$x \rightarrow y$

Path Models I

A different way of thinking about statistical models

Path Models I

Today's goal:

Teach you the basics of path models

Outline:

- Model specification: types of models
- Model identification

$x \rightarrow y$

Model specification

the types of models we can test

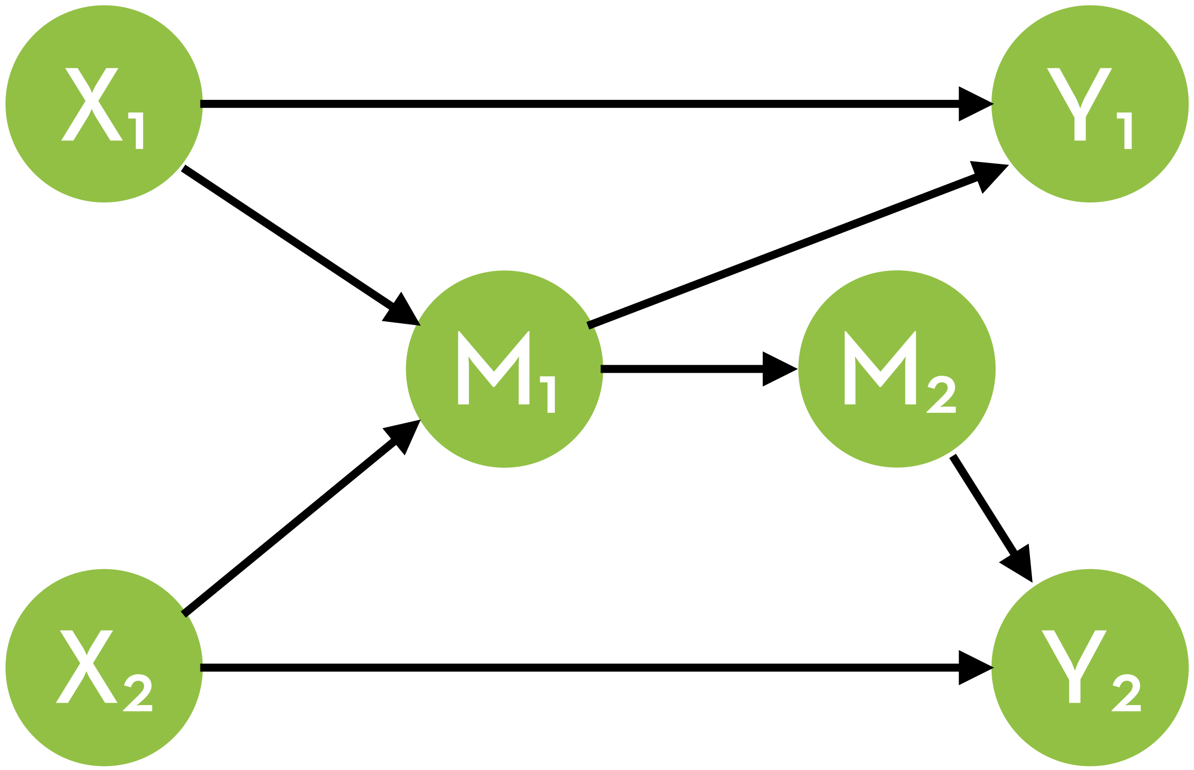
Model specification

In regression, you have created models with one Y and several Xes

- In M&E I we talked about selecting suitable Xes

In path models, you can have many interconnected Xes and Ys

- Models can get very complicated



Model specification

To prevent problems, you will have to **specify** your model

Do this **before** you do your study!

Motivate expected effects, based on:

- previous work
- theory
- common sense

If in doubt, create alternate specifications!

Model specification

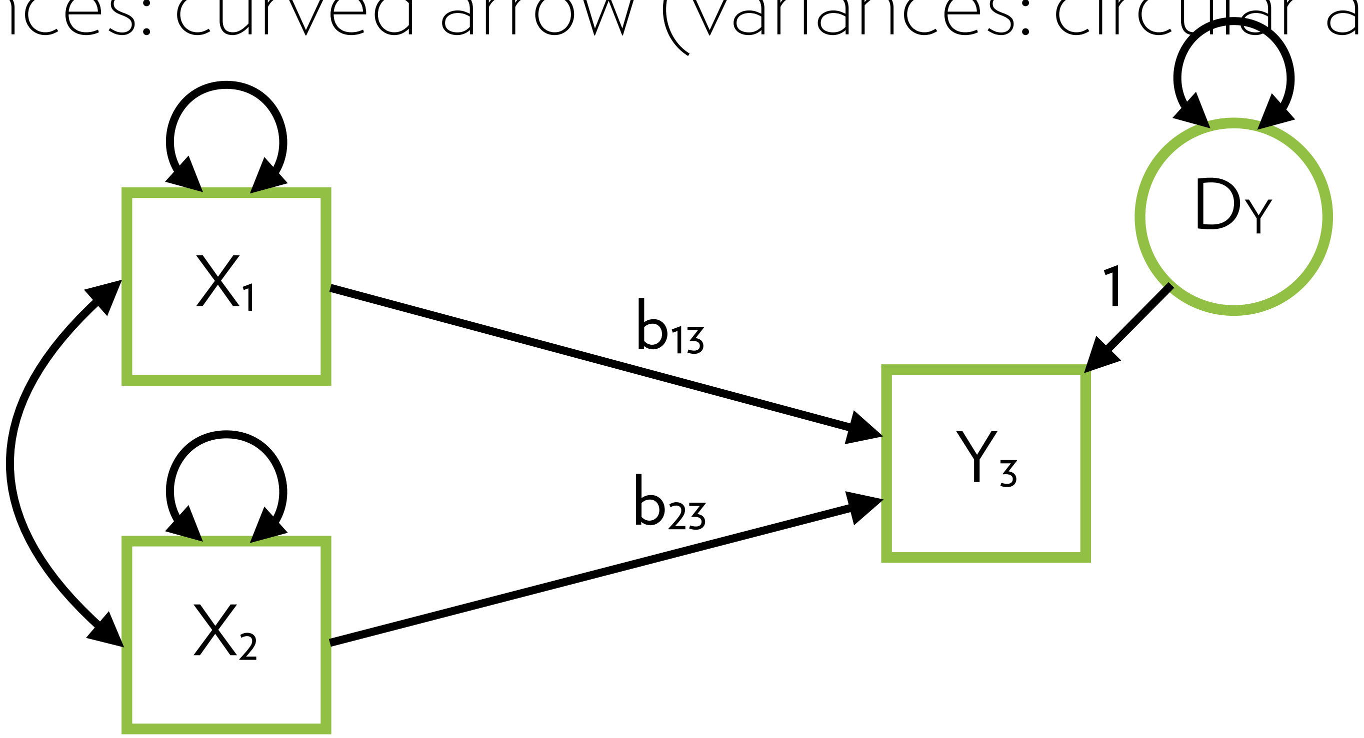
Research steps:

- Specify a model
- Make sure the model is identified
- Run the study and test the model
- If it didn't fit, respecify the model (or start over)
- If it did fit, interpret and report on the results
- (try alternative specifications)

Terminology

Linear regression: $Y_3 \sim b_{13}X_1 + b_{23}X_2 + e$

- Observed variables: square/rectangle:
- Latent variables: circle/ellipse
- Causal effects: arrow
- Covariances: curved arrow (variances: circular arrow)



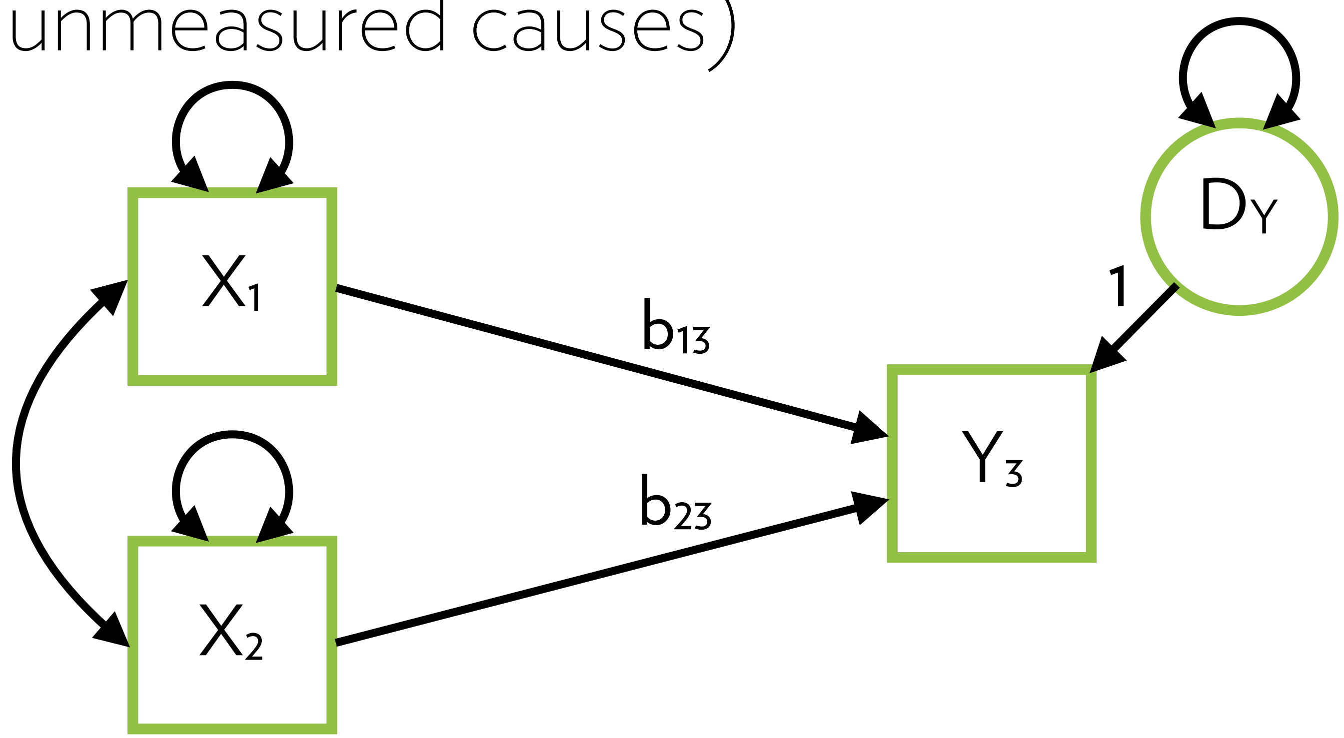
Terminology

Exogenous variables (x's only):

Are free to vary, and always correlated with each other

Endogenous variables (anything that is a y):

Have a "disturbance" (kind of like the "error" in regression; includes unmeasured causes)

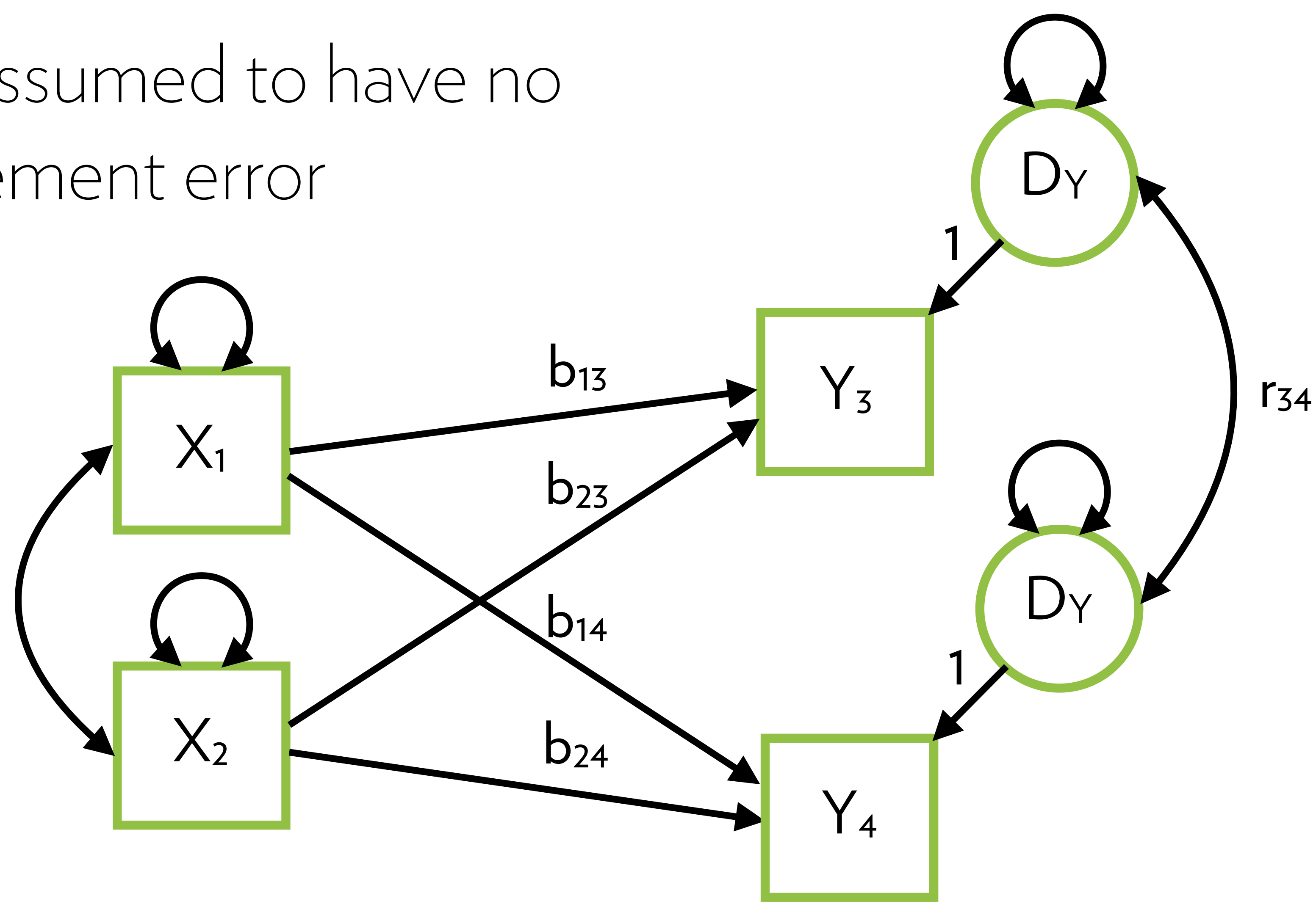


Terminology

More complex:

note: disturbances have a fixed effect on Ys (hence the 1), but are free to vary (hence the circular arrow)

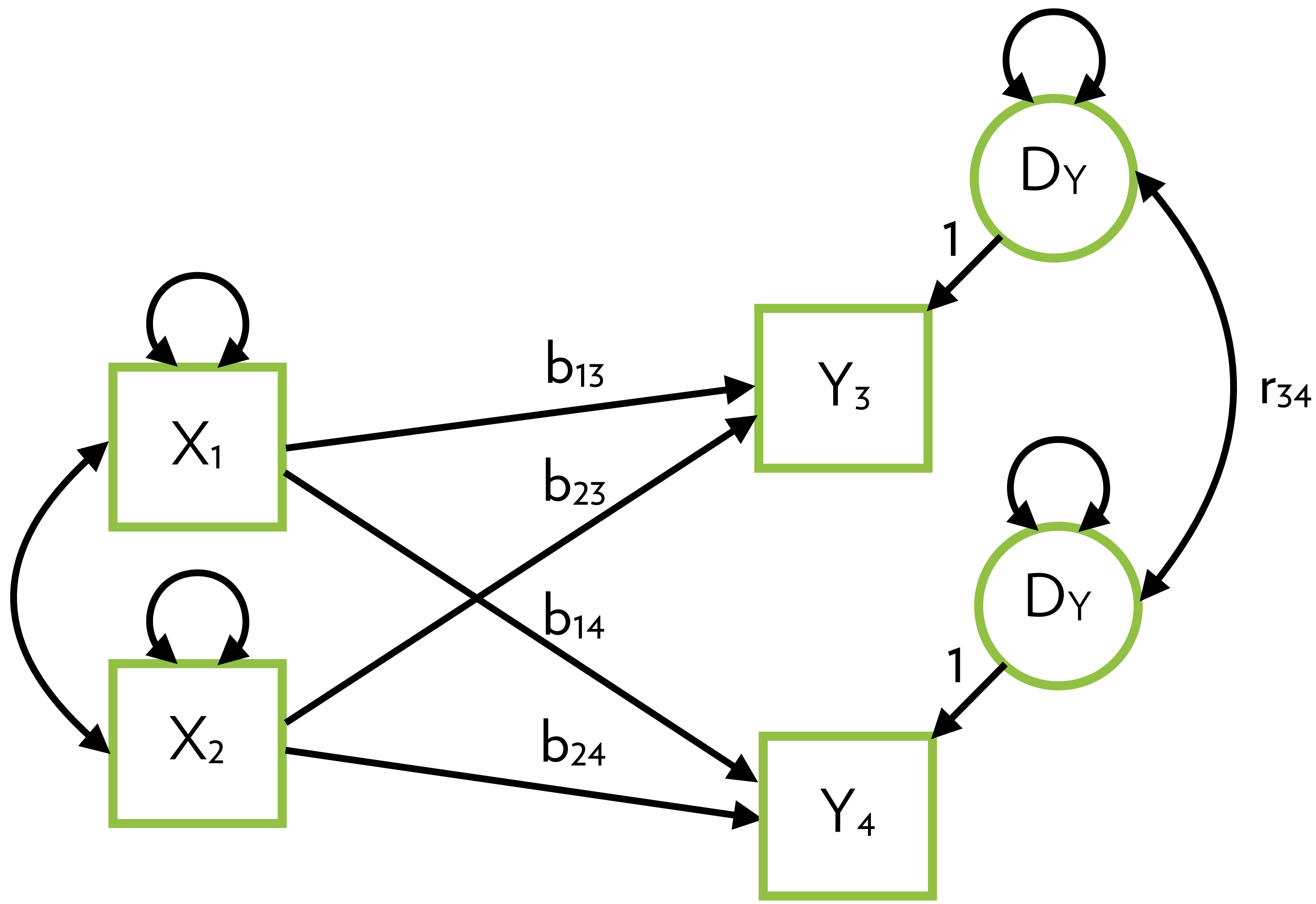
Xs are assumed to have no measurement error



Terminology

Covariance of disturbance terms is not assumed!

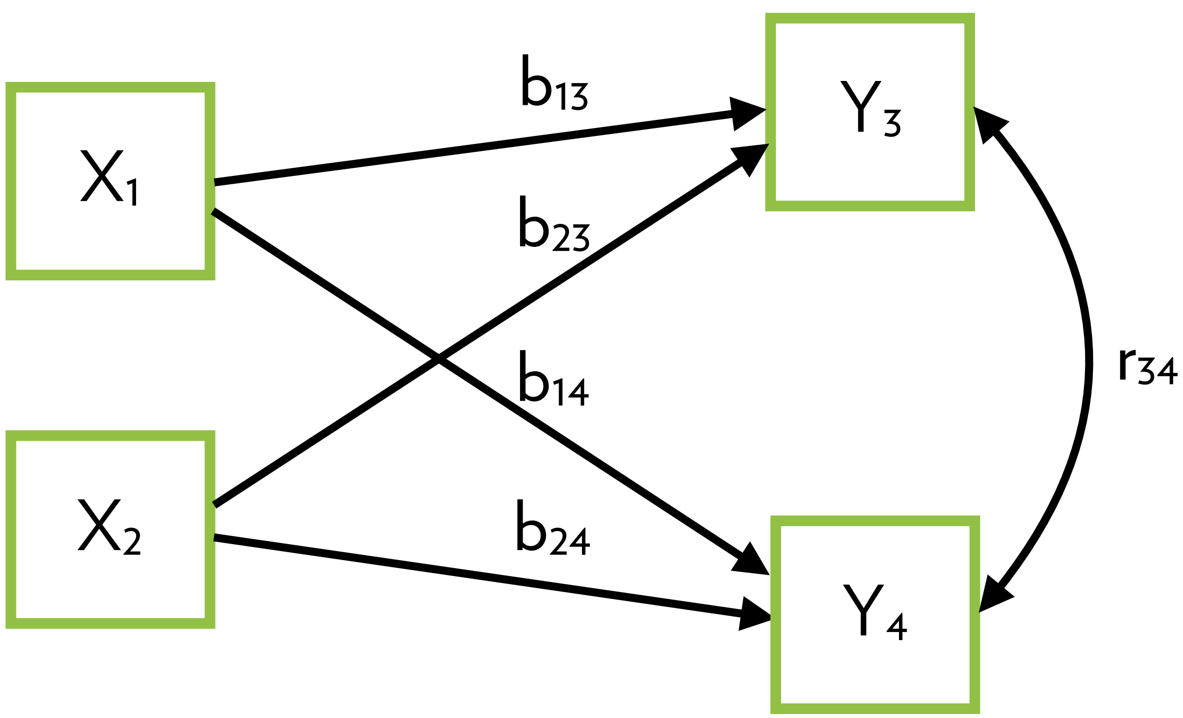
- If you model this, it means you think Y_3 and Y_4 share unmeasured causes



Terminology

Shorthand notation:

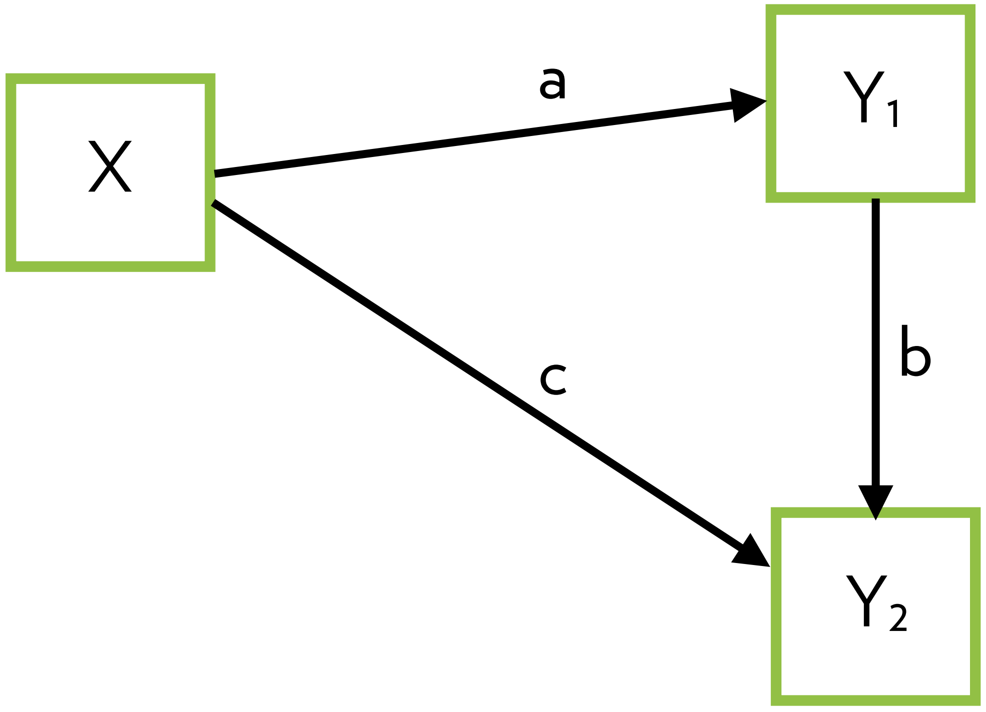
- Hide correlations between Xs (keep for the Ys)
- Hide disturbances
- Hide variances



Easy mediation

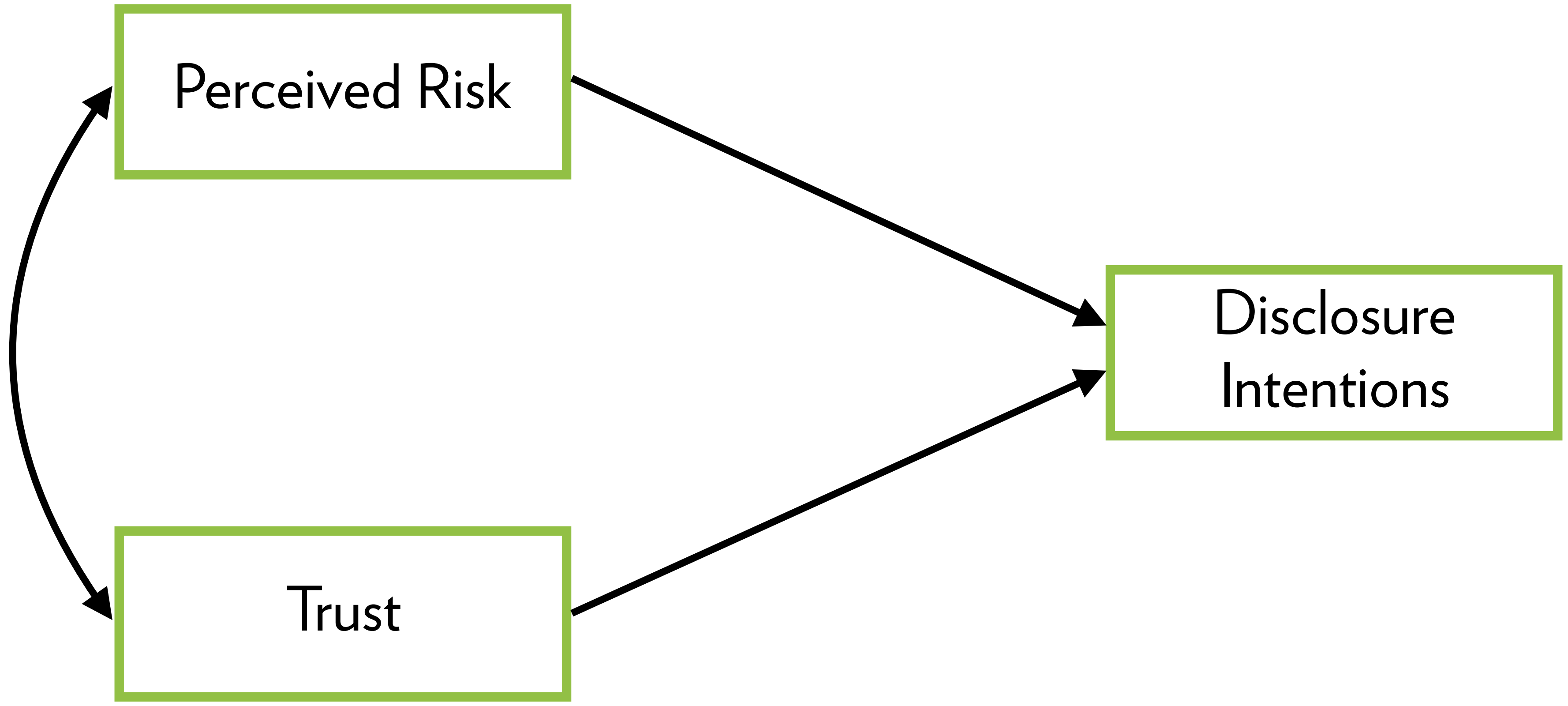
Mediation

- Direct effect: c
- Indirect effect: a*b
- Total effect: a*b + c

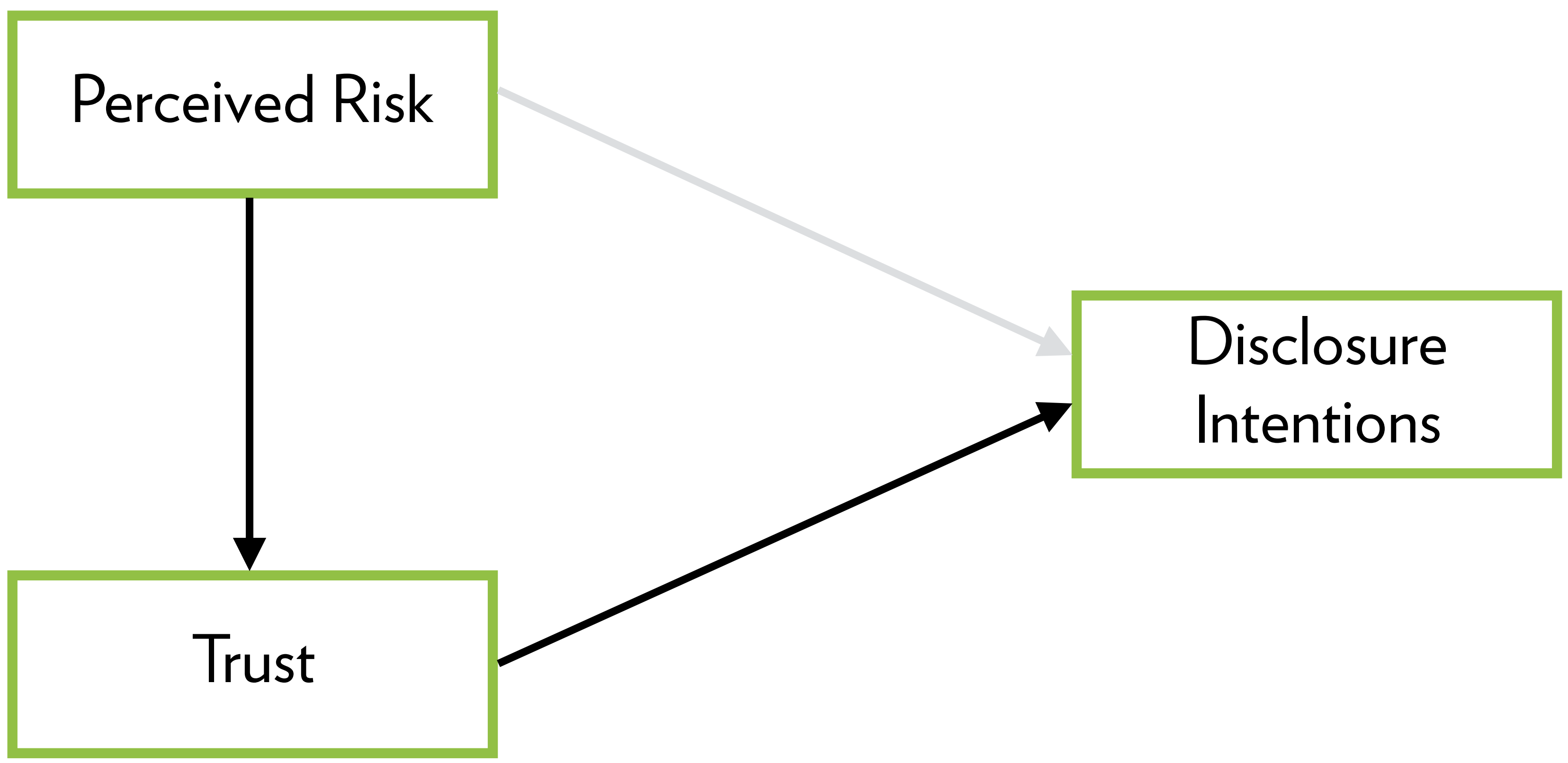


Causality

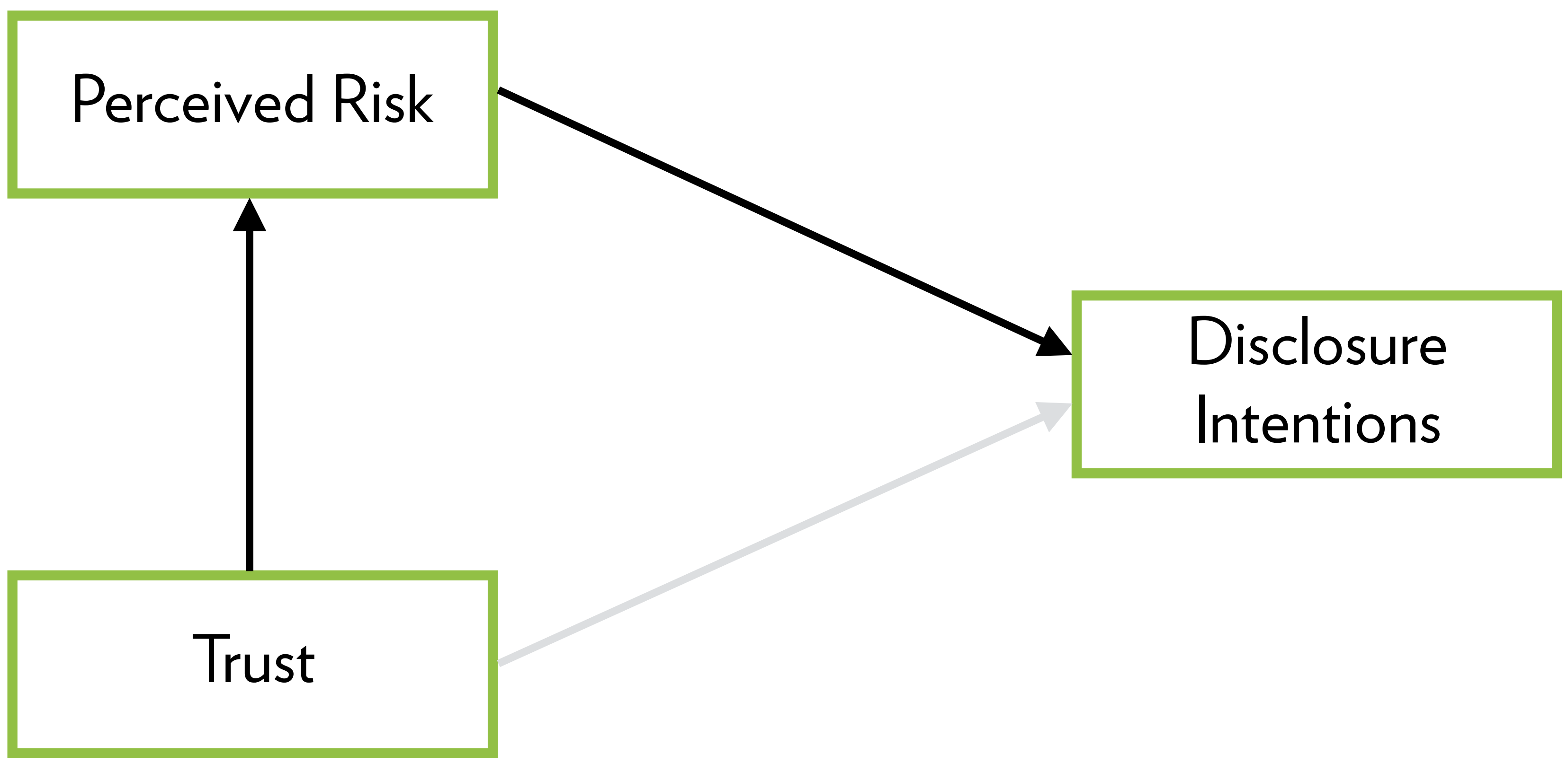
Which model is correct?



Causality



Causality



Causality problems

X = Facebook use

Y = Depression

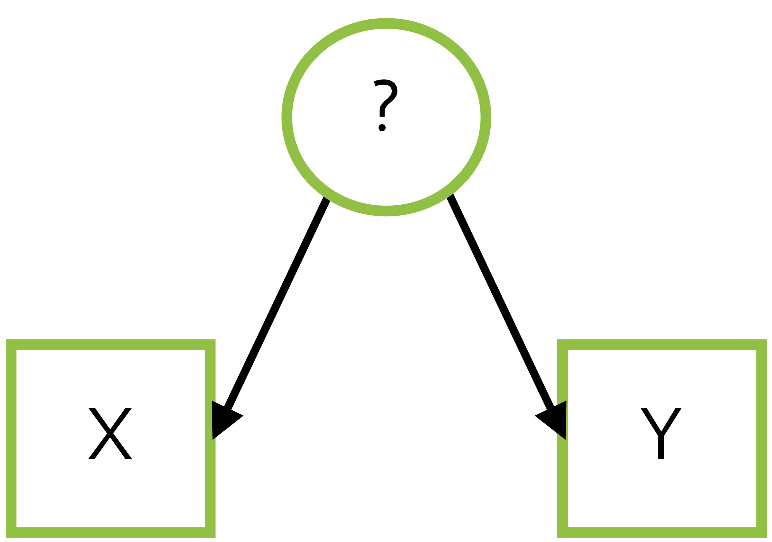
What causes what?

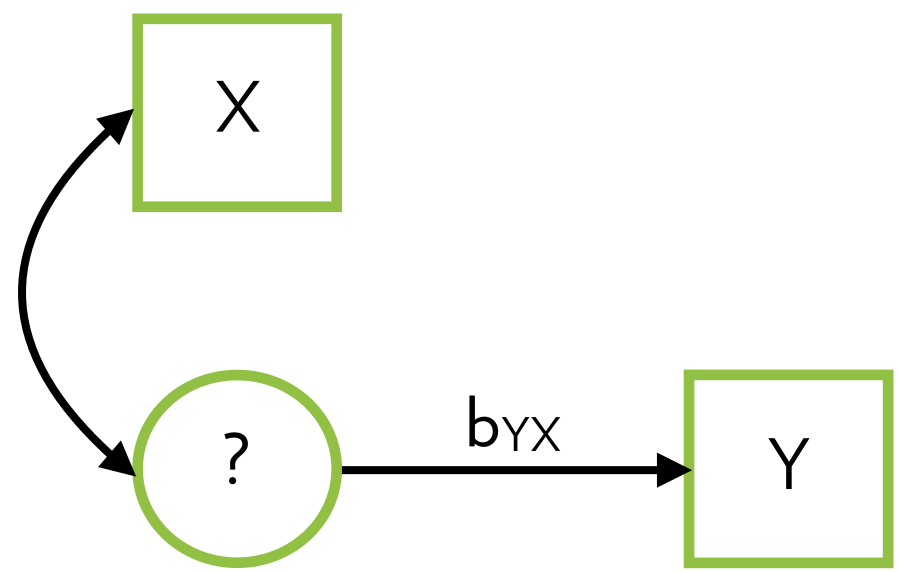
- Models are equivalent!
- We can't determine the "right" model based on the data

Causality problems

Other options:

- A third variable causing both (e.g. bad weather?)
- A third variable causing depression, correlated with Facebook use (e.g. boredom?)





$x \to y$ Causality rules

X causes Y if:

- Temporal precedence: X happens before Y
- Counterfactual: There is a control group of “not X” to compare to
- Isolation: There is no other plausible explanation (third variable)

This is all true when X is a manipulation!

This is why experiments are so awesome

$x \rightarrow y$ Causality rules

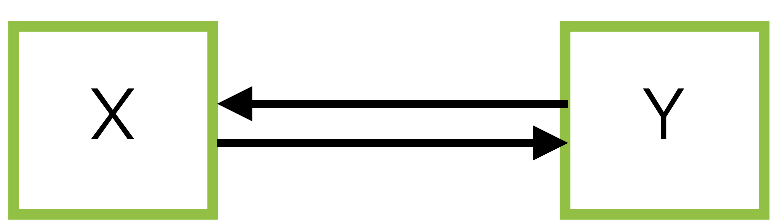
What if X and Y are both measured variables?

- Reason from theory (what have others shown?)
- Measure Y **after** X (using a time-referent or longitudinal study)
- Specify a relation without assuming causality:

Causality rules

Note: a reciprocal relationship assumes that X and Y influence each other instantaneously

- Probably not true for depression and Facebook use
- Might be true for e.g. my mood and your mood



$x \rightarrow y$ Other assumptions

Other assumptions of path models:

- Uncorrelated observations (we can do multi-level though)
- No interactions (we can specify some types, but difficult)
- Parts of the model hold up in isolation
- Good measurement reliability (more on this later)
- Include all causes (Xs) for each outcome (Ys)

Model specification

You can go crazy specifying your model, but note that there are **two types** of models:

- Recursive models (easy)
- Nonrecursive models (hard)

Recursive models

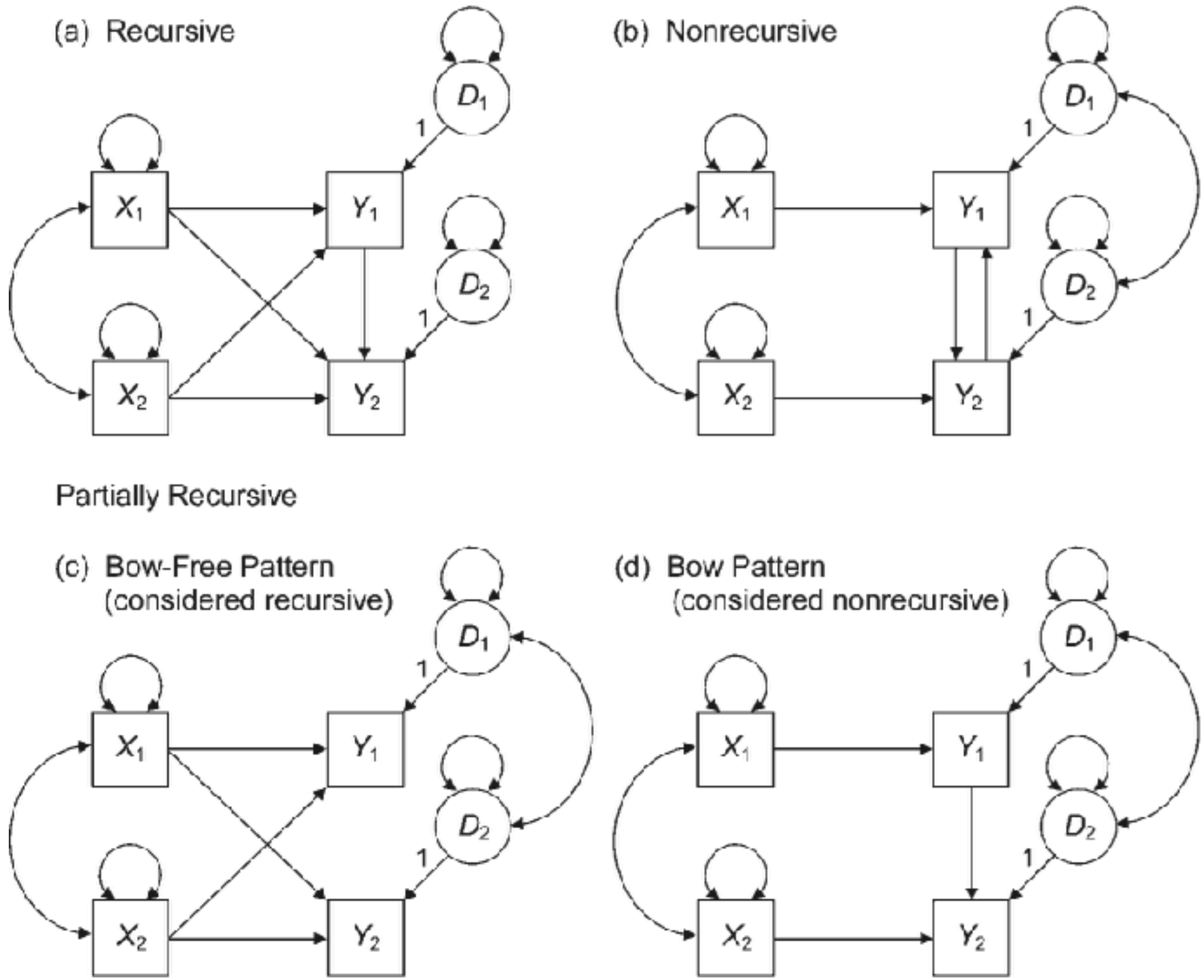
No direct or indirect loops **and**

No correlations between Ys

(exception: a correlation between Ys without a direct causal effect between the same Ys)

Otherwise: nonrecursive!

Recursive models



Model identification

figuring out if a model can be evaluated or not

Model identification

Identification:

Is it theoretically possible to fit this model?

(may still not yield a practical result!)

A model is identified if:

- the model degrees of freedom is at least zero
- every latent variable is "scaled" (this is the 1 on the disturbance arrow)
- additional rules for nonrecursive models (I told you these would be harder!)

Degrees of freedom

Let's say you have four variables… How many regressions can you test?

In path models, this depends on the number of "**observations**"

This is **not** N, but the number of variances and covariances!

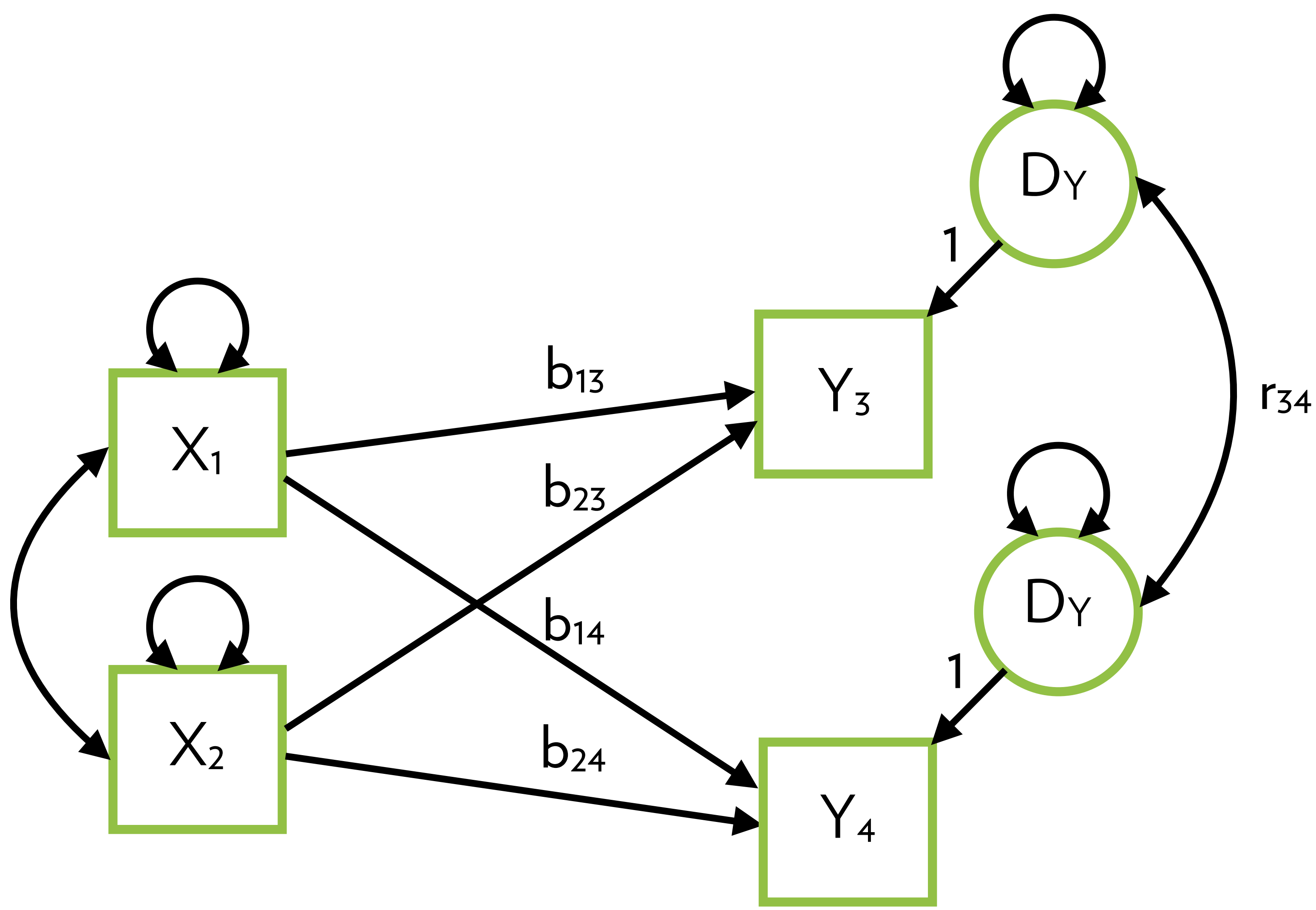
v(v+1)/2 where v = # of variables

p1	a1	a2	a3	a4
a1	d_1^2	d_{12}	d_{13}	d_{14}
a2	d_{12}	d_2^2	d_{23}	d_{24}
a3	d_{13}	d_{23}	d_3^2	d_{34}
a4	d_{14}	d_{24}	d_{34}	d_4^2

Degrees of freedom

Number of **estimated parameters**: count the number of arrows (except the ones with a 1 on them)

For this model: 10



Degrees of freedom

Degrees of freedom = # observations - # estimated parameters

- df = 0: model is **just-identified** and has a perfect fit (not very useful for testing)
- df > 0: model is **overidentified**; there are more observations than parameters (this allows for testing)
- df < 0: model is **underidentified** and cannot be calculated

Overidentified models are good; the larger the df, the simpler the model!

An analogy

Let's say you have the following "observation":

(1) a + b = 6

Can you determine the value of a and b?

No! You have fewer observations than parameters!

This "model" is underidentified

An analogy

How about the following two observations:

(1) a + b = 6

(2) 2a + b = 10

Can you determine the value of a and b?

(2) - (1): a = 4

fill in (2): b = 2

There is a single "perfect" answer

This model is just identified

An analogy

How about the following three observations:

(1) $a + b = 6$

(2) $2a + b = 10$

(3) $3a + b = 12$

Can you determine the value of a and b?

There is no single perfect solution, but you can try to get as close as possible, with some error (e.g. $a = 3$, $b = 3.33$)

This model is overidentified

An analogy

If # observations >= # parameters, then the model is identified

However, sometimes the observations are not unique, e.g. a + b = 6 and 2a + 2b = 12

- This is why multicollinearity can lead to problems!

Also, you may only be able to identify a subset of the parameters, e.g. a + b + c = 6, 2a + b + c = 10, b + c = 2

- If certain parameters cannot be identified, then the model is not identified either

Degrees of freedom

Identified models can be fitted, but the fit will always be perfect

- Not very useful

Overidentified models are better; the larger the df, the simpler the model!

- But with fewer paths, your fit will be lower

We have to balance simplicity and fit

Degrees of freedom

Adding parameters (paths) to a model will always increase the fit

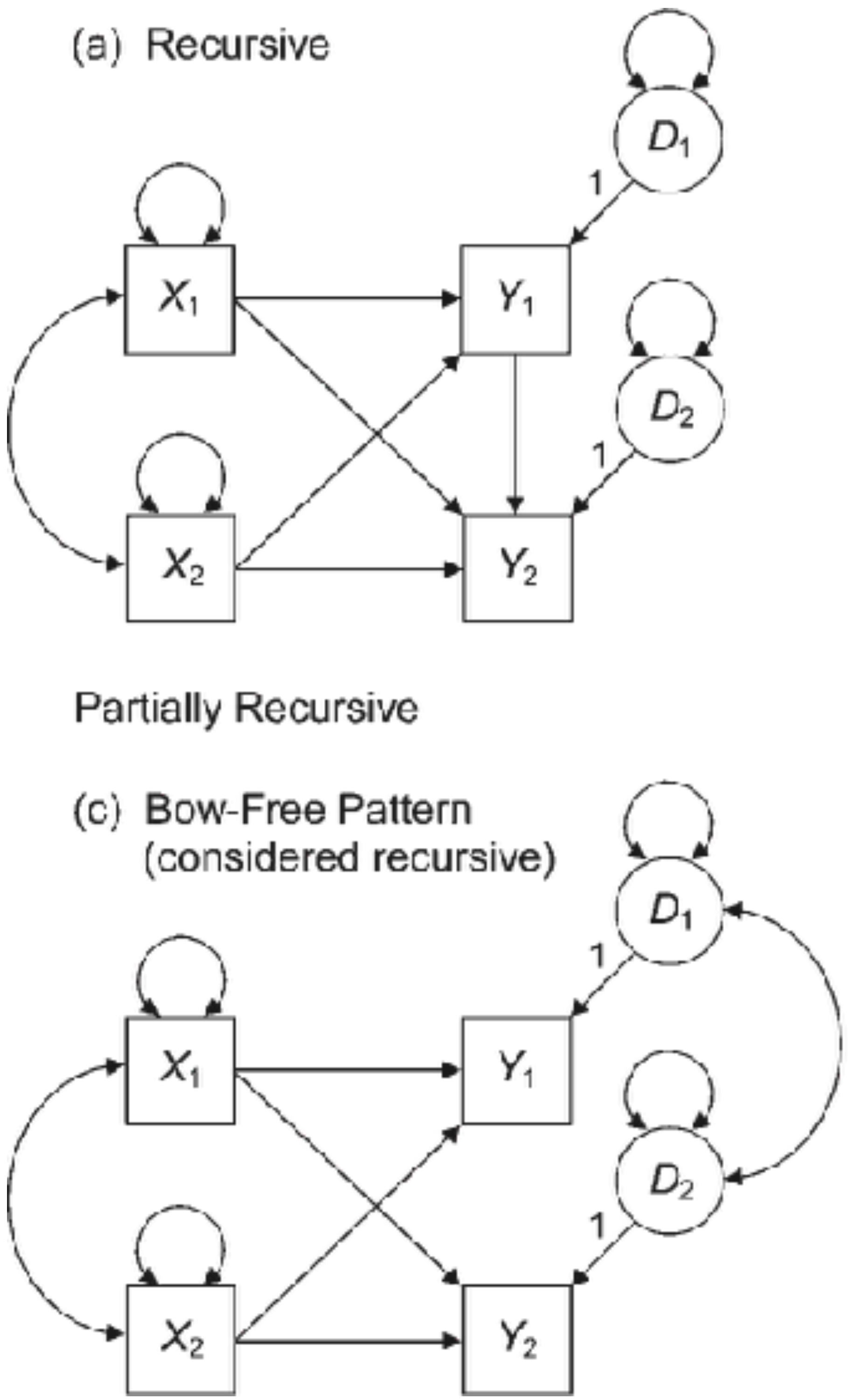
- The new model is "nested" within the old one

This is similar to adding an X to a regression

- This will always increase the R^2

You can test these "nested models" to find out whether the addition was justified

Recursive models



Recursive models

Recursive path models are always identified

Saturated recursive path models (with all possible arrows) are just-identified

- In other words: you can't really test everything, because it will just show a perfect fit

For each arrow you remove from the model, you will increase the df by 1

- This will reduce the fit, but the model will get simpler

Nonrecursive

The identification rules for nonrecursive models are more complicated

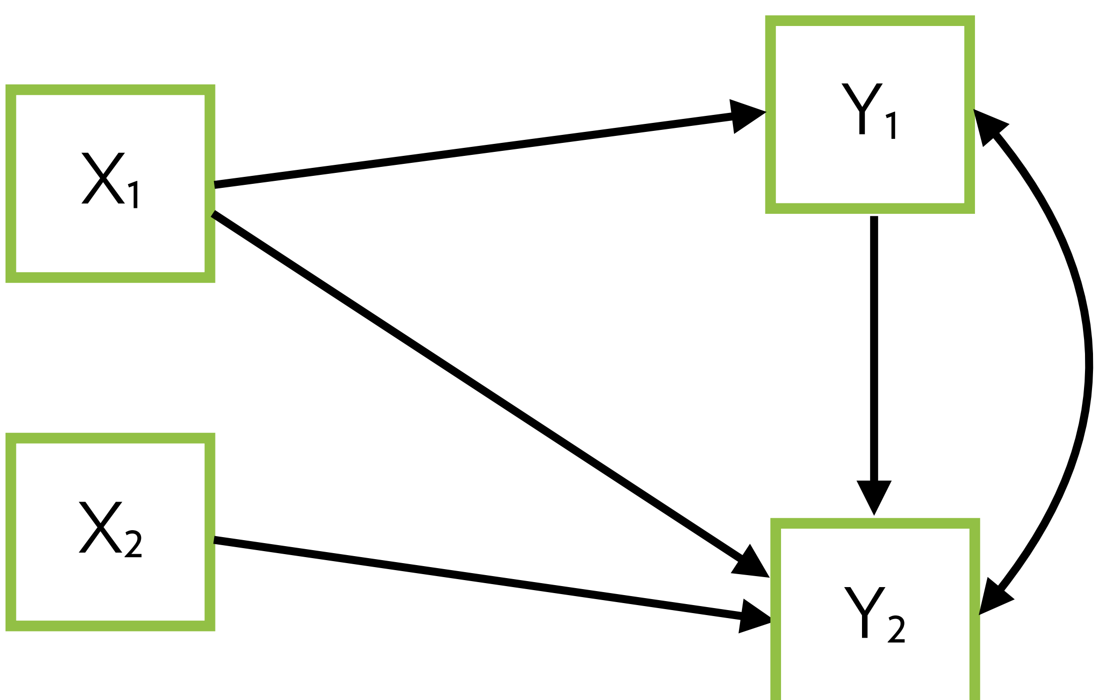
- Pro tip: avoid nonrecursive models wherever possible!

Nonrecursive

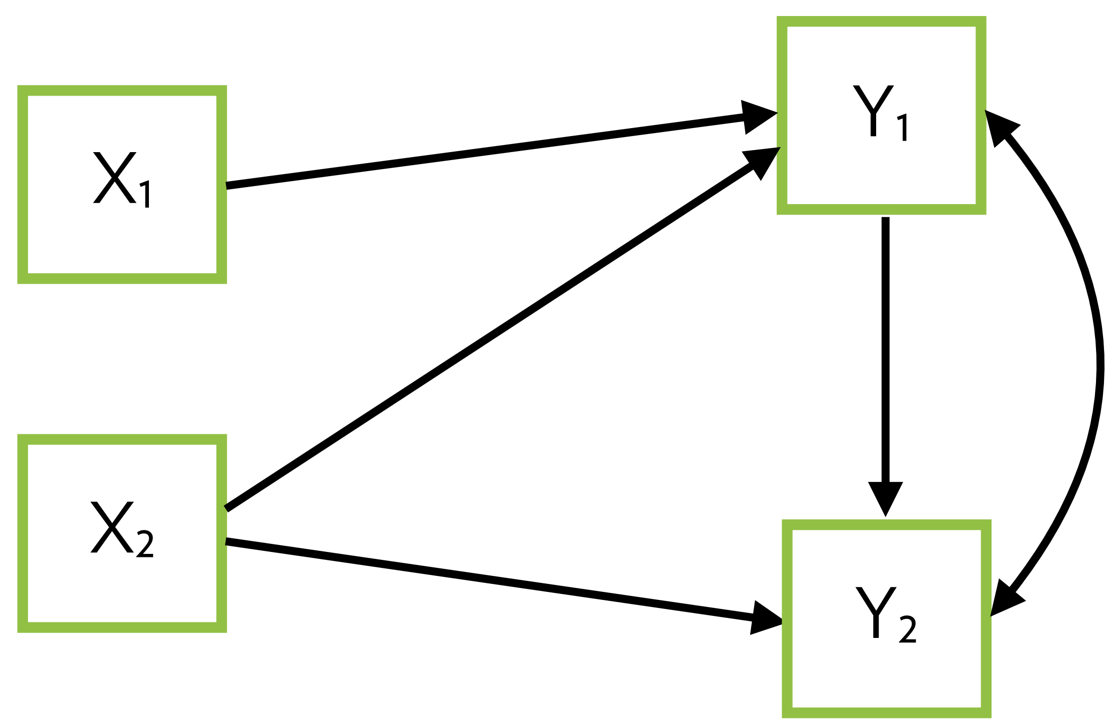
Correlation between Ys that also have a causal path (bow pattern):

- Identified if Y_1 has at least one X that is not an X of Y_2
- This X is called an “instrument”

Not identified



Identified (X_1 only goes to Y_1)

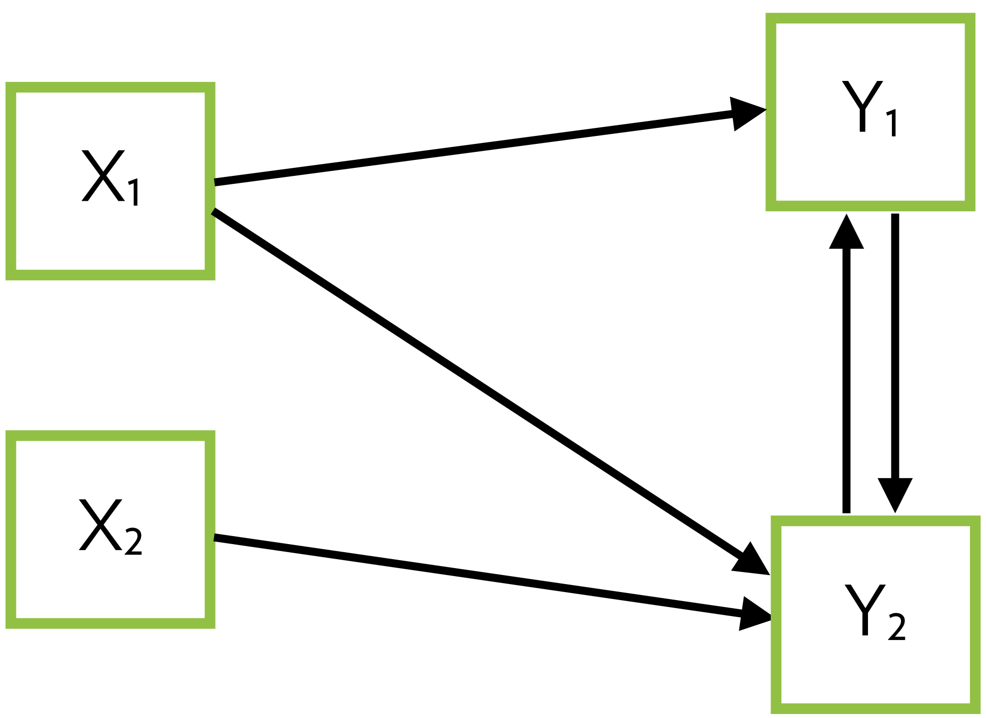


Nonrecursive

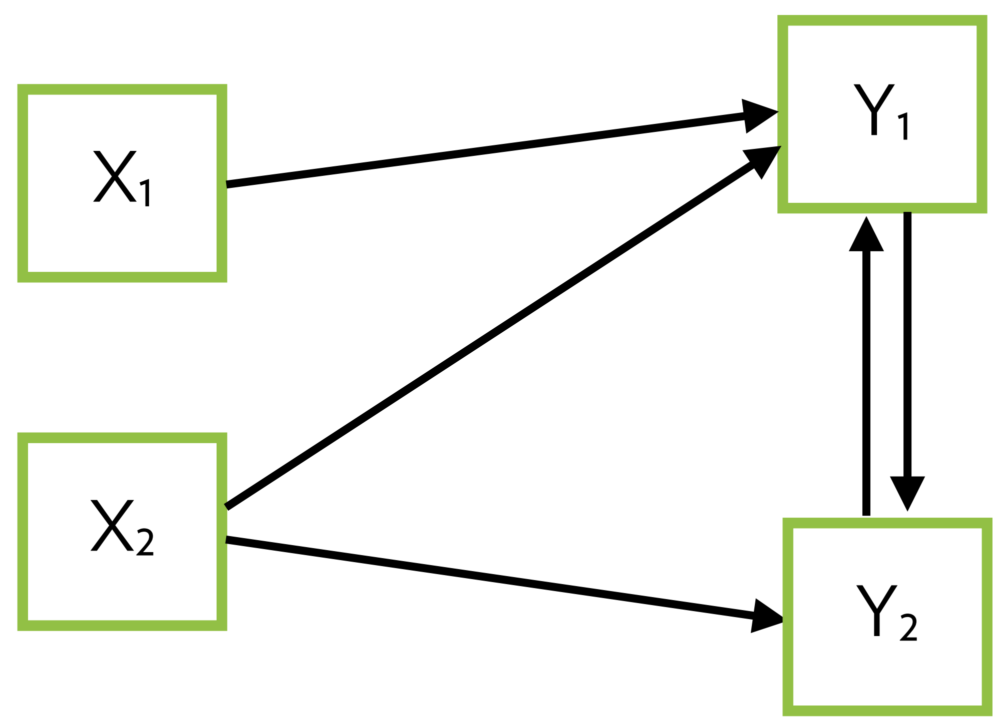
Feedback loop between Ys:

> Identified if **either** Y has at least one X that is not an X of the other Y

Identified (X_2 only goes to Y_2)



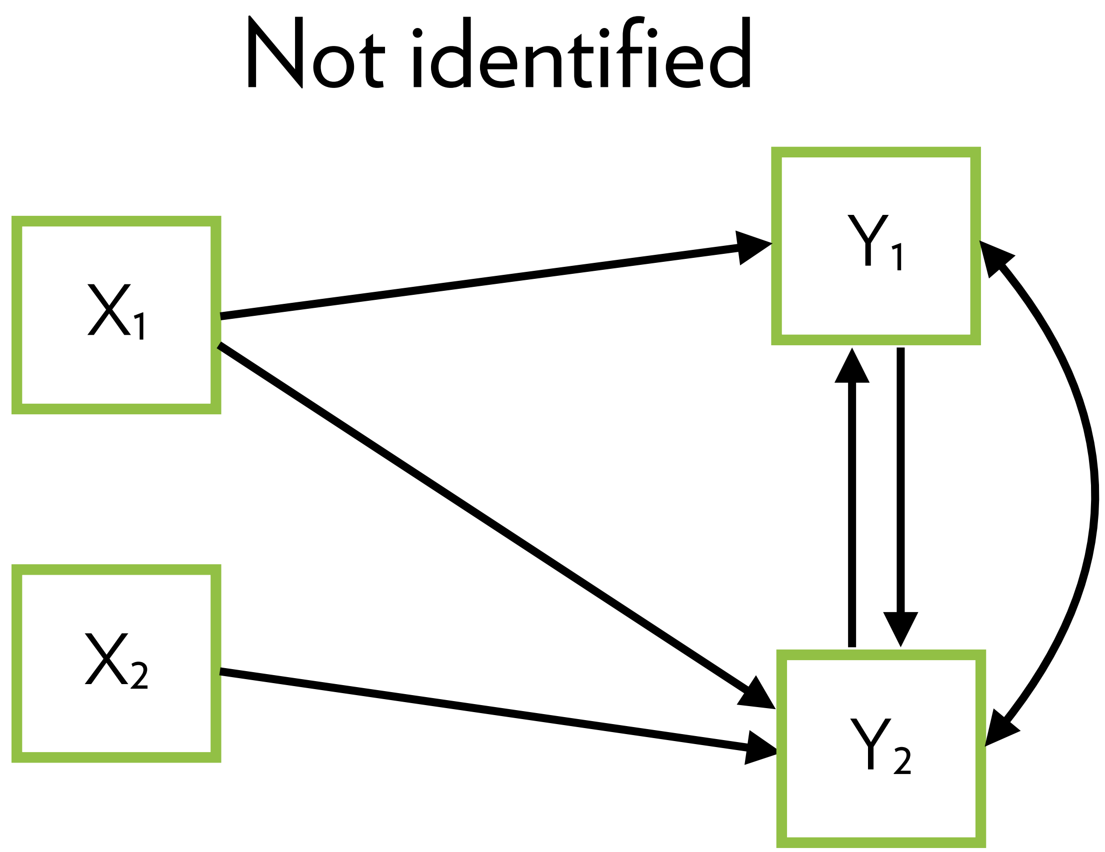
Identified (X_1 only goes to Y_1)

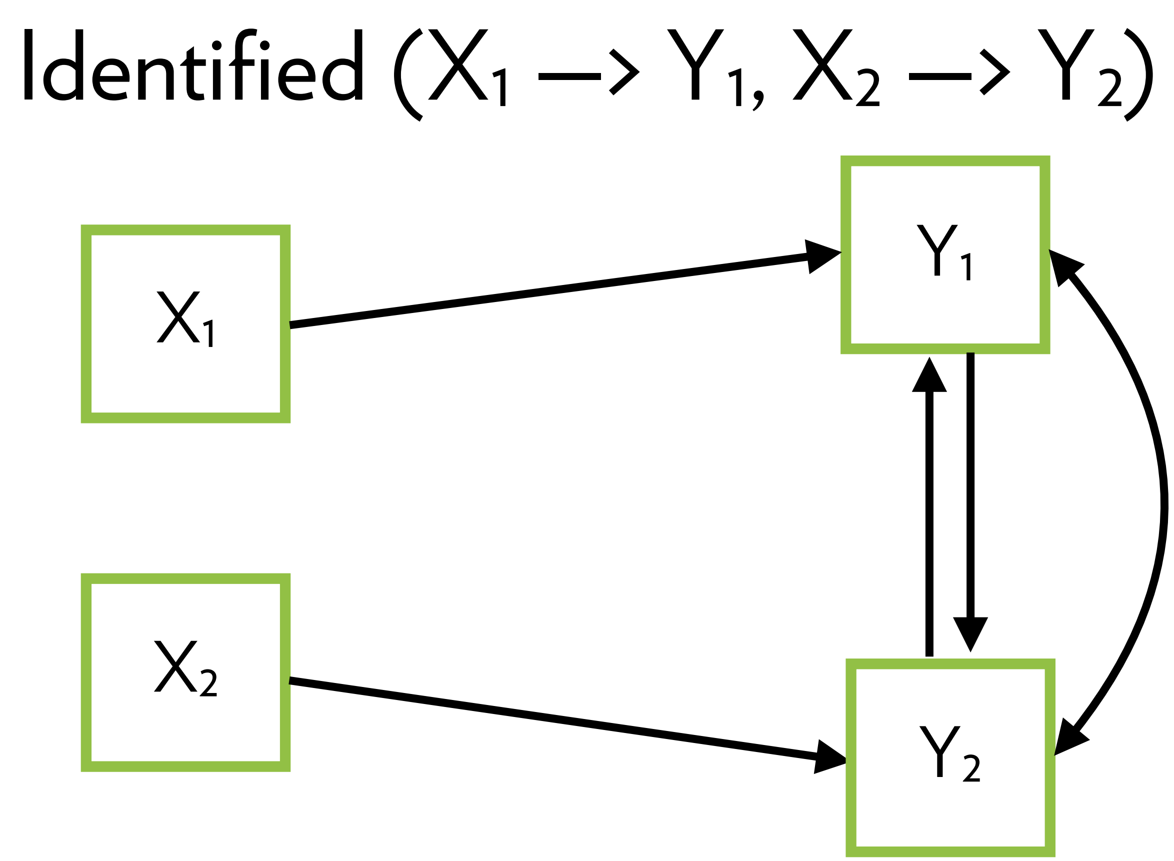


Nonrecursive

Feedback loop and correlation between Ys:

Identified if **both** Ys have at least one X that is not an X of the other Y

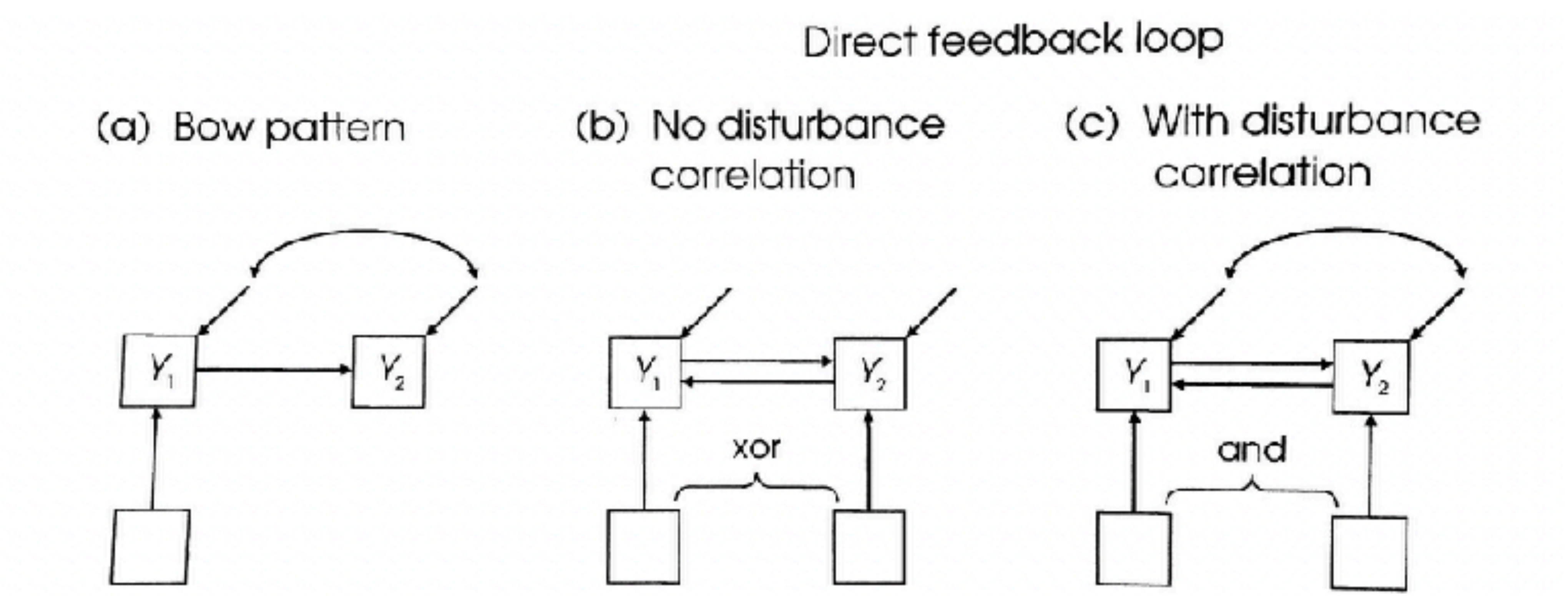




Nonrecursive

Other requirements:

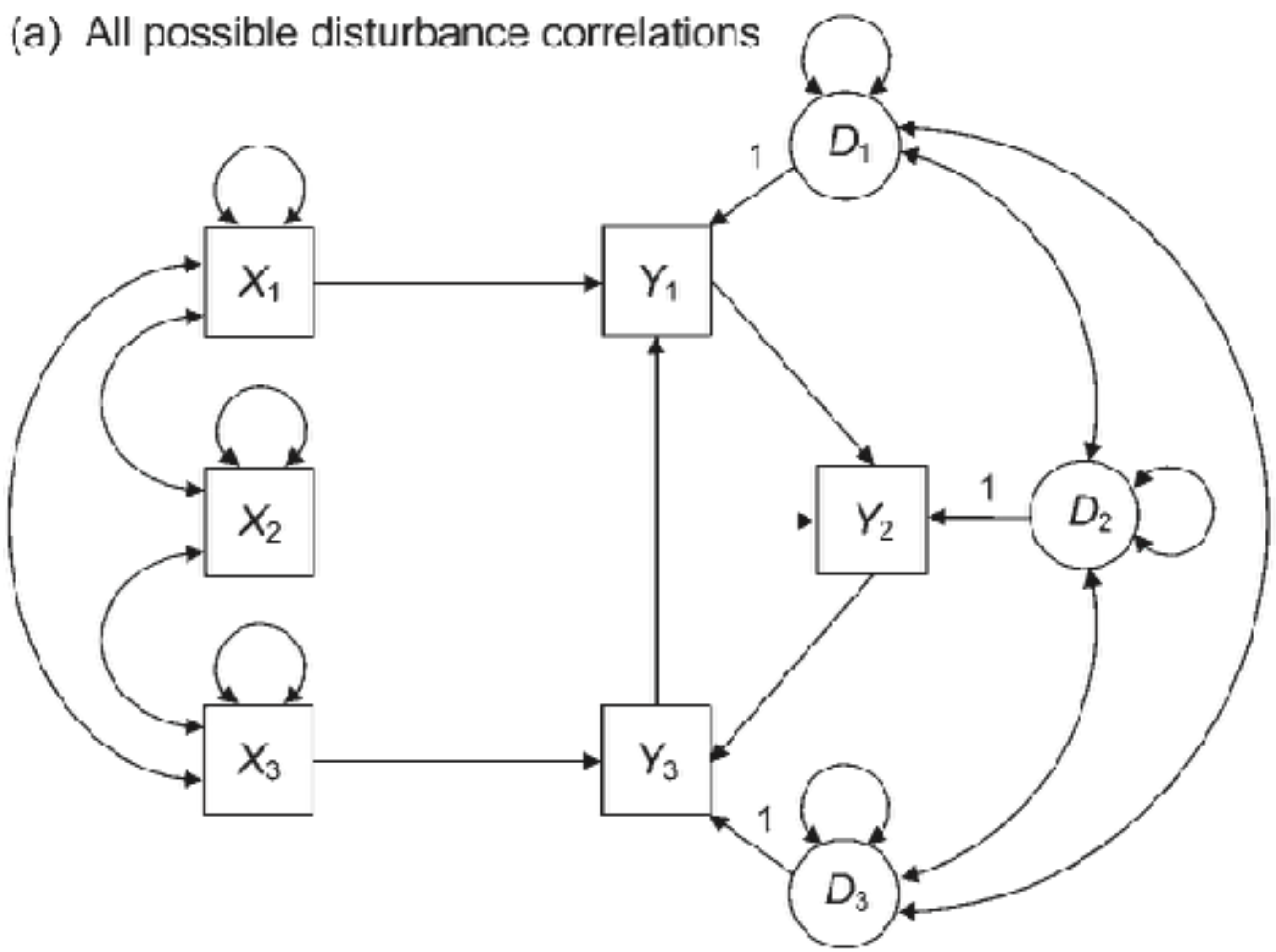
- Neither of the Ys causes the instrument X
- No other variable (indirectly) causes both the instrument X and the other Y
- The instrument X can be a correlation instead of a cause



Nonrecursive

Another example:

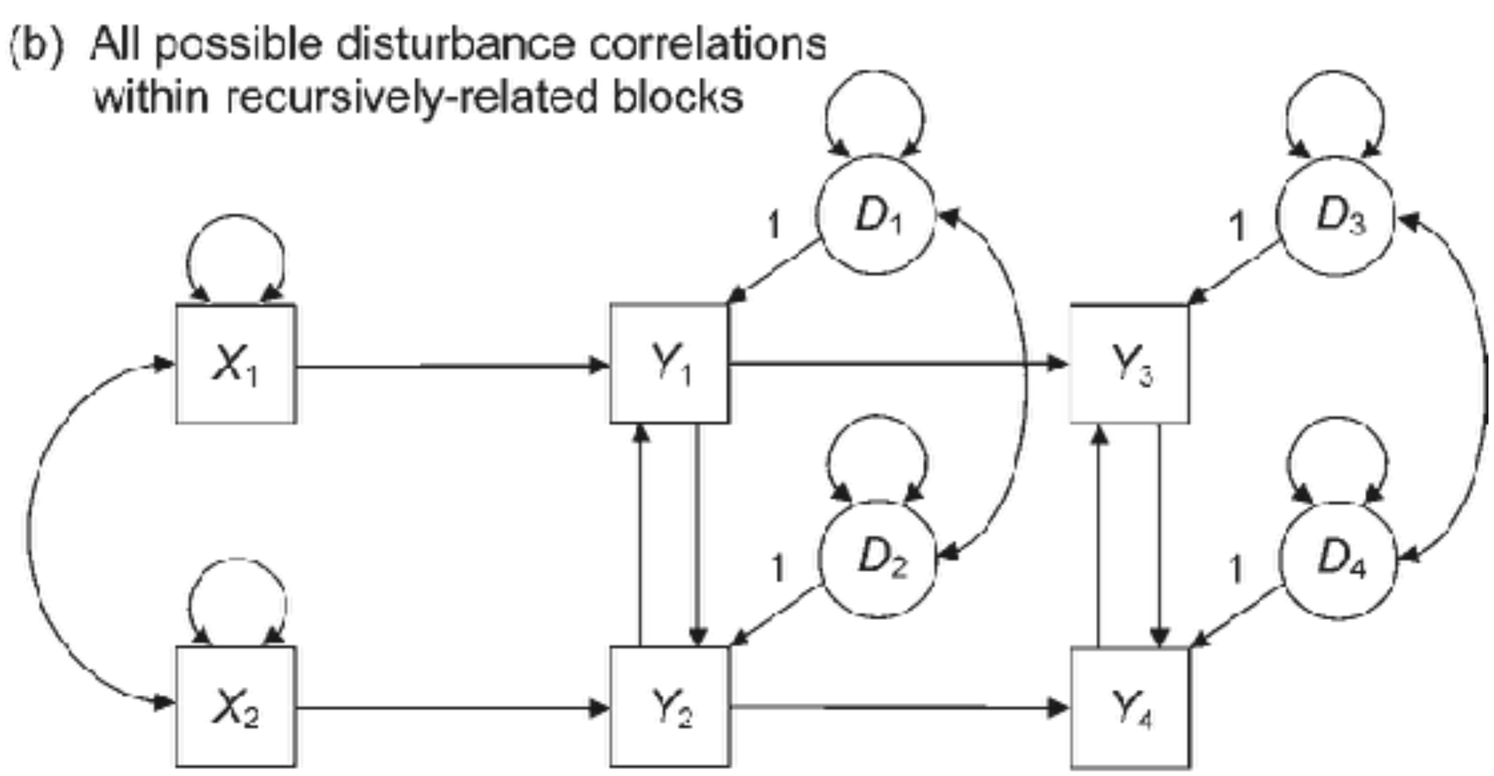
What are the instruments for Y_1 -> Y_2? For Y_2 -> Y_3? And for Y_3 -> Y_1?



Nonrecursive

Another example:

What are the instruments for Y_1 -> Y_2? Y_2 -> Y_1? Y_3 -> Y_4? Y_4 -> Y_3?



Not identified?

What if your model is not identified?

This model cannot be tested!

Solutions:

- Remove effects to make the model identified (if theoretically justifiable)
- Add additional instruments (if theoretically sound)

“It is the mark of a truly intelligent person to be moved by statistics.”

T H A N K S !

George Bernard Shaw