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?

$\delta_x \rightarrow X \xrightarrow{\lambda_x} \xi$

$\xi$: A latent variable is a variable that is unmeasured, but is hypothesized to exist

$\lambda_x$: The relationship between a latent variable and its observed counterpart
What is a latent variable?

\[ \delta_x \rightarrow X \xrightarrow{\lambda_x} \xi \]

\[ \delta_x : \text{The error in the measurement of } x \text{ by } \xi \]

Latent Exogenous Variables

\[ \delta_x \rightarrow X \xrightarrow{\lambda_x} \xi \]

\[ X = \lambda_x \xi + \delta_x \]

Latent Endogenous Variables

\[ \epsilon_y \rightarrow Y \xrightarrow{\lambda_y} \eta \]

\[ y = \lambda_y \eta + \epsilon_y \]

Latent Endogenous Variables

\[ \epsilon_y \rightarrow Y \xrightarrow{\lambda_y} \eta \]

\[ \zeta : \text{Variance in response to predictors} \]
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

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.”
Indicators may Covary for Other Reasons

\[ X_1 \xrightarrow{\delta_{x1}} X_1 \xrightarrow{\lambda_{x1}} \xi \]
\[ X_2 \xrightarrow{\delta_{x2}} X_2 \xrightarrow{\lambda_{x2}} \xi \]
\[ X_3 \xrightarrow{\delta_{x3}} X_3 \xrightarrow{\lambda_{x3}} \xi \]

Indicators may Covary for Causal Reasons

\[ X_1 \xrightarrow{\delta_{x1}} X_1 \xrightarrow{\lambda_{x1}} \xi \]
\[ X_2 \xrightarrow{\delta_{x2}} X_2 \xrightarrow{\lambda_{x2}} \xi \]
\[ X_3 \xrightarrow{\delta_{x3}} X_3 \xrightarrow{\lambda_{x3}} \xi \]

Maybe respecify your model?

What if One Measurement Alone isn't Very Good?

\[ \delta_1 \xrightarrow{\text{Inaccurate Machine}} \delta_2 \xrightarrow{\text{N Strip Color}} \delta_3 \xrightarrow{\text{Stu's Tasted N}} \]

\[ \text{Nitrogen} \]

Different Measurements

\[ \delta_{x1} \xrightarrow{\% Cover} \]
\[ \delta_{x2} \xrightarrow{\text{Quadrat Densities}} \]
\[ \delta_{x3} \xrightarrow{\text{Band Transects}} \]

\[ \text{Algal Cover} \]
Multiple Properties

- $\delta_{x1}$: Algae Lost
- $\delta_{x2}$: Change in Grazer Density
- $\delta_{x3}$: Bites per alga

Grazing Disturbance

Repeated Measures

- $\delta_{x1}$: Fish at t1
- $\delta_{x2}$: Fish at t2
- $\delta_{x3}$: Fish at t3

Fish Density

Multi-Sample

- $\delta_{x1}$: Count by Observer 1
- $\delta_{x2}$: Count by Observer 2
- $\delta_{x3}$: Count by Observer 3

Fish Density

Latent Variables as Theoretical Constructs

Storm Disturbance
Latent Variables as Theoretical Constructs

\[ \delta_1 \rightarrow \text{Wave Heights} \]
\[ \delta_2 \rightarrow \text{Precipitation Intensity} \]
\[ \delta_3 \rightarrow \text{Onshore Detritus} \]

\(\delta\)'s are variation in observations not explained by storm disturbance

---

Concept Validity: The Fallacy of Naming

\[ \delta_1 \rightarrow \text{Sun Star Density} \]
\[ \delta_2 \rightarrow \text{Lobster Density} \]
\[ \delta_3 \rightarrow \text{Sheephead Density} \]

\(\delta\)'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
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

Evaluating Whether Indicators Will Make a Good Latent Variable

<table>
<thead>
<tr>
<th>Correlations among candidate indicators tell us whether data is consistent with what is implied by our model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note correlations are all strong, but not all equally strong. This shows us that these are not redundant indicators that are completely interchangeable.</td>
</tr>
<tr>
<td>In particular, variables y4 and y5 are more strongly correlated with each other than with the other vars.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observed Correlations:</th>
<th>y1</th>
<th>y2</th>
<th>y3</th>
<th>y4</th>
<th>y5</th>
</tr>
</thead>
<tbody>
<tr>
<td>y1</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y2</td>
<td>0.933</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y3</td>
<td>0.813</td>
<td>0.834</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>y4</td>
<td>0.773</td>
<td>0.728</td>
<td>0.693</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>y5</td>
<td>0.730</td>
<td>0.646</td>
<td>0.603</td>
<td>0.969</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Fixing Parameters for Identifiability

1. 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.

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.

NOT IDENTIFIED.

What drives the evolution of warning coloration?

Toxicity?  Diet?  Body condition?

Example: Aposematism in Poison Dart Frogs


A Phylogenetic Approach to SEM using Independent Contrasts

Santos and Cannatella 2011 PNAS
**Aposematism as a Latent Variable**

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**

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

santosSize<-'
santosSizeFit<-sem(santosSize, sample.cov=santosCov, sample.nobs=21)
Exercise: Fitting Latent Variables

Latent variables:
- Size
- Log.Mass
- Log.RMR
- Log.Scope

| 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   | 10.771  | 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.

Phosphorous
Nitrogen
Carbon
pH
Grain Size
Humus
Soil Factor 1
Soil Factor 2

Principle Components Analysis

- Factors do not correlate, no error estimated
- Note change in direction of causality

Phosphorous
Nitrogen
Carbon
pH
Grain Size
Humus
Soil Factor 1
Soil Factor 2

Confirmatory Factor Analysis

- Tests whether variables separate into groups
- Useful to test a "measurement model" for SEM
- $\Phi_{12}$ 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

\[
\begin{align*}
\delta_{x_1} & \rightarrow x_1 \\
\delta_{x_2} & \rightarrow x_2 \\
\delta_{x_3} & \rightarrow x_3 \\
\delta_{x_4} & \rightarrow x_4 \\
\lambda_{x_1} & \rightarrow \xi_1 \\
\lambda_{x_2} & \rightarrow \xi_2 \\
\lambda_{x_3} & \rightarrow \xi_2 \\
1 & \rightarrow \xi_1 \\
1 & \rightarrow \xi_2
\end{align*}
\]

Shared Indicator Rule - NECESSARY

Empirical Underidentification Still Possible

\[
\begin{align*}
\delta_{x_1} & \rightarrow x_1 \\
\delta_{x_2} & \rightarrow x_2 \\
\delta_{x_3} & \rightarrow x_3 \\
1 & \rightarrow \xi
\end{align*}
\]

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, 
  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 Example: The general performance of transplanted plants as a function of their genetic dissimilarity to local populations.

From:

The Theory Driving the Modeling

Theory suggests following for transplanted Spartina.

Path Distance
Latitude
Genetic Distance
Performance
Distance effects
Gene Flow
But what do we mean by performance?

Performance as a latent construct

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

Simply linking a bunch of measures to a latent variable does not mean you have correctly specified the model.

Stay focused on the causes of an indicator to aid latent variable model specification.

Model hypothesizes 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.
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, Latent Indicators, and Regression

```lavaan
clonediam -> performance
performance =~ clonediam
```

Assessing Fit

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

Coefficients

```
Estimate Std.err Z-value P(>|z|) Std.lv Std.all
Latent variables:
performance =~
  clonediam 1.000
  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**

---

**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

```r
genericModel <- readRDS('genericModel.rds')

#Fit model
set.seed(123)
set.seed(123)

#Fit meanht ~ latitude
meanht ~ latitude

#Fit meanwidth ~ latitude
meanwidth ~ latitude
```
Latent Variables and Measurement Error

True Cover Of Algae → 1 → Estimated Cover of Algae → Error in Measurement of Cover

LANDSAT Measurements of Kelp Canopy

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

LANDSAT Kelpiness vs. Canopy Biomass

Canopy biomass = kelpiness * 154.89 - 68.62

\[ R^2 = 0.62 \]
We can transform satellite data to canopy biomass, and fix the unstandardized loading to 1.

![Diagram of LANDSAT Kelpiness vs. Canopy Biomass]

But what about error?

We know that $R^2 = 1 - \text{estimated var/observed var}

$$\delta_x = (1 - R^2)$$

Unstandardized Measurement Error = $\delta_x \cdot \text{var(Measured Canopy Biomass)}$

Let's Look at the LTER data: Data Prep

```r
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 $y = 0.08x + 0.01 \text{ } r^2 = 0.79$
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)
```

Fit this Model!

![Diagram of LANDSAT Kelpiness vs. Canopy Biomass]

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

(unstandardized coefficients)
Incorporating measurement error in the predictor increases the coefficient size and amount of variance explained.

Exercise: Code this model!

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.
OMG!!
I has doodlebobbers!