Looking at Familiar Statistical Concepts in a New SEM Light

Jarrett E. K. Byrnes

Data from:
Biodiversity and complex environmental forcing of ecosystem functioning in the marine foundation species, eelgrass:
Matt Whalen, J. Emmett Duffy, Jim Grace

Old Wine in a New Bottle

1. ANOVA and ANCOVA in an SEM context
2. Multiple categorical predictors
3. Nonlinear effects
   - Squared
   - Interactions

York River, Virginia:
Major herbivores are invert crustaceans - these grazers control epiphytes and promote the eelgrass
Preliminary Study: Virginia site

Experimental Design:
- Pesticide to reduce crustacean grazers
- Nutrient addition
- Combination
- Controls
8 reps @ 5 trts = 40 plots

Pesticide effects:
- Crustaceans: reduced 58-96%
- Algal biomass: increased 130-748%
- Nutrients: inconsistent effects

Matt Whalen

Using Summarized Information

Carbaryl block

50 cm

PVC Anchor Poles

50 cm

Nutrient Diffuser

ASU

Using Summarized Information

Seagrass ANOVA Model

Seagrass ANOVA Model

- `whalenMeans <- c(0.4, 0.702, 1.374, 1.044, 2.374, -0.254)`
- `whalenN <- 40`

- `anovaModel<-'lnchla ~ pesticide'
- `anovaFit<-sem(anovaModel, sample.cov=whalenCov, sample.mean=whalenMeans, sample.nobs=whalenN)`

```
lower <= c(0.24, -0.104, 3.426, -0.0085, -0.163, 0.412, -0.207, 0.871, 0.062, 0.893, -0.527, 1.203, 0.098, 0.766, 1.92, 0.239, -0.312, 0.103, -0.35, -0.651, 0.466)
whalenCov <- getCov(lower, names=c("pesticide", "macroalgae", "grass", "LNCaprell", "LNGamm", "lnchla"))
```
Seagrass ANCOVA Model

AMOS v. lavaan

```
ancovaModel <- 'lnchla ~ pesticide + macroalgae + grass
pesticide ~~ 0*macroalgae + 0*grass'
ancovaFit <- sem(ancovaModel, sample.cov=whalenCov, sample.mean=whalenMeans, sample.nobs=whalenN, fixed=x=F)
```

The Models

```
fullModel <- 'lnchla ~ macroalgae + grass + LNGamm
LNGamm ~ macroalgae + grass + pesticide
pesticide ~~ 0*macroalgae + 0*grass'
partialModel <- 'lnchla ~ macroalgae + grass + LNGamm + pesticide
LNGamm ~ macroalgae + grass + pesticide
pesticide ~~ 0*macroalgae + 0*grass'
```

Mediation Exercise 2!

```
partialFit <- 'lnchla ~ macroalgae + grass + pesticide
LNGamm ~ macroalgae + grass + pesticide
pesticide ~~ 0*macroalgae + 0*grass'
```

Likelihood Ratio Comparison

```
> anova(fullFit, partialFit)
Chi Square Difference Test

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>Chisq</th>
<th>Chisq diff</th>
<th>Df</th>
<th>diff Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>partialFit</td>
<td>3</td>
<td>439.17</td>
<td>459.44</td>
<td>1.3654</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fullFit</td>
<td>4</td>
<td>441.71</td>
<td>460.29</td>
<td>5.9076</td>
<td>4.5422</td>
<td>1</td>
<td>0.03307 *</td>
</tr>
</tbody>
</table>
```

They are different
Partial Mediation Favored
Model selection based on AICc:

<table>
<thead>
<tr>
<th>K</th>
<th>AICc</th>
<th>Delta_AICc</th>
<th>AICcWt</th>
<th>Cum.Wt</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>partial</td>
<td>12</td>
<td>442.67</td>
<td>0.00</td>
<td>0.69</td>
<td>-207.58</td>
</tr>
<tr>
<td>full</td>
<td>11</td>
<td>444.26</td>
<td>1.59</td>
<td>0.31</td>
<td>-209.86</td>
</tr>
</tbody>
</table>

Model selection based on AIC:

<table>
<thead>
<tr>
<th>K</th>
<th>AIC</th>
<th>Delta AIC</th>
<th>AICWt</th>
<th>Cum.Wt</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>partial</td>
<td>12</td>
<td>439.17</td>
<td>0.00</td>
<td>0.78</td>
<td>-207.58</td>
</tr>
<tr>
<td>full</td>
<td>11</td>
<td>441.71</td>
<td>2.54</td>
<td>0.22</td>
<td>-209.86</td>
</tr>
</tbody>
</table>

Summary

- Information criteria are suggestive
- Both state that the direct link is a better model, but hard to say...
- LR Test shows that partial mediation model is a better fit to the data
- Given that we may have a 2nd mediator (caprellids), this may lead to weaker performance of AICs

Old Wine in a New Bottle

1. ANOVA and ANCOVA in an SEM context
2. Multiple categorical predictors
3. Nonlinear effects
   - Squared
   - Interactions
What about experiment with more than 2 levels of treatment?

1. Can you make the treatment continuous?
   - E.g. nutrient levels

2. Or, treat each level as being present/absent
   \[ y = \gamma_1 x_1 + \gamma_2 x_2 + \zeta \]
   where \( x_i = 0 \) or \( 1 \)

Experiment with 3 Levels

- Exogenous covariance no longer 0.
  \[
  \begin{array}{ccc}
  x_1 & x_2 & x_3 \\
  1.0 & -0.5 & -0.5 \\
  \\
  x_2 & -0.5 & 1.0 & -0.5 \\
  \\
  x_3 & -0.5 & -0.5 & 1.0 \\
  \end{array}
  \]

Cannot Include All 3 Variables

- This matrix is singular
  \[
  \begin{array}{ccc}
  x_1 & x_2 & x_3 \\
  1.0 & -0.5 & -0.5 \\
  \\
  x_2 & -0.5 & 1.0 & -0.5 \\
  \\
  x_3 & -0.5 & -0.5 & 1.0 \\
  \end{array}
  \]

- If you know \( x_1 \) and \( x_2 \), you know the state of \( x_3 \)

Coefficient judged relative to effect of missing variable
Does Diet Affect Urchin Gonad Development?

- Rhodymenia
- Macrocystis Mixture
- Urchins feeding measured over 6 months
- Gonads and body size assessed at end
- All consumption rates converted to g dry carbon

Urchin Gonad Development Model

- Note that the polyculture is not included.
- Results judged relative to polyculture.

How do we use a categorical variable?

```r
> urchinData <- read.csv("./urchin_ex_sem.csv")
> summary(urchinData)
```

Box: treatment
- Min. :1 MAPY:7
- 1st Qu.:10 POLY:7
- Median :18 R :7
- Mean :18
- 3rd Qu.:26
- Max. :35
#Make treatment into a series of binary variables
source("./makeBinaryTreatments.R")

binTrt <- makeBinaryTreatments(urchinData, "treatment")

head(binTrt)

MAPY POLY R
1 1 0 0
2 0 0 1
3 0 1 0
4 1 0 0
5 0 0 1
6 0 1 0

> cor(binTrt)
   MAPY POLY R
MAPY 1.0 -0.5 -0.5
POLY -0.5 1.0 -0.5
R -0.5 -0.5 1.0

> solve(cor(binTrt))
Error in solve.default(cor(binTrt)):
  Lapack routine dgesv: system is exactly singular

#add new columns to data frame
urchinData <- cbind(urchinData, binTrt)

urchinModel <-
  "Feeding.rate.dry ~ MAPY + R ~ MAPY + R + Feeding.rate.dry",

urchinSEM <- sem(urchinModel, data=urchinData)
Fitting the Model

```r
# Add new columns to data frame
urchinData <- cbind(urchinData, binTrt)

urchinModel <-
  Feeding.rate.dry ~ MAPY + R
  GONAD_INDEX ~ MAPY + R + Feeding.rate.dry

urchinSEM <- sem(urchinModel, data = urchinData)
```

lavaan (0.5-17) converged normally after 40 iterations

```
Number of observations Used  Total
20     21

Estimator ML
Minimum Function Test Statistic 0.000
Degrees of freedom 0
```

The Fit

```
MAPY ~ MAPY
R ~ R
GONAD_INDEX ~ MAPY
Feeding.Rate.dry ~ Feeding.Rate.dry
```

| Regression | Estimate | Std.err | Z-value | P(>|z|) | Std.lv | Std.all |
|------------|----------|---------|---------|---------|--------|---------|
| Feeding.Rate.dry ~ MAPY | -0.001 | 0.001 | -1.083 | 0.279 | -0.001 | -0.229 |
| R ~ MAPY | 0.002 | 0.001 | 2.013 | 0.044 | 0.002 | 0.425 |
| GONAD_INDEX ~ MAPY | -0.009 | 0.008 | -1.038 | 0.299 | -0.009 | -0.171 |
| R ~ MAPY | -0.041 | 0.009 | -4.644 | 0.000 | -0.041 | -0.814 |
| Feeding.rt.dry ~ MAPY | -1.027 | 2.218 | -0.463 | 0.643 | -1.027 | -0.078 |
```

Interpretation Relative to Mixture

1. *Rhodymenia* is not good food.
   - Urchins eat more, but produce less gonad
2. Performance is similar with *Macrocystis* or Mixture diet
1. *Rhodymenia* is not good food.
   - Urchins eat more, but produce less gonad
2. Performance is similar with *Macrocystis* or Mixture diet (and = intercept)

> urchinSEM2<-'
  Feeding.rate.dry ~ MAPY + R
  GONAD_INDEX ~ MAPY + R
  ,
  urchinSEM2<-sem(urchinModel2, data=urchinData)
lavaan Default Behavior

> urchinSEM2
> lavaan (0.5-17) converged normally after 73 iterations

Number of observations	20          21
Estimator	ML
Minimum Function Test Statistic	0.000
Degrees of freedom	0
P-value (Chi-square)	1.000

One Solution: Turn Off Behavior

> urchinSEM2 <- lavaan(urchinModel2, data = urchinData, auto.cov.y = FALSE, auto.var = TRUE)

Or, Code the Lack of Covariance

urchinModel2a <- 

Feeding.rate.dry ~ MAPY + R
GONAD_INDEX ~ MAPY + R
GONAD_INDEX ~~ 0*Feeding.rate.dry

OK – let’s Test for Mediation!

> anova(urchinSEM, urchinSEM2a)
Chi Square Difference Test

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>Chiq</th>
<th>Chiq diff</th>
<th>Diff Pr (&gt;Chiq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>urchinSEM</td>
<td>-254.32</td>
<td>-247.35</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>urchinSEM2a</td>
<td>-256.10</td>
<td>-250.12</td>
<td>0.2245</td>
<td>0.22447</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Old Wine in a New Bottle

1. ANOVA and ANCOVA in an SEM context
2. Multiple categorical predictors
3. Nonlinear effects
   - Squared
   - Interactions

Nonlinearities in Observed Variable SEM

- Nonlinearities are just another variable
- But, nonlinearities may be collinear with their predictor
- Incorporate collinearities into path structure (simple for exogenous variables)
- If necessary ($r>0.9$), consider centering variables before transforming
- However, best solution for nonlinearities is a larger sample size & range of values!

Does Diversity = Productivity of vice-versa?

A Multi-Stream Experiment

Cardinale et al. 2009
Nonlinear Relationship Between Nutrient Addition and Richness

Create a Nonlinear Variable

# read in the data
cards <- read.csv("./cardinale_et_al_2009.csv")

# make a new nonlinear column
cards$logN2 <- cars$logN^2

Nonlinear Nutrient Effect on Richness

Note that Treatment’s Doesn’t Covary with Regional Richness

# read in the data
cards <- read.csv("./cardinaleet_al_2009.csv")

# make a new nonlinear column
cards$logN2 <- cards$logN^2

cardModel <- 'SA ~ logN + logN2 + SR
logChl ~ SA + logN
logN ~ logN2
'
cardFit <- sem(cardModel, data=cards, fixed.x=F)

Cardinale et al 2009
Model Fits Quite Well

```
lavaan (0.4-12) converged normally after 64 iterations

Number of observations                           127
Estimator                                         ML
Minimum Function Chi-square                    0.545
Degrees of freedom                                 4
P-value                                        0.969
Cardinale et al 2009
```

But...no Nutrient Effect?

Centering helps remove collinearities.

Warning: it does change interpretation of interaction effects.

```
Regressions:
logN ~ SA                0.097    0.023    4.194    0.000    0.097    0.321
logN2                  -0.475    0.456   -1.041    0.298   -0.475   -0.506
SR                      0.384    0.035   10.859    0.000    0.384    0.688
logChl ~ SA                0.020    0.005    3.768    0.000    0.020    0.393
logN                      0.097    0.023    4.194    0.000    0.097    0.321
Cardinale et al 2009
```

Refit with Centered Nutrients

```
Refit with Centered Nutrients

#make a new nonlinear column
cards$logN2Cen <- (cards$logN-mean(cards$logN))^2

Regressions:
logN ~ centered logN2Cen  0.368    0.452    0.815    0.415    0.368    0.062
logN2Cen               -0.475    0.240   -1.976    0.048   -0.475   -0.147
SR                      0.384    0.035   10.859    0.000    0.384    0.688
logChl ~ SA                0.020    0.004    4.667    0.000    0.020    0.393
logN                      0.097    0.023    4.194    0.000    0.097    0.321
Cardinale et al 2009
```
What about Interactions?

What if Age’s Effect was Mediated *AND* Moderated?

\[
\text{partialMedModel_int<-'} \text{firesev} \sim \text{age} \\
\text{cover} \sim \text{firesev} + \text{age} + \text{firesev:age'}
\]

Well That Doesn’t Fit…

Although a Significant Interaction

Number of observations                            90
Estimator                                         ML
Minimum Function Test Statistic              121.892
Degrees of freedom                                 1
P-value (Chi-square)                           0.000

Regressions:

|                  | Estimate | Std.Err | z-value | P(>|z|) |
|------------------|----------|---------|---------|---------|
| firesev ~        |          |         |         |         |
| age              | 0.060    | 0.012   | 4.832   | 0.000   |
| cover ~          |          |         |         |         |
| firesev          | -0.015   | 0.020   | -0.751  | 0.453   |
| age              | 0.005    | 0.005   | 0.941   | 0.347   |
| firesev:age      | -0.002   | 0.001   | -3.073  | 0.002   |
We Have a Collinearity Problem

```
> modificationindices(partialMedFit_int)
   lh  op     rhs         mi     epc  sepc.lv  sepc.all  sepc.nox
... 
12 firesev ~ firesev:age 66.770  0.030  0.030    1.550    0.018
16 firesev:age ~ firesev 53.016  19.871  19.871    0.380    0.380
```

We Have a Collinearity Problem

```
> cor(keeley$firesev, keeley$firesev*keeley$age)
[1] 0.7743624
```

Centering an Interaction

```
keeley$fire_age_int <- with(keeley, 
  scale(firesev, scale=FALSE) * 
  scale(age, scale=FALSE))

#model
partialMedModel_int2<-' firesev ~ age
  cover ~ firesev + age + fire_age_int'
```

Centered Interaction Fits

```
lavaan (0.5-22) converged normally after  31 iterations
Number of observations                            90
Estimator                                         ML
Minimum Function Test Statistic                0.121
Degrees of freedom                                 1
P-value (Chi-square)                           0.728
```
But...

Regressions:

|            | Estimate | Std.Err | z-value | P(|z|) |
|------------|----------|---------|---------|--------|
| firesev ~ age | 0.060    | 0.012   | 4.832   | 0.000  |
| cover ~ firesev | -0.068  | 0.020   | -3.455  | 0.001  |
| age        | -0.005   | 0.003   | -1.923  | 0.055  |
| fire_age_int | -0.002  | 0.001   | -1.562  | 0.118  |

Also, Careful in Interpreting Results of Centering

- In uncentered model, additive paths estimate the effect of one variable in the absence of the other.
- In centered model, additive paths estimate the effect of one variable at the average level of the other.
- E.g., fire severity has no effect when age = 0 versus fire severity has an effect at the mean level of age

Questions?

Afternoon Lab...

1. With your data, draw up some models.
   - See me if you don’t have data to work with
   - First, try a conceptual model
   - Second, try some models using the data you have at hand
2. Evaluate what type of model you have and if it is identified.
3. Try and fit one of your models using either piecewise or ML techniques.