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?

$\xi$: 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?

\[ \delta_x \rightarrow X \rightarrow \lambda_x \rightarrow \xi \]

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

Latent Exogenous Variables

\[ \delta_x \rightarrow X \rightarrow \lambda_x \rightarrow \xi \]

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

Latent Endogenous Variables

\[ \varepsilon_y \rightarrow Y \rightarrow \lambda_y \rightarrow \eta \]

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

Latent Endogenous Variables

\[ \varepsilon_y \rightarrow Y \rightarrow \lambda_y \rightarrow \eta \]

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

**What if One Measurement Alone isn't Very Good?**

![Diagram](image)

**Latent Variables as Theoretical Constructs**

![Image of 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
“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

Latent Variables

1. What is a latent variable?
2. Latent variables with multiple indicators
3. Factor Analysis
4. Latent Variables as a Response
5. Coping with measurement error
Latent Variables with Multiple Indicators

\[ \delta_{x1} \rightarrow X1 \rightarrow \lambda_{x1} \]
\[ \delta_{x2} \rightarrow X2 \rightarrow \lambda_{x2} \]
\[ \delta_{x3} \rightarrow X3 \rightarrow \lambda_{x3} \]

Latent variables represent 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.”

Latent variables represent shared information of indicators

Indicators may covary for other reasons
Indicators may Covary for Causal Reasons

\[ \delta_x \rightarrow X1 \rightarrow \xi \]
\[ \delta_x \rightarrow X2 \rightarrow \xi \]
\[ \delta_x \rightarrow X3 \rightarrow \xi \]

Maybe respecify your model?

Different Measurements

\[ \delta_x \rightarrow %\ Cover \rightarrow Algal\ Cover \]
\[ \delta_x \rightarrow Quadrat\ Densities \rightarrow \]
\[ \delta_x \rightarrow Band\ Transects \rightarrow \]

Multiple Properties

\[ \delta_x \rightarrow Algae\ Lost \]
\[ \delta_x \rightarrow Change\ in\ Grazer\ Density \]
\[ \delta_x \rightarrow Bites\ per\ alga \]

Grazing Disturbance

Repeated Measures

\[ \delta_x \rightarrow Fish\ at\ t1 \rightarrow Fish\ Density \]
\[ \delta_x \rightarrow Fish\ at\ t2 \rightarrow \]
\[ \delta_x \rightarrow Fish\ at\ t3 \rightarrow \]
Multi-Sample

\[ \delta_{x1} \rightarrow \text{Count by Observer 1} \]
\[ \delta_{x2} \rightarrow \text{Count by Observer 2} \]
\[ \delta_{x3} \rightarrow \text{Count by Observer 3} \]

Fish Density

Evaluating Whether Indicators Will Make a Good Latent Variable

**Observed Correlations:**

<table>
<thead>
<tr>
<th></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>

1. Correlations among candidate indicators tell us whether data is consistent with what is implied by our model.
2. Note correlations are all strong, but not all equally strong. This shows us that these are not redundant indicators that are completely interchangeable.
3. In particular, variables y4 and y5 are more strongly correlated with each other than with the other vars.

Fixing Parameters for Identifiability

1. Note 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 known 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.
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

Example: Aposematism in Poison Dart Frogs


What drives the evolution of warning coloration?

Toxicity? Diet? Body condition?
A Phylogenetic Approach to SEM using Independent Contrasts

Santos and Canestrelli 2011 PNAS

SantosCov <- read.table("santosCov.txt", na.strings=".")
SantosCov <- as.matrix(santosCov)
santosCFA1 <- "Aposematism =~ Alkaloid.quantity + Alkaloid.diversity + Conspicuous.coloration"
santosFit1 <- sem(santosCFA1, sample.cov = SantosCov, sample.nobs = 21)

Estimate Std.err  Z-value  P(>|z|)
Latent variables:
Aposematism 1.000
Alkaloid.quantity 2.358 0.558 4.223 0.000
Alkaloid.diversity 0.753 0.215 3.502 0.000
Conspicuous.coloration 0.593 0.225 2.625 0.009

Variances:
Alkaloid.quantity -0.000 0.000 -0.190 0.849
Alkaloid.diversity 0.004 0.002 2.376 0.018
Conspicuous.coloration 0.001 0.000 2.905 0.004
Aposematism 0.001 0.000 2.915 0.004

Aposematism as a Latent Variable

Santos & Canestrelli 2011 PNAS

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
Exercise: Fitting Latent Variables

\[
\begin{align*}
\delta_{x1} & \rightarrow \text{Log.Mass} \\
\delta_{x2} & \rightarrow \text{Log.RMR} \\
\delta_{x3} & \rightarrow \text{Log.Scope}
\end{align*}
\]

\[
\lambda_{x1} \quad \lambda_{x2} \quad \lambda_{x3}
\]

\[
\text{Size}
\]

\[
santosSize <- \text{`Size =~ Log.Mass + Log.RMR + Log.Scope'}
\]

\[
santosSizeFit <- \text{sem(santosSize, sample.co=santosCov, sample.nobs=21)}
\]

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)

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

Exploratory Factor Analysis

- Observed variables explained by correlated factors
- Error in observed variables
- But...exploratory - NOT HYPOTHESIS TESTING.

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

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!

Empirical Underidentification Still Possible

{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

Performance as a latent construct

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

Be aware that simply linking a bunch of measures to a latent variable does not mean you have correctly specified the model.

One needs to stay focused on the question of “what actually causes the responses I see?” This is how you arrive at proper model specification, not simply through applying the model structures to your data.

Note this model hypothesizes we have 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.

Also, performance is represented as a unidimensional cause (not a collection of disparate causes).
lavaan, Latent Indicators, and Regression

```
clonediam ~ performance
performance ~ clonediam
```

Assessing Fit
```
> spartinaFit<-sem(spartinaModel, data=spartina)
> summary(spartinaFit, standardized=T, rsquare=T)
lavaan (0.5-12) converged normally after 154 iterations

Number of observations                       23
Estimator                                     ML
Minimum Function Test Statistic               12.237
Degrees of freedom                            8
P-value (Chi-square)                          0.141
```

Using Latent Variables in SEM
```
geneticdist ~ performance =~ clonediam + numbstems + numbinfl + meanht + meanwidth
meanht -- meanwidth
```

Coefficients
```
| Latent variables:   | Estimate | Std.err | z-value | P(>|z|) | Std.lv | Std.all |
|---------------------|----------|---------|---------|---------|--------|---------|
| performance = geneticdist | -57.134  | 12.465  | -4.584  | 0.000   | -3.322 | -0.708  |

| Regressions:        | Estimate | Std.err | z-value | P(>|z|) | Std.lv | Std.all |
|---------------------|----------|---------|---------|---------|--------|---------|
| performance =~ clonediam | 1.000    | 17.199  | 0.969   | 0.332   | 15.555 | 0.962   |
| 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   |
```
Fit of Variables

- geneticdist
- performance
  - clonediam
  - numbstems
  - numbinfl
  - meanht
  - meanwidth

R-Square:

- clonediam: 0.938
- numbstems: 0.925
- numbinfl: 0.727
- meanht: 0.617
- meanwidth: 0.515
- performance: 0.502

A Latent Exercise

- geneticdist
- performance
  - clonediam
  - numbstems
  - numbinfl
  - meanht
  - meanwidth

- latitude

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

spartinaModel2<-paste(spartinaModel, 'meanht ~ latitude
meanwidth ~ latitude', sep="\n")
Latent Variables and Measurement Error

- True Cover Of Algae
- 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

$R^2 = 0.62$

Canopy biomass = kelpiness * 154.89 - 68.62
LANDSAT Kelpiness vs. Canopy Biomass
We can transform satellite data to canopy biomass, and fix the unstandardized loading to 1.

\[ \delta_x \rightarrow \text{Measured Canopy Biomass} \rightarrow \text{True Canopy Biomass} \]

But what about error?
We know that \( R^2 = 1 \)-estimated var/observed var

\[ \delta_v = (1 - R^2) \]

Unstandardized Measurement Error = \( \delta_v \times \text{var(Measured Canopy Biomass)} \)

---

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 \( y = 0.08x + 0.01 \) \( r^2 = 0.79 \)
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)
```

---

LANDSAT Kelpiness vs. Canopy Biomass
Fit this Model!

```
noerror <- 'summer_kelp_biomass ~ landsat_spring_biomass'

r2 <- var(lter$landsat_spring_biomass, na.rm = T) * (0.38)

[1] 3.762301

errorCanopy <- c(
  true_spring_biomass <- 1 * landsat_spring_biomass,
  summer_kelp_biomass <- true_spring_biomass
)

Error
landsat_spring_biomass <- 3.762301 * landsat_spring_biomass
```

---

LANDSAT Kelpiness vs. Canopy Biomass

```
landsat_spring_biomass <- 154.89 * spring_canopy + 68.62

summer_kelp_biomass <- 0.08 * kelp + 0.01

summer_kelp_biomass <- log(summer_kelp_biomass + 1)

landsat_spring_biomass <- log(landsat_spring_biomass + 1)
```

---
Incorporating measurement error in the predictor increases the coefficient size and amount of variance explained.