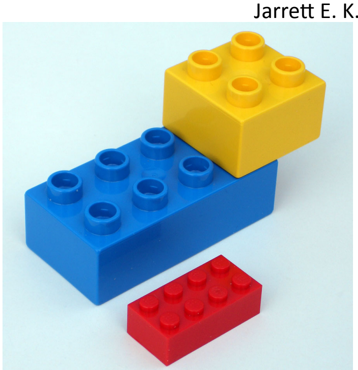
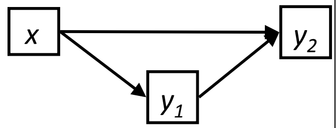
Piecewise Fitting and Evaluation of SEMs with Observed Variables

Jarrett E. K. Byrnes





lm(y2 ~ x+y1)

lm(y1 ~ x)

"There are no routine statistical questions, only questionable statistical routines"

- Sir David Cox

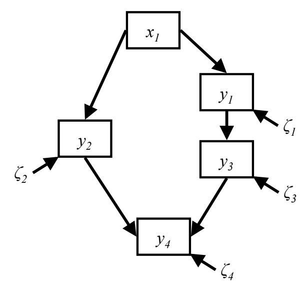
Outline

1. What is piecewise SEM?
2. Example using Keeley's fire data
   – Pre-SEM data screening
   – Combining models into an SEM
   – Simple evaluation of mediation
3. Evaluating model fit
   – D-Separation
4. D-Separation in R

The Two Goals of Fittings SEMs

(a) Evaluation of the network

(b) Estimation of the parameters



Major questions about network:

P & ML 1. Does theory support the causal structure?

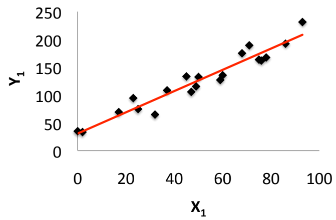
P & ML 2. Are there any missing connections?

P & ML 3. Are there any unnecessary connections?

ML 4. Are the data consistent with the model structure?

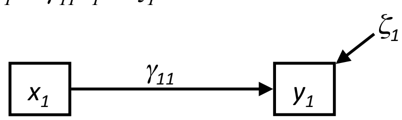
ML 5. Any error correlations indicating missing processes?

Linear Regression & SEM in 1 Slide

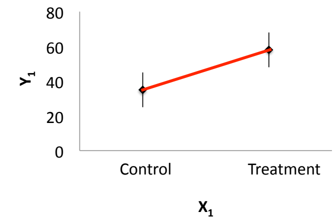


equation form $y_1 = \gamma_{11}x_1 + \zeta_1$

graphical form



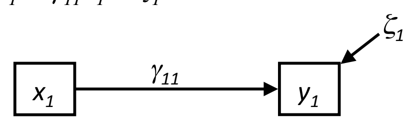
ANOVA *IS* Linear Regression



F-Tests assess whether variable contributes to observed variation in data

equation form $y_1 = \gamma_{11}x_1 + \zeta_1$

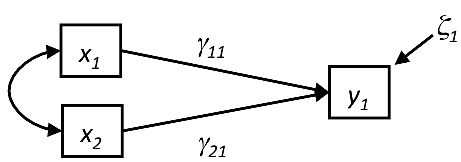
graphical form



Multiple Regression: Hidden Assumptions

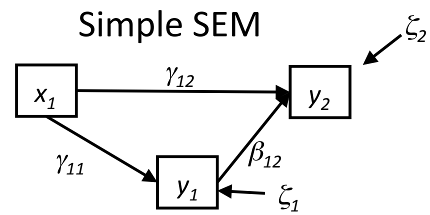
equation form $y_1 = \gamma_{11}x_1 + \gamma_{21}x_2 + \zeta_1$

graphical form



- To estimate γ11 and γ21 you need to control for the unanalyzed correlation between x1 and x2.
- In a factorial experiment, this correlation is 0!

Moving from Multiple Regression to Simple SEM



equation form

$$y_1 = \gamma_{11}x_1 + \zeta_1$$

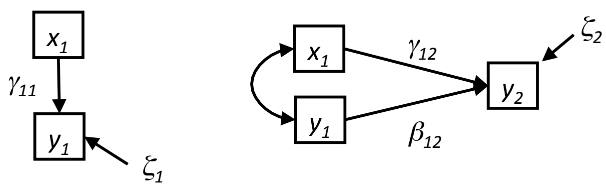
$$y_2 = \gamma_{12}x_1 + \beta_{12}y_1 + \zeta_2$$

Moving from Multiple Regression to Simple SEM

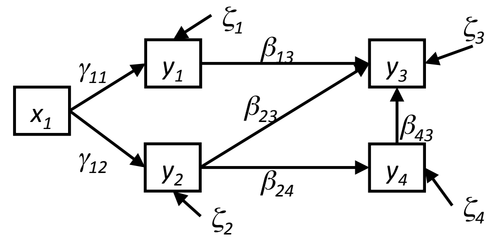
equation form

$$y_1 = \gamma_{11} x_1 + \zeta_1$$

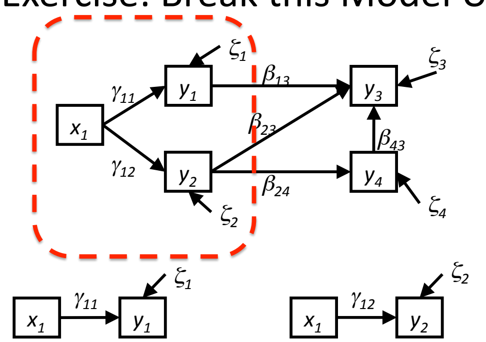
$$y_2 = \gamma_{12} x_1 + \beta_{12} y_1 + \zeta_2$$



Exercise: Break this Model Up

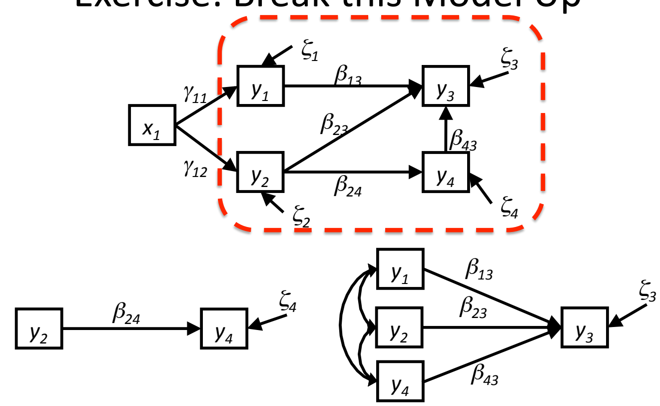


Exercise: Break this Model Up



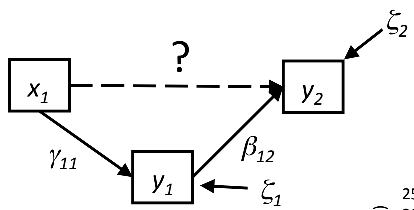
2 Simple Linear Regressions

Exercise: Break this Model Up

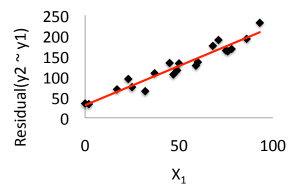


1 Simple, 1 Multiple Regression

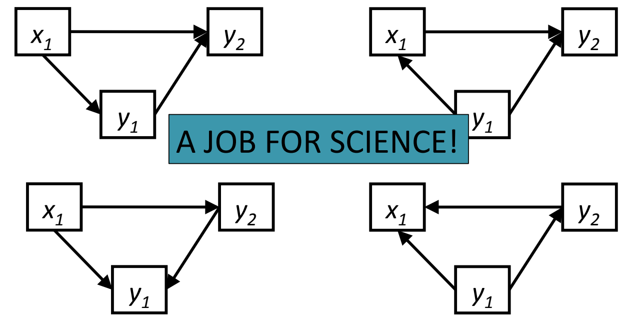
Evaluation of Missing Paths



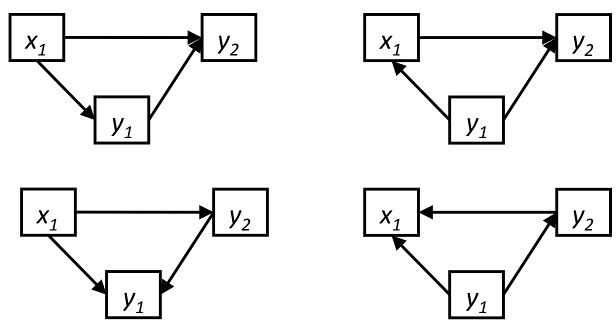
1. Fit model with and without path. Evaluate if path is significant.
2. Evaluate relationship between residual of y2 ~ y1 relationship and x1



The Problem of Equivalent Models



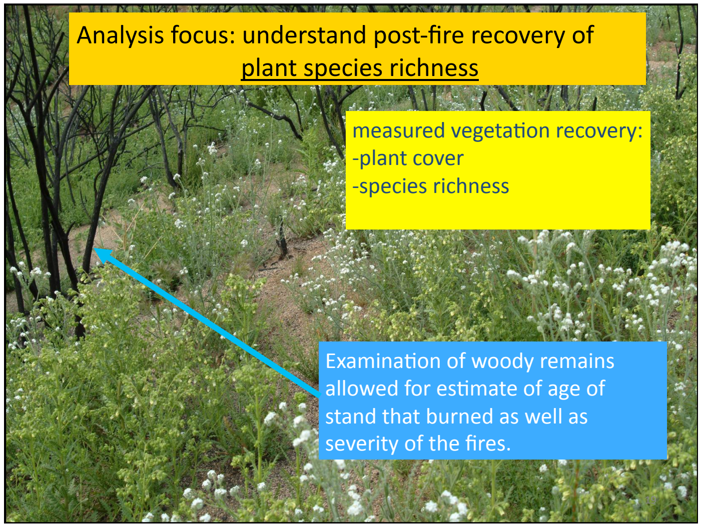
Piecewise is ONLY for Recursive Models (Directed Acyclic Graphs – DAGs)

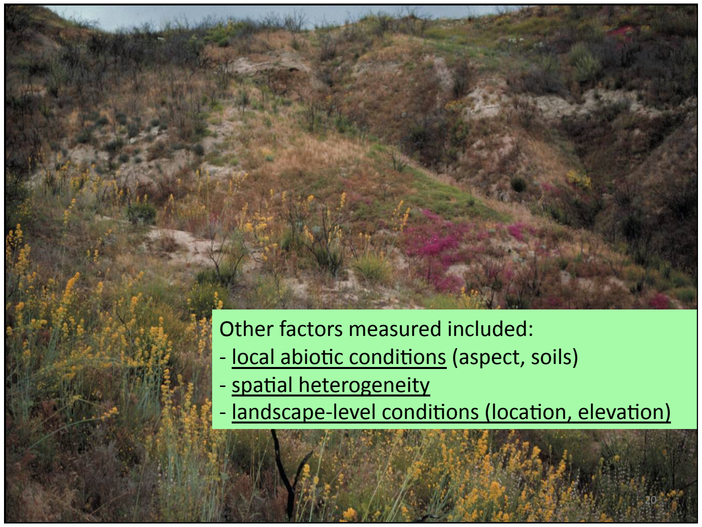


Questions?

Outline

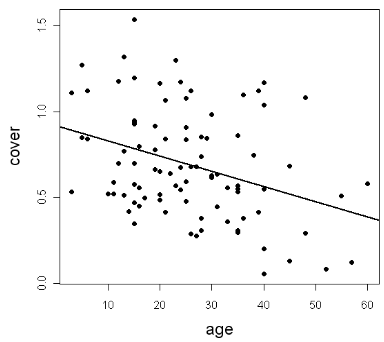
1. What is piecewise SEM?
2. Example using Keeley's fire data
   – Pre-SEM data screening
   – Combining models into an SEM
   – Simple evaluation of mediation
3. Evaluating model fit
   – D-Separation
4. D-Separation in R



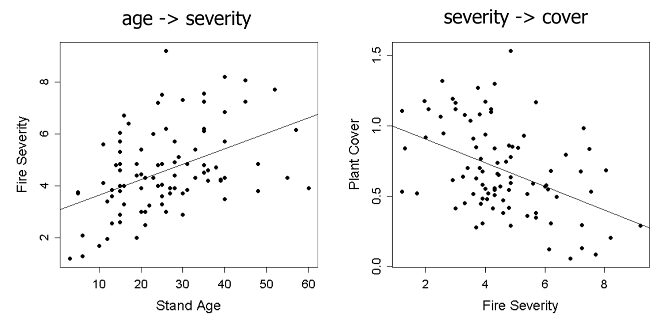


Post-fire Vegetation Recovery Example

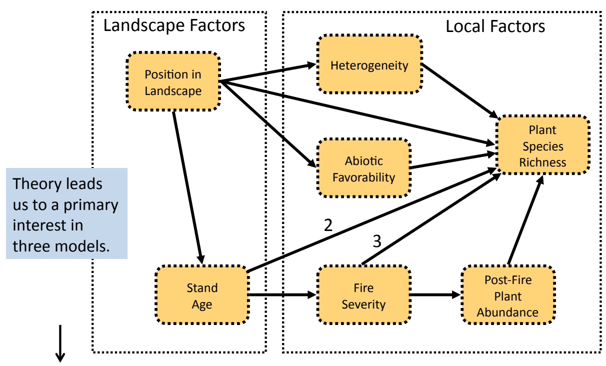
Observation: Post-fire Cover Declines with Age of Stand that Burned



Post-fire Vegetation Recovery Example (cont.):

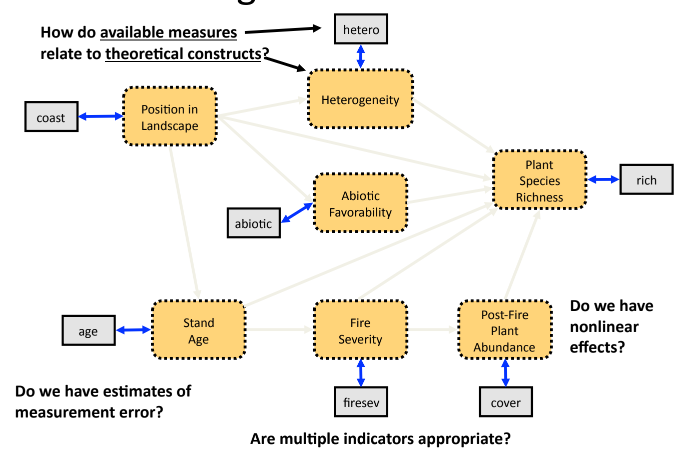


The SEMM

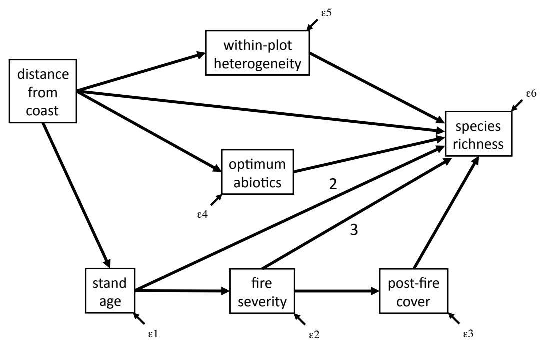


model 1 - includes all unnumbered paths
model 2 - possible loss of species from seedbank in older stands
model 3 - possible selective destruction of seeds in seedbank in severe fires

Matching the SEMM to Data

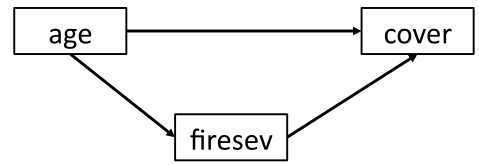


Realized Models with Data



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

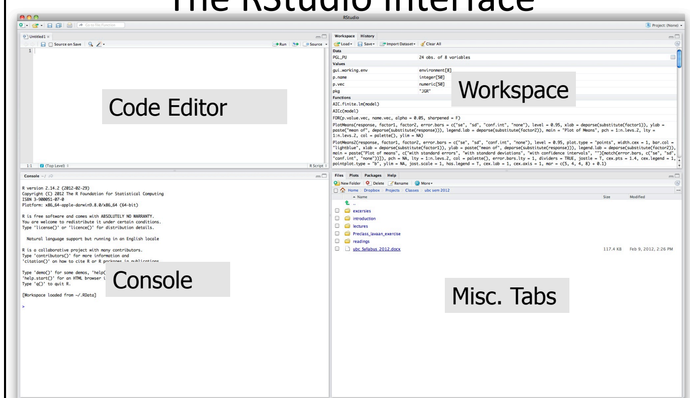
Our Simple Model for Today

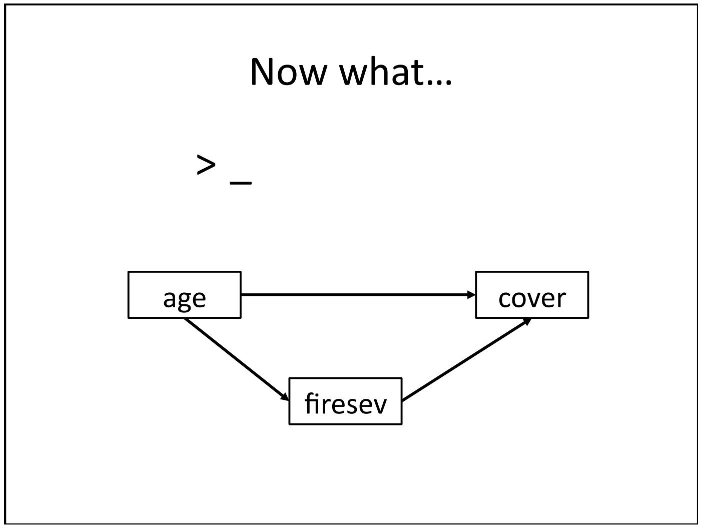


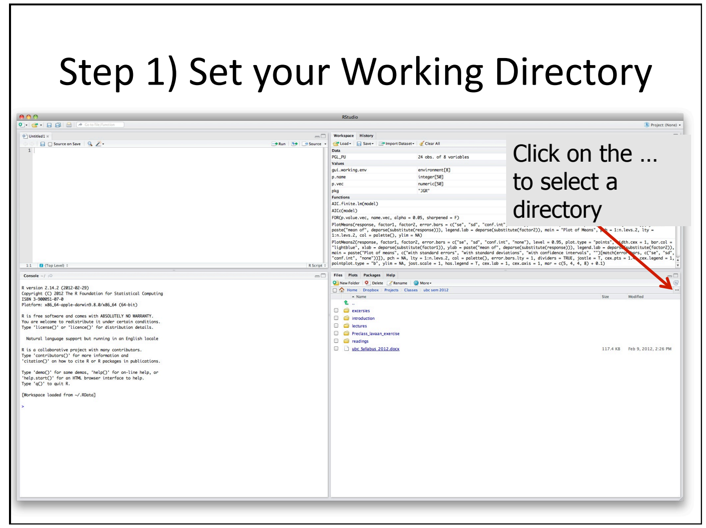
What Will You Need

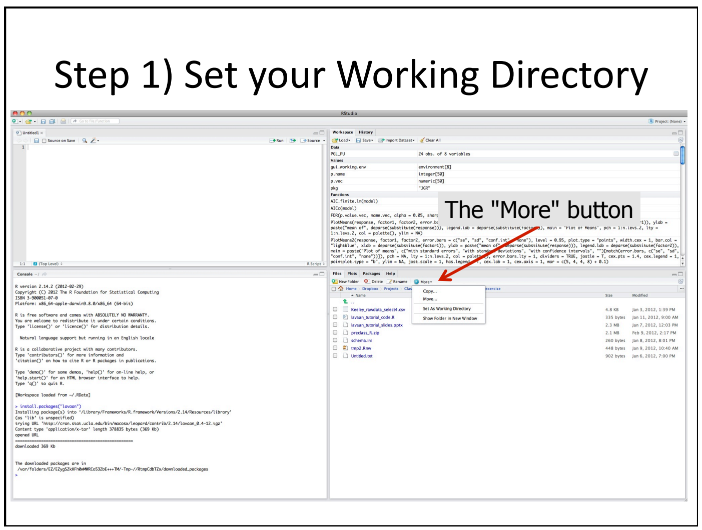
- R
  - Open Source Statistical Software
  - http://www.r-project.org
- Rstudio
  - A Great Integrated Development Environment
  - http://rstudio.r-project.org
- lavaan, car, QuantPsyc & ggm libraries
  - See tutorial for how to install a library

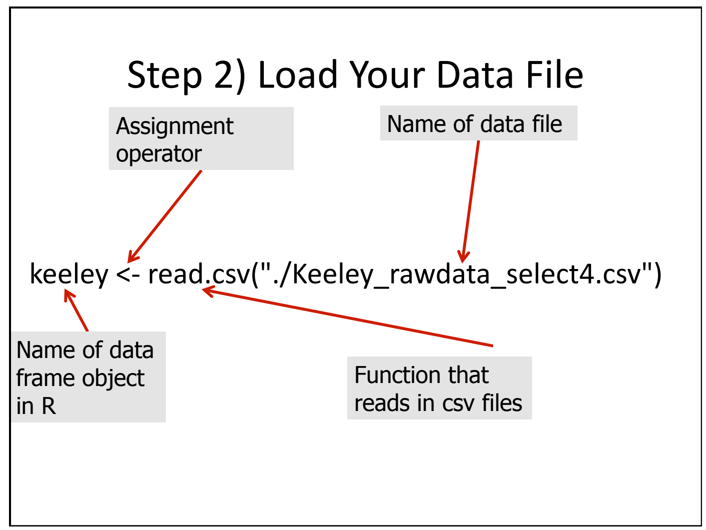
The RStudio Interface




Now what…
> _
age
cover
firesev


Step 1) Set your Working Directory
Click on the … to select a directory


Step 1) Set your Working Directory
The "More" button


Step 2) Load Your Data File
Assignment operator
Name of data file
keeley <- read.csv("./Keeley_rawdata_select4.csv")
Name of data frame object in R
Function that reads in csv files

Step 3) Check your Data in R

Column name, which we'll use as a variable later

```
> head(keeley)

  distance elev  abiotic age   hetero firesev     cover rich
1 53.40900 1225 60.67103  40 0.757065    3.50 1.0387974   51
2 37.03745   60 40.94291  25 0.491340    4.05 0.4775924   31
3 53.69565  200 50.98805  15 0.844485    2.60 0.9489357   71
4 53.69565  200 61.15633  15 0.690847    2.90 1.1949002   64
5 51.95985  970 46.66807  23 0.545628    4.30 1.2981890   68
6 51.95985  970 39.82357  24 0.652895    4.00 1.1734866   34
```

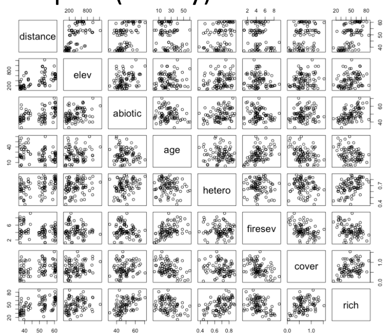
Step 3) Check your Data

```
> summary(keeley)

    distance          elev           abiotic            age            hetero      
 Min.   :37.04   Min.   :  60.0   Min.   :32.59   Min.   : 3.00   Min.   :0.3842  
 1st Qu.:39.46   1st Qu.: 202.5   1st Qu.:43.81   1st Qu.:15.00   1st Qu.:0.6246  
 Median :51.77   Median : 400.0   Median :48.04   Median :25.00   Median :0.6843  
 Mean   :49.23   Mean   : 424.7   Mean   :49.24   Mean   :25.57   Mean   :0.6833  
 3rd Qu.:58.40   3rd Qu.: 630.0   3rd Qu.:54.90   3rd Qu.:35.00   3rd Qu.:0.7684  
 Max.   :60.72   Max.   :1225.0   Max.   :70.46   Max.   :60.00   Max.   :0.8779  
    firesev          cover              rich          coverSQ            abioticSQ   
 Min.   :1.200   Min.   :0.05558   Min.   :15.00   Min.   :0.003089   Min.   :1062  
 1st Qu.:3.700   1st Qu.:0.48769   1st Qu.:37.00   1st Qu.:0.237873   1st Qu.:1920  
 Median :4.300   Median :0.63712   Median :50.00   Median :0.405919   Median :2307  
 Mean   :4.565   Mean   :0.69123   Mean   :49.23   Mean   :0.577321   Mean   :2483  
 3rd Qu.:5.550   3rd Qu.:0.91468   3rd Qu.:62.00   3rd Qu.:0.836645   3rd Qu.:3014  
 Max.   :9.200   Max.   :1.53541   Max.   :85.00   Max.   :2.357478   Max.   :4964  
```

- **Any missing values?**
- **Information on data types**

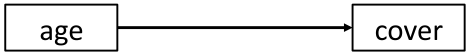
Step 4) View Your Data

> pairs(keeley)



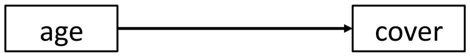
- Anything odd?
- Linear relationships?
- Normal distribution?
- Outliers?

Coding a Regression in R




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

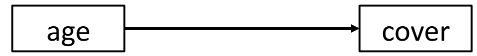
Does a Predictor Explain a Response?




```
#evaluation
library(car)
Anova(aLM)
```

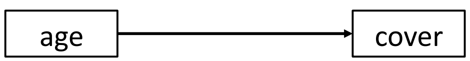
Venables, W. (1998). Exegeses on linear models. *S-Plus User's Conference, Washington DC.*

F-Table Output



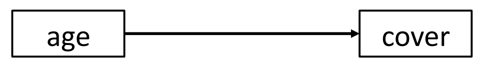

```
Anova Table (Type II tests)

Response: cover
          Sum Sq Df F value    Pr(>F)    
age       1.0998  1  12.318 0.0007097 ***
Residuals 7.8570 88                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Evaluating Coefficients



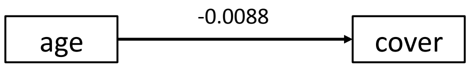

```
#getting coefficients
summary(aLM)
```

Evaluating Coefficients 1



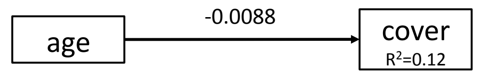

```
Call:
lm(formula = cover ~ age, data = keeley)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.50798 -0.24998 -0.03638  0.18407  0.75070 
```

Evaluating Coefficients 1



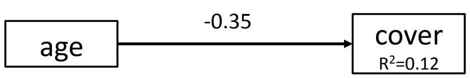

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.917395   0.071726   12.79  < 2e-16 ***
age         -0.008846   0.002520   -3.51  0.00071 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Evaluating Coefficients 3



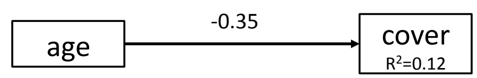

```
Residual standard error: 0.2988 on 88 degrees of freedom
Multiple R-squared: 0.1228,   Adjusted R-squared: 0.1128
F-statistic: 12.32 on 1 and 88 DF,  p-value: 0.0007097
```

Calculating a Standardized Coefficient



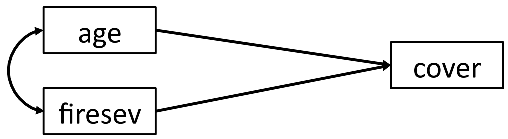

```
#standardized
> coef(aLM)[2]*sd(keeley$age)/sd(keeley$cover)

 age
-0.3504073
```

Calculating a Standardized Coefficient



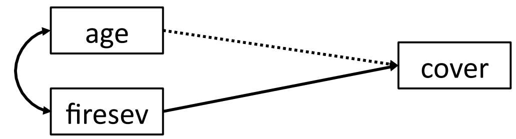

```
#a different approach
library(QuantPsyc)
lm.beta(aLM)

 age
-0.3504073
```

Multiple Regression



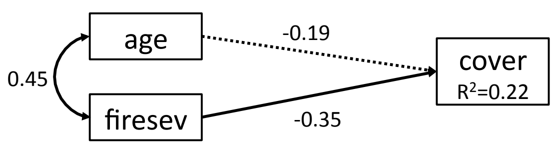

```
#regression
aLM2<-lm(cover ~ age+firesev,
  data=keeley)
```

Evaluating Effects in Multiple Regression



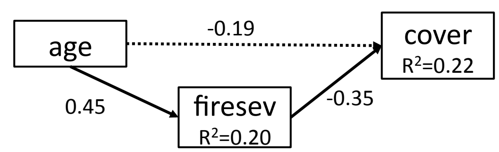

```
Anova Table (Type II tests)

Response: cover
          Sum Sq Df F value   Pr(>F)
age       0.2606  1  3.2466 0.075034 .
firesev   0.8724  1 10.8667 0.001418 **
Residuals 6.9846 87
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Standardized Coefficients



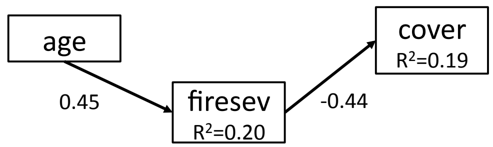

```
#Standardized Coefficients
cor(keeley$age, keeley$firesev)

summary(aLM2)$r.squared

lm.beta(aLM2)
```

Moving to SEM



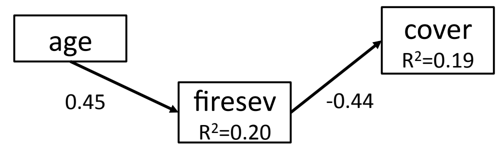

```
#Finish the SEM
aLM3<-lm(firesev ~ age, data=keeley)

Anova(aLM3)

summary(aLM3)$r.squared
```

Evaluating Full Mediation



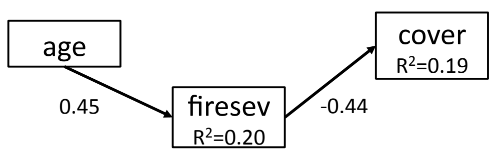

```
#Refit the new cover relationship
fullMedLM<-lm(cover ~ firesev, data=keeley)

summary(fullMedLM)$r.squared

lm.beta(fullMedLM)
```

Evaluating Full Mediation

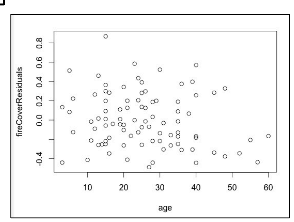



```
#evaluate the residual relationship
keeley$fireCoverResiduals<-residuals(fullMedLM)

residLM<-lm(fireCoverResiduals ~ age,
  data=keeley)

Anova(residLM)
```

Evaluating Full Mediation

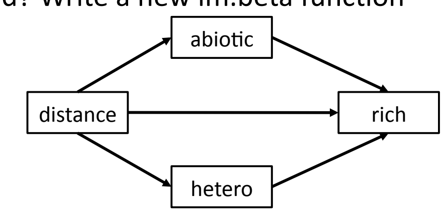



```
Anova Table (Type II tests)

Response: fireCoverResiduals
          Sum Sq Df F value Pr(>F)
age       0.2070  1  2.5876 0.1113
Residuals 7.0383 88
```

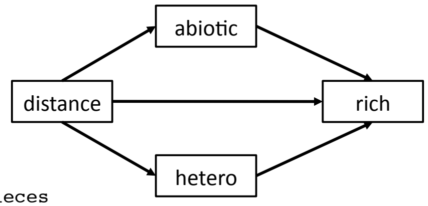


Exercise!

1. Fit and evaluate the following model
2. Fill in the standardized coefficients
3. Test for mediation for the distance -> richness relationship
4. Bored? Write a new lm.beta function

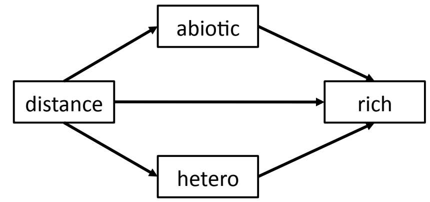


Fitting



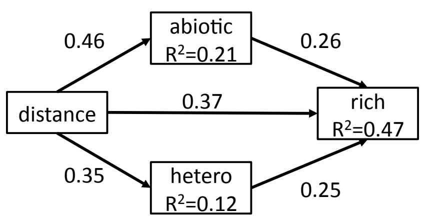

```
#fit the pieces
distanceLM <- lm(abiotic ~ distance, data=keeley)

heteroLM <- lm(hetero ~ distance, data=keeley)

richnessLM <- lm(rich ~ abiotic + distance + hetero,
  data=keeley)
```

Evaluation…



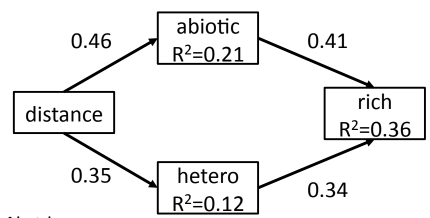

```
#evaluate the pieces
Anova(distanceLM)
Anova(heteroLM)
Anova(richnessLM)
```

Coefficients



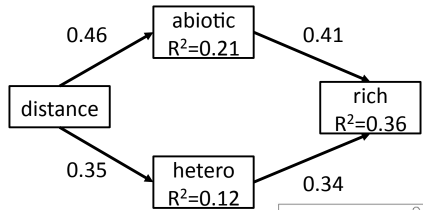

```
#standardized coefficients
lm.beta(distanceLM)
lm.beta(heteroLM)
lm.beta(richnessLM)
```

Distance -> Richness?

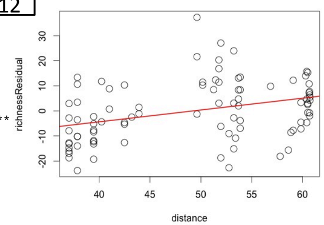



```
#Evaluate Mediation
richnessLM2 <- lm(rich ~ abiotic  + hetero, data=keeley)
keeley$richnessResidual<-residuals(richnessLM2)

richResidLM<-lm(richnessResidual ~ distance, data=keeley)
Anova(richResidLM)
```

Distance -> Richness?



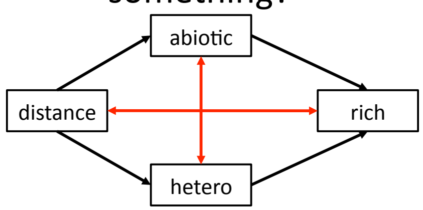

```
Anova Table (Type II tests)

Response: richnessResidual
          Sum Sq Df F value    Pr(>F)
distance    1545  1  12.016 0.0008177***
Residuals  11314 88
```



Outline

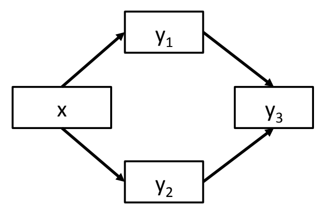
1. What is piecewise SEM?
2. Example using Keeley's fire data
   – Pre-SEM data screening
   – Combining models into an SEM
   – Simple evaluation of mediation
3. Evaluating model fit
   – D-Separation
4. D-Separation in R

How do we know we aren't missing something?



- Testing for 1 missing association simple
- How many possible additional paths are there?

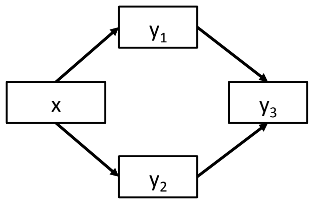
Directed Separation



- Concept from Graph Theory
- Two nodes are d-separated if they are conditionally independent
- $v_i \perp v_j \mid \{\mathbf{Pa_i} \cup \mathbf{Pa_j}\}; \mathbf{i > j}$
- $x \perp y3 \mid (y1, y2)$

Shipley, B. (2004). Analysing the allometry of multiple interacting traits. *Perspect Plant Ecol, 6, 235–241.*

Example: $x \perp y_3 \mid (y_1, y_2)$



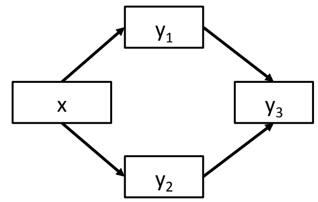
For example, for x and y_3:
(1) There are no common ancestors.
(2) We include the parents of y_3 that are part of mediating pathways: y_1 and y_2
(3) There are no common descendants to worry about.

Thus, their residuals are predicted to be uncorrelated.

The d-separation criterion for any pair of variables involves:

1. Controlling for common ancestors that could generate correlations between the pair
2. Controlling for causal connections through multi-link directed pathways via parents
3. Not controlling for common descendent variables.

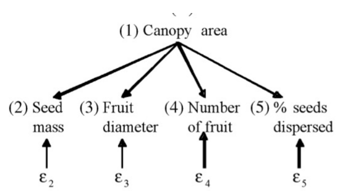
Which relationships to test: the *basis set*



The basis set is the smallest possible set of d-separation relationships from a graph.

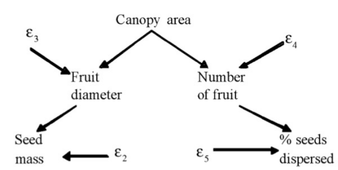
1. $x \perp y_3 \mid (y_1, y_2)$
2. $y_1 \perp y_2 \mid (x)$

Exercise: What is the basis set?



1. mass $\perp$ dia | (canopy)
2. mass $\perp$ # | (canopy)
3. mass $\perp$ %| (canopy)
4. dia $\perp$ # | (canopy)
5. dia $\perp$ % | (canopy)
6. % $\perp$ # | (canopy)

Exercise: What is the basis set?



1. canopy $\perp$ % | (#)
2. canopy $\perp$ mass | (dia)
3. dia $\perp$ # | (canopy)
4. dia $\perp$ % | (canopy, #)
5. mass $\perp$ #| (dia, canopy)
6. mass $\perp$ % | (dia, #)

Combining D-Sep Tests with Fisher's C

• An omnibus test for conditional independences across the entire model.

The test statistic is $C = -2*\sum \ln(p_i)$

where p_i = the p-values of all tests of conditional independence for all pairs of variables.

• p can come from various statistics. Typically Pearson or Spearman partial correlation, but can get more involved. See Shipley 2000.

• The statistic has a chi-square distribution on 2k degrees of freedom where k is the number of elements of the basis set. 65

What if p<0.05?

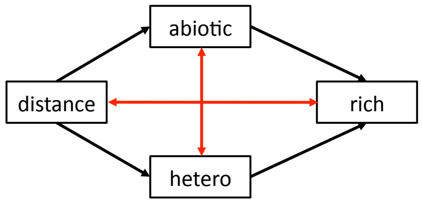
- You are likely missing some associations
- You can reject this model
- The way forward: adding links or different model structure?
- To re-iterate, p≥0.05 is GOOD

Questions?

Outline

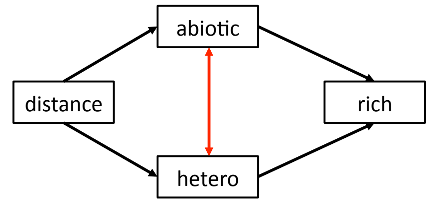
1. What is piecewise SEM?
2. Example using Keeley's fire data
   - Pre-SEM data screening
   - Combining models into an SEM
   - Simple evaluation of mediation
3. Evaluating model fit
   - D-Separation
4. D-Separation in R

Test of D-Separation, by hand



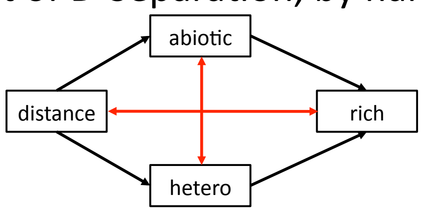

```
> summary(richnessLM)$coef
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -30.8880109  9.5340287 -3.239765 1.701573e-03
abiotic       0.5232920  0.1756398  2.979348 3.754034e-03
distance      0.6404318  0.1564575  4.093329 9.564005e-05
hetero       33.4010417 11.1186768  3.004048 3.489151e-03
```

Test of D-Separation, by hand




```
> bs2LM<-lm(abiotic ~ hetero + distance, data=keeley)
> summary(bs2LM)$coef
              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 25.4282211 5.14157086 4.945613 3.663633e-06
hetero       8.9334404 6.71896688 1.329585 1.871306e-01
distance     0.3596337 0.08737312 4.116068 8.735290e-05
```

Test of D-Separation, by hand



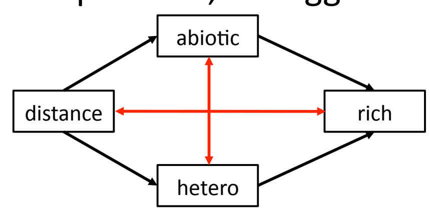

```
> #calculate C
> fisherC <- -2*(log(9.56e-05) + log(0.187))

> #the test
> 1-pchisq(fisherC, 4)
[1] 0.0002133105
```

p<0.05 means model missed something!

We would reject this model as inadequate

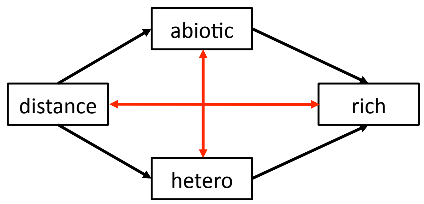
What happens next depends on goal of analysis

Test of D-Separation, with ggm: the DAG



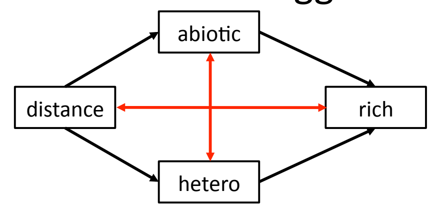

```
library(ggm)

#code the model into a matrix
modelMat<-DAG(abiotic ~ distance,
    hetero ~ distance,
    rich ~ abiotic + hetero)
```

Test of D-Separation, with ggm: the DAG



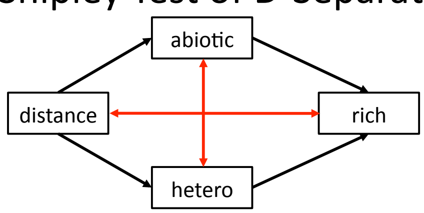

```
> modelMat
         abiotic distance hetero rich
abiotic        0        0      0    1
distance       1        0      1    0
hetero         0        0      0    1
rich           0        0      0    0
```

Basis Set with ggm



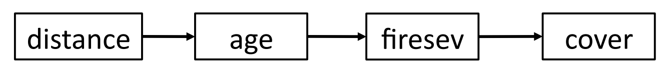

```
> basiSet(modelMat)
[[1]]
[1] "distance" "rich"     "hetero"   "abiotic"

[[2]]
[1] "hetero"   "abiotic"  "distance"
```

The Shipley Test of D-Separation



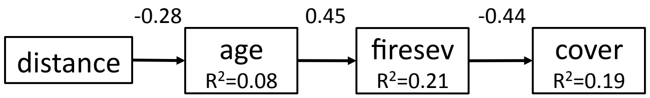

```
> shipley.test(modelMat, cov(keeley), n=nrow(keeley))
$ctest
[1] 21.86173

$df
[1] 4

$pvalue
[1] 0.0002135289
```

Exercise



1. **Examine for D-Separation**
2. **Evaluate model and coefficients**

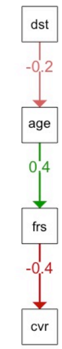
Solution




```
$ctest
[1] 12.34181

$df
[1] 6

$pvalue
[1] 0.05476368
```

semPlot

```
> library(semPlot)

> semPaths(lm1+lm2+lm3, layout="tree2",
  "std", intercepts=FALSE)
```



Take a break! Then come back and work on your own data!