Piecewise Fitting and Evaluation of SEMs with Observed Variables

Jarrett E. K. Byrnes

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

"There are no routine statistical questions, only questionable statistical routines"
- Sir David Cox

The Two Goals of Fittings SEMs

(a) Evaluation of the network
(b) Estimation of the parameters

Major questions about network:

1. Does theory support the causal structure?
2. Are there any missing connections?
3. Are there any unnecessary connections?
4. Are the data consistent with the model structure?
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:**

- \( x_1 \) to \( y_1 \)
- \( \gamma_{11} \) arrow from \( x_1 \) to \( y_1 \)
- \( \zeta_1 \) arrow from \( x_1 \) to \( y_1 \)

**ANOVA IS Linear Regression**

**Equation form:**

\[ y_1 = \gamma_{11} x_1 + \zeta_1 \]

**Graphical form:**

- \( x_1 \) to \( y_1 \)
- \( \gamma_{11} \) arrow from \( x_1 \) to \( y_1 \)
- \( \zeta_1 \) arrow from \( x_1 \) to \( y_1 \)

**Multiple Regression: Hidden Assumptions**

**Equation form:**

\[ y_1 = \gamma_{11} x_1 + \gamma_{21} x_2 + \zeta_1 \]

**Graphical form:**

- \( x_1 \) to \( y_1 \)
- \( x_2 \) to \( y_1 \)
- \( \gamma_{11} \) arrow from \( x_1 \) to \( y_1 \)
- \( \gamma_{21} \) arrow from \( x_2 \) to \( y_1 \)
- \( \zeta_1 \) arrow from \( x_1 \) to \( y_1 \)

*To estimate \( \gamma_{11} \) and \( \gamma_{21} \) you need to control for the unanalyzed correlation between \( x_1 \) and \( x_2 \).
*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

\[ y_1 = \gamma_{11}x_1 + \zeta_1 \]
\[ y_2 = \gamma_{12}x_1 + \beta_{13}y_1 + \zeta_2 \]

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 $y_2 \sim y_1$ relationship and $x_1$

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

---

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

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

- Measured vegetation recovery:
  - plant cover
  - species richness

---

Example: 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.)

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

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?

Do we have estimates of measurement error?

Are multiple indicators appropriate?

Do we have nonlinear effects?
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

What Will You Need

- R
  - Open Source Statistical Software
    - [http://www.r-project.org](http://www.r-project.org)
  
- Rstudio
  - A Great Integrated Development Environment
    - [http://rstudio.r-project.org](http://rstudio.r-project.org)

- lavaan, car, QuantPsyc & ggm libraries
  - See tutorial for how to install a library

Our Simple Model for Today

The RStudio Interface
Now what...

> _

age

cover

firesev

Step 1) Set your Working Directory

Click on the ... to select a directory

Step 2) Load Your Data File

```r
keeley <- read.csv("./Keeley_rawdata_select4.csv")
```

Assignment operator

Name of data file

Function that reads in csv files

Name of data frame object in R
Step 3) Check your Data in R

> head(keeley)

<table>
<thead>
<tr>
<th>distance</th>
<th>elev</th>
<th>abiotic</th>
<th>age</th>
<th>firesev</th>
<th>cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>51.40900</td>
<td>1225</td>
<td>60.67103</td>
<td>40</td>
<td>0.757065</td>
<td>1.0387974</td>
</tr>
<tr>
<td>37.03745</td>
<td>60</td>
<td>40.94291</td>
<td>25</td>
<td>0.491340</td>
<td>0.4775924</td>
</tr>
<tr>
<td>53.69565</td>
<td>200</td>
<td>50.98805</td>
<td>15</td>
<td>0.844485</td>
<td>0.9489357</td>
</tr>
<tr>
<td>53.69565</td>
<td>200</td>
<td>61.15633</td>
<td>15</td>
<td>0.690847</td>
<td>1.1949002</td>
</tr>
<tr>
<td>51.95985</td>
<td>970</td>
<td>46.66807</td>
<td>23</td>
<td>0.545628</td>
<td>1.2981890</td>
</tr>
<tr>
<td>51.95985</td>
<td>970</td>
<td>39.82357</td>
<td>24</td>
<td>0.652895</td>
<td>1.1734866</td>
</tr>
</tbody>
</table>

Step 3) Check your Data

> summary(keeley)

   distance    elev    abiotic  age     firesev
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 :0.6842
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
data types

   firesev    cover    rich
Min.   :1.200  Min.   :0.05558  Min.   :15.00
1st Qu.:3.700  1st Qu.:0.48769  1st Qu.:37.00
Median :4.300  Median :0.63712  Median :50.00
Mean   :4.565  Mean   :0.69123  Mean   :49.23
3rd Qu.:5.550  3rd Qu.:0.91448  3rd Qu.:62.00
Max.   :9.200  Max.   :1.53541  Max.   :85.00

• 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)


Evaluating Coefficients

#getting coefficients
summary(aLM)

F-Table Output

Anova Table (Type II tests)

Response: cover

<table>
<thead>
<tr>
<th>Sum Sq</th>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td>12.318</td>
<td>0.0007097 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>88</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Evaluating Coefficients 1

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

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.50798</td>
<td>-0.24998</td>
<td>-0.03638</td>
<td>0.18407</td>
<td>0.75070</td>
</tr>
</tbody>
</table>
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

Calculating a Standardized Coefficient

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

age
-0.3504073

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

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

age
-0.3504073
Multiple Regression

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

Evaluating Effects in Multiple Regression

```
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

```
cor(keeley$age, keeley$firesev)
summary(aLM2)$r.squared
lm.beta(aLM2)
```

Moving to SEM

```
aLM3<-lm(firesev~age, data=keeley)
Anova(aLM3)
summary(aLM3)$r.squared
```
Evaluating Full Mediation

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

summary(fullMedLM)$r.squared

lm.beta(fullMedLM)
```

Evaluating Full Mediation

- **Response:** fireCoverResiduals
  
  - **Sum Sq:** 0.2070
  - **Df:** 1
  - **F value:** 2.5876
  - **Pr(>F):** 0.1113
  - **Residuals:** 7.0383

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

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

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

Anova(residLM)
```
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(keeley$richnessResidual ~ distance, data=keeley)
Anova(richResidLM)
**Distance -> Richness?**

- $R^2 = 0.36$ for distance
- $R^2 = 0.12$ for hetero
- $R^2 = 0.21$ for abiotic

**Anova Table (Type II tests)**

<table>
<thead>
<tr>
<th>Sum Sq</th>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance</td>
<td>1545</td>
<td>12.016</td>
<td>0.000817</td>
</tr>
</tbody>
</table>

**How do we know we aren't missing something?**

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

**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

**Directed Separation**

- Concept from Graph Theory
- Two nodes are $d$-separated if they are conditionally independent
  - $v_i \perp v_j | \{Pa_i, Pa_j\}; i > j$
  - $x \perp y_3 | (y_1, y_2)$

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

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.

Thus, their residuals are predicted to be uncorrelated.

Exercise: What is the basis set?
1. mass $\perp$ dia $\mid$ (canopy)
2. mass $\perp$ # $\mid$ (canopy)
3. mass $\perp$ % $\mid$ (canopy)
4. dia $\perp$ # $\mid$ (canopy)
5. dia $\perp$ % $\mid$ (canopy)
6. % $\perp$ # $\mid$ (canopy)

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

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, $p\geq0.05$ is GOOD

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 | 5.1415708 | 4.945613 | 3.663633e-06 |
| hetero | 8.9334404 | 6.7189668 | 1.329585 | 1.871306e-01 |
| distance | 0.3596337 | 0.0873731 | 4.116068 | 8.735290e-05 |

> #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-Separa?on, with ggm: the DAG

\[
\begin{align*}
\text{abiotic} & \quad \text{distance} & \quad \text{hetero} \\
\text{distance} & \quad 0 & \quad 0 & \quad 1 \\
\text{hetero} & \quad 0 & \quad 0 & \quad 1 \\
\text{rich} & \quad 0 & \quad 0 & \quad 0 \\
\end{align*}
\]

Basis Set with ggm

\[
\begin{align*}
\text{abiotic} & \quad \text{distance} & \quad \text{hetero} & \quad \text{rich} \\
[1] & \quad \text{distance} & \quad \text{rich} & \quad \text{hetero} & \quad \text{abiotic} \\
[2] & \quad \text{hetero} & \quad \text{abiotic} & \quad \text{distance} \\
\end{align*}
\]

The Shipley Test of D-Separa?on

\[
\begin{align*}
\text{abiotic} & \quad \text{distance} & \quad \text{hetero} \\
\text{distance} & \quad \text{age} & \quad \text{firesev} & \quad \text{cover} \\
\end{align*}
\]

\[
\begin{align*}
\text{distance} & \quad \text{age} & \quad \text{firesev} & \quad \text{cover} \\
\end{align*}
\]

Exercise

1. Examine for D-Separa?on
2. Evaluate model and coefficients
Solution

\[ R^2 = 0.08 \]
\[ R^2 = 0.19 \]
\[ R^2 = 0.21 \]

```r
$ctest
[1] 12.34181$

```dfl
[1] 6

$\text{pvalue}$
[1] 0.05476368

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

```r
semPlot

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

```