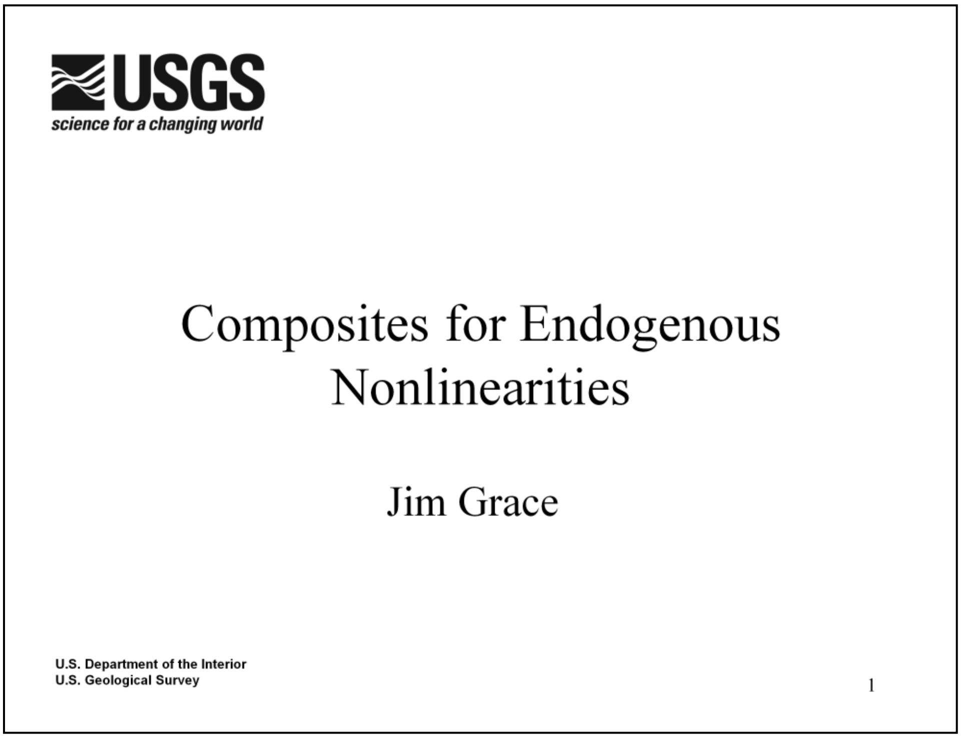


Composites for Endogenous Nonlinearities

Jim Grace

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

1

In this module I again consider an important “third” variable type, the composite. This module follows one entitled, “Composites and Formative Indicators”, which should be reviewed first.

In this module, I deal with a special situation where there is an endogenous link that is nonlinear and one we wish to model using a polynomial approach.

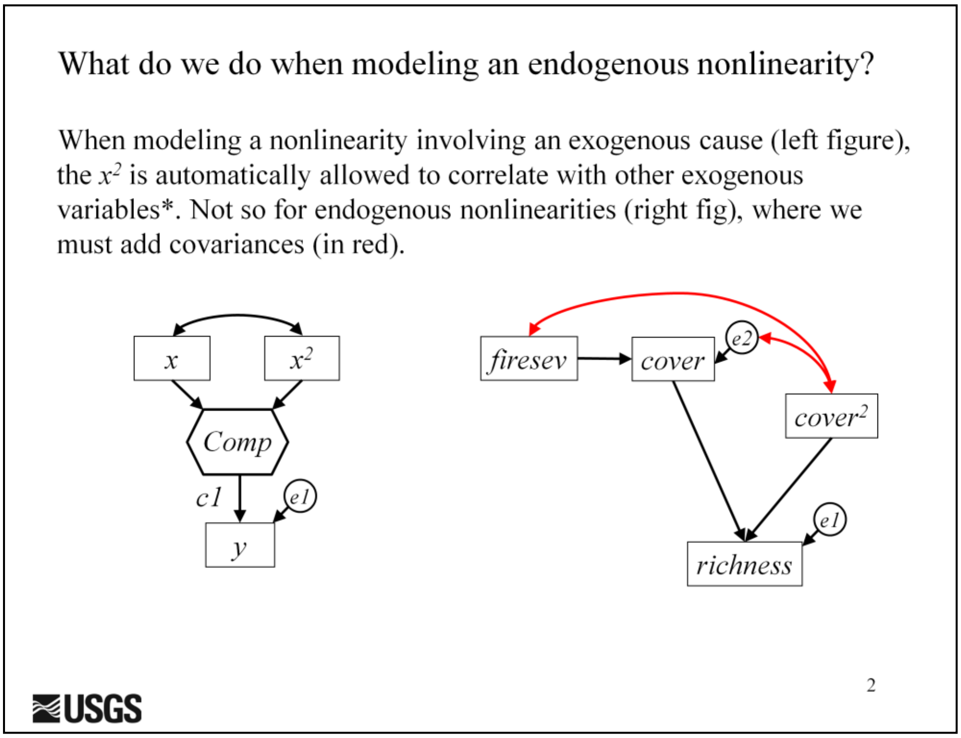
An appropriate citation for this material is

Grace, J.B. and Bollen, KA. 2008. Representing general theoretical concepts in structural equation models: the role of composite variables. Environmental and Ecological Statistics 15:191-213. (http://link.springer.com/article/10.1007/s10651-007-0047-7)

(http://www.odum.unc.edu/content/pdf/Bollen%20Grace%20Bollen%20(preprint%202008)%20Environ%20and%20Ecol%20Stats**.pdf**)

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 20141216. Questions about this

material can be sent to sem@usgs.gov.



What do we do when modeling an endogenous nonlinearity?

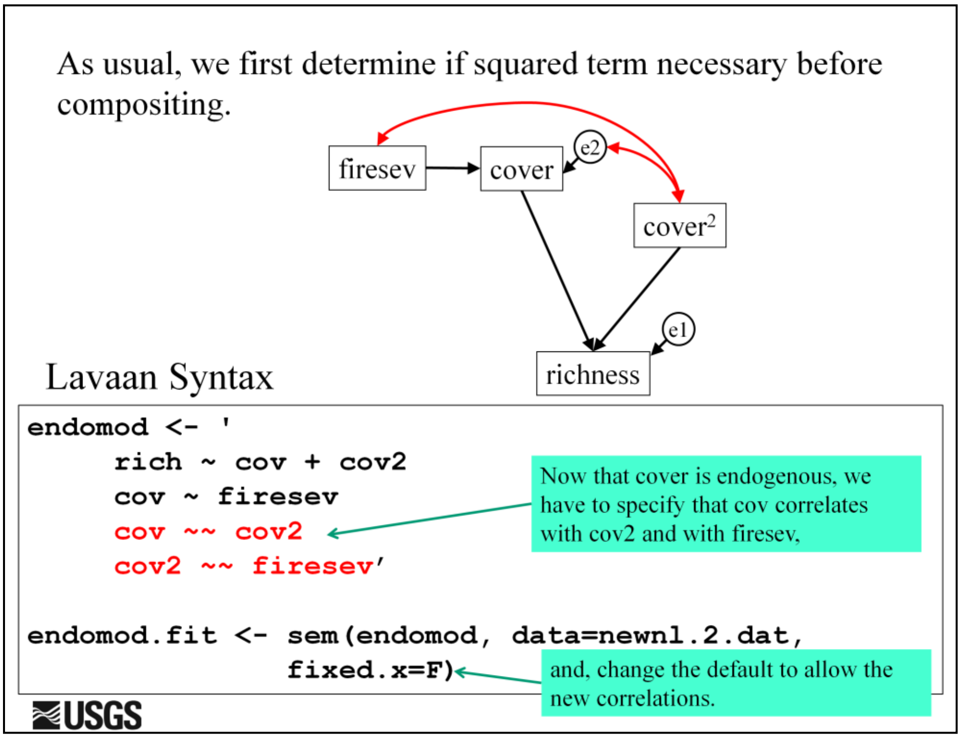
When modeling a nonlinearity involving an exogenous cause (left figure), the x^2 is automatically allowed to correlate with other exogenous variables*. Not so for endogenous nonlinearities (right fig), where we must add covariances (in red).

USGS

The left figure shows the case where we have created a composite “Comp” that captures the collective effect of x and x-square on y. This example is from the other module on composites “Composites and Formative Indicators”.

The right figure is a new situation where the nonlinear effect of cover on richness is endogenous. Stated in another way, the causes of the composite variable we intend to create, cover and cover2, are endogenous.

What is going on here is we have to deal with “cover2”, which is really not a true variable in our model, but just a device we are using. We wish to isolate any stray correlations cover2 may have with other parts of the model, in this case, with the exogenous variable firesev.



As usual, we first determine if squared term necessary before compositing.

Lavaan Syntax

```
endomod <- '
     rich ~ cov + cov2
     cov ~ firesev
     cov ~~ cov2
     cov2 ~~ firesev'

endomod.fit <- sem(endomod, data=newnl.2.dat,
               fixed.x=F)
```

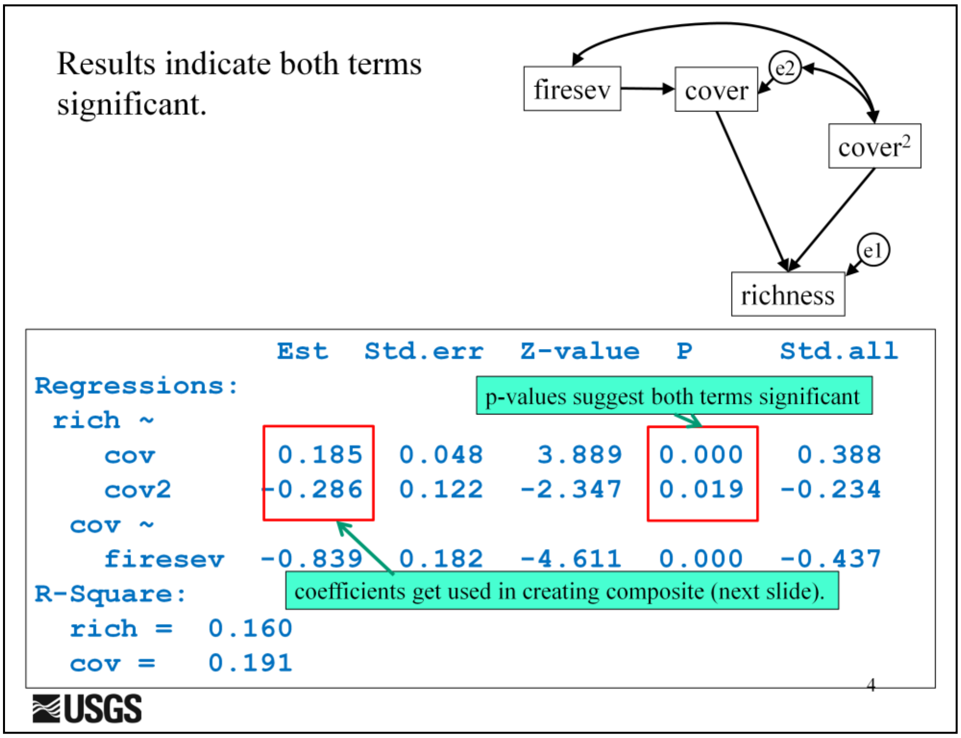
Now that cover is endogenous, we have to specify that cov correlates with cov2 and with firesev,

and, change the default to allow the new correlations.

It is not mandatory that we confirm the terms in the composite are all individually significant influences. The exception is where we have a theoretical reason for leaving in variables, say because we are comparing situations and we want comparable predictors.

In this case, however, we wish to confirm that cover and cover2 both contribute to explaining richness.

Note in red are commands used to specify the correlations between cover2 we need to include (arrows in red).



Results indicate both terms significant.

firesev
cover
e2
cover²
e1
richness

```
                Est   Std.err  Z-value  P      Std.all
Regressions:
 rich ~
    cov         0.185  0.048    3.889  0.000   0.388
    cov2       -0.286  0.122   -2.347  0.019  -0.234
  cov ~
    firesev    -0.839  0.182   -4.611  0.000  -0.437
R-Square:
  rich =   0.160
  cov =    0.191
```

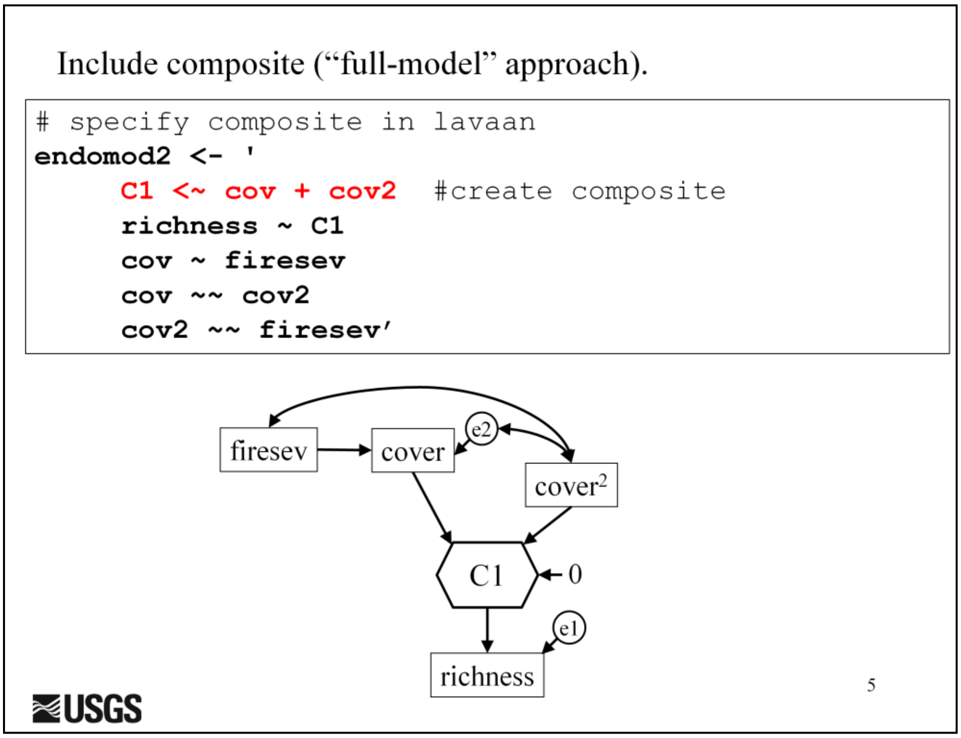
p-values suggest both terms significant

coefficients get used in creating composite (next slide).

4

USGS

Of course p-values for parameters are not the ultimate arbiters. In covariance-based SEM, the overall model fit is meant to be the authority. However, when p-values are very small, as they are here, we can be pretty confident we will confirm their contributions to the model using more formal means of model comparison (see module “Model Evaluation” if need be).



Include composite ("full-model" approach).

```
# specify composite in lavaan
endomod2 <- '
     C1 <~ cov + cov2  #create composite
     richness ~ C1
     cov ~ firesev
     cov ~~ cov2
     cov2 ~~ firesev'
```

5

USGS

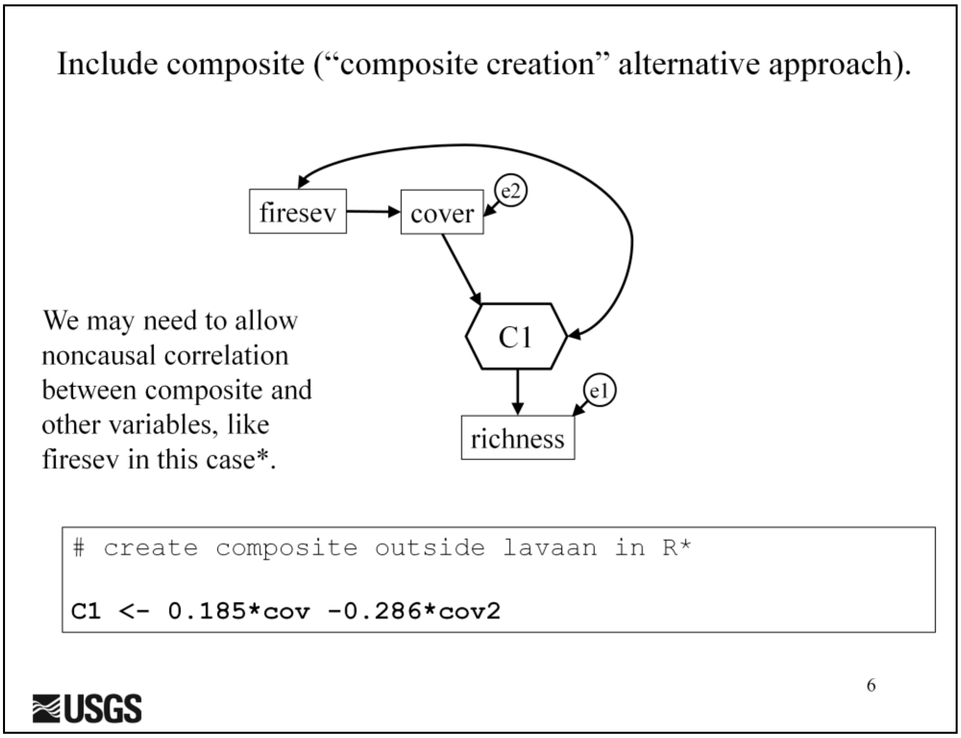
Here is the original method presented by Grace and Bollen. This is the more general case and the one that extends to more theoretical situations, such as latent composites.

Lavaan, based on our work on composites, and with the help of Jarrett Byrnes, is the only software with a separate operator for specifying composite variables, which is "<~".

In this case we specify the composite and then the other relations. Lavaan automatically creates a zero-variance composite.

Note 1: cover squared (cover2) is being used as a "device" here to allow us to fit curvilinear relationships. When we bring such variables into models, we should treat them as exogenous and control for their correlations with other variables (i.e., include correlations, such as with the error of cover and with firesev).

Note 2: it is not uncommon for this model type to have problems converging. For this reason, on the next page I show a different approach that will avoid convergence problems.



Include composite (“composite creation” alternative approach).

We may need to allow noncausal correlation between composite and other variables, like firesev in this case*.

```
# create composite outside lavaan in R*

C1 <- 0.185*cov -0.286*cov2
```

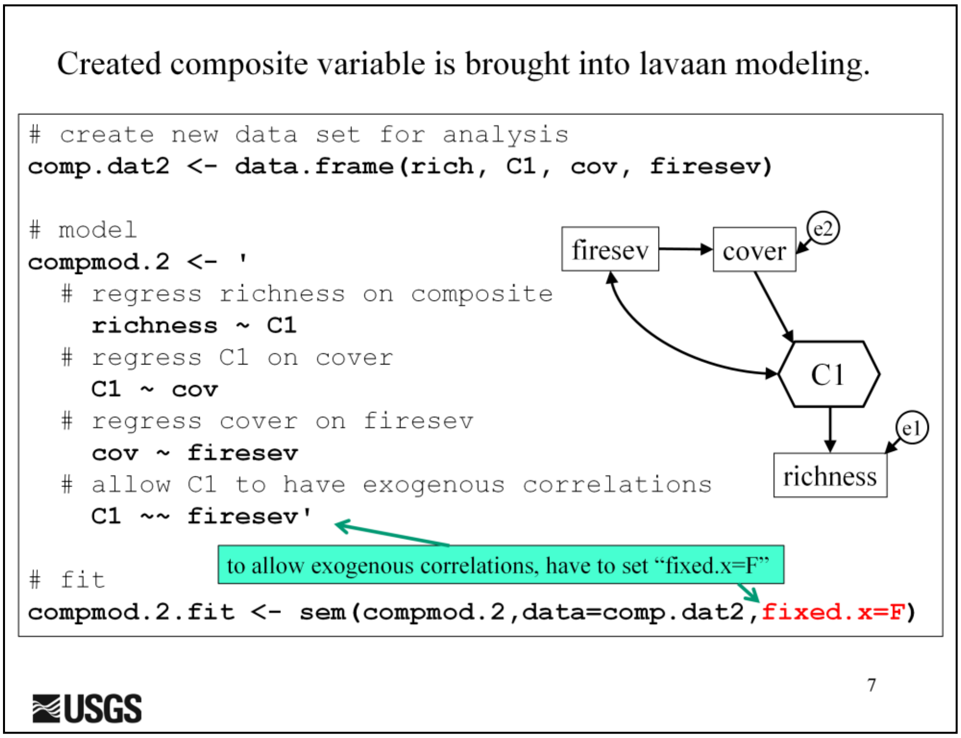
6

In this alternative approach, we use the results from the model omitting the composite (see previous slides) to compute composite scores outside of lavaan.

So, we can simply compute values for C1 using coefficients from prior lavaan model.

Through this we see that the composite is essentially the multiple regression predicted scores for the effects of cover and cover2 on richness. It can get more complicated than that, but isn’t in this case.

Note: The exogenous correlation between firesev and cover2 in the last slide is now represented by a correlation between C1 and firesev. Again, this is just a bit of statistical control being used to manage features that come along with nonlinear modeling.



Created composite variable is brought into lavaan modeling.

```
# create new data set for analysis
comp.dat2 <- data.frame(rich, C1, cov, firesev)

# model
compmod.2 <- '
  # regress richness on composite
    richness ~ C1
  # regress C1 on cover
    C1 ~ cov
  # regress cover on firesev
    cov ~ firesev
  # allow C1 to have exogenous correlations
    C1 ~~ firesev'

# fit
compmod.2.fit <- sem(compmod.2,data=comp.dat2,fixed.x=F)
```

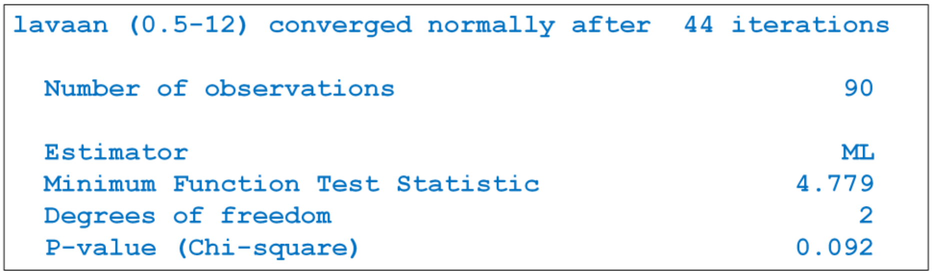
to allow exogenous correlations, have to set "fixed.x=F"

USGS

The first command in the slide creates a new data set that includes the composite scores (C1) and omits the cover2 variables.

Now we are modeling with the composite and its causal source, "cover".

Here are results from the alternative approach.



```
lavaan (0.5-12) converged normally after  44 iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic                4.779
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.092
```

Results suggest model fit is OK, which basically means there is no additional direct link from firesev to richness required in the model. The other degree of freedom is because there is no direct link from cover to richness, which would not make any sense (though if it were indicated would suggest a problem with the construction of C1).

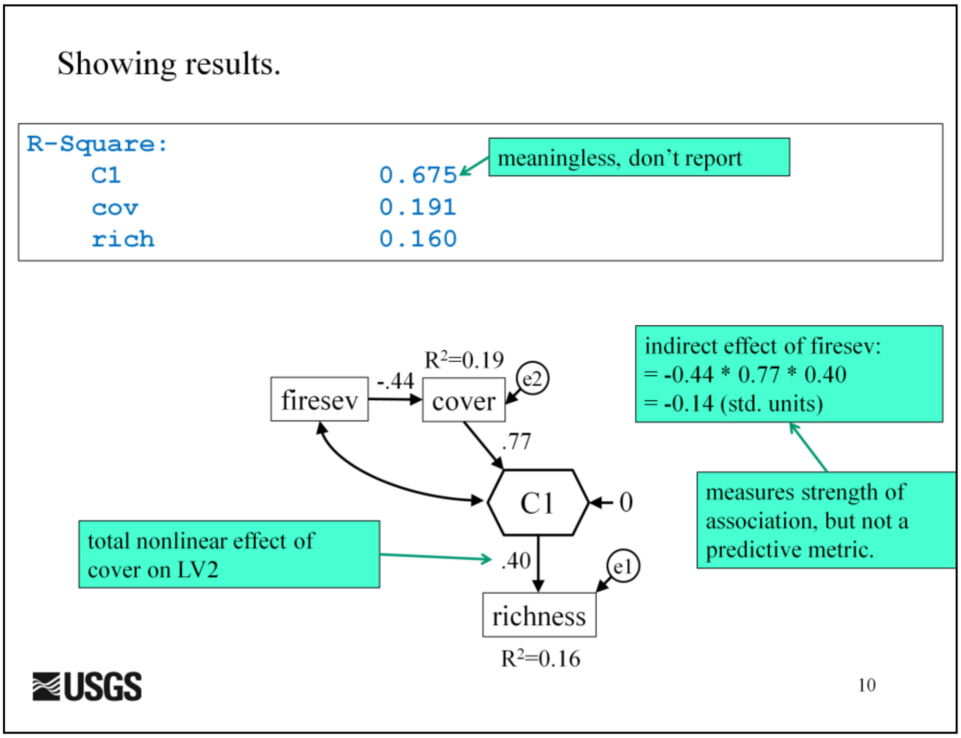
More results from the alternative approach.

```
               Estimate  Std.err  Z-value  P(>|z|)   Std.all

Regressions:
  richness ~
    C1            1.000    0.242    4.137    0.000     0.400
  C1 ~
    cover         0.146    0.012   11.793    0.000     0.769
  cover ~
    firesev      -0.839    0.182   -4.611    0.000    -0.437

Covariances:
  C1 ~~
    firesev      -0.001    0.001   -1.831    0.067    -0.218
```

we typically control for exogenous correlations even if ns

USGS

Now, here are results for our composite model.



Showing results.

```
R-Square:
    C1                 0.675
    cov                0.191
    rich               0.160
```

meaningless, don’t report

indirect effect of firesev:
= -0.44 * 0.77 * 0.40
= -0.14 (std. units)

measures strength of association, but not a predictive metric.

total nonlinear effect of cover on LV2

USGS

10

And here are some useful computations and how they would be made in this situation.