Kline's Rules for How to Fool Yourself with SEM

Don’t Have a Plan

(ok, this one is mine)

Amelia Hoover I CAN HAS ESTIMITS!!!
Thurs at 13:02 · Comment · Unlike

You and 2 others like this.

Rob Person BUT IZ THEY CANS BE KONSISTENT???
Thurs at 13:50

Amelia Hoover KONFUDUNS INNRVAL KONSISTENTLY INKLOOD ZERO. DO OVERS!!!
Thurs at 13:52

David Mister ..GEEKcats?
Thurs at 14:00

David January All I can say is "wow".
Thurs at 14:55

Amelia Hoover update: I CAN HAS INTERACKSHIN TURM!
Thurs at 17:35

Emily Clough BAYESMENT CAT ATTACK!
Thurs at 22:23

Amelia Hoover oh snap.
Thurs at 22:26

Branen Salmon This is the most amazing thing I have ever witnessed.
Fri at 00:16

Amelia Hoover oh, this is nothing. a couple of years ago I wrote a whole program in LOLpython.
Fri at 02:03

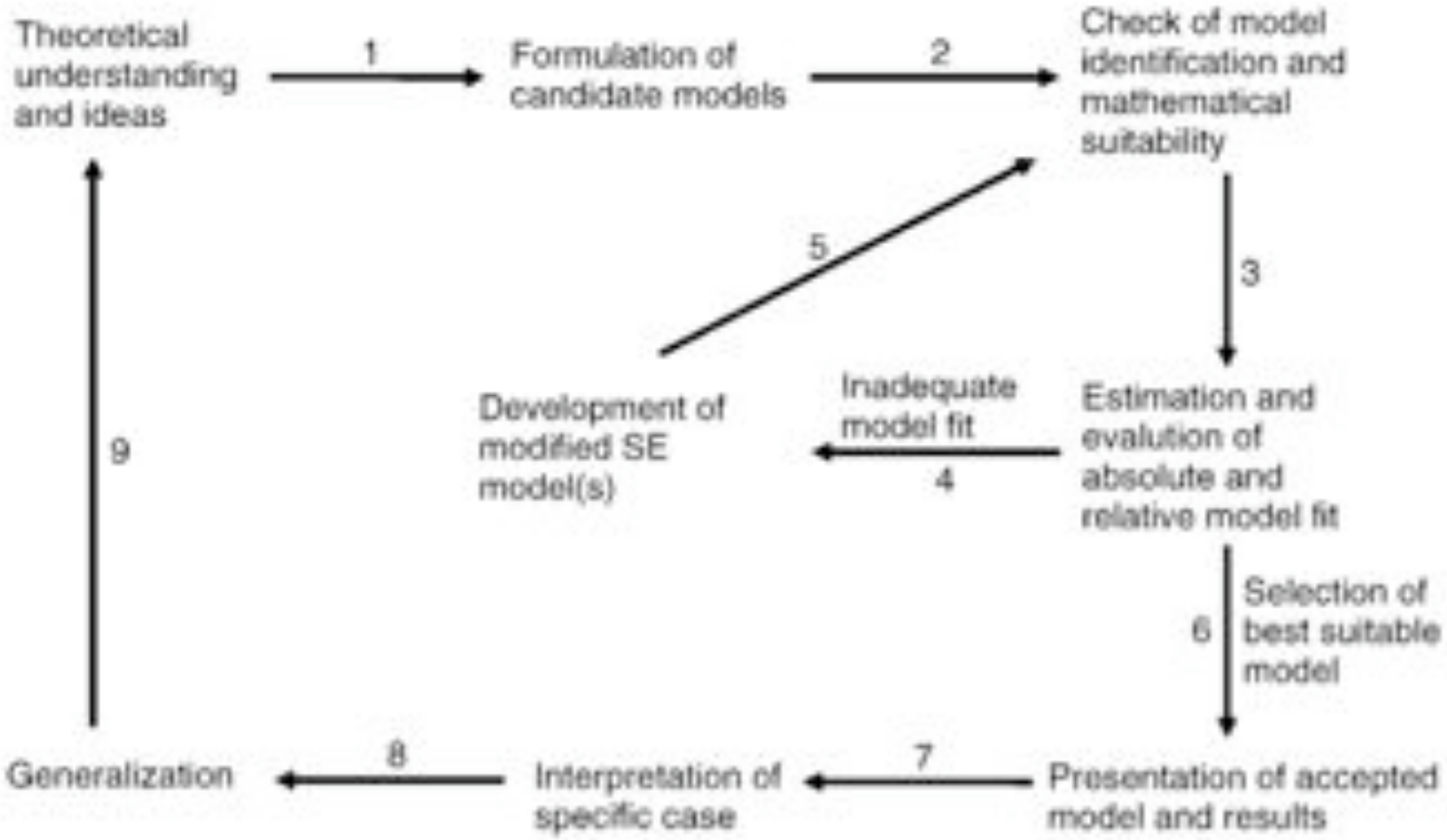
Rob Person OOPS I HAZ MADE P-VALYOO ON TEH KARPITZ.
Fri at 02:23

Jarrett Byrnes IM IN UR AICZ, MAXIMIZIN MAH LIKLIHOODZ!
Fri at 11:34 · Delete

Amelia Hoover i will cherish this comment thread forever.
Fri at 12:16

Don’t Have a Plan

(ok, this one is mine)



Grace et al 2010 Ecological Monographs

Tripping at the Starting Line: Model Specification

Specify the model after the data are collected

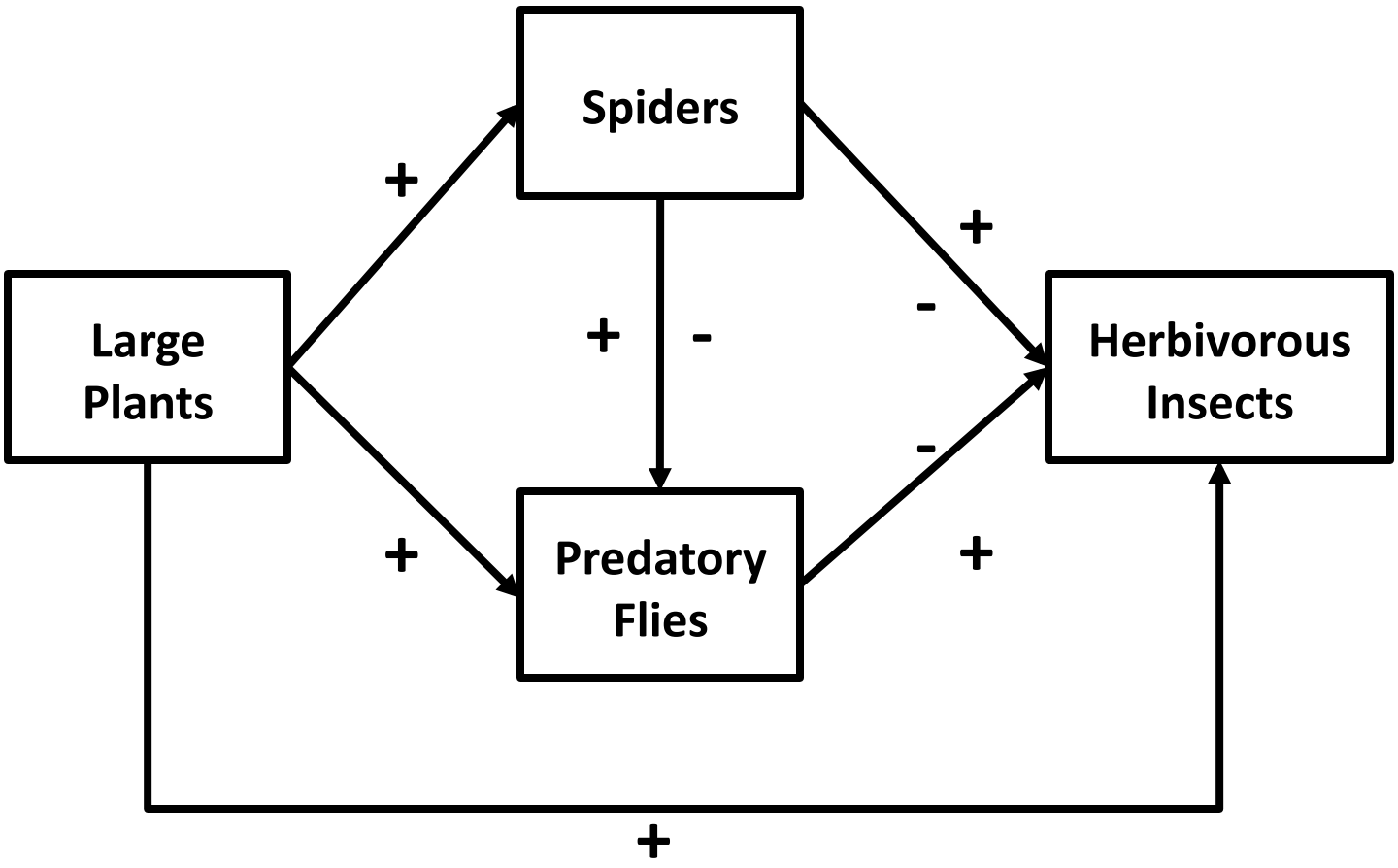
You must ask:

Will the data provided by adequate?

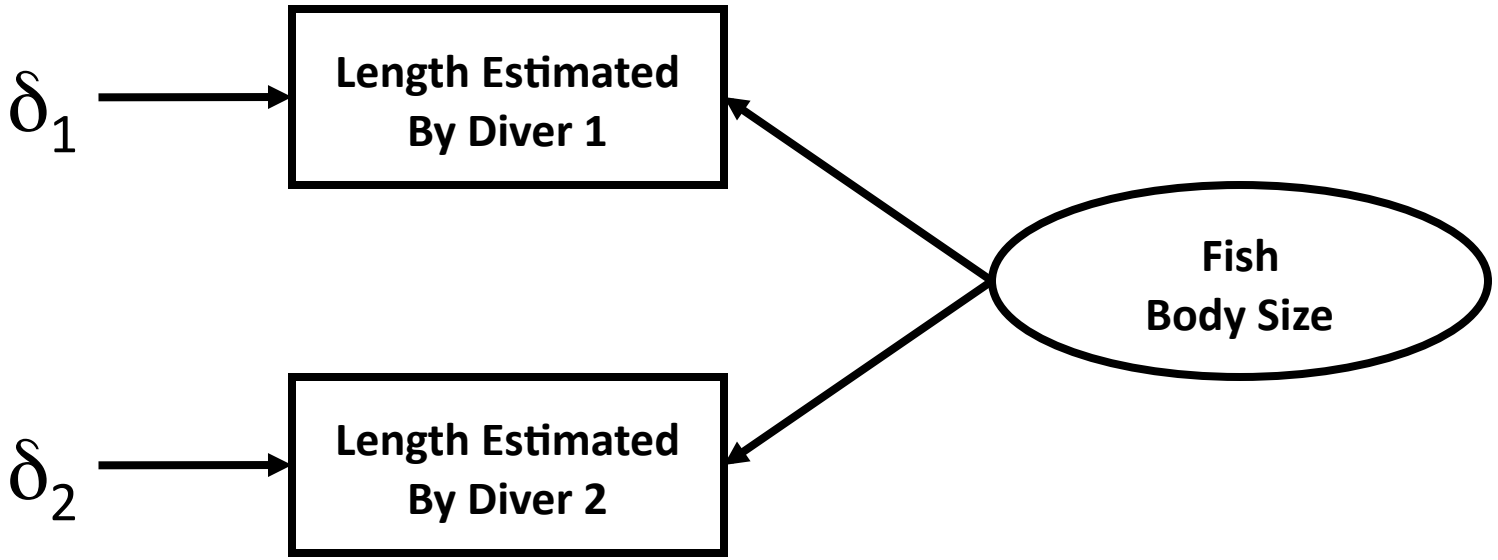
What did they miss?

Can it be modeled?

Omit causes that are correlated with variables in the structural model

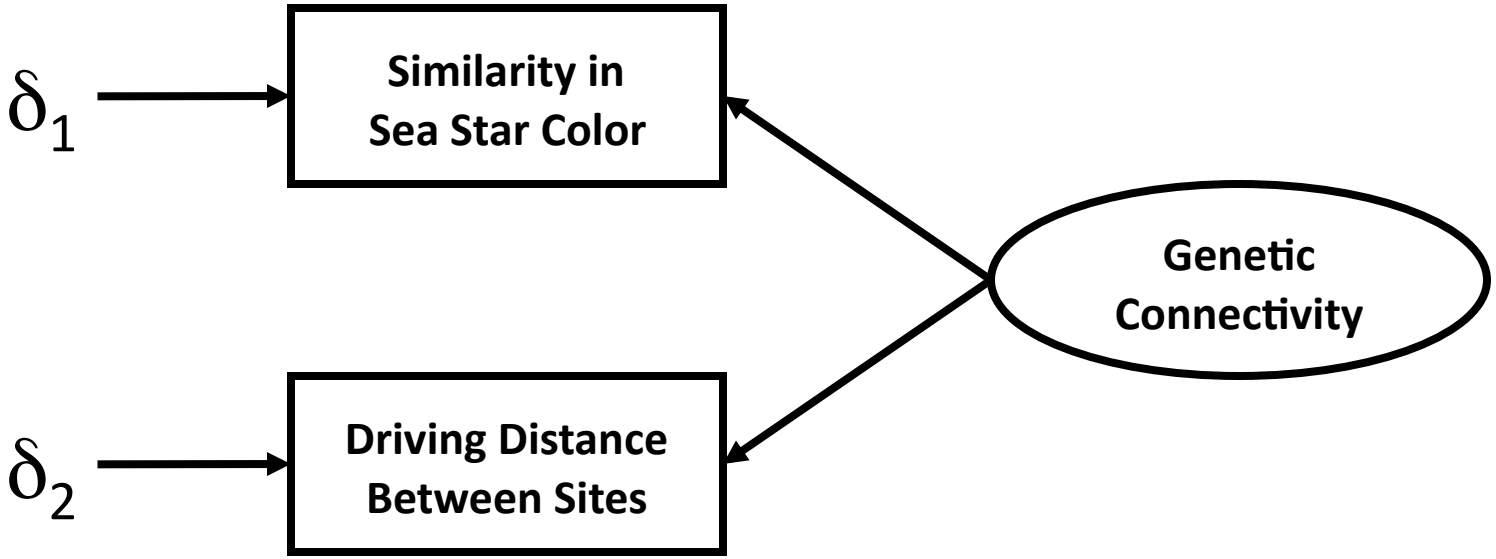


Fail to have a sufficient number of indicators for latent variables



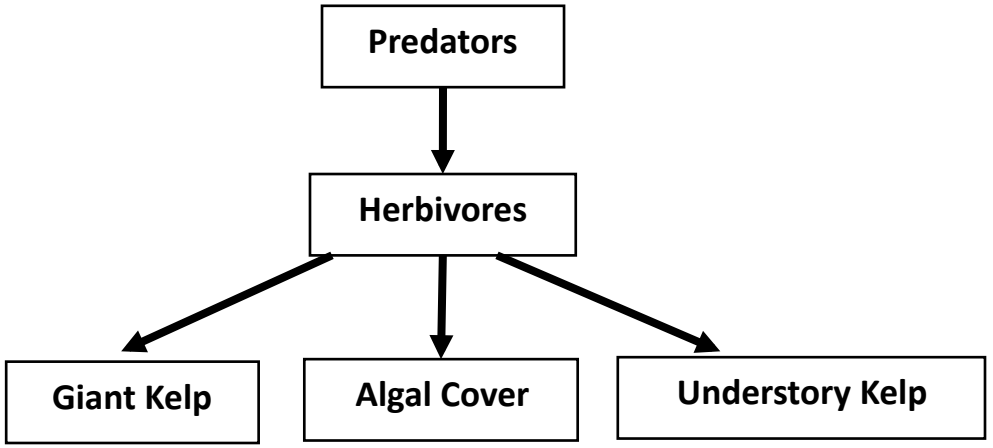
What if these estimates are wildly divergent? Solution becomes unstable.

Use indicator variables that do not match concepts contained in latent variables

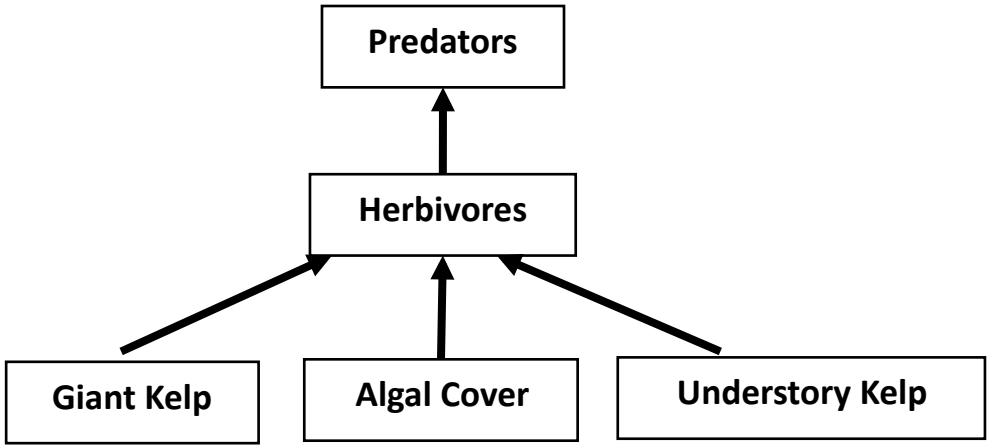


Fail to give careful consideration to directionality

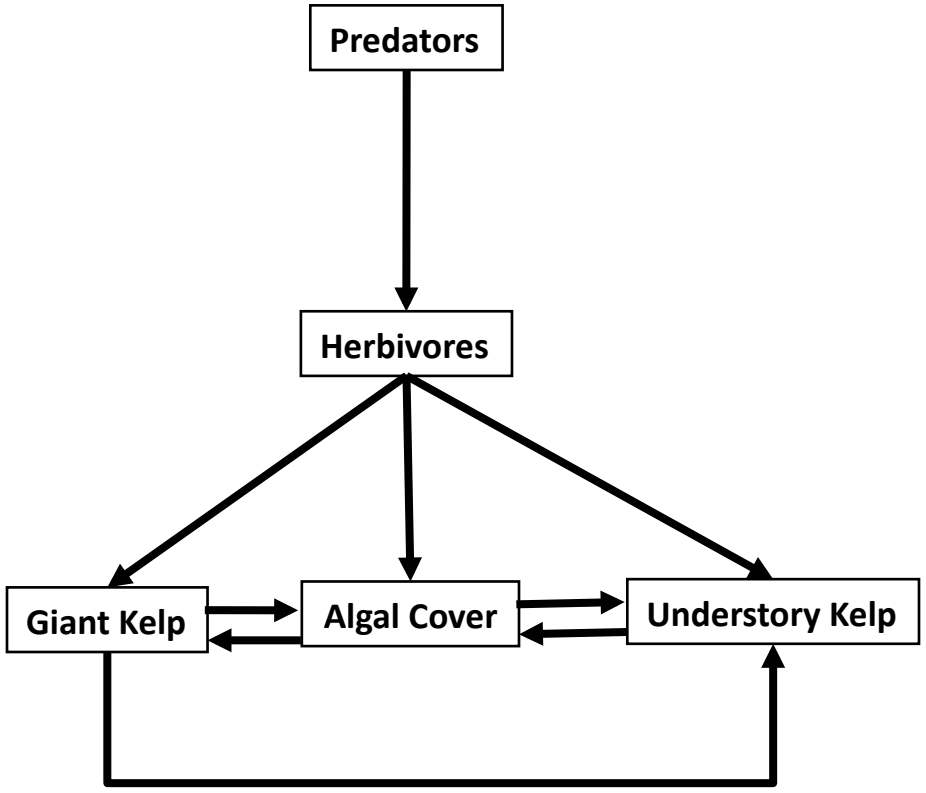
- These models describe completely different phenomena

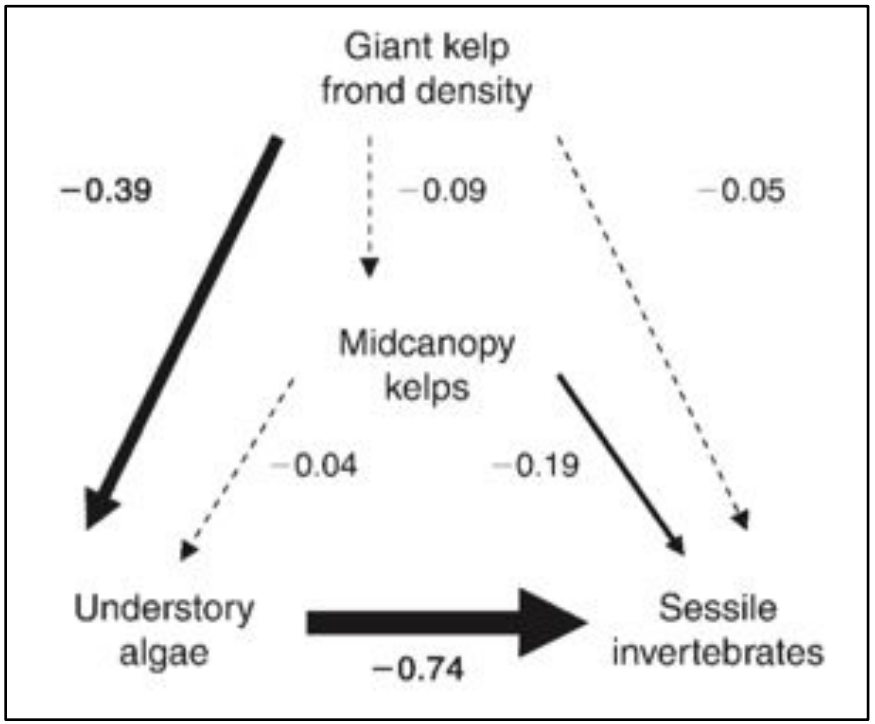


- If these models have the same coefficient estimates, what does that mean?



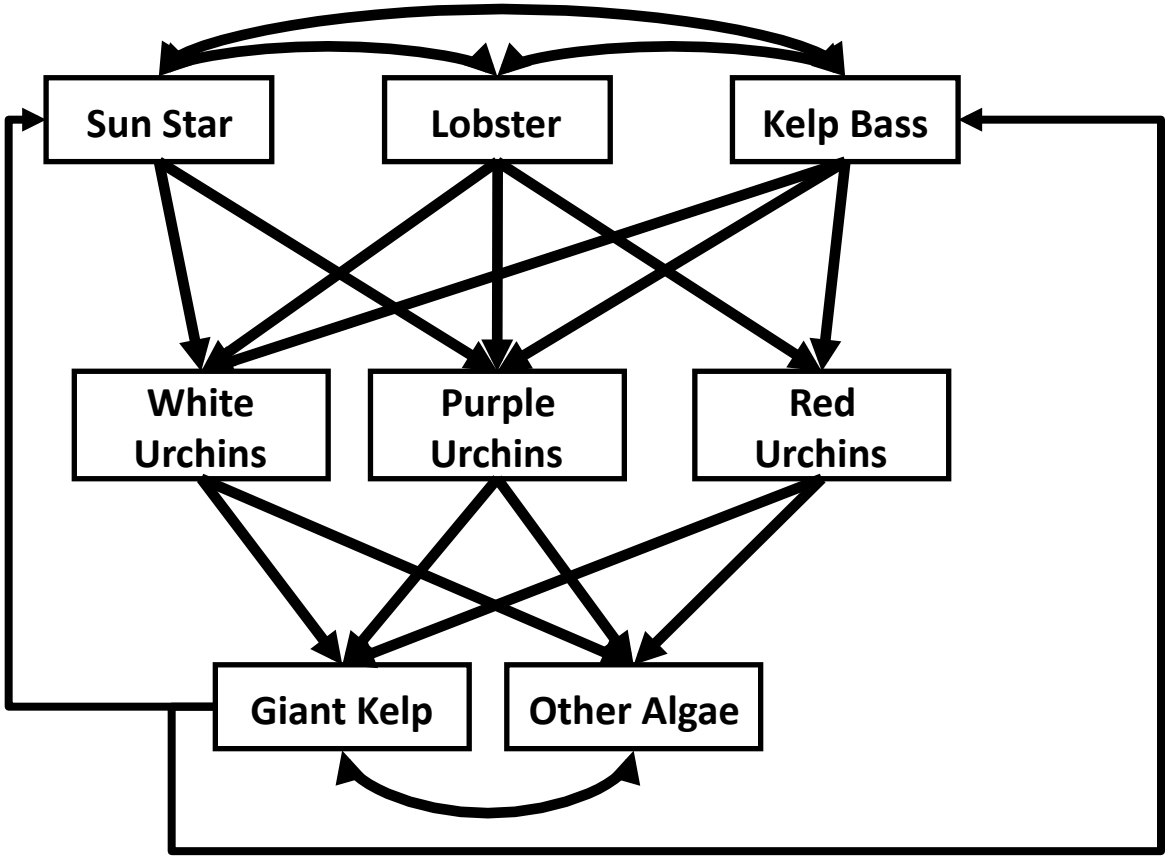
Specify feedbacks as a way to mask uncertainty about directionality





Arkema et al. 2010

Forget the Goal of Parsimony



What will we learn from this model?

How is it being a multivariate model useful?

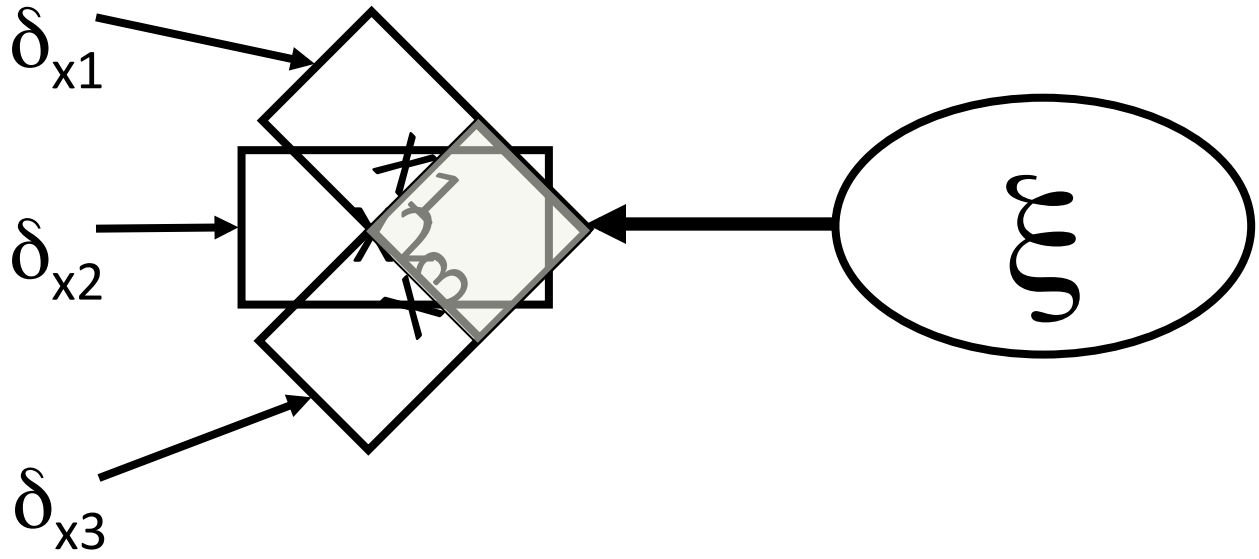
Would a simpler model better represent processes we can detect given our data?

Will including all of these paths lead to excessive parameter uncertainty?

Add unexplained covariances without substantive reasoning

- Yes, everything in the world is correlated.
- You will miss some substantive correlations in any model you create.
- But…if you have no reason to include correlated errors, why should you?

Specify that indicators load on more than one latent variable without a substantive reason



If one of these variables is an indicator of another latent variable, it must be from the non-shared variation. Otherwise, what does this variable mean? Solution potentially unstable.

Improper Care and Feeding: Data

Don’t check the accuracy of data inputs or coding

YEAR	SITE	TRANSECT	QUAD	FRONDS	HLD_DIAM
2000	BULL	1	20	2	-99999
2000	BULL	1	20	4	7
2000	BULL	1	20	2	-99999
2000	BULL	1	20	2	-99999
2000	BULL	1	20	2	-99999
2000	BULL	1	20	2	-99999
2000	BULL	1	40	2	-99999

- -99999 was the code for missing data.
- I had no clue.
- My models gave some ***very*** strange estimates.

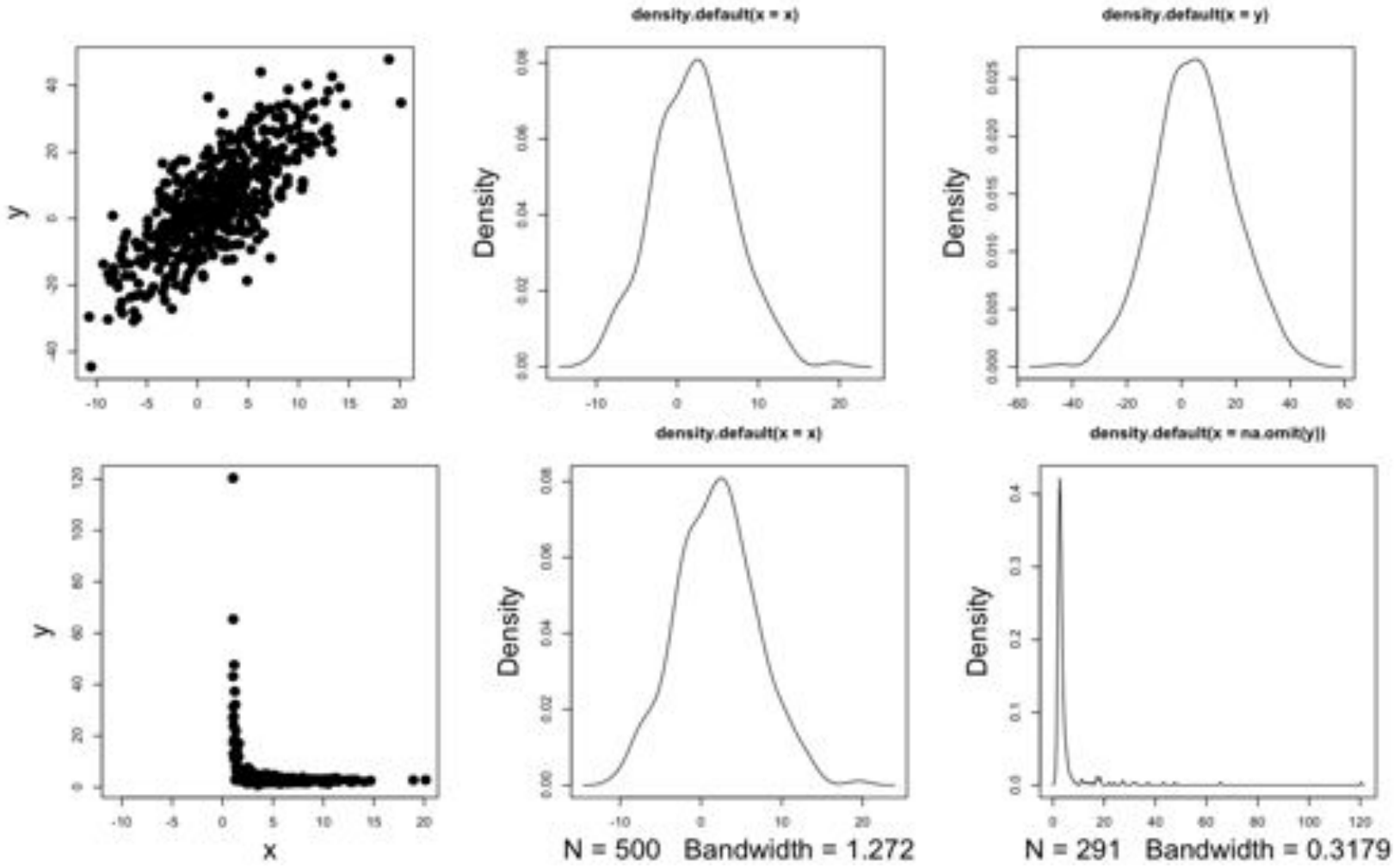
Ignore whether the pattern of missing data loss is random or systematic

YEAR	SITE	MAX_WAVE_HEIGHT_M
1996	BULL	1.452661106
1997	BULL	NA
1998	BULL	2.614931561
1999	BULL	2.505335617
2000	BULL	1.498685958
1996	NAPL	1.461412561
1997	NAPL	NA
1998	NAPL	2.660347776
1998	NAPL	2.540209549
2000	NAPL	3.940096761
1996	MOHK	1.383051685
1997	MOHK	NA
1998	MOHK	2.145915443
1998	MOHK	1.778718707
2000	MOHK	1.738998659

YEAR	SITE	MAX_WAVE_HEIGHT_M
1996	BULL	1.497169396
1997	BULL	NA
1998	BULL	2.289101797
1999	BULL	2.396824321
2000	BULL	3.003874137
1996	NAPL	NA
1997	NAPL	3.815917303
1998	NAPL	3.158332091
1998	NAPL	1.138132847
2000	NAPL	3.990087331
1996	MOHK	NA
1997	MOHK	1.017746133
1998	MOHK	2.337630923
1998	MOHK	NA
2000	MOHK	1.592523562

- There is real information loss when the data is missing systematically.
- Estimates will not encompass true range of variation.
- Missing data imputation is a **big** field.

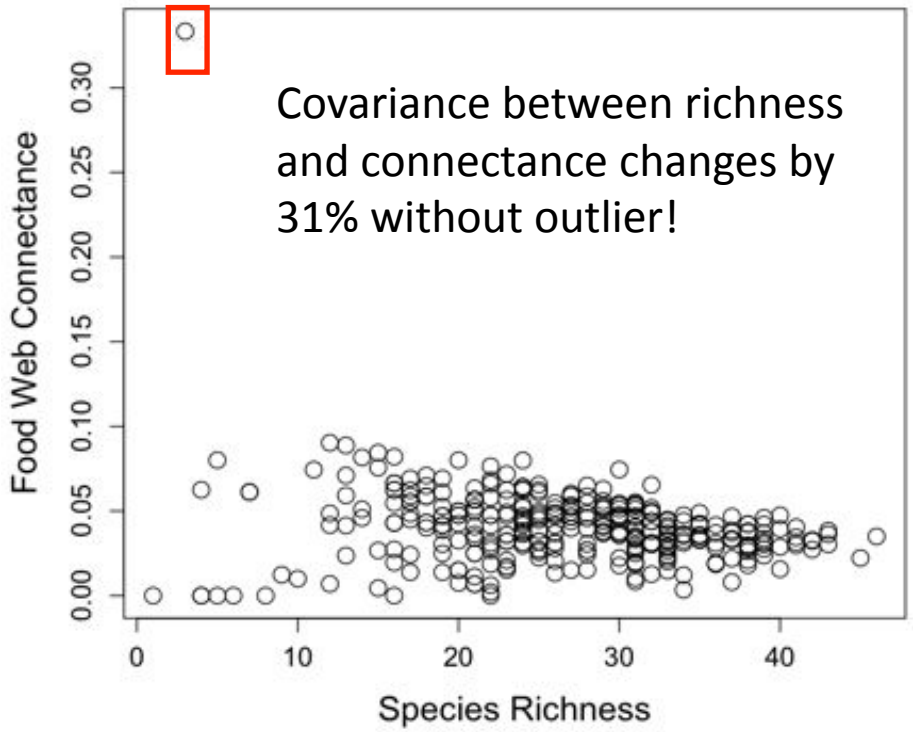
Failing to examine data for multivariate normality



χ2 for TopDown+BottomUp model=58.11, p<0.0001

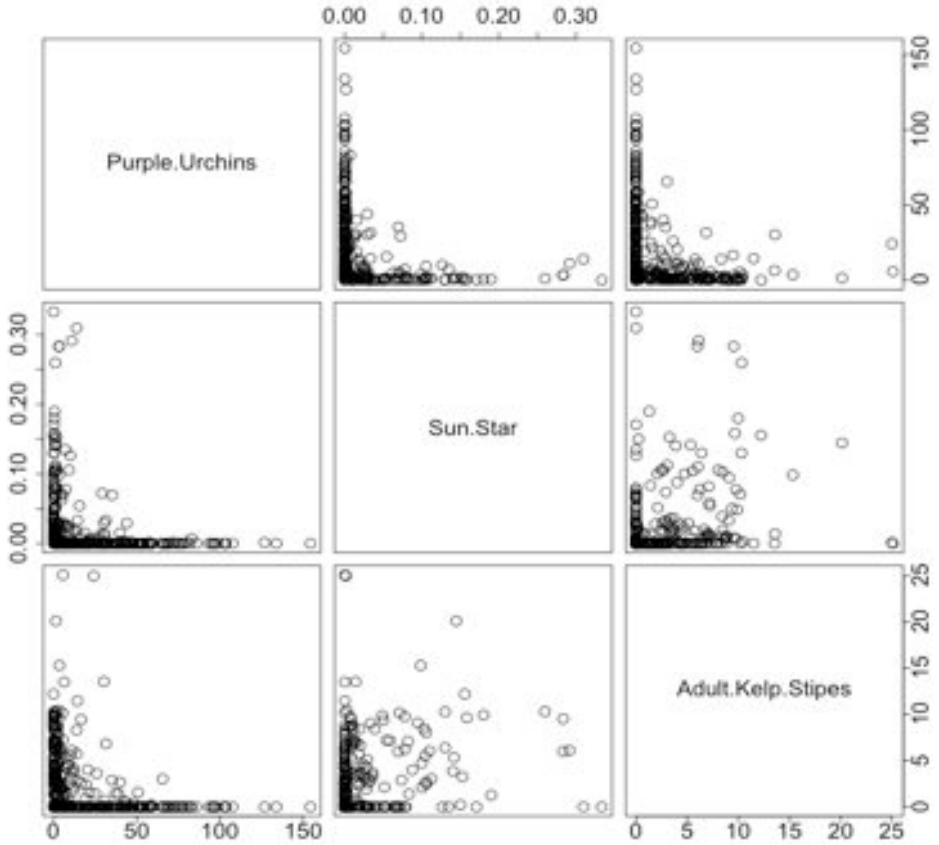
χ2 for TopDown+BottomUp model after SB correction=25.49, p=0.18

Don’t screen for outliers



- Can greatly influence variances and covariances – influences paths in which variable with outlier is not directly involved.
- But…there’s information in them thar’ outliers!

Assume that all relations are linear without checking

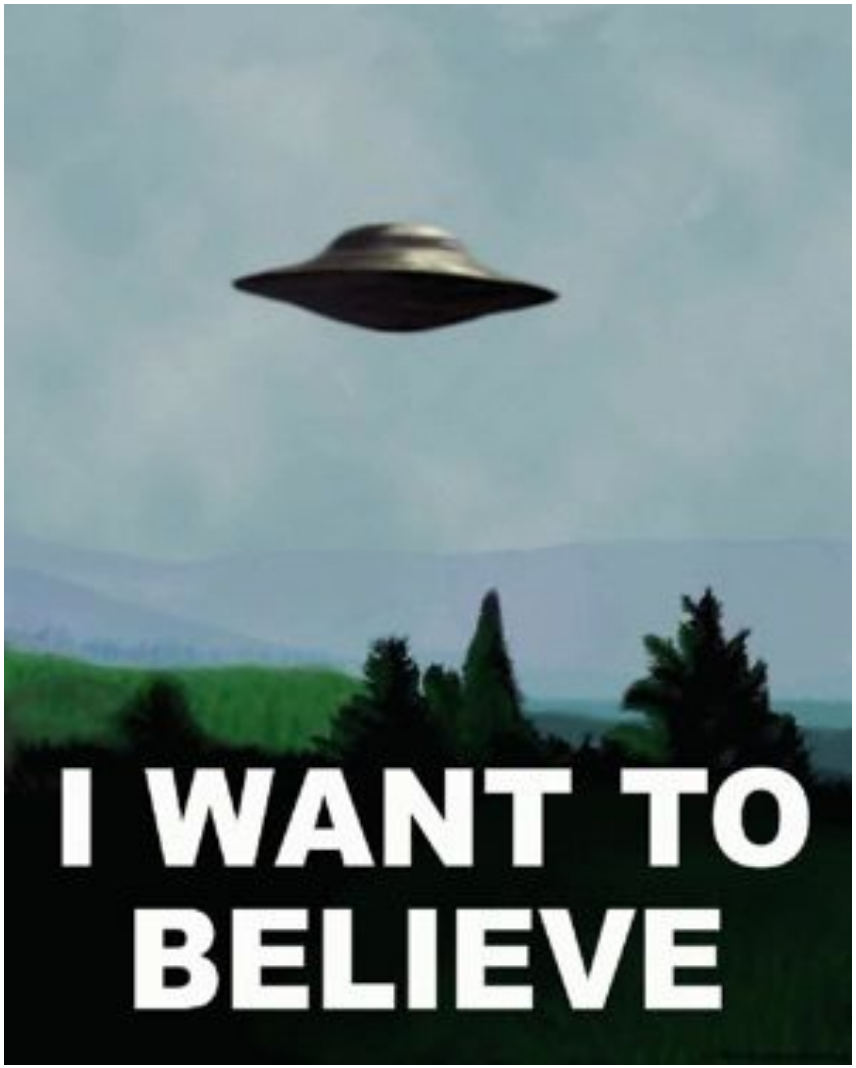


- If do not want to transform your data (and there are some very good reasons for this), use piecewise approaches – maybe Bayes!

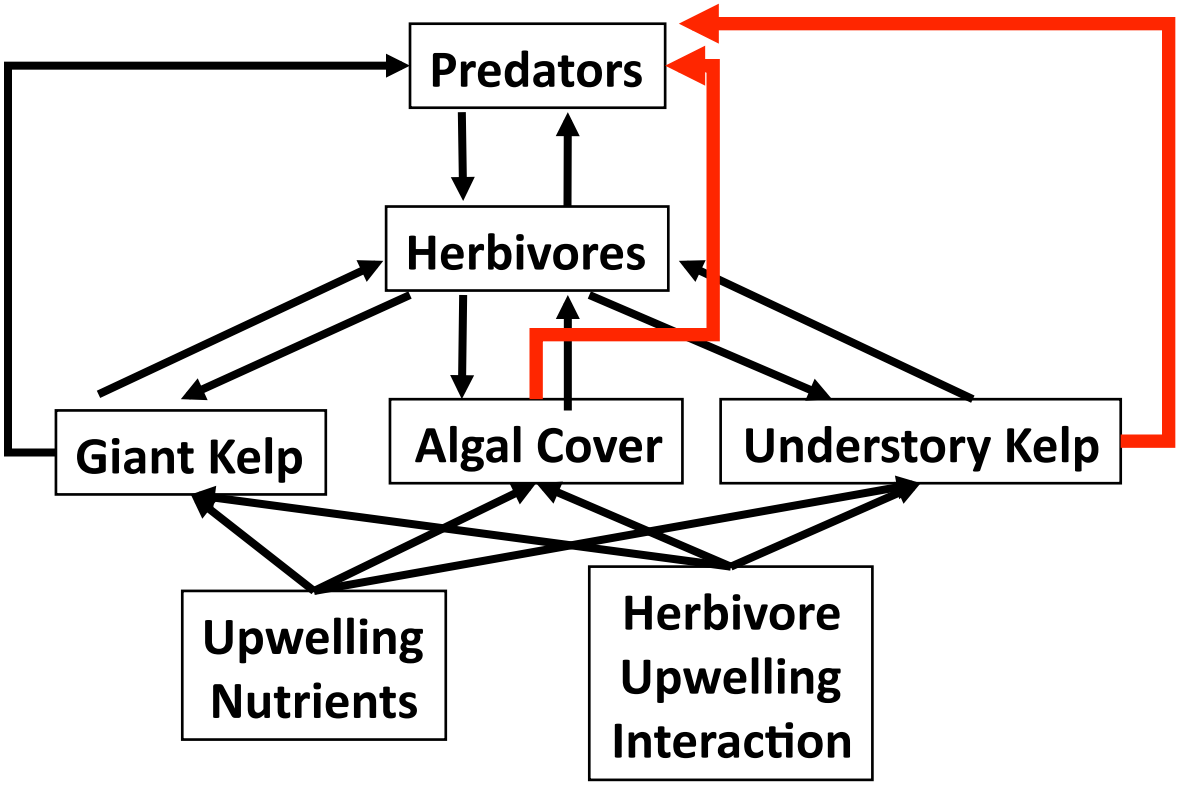
Ignore the lack of independence between observations

Solutions: Autoregressive models, Incorporate hierarchies explicitly, Multilevel SEM, Moran's I

Checking Critical Thinking at the Door: Analysis and Respecification



Respecify a model based entirely on statistical criteria



Lagrange multipliers can be dangerous things…

Fail to check the accuracy of your computer code

```
pred.model<-specify.model()
Sun.Stars -> Purple.Urchins, star.purp, NA
Kelp.Bass -> Purple.Urchins, kelpbass.purp, NA
Lobster -> Purple.Urchins, lobster.purp, NA
Purple.Urchins -> Adult.Kelp.Stipes, purp.kelp, NA
Purple.Urchins <-> Purple.Urchins, urchin.var, NA
Adult.Kelp.Stipes<->Adult.Kelp.Stipes, kelp.var, NA
Rock.Percent -> Adult.Kelp.Stipes, rock.kelp, NA
Rock.Percent -> Purple.Urchins, rock.purp, NA
```

```
> pred.fit<-sem(pred.model, pred.cov, N=length(kfm[,1]), fixed.x=c("Kelp.Bass", "Lobster",
"Sun.Star", "Rock.Percent"), debug=T)

 observed variables:
[1] "1:Purple.Urchins"    "2:Kelp.Bass"         "3:Lobster"           "4:Adult.Kelp.Stipes"
"5:Rock.Percent"

 latent variables:
[1] "6:Sun.Stars"
```

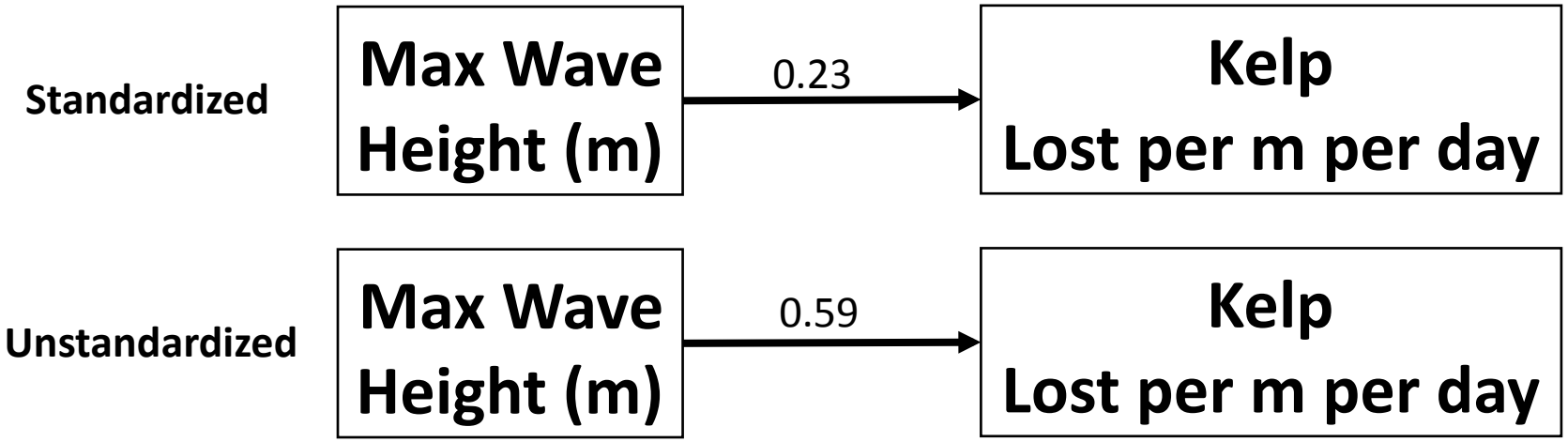
Fail to carefully inspect the solution for problems

1. Do you have any Heywood cases (estimates of variance <0)?

2. How well explained are your endogenous variables?

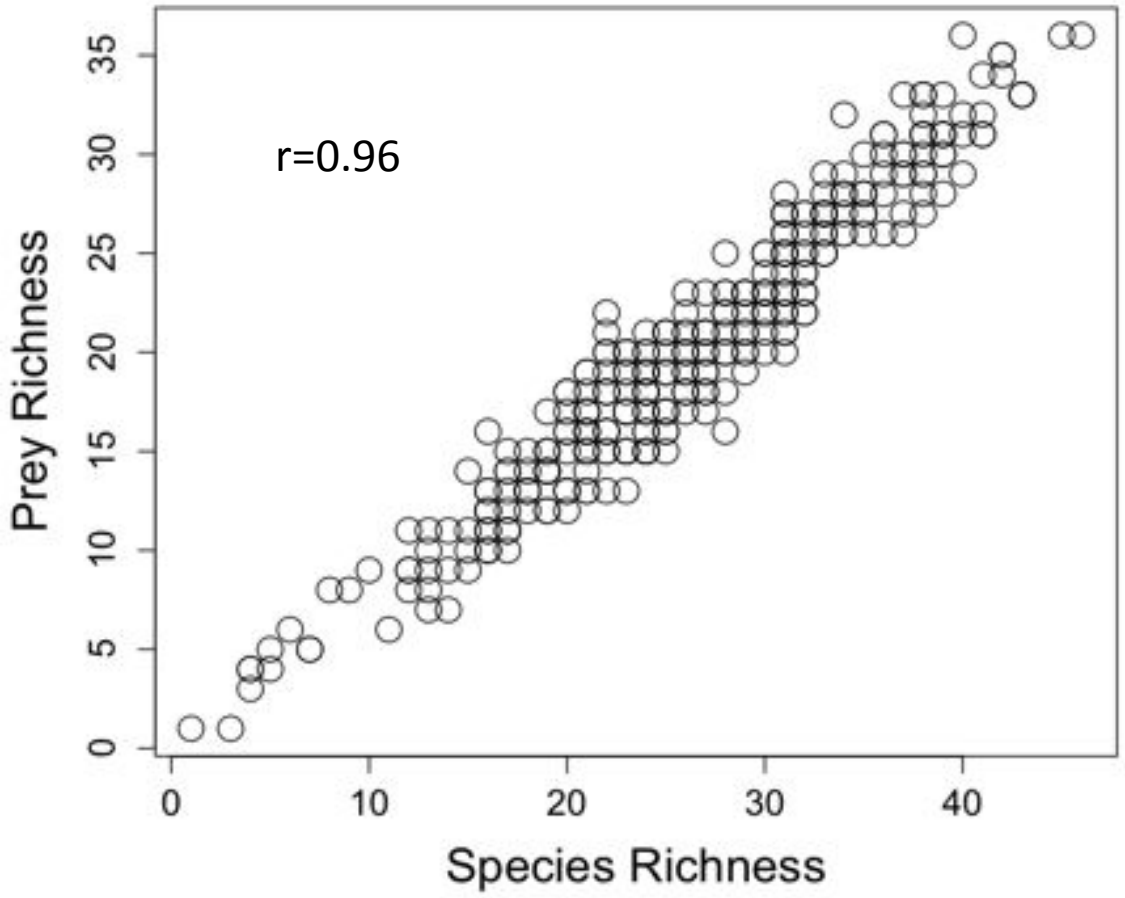
3. Any “surprises”? And not the kind that leads to birthday cake.

Report only standardized estimates



- $r_{xy}=B_{xy}*sd(x)/sd(y)$ -> what is the range of variability with your x and y? (see relative range standardizations)
- Standardized coefficients useful for comparisons within a model.
- If you are using your model for predictive purposes, the unstandardized estimate is more useful.

Analyze variables so highly correlated the solution is unstable

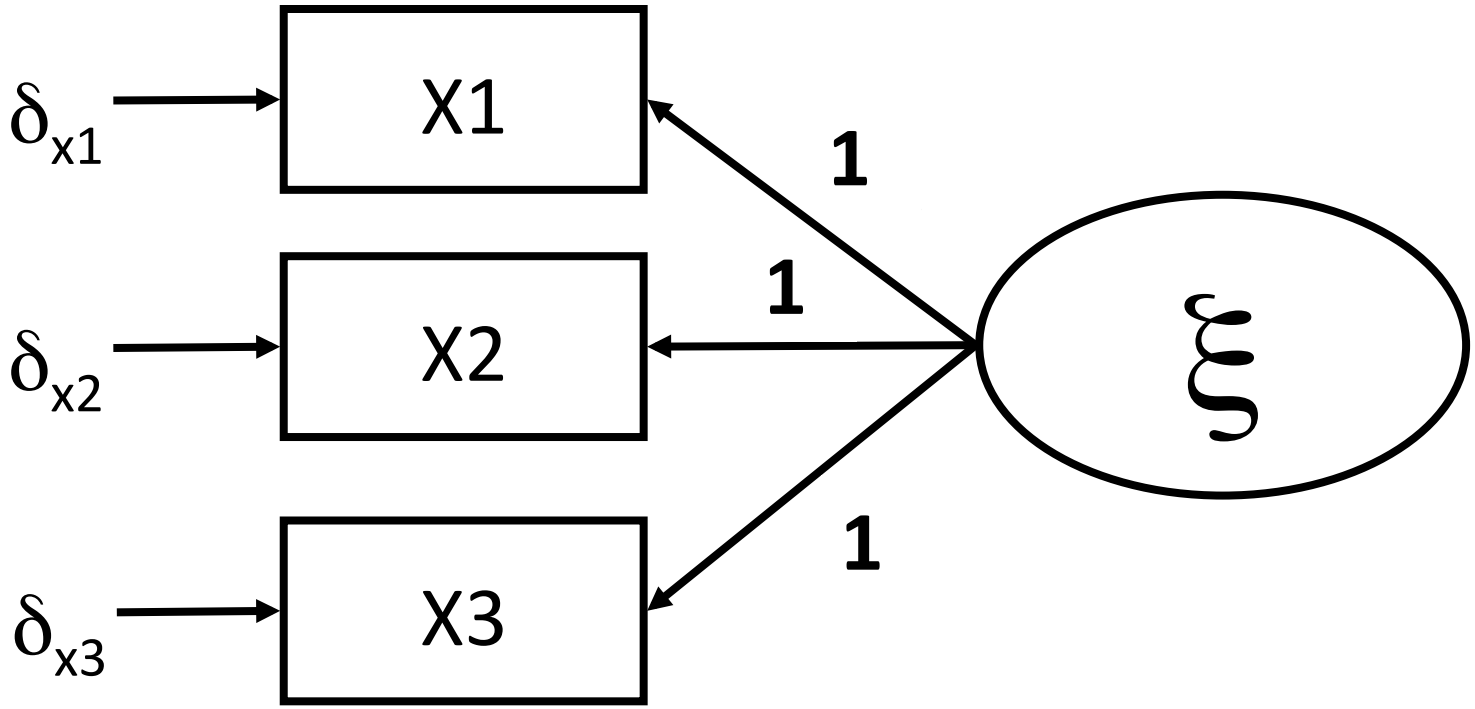


Neither variable provides unique information.

Estimate a complex model with a small sample size

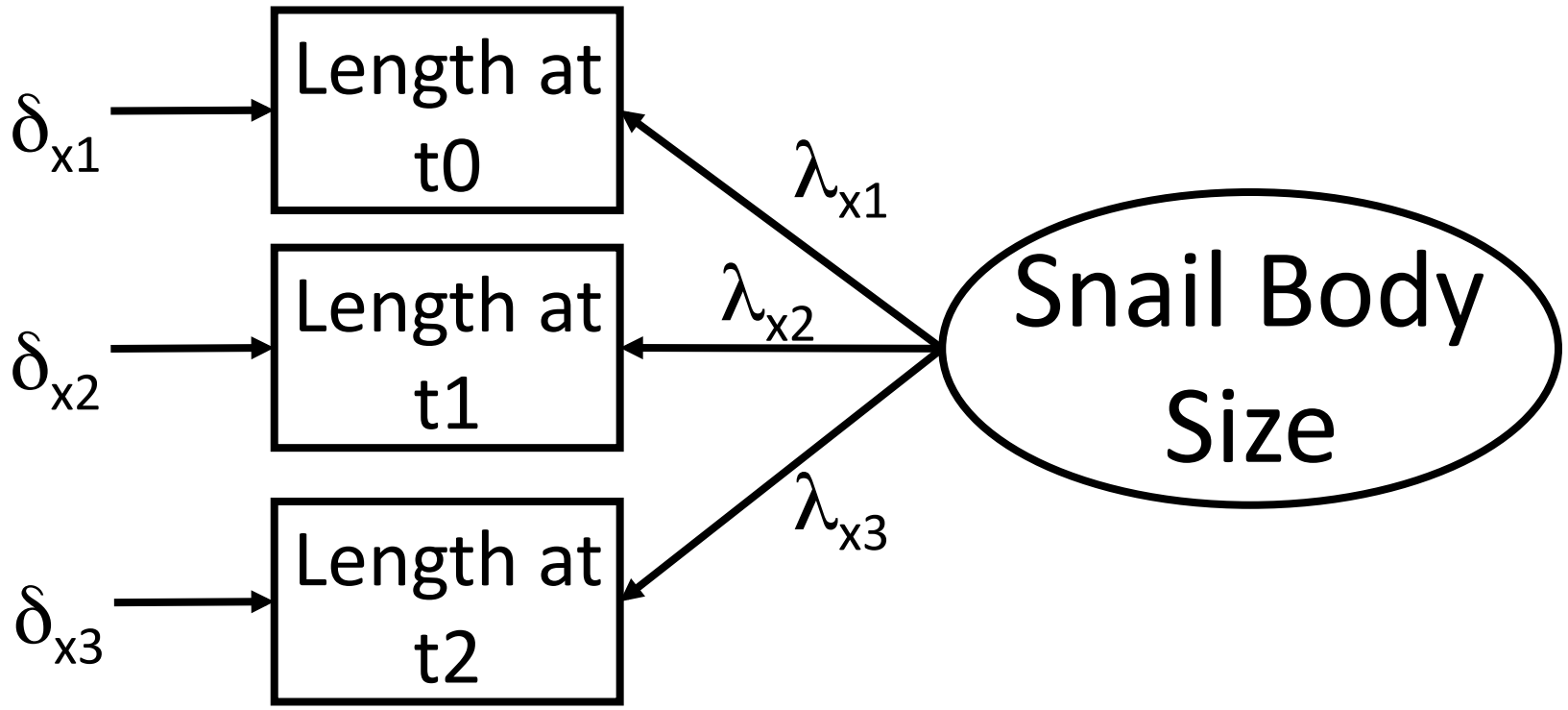
- N=5-10 per parameter
- Aids in stable solutions
- Influences significance testing of individual paths
  - But beware the fallacy of too large of a sample size = highly significant correlations
- This is even more of a problem for fitting criteria other than F_{ML}

Set scales for latent variables inappropriately



Setting a scale determines the definition of your latent variable.

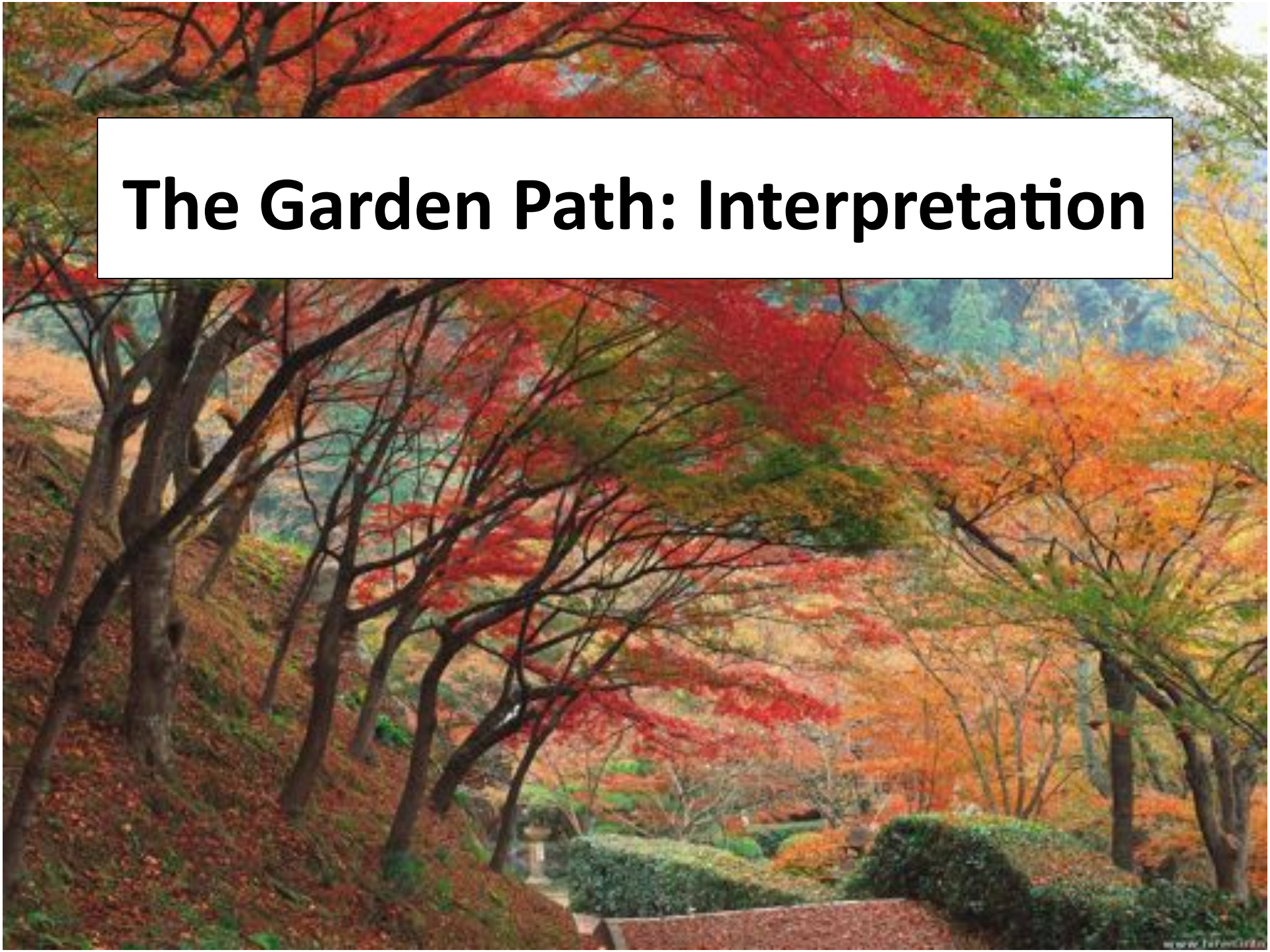
Set scales for latent variables inappropriately



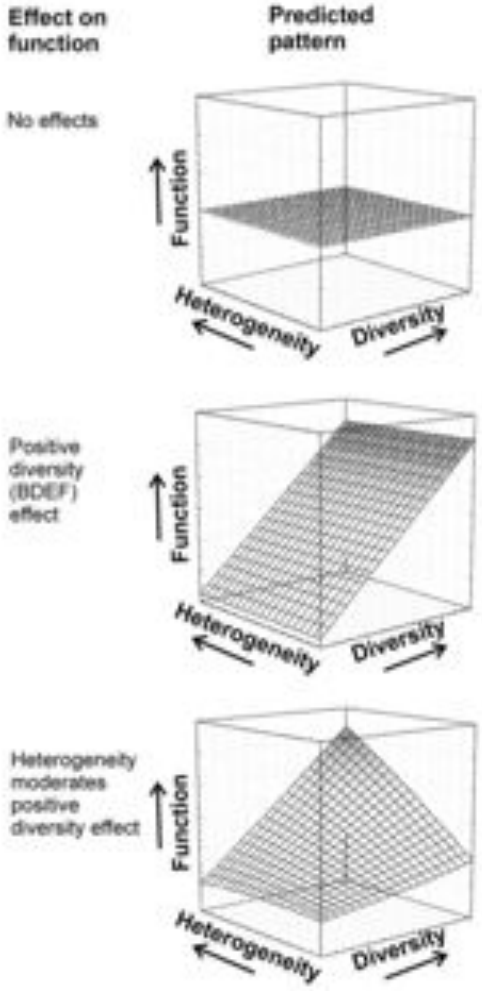
For example, what does setting a scale mean for repeated measures?

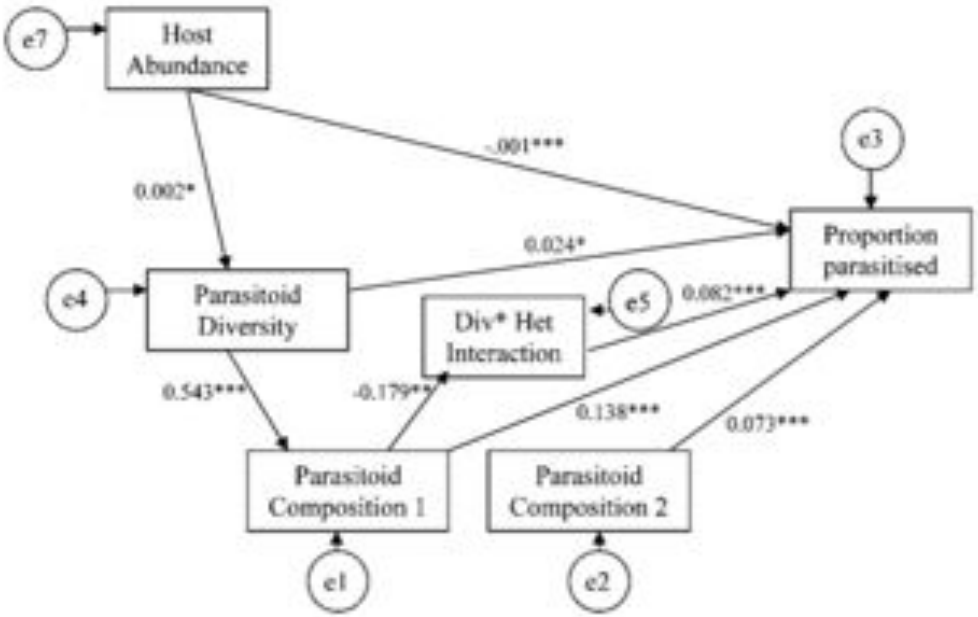
Ignore the problem of start values, or provide bad start values

- Start values close to your solution will speed convergence.
- Start values far from your solution will delay convergence, and can lead you into local solutions.
- If your intuition says that a bad fit is incorrect, try varying start values -> your results should be robust anyway!


The Garden Path: Interpretation

Look Only at Path Significance and Not Overall Fit (yes, people do this)





No fit indices reported.

Email with author (I had some questions about the nonlinearity) revealed that model did not fit the data.

(however, SEM was tacked on to support regression results due to a reviewer – I still believe the message in this paper and think it's a pretty cool attempt to get at some complicated concepts)

Interpret good fit as the model being "proved"

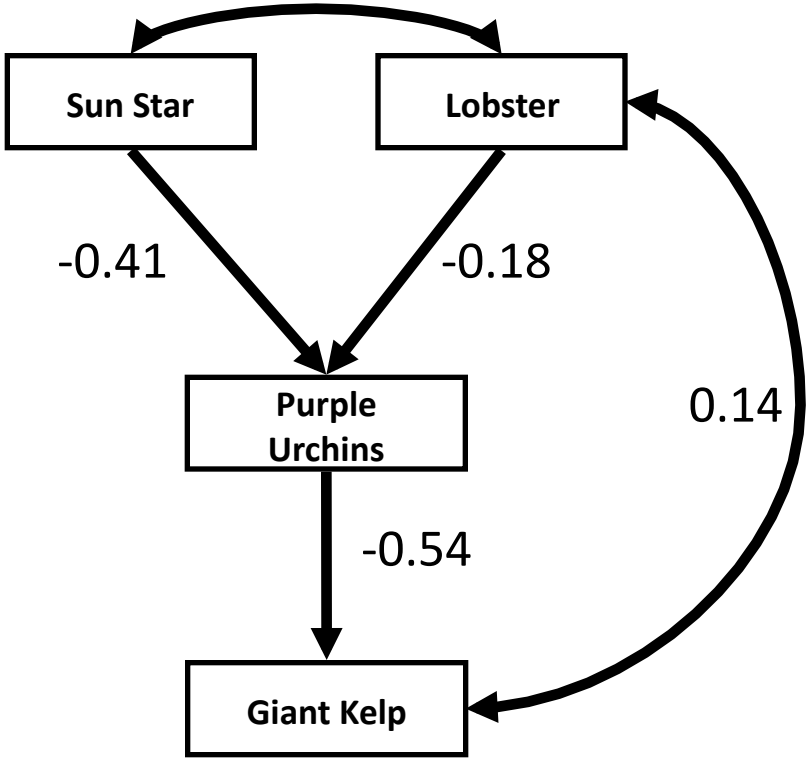
- P(Data | Hypothesis) = Probability of observing the data at hand given a hypothesis proposed.

- High p values **DOES NOT** equal good support for a hypothesis.

- P values are useful only as support for **rejecting** a hypothesis.

- Data can be consistent with one hypothesis and still conform to many others.

Look only at indexes of overall fit and ignore other information about fit; interpret fit indices as meaning that endogenous variables are strongly predicted



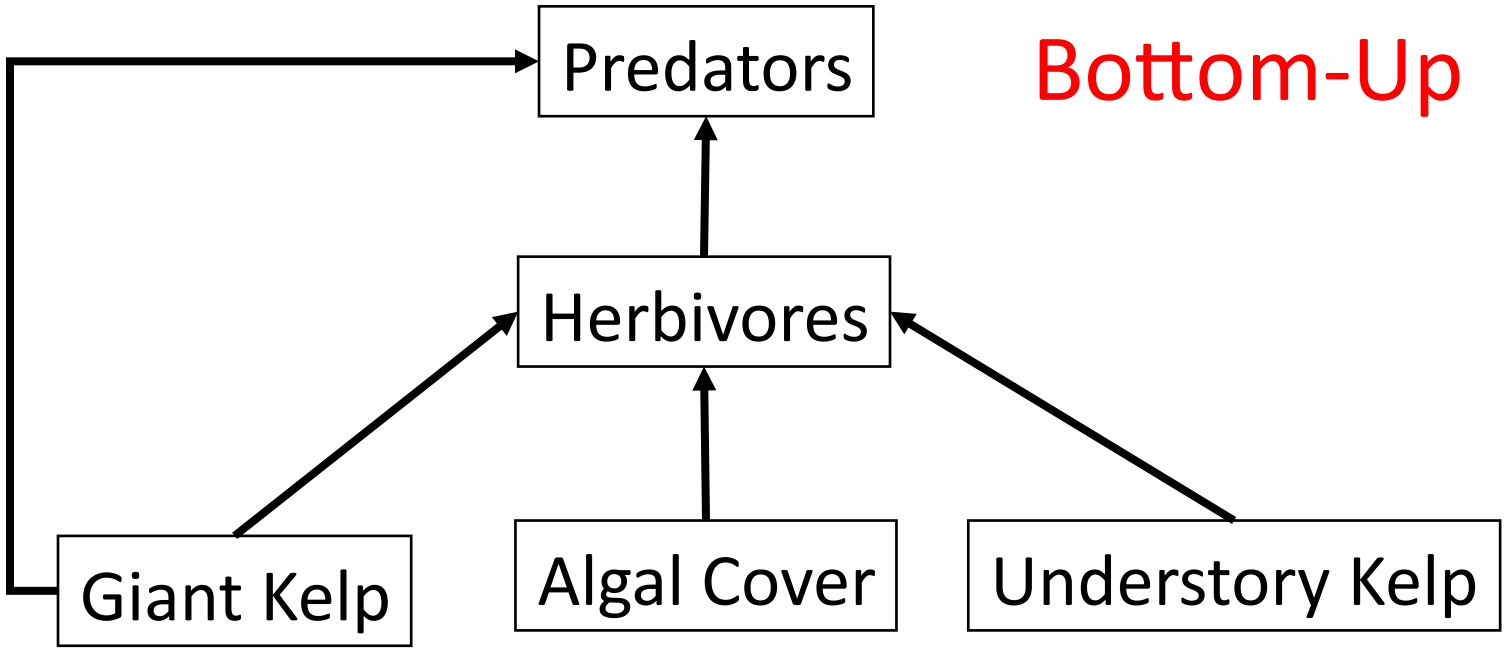
$\chi^2 = 2.58$, df = 1, P = 0.108

$R^2_{lobster} = 0.09$

$R^2_{urchins} = 0.17$

$R^2_{kelp} = 0.31$

Rely solely on statistical criteria in model evaluation



What if you fit this model, and all paths were significant. You have good model fit. Variables were even well predicted, but…

ALL OF THE PATH COEFFICIENTS WERE NEGATIVE

Interpret the standardized solution in inappropriate ways

Smith et al. 1997

TABLE 3

CALCULATED PATH EQUATIONS OF THE EFFECT OF KANGAROO RATS ON OTHER RODENTS

Treatment	Path Equation	e_y
Harvest mice:		
A	$-.303 = -.230 - .073$	.85
B	$-.710 = -.723 + .014$	.66
C	$-.760 = -.450 - .310$	.57
Pocket mice:		
A	$-.212 = -.150 - .062$	.91
B	$-.554 = -.561 + .006$	.82
C	$-.362 = .045 - .407$	.67

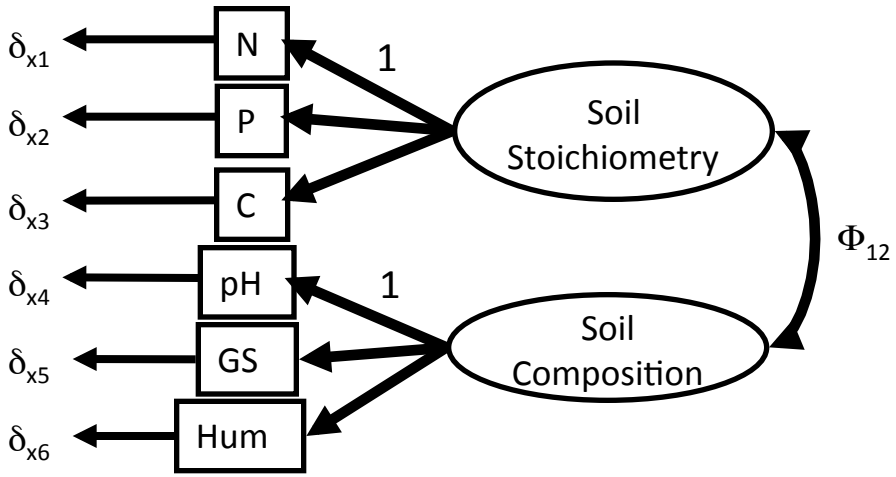
Groups differed in variation, producing different standardized coefficients!

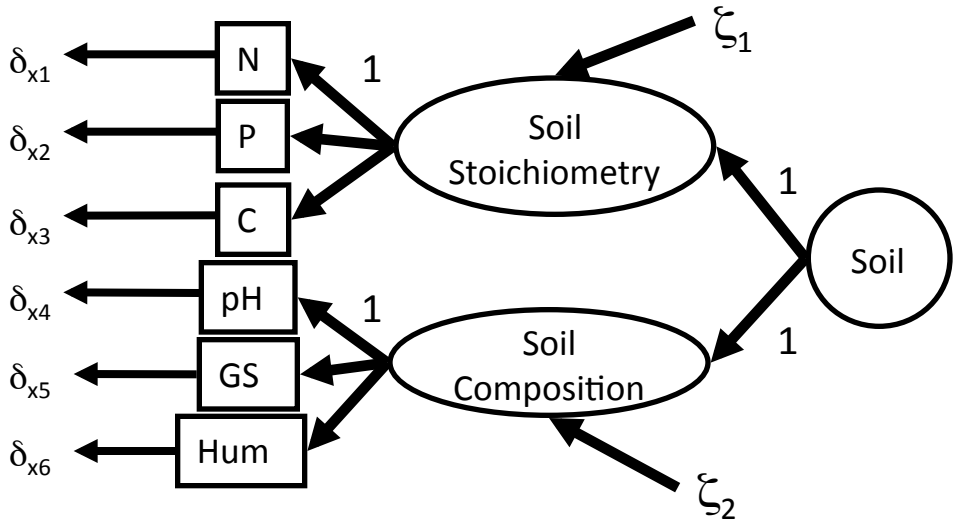
versus

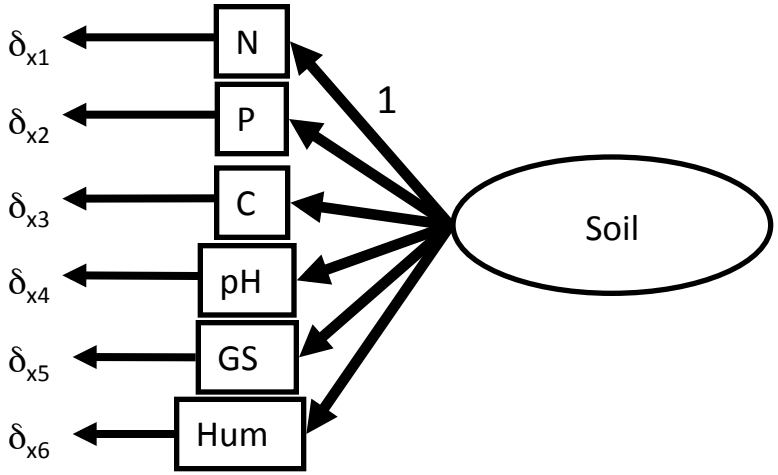
Grace and Pugesek 1998

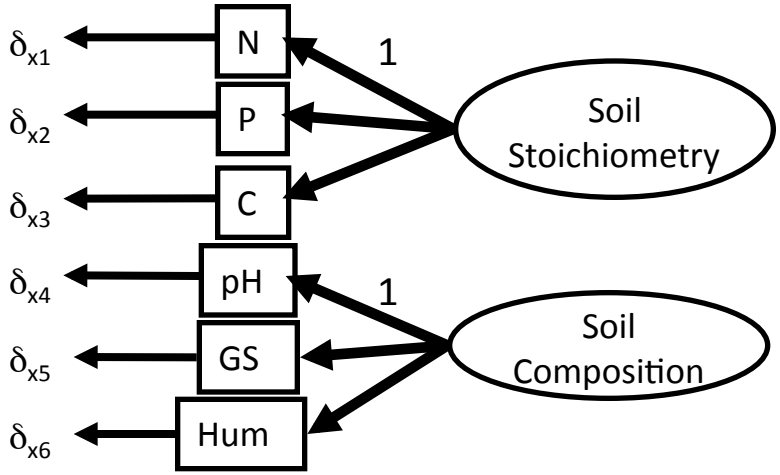
Harvest mice	*Pocket mice*
A. Multigroup goodness-of-fit statistic	
N for multigroup analysis = 38	N for multigroup analysis = 38
χ^2 with 9 df = 7.8400 (P = .5503)	χ^2 with 8 df = 9.6308 (P = .2919)
Group A χ^2 =3.2771	Group A χ^2 =4.6998
Group B χ^2 =1.7419	Group B χ^2 =2.2934
Group C χ^2 =2.8210	Group C χ^2 =2.6376

Fail to consider equivalent models

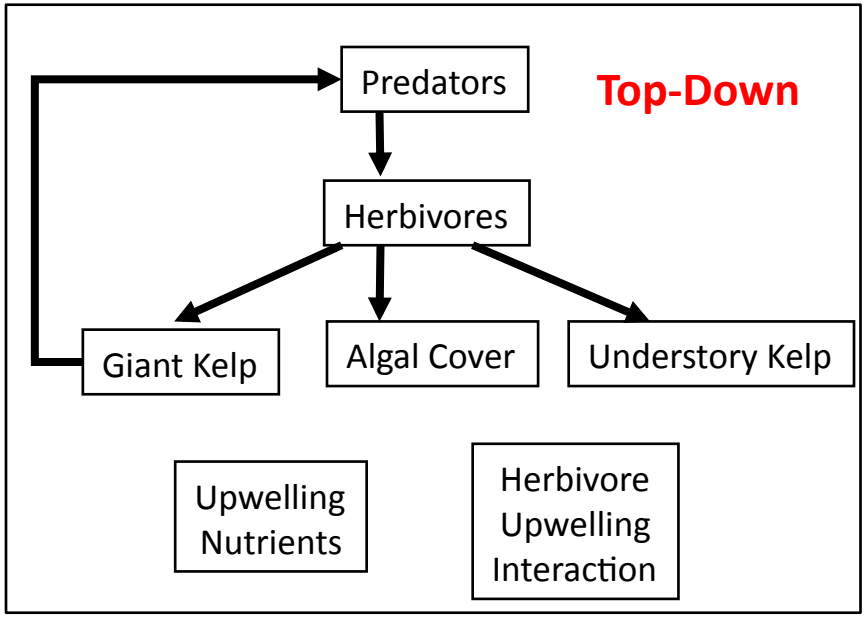


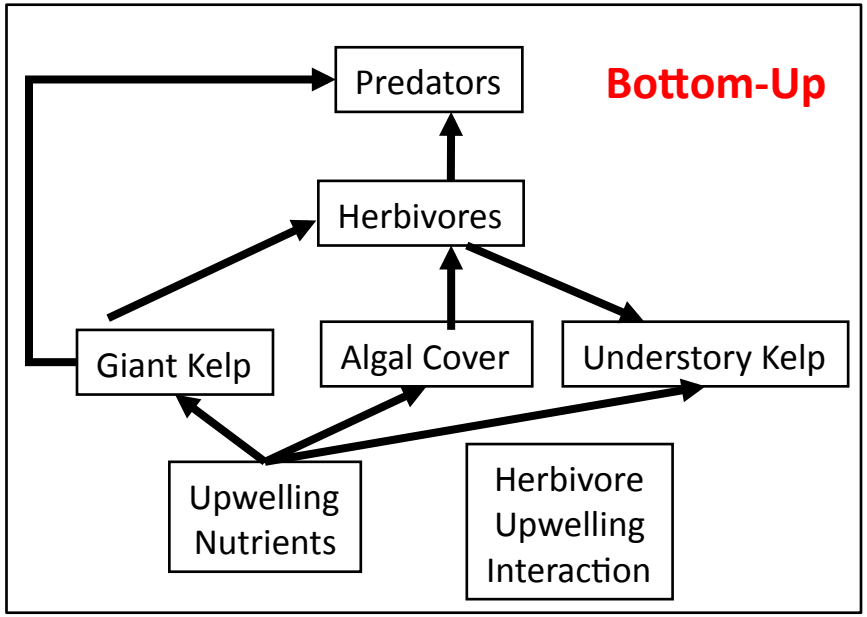


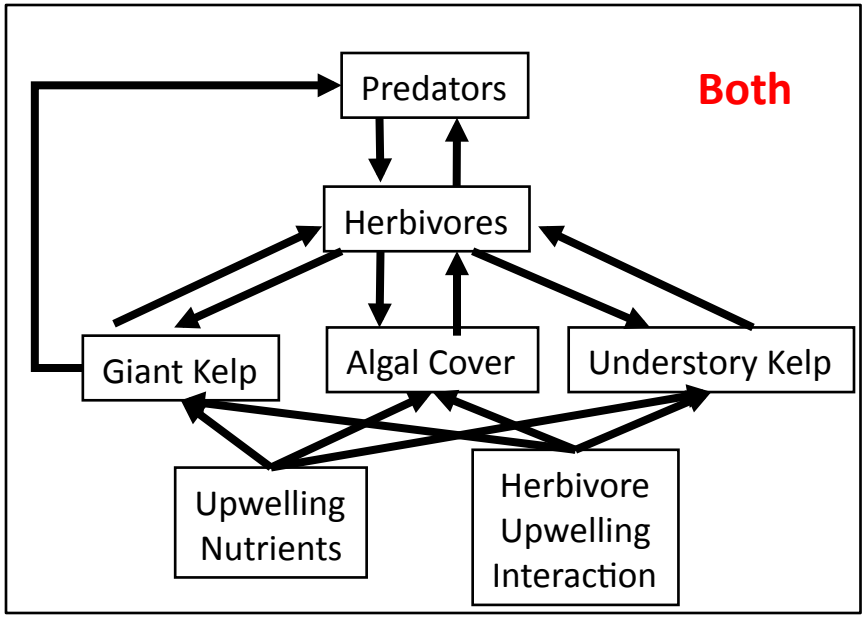




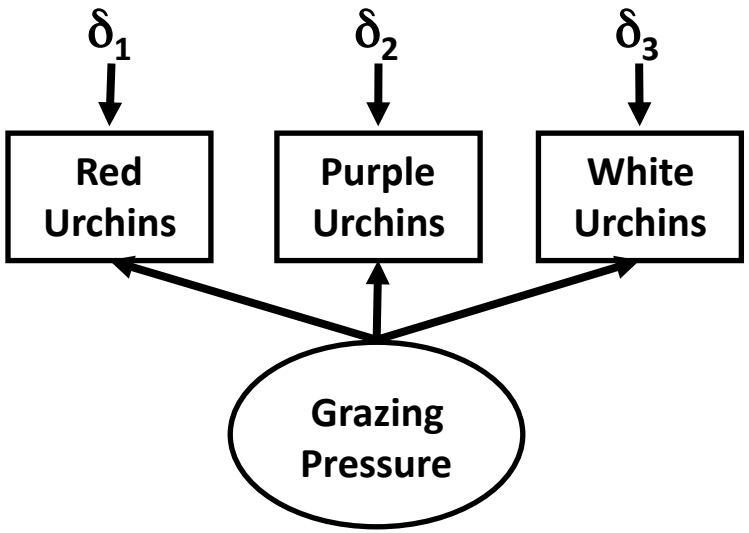
Fail to consider non-equivalent alternative models

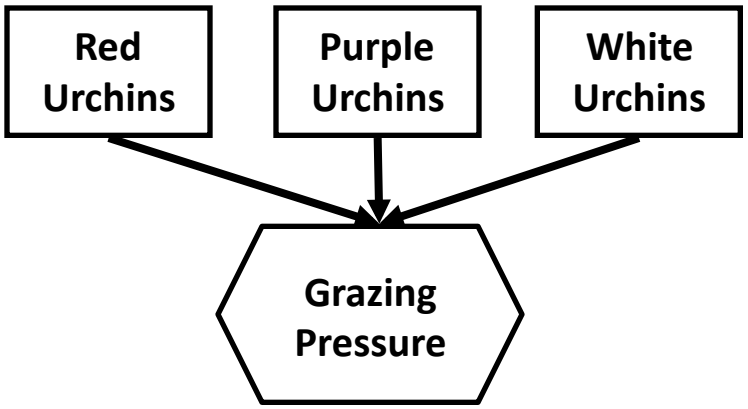






Believe that naming a latent variable with a concept makes it so





What information does this latent variable really convey? How did it differ from the composite variable we came to believe was correct?

Beware of the fallacy of naming!

Believe that a strong SEM analysis can compensate for a poor study

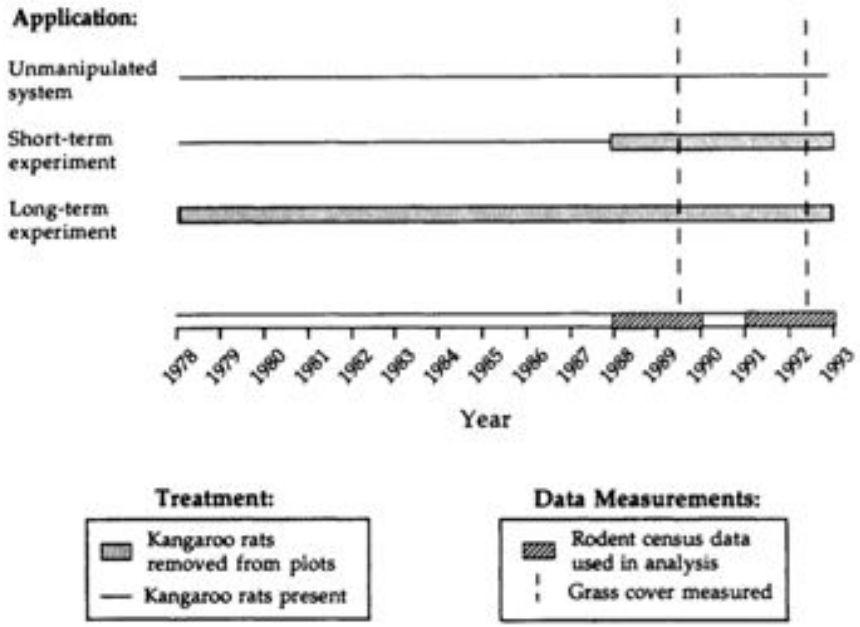


TABLE 3

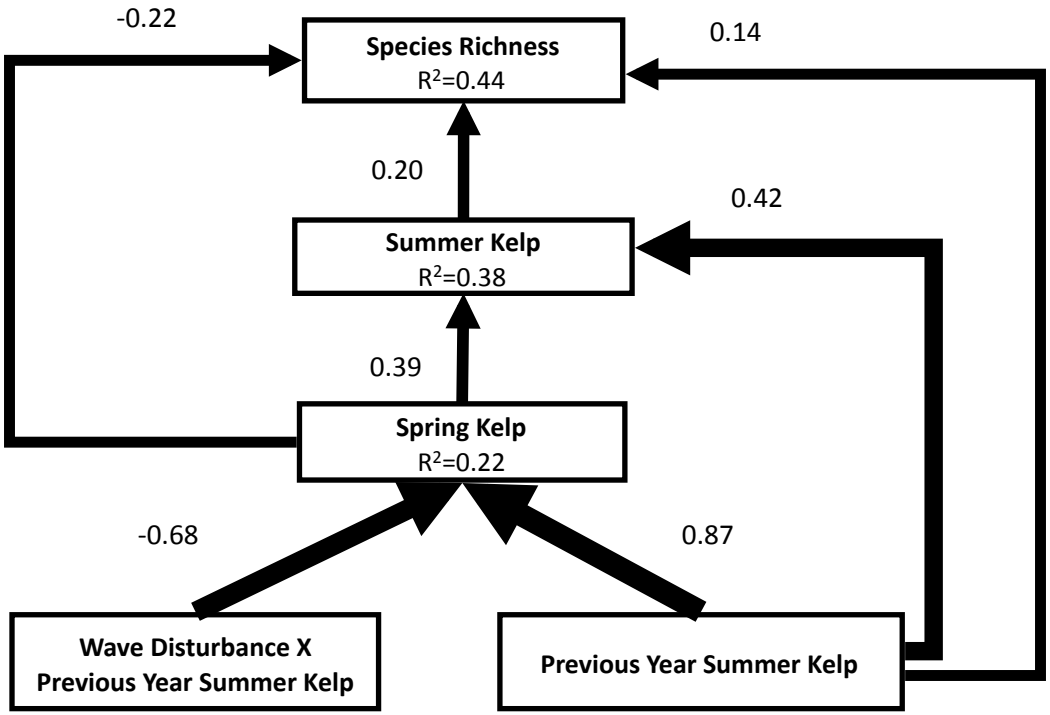
CALCULATED PATH EQUATIONS OF THE EFFECT OF KANGAROO RATS ON OTHER RODENTS

Treatment	Path Equation	e_y
Harvest mice:		
A	$-.303 = -.230 - .073$	.85
B	$-.710 = -.723 + .014$	.66
C	$-.760 = -.450 - .310$	.57
Pocket mice:		
A	$-.212 = -.150 - .062$	.91
B	$-.554 = -.561 + .006$	.82
C	$-.362 = .045 - .407$	.67

- Standardized Coefficients varied wildly
- Indirect effects not detected in 2 of 3 path models

Smith et al. 1997

Fail to report enough information so that others can reproduce your model



- If you have a simplified conceptual diagram, include the gory details.

What should I report?

- A clear path diagram
- Relevant fit statistics (χ^2)
- Unstandardized path coefficients and evaluation of whether they are different from zero
- Covariance matrix and/or correlation, standard deviation, and means

Interpret large direct effects as “proof” of causality

1. Question causal assumptions.

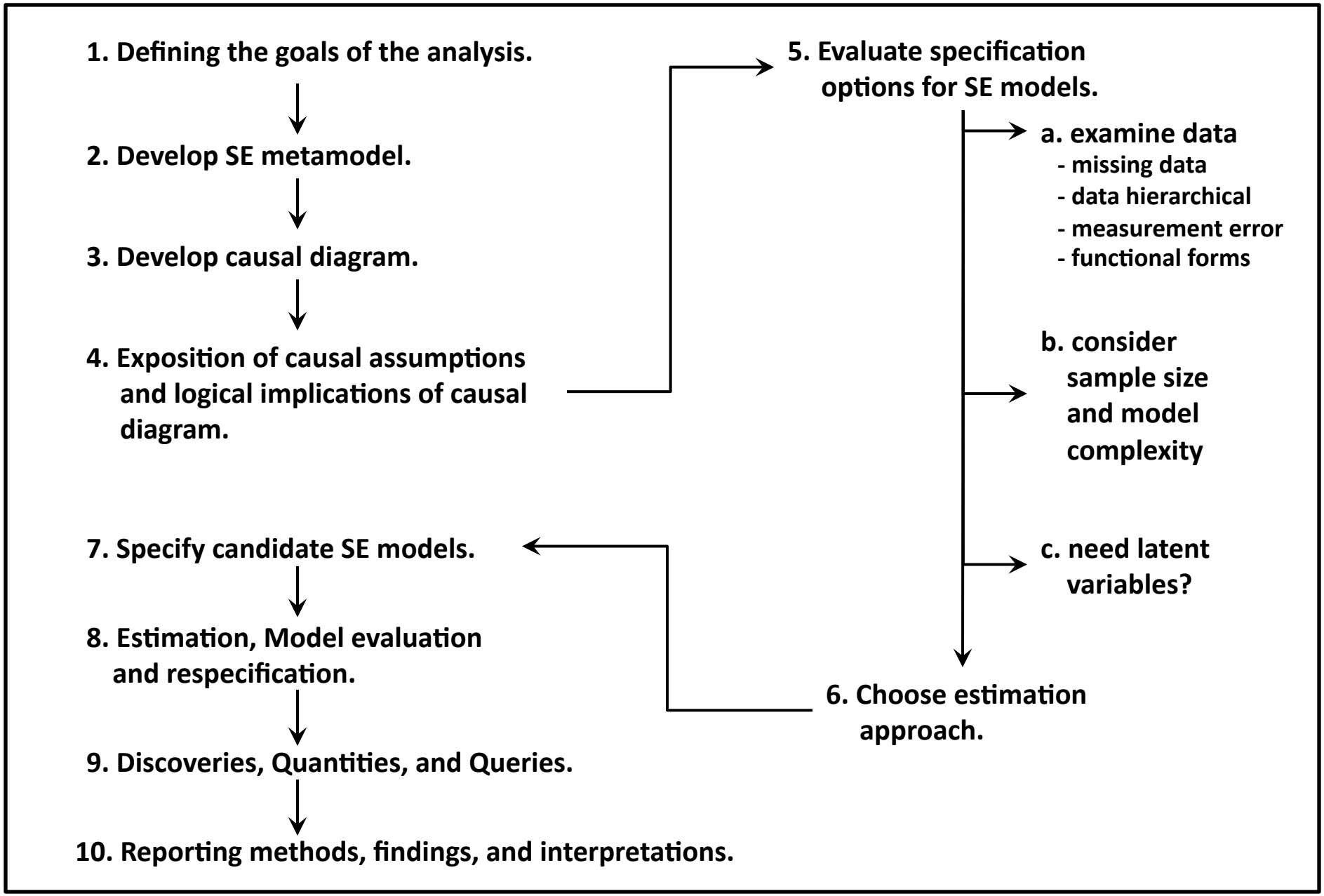
2. Question directionality.

3. Question whether there you have all of the relevant variables included.

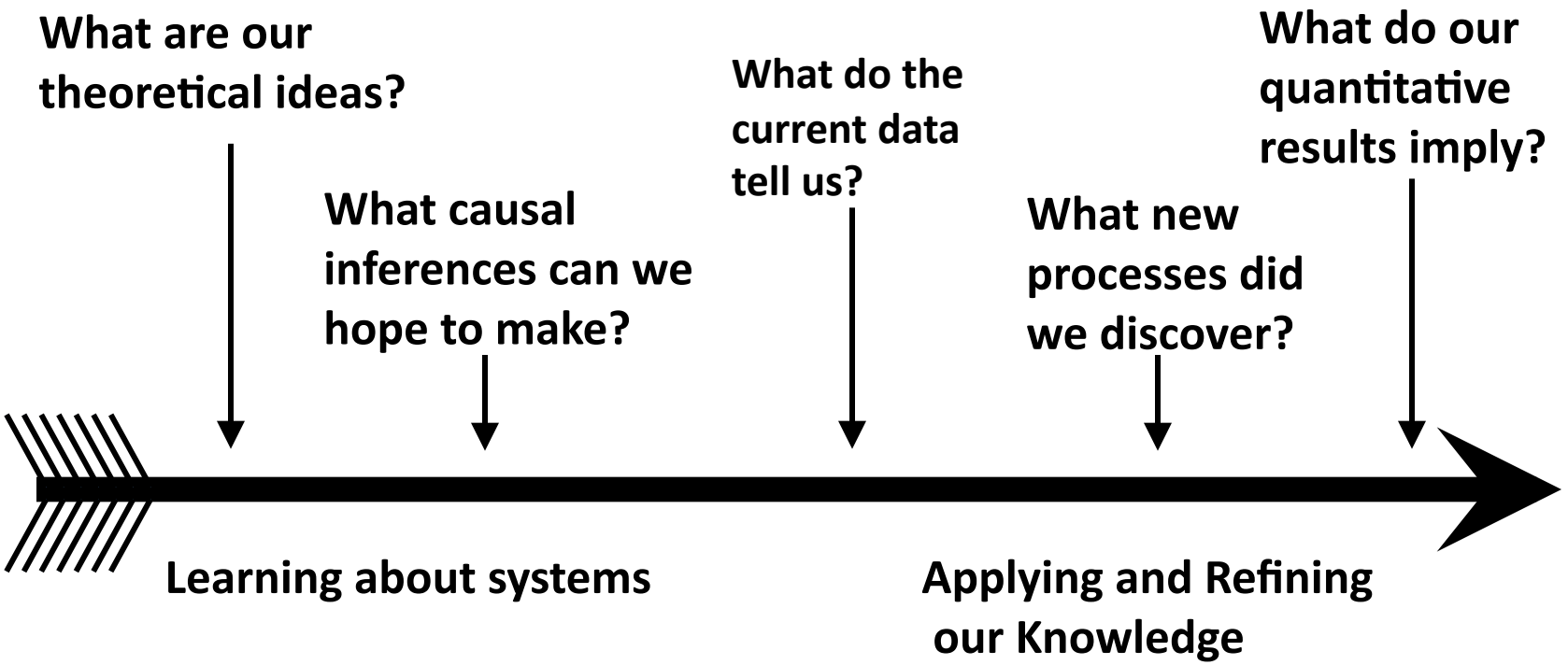
4. Could a misspecified model have led to the same result?

Welcome to the SEM Brigade

A New Framework



An expanded causal inference process



It is likely that no one ever masters anything in which he has not known impotence; and if you agree, you will also see that this impotence comes not at the beginning of or before the struggle with the subject, but in the heart of it.

– Walter Benjamin