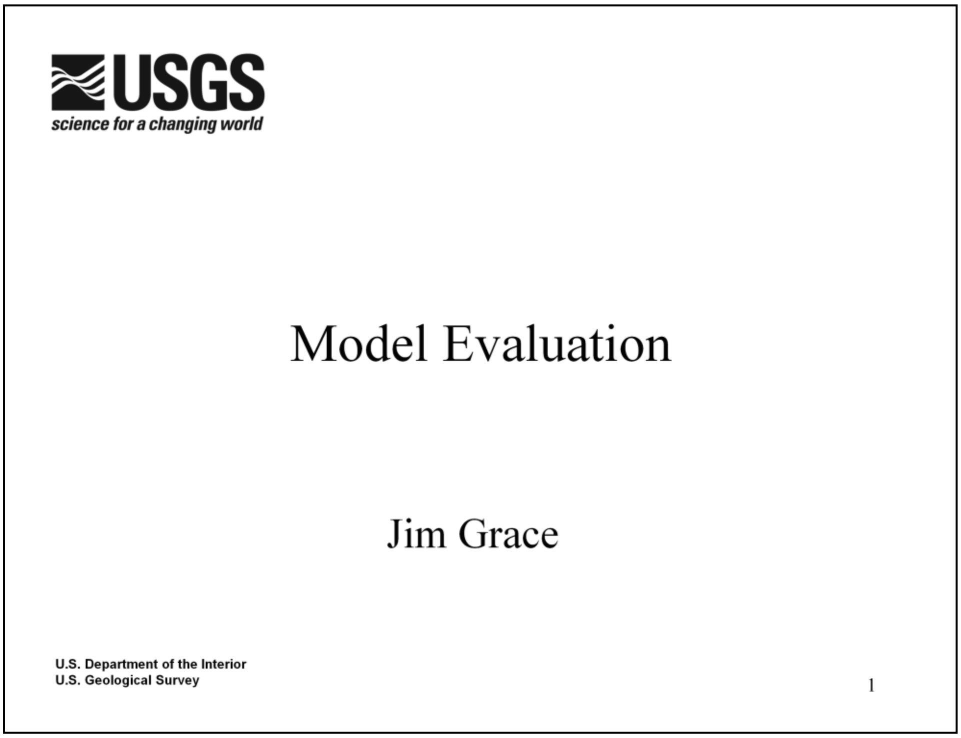


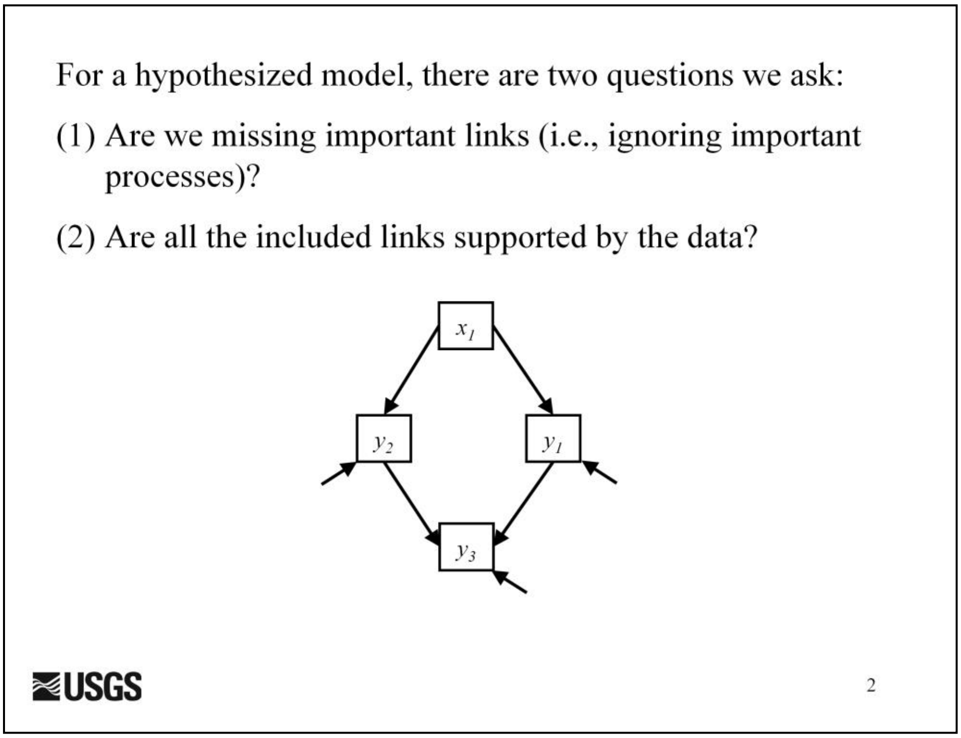
Model Evaluation

Jim Grace

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

This module deals with the important topic of how one evaluates an estimated model, determines if it is sufficient, and evaluates the support from data.

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. This is version 1.1 (last revised 20150319)



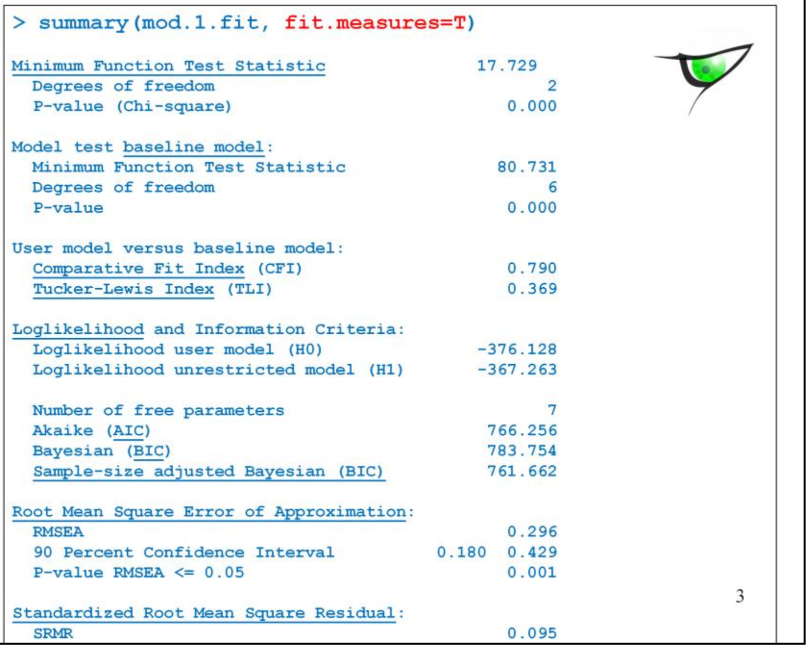
For a hypothesized model, there are two questions we ask:

(1) Are we missing important links (i.e., ignoring important processes)?

(2) Are all the included links supported by the data?

For unsaturated networks, we have two different, though related, issues to consider, the first being a bit novel.

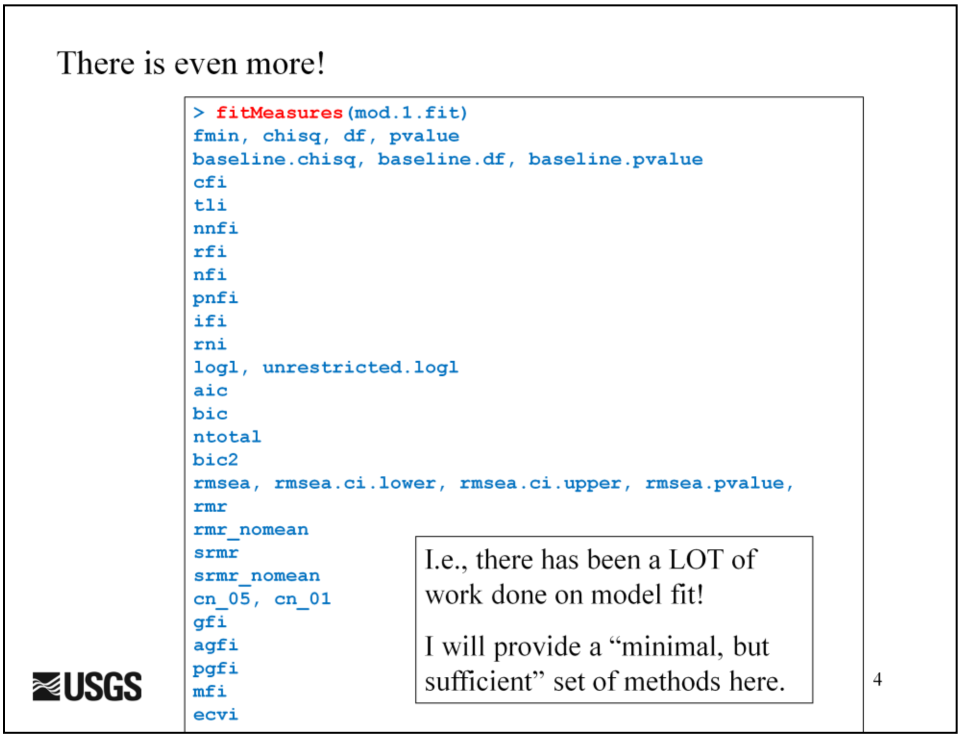
Many indices have been developed for evaluating model-data fit.



```
> summary(mod.1.fit, fit.measures=T)

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

Model test baseline model:
  Minimum Function Test Statistic             80.731
  Degrees of freedom                               6
  P-value                                      0.000

User model versus baseline model:
  Comparative Fit Index (CFI)                  0.790
  Tucker-Lewis Index (TLI)                     0.369

Loglikelihood and Information Criteria:
  Loglikelihood user model (H0)             -376.128
  Loglikelihood unrestricted model (H1)     -367.263

  Number of free parameters                        7
  Akaike (AIC)                               766.256
  Bayesian (BIC)                             783.754
  Sample-size adjusted Bayesian (BIC)        761.662

Root Mean Square Error of Approximation:
  RMSEA                                        0.296
  90 Percent Confidence Interval          0.180  0.429
  P-value RMSEA <= 0.05                        0.001

Standardized Root Mean Square Residual:
  SRMR                                         0.095
```

3

I don’t expect you to be able to read all this. For now I just want to make the point that lavaan will give you a great deal of information if you ask for it using “fit.measures=T” in the summary command.



There is even more!

```
> fitMeasures(mod.1.fit)
fmin, chisq, df, pvalue
baseline.chisq, baseline.df, baseline.pvalue
cfi
tli
nnfi
rfi
nfi
pnfi
ifi
rni
logl, unrestricted.logl
aic
bic
ntotal
bic2
rmsea, rmsea.ci.lower, rmsea.ci.upper, rmsea.pvalue,
rmr
rmr_nomean
srmr
srmr_nomean
cn_05, cn_01
gfi
agfi
pgfi
mfi
ecvi
```

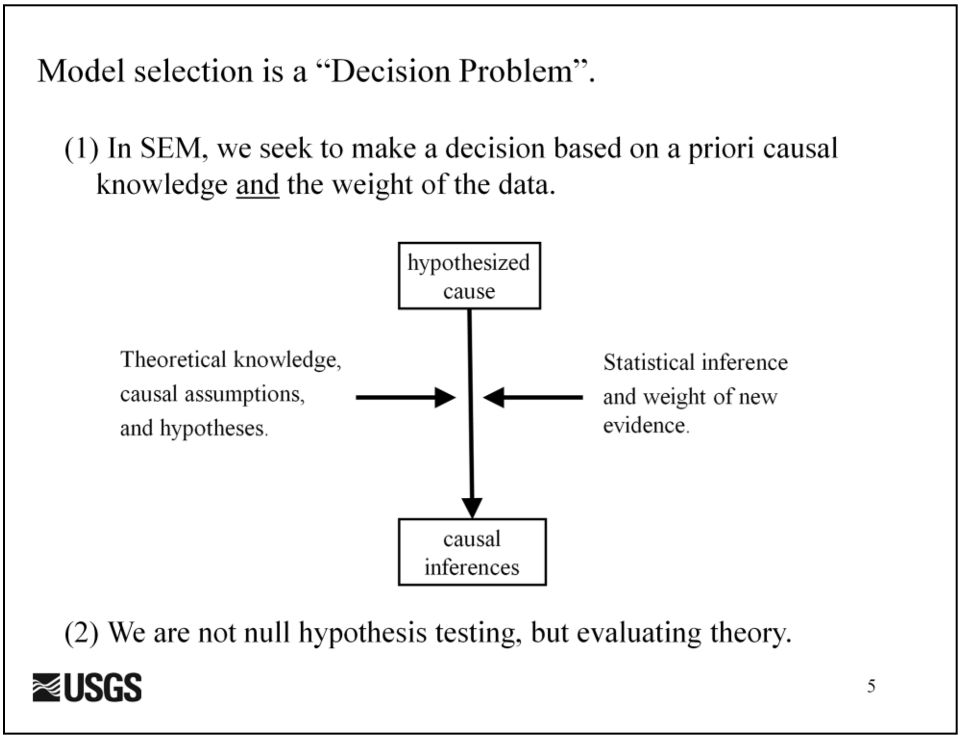
I.e., there has been a LOT of work done on model fit!

I will provide a “minimal, but sufficient” set of methods here.

Actually, lavaan calculates even more fit indices, which requires a different command to obtain “fitMeasures(fit.object)?

Again, I don’t expect you to read all this. For now just know that SEMers have been preoccupied with this topic since the early 1970s.

Note, in this module I will provide you with a practical approach to evaluating models. There are many different approaches that can be used and I fear it will be distracting to you if I cover a broad variety of perspectives. So, I am, for now, ignoring many of the options, while providing a sufficient set of test procedures for the basic task.



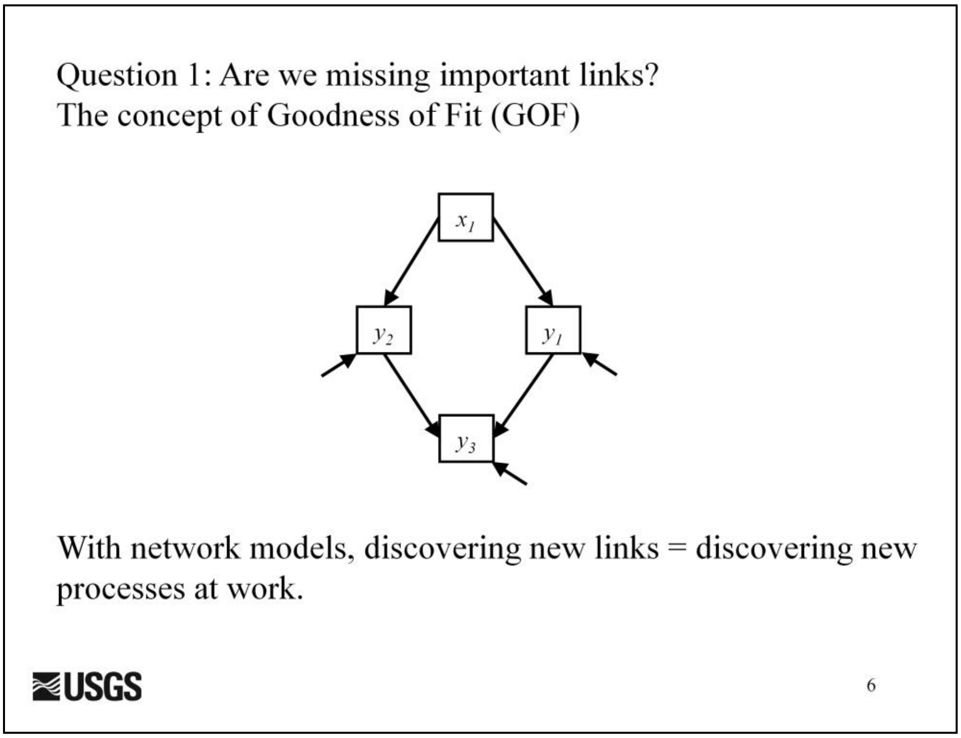
Model selection is a "Decision Problem".

(1) In SEM, we seek to make a decision based on a priori causal knowledge and the weight of the data.

hypothesized cause
Theoretical knowledge, causal assumptions, and hypotheses.
Statistical inference and weight of new evidence.
causal inferences

(2) We are not null hypothesis testing, but evaluating theory.

We should not expect typical data sets to be our definitive source of conclusions for causal models. This seems like a radical idea, but the current data set is not our total knowledge on a subject. Causal modeling has to be built through knowledge accumulation.

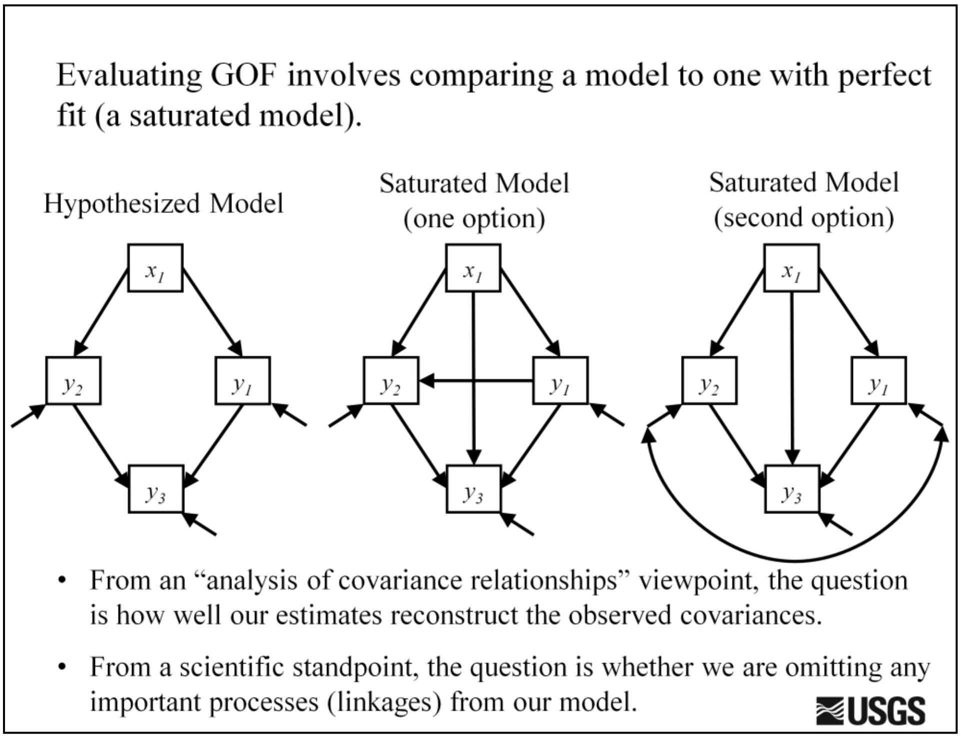


Question 1: Are we missing important links?
The concept of Goodness of Fit (GOF)

With network models, discovering new links = discovering new processes at work.

USGS

A minimum requirement for network models under global estimation is that no links of major importance are omitted. If they are we should not believe the parameter estimates. DON’T use such models for interpretation.



Evaluating GOF involves comparing a model to one with perfect fit (a saturated model).

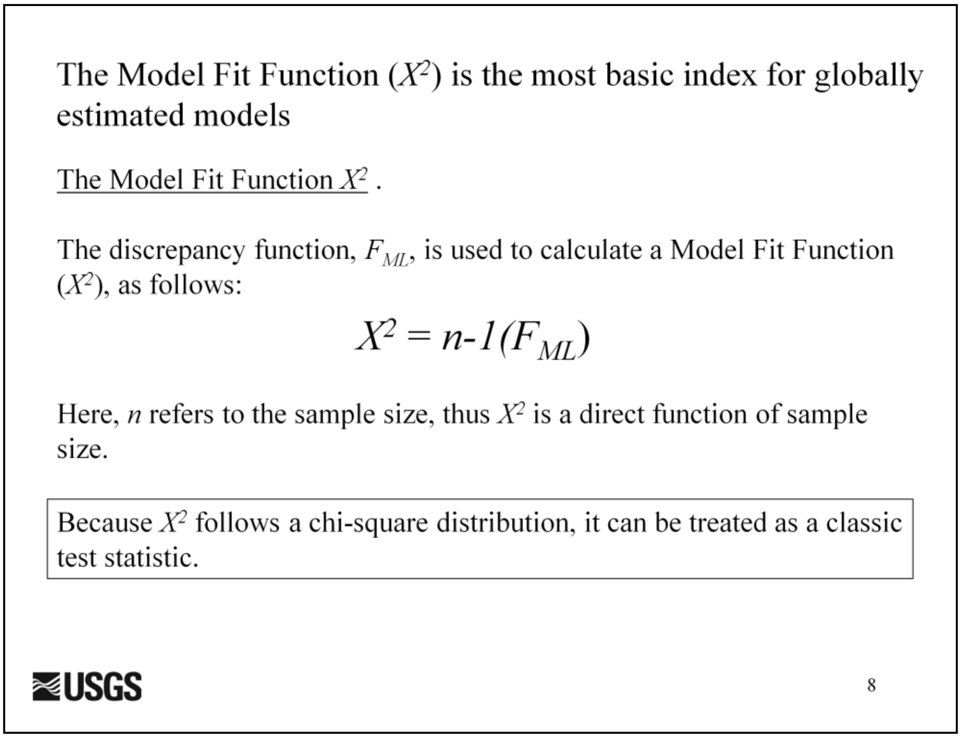
- From an “analysis of covariance relationships” viewpoint, the question is how well our estimates reconstruct the observed covariances.
- From a scientific standpoint, the question is whether we are omitting any important processes (linkages) from our model.

USGS

Under global estimation, our comparison is to a saturated model. The reason is that saturated models permit every covariance to be explained, so our Fml fit function goes to zero.

For an unsaturated model to be “sufficient” it needs to allow things to “mostly add up”.

From a local-estimation approach our question is only, “do the implied conditional independences hold”. These two perspectives ought to converge in the majority of cases.



The Model Fit Function (X^2) is the most basic index for globally estimated models

The Model Fit Function X^2 .

The discrepancy function, F_{ML}, is used to calculate a Model Fit Function (X^2), as follows:

$$X^2 = n\text{-}1(F_{ML})$$

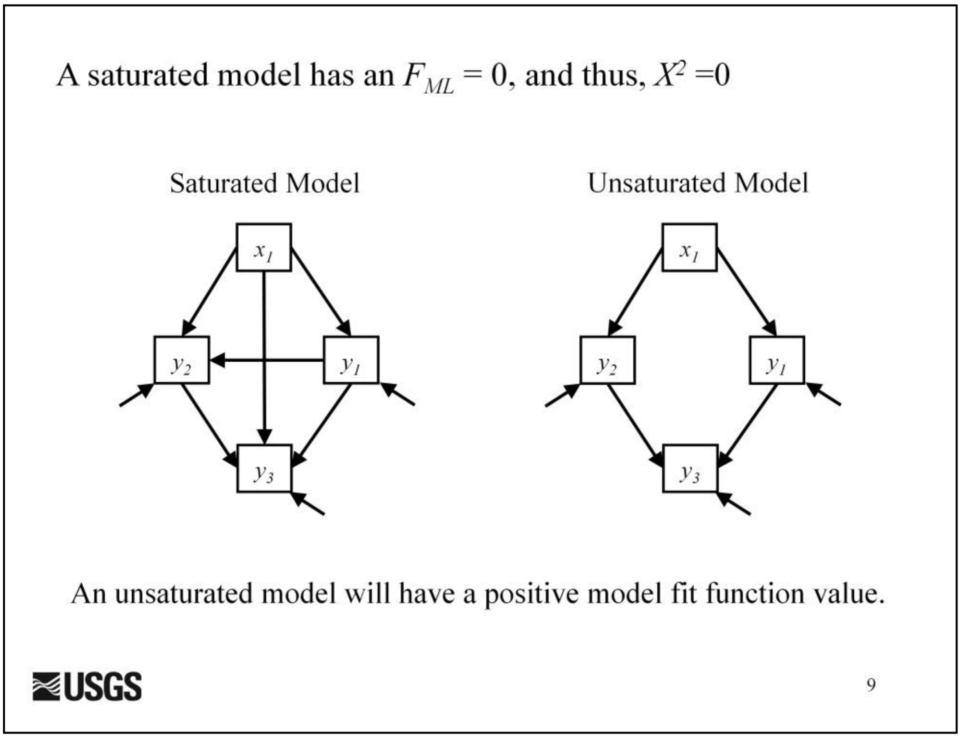
Here, *n* refers to the sample size, thus X^2 is a direct function of sample size.

Because X^2 follows a chi-square distribution, it can be treated as a classic test statistic.

USGS

It was noted early on in modern SEM that the fit function follows a chi-square distribution. We calculate our model “Chi-square statistic” directly from the model discrepancy, summed over the samples.

Important here is the single-degree-of-freedom criterion, which is a handly device in SEM.



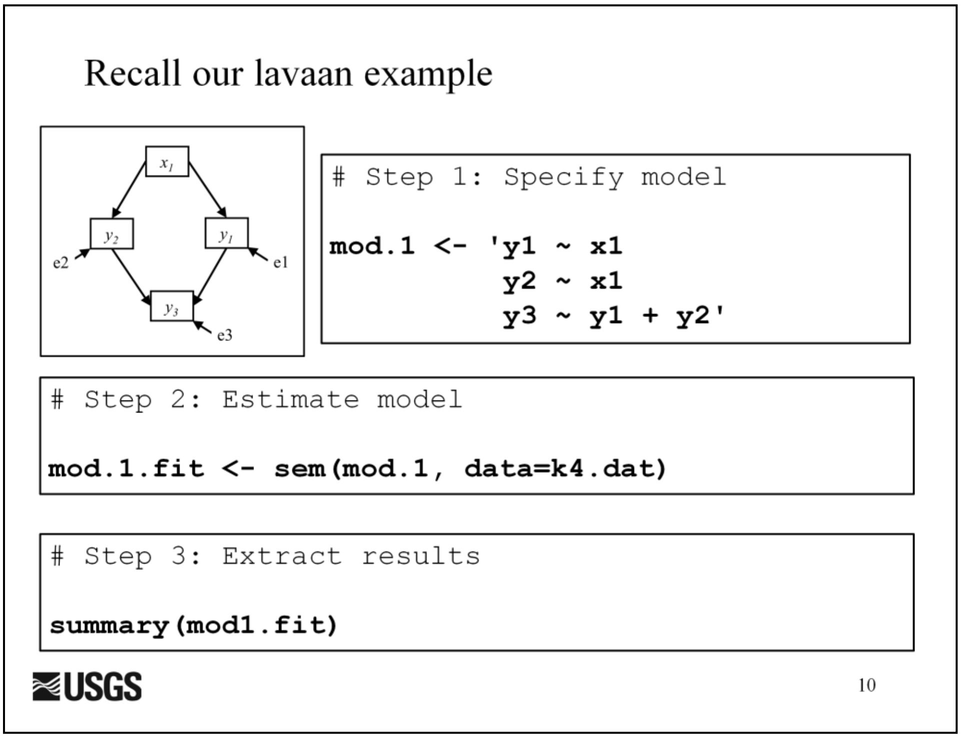
A saturated model has an F_{ML} = 0, and thus, X^2 =0

Saturated Model

Unsaturated Model

An unsaturated model will have a positive model fit function value.

This slide serves as a reminder that overall goodness of fit involves reference to a saturated model and for the classic-type model we are dealing with, such models have zero discrepancy.



Recall our lavaan example

```
# Step 1: Specify model

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

```
# Step 2: Estimate model

mod.1.fit <- sem(mod.1, data=k4.dat)
```

```
# Step 3: Extract results

summary(mod1.fit)
```

USGS

10

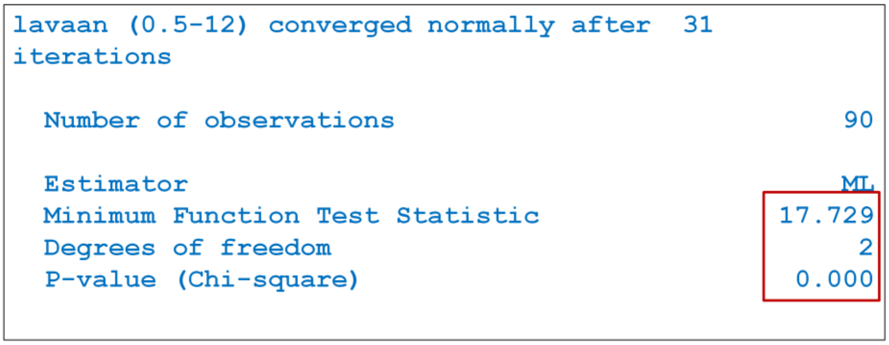
Here, once again, is the generic example model from the module “Brief Intro to Lavaan”. Again, three steps in lavaan:

(1) specify model using lavaan’s code,

(2) use the “sem” function to estimate the parameters (“fit”) for the specified model, and

(3) extract information from the fitted object.

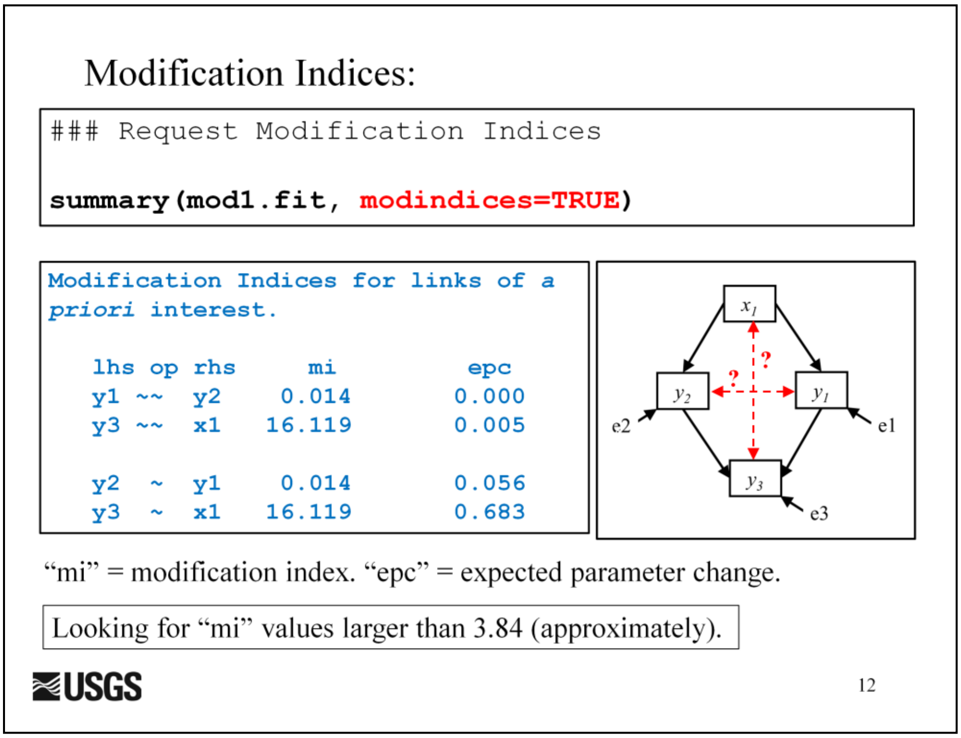
Recall lavaan Results.



```
lavaan (0.5-12) converged normally after  31
iterations

  Number of observations                            90

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

We seek a p-value for the model that is GREATER than 0.05, which would mean no significant discrepancy between model and data. Here, by any usual measure, this discrepancy is large and model fit is poor!

Here in blue is output in R. Now we understand that a chi-square of 17 with 2 df and a p-value less than 1 in 1000 indicates a large discrepancy between data and model-implied covariances.



Modification Indices:

```
### Request Modification Indices

summary(mod1.fit, modindices=TRUE)
```

```
Modification Indices for links of a
priori interest.

   lhs op rhs      mi         epc
   y1 ~~  y2    0.014       0.000
   y3 ~~  x1   16.119       0.005

   y2  ~  y1    0.014       0.056
   y3  ~  x1   16.119       0.683
```

x_1
?
?
y_2
y_1
e2
e1
y_3
e3

“mi” = modification index. “epc” = expected parameter change.

Looking for “mi” values larger than 3.84 (approximately).

USGS

12

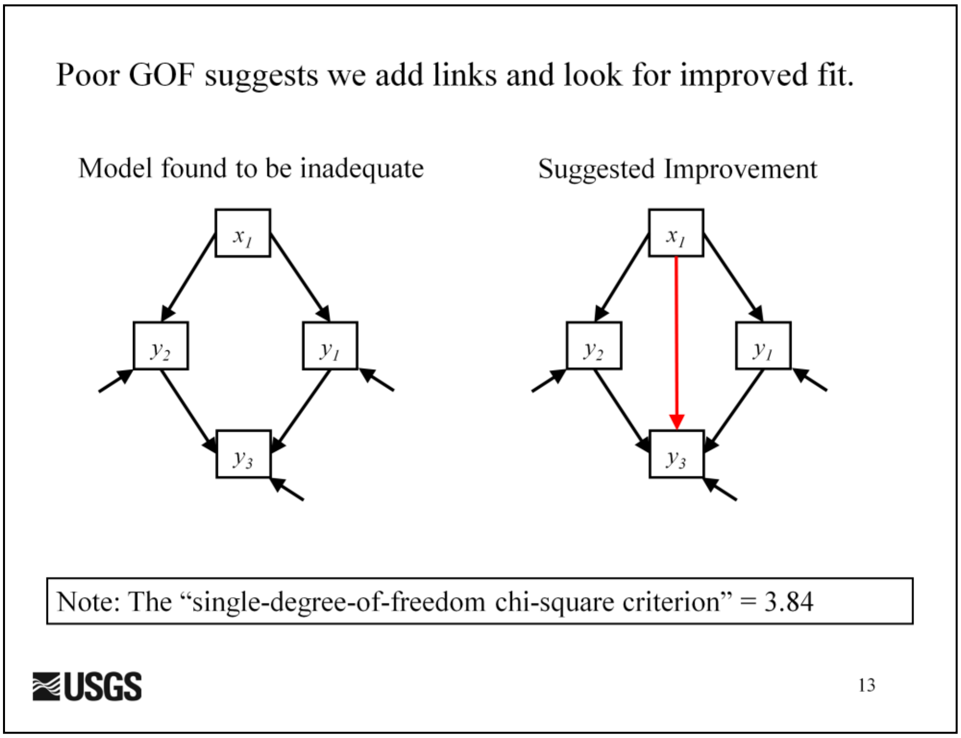
When we have a discrepant model, we need help sometimes in knowing where to look for suggested changes. Lavaan like all sem software can compute “modification indices”. These are uninformed suggestions that should not be followed blindly.

In this case, we already know what the logical candidates for additional paths are. If we have some experience with SEM and with this system, we will have already thought about some of the reasons variables might not be conditionally independent.

Of course we could just saturate our models at the outset, looking for everything and then just prune paths. That is fine for exploratory modeling, but not good practice for confirmatory modeling.

Here I show results for the two missing connections. Shown are both correlations and directed relationships, both of which would yield similar improvements in model fit, though mean very different things scientifically.

We can ignore the epc, but the mi is an estimated reduction in model chi-square we would presumably obtain by including that link.



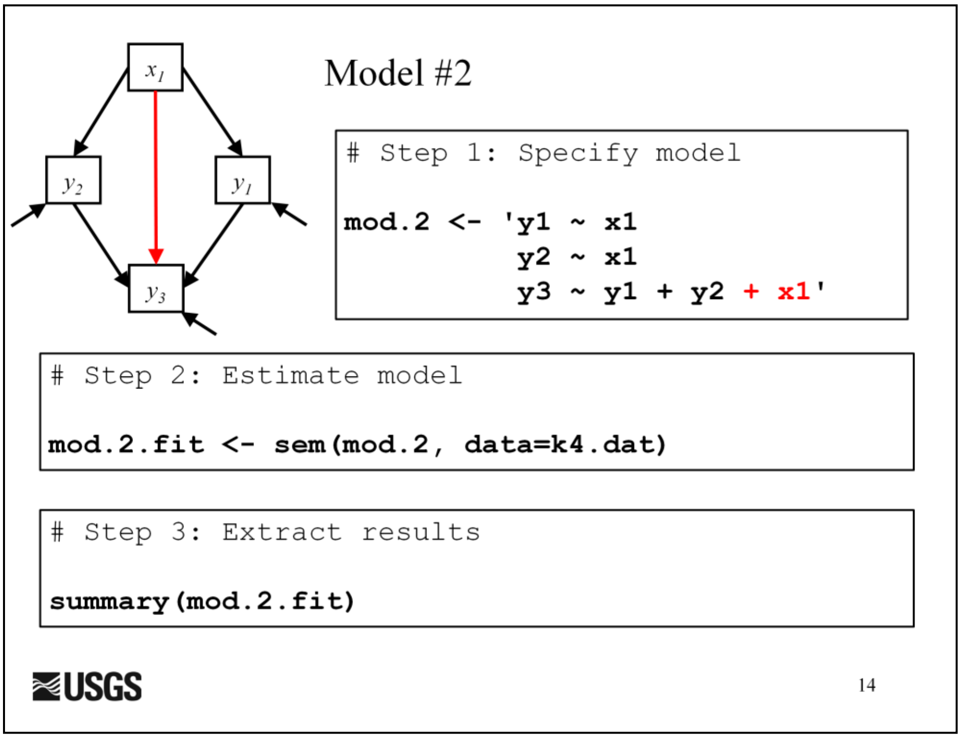
Poor GOF suggests we add links and look for improved fit.

Model found to be inadequate

Suggested Improvement

Note: The “single-degree-of-freedom chi-square criterion” = 3.84

Based on what we have seen and what we know about the variables, we would be likely to add a directed path from x1 to y3. This is still not a saturated model, so we will still get a test as to whether it is as good as a saturated model.



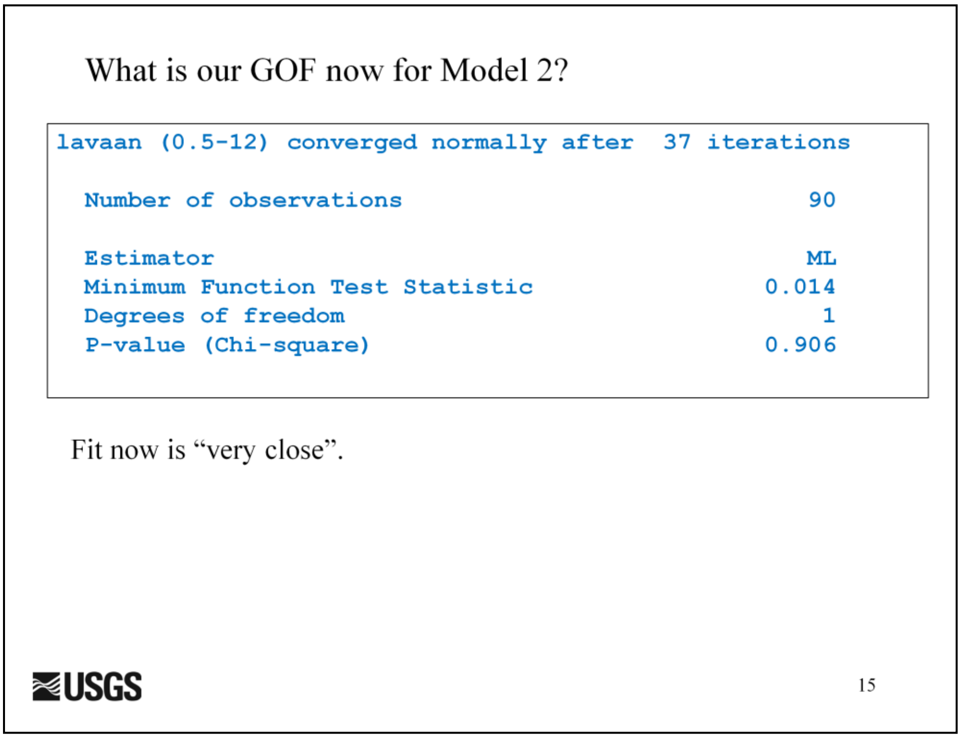
Model #2

```
# Step 1: Specify model

mod.2 <- 'y1 ~ x1
          y2 ~ x1
          y3 ~ y1 + y2 + x1'
```

```
# Step 2: Estimate model

mod.2.fit <- sem(mod.2, data=k4.dat)
```

```
# Step 3: Extract results

summary(mod.2.fit)
```

For Model 2, we add the bit in red to complete the code.

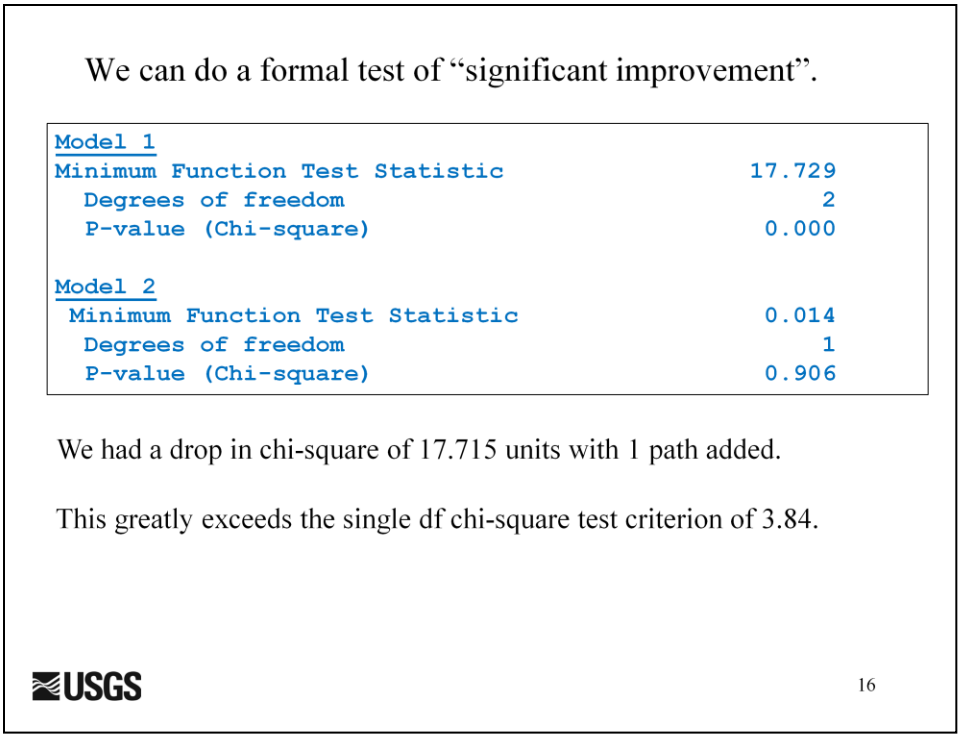


What is our GOF now for Model 2?

```
lavaan (0.5-12) converged normally after  37 iterations

  Number of observations                            90

  Estimator                                         ML
  Minimum Function Test Statistic                0.014
  Degrees of freedom                                 1
  P-value (Chi-square)                           0.906
```

Fit now is “very close”.

Adding the link made a big different. Our results imply y1 and y2 cannot explain all effects of x1 on y3.



We can do a formal test of “significant improvement”.

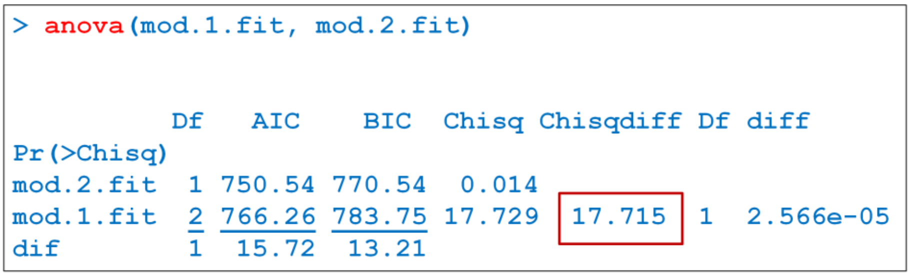
```
Model 1
Minimum Function Test Statistic                   17.729
  Degrees of freedom                                   2
  P-value (Chi-square)                             0.000

Model 2
 Minimum Function Test Statistic                   0.014
  Degrees of freedom                                   1
  P-value (Chi-square)                             0.906
```

We had a drop in chi-square of 17.715 units with 1 path added.

This greatly exceeds the single df chi-square test criterion of 3.84.

Formally, here is where we would rely on the single-degree-of-freedom chi-square test. The observed drop of 17.715 greatly exceeds the criterion of 3.84 for a classical, p-value based hypothesis test.

Option 2: Using ANOVA function to compare models.

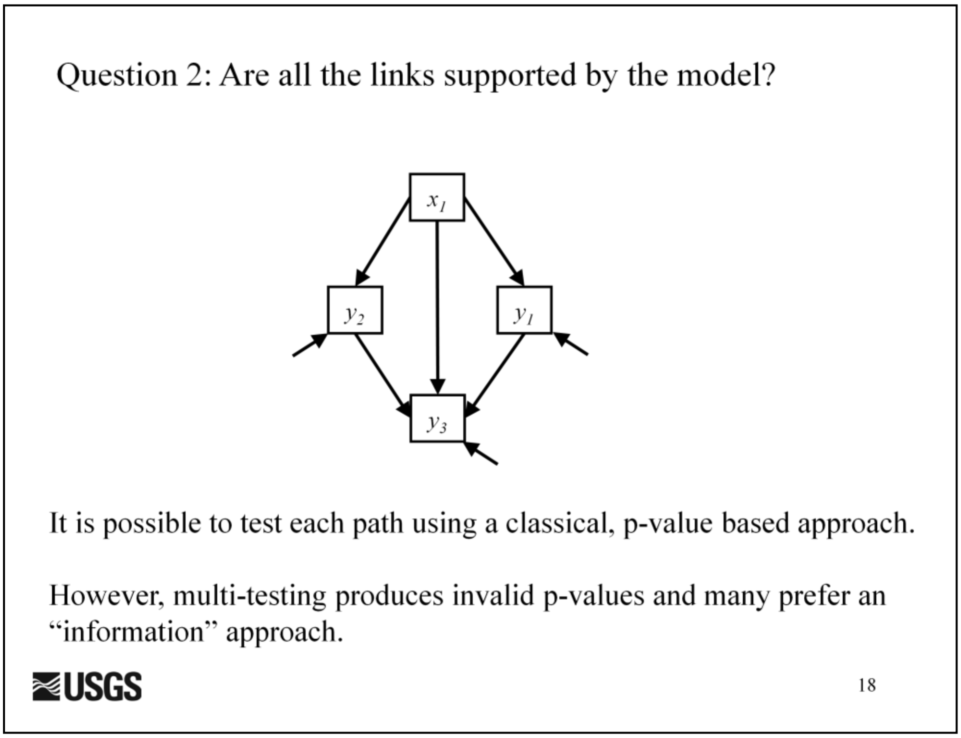


```
> anova(mod.1.fit, mod.2.fit)


           Df   AIC    BIC  Chisq Chisqdiff Df diff
Pr(>Chisq)
mod.2.fit   1 750.54 770.54  0.014
mod.1.fit   2 766.26 783.75 17.729  17.715  1   2.566e-05
dif         1  15.72  13.21
```

Note: “anova” also gives us AIC and BIC comparisons.

Gives us same chi-square test result as our by-hand comparison.

We can perform likelihood-ratio tests for a set of models using the ‘anova’ function with a lavaan object. Aside from giving AIC values for models to compare, it also automates a chi-square comparison between models.



Question 2: Are all the links supported by the model?

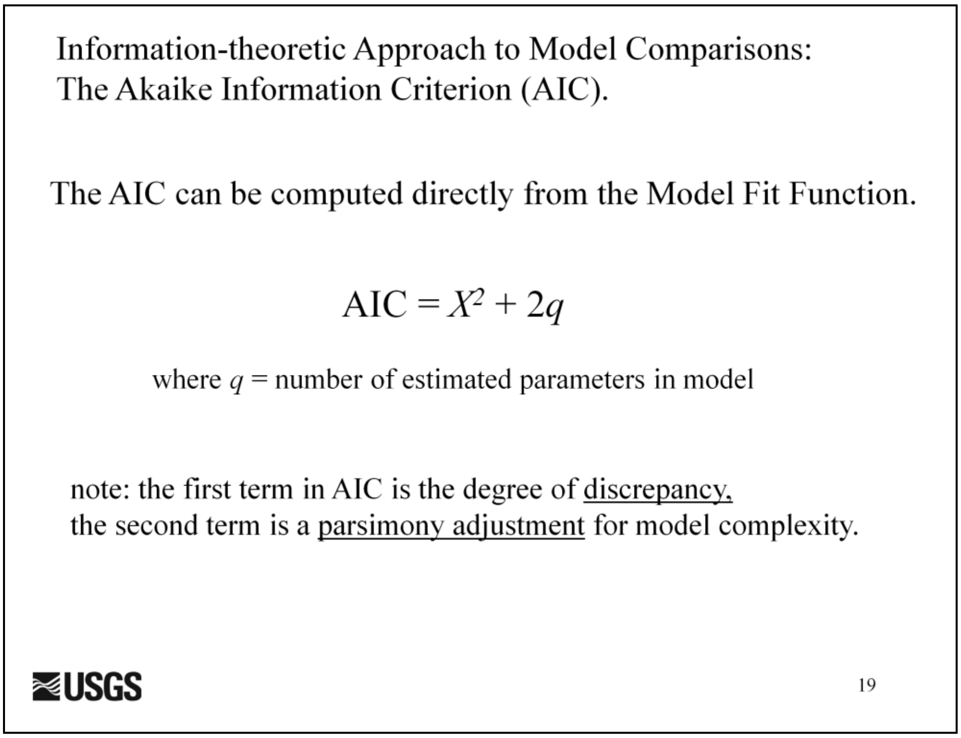
It is possible to test each path using a classical, p-value based approach.

However, multi-testing produces invalid p-values and many prefer an "information" approach.

USGS

18

OK, we have made sure we are not missing major links in our model. Only now can we begin to address the question of whether we have paths in the model that are not supported by the data.



Information-theoretic Approach to Model Comparisons: The Akaike Information Criterion (AIC).

The AIC can be computed directly from the Model Fit Function.

$$\text{AIC} = X^2 + 2q$$

where q = number of estimated parameters in model

note: the first term in AIC is the degree of discrepancy, the second term is a parsimony adjustment for model complexity.

USGS

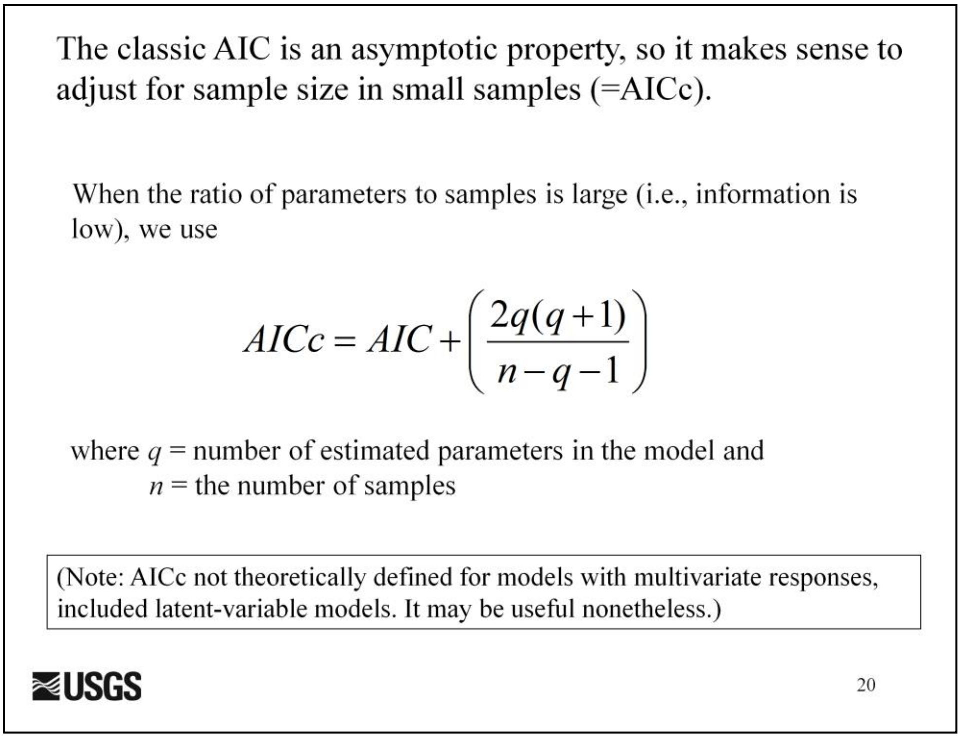
Interest in AIC for model comparison has grown in biology. There are lots of important points that could be made about this, but I will save that for another time and place. What is most relevant, it to understand that since classical SEM is a model-centered, likelihood-based approach, AIC is a direct computation from the Model Fit Function.

A key reference for the use of AIC is

Burnham, K.P. and Anderson , D.R. 2002. Model Selection and Multimodel Inference. 2nd Ed. Springer Verlag.

There is a Forum of viewpoints provided in Ecology recently that discuss the merits of p-values and AIC.

http://www.esajournals.org/toc/ecol/95/3



The classic AIC is an asymptotic property, so it makes sense to adjust for sample size in small samples (=AICc).

When the ratio of parameters to samples is large (i.e., information is low), we use

$$AICc = AIC + \left(\frac{2q(q+1)}{n-q-1} \right)$$

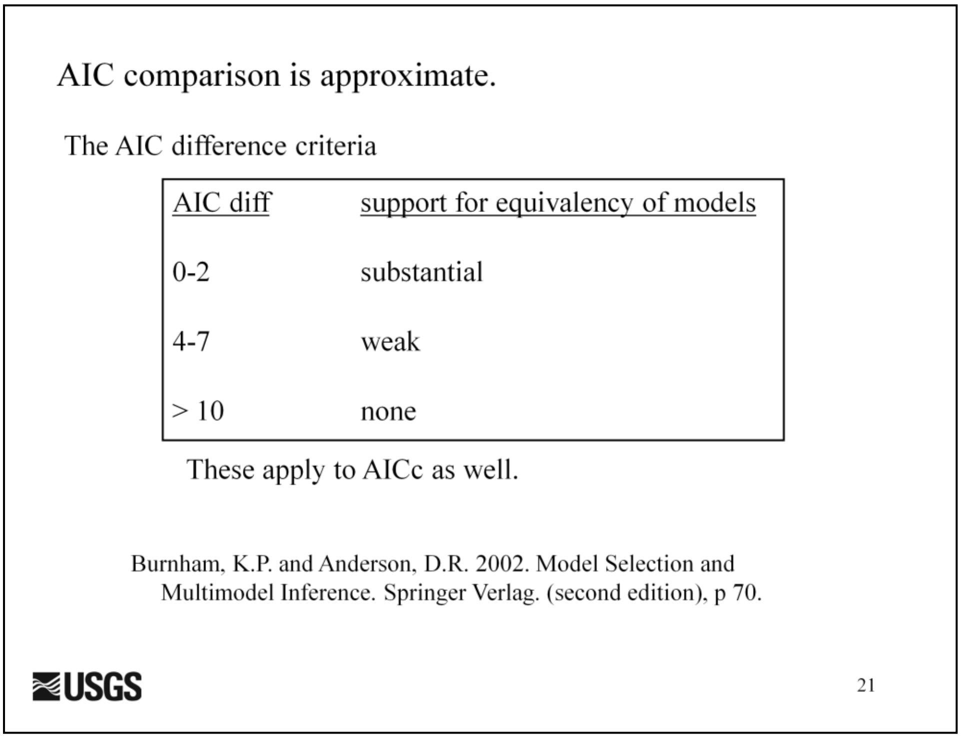
where q = number of estimated parameters in the model and
n = the number of samples

(Note: AICc not theoretically defined for models with multivariate responses, included latent-variable models. It may be useful nonetheless.)

At the moment, for moderate to small samples (250 samples or less), the AICc seems like a good choice for model selection.

With the AICc, we adjust the AIC for the ratio of information/samples to parameters in the model. This is a reasonable suggestion because it takes information to estimate parameters and the ratio of information to parameters is a handy way to discuss sample size recommendations, though such things are truly not simple.

Anyway, the AICc has a more complex parsimony correction term than does the AIC (which is just 2q), as shown in the slide.



AIC comparison is approximate.

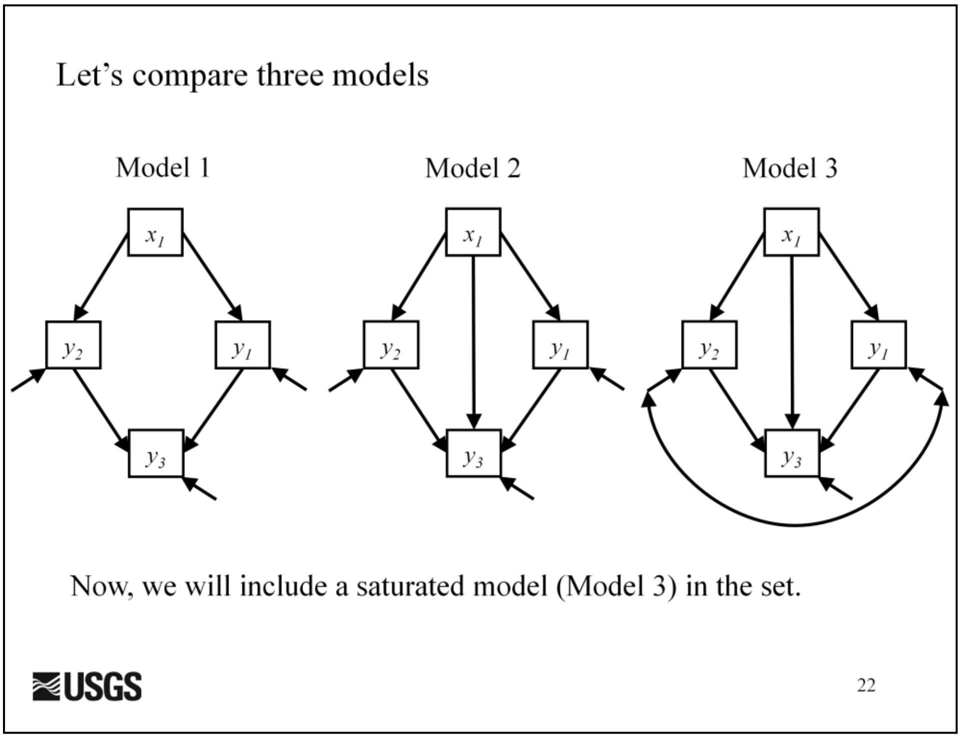
The AIC difference criteria

AIC diff	support for equivalency of models
0-2	substantial
4-7	weak
> 10	none

These apply to AICc as well.

Burnham, K.P. and Anderson, D.R. 2002. Model Selection and Multimodel Inference. Springer Verlag. (second edition), p 70.

Whether one is using the AIC, AICc, or BIC, the guidelines for simple interpretations are the same.



Let’s compare three models

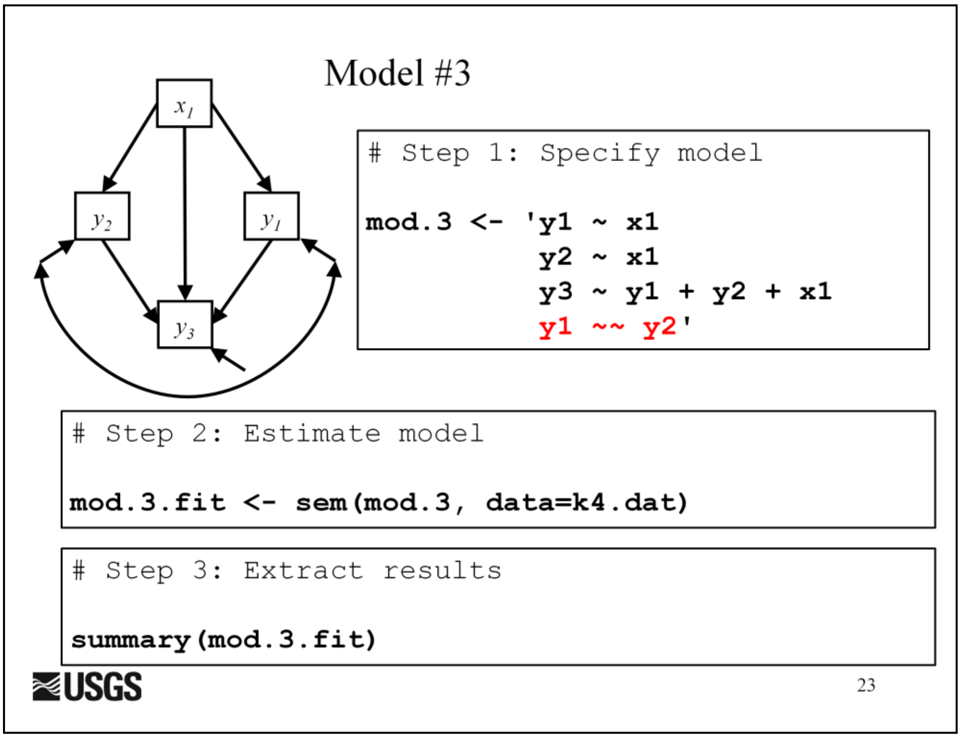
Model 1

Model 2

Model 3

Now, we will include a saturated model (Model 3) in the set.

Since the utility of AIC shines when comparing a set of models, here we throw in a third model, which allows y1 and y2 to have correlated errors.



Model #3

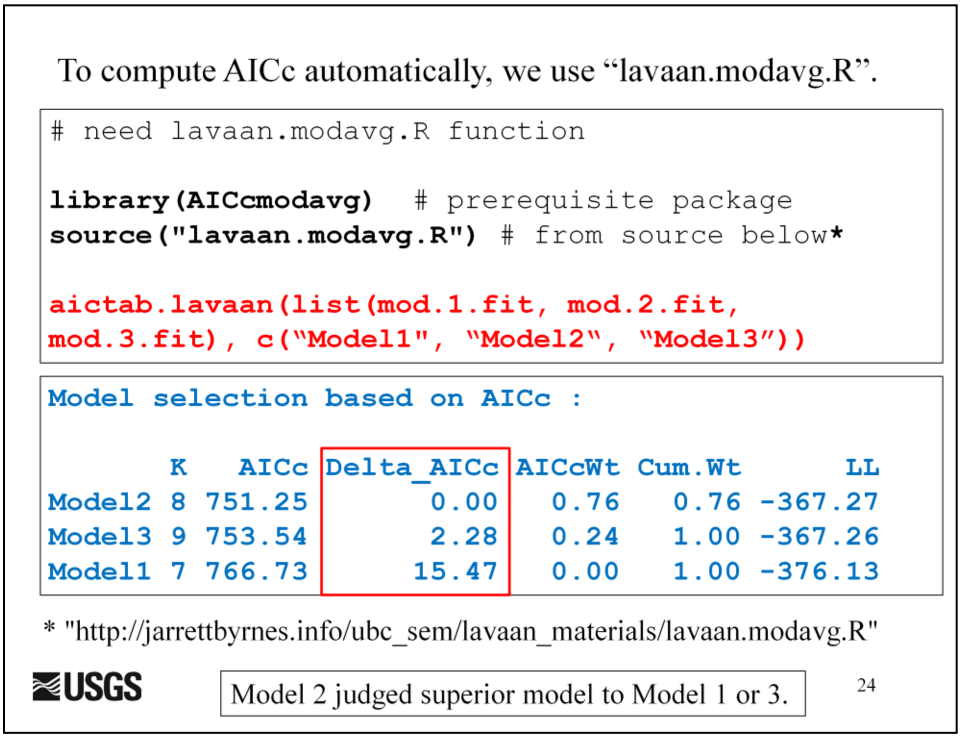
```
# Step 1: Specify model

mod.3 <- 'y1 ~ x1
          y2 ~ x1
          y3 ~ y1 + y2 + x1
          y1 ~~ y2'
```

```
# Step 2: Estimate model

mod.3.fit <- sem(mod.3, data=k4.dat)
```

```
# Step 3: Extract results

summary(mod.3.fit)
```

23

For Model 3, we add the bit in red to specify an error correlation.



To compute AICc automatically, we use “lavaan.modavg.R”.

```
# need lavaan.modavg.R function

library(AICcmodavg)   # prerequisite package
source("lavaan.modavg.R") # from source below*

aictab.lavaan(list(mod.1.fit, mod.2.fit,
mod.3.fit), c(“Model1", “Model2“, “Model3”))
```

```
Model selection based on AICc :

        K    AICc Delta_AICc AICcWt Cum.Wt      LL
Model2 8 751.25       0.00   0.76   0.76 -367.27
Model3 9 753.54       2.28   0.24   1.00 -367.26
Model1 7 766.73      15.47   0.00   1.00 -376.13
```

* "http://jarrettbyrnes.info/ubc_sem/lavaan_materials/lavaan.modavg.R"

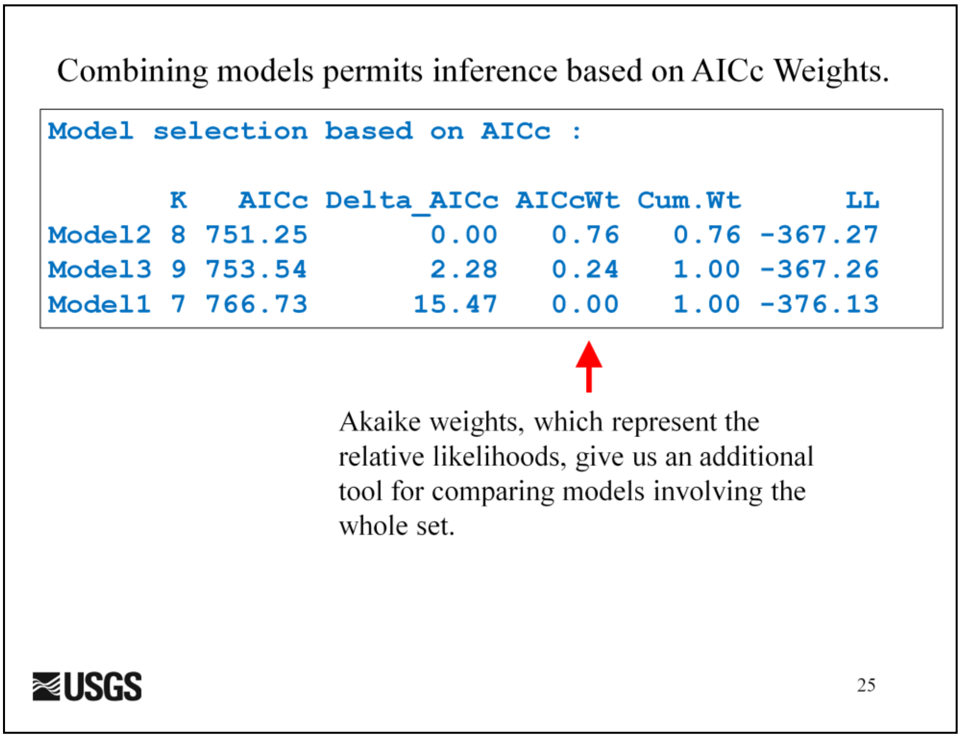
Model 2 judged superior model to Model 1 or 3.

Jarrett Byrnes from Univ. Mass at Boston has developed a function for computing AICc for lavaan models. It can be obtained from his website at:

http://jarrettbyrnes.info/ubc_sem/lavaan_materials/lavaan.modavg.R

Using this function allows us to simply compute the AICc differences for a candidate set.

The simplest way to evaluate a set of models is to rank them by AICc values and look for the smallest value. For a difference between models to be “important-ish”, the Delta-AICc should be greater then 2.0. If the difference is less than that, I might be inclined to choose the model with the fewer paths, since in a situation the evidence suggests that adding a path does not improve the model substantially.



Combining models permits inference based on AICc Weights.

```
Model selection based on AICc :

       K   AICc Delta_AICc AICcWt Cum.Wt      LL
Model2 8 751.25       0.00   0.76   0.76 -367.27
Model3 9 753.54       2.28   0.24   1.00 -367.26
Model1 7 766.73      15.47   0.00   1.00 -376.13
```

Akaike weights, which represent the relative likelihoods, give us an additional tool for comparing models involving the whole set.

There is a bit more that can be done beyond computing delta AICc. Here I show output from the command given on the previous slide. There are a few things here, including:

AICc

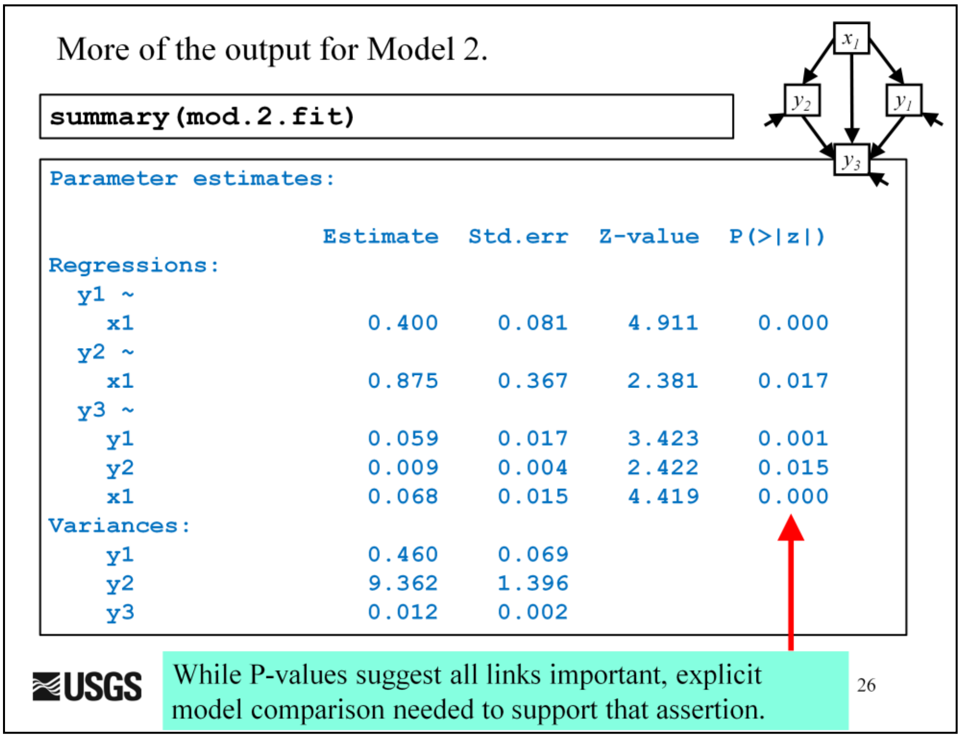
Delta_AICc

AICcWt

Other things presented in upper table include Cumulative weight and the raw log likelihoods, neither of which are important for our discussion here.

Reference for this material.

Anderson, D. (2008) Model Based Inference in the Life Sciences: A Primer on Evidence. Springer Verlag. (p 89)



More of the output for Model 2.

```
summary(mod.2.fit)
```

```
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                0.059    0.017    3.423    0.001
    y2                0.009    0.004    2.422    0.015
    x1                0.068    0.015    4.419    0.000
Variances:
    y1                0.460    0.069
    y2                9.362    1.396
    y3                0.012    0.002
```

USGS

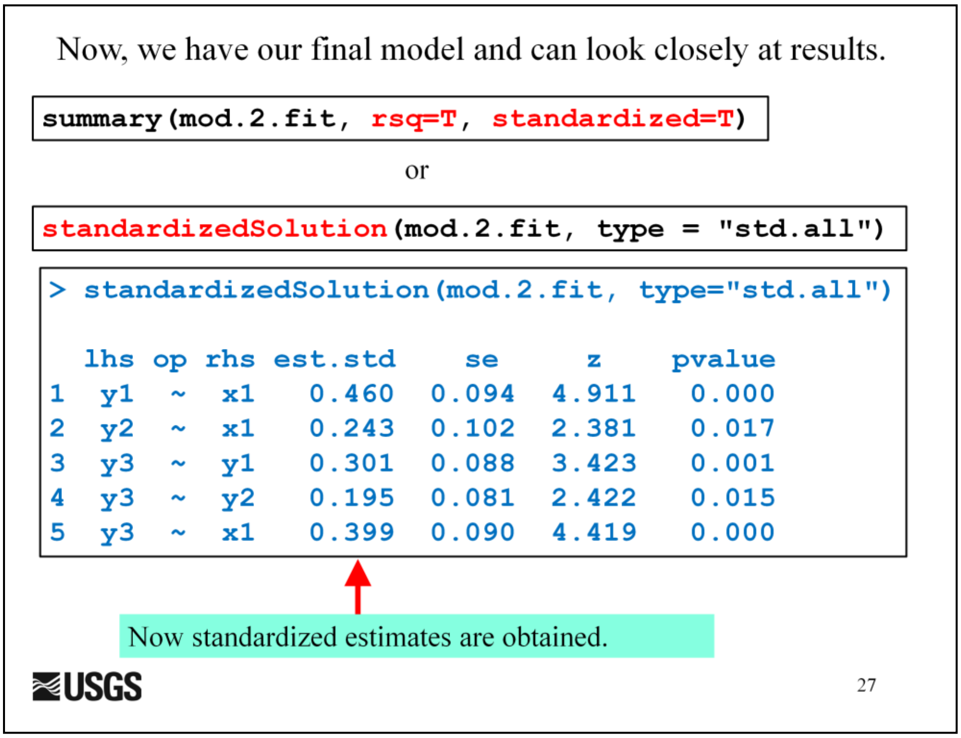
While P-values suggest all links important, explicit model comparison needed to support that assertion.

26

Here are the basic parameter results. Note you can count the 8 parameters here that are shown as K for Model 2 on the previous page.

We now can begin to interpret results. Generally not a good idea to worry about interpretation before selecting a model that (1) contains all important links and (2) is found to be the best choice amongst model options.

Glancing at the p-values for the parameters provides us with another diagnostic, much like the modification indices. WE SHOULD NOT TRUST THE PARAMETER P-VALUES AS A SOLE SOURCE OF INFORMATION. To elaborate, the overall model has a behavior that we are evaluating. It is not uncommon for parameter p-values to be greater than 0.05, but then when you remove the associated link from the model, overall discrepancy goes up significantly. TO REPEAT, PARAMETER P-VALUES ARE ONLY DIAGNOSTICS, NOT DEFINITIVE SOURCES FOR DECISIONS.



Now, we have our final model and can look closely at results.

```
summary(mod.2.fit, rsq=T, standardized=T)
```

or

```
standardizedSolution(mod.2.fit, type = "std.all")
```

```
> standardizedSolution(mod.2.fit, type="std.all")

  lhs op rhs est.std    se     z   pvalue
1  y1  ~  x1   0.460 0.094 4.911   0.000
2  y2  ~  x1   0.243 0.102 2.381   0.017
3  y3  ~  y1   0.301 0.088 3.423   0.001
4  y3  ~  y2   0.195 0.081 2.422   0.015
5  y3  ~  x1   0.399 0.090 4.419   0.000
```

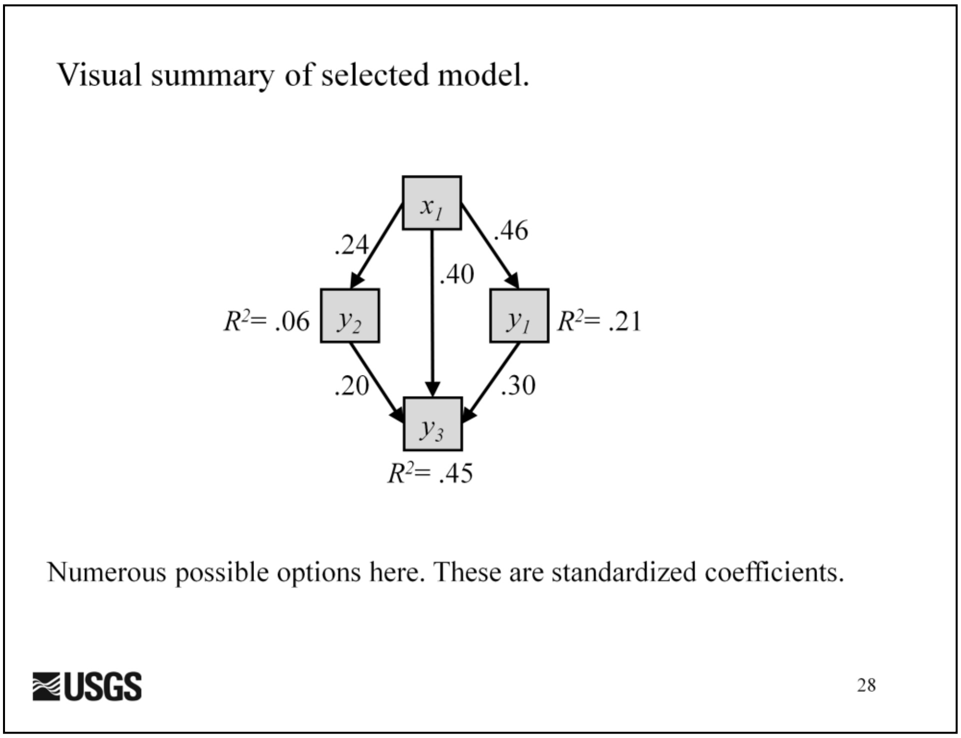
Now standardized estimates are obtained.

Now we are ready for the scientific interpretation of quantitative parameter estimates. In this case, we wish to look at the relative magnitudes of the standardized parameters.

Another module discusses the interpretability of standardized solutions and so you should know that material before going too far into interpretations. Here I adopt a simple approach using classic standardized parameters.

There are a couple of different ways we can ask for standardized parameters.

Note that lavaan computes “std.lvs” which you want to ignore in favor of “std.all”.



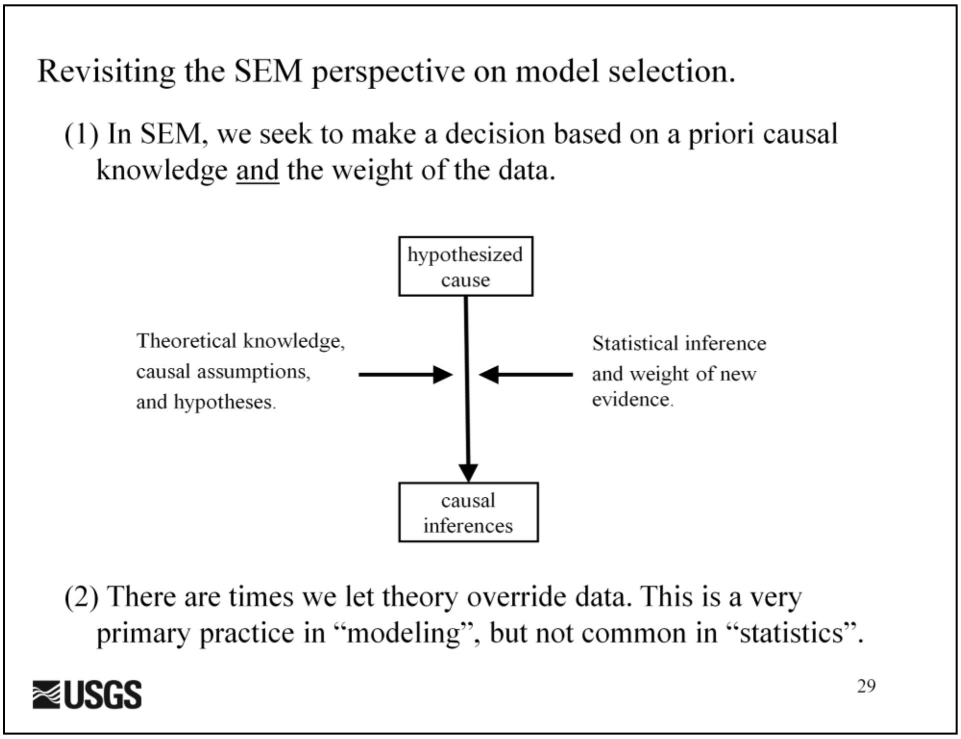
Visual summary of selected model.

Numerous possible options here. These are standardized coefficients.

Here is one of the many ways you might present your results, along with various tables of total effects in the models or scenario findings.

You should also present model fit results.

A separate module will be developed soon for a more complete discussion of results presentation. For now, consulting some of the published papers for examples would be a good idea.



Revisiting the SEM perspective on model selection.

(1) In SEM, we seek to make a decision based on a priori causal knowledge and the weight of the data.

hypothesized cause

Theoretical knowledge, causal assumptions, and hypotheses.

Statistical inference and weight of new evidence.

causal inferences

(2) There are times we let theory override data. This is a very primary practice in "modeling", but not common in "statistics".

29

OK, after considering all this quantitative weighing of evidence in your data set, I want to again emphasize you still have to consider many things in decide what you believe about the system under investigation from all your information sources. The next slide gives one example of the latitude I suggest you give yourself.

Sometimes a model retains a link even though the parameter is not consistently different from zero.

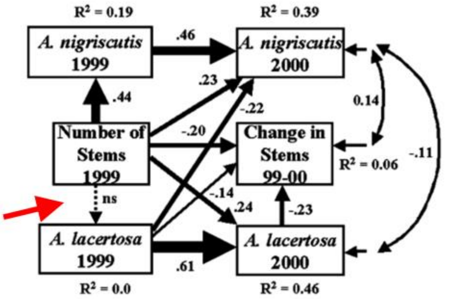
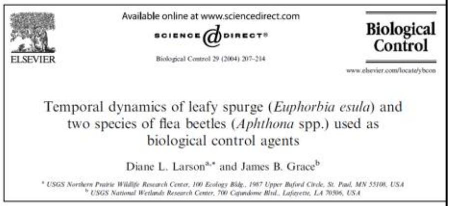


Fig. 2. Model results for 1999–2000. Single-pointed arrows indicate causal paths. Double-pointed arrows represent correlations, which were modeled as correlated errors. Path coefficients are standardized and all solid arrows indicate significant paths at $P < 0.05$; size of the arrow correlates with the magnitude of the path coefficient. The dashed arrow indicates a nonsignificant path that remains in the model. R^2 values are shown for dependent variables.



Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Biological Control 29 (2004) 207–214

ELSEVIER

Biological Control

www.elsevier.com/locate/ybcon

Temporal dynamics of leafy spurge (*Euphorbia esula*) and two species of flea beetles (*Aphthona* spp.) used as biological control agents

Diane L. Larson[a,*] and James B. Grace[b]

[a] USGS Northern Prairie Wildlife Research Center, 100 Ecology Bldg., 1987 Upper Buford Circle, St. Paul, MN 55108, USA
[b] USGS National Wetlands Research Center, 700 Cajundome Blvd., Lafayette, LA 70506, USA

USGS

Here is one of our examples where a “non-significant” path represents the dependence of a obligate biocontrol agent on its only food source. Leaving the path in the model, versus taking it out, has very small effects on other parameters. Scientifically, though, it would be a radical claim for either the causal processes operating in this system or its future behavior to claim no causal effect of the food on the herbivore. So, the path was left in the model, but noted as non-significant in conventional tests.

Later studies on this system validated the approach we used in this case.