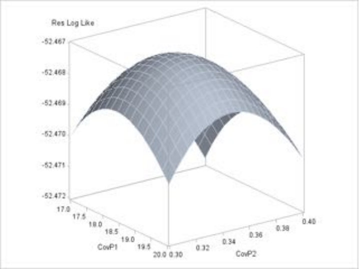
Introduction to Likelihood Methods for SEM

Jarrett E. K. Byrnes
University of Massachusetts Boston



$$\Sigma = \Sigma(\Theta)$$

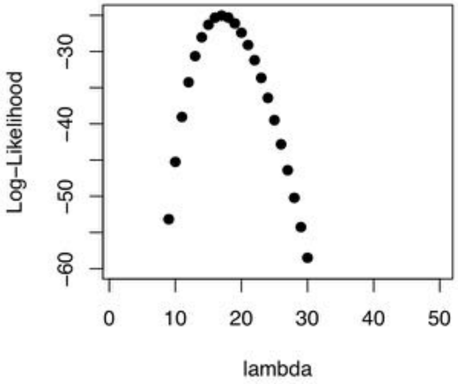
What is Covariance-Based SEM Estimation with Likelihood?

- Estimation of parameters given covariance of the data
- Equivalent to Linear Regressions, but…
- Estimation of each parameter influences the others
- Can accomodate unobserved (latent) variables and feedbacks

A Likely Outline

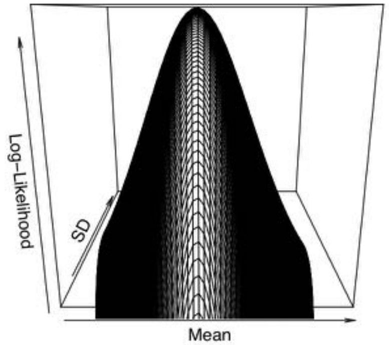
1. What SEM using likelihood and covariance matrices?
2. Model Identifiability
3. Sample Size for SEM
4. Introduction to `lavaan`

Maximizing Likelihood with One Parameter



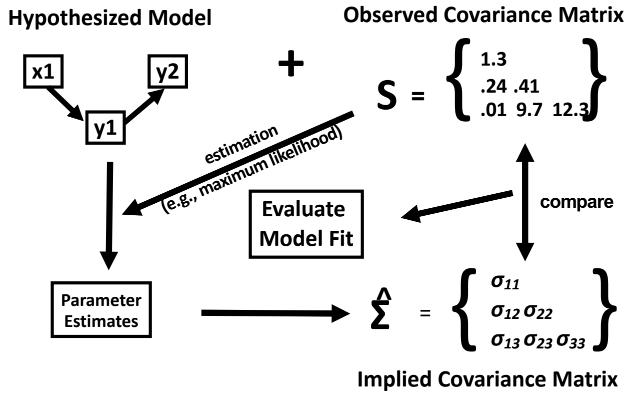
Iteration over possible values simple

Likelihood with Two Parameters



- Algorithms used to search parameter space
- Integrate answer over all data points
  - difficult computationally!

How does ML Estimation Work?


Hypothesized Model
Observed Covariance Matrix
x1
y2
y1
+
S =
1.3
.24 .41
.01 9.7 12.3
estimation
(e.g., maximum likelihood)
Evaluate
Model Fit
compare
Parameter
Estimates
$\hat{\Sigma}$ =
σ_{11}
$\sigma_{12}\ \sigma_{22}$
$\sigma_{13}\ \sigma_{23}\ \sigma_{33}$
Implied Covariance Matrix

What we're used to with ML

Data Generation: $\mu_i = a + bX_i$

Likelihood Function: $F_r = Y_i \sim \mathrm{dnorm}(\mu_i, \sigma)$

We minimize the likelihood function, F_r

It's…More Complicated with SEM

Data Generation:

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} = \begin{pmatrix} \Lambda_y (\mathrm{I}-\mathrm{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(\mathrm{I}-\mathrm{B})^{-1'}\Lambda_y' + \theta_\varepsilon & \Lambda_y (\mathrm{I}-\mathrm{B})^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x \Phi\Gamma'(\mathrm{I}-\mathrm{B})^{-1'} & \Lambda_x\Phi\Lambda_x' + \theta_\delta \end{pmatrix}$$

Likelihood Function:

$$F_{ML} = \log\left|\hat{\Sigma}\right| - \log\left|\mathbf{S}\right| + tr\left(\mathbf{S}\hat{\Sigma}^{-1}\right) - (p+q)$$

The Maximum Likelihood Fitting Function

$$F_{ML} = \log\left|\hat{\Sigma}\right| - \log\left|\mathbf{S}\right| + tr\left(\mathbf{S}\hat{\Sigma}^{-1}\right) - \left(p+q\right)$$

S = Sample covariance matrix
S = Fit covariance matrix
p = endogenous variables
q = exogenous variables

Linear Algebra Review

Det(A) = scalar number

$A*A^{-1}$ = Diagonal matrix of ones

- **If S =Σ, term 1 - 2 = 0 and terms 3 - 4 = 0.**
- **F_{ML} = 0 with perfect fit**

Assumptions Behind F_{ml}

- Multivariate normality
  - Fairly robust (non-normality of residuals bigger problem)
  - Test with multivariate Shapiro-Wilk's Test (library mvnormtest)
  - In particular, no skew
  - Severe violations bias parameter error and tests of model fit
- No missing data in calculation of S
  - Biases your estimates with pairwise corrections
- No redundant variables
  - S must be positive definite
- Sample size is "large" (more soon)

A Likely Outline

1. What is different about fitting using likelihood and covariance matrices?
2. Identifiability
3. Sample Size (for likelihood and piecewise approaches)
4. Introduction to `lavaan`

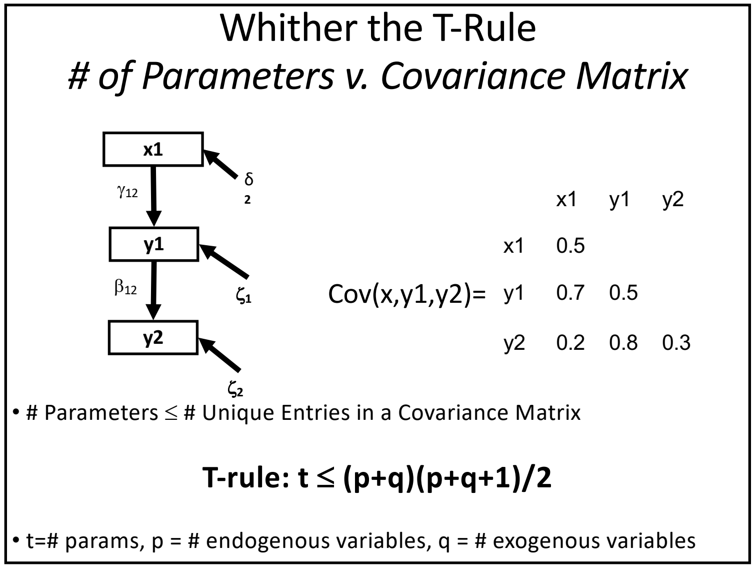
Identifiability

1. To fit a model, it must be identified

2. We need as much unique information as parameters

3. What can make a model non-identified?
   - Too many paths relative to # of variables
   - Certain model structures
   - High multicollinearity (r>0.9)
   - Complex model & small sample

4. How do I know if my model is identified?



Whither the T-Rule

of Parameters v. Covariance Matrix

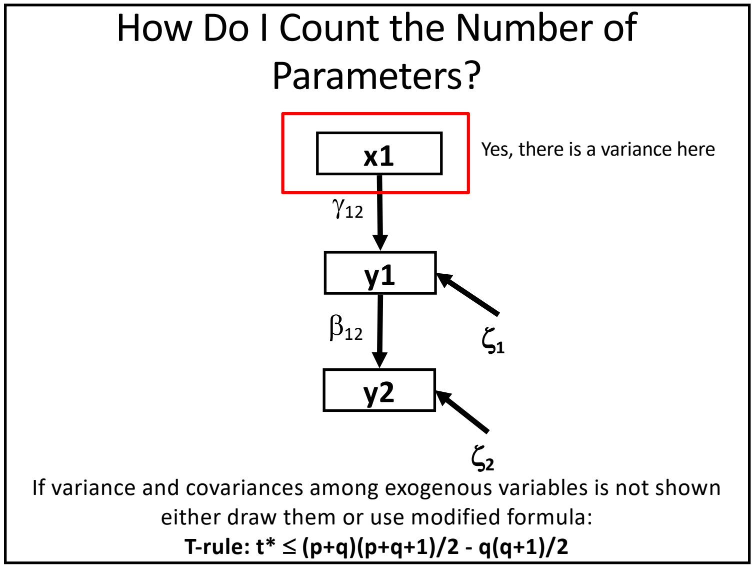
Cov(x,y1,y2)=

	x1	y1	y2
x1	0.5		
y1	0.7	0.5	
y2	0.2	0.8	0.3

- # Parameters ≤ # Unique Entries in a Covariance Matrix

T-rule: $t \leq (p+q)(p+q+1)/2$

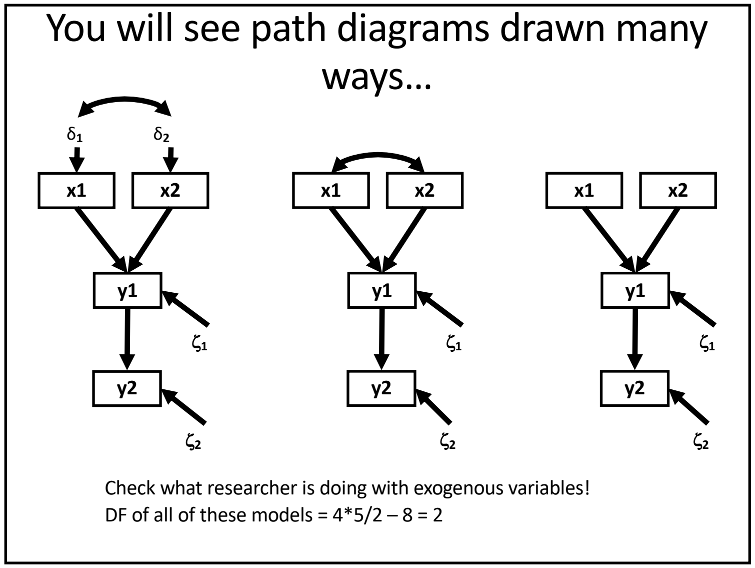
- t=# params, p = # endogenous variables, q = # exogenous variables



How Do I Count the Number of Parameters?

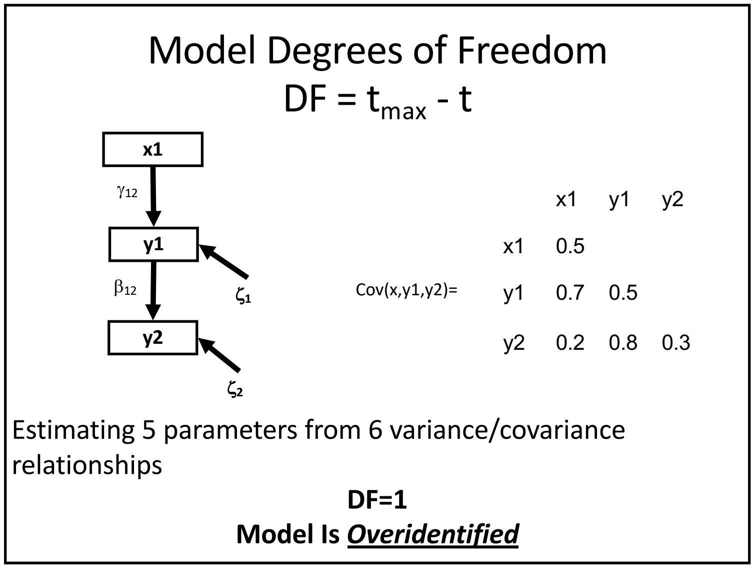
If variance and covariances among exogenous variables is not shown either draw them or use modified formula:

T-rule: $t^* \leq (p+q)(p+q+1)/2 - q(q+1)/2$



You will see path diagrams drawn many ways…

Check what researcher is doing with exogenous variables!
DF of all of these models = 4*5/2 – 8 = 2



Model Degrees of Freedom

DF = t_{max} - t

Cov(x,y1,y2)=

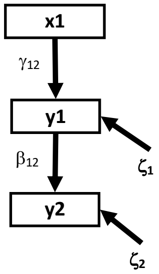
	x1	y1	y2
x1	0.5		
y1	0.7	0.5	
y2	0.2	0.8	0.3

Estimating 5 parameters from 6 variance/covariance relationships

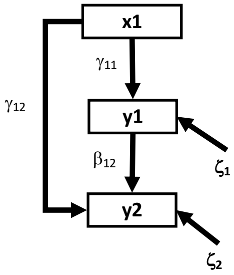
DF=1

Model Is *<u>Overidentified</u>*

Identification in SEM
# of Parameters v. Covariance Matrix



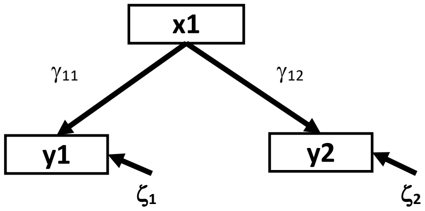
Overidentified



Just Identified

Just Identified models have no DF to evaluate fit

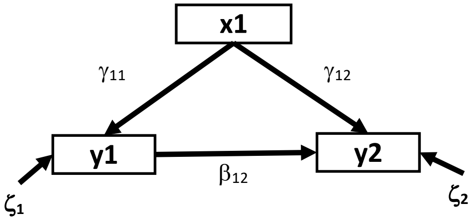
Identification in SEM
Many Regressions



Yes: There are no relationships between endogenous variables

SUFFICIENT CONDITION

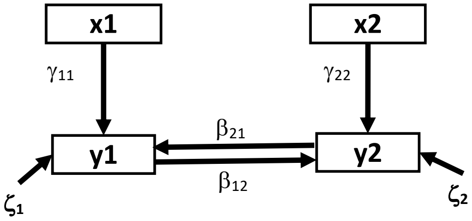
Identification in SEM
No Feedbacks



Yes: Model is Recursive

SUFFICIENT CONDITION

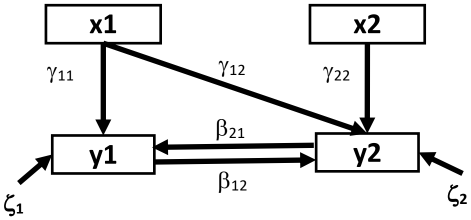
Identification in SEM
Feedbacks with Different Causes



YES: Model is Non-recursive, but y's have unique information

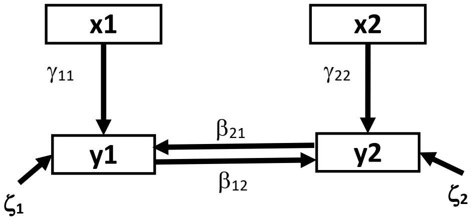
NECESSARY CONDITION

Identification in SEM
Is this model identified?



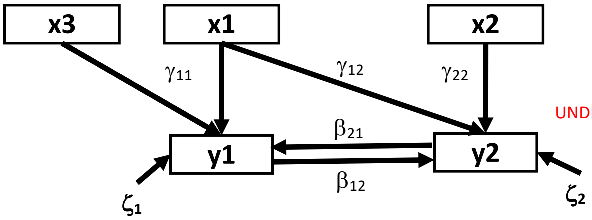
NO! Model is Non-recursive
AND not enough information for unique solution

Identification in SEM
The Order Condition



- G = # incoming paths
- H = # of exogenous vars+ # indirectly connected endogenous vars
- G ≤ H: Enough information per variable!
- **NECESSARY CONDITION**

Identification in SEM
The Rank Condition



EMPIRICAL
UNDERIDENTIFICATION

Everything that affects y1 affects y2 – Fails *Rank Test*

SUFFICIENT CONDITION

Rules of Identification

Necessary

- Fewer parameters than entries in covariance diagonal matrix (T-Rule)
- Fewer incoming paths than # of variables connected to (Order condition for non-recursive models)

Sufficient

- No paths between endogenous variables
- Model is recursive
- Unique effects on endogenous variables in a feedback (Rank Condition)

A Likely Outline

1. What is different about fitting using likelihood and covariance matrices?
2. Identifiability
3. Sample Size
4. Introduction to `lavaan`

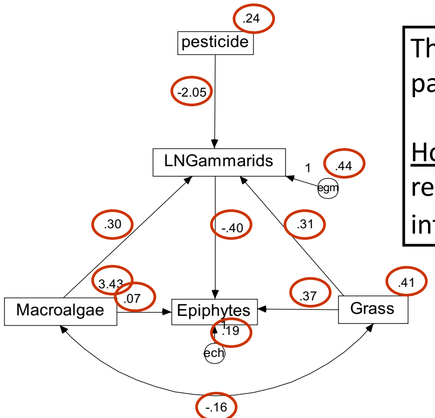
Sample Size

1. The further you are in a model from an exogenous data-generating, the weaker it's influence.
2. Our ability to detect these tapering effect sizes is proportional to our information (especially sample size) and the number of parameters being estimated.
3. Sample size sets an upper limit for the complexity of the model we can obtain.
4. Sample Size influences our ability to detect lack of model fit
   - This might not be a benefit…

So…What's my Sample Size?

1. Rules of thumb for sample size - at least 5 samples per estimated parameter
   - prefer 20 samples per parameter
   - Really, $p^{3/2}/n$ should approach 0 (Portnoy 1988)
2. Path coefficients add to our parameter list, not the variances

Number of Estimated Parameters

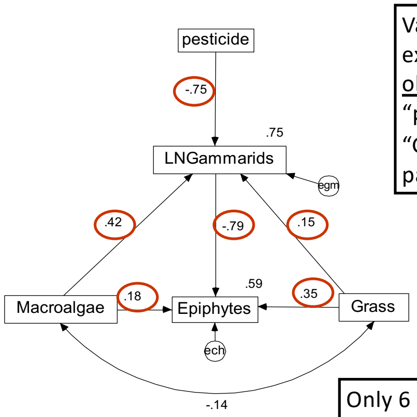


Chi-sqr = 5.147; df = 3; p = .161
AIC = 29.147; NPAR = 12

There are a total of 12 parameters shown.

However, only 6 of these require unique information…

Parameters Needing Unique Information



Chi-sqr = 5.147; df = 3; p = .161
AIC = 29.147; NPAR = 12

Variances & covariance of exogenous variables can be obtained from the data. For "pesticide", "Macroalgae", and "Grass", this removes 4 parameters.

Error variances (and R^2) for endogenous variables are calculated from other parameters. This removes 2 parameters.

Only 6 parameters require unique information.
Samples/parameters = 40/6 = 6.7.

A Likely Outline

1. What is different about fitting using likelihood and covariance matrices?
2. Identifiability
3. Sample Size (for likelihood and piecewise approaches)
4. Introduction to `lavaan`

What is `lavaan`?

- Stands for LAtent VAriable Analaysis
- Written by Yves Roseel in 2010
- Currently in version 5, but 6 coming soon
- Uses R `lm` syntax

A Reminder

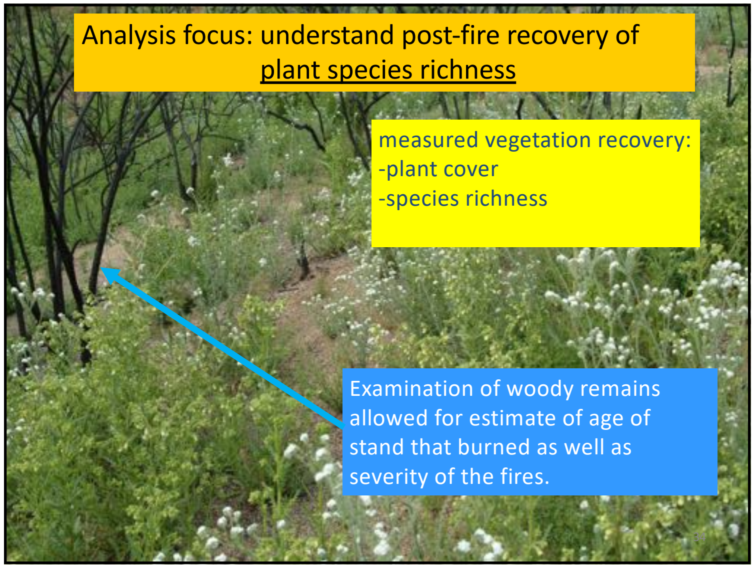
1. SOFTWARE IS A TOOL

2. IT IS NOT PERFECT

3. ALWAYS MAKE SURE IT IS DOING WHAT YOU THINK IT IS DOING!

Mediation in Analysis of Post-Fire Recovery of Plant Communities in California Shrublands*

*Five year study of wildfires in Southern California in 1993. 90 plots (20 x 50m), (data from Jon Keeley et al.)

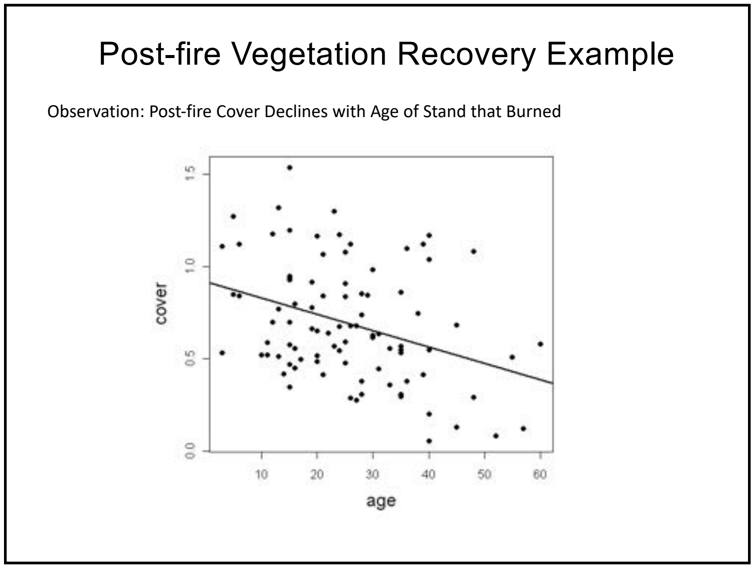


Analysis focus: understand post-fire recovery of plant species richness

measured vegetation recovery:
-plant cover
-species richness

Examination of woody remains allowed for estimate of age of stand that burned as well as severity of the fires.

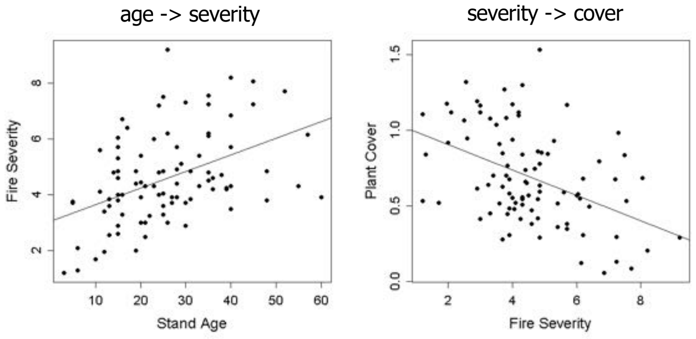
Other factors measured included:
- local abiotic conditions (aspect, soils)
- spatial heterogeneity
- landscape-level conditions (location, elevation)



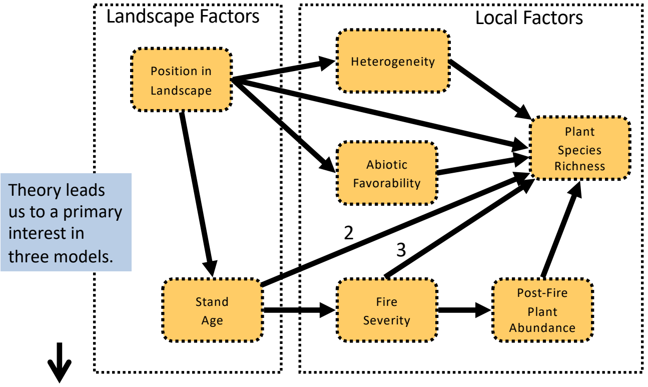
Post-fire Vegetation Recovery Example

Observation: Post-fire Cover Declines with Age of Stand that Burned

Post-fire Vegetation Recovery Example (cont.):



The SEMM

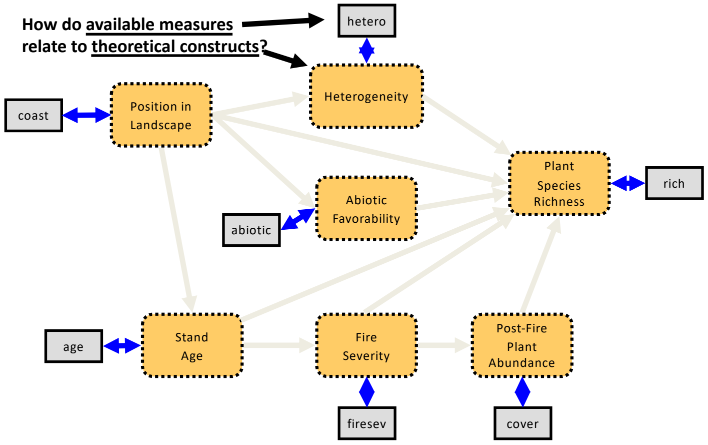


Theory leads us to a primary interest in three models.

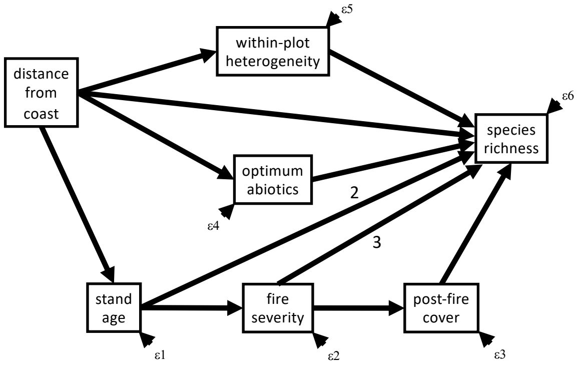
model 1 - includes all unnumbered paths
model 2 - possible loss of species from seedbank in older stands
model 3 - possible selective destruction of seeds in seedbank in severe fires

Matching the SEMM to Data

How do available measures relate to theoretical constructs?

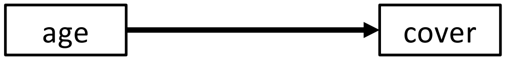


Realized Models with Data



Model 1 - all unnumbered paths
Model 2 – includes additional mechanism: loss of species from seedbank in older stands
Model 3 – includes selective destruction of seeds in seedbank in severe fires

Coding a Regression versus SEM



```
#regression
aLM<-lm(cover ~ age, data=keeley)

#sem
library(lavaan)
aSEM<-sem('cover ~ age', data=keeley)
```

summary(aSEM)

The model converged!

```
lavaan (0.5-23.1097) converged normally after  10 iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic                0.000
  Degrees of freedom                                 0

Parameter estimates:

  Information                                 Expected
  Standard Errors                             Standard

                   Estimate  Std.err  Z-value  P(>|z|)
Regressions:
  cover ~
    age              -0.009    0.002   -3.549    0.000

Variances:
    .cover            0.087    0.013
```

Model is saturated so, χ2 test has no df

Compare to Regression

```
                   Estimate  Std.err  Z-value  P(>|z|)
Regressions:
  cover ~
    age              -0.009    0.002   -3.549    0.000

Variances:
    .cover            0.087    0.013

> summary(aLM)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.917395   0.071726   12.79  < 2e-16 ***
age         -0.008846   0.002520   -3.51  0.00071 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2988 on 88 degrees of freedom
```

Compare to Residual SE
sqrt(0.087)=0.295

But what about the intercept?

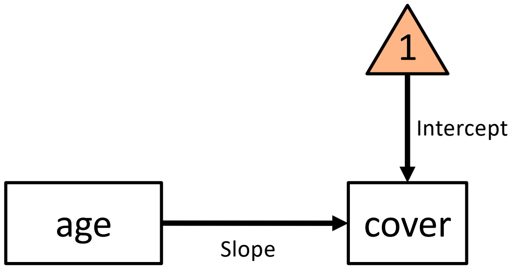
Intercepts Estimated with Mean Structure

```
> aMeanSEM<-sem('cover ~ age',
  data=keeley, meanstructure=T)
```

```
> summary(aMeanSEM)
...
                   Estimate  Std.err  Z-value  P(>|z|)
Regressions:
  cover ~
    age              -0.009    0.002   -3.549    0.000

Intercepts:
    .cover            0.917    0.071   12.935    0.000

Variances:
    .cover            0.087    0.013
```

Intercepts Estimated with Mean Structure

```
> aMeanSEM<-sem('cover ~ age',
  data=keeley, meanstructure=T)
```



1

Intercept

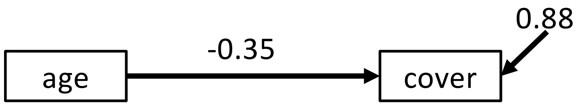
age

Slope

cover

Standardized Coefficients

```
>standardizedSolution(aSEM)
    lhs op   rhs est.std    se      z pvalue
1 cover  ~   age  -0.350 0.090 -3.912      0
2 cover ~~ cover   0.877 0.063 13.973      0
3   age ~~   age   1.000 0.000     NA     NA
```



0.88

age

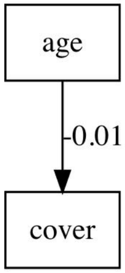
-0.35

cover

Also: `summary(aSEM, standardized=T, rsq=T)`

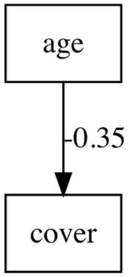
Can I See It?

```
library(lavaanPlot)
lavaanPlot(model = aSEM, coefs = TRUE)
```

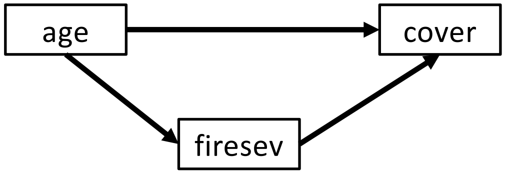


Can I See It?

```
lavaanPlot(model = aSEM, coefs = TRUE,
           stand=TRUE)
```

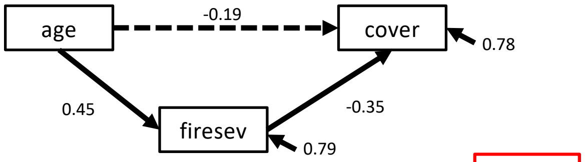


Indirect Effects and Fire



```
partialMedModel<-' firesev ~ age
                  cover ~ firesev + age'

partialMedSEM<-sem(partialMedModel,
                data=keeley)
```

summary(partialMedSEM, rsquare=T, standardized=T)



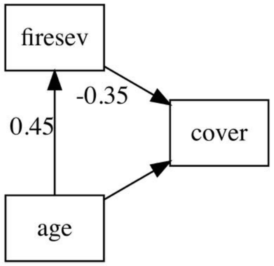
```
                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Regressions:
  firesev ~
    age               0.060    0.012    4.832    0.000    0.060    0.454
  cover ~
    firesev          -0.067    0.020   -3.353    0.001   -0.067   -0.350
    age              -0.005    0.003   -1.833    0.067   -0.005   -0.191

Variances:
    .firesev           2.144    0.320                      2.144    0.794
    .cover             0.078    0.012                      0.078    0.780

R-Square:

    firesev           0.206
    cover             0.220
```

Plotting… and it's Limits

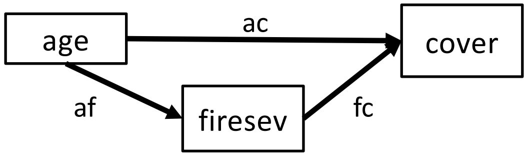
```
lavaanPlot(model = partialMedSEM, coefs = TRUE,
           stand = TRUE,
           graph_options = list(layout = "circo"),
           sig = 0.05)
```

Only shows coefs p≤0.05

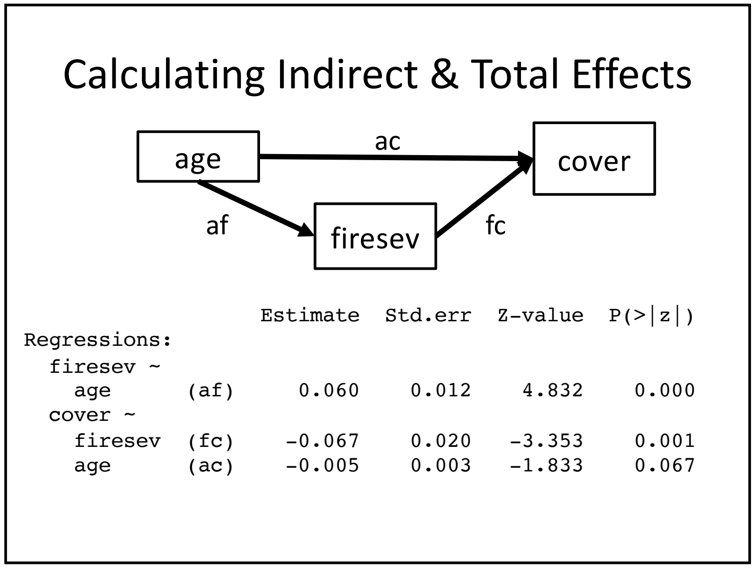
Better layout for this model



Calculating Indirect & Total Effects

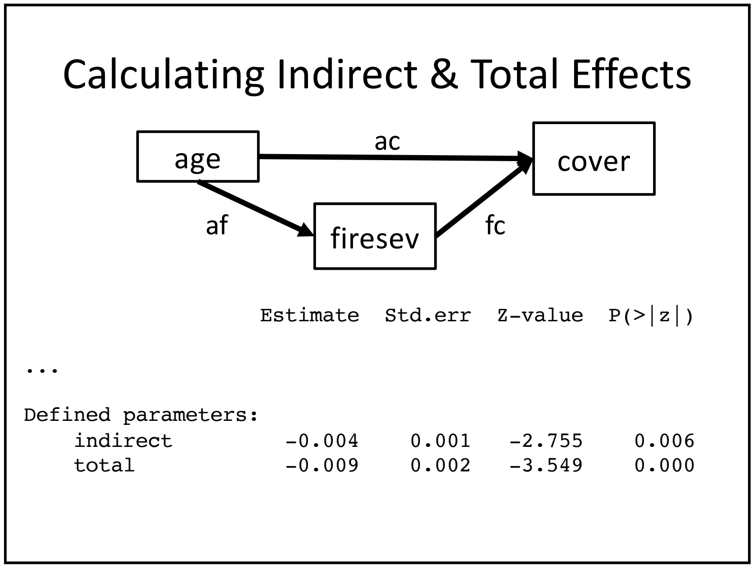


```
partialMedModelInd <-'

  #model
  firesev ~ af*age
  cover ~ fc*firesev + ac*age

  #Derived Calcuations
  indirect := af*fc
  total := ac + (af*fc)
'
```



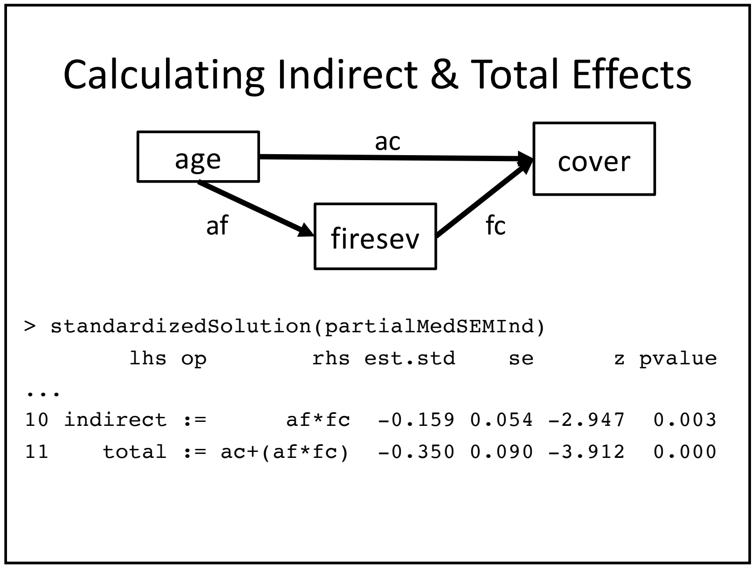
Calculating Indirect & Total Effects

```
                   Estimate  Std.err  Z-value  P(>|z|)
Regressions:
  firesev ~
    age      (af)     0.060    0.012    4.832    0.000
  cover ~
    firesev  (fc)    -0.067    0.020   -3.353    0.001
    age      (ac)    -0.005    0.003   -1.833    0.067
```



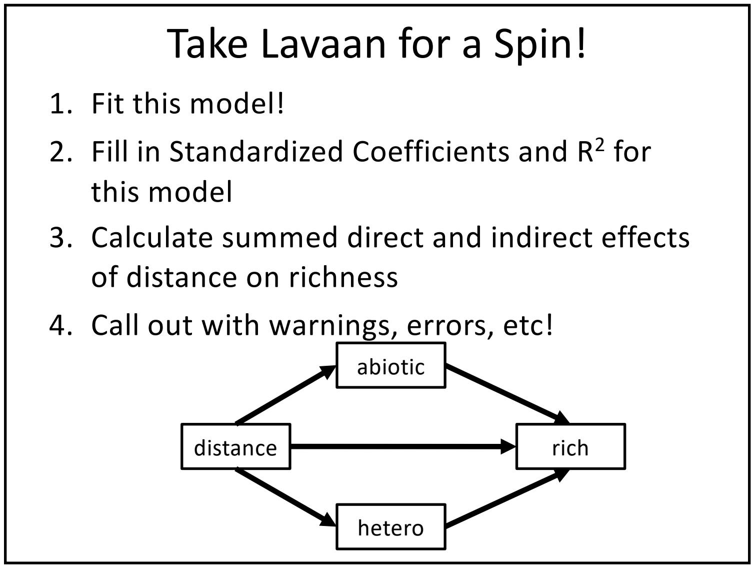
Calculating Indirect & Total Effects

```
                   Estimate  Std.err  Z-value  P(>|z|)
...

Defined parameters:
    indirect         -0.004    0.001   -2.755    0.006
    total            -0.009    0.002   -3.549    0.000
```



Calculating Indirect & Total Effects

```
> standardizedSolution(partialMedSEMInd)
        lhs op         rhs est.std    se      z pvalue
...
10 indirect :=       af*fc  -0.159 0.054 -2.947  0.003
11    total := ac+(af*fc)  -0.350 0.090 -3.912  0.000
```



Take Lavaan for a Spin!

1. Fit this model!
2. Fill in Standardized Coefficients and R^2 for this model
3. Calculate summed direct and indirect effects of distance on richness
4. Call out with warnings, errors, etc!

The dreaded variance warning!

```
Warning message:
In lav_data_full(data = data, group =
group, cluster = cluster,   :
  lavaan WARNING: some observed
variances are (at least) a factor
1000 times larger than others; use
varTable(fit) to investigate
```

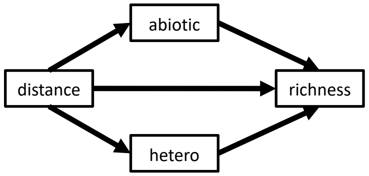
Diagnosing Error Issues

```
> inspect(distFit, "obs")
$cov
         rich    hetero  abiotc  distnc
rich     225.646
hetero     0.784   0.013
abiotic   58.312   0.241  58.314
distance  77.089   0.347  30.824  77.094
```

Is this OK?

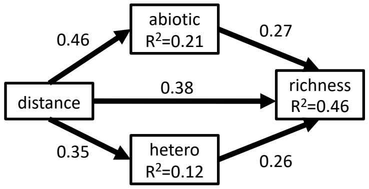
1. Does it indicate an outlier or data problem?
2. This is a likelihood algorithm problem – can be fine!
3. If you are worried, rescale by 10s, see if answers change

Solution 1: The Model



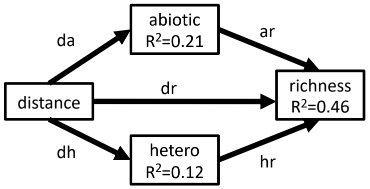
```
#The Richness Partial Mediation Model
distModel <- 'rich ~ distance + abiotic + hetero
              hetero ~ distance
              abiotic ~ distance'

distFit <- sem(distModel, data=keeley)

standardizedSolution(distFit)
```

Solution 2: Coefficients



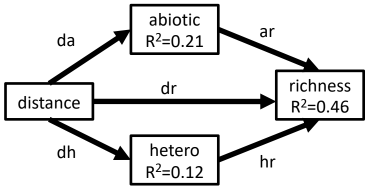
```
       lhs op      rhs est.std    se     z pvalue
1     rich  ~ distance   0.377 0.092 4.117  0.000
2     rich  ~  abiotic   0.268 0.087 3.079  0.002
3     rich  ~   hetero   0.256 0.082 3.104  0.002
4   hetero  ~ distance   0.346 0.099 3.498  0.000
5  abiotic  ~ distance   0.460 0.094 4.911  0.000
6     rich ~~     rich   0.539 0.080 6.708  0.000
7   hetero ~~   hetero   0.880 0.131 6.708  0.000
8  abiotic ~~  abiotic   0.789 0.118 6.708  0.000
9 distance ~~ distance   1.000    NA    NA     NA
```

Solution 3: Direct and Indirect



```
distModelEff <- '
rich ~ dr*distance + ar*abiotic + hr*hetero
hetero ~ dh*distance
abiotic ~ da*distance

#The effects
direct := dr
indirect := dh*hr + da*ar
total := direct + indirect
'
```

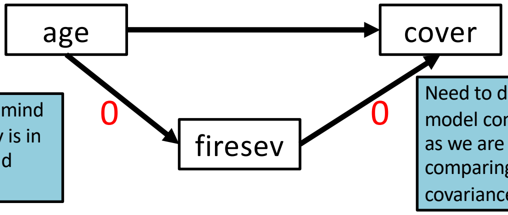
Solution 3: Direct and Indirect



```
> standardizedSolution(distFitEff)
        lhs op             rhs est.std    se      z pvalue
...
10   direct :=              dr   0.377 0.086  4.390  0.000
11 indirect :=    dh*hr+da*ar   0.212 0.055  3.835  0.000
12    total := direct+indirect   0.589 0.062  9.433  0.000
```

What would you say about direct and indirect effects in this system?

What if we know better?



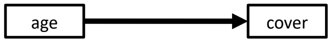
Fill in 0's to remind us that firesev is in the model, and fixed to 0

Need to do this for model comparison, as we are comparing covariance matrices

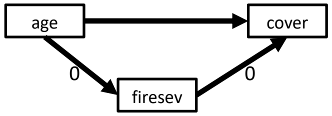
```
zeroMedModel<-' firesev ~ 0*age
               cover ~ 0*firesev + age'

zeroMedFit<-sem(zeroMedModel,
                data=keeley)
```

What lavaan sees…

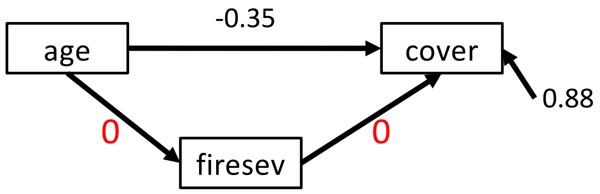
```
> inspect(aSEM, "obs")
$cov
      cover   age
cover   0.100
age    -1.381 156.157
...
```



```
> inspect(zeroMedFit, "obs")
$cov
        firesv  cover    age
firesev   2.700
cover    -0.227  0.100
age       9.319 -1.381 156.157
...
```

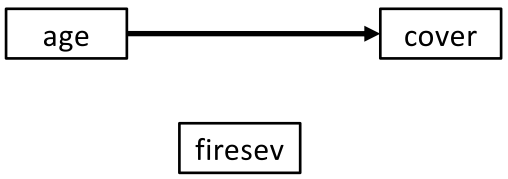


standardizedSolution(zeroMedFit)



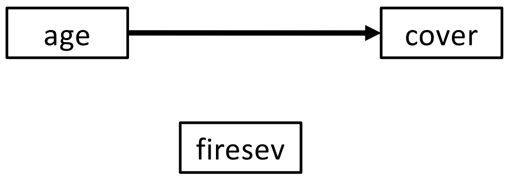
```
    lhs op     rhs est.std    se      z pvalue
1 firesev  ~     age   0.000    NA     NA     NA
2   cover  ~ firesev   0.000    NA     NA     NA
3   cover  ~     age  -0.350 0.099 -3.549      0
4 firesev ~~ firesev   1.000 0.149  6.708      0
5   cover ~~   cover   0.877 0.131  6.708      0
6     age ~~     age   1.000    NA     NA     NA
```

Or… Just use intercepts!



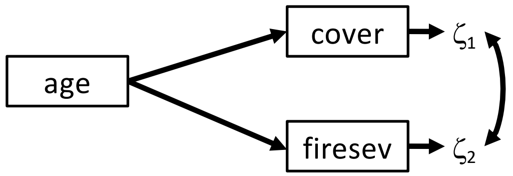
```
zeroMedModel2<-'
  firesev ~ 1
  cover ~ age
'
```

Or… Just use intercepts!



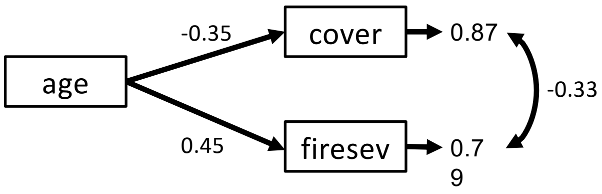
```
      lhs op     rhs est.std    se      z pvalue
1 firesev ~1              2.778 0.232 11.956      0
2   cover  ~     age     -0.350 0.090 -3.912      0
3   cover ~~   cover      0.877 0.063 13.973      0
```

What about Correlated Error?



```
#what about correlations
corModel <-'firesev ~ age
            cover ~ age
            cover ~~ firesev'

corFit <- sem(corModel, data=keeley)
```

What about Correlated Error?

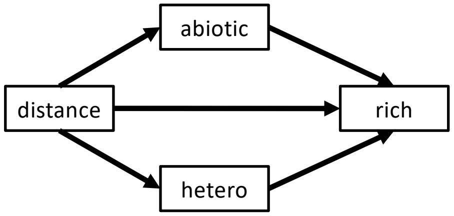


```
> standardizedSolution(corFit)
      lhs op     rhs est.std    se      z pvalue
1 firesev  ~     age   0.454 0.094  4.832      0
2   cover  ~     age  -0.350 0.099 -3.549      0
3 firesev ~~   cover  -0.333 0.094 -3.556      0
4 firesev ~~ firesev   0.794 0.118  6.708      0
5   cover ~~   cover   0.877 0.131  6.708      0
6     age ~~     age   1.000    NA     NA     NA
```

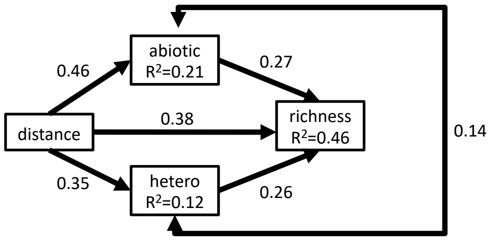
Final Exercise

1. How does this model differ if the abiotic and hetero error correlate?

2. Fit assuming that there is a 1:1 (think 1 instead of 0) relationship between distance and richness
   - No error correlation please



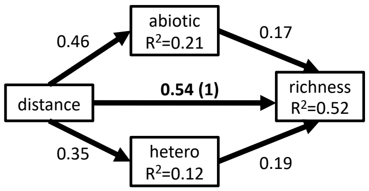
Solution 1: Error Correlation



```
corErrorModel <- '
  rich ~ distance + abiotic + hetero
  hetero ~ distance
  abiotic ~ distance

  abiotic ~~ hetero
'
```

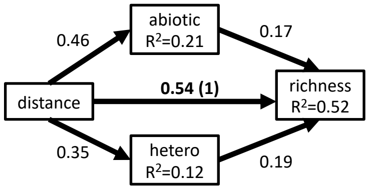
Coefficients unaffected

Solution 2: The New Model



```
oneDistModel <- 'rich ~ 1*distance + abiotic + hetero
                 hetero ~ distance
                 abiotic ~ distance'

oneFit<-sem(oneDistModel, data=keeley)
summary(oneFit, stdandardized=T, rsquare=T)
```

Solution 2: The New Model



distance

abiotic R²=0.21

hetero R²=0.12

richness R²=0.52

0.46

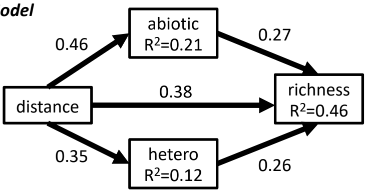
0.35

0.54 (1)

0.17

0.19

Unconstrained Model



distance

abiotic R²=0.21

hetero R²=0.12

richness R²=0.46

0.46

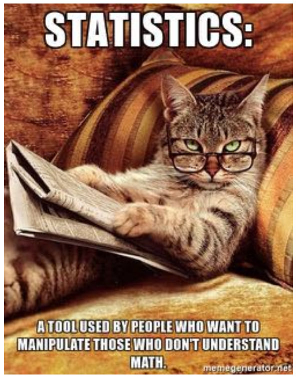
0.35

0.38

0.27

0.26

Now that you're armed and dangerous…


STATISTICS:
A TOOL USED BY PEOPLE WHO WANT TO MANIPULATE THOSE WHO DON'T UNDERSTAND MATH.
memegenerator.net

Fit your data to a ***SIMPLE*** model with lavaan