SEM Essentials: Path Rules

Jim Grace

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

This module seeks to illustrate certain basic ideas related to SEM through a description of "path rules". These rules were developed by Sewall Wright many years ago, but still represent some fundamental ideas important for SEM practitioners to understand.

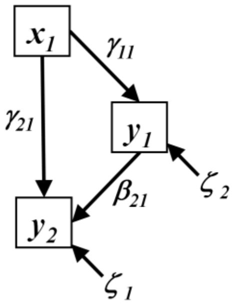
A citation that can be used for the information included in this module is:

Grace, J.B. (2006) Structural Equation Modeling and Natural Systems. Cambridge University Press

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.

We can illustrate some key properties of SE models in terms of model-implied versus observed correlations/covariances.



Standardized covariances (correlations)

	x_1	y_1	y_2
x_1	**1.0**		
x_2	**0.40**	**1.0**	
y_1	**0.50**	**0.60**	**1.0**

2

For this presentation, it is helpful to adopt a linear Gaussian approximation. One thing this allows us to do is to summarize our data in the form of a covariance matrix, which when standardized, is a correlation matrix.

Judea Pearl uses the same approach to connect historical SEM to his nonparametric, generalized model in

Pearl, J. 2013. Linear models: a useful "microscope" for causal analysis. *Journal of Causal Inference*. 1:155-170.

Note that the principles that apply to path coefficients extend to more general and more complex (e.g., non-linear, non-Gaussian) functional relationships.

While I will present the path rules using correlations, actual analyses use the raw covariances.

Raw Covariance Matrix

	x_1	x_2	y_1
x_1	**3.2**		
x_2	**0.65**	**0.8**	
y_1	**1.98**	**1.19**	**4.8**

variance covariance

Standardized Covariance Matrix

	x_1	x_2	y_1
x_1	**1.0**		
x_2	**0.40**	**1.0**	
y_1	**0.50**	**0.60**	**1.0**

correlation

The computational relationship between correlations (r_{xy}) and covariances (COV_{xy}) is quite simple.

$$r_{xy} = \frac{COV_{xy}}{SD_x \times SD_y}$$

r_{xy} = correlation
COV = covariances (ave cross product)
SD = std. dev. = $VAR^{\wedge 1/2}$

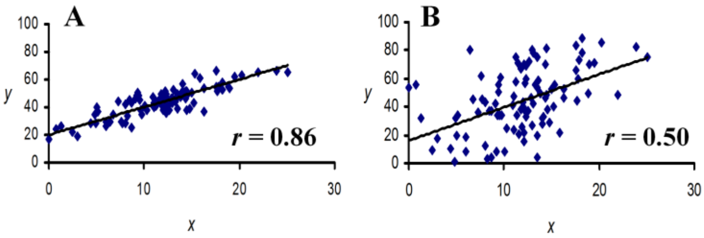
3

It is important to make the distinction between covariances and correlation, but also to show the simple relationship between the two.

Why work with covariances instead of correlations?

Analysis of correlations represents loss of information

Illustration of regressions having same slope and intercept, but different degrees of scatter around the line.



Analysis of covariances allows for estimation of both standardized and unstandardized parameters. This extends to the issue of unstandardized versus standardized coefficients.

USGS

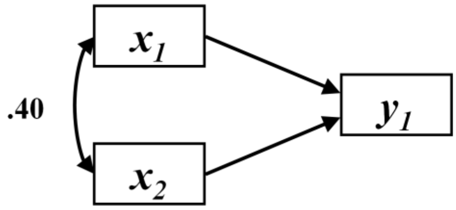
4

We need to understand that the raw values matter when it comes to coefficients. Some statisticians don’t like standardized parameter estimates because they involve combining multiple quantities. This issue is discussed and resolved in

Grace, J.B. and Bollen, K.A. (2005) Interpreting the results from multiple regression and structural equation models. *Bulletin of the Ecological Society of America* 86:283–295. http://dx.doi.org/10.1890/0012-9623(2005)86[283:ITRFMR]2.0.CO;2

1. First rule of path coefficients* is . . .

The path coefficients for unanalyzed relationships (curved arrows) between exogenous variables are simply the correlations (standardized form) or covariances (unstandardized form).



	x_1	x_2	y_1
x_1	1.0		
x_2	0.40	1.0	
y_1	0.50	0.60	1.0

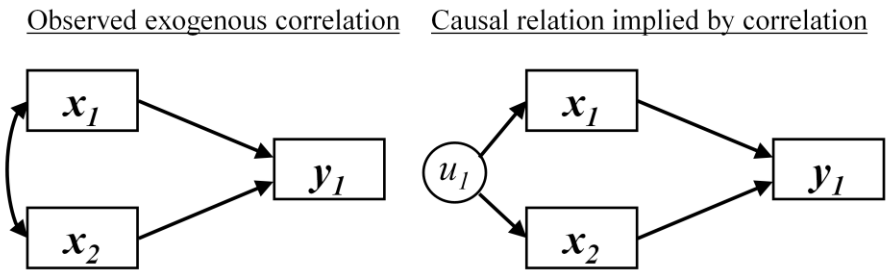
USGS

5

It is useful to know that exogenous correlations/covariance are not computed by SEM packages. Rather, we can pull them directly from a covariance matrix. This practice is predicated on the first rule of path coefficients.

A digression about correlations in SE models.

We should think of correlations in models as latent joint effects.



USGS

6

Before moving on, we should realize that correlations in models actually imply some causal process involving hidden variables. Essentially these are latent processes of some type, but we generally don't go so far as to represent them that way, UNLESS, we are doing latent variable modeling (see separate modules for that).

2. Second rule of path coefficients is . . .

When variables are connected by a single causal path, the path coefficient is the simple correlation/covariance.

$$x_1 \xrightarrow{\gamma_{11} = .50} y_1 \xrightarrow{\beta_{21} = .60} y_2$$

slightly different data matrix from the one on the previous slide

	x_1	y_1	y_2
x_1	**1.0**		
y_1	**0.50**	**1.0**	
y_2	**0.30**	**0.60**	**1.0**

γ (gamma) used to represent effect of exogenous on endogenous.

β (beta) used to represent effect of endogenous on endogenous.

USGS

7

When variables are connected by a single causal pathway, coefficients are “simple”. This means they will correspond with the observed correlations.

3. Third rule of path coefficients is . . .

Strength of a compound path is the product of the coefficients along the path.

	x_1	y_1	y_2
x_1	**1.0**		
y_1	**0.50**	**1.0**	
y_2	**0.30**	**0.60**	**1.0**

$$x_1 \xrightarrow{.50} y_1 \xrightarrow{.60} y_2$$

Thus, in this example the predicted effect of x_1 on y_2 = **0.5 * 0.6 = 0.30**

Since the indirect path from x_1 to y_2 equals the correlation between x_1 and y_2, we say x_1 and y_2 are **conditionally independent given** $\boldsymbol{y_1}$.

USGS

8

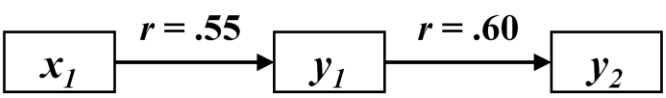
We multiply coefficients along compound pathways to get the expected net relationship. Since the second rule tells us that for this model the path coefficients should be the observed correlations, we can then compute the indirect effect of x_1 on y_2, which also represents the implied correlation between these two variables.

This then allows us to introduce and illustrate a critically-important concept “conditional independence”. If the model holds here, then once we know the values for y_1, we don’t need the values of x_1 to predict y_2. This is why we say that x_1 and y_2 are independent, once we condition on y_1.

What does it mean when two separated variables are not conditionally independent?

(different data matrix)

	x_1	y_1	y_2
x_1	**1.0**		
y_1	**0.55**	**1.0**	
y_2	**0.50**	**0.60**	**1.0**

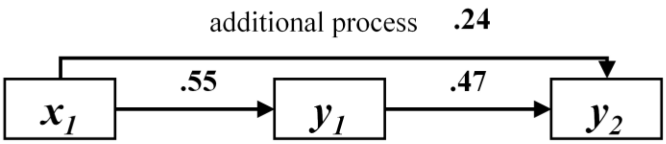


0.55 * 0.60 = 0.33, which is not equal to 0.50

We can reject the conditional independence model (i.e., a purely indirect effect) by seeing that things don't add up.

What does it mean when two separated variables are not conditionally independent? (continued)

The inequality implies that the true model is



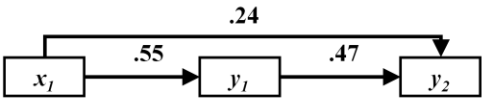
USGS

10

Failure to find conditional independence implies some other process at work. Maybe it is something you didn't even suspect was important! Note, that now that we have changed the model, some of the parameter estimates are different from those in the previous model. When you complete this module, you should be able to look at a model and know which paths will be simple covariances and which ones will be complex parameters representing "partial" effects.

4. Fourth rule of path coefficients is . . .

When variables are connected by more than one causal pathway, the path coefficients represent "partial" effects.



	x_1	y_1	y_2
x_1	**1.0**		
y_1	**0.55**	**1.0**	
y_2	**0.50**	**0.60**	**1.0**

Q: Which pairs of variables are connected by two causal paths?

answer: x_1 and y_2 (obvious one), but also y_1 and y_2, which are connected by the joint influence of x_1 on both of them.

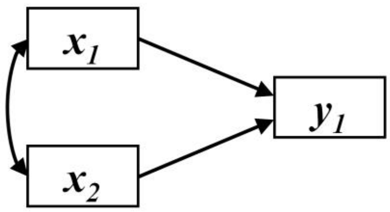
USGS

11

Continuing, when two variables are connected by more than one causal pathway, parameters/coefficients are not simple anymore.

Note that we cannot pick the values out of the covariance matrix when there are partial effects in the model.

The fourth rule of path coefficients also applies when variables are indirectly connected through correlations.

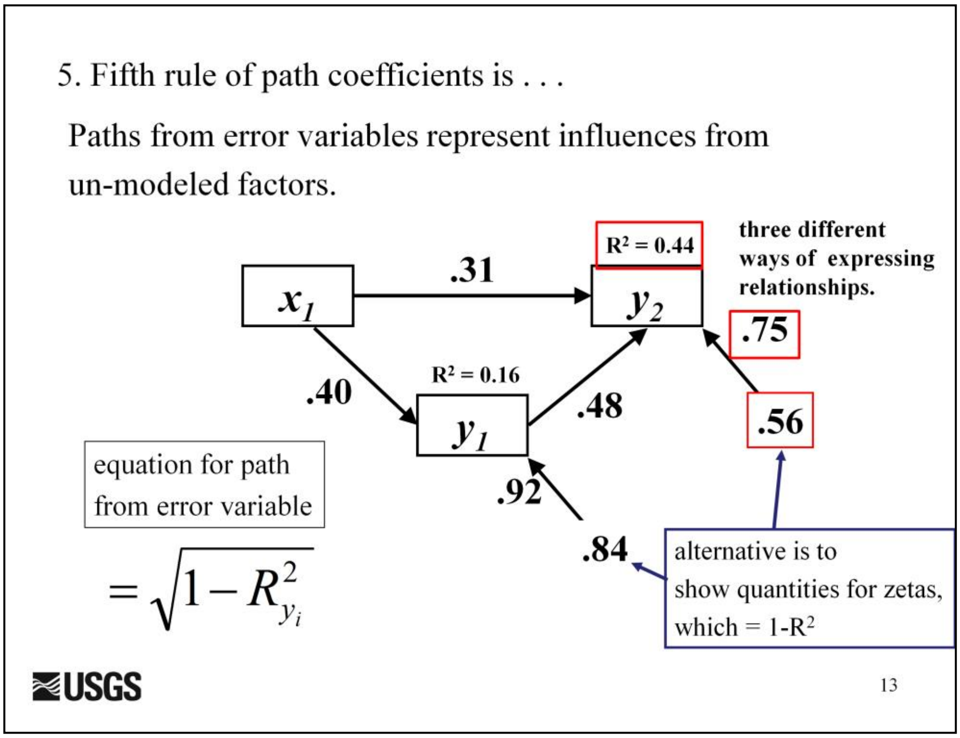


A case of shared causal influence: the unanalyzed relation between x_1 and x_2 represents the effects of an unspecified joint causal process. Therefore, x_1 and y_1 are connected by two causal paths. x_2 and y_1 likewise.

USGS

12

Returning to those pesky exogenous correlations, they have causal interpretations, but not clean ones. There are two causal connections between *x*s and *y*, but the effects on *y* are shared across the *x*s in this model.



A key part of our model is the "influences of other factors" or errors of prediction. The estimates for these quantities are computed from the other information. They can be represented in different ways and these are important to know for a variety of reasons.

First, we can express the variance explained for an endogenous variable by giving the R-square. This represents 1-error variance in standardized terms.

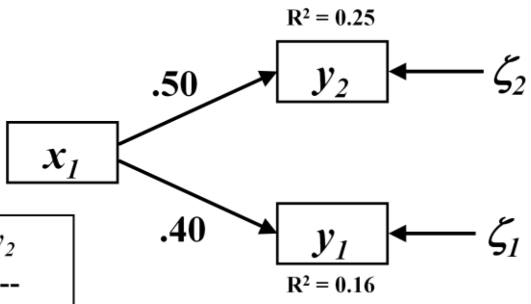
Second, we can actually present the quantity of error variances or zetas, either in raw or standardized units.

Third, if we wish to treat error variables like true causal influences, then we might use path coefficients for their effects. These = the square roots of the error variances (e.g., sqrt(0.84) = 0.92) or alternatively, the square roots of 1-R-square.

6. Sixth rule of path coefficients is . . .

Unanalyzed residual correlations between endogenous variables are partial correlations or covariances (correlations between the residuals).

Now, imagine y_1 and y_2 are joint responses.



	x_1	y_1	y_2
x_1	**1.0**		
y_1	**0.40**	**1.0**	
y_2	**0.50**	**0.60**	**1.0**

Implied correlation between y_1 and y_2 = **0.50 ∗ 0.40 = 0.20**.

USGS

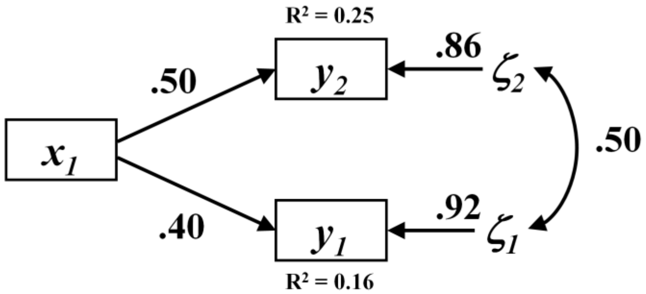
14

Considering models with joint responses are quite instructive for multiple reasons. Here our focus is on estimating residual correlations/covariances.

The basic point is that the implied correlation between two joint responses is the product of the path coefficients connecting them.

In this case, the observed correlation is 0.60 while the model-implied correlation is 0.20. This is a pretty big difference, implying a residual correlation between zeta1 and zeta2.

6. Sixth rule of path coefficients (continued).



the partial correlation between y_1 and y_2 is typically represented as a **correlation between errors**.

This implies that some other factor is influencing y_1 and y_2

Note that total correlation between y_1 and y_2 =

(0.50 ∗ 0.40) + (0.86 ∗ 0.92 ∗ ?) = 0.60 (the observed corr)

15

When the observed correlation is non-trivially different from that implied by their joint responses to x1, in this case, it implies some other (latent) factor. This additional connection is expressed in the form of a residual correlation among errors.

Our goal here is to compute the residual correlation by hand.

Step 1: Compute the R-squares for the two responses (0.16 and 0.25 in this case.

Step 2: Compute the path coefficients for the error influences on the variables. Here we subtract the R-squares from 1.0 to get the error variances, then take the square roots of those to get the path coefficients (0.92 and 0.86 in this case).

Step 3: The total correlation between y1 and y2 will be the sum of the two compound pathways connecting them.

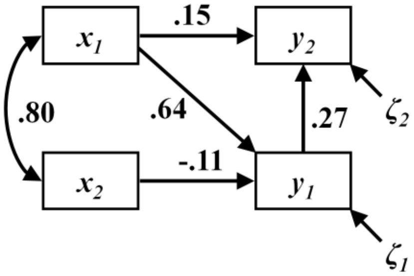
Path 1: 0.50*0.40 = 0.2

Path 2: 0.92*0.86*? = the difference

This leads us to an estimate of 0.50 for the standardized error correlation.

7. Seventh rule of path coefficients . . .

The **<u>total effect in the model</u>** one variable has on another equals the sum of its direct and indirect effects.



Total Effects in the Model:

	x_1	x_2	y_1
y_1	0.64	-0.11	---
y_2	0.32	-0.03	0.27

USGS

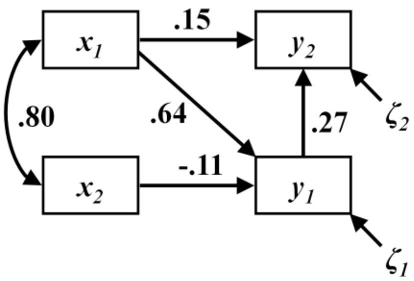
Total effect is an important concept. On the next slide we will contrast “total effect” with “total correlation”. For the moment, we simply consider that the total effect represents the causal influence of a predictor on a response when all mediating variables are allowed to change in response to changes in the predictor.

In this case, let’s consider the total effect of x1 on y2. There is a direct component = 0.15. There is also one indirect path through y1 whose effect is 0.64*0.27.

So, the total effect is the sum of the paths, or 0.15 + (0.64*0.27) = 0.32.

8. Eighth rule of path coefficients . . .

The sum of all pathways between two variables (directed and undirected) equals the correlation/covariance.



Model-implied correlation between x_1 and y_1 =
0.64 - 0.80∗0.11 = 0.55

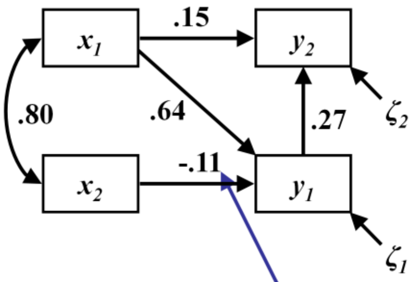
USGS

17

We already introduced this basic concept, but the total correlation between two variables represents the sum of all pathways connecting them, including causal and non-causal paths. Here, the correlation between x1 and x2 is considered a non-causal path because we have not included in our model a variable to explain that effect.

The computations for partial effects are defined as being those that will recover the net correlations observed in the data.

Suppression Effect (aka Simpson's paradox) - when presence of another variable causes path coefficient to strongly differ from bivariate correlation.



	x_1	x_2	y_1	y_2
x_1	1.0			
x_2	0.80	1.0		
y_1	0.55	0.40	1.0	
y_2	0.30	0.23	0.35	1.0

path coefficient for x_2 to y_1 very different from correlation, (results from overwhelming influence from x_1.)

USGS

18

Simpson's paradox was the observation that adding variables changes the partial effects of other variables, often in radical ways.

Here we illustrate the common case where the observed correlation is very different from the calculated causal connection. There is no real surprise here except that we often try to understand causal relations between variables using bivariate patterns when we know that is the wrong way to approach the problem.