USGS
science for a changing world

SEM in R:
Local Estimation

Jim Grace

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

Here I provide just a very brief introduction to the concept of local estimation.

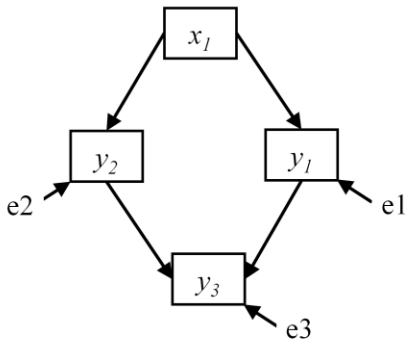
A citation that can be used for the information included in this module is:

Grace, J.B., Schoolmaster, D.R. Jr., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M., and Schweiger, E.W. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere* 3(8): article 73 (44 pages).

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 15.04.05.

How would be evaluate this model using local estimation methods?

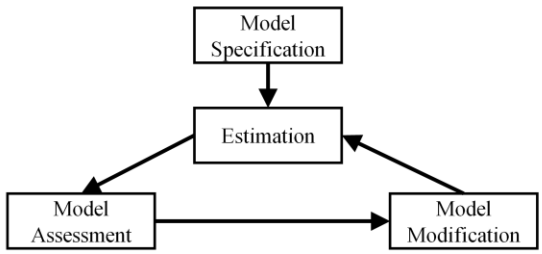


USGS

Here we consider the same model we did in our brief "Intro to Lavaan". This model represents the hypothesis that the effect of x1 on y3 occurs because of two processes, one propagated through y1 and the other through y2.

Again, we focus on the mechanics of

- specification
- estimation
- model assessment.

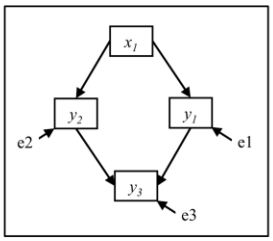


USGS

As before, we will want to

1. specify our model,
2. estimate the parameter values
3. assess how well our data correspond to our model.

Identify conditional independences as a way of thinking about specification alternatives.



What are the conditional independence claims?

(1) $y_2 \perp y_1 \mid x_1$

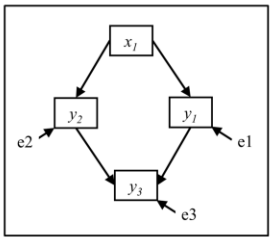
(2) $y_3 \perp x_1 \mid y_1, y_2$

USGS

It is handy to consider the possible alternative models up front when one is doing local estimation, since we will have to check our models by hand rather than having software do that for us. As discussed in the module on Model Evaluation, the first order of business is to determine if there are any important omitted links. The reason this is the first order of business is because when links that are important in the data generating process are omitted from the model, the estimates for other links can be way off. In contrast, including unimportant links in models has a comparatively smaller effect on the estimates for other links.

Regarding how we evaluate missing links, I introduce the concept of "conditional independence" in the module "SEM Essentials – Basics of Estimation". It is also covered in greater depth in "SEM Essentials – Path Rules". In this example, there are two implied independences in our model. There is no link directly from x1 to y3 and none connecting y1 and y2. We need to know if those pairs are indeed conditionally independent.

Specification and estimation of equations/submodels in R:




```
# Specification and estimation of submodels

y1.mod <- lm(y1 ~ x1,      data=fourvars.dat)
y2.mod <- lm(y2 ~ x1,      data=fourvars.dat)
y3.mod <- lm(y3 ~ y1 + y2, data=fourvars.dat)
```

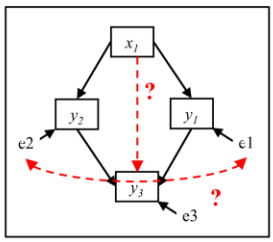
USGS

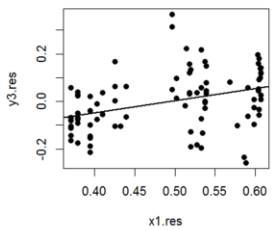
Here I show a simple “by-hand” approach. Recall that a network model can be represented by an equation for each endogenous variable. We can use the base function “lm” to model each of our endogenous variables as a function of its parent predictors, creating three model objects that collectively summarize the network hypothesis. This is an example of local or piecewise approach to estimation.

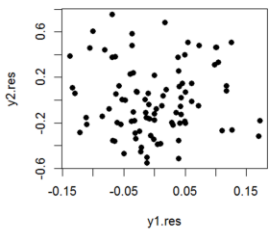
Model assessment.

```
# Capture residuals

x1.res <- x1
y1.res <- resid(y1.mod)
y2.res <- resid(y2.mod)
y3.res <- resid(y3.mod)
```







USGS

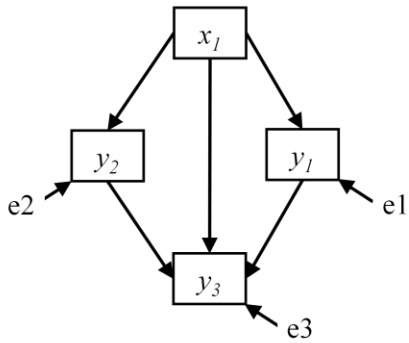
6

Missing links are detected through the existence of residual relationships. We can look at this directly by obtaining residuals for our endogenous variables with the function "resid". Note that for x1, the residuals are the raw values, as this exogenous variable has no predicted scores.

Examining residual relationships simply involves looking to see if the residuals for two unconnected variables are significantly related.

When we test for significance in this way, we find evidence that perhaps x1 should be in the equation for y3. This evidence should not be considered the final "test" for inclusion, but only a diagnostic. Also, this scatterplot approach permits us to find nonlinear residual relationships, which would require special modeling techniques to include in equations.

Model respecification.




```
# Specification and estimation of submodels

y1.mod <- lm(y1 ~ x1,           data=fourvars.dat)
y2.mod <- lm(y2 ~ x1,           data=fourvars.dat)
y3.mod2 <- lm(y3 ~ y1 + y2 + x1,data=fourvars.dat)
```

USGS

The way we confirm missing links is by adding them to the local equations and evaluating. In this case, we add a path from x1 to y3 in our model by including the predictor x1 in the equation for y3 (shown in red).

USGS

After a modification is made to the model, residual checking would need to be done again.

```
# Capture residuals

x1.res <- x1
y1.res <- resid(y1.mod)
y2.res <- resid(y2.mod)
y3.res2 <- resid(y3.mod2)
```

```
# Plot residual relationships

plot(y1.res, y2.res)
```

(Results from these further evaluations not shown here.)

USGS

8

Once we add a link to our model (i.e., add a predictor to a submodel), we need to retest. This set of procedures is continued until no further additions are indicated.

Ultimately, model assessment involves considering both whether included paths are justified as well as whether any additional paths are omitted.

```
# Examine parameters support for revised model
summary(y1.mod)
summary(y2.mod)
summary(y3.mod2)
```

USGS

Once we believe all necessary links are included in our model, we consider model complexity and whether some pruning is in order. In local estimation, we get our first hint about expendable paths by looking at the model parameter contributions. Here I use the summary command to obtain those.

Results for first submodel (y1)

```
> summary(y1.mod)

Call:
lm(formula = y1 ~ x1, data = t.dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29554    0.04118   7.177 2.14e-10 ***
x1           0.39983    0.08233   4.856 5.16e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.06858 on 88 degrees of freedom
Multiple R-squared:  0.2113,    Adjusted R-squared:  0.2024
F-statistic: 23.58 on 1 and 88 DF,  p-value: 5.158e-06
```

USGS

Here we see the results for submodel y1. Its only predictor x1 has a very low p-value, suggesting any test would lead to the conclusion this link should be included in the model.

Results for second submodel (y2)

```
Call:
lm(formula = y2 ~ x1, data = t.dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2606     0.1858   1.403   0.1643
x1            0.8747     0.3715   2.355   0.0208 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.3094 on 88 degrees of freedom
Multiple R-squared:  0.05927,   Adjusted R-squared:  0.04858
F-statistic: 5.544 on 1 and 88 DF,  p-value: 0.02077
```

USGS

Our second submodel is less strong regarding evidence supporting that link, though still significant by conventional standards. In a slide coming up, more formal testing of alternative hypotheses is presented.

Results for third submodel (y3)

```
Call:
lm(formula = y3 ~ y1 + y2 + x1, data = t.dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.19931    0.08666  -2.300  0.02388 *
y1           0.59252    0.17711   3.345  0.00122 **
y2           0.09294    0.03925   2.368  0.02015 *
x1           0.68174    0.15761   4.325 4.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.1139 on 86 degrees of freedom
Multiple R-squared:  0.4503,    Adjusted R-squared:  0.4311
F-statistic: 23.48 on 3 and 86 DF,  p-value: 3.404e-11
```

USGS

12

Results for the revised submodel for y3 gives support for all three predictors. Again, more formal analysis will be presented next.

All submodels should be thoroughly checked and justified (see module on Model Evaluation). Here I illustrate for y3 submodel.

```
### Illustrate evaluation of y3 submodel:
### Compare revised model against simpler models
# Create suite of alternative models
y3.mod2 <- lm(y3 ~ y1 + y2 + x1, data=t.dat)
y3.mod1 <- lm(y3 ~ y1 + y2,      data=t.dat)
y3.mod3 <- lm(y3 ~ y1 + x1,      data=t.dat)
y3.mod4 <- lm(y3 ~ y2 + x1,      data=t.dat)

# Compare using AICc
library(AICcmodavg)
aictab(list(y3.mod2,y3.mod1,y3.mod3,y3.mod4),
modnames=c("y3.mod2","y3.mod1","y3.mod3","y3.mod4"))
```

13

One approach to model evaluation is to compare the weight of evidence for alternative models using information-theoretic measures, like AICc. I illustrate this and other methods for globally-estimated SEMs in the module “Model Evaluation”. Here I perform the same procedure, creation of an AIC table, for submodel 3.

AIC table results.

```
Model selection based on AICc :

         K    AICc Delta_AICc AICcWt Cum.Wt    LL
y3.mod2  5 -128.95       0.00   0.84   0.84 69.83
y3.mod3  4 -125.51       3.44   0.15   0.99 66.99
y3.mod4  4 -120.18       8.77   0.01   1.00 64.33
y3.mod1  4 -113.48      15.47   0.00   1.00 60.97
```

Our modified model is best of this set, with AICcWt of 0.84.
In combination with other diagnostics, we chose this model as best for this local node.

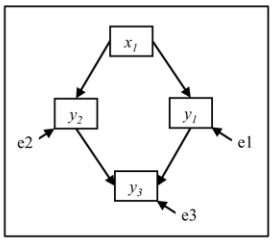
14

The package "AICcmodavg" has many capabilities. Its default is to create a table as shown here, reporting the corrected AIC, "AICc".

If we look at the Delta_AICc column, we see that our revised model is more than 2 units better than the second best model, suggesting that it is distinct from the others. Weight of evidence for this model is 0.84, providing a more specification of its support from the data.

An alternative way to test for missing links – the "d-sep" test.

Verma and Pearl (1988) defined the general criterion whereby one may identify the conditional independence claims for Graphs that are Directed and Acyclic (dags). This is referred to as the "d-separation criterion".



Building upon this theorem, Shipley (2000) developed a quantitative test, given a graph and data, called the "d-sep test".

USGS

15

An alternative approach to evaluating network models is referred to as "d-separation". This concept was developed by Judea Pearl and his colleagues as a theoretical basis for evaluating networks. The somewhat abstract nature of the theoretical principle comes from the fact that is is designed for the development of algorithms that might be used in artificially intelligent systems. For those of us working on SEMs "by hand", it is an intuitive extension of the principle of conditional independence.

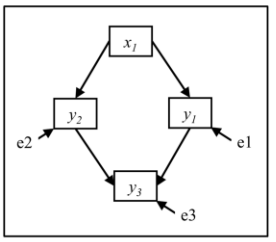
In 2000, Bill Shipley developed empirical tests based on the d-separation principle, which he refers to as the "d-sep test". This has come to be known also as the Shipley test (for example, in the ggm package).

The "d-sep" test continued.

A set of *k* conditional independence claims, e.g.,

$$y_2 \perp y_1 \mid x_1$$

$$y_3 \perp x_1 \mid y_1, y_2$$



can be turned into a set of individual tests, whose p-values can be accumulated into a C-score.

$$C = -2\sum_{1}^{k} \ln(p)$$

This C score follows a chi-square distribution, which permits an overall test of model-data fit.

USGS

The d-sep test statistic, which is Fisher's C statistic, accumulates the p-values for a model into a single index. This overall index then represents average fit (more or less) for the model.

Executing the "d-sep" test.

```
### Use ggm to run Shipley (d-sep) test
library(ggm)

cov.dat <- as.matrix(cov(t.dat))

dag <- DAG(y1 ~ x1, y2 ~ x1, y3 ~ y1 + y2)

shipley.test(dag, S=cov.dat, n=90)
```

Using the ggm package, we can run the d-sep test.

This requires turning our data into a covariance matrix.

```
$ctest = 20.40476
$df = 4                    # true df = 2
$pvalue = 0.000415407   # indicates discrepancy
```

USGS

We can implement the d-separation test using the library ggm. For this package, we need to independently feed in the form of the graph (or DAG). In this case, we are examining the original model again.

Here we get a C score of 20.4, which is highly significant. As with SEM chi-square statistics, a significant p-value here means important descrepancies between the data and the model.

Note that the df returned by ggm apparently evaluates multiple hypotheses about the directionality of omitted links, returning an value of df=4, when from a pure conditional-independence standpoint is df=2.

All this (and more) is now automated in the "piecewiseSEM" package developed by Jon Lefcheck.

```
### Install Lefcheck's piecewiseSEM package
library(devtools)
install_github("jslefche/piecewiseSEM")
library(piecewiseSEM)

# Original Model
modlist1 = list(
  lm(y1 ~ x1, data=t.dat),
  lm(y2 ~ x1, data=t.dat),
  lm(y3 ~ y1 + y2, data=t.dat))

# derive the fit statistics:
get.sem.fit(modlist1, t.dat)
```

USGS

18

A very new package developed by Jon Lefcheck is called "piecewiseSEM". This package handles the bundling of individual local models and d-sep testing for us.

The code shown in the slide illustrates how to download the package from its current location at github (as of April 2015) .

The SEM is presented as a list of R models. Here I only show the simplest sort, "lm" models. The real utility of this package, however, is that it can bundle many different sorts of R models, such as glms, lmers, and more.

Results.

```
> get.sem.fit(modlist1, t.dat)

$missing.paths
  missing.path estimate std.error DF crit.value p.value
1     y3 <- x1    0.682     0.158 NA       4.325   0.000
2     y1 <- y2    0.003     0.024 NA       0.116   0.908

$Fisher.C
Fisher.C        k        P
  20.405    2.000    0.000

$AIC
   AIC   AICc      K      n
40.405 43.190 10.000 90.000
```

The output includes p-values for the individual tests of missing links, as well as the overall *C*-score and its *p*-value.

In addition, Shipley has now generalized *C* to a model *AIC*.

USGS

19

The get.sem.fit function returns a listing of each omitted linkage and an evaluation of it. We again get the Fisher's C score for the total model, though our real focus (I would argue) should be on the individual path results. While a non-significant C is sought, I would argue the real criterion is "O omitted links", meaning no important links are left out of our model.

Shipley has generalized C to an model AIC to stay up with the times.

Shipley, B. 2013. The AIC model selection method applied to path analytic models compared using a d-separation test. Ecology 94:560-564.

Piecewise (local) estimation opens up many possibilities for modeling more complex specifications.

20

For additional illustrations of local estimation in SEMs, see

Grace, J.B., Schoolmaster, D.R. Jr., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M., and Schweiger, E.W. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere* 3(8): article 73

Available at http://www.esajournals.org/doi/pdf/10.1890/ES12-00048.1

Also,

Shipley, B. 2009. Confirmatory path analysis in a generalized multilevel context. Ecology 90: 363-368.