USGS
science for a changing world

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

Jim Grace

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

This module offers a very brief introduction to the lavaan R package for SEM. The assumption here is that the user

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. Questions about this material can be sent to sem@usgs.gov.

(last revised 15.03.28)

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 classical regression methods augumented by graph-theoretic analyses,

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

USGS

3

Note that we recognize 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.

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

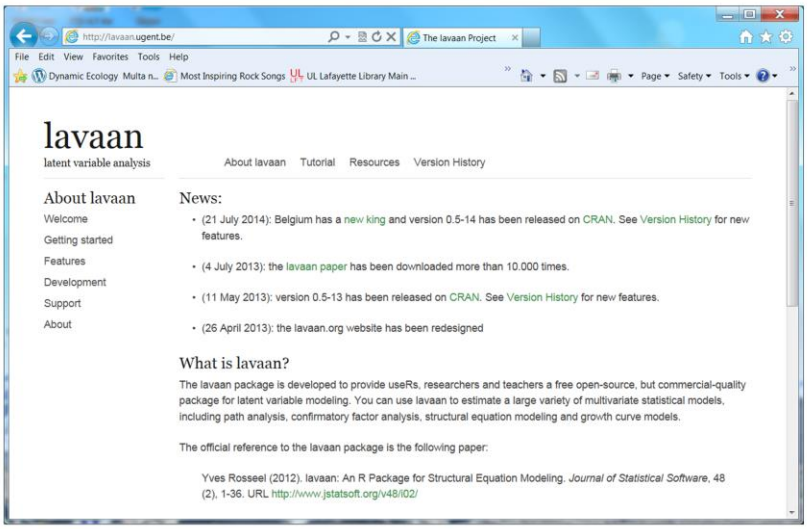
My very introductory tutorials etc. are at:
http://www.nwrc.usgs.gov/test/sem.html

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

Jarrett Byrnes has some good material at:
http://jarrettbyrnes.info/ubc_sem/

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.

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

R code: (we will **bold** command lines)

```
# Set working directory and load data

setwd("C:/Documents/LavaanTutorial")
```

```
# Read in data

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

```
# Load lavaan library

library(lavaan)
```

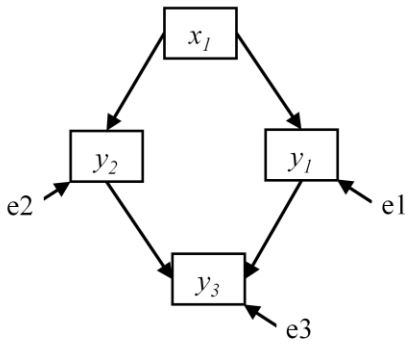
USGS

6

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

The data file will be provided along with this tutorial.

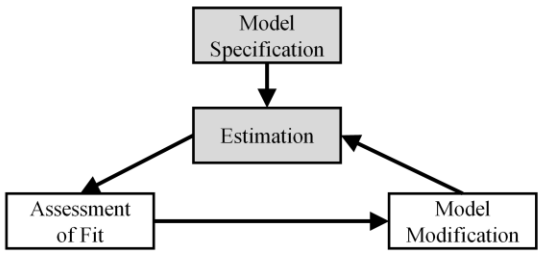
Choose a model to code.



USGS

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.

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



USGS

8

The module "SEM Essentials - Summary Points" presents a multi-step outline of the modeling process. Here we illustrate just Model Specification and Estimation using lavaan.

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

Step 1: Specify Model.

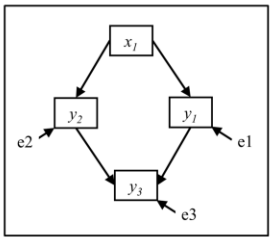
Step 2: Estimate aka “fit” Model.

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(mod1.fit)
```

USGS

Specifying a model simply involves an equation for each response variable in the model.

The "sem" function is used to "fit" the model. 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 variables to roughly same scale
x1 <- x1/10
y1 <- y1/10
y2 <- y2*10
y3 <- y3/100

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 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).

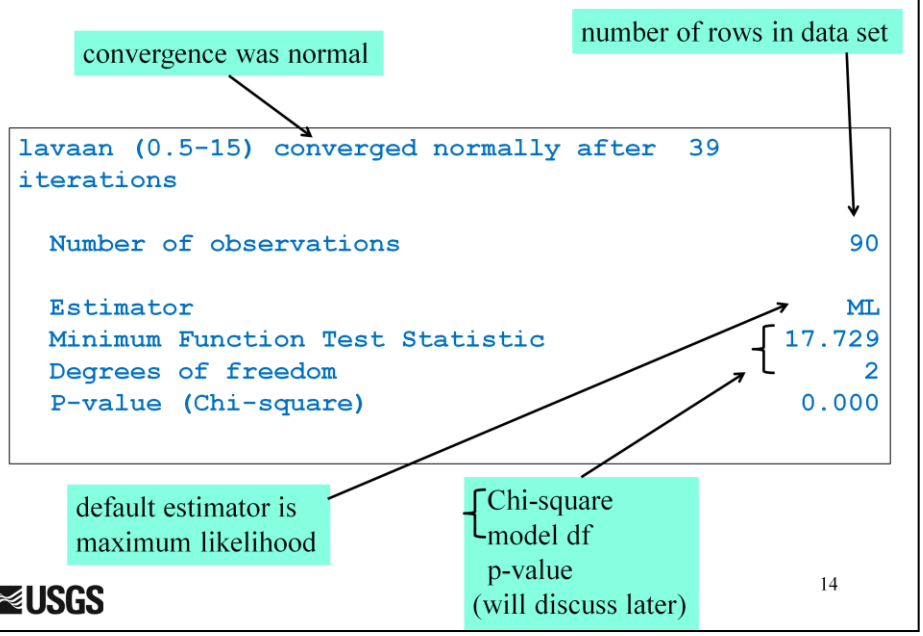
Results Summary.

```
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 13

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

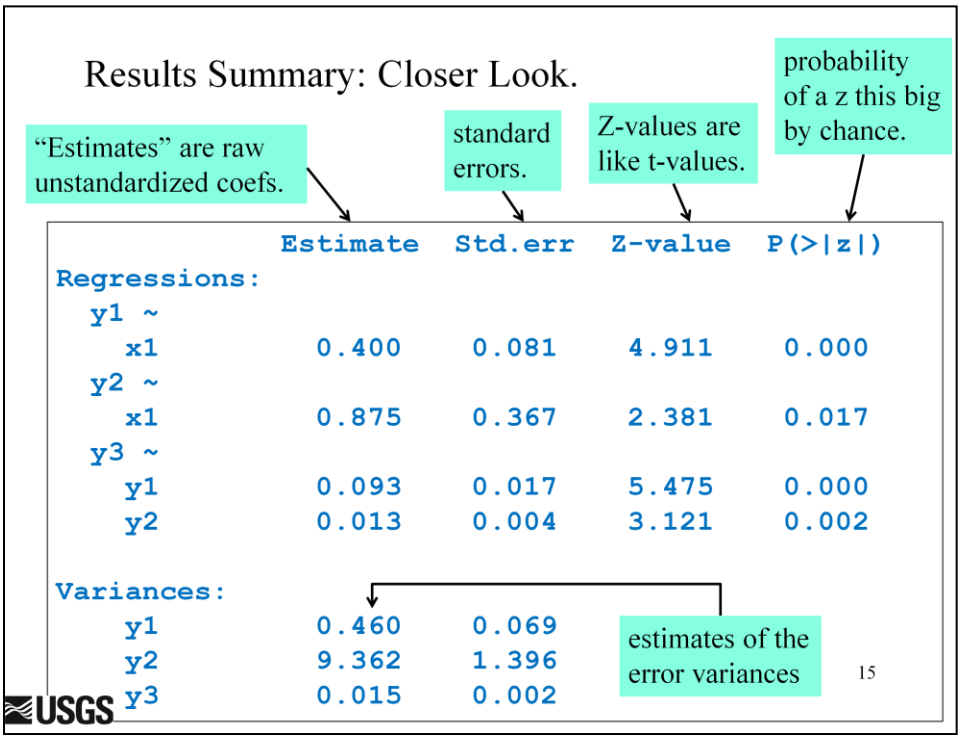
Convergence is necessary, so good to see it was successful.

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.



"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".