USGS
science for a changing world

# SEM in R:
# A Brief Introduction to the `Lavaan` R Package

U.S. Department of the Interior
U.S. Geological Survey

This module offers a very brief introduction to the lavaan R package for SEM.

A citation that can be used for the information included in this module is:

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

Notes: IP-056512; Support provided by the USGS Climate & Land Use R&D and Ecosystems Programs. I would like to acknowledge formal review of this material by Jesse Miller and Phil Hahn, University of Wisconsin. Many helpful informal comments have contributed to the final version of this presentation. The use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Last revised 18.07.11.

Source: https://www.usgs.gov/centers/wetland-and-aquatic-research-center/science/quantitative-analysis-using-structural-equation

## The R environment

For those not yet using R, a few basic resources are listed here for convenience. Links to additional resources can be found in the first two, while the third one is self-contained.

- **The Main Page for R**: (http://www.r-project.org/)

- **A Wiki for getting started**: (http://scs.math.yorku.ca/index.php/R:_Getting_started_with_R)

- **Quick-R resource**: http://www.statmethods.net/

Cite as: R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

USGS



The use of R is so wide-spread at this point and there is so much information for getting started I simply direct the reader to some of this material.

The Main Page for R: http://www.r-project.org/

A Wiki for getting started:
http://scs.math.yorku.ca/index.php/R:_Getting_started_with_R

Quick-R resource:http://www.statmethods.net/

The R environment permits several different ways to implement SEM.

Three primary implementations within the R environment:

(1) Global estimation using `lavaan` or `sem` packages,

(2) Local estimation using the `piecewiseSEM` package,

(3) Local estimation 'by hand' using classical regression methods augumented by graph-theoretic analyses,

(4) Local estimation using Markov chain Monte Carlo methods associated with Bayesian implemention.

(5) Bayesian estimation using `blavaan`, which uses lavaan syntax.

USGS

3

Note that there are several good software packages for SEM. The modules on Model Specifications and Estimation Methods provide discussions of alternative software packages. I am at present teaching using R and R-based implementations because they are free for users and R is widely used amongst natural scientists. The 'lavaan' and 'piecewiseSEM' packages are the two I will discuss. Each has its own strengths and limits, so they make a good pair of packages for most applications.

This tutorial briefly introduces the SEM R package known as **lavaan** ("latent variable analysis").

Url for the home page: http://lavaan.ugent.be/?q=node/2

My tutorials and exercises, some of which are specific to lavaan, are at: http://bit.ly/graceSEM

Yves Rosseel's latest (authoratative) tutorial is at: http://lavaan.ugent.be/tutorial/tutorial.pdf

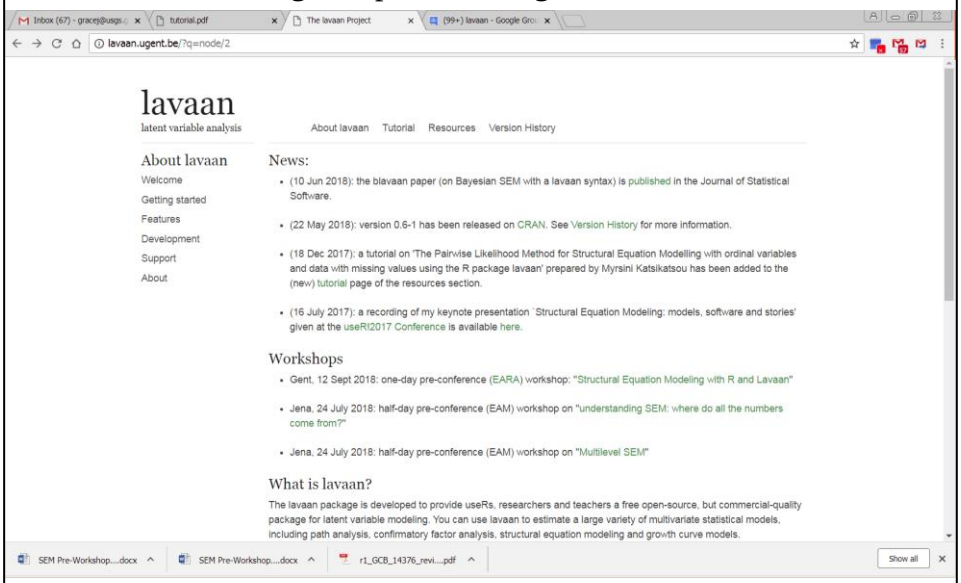
There is a Google Group for lavaan users to post questions at: https://groups.google.com/forum/#!forum/lavaan

USGS



I will initially focus on how to do SEM using lavaan for simplicity.

Lavaan Home Page: http://lavaan.ugent.be/




USGS

5

It is useful to visit the lavaan homepage for information and resources. The examples are very much from the sociometric and psychometric traditions.

Getting started in R: First read in data and load library.

R code: (command lines are in **bold**)

```
# Set working directory and load data

setwd("C:/your directory here")
```

```
# Read in data

dat <- read.csv("SEM.2.1_data.csv")
```

```
# Load lavaan library

library(lavaan)
```

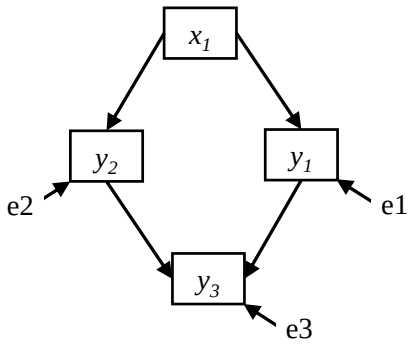
USGS



Only a very minimal use of R is required to work in lavaan.

The data file will be provided along with this tutorial.

Choose a model to code.




USGS

7

Here we have a model that represents the idea that x1 affects y3 indirectly in the model through y1 and y2. This could be called a "full mediation" model because effects of x1 on y3 are fully mediated or conveyed through y1 and y2.

For example, let

x1 = amount of human activity in the area around a wetland,

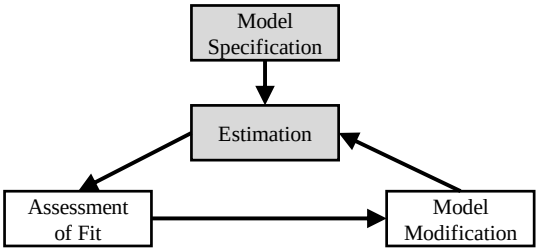
y1 = degree of changes to the hydrology of a wetland,

y2 = degree of nutrient inflow to a wetland

y3 = ecological integrity index for a wetland (based on community composition deviations from 'reference' condition).

In this example, the diagram represents the hypothesis that the effects of human activities around a wetland can be explained entirely by associated alterations in hydrology and nutrient inputs.

Here we illustrate just two steps in the overall modeling process: Model Specification and Estimation.




USGS

8

There are other tutorials that illustrate the full SEM workflow process. Here I illustrate just Model Specification and Estimation using lavaan.

In lavaan, there are three steps we will need to take.

Step 1: Specify Model.

Step 2: Estimate aka “fit” Model.

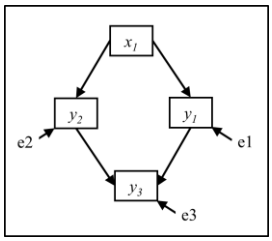
Step 3: Extract Results (both estimates and assessment of fit).

USGS



You will find the basic specifications in lavaan quite simple. Because it is syntax-based procedure, however, you will want to keep a visual representation of your models handy for keeping track of the causal logic of various versions of the model.

Three steps for working in lavaan - illustrated:




```
# Step 1: Specify model

mod.1 <- 'y1 ~ x1
          y2 ~ x1
          y3 ~ y1 + y2'
```

```
# Step 2: Estimate model using the 'sem' function

mod.1.fit <- sem(mod.1, data=dat)
```

```
# Step 3: Extract results

summary(mod.1.fit)
```

USGS



Specifying a model simply involves an equation for each response variable in the model. In this context, a 'response' variable is one with an incoming arrow. In this example, y1, y2, and y3 are response variables.

The "sem" function is used to "fit" or estimate the parameters in the model. In typical R fashion, this process creates a "fit object" from which summary and other information can be extracted.

Oops, there is a problem!

```
Warning message:
lavaan WARNING: some observed variances are (at least) a factor
1000 times larger than others; use varTable(fit) to investigate
```

So, we follow the output advice and request "varTable(fit object)"

```
> varTable(mod.1.fit)
  name idx nobs    type exo user   mean     var nlev lnam
1   y1   2   90 numeric   0    0 49.239  58.969    0
2   y2   3   90 numeric   0    0  0.691   0.101    0
3   y3   4   90 numeric   0    0 49.233 228.181    0
4   x1   1   90 numeric   1    0 49.235  77.960    0
>
```

USGS



Lavaan is fussy about data scales, since they impact the internal matrix manipulations. This seems to vary from version to version with lavaan, though lavaan will let you know if it has a problem with your data.

Recode the data and try again.

```
### Recode vars to roughly same scale
x1 <- dat$x1/100
y1 <- dat$y1/100
y2 <- dat$y2
y3 <- dat$y3/100

### Create Transformed Dataset
# overwrite file with recoded data
t.dat <- data.frame(x1, y1, y2, y3)

# Repeat Step 2: Estimate model
mod.1.fit <- sem(mod.1, data=t.dat)
```

Now, no error message this time, so now we can ask for results summary.

```
# Step 3: Extract results
summary(mod.1.fit)
```

USGS

12

The information we obtained about variances in the previous slide helps us to appropriately recode the variables. We may need this information later to decode things (though often that is not important unless one wants to talk about raw units).

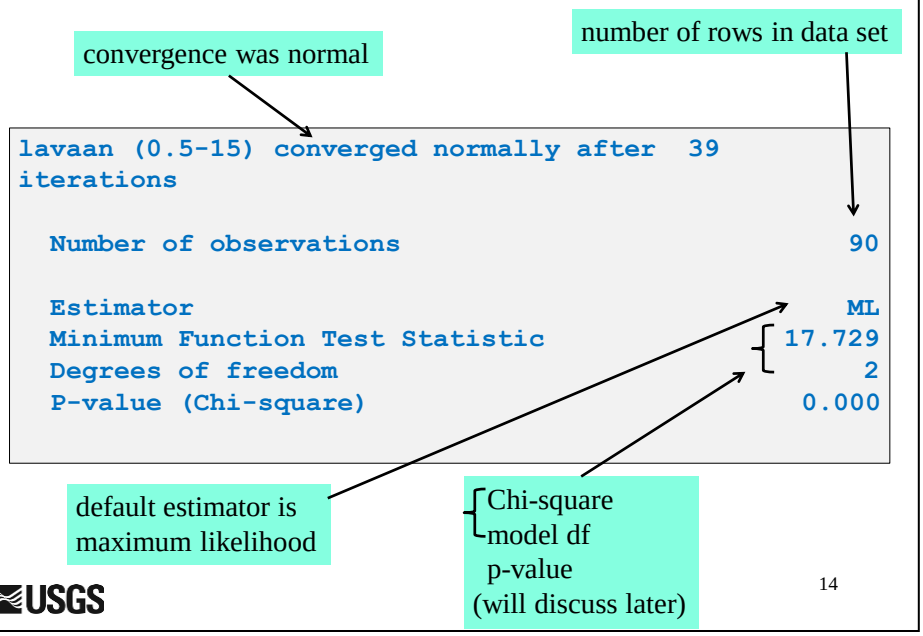
## Results Summary.

```
lavaan (0.5-15) converged normally after  39 iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic               17.729
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.000

Parameter estimates:
                   Estimate  Std.err  Z-value  P(>|z|)
Regressions:
  y1 ~
    x1                0.400    0.081    4.911    0.000
  y2 ~
    x1                0.875    0.367    2.381    0.017
  y3 ~
    y1                0.093    0.017    5.475    0.000
    y2                0.013    0.004    3.121    0.002
Variances:
    y1                0.460    0.069
    y2                9.362    1.396
    y3                0.015    0.002
```

USGS

Here are some of the results generated by the "summary" command. In the next two slides we zoom in on these results. First we will look at the model fit information at the top, then the estimates table at the lower part of the page.

## Results Summary: Closer Look.




USGS

14

Convergence is necessary, so good to see it was successful.

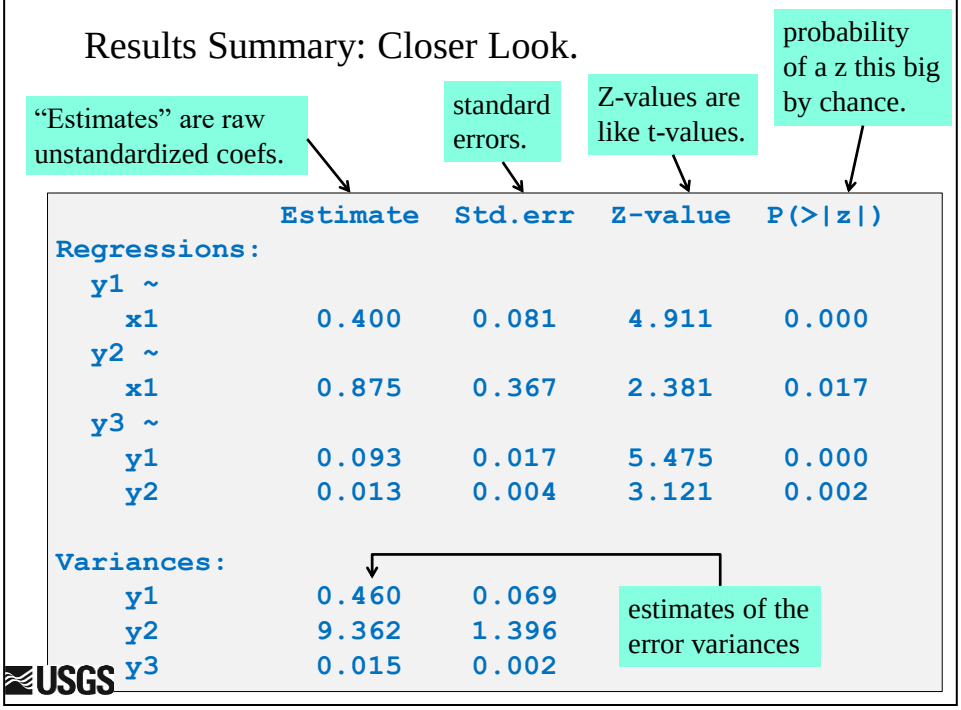
The "Minimum Function Test Statistic" is a long way of saying what is usually called the "Model Chi-square".

The "Degrees of freedom" represents the number of paths omitted from the model. These provide us with a capacity to test the architecture of the model.

The P-value refers to the probability of the data given our model. In this case the probability is very low, suggesting our model is inconsistent with the data and changes will need to be made.

Interpretation of this information is discussed in a later section.

## Results Summary: Closer Look.




```
                   Estimate  Std.err  Z-value  P(>|z|)
Regressions:
  y1 ~
    x1              0.400    0.081    4.911    0.000
  y2 ~
    x1              0.875    0.367    2.381    0.017
  y3 ~
    y1              0.093    0.017    5.475    0.000
    y2              0.013    0.004    3.121    0.002

Variances:
    y1              0.460    0.069
    y2              9.362    1.396
    y3              0.015    0.002
```

"Estimates" refer to parameter estimates. These are the coefficients for the equations. We can assign names to the parameters. Lavaan uses the string 'y1 ~ x1" as the label for the parameter whose value is 0.400.

Since the estimates are arrived at through maximum likelihood methods we get a "Z-value" instead of a "t-value".

Note that the estimates of "Variances" are actually error variances. Recall that error variances were discussed in the module "SEM Essentials – Model Anatomy".

# More Lavaan Syntax

USGS

16

Here we visit the issue of syntax available in lavaan.

1. lavaan has a number of operators and syntax options.

| formula type | operator | operator stands for |
|---|---|---|
| regression | ~ | "regressed on" |
| correlation | ~~ | "correlated with" |
| intercept | ~ 1 | "estimates intercept" |
| latent variable definition | =~ | "is measured by" |
| create a composite | <~ | "is caused by" |

As we shall see, this is not a complete list of operators.

USGS

17

Yves Rosseel's latest (authoratative) tutorial is at:

http://lavaan.ugent.be/tutorial/tutorial.pdf

(which is hyperlinked here: http://lavaan.ugent.be/tutorial/tutorial.pdf)

2. You can work with individual parameters by naming them.

Lavaan names parameters as shown in output, e.g., **"y1 ~ x1"**.

We can assign names by pre-multiplying a predictor with the name being assigned. This will allow us to manipulate the parameters.

```
model.2a <- 'y1 ~ b1*x1
```

Note, parameter labels must start with a letter!

USGS



Naming parameters is a key step in many operations.

## 2. (cont.) Working with Named Parameters – Indirect Effects.

We can compute the indirect effect along a pathway using the named parameters.

```
##### Demonstrate the naming of parameters

# Step 1: Specify model adding labels
mod.2 <- 'y1 ~ b1*x1
          y2 ~ b2*x1
          y3 ~ b3*y1 + b4*y2
          # compute indirect effect
          IE1 := b1*b3'
```

Note we use '**:=**' to define a computed quantity

USGS



Naming parameters is a key step in many operations.

## Results Summary.

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  y1 ~
    x1        (b1)    0.400    0.081    4.911    0.000
  y2 ~
    x1        (b2)    0.875    0.367    2.381    0.017
  y3 ~
    y1        (b3)    0.935    0.171    5.475    0.000
    y2        (b4)    0.129    0.041    3.121    0.002

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .y1                0.005    0.001    6.708    0.000
   .y2                0.094    0.014    6.708    0.000
   .y3                0.015    0.002    6.708    0.000

Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)
    IE1                0.374    0.102    3.656    0.000
```

USGS

20

We now get to see the labels we added on the output, along with the Defined Parameters.

## 3. Assigning values to parameters by naming them.

### c. Fixing Parameter Values to Specific Quantities

There are times when we want to be able to specify that particular parameters have fixed quantitative values. Lavaan allows us to do this using various options. Here is one approach:

```
model.2b <- 'y1 ~ 0*x1 + x2
             y2 ~ x2
             y1 ~~ y2'
```

In this model statement, x1 is pre-multiplied by zero to set its value to zero. We can also accomplish this using a more elaborate and more flexible approach:

```
model.2c <- 'y1 ~ b1*x1 + x2
             y2 ~ x2
             y1 ~~ y2
             b1 == 0'
```

Now we have labeled the parameter "b1" and then assigned it a value of 0 in a separate statement. This second specification will actually result in an explicit test of the constraint.

USGS

21

Assigning values is also important.

4. Correlations/Covariances between exogenous variables are not usually estimated, but we can.

```
#estimating the model

model.2d.ests <- sem(model.2, data = data.mod1, fixed.x=FALSE)
```

Now we obtain an estimate of the covariance in our lavaan output, as shown in bold below.

```
                   Estimate  Std.err  Z-value  P(>|z|)
Regressions:
  y1 ~
    x1               -0.003    0.004   -0.763    0.446
    x2               -0.087    0.019   -4.643    0.000
  y2 ~
    x2               -3.363    0.896   -3.752    0.000

Covariances:
  y1 ~~
    y2                0.945    0.432    2.189    0.029
  x1 ~~
    x2               -2.651    1.352   -1.961    0.050
```

USGS

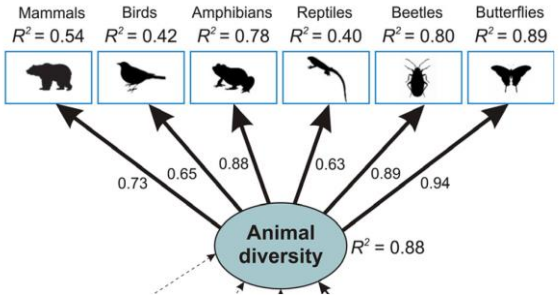
22

A default of lavaan, like all software except Amos, is to just take the exogenous correlations/covariances directly from the data and not treat them as estimated parameters. The module on my website

"SEM Essentials – Path Rules"

explains how that is possible. Anyway, sometime we want or need to treat those as estimated parameters, so the command above shows how.

5. Lavaan creates latent variables by declaring them in the absence of any known values.




```
mod1 <- ' # Latent variable definition
        Diversity =~ Mammals +Birds +Amphibians
                     +Reptiles +Beetles +Butterflies
        '
```

(Note figure is just showing part of a larger model.)

USGS

23

This model type is covered in depth in the module on latent variable modeling. This example is from:

Jimenez et al. 2016. Ecology and Evolution. 6(5): 1515–1526.

# For Further Information

Beaujean, A.A., 2014. Latent variable modeling using R: A step-by-step guide. Routledge. (primarily `lavaan`-based)

Book Website: http://blogs.baylor.edu/rlatentvariable

USGS