. Fitting a latent variable
4. Factor Analysis
5. Latent Variables as a Response
6. Coping with measurement error

What is a latent variable?



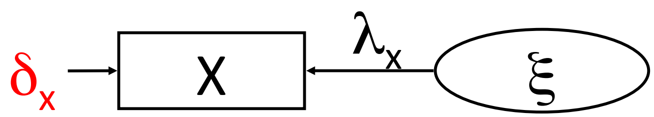
ξ : A latent variable is a variable that is unmeasured, but is hypothesized to exist

What is a latent variable?



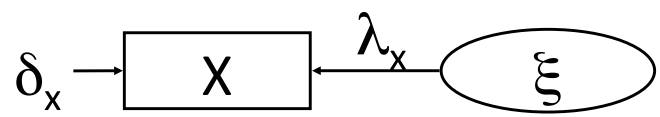
λ_x : The relationship between a latent variable and its observed counterpart

What is a latent variable?



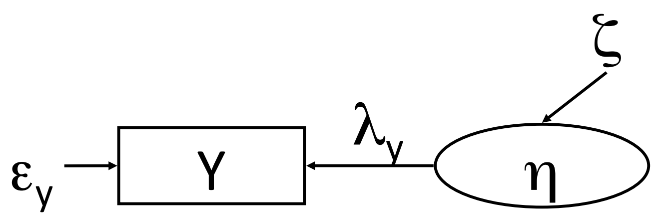
δ_x : The error in the measurement of x by ξ

Latent Exogenous Variables



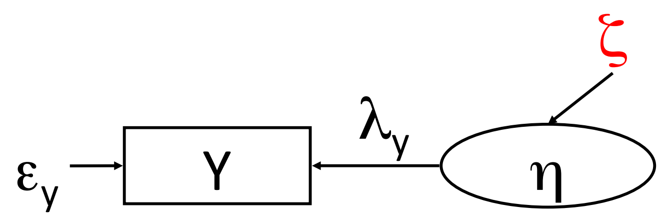
$$X = \lambda_x \xi + \delta_x$$

Latent Endogenous Variables



$$y = \lambda_y \eta + \varepsilon_y$$

Latent Endogenous Variables

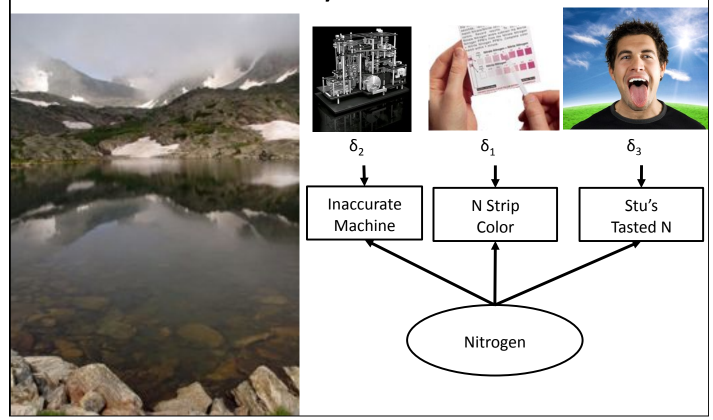


ζ: Variance in response to predictors

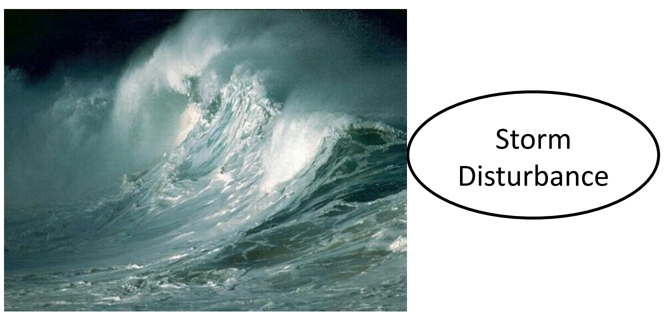
Latent Variables

1. What is a latent variable?
2. Latent variables with multiple indicators
3. Fitting a latent variable
4. Factor Analysis
5. Latent Variables as a Response
6. Coping with measurement error

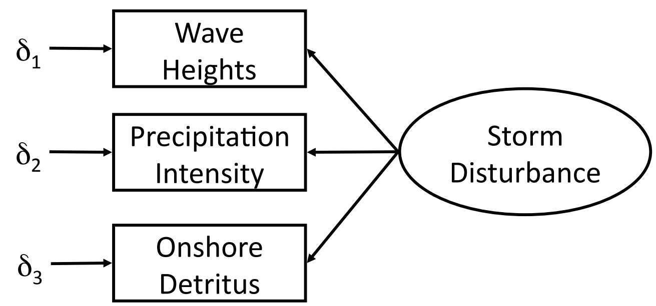
What if One Measurement Alone isn't Very Good?



Latent Variables as Theoretical Constructs

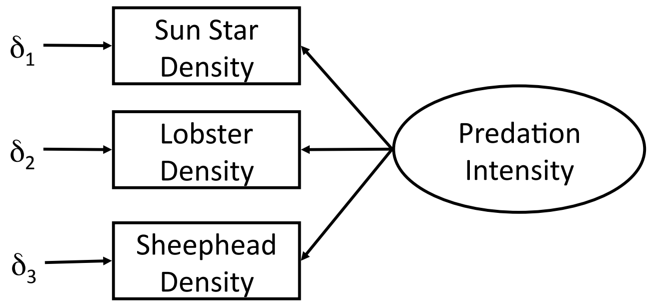


Latent Variables as Theoretical Constructs



δ' s are variation in observations not explained by storm disturbance

Concept Validity: The Fallacy of Naming



δ' s are variation in observations not explained by predation intensity???

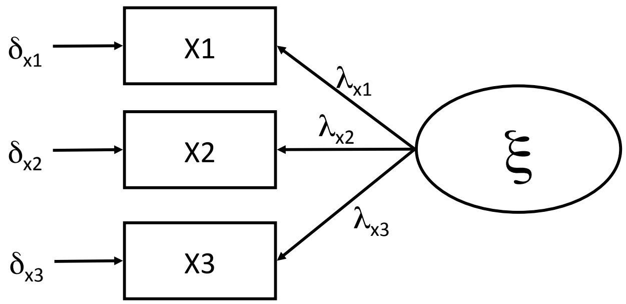
"The skepticism regarding 'latent variables' among many statisticians can probably be attributed to the metaphysical status of hypothetical constructs. On the other hand ... the concept of a 'good statistician' is not real, but nevertheless useful …"

- Skrondal and Rabe-Hesketh

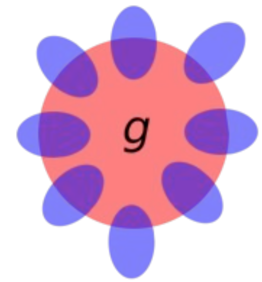
Latent Variables

1. What is a latent variable?
2. Latent variables with multiple indicators
3. Factor Analysis
4. Latent Variables as a Response
5. Coping with measurement error

Latent Variables with Multiple Indicators

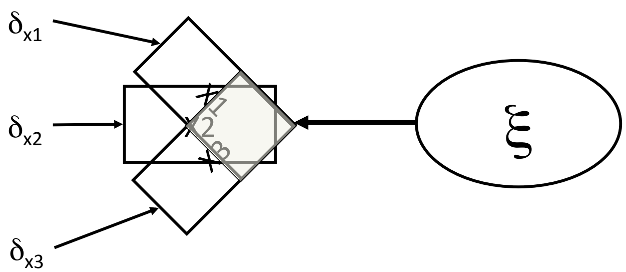


Latent variables represents shared information of indicators

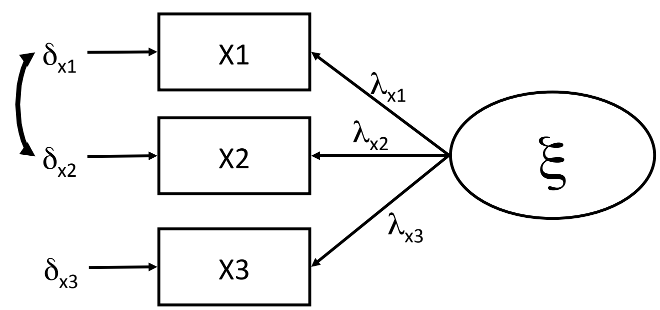


Common conceptual diagram of Spearman's analysis of "G-Theory," the idea of a generalized intelligence factor underlying test performance. Note shared variance of tests indicate "g."

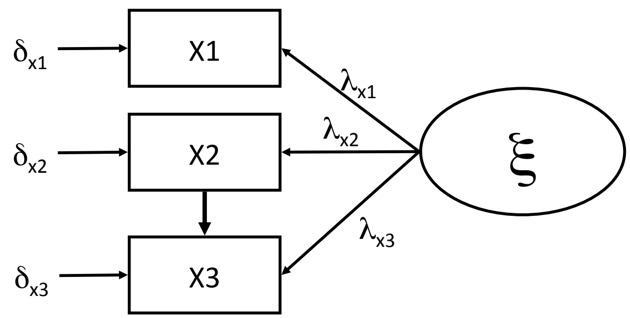
Latent variables represents shared information of indicators



Indicators may Covary for Other Reasons

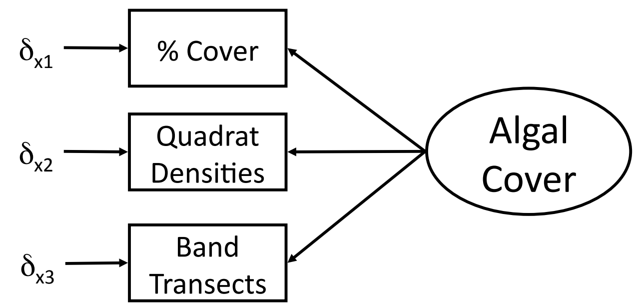


Indicators may Covary for Causal Reasons

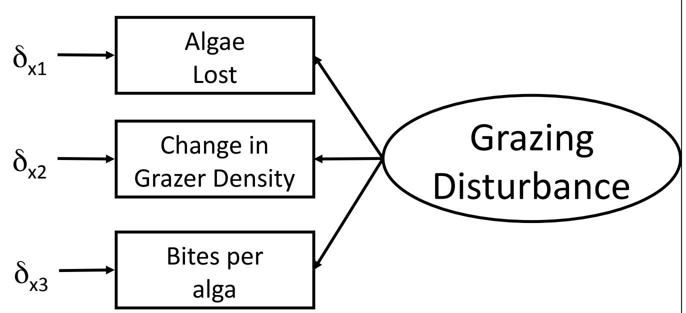


Maybe respecify your model?

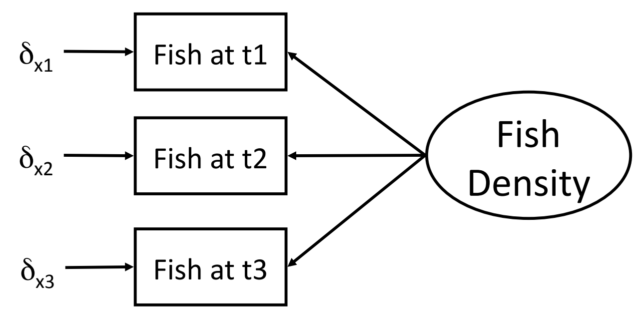
Different Measurements



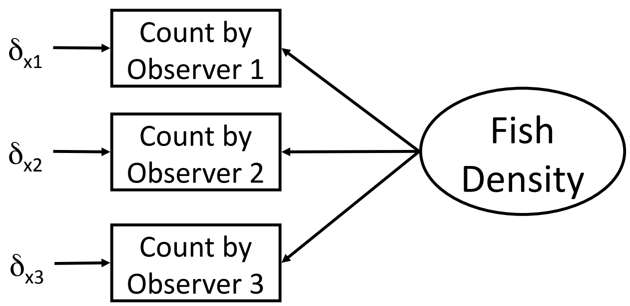
Multiple Properties



Repeated Measures



Multi-Sample



Evaluating Whether Indicators Will Make a Good Latent Variable

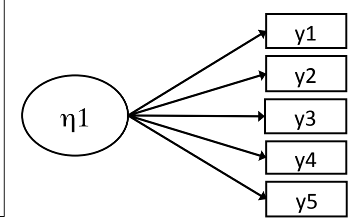
Observed Correlations:

	y1	y2	y3	y4	y5
y1	1.000				
y2	0.933	1.000			
y3	0.813	0.834	1.000		
y4	0.773	0.728	0.693	1.000	
y5	0.730	0.646	0.603	0.969	1.000

(1) Correlations among candidate indicators tell us whether data is consistent with what is implied by our model.

(2) Note correlations are all strong, but not all equally strong. This shows us that these are not redundant indicators that are completely interchangeable.

(3) In particular, variables y4 and y5 are more strongly correlated with each other than with the other vars.



Fixing Parameters for Identifiability

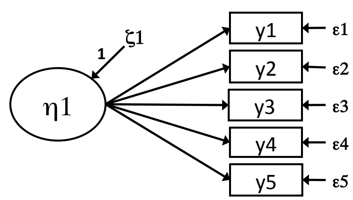
(1) Note we need to "fix" some parameters (specify their values) for identifiability.

(2) In this case, I chose to set variance of latent variable = 1.0.

(3) The other choice would be to fix one of the path coefficients to 1.0.

(4) Fixing a loading to 1 puts the latent variable on the scale of that indicator.

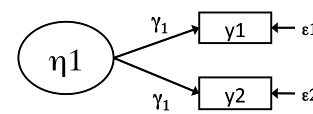
(5) Test model with different paths fixed to 1 to ensure that your latent variable is good



Latent Variable with Two Indicators

1. Problem - we have only one piece of information about y1 and y2 - their correlation (= 0.933).
2. Model has two path coefficients, plus the variance of our latent variable.
3. We can fix the value of our LV to 1, but that still leaves us with one know and two unknowns.

NOT IDENTIFIED.



One Solution: when there are only two indicators, they have equal weight in the estimation of the LV (absent other information).

So, we can standardize the two measures, and only estimate a single parameter for both paths.

Why Use Latent Variables with Multiple Indicators?

1. Better accuracy in measurement of relationships due to shared variation between indicators.
2. You cannot measure a theoretical construct!

Latent Variables

1. What is a latent variable?
2. Latent variables with multiple indicators
3. Fitting a latent variable
4. Factor Analysis
5. Latent Variables as a Response
6. Coping with measurement error

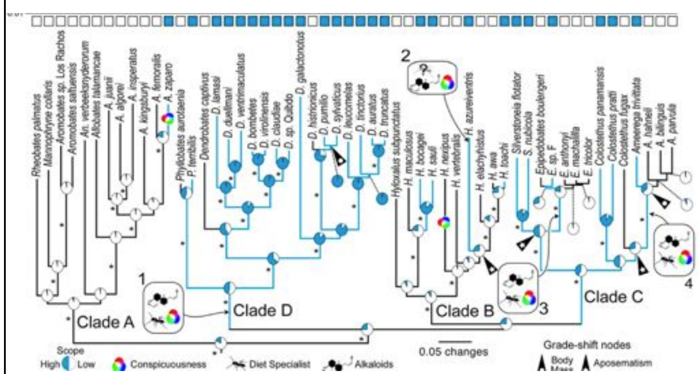
Example: Aposematism in Poison Dart Frogs

Santos, J.C. & Cannatella, D.C. (2011). Phenotypic integration emerges from aposematism and scale in poison frogs. *Proc. Natl. Acad. Sci. U.S.A., 108, 6175–6180.*

What drives the evolution of warning coloration?

Toxicity? **Diet?** **Body condition?**

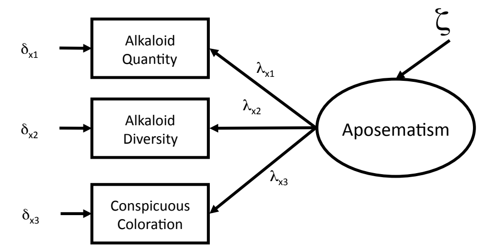
A Phylogenetic Approach to SEM using Independent Contrasts



Santos and Cannatella 2011 PNAS

Aposematism as a Latent Variable

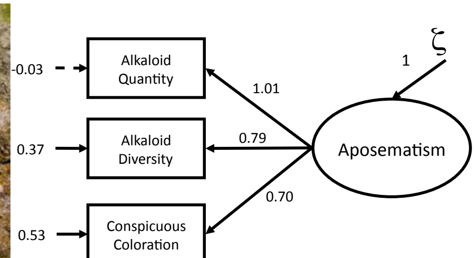

Santos & Cannatella 2011 PNAS




```
santosCov <-read.table("santosCov.txt", na.strings=".")
santosCov <-as.matrix(santosCov)

santosCFA1<-'
     Aposematism =~ Alkaloid.quantity + Alkaloid.diversity +
                    Conspicuous.coloration'

santosFit1<-sem(santosCFA1, sample.cov=santoCov, sample.nobs=21)
```

Aposematism as a Latent Variable

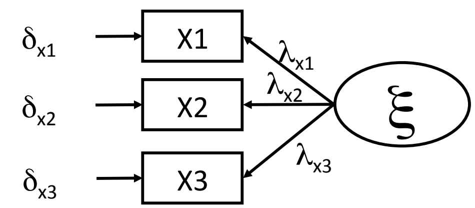

Santos & Cannatella 2011 PNAS




```
                     Estimate  Std.err  Z-value  P(>|z|)
Latent variables:
  Aposematism =~
    Alkaloid.quan       1.000
    Alkaloid.dive       2.358    0.558    4.223    0.000
    Conspicuous.c       0.753    0.215    3.502    0.000

Variances:
    Alkaloid.quan      -0.000    0.000   -0.190    0.849
    Alkaloid.dive       0.004    0.002    2.376    0.018
    Conspicuous.c       0.001    0.000    2.905    0.004
    Aposematism         0.001    0.000    2.915    0.004
```

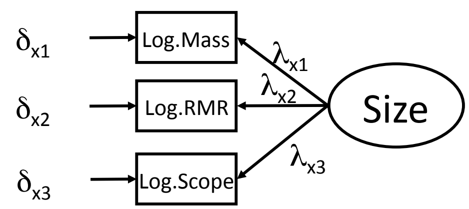
Haywood case: negative estimate of variance. Can indicate overfit indicator. All within rounding error here, as indicator was fixed, and variance not different from 0

Exercise: Fitting Latent Variables



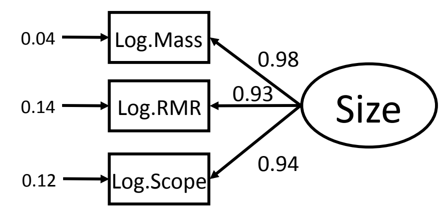
- The Santos covariance matrix has many other variables related to frog diet and frog size – try out 'body size' as a latent variable

Exercise: Fitting Latent Variables




```
santosSize<-'
  Size =~ Log.Mass + Log.RMR + Log.Scope'

santosSizeFit<-sem(santosSize,
sample.co=santosCov, sample.nobs=21)
```

Exercise: Fitting Latent Variables



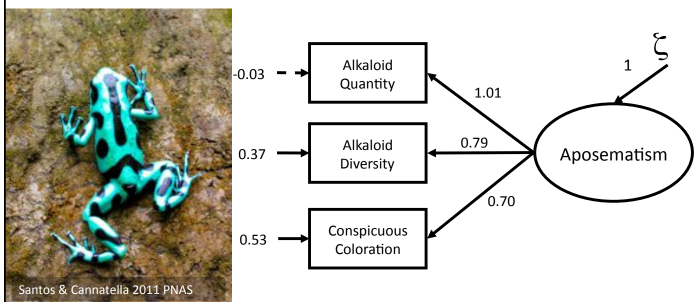
	Estimate	Std.err	Z-value	P(>|z|)	Std.lv	Std.all
Latent variables:						
Size =~						
Log.Mass	1.000				0.096	0.981
Log.RMR	0.815	0.083	9.771	0.000	0.078	0.930
Log.Scope	0.861	0.084	10.228	0.000	0.082	0.938

Questions?

Latent Variables

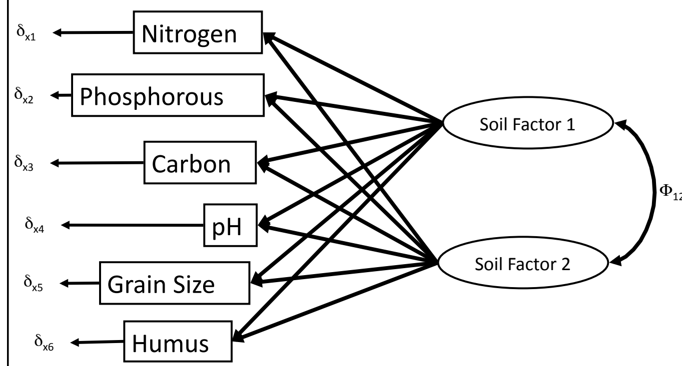
1. What is a latent variable?
2. Latent variables with multiple indicators
3. Fitting a latent variable
4. Factor Analysis
5. Latent Variables as a Response
6. Coping with measurement error

Example: Phylogenetic CFA!



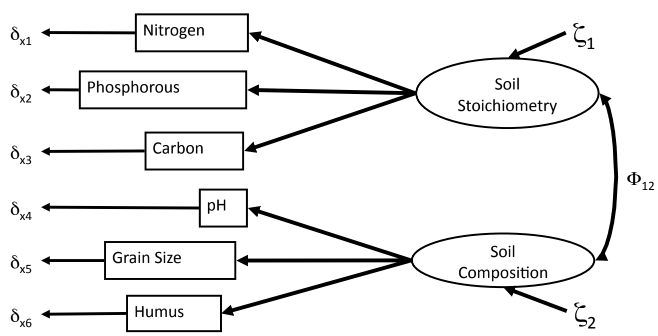
(Confirmatory Factor Analysis)

Exploratory Factor Analysis



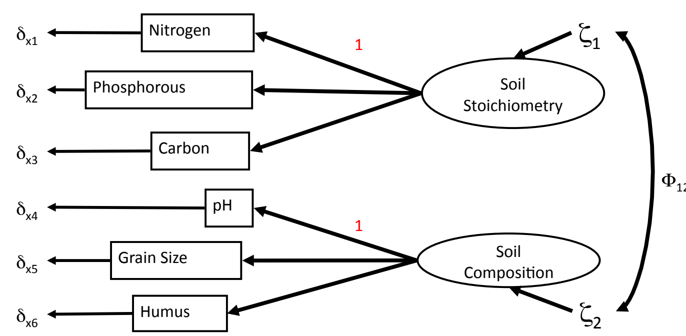
- Observed variables explained by correlated factors
- Error in observed variables
- But…exploratory - NOT HYPOTHESIS TESTING.

Confirmatory Factor Analysis



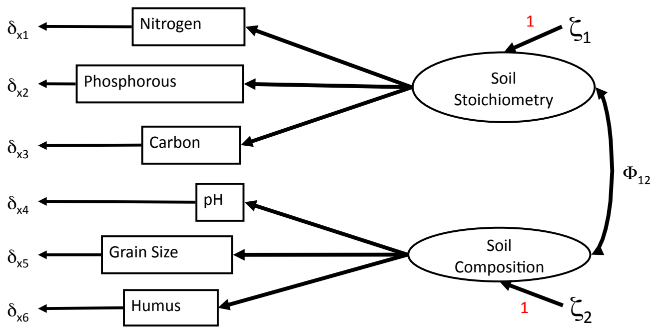
- Tests whether variables separate into groups
- Useful to test a "measurement model" for SEM
- Φ_{12} represents common variation due to other factors

Identification: Fixing Scale for a Standardized Model



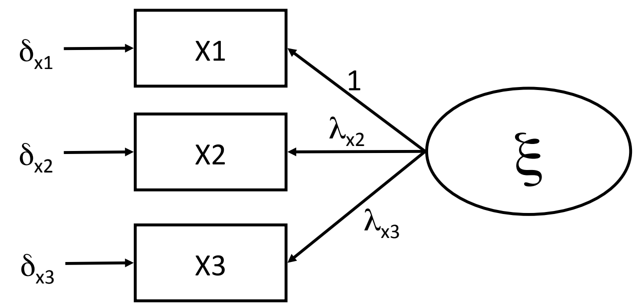
- In most cases, we need to provide a scale for our latent variables.
- Test that results don't change if you change scale.

Identification: Fixing Variance for an Unstandardized Model



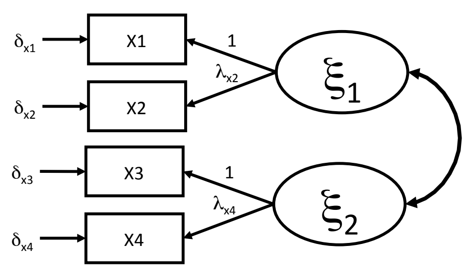
- Check to make sure method of identification doesn't alter results of CFA
- Standardized or Unstandardized approach NECESSARY for identification

Identification: All Latent Variables have Three Indicators without Correlated Error



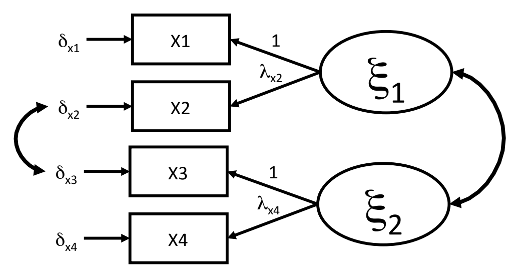
Three Indicator Rule - SUFFICIENT

Identification: Two Indicators per Latent, Multiple Latents, Uncorrelated Indicator Variance



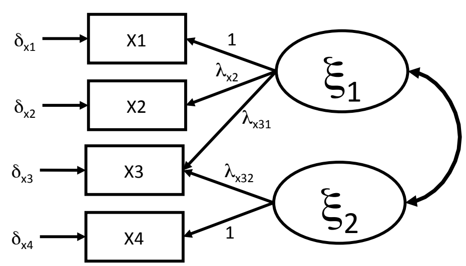
Two Indicator Rule - SUFFICIENT

Identification: If Some Indicators Covary, Each Must Have At Least One Uncorrelated Indicator



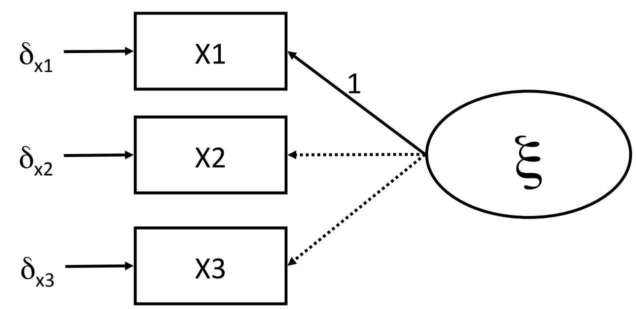
Correlated Indicator Rule - NECESSARY

Identification: If Indicators Shared, Each Latent Needs One Unique Indicator



Shared Indicator Rule - NECESSARY

Empirical Underidentification Still Possible

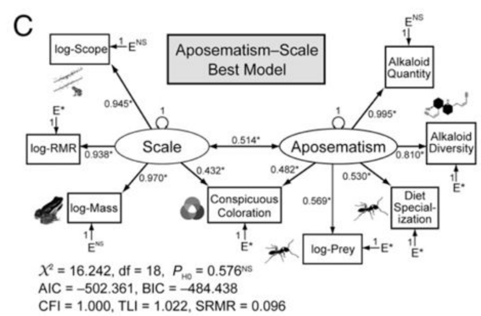


General Rules for Identification

1. T-rule still holds – necessary
2. Standardization - necessary
3. Three indicator rule – sufficient
4. Two Indicator rule – sufficient
5. Correlated Indicator rule – necessary
6. Shared Indicator Rule - necessary

N.B. None of these are both necessary and sufficient!

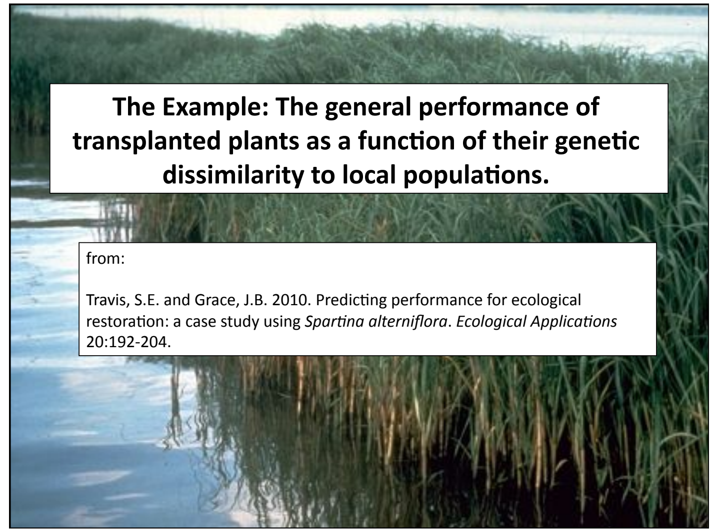
Exercise: Phylogenetic CFA!




```
santosCFA2<-paste(santosCFA1,
'
   Aposematism =~ Ant.Mite.Specialization+log.Prey

   Scale =~ Log.Mass+Log.RMR+Log.Scope+Conspicuous.coloration',

   sep="\n")
```

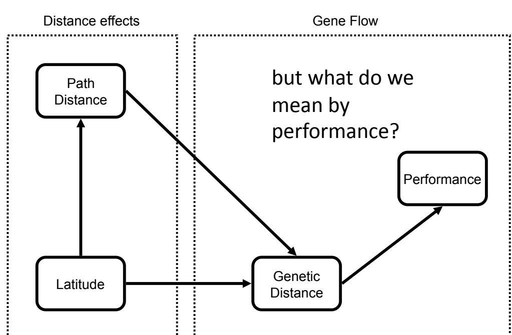
Latent Variables

1. What is a latent variable?
2. Latent variables with multiple indicators
3. Fitting a latent variable
4. Factor Analysis
5. Latent Variables as a Response
6. Coping with measurement error



The Theory Driving the Modeling

Theory suggests following for transplanted *Spartina*.



Performance as a latent construct

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

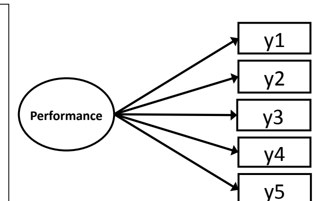
Be aware that simply linking a bunch of measures to a latent variable does not mean you have correctly specified the model.

One needs to stay focused on the question of "what actually causes the responses I see?". This is how you arrive at proper model specification, not simply through applying the model structures to your data.

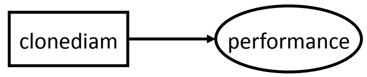
Note this model hypothesizes we have five observed responses whose intercorrelations are consistent with a single underlying cause.

There may be other things that influence y1-y5 and affect their observed intercorrelations.

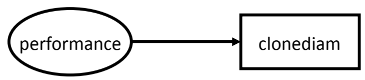
Also, performance is represented as a unidimensional cause (not a collection of disparate causes).



lavaan, Latent Indicators, and Regression

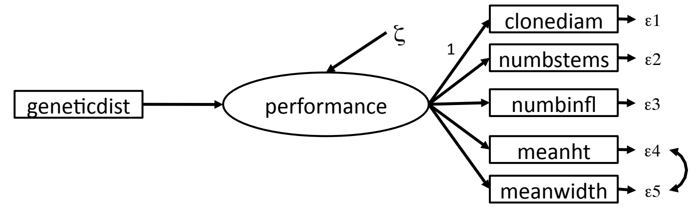



```
performance ~ clonediam
```



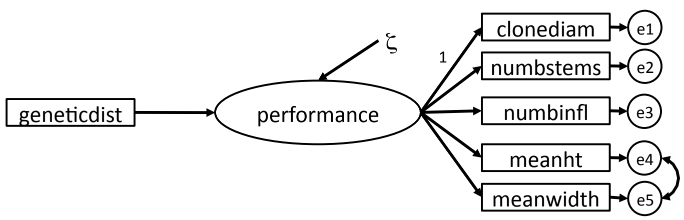

```
performance =~ clonediam
```

Using Latent Variables in SEM



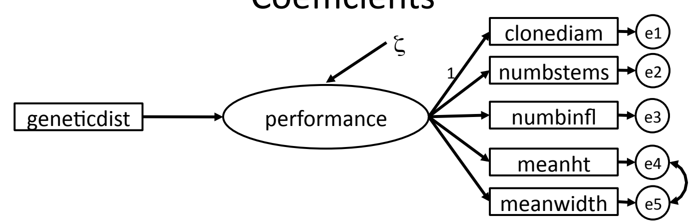

```
spartinaModel<-'performance =~  clonediam +
               numbstems + numbinfl + meanht +
               meanwidth

               meanht ~~ meanwidth

               performance ~ geneticdist'
```

Assessing Fit



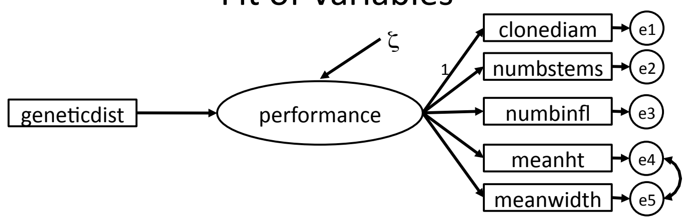

```
> spartinaFit<-sem(spartinaModel, data=spartina)
> summary(spartinaFit, standardized=T, rsquare=T)
lavaan (0.5-12) converged normally after 154 iterations

  Number of observations                            23

  Estimator                                         ML
  Minimum Function Test Statistic               12.237
  Degrees of freedom                                 8
  P-value (Chi-square)                           0.141
```

Coefficients



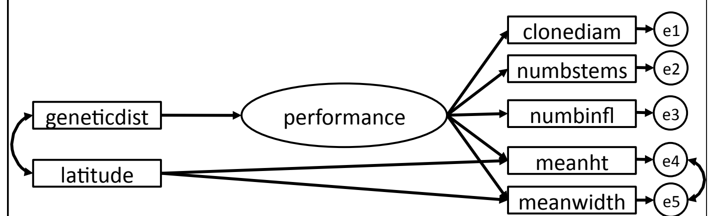
	Estimate	Std.err	Z-value	P(>|z|)	Std.lv	Std.all
Latent variables:						
performance =~						
clonediam	1.000				17.199	0.969
numbstems	0.904	0.079	11.508	0.000	15.555	0.962
numbinfl	0.106	0.015	7.030	0.000	1.822	0.853
meanht	0.643	0.114	5.654	0.000	11.066	0.785
meanwidth	0.076	0.016	4.680	0.000	1.308	0.718
Regressions:						
performance ~						
geneticdist	-57.134	12.465	-4.584	0.000	-3.322	-0.708

Fit of Variables

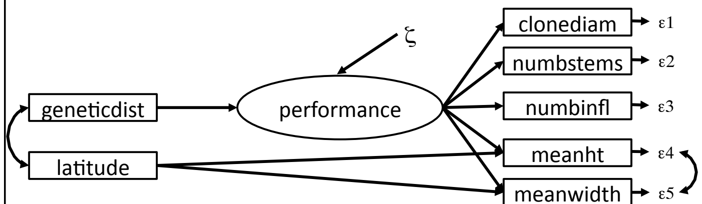



```
R-Square:

    clonediam          0.938
    numbstems          0.925
    numbinfl           0.727
    meanht             0.617
    meanwidth          0.515
    performance        0.502
```

A Latent Exercise



A Latent Exercise

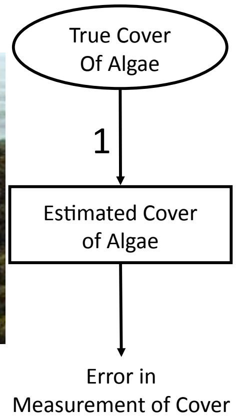



```
spartinaModel2<-paste(spartinaModel, '
                      meanht ~ latitude
                      meanwidth ~ latitude', sep="\n")
```

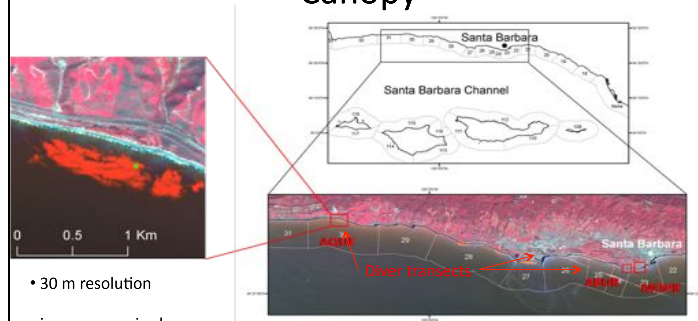
Latent Variables

1. What is a latent variable?
2. Latent variables with multiple indicators
3. Fitting a latent variable
4. Factor Analysis
5. Latent Variables as a Response
6. Coping with measurement error

Latent Variables and Measurement Error

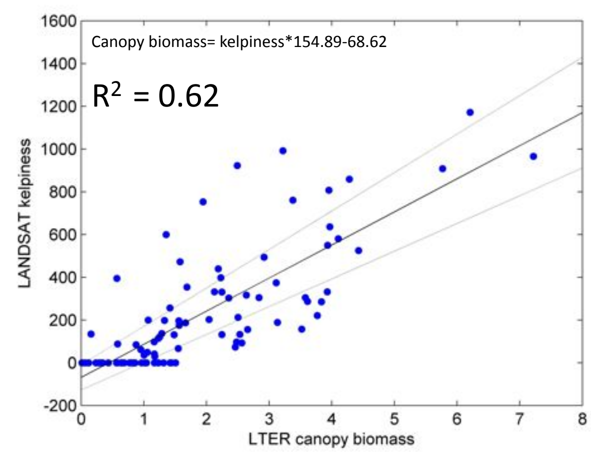


LANDSAT Measurements of Kelp Canopy



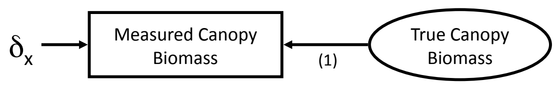
- 30 m resolution
- imagery acquired approx. every 2 weeks from 1984-present

LANDSAT Kelpiness vs. Canopy Biomass



LANDSAT Kelpiness vs. Canopy Biomass

We can transform satellite data to canopy biomass, and fix the unstandardized loading to 1.



But what about error?

We know that R^2=1-estimated var/observed var

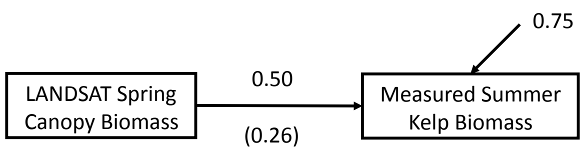
δ_x= (1-R^2)

Unstandardized Measurement Error =δ_x*var(Measured Canopy Biomass)

Let's Look at the LTER data: Data Prep

```
library(lavaan)
lter<-read.csv("./lter_kelp.csv")

#1) Calculate fitted values for spring biomass
#landsat observations to biomass
lter$landsat_spring_biomass<-154.89*lter$spring_canopy+68.62

#2) Calculate fitted values for summer biomass
#summer kelp counts to biomass y=0.08x+0.01 r^2=0.79
lter$summer_kelp_biomass<-0.08*lter$kelp+0.01


#3) Transform fitted values for easier fitting
#transformation for easier fitting
lter$summer_kelp_biomass<-log(lter$summer_kelp_biomass+1)
lter$landsat_spring_biomass <-log(lter$landsat_spring_biomass +1)
```

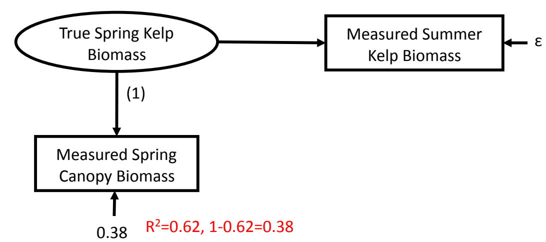
LANDSAT Kelpiness vs. Canopy Biomass
Fit this Model!




```
noerror<-'summer_kelp_biomass ~ landsat_spring_biomass'
```

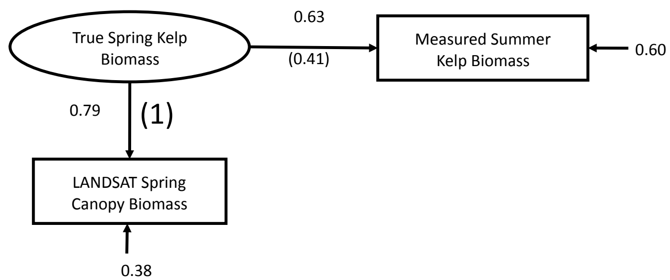
(unstandardized coefficients)

LANDSAT Kelpiness vs. Canopy Biomass



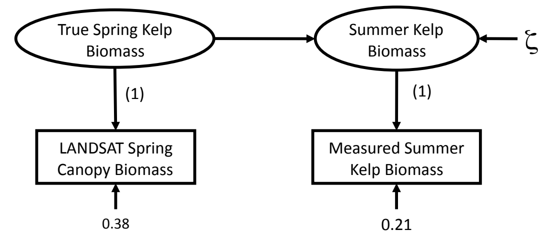

```
var(lter$landsat_spring_biomass, na.rm=T)*(0.38)

[1] 3.762301

errorCanopy<-'
    true_spring_biomass =~ 1*landsat_spring_biomass
    summer_kelp_biomass ~ true_spring_biomass

    #error
    landsat_spring_biomass ~~ 3.762301*landsat_spring_biomass
'
```

LANDSAT Kelpiness vs. Canopy Biomass



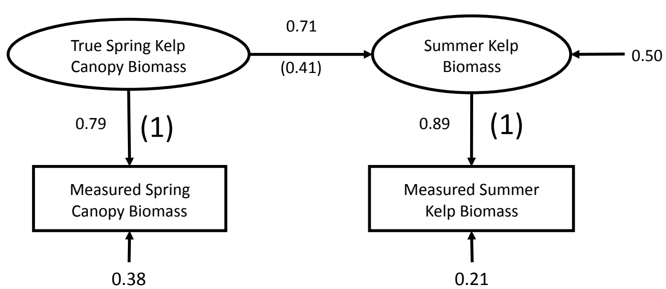
Incorporating measurement error in the predictor increases the coefficient size and amount of variance explained

Exercise: Code this model!




```
var(lter$summer_kelp_biomass, na.rm=T)*(0.21)
[1] 0.5495345

error_both<-'
    true_spring_biomass =~ 1*landsat_spring_biomass
    true_summer_biomass =~ 1*summer_kelp_biomass

    true_summer_biomass ~ true_spring_biomass

    #error
    landsat_spring_biomass ~~ 3.762301*landsat_spring_biomass
    summer_kelp_biomass ~~ 0.5495345* summer_kelp_biomass
'
```

LANDSAT Kelpiness vs. Canopy Biomass



Measurement error in the response affects standardized coefficients and estimates of error, but not the unstandardized coefficient.

Reasons to Think about Measurement Error

1. We know our measurements are not perfect!
2. Increased accuracy in estimating relationships between variables.
3. Increasing explanatory power of your hard-earned measurements.