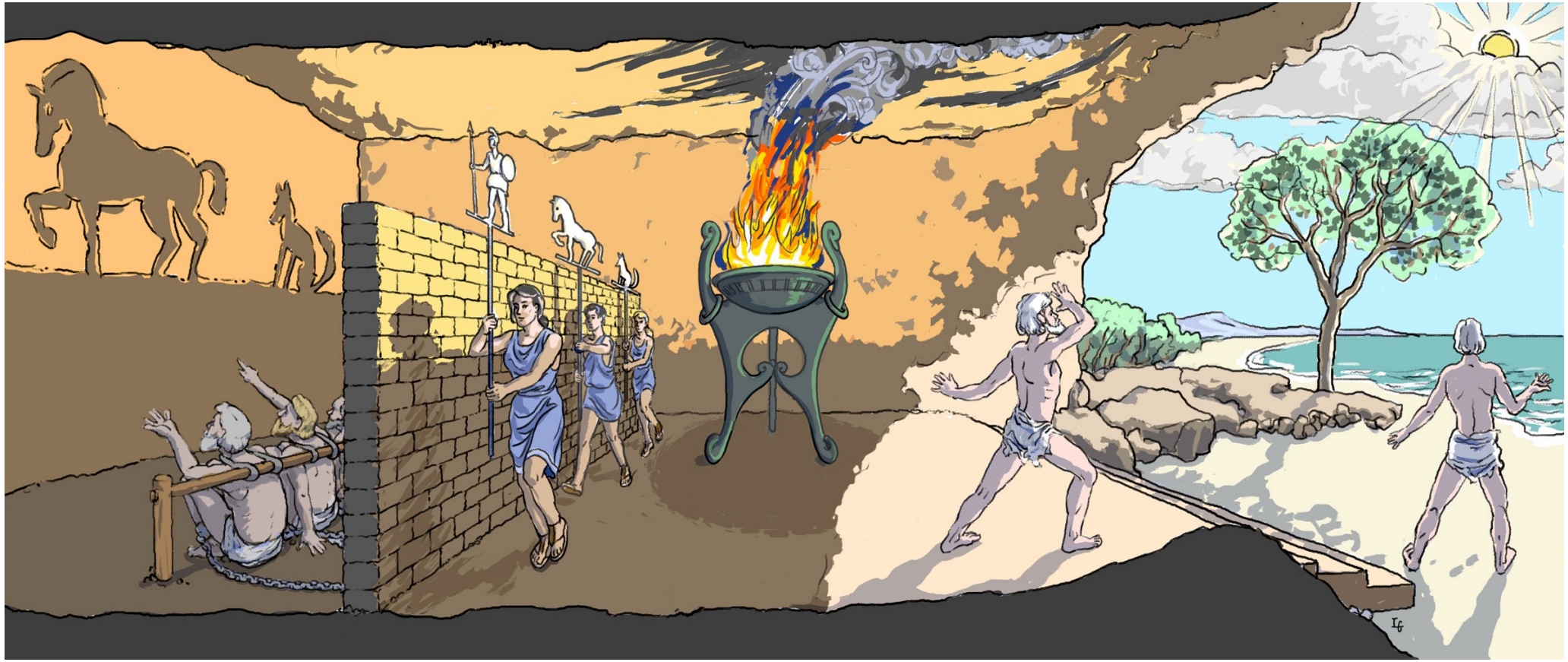
Introduction to Structural Equation Modeling: Theory

April 26th, 2024

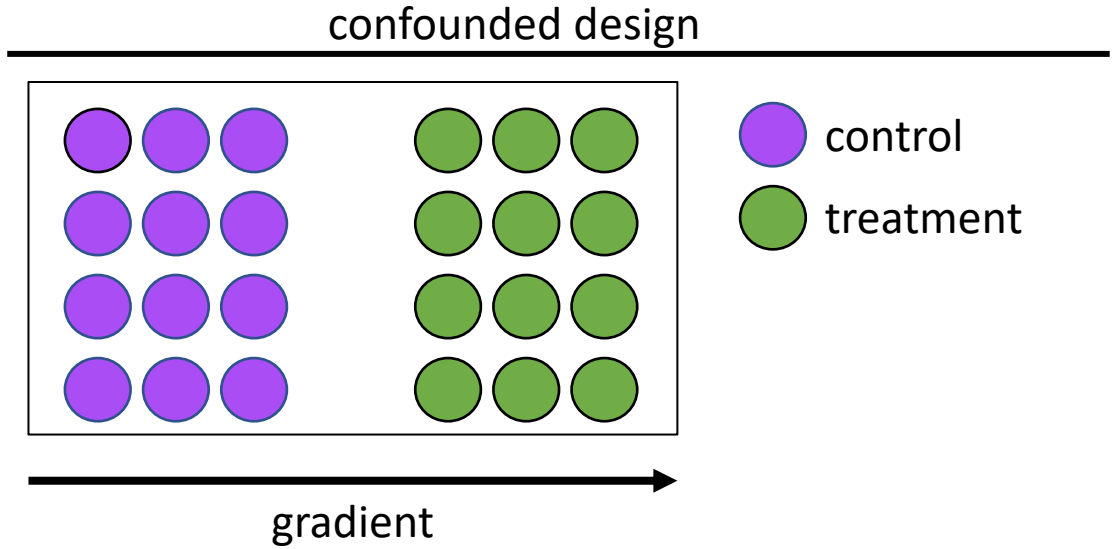
Jordan Croy

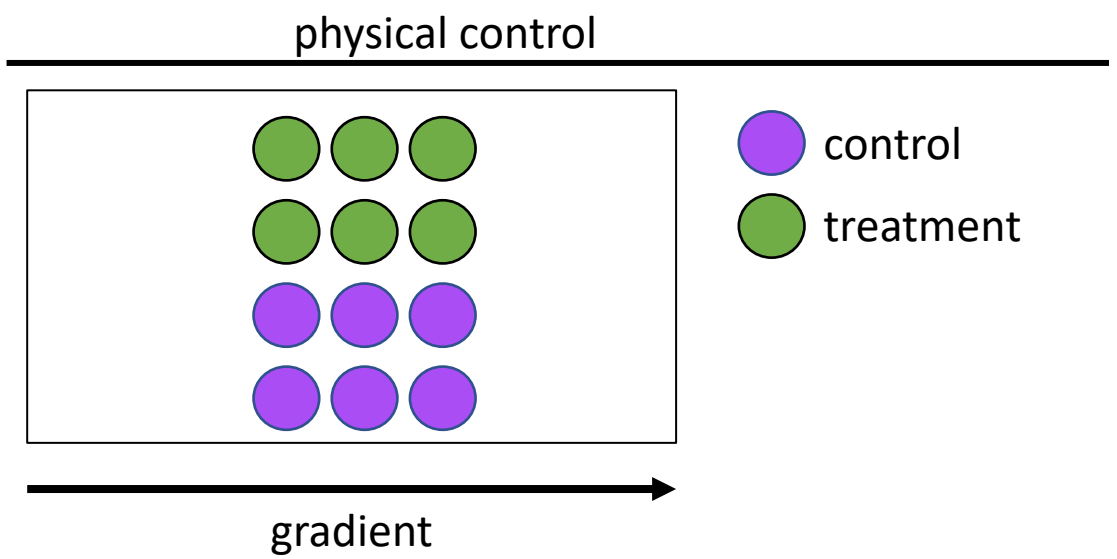
Theory: SEM Philosophy

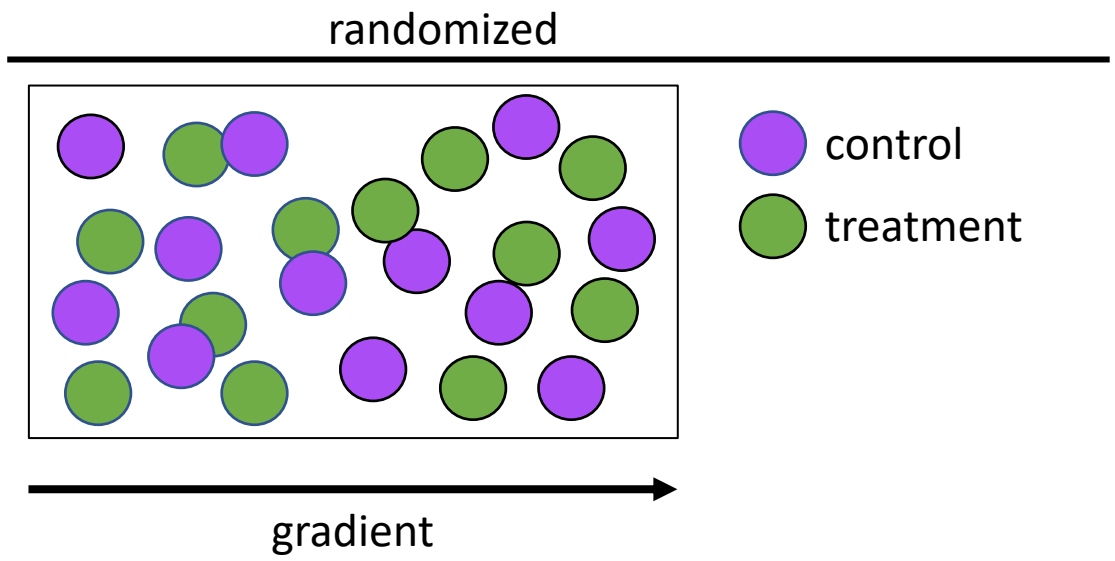
1. causal structures produce correlation structures
2. multiple causal structures can produce the same correlation structure
3. prior knowledge can be used to generate hypotheses about causal structures

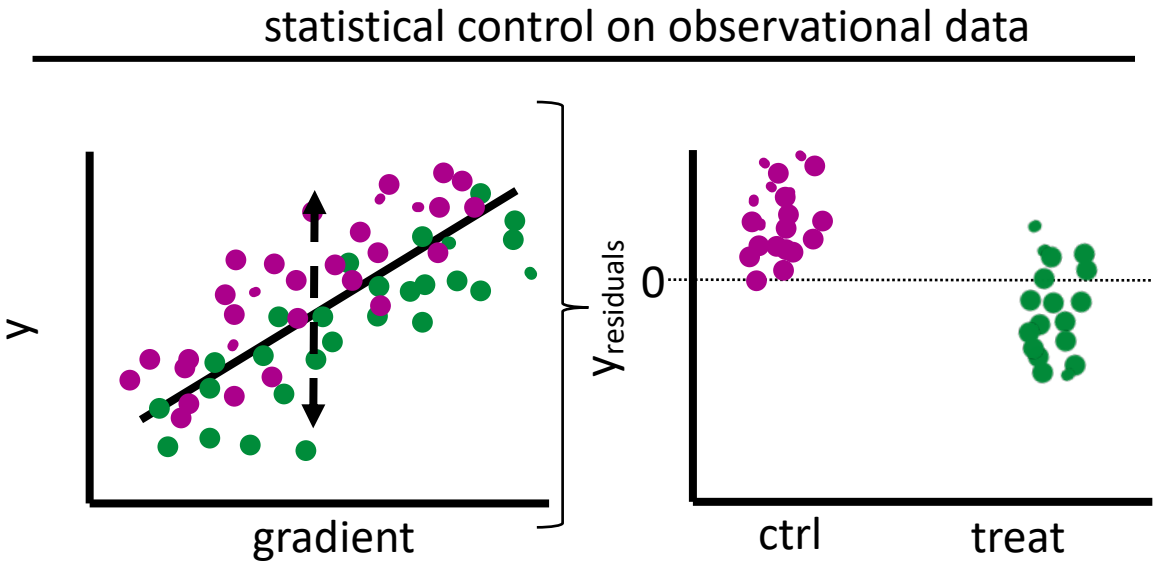


Theory: determining causality

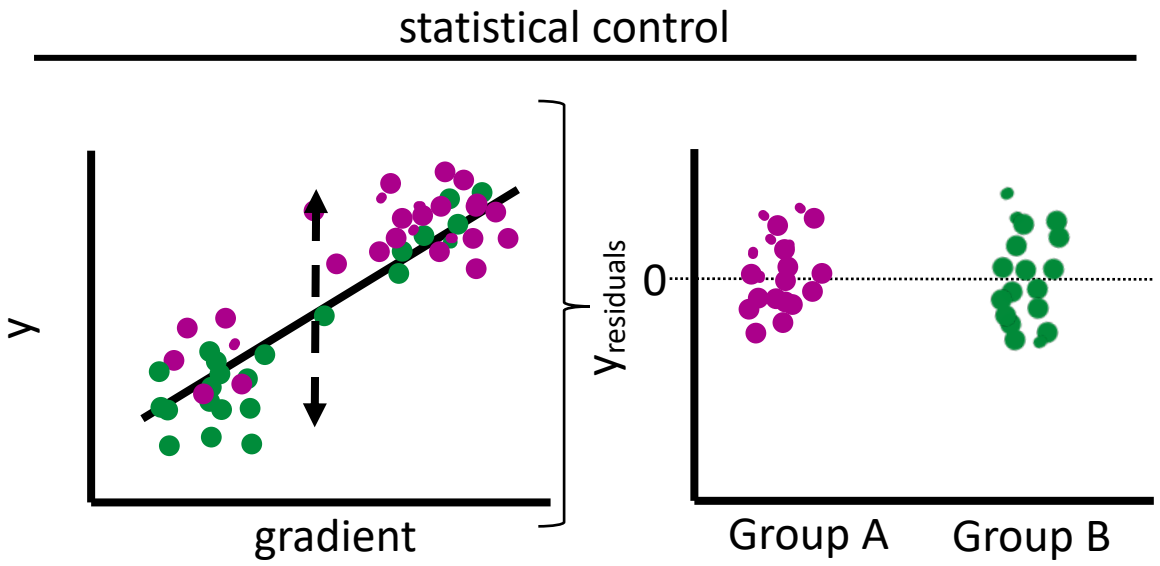








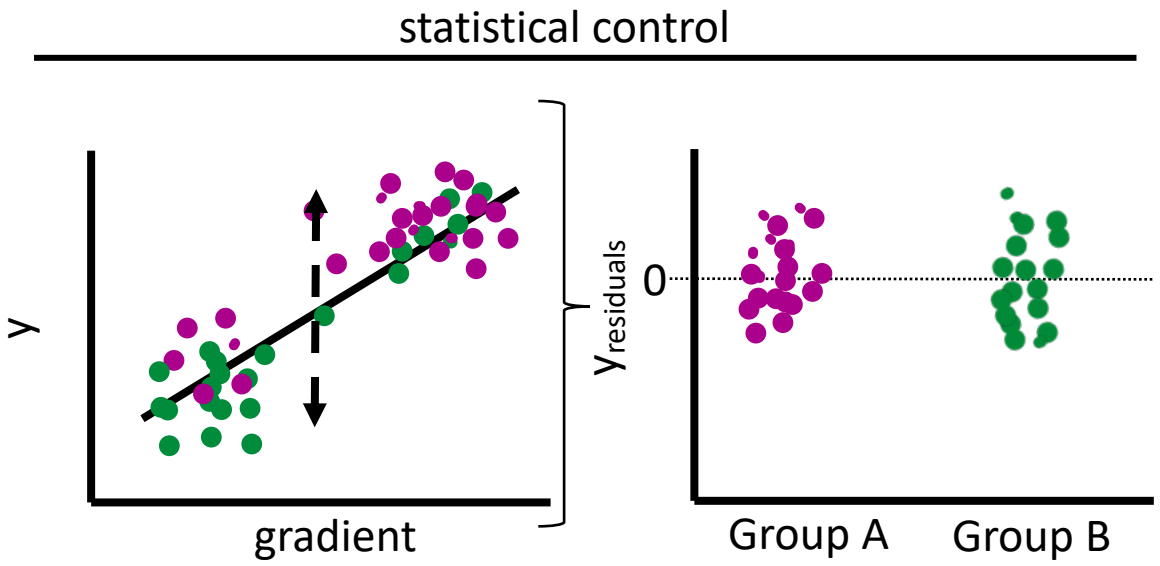
Theory: determining causality



What if there was no difference between two groups, but one happened to be concentrated on one end of a gradient and the gradient influenced y? The average of Group A in this case would be higher than Group B.

What would be the regression model syntax to test for treatment effects? In other words, how do we ***statistically control*** for the effects of gradient on y to assess treatment effects on y?

Theory: determining causality



What if there was no difference between two groups, but one happened to be concentrated on one end of a gradient and the gradient influenced y? The average of Group A in this case would be higher than Group B.

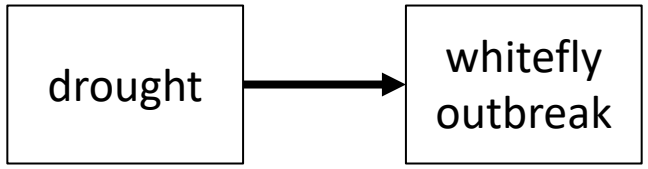
What would be the regression model syntax to test for treatment effects? In other words, how do we ***statistically control*** for the effects of gradient on y to assess treatment effects on y?

y ~ gradient + treatment

Without gradient in the analysis, you might find Group A to have higher y values than Group B.

Theory: directed acyclic graphs (DAGs)

these are causal path diagrams depicting the causal effect of one variable on another variable (we do this all the time)



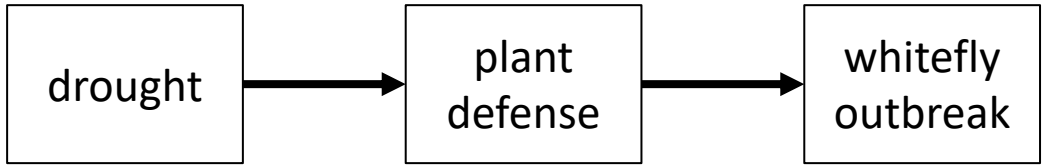
this is read as ”drought causes whitefly outbreak” and not “whitefly outbreak causes drought”

if you believe that, in the strictest sense, “correlation does not equal causation,” then you would not be able to rule out the possibility that whitefly outbreaks cause drought.

A correlation coefficient then would leave you with two possible interpretations. This is where SEM philosophy is useful and practical. The direction of causality is known through reasoning and prior experience, allowing us to interpret correlations as causes

Theory: directed acyclic graphs (DAGs)

plant defense in this DAG is a *mediator*, meaning, the effect of drought on whitefly outbreaks is entirely due to changes in plant defenses

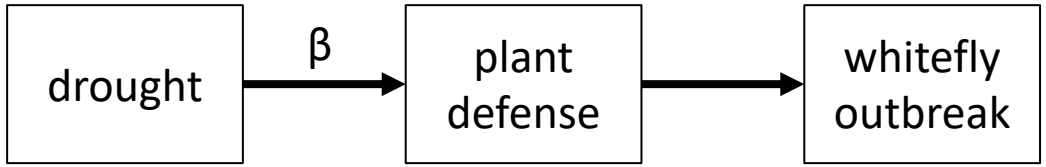


How to read this DAG–
drought is a *direct cause* of plant defense
plant defense is a *direct cause* of whitefly outbreaks
drought is an *indirect cause* of whitefly outbreaks
plant defense is a *causal child* of drought (and drought is a *causal parent* of plant defense)
whitefly outbreaks are a *causal descendant* of drought (and drought is a *causal ancestor* of whitefly outbreaks)

note that these terms are relative to your DAG—e.g., in the previous DAG (without plant defense), outbreaks were directly caused by drought.

Theory: directed acyclic graphs (DAGs)

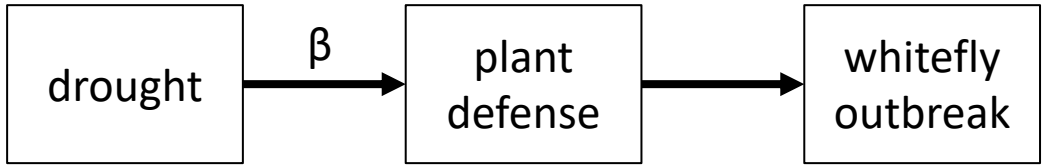
let's illustrate why DAGs are useful in general for multivariate regression models



What is the regression model syntax for estimating drought effects on plant defense (β)?

Theory: directed acyclic graphs (DAGs)

let's illustrate why DAGs are useful in general for multivariate regression models



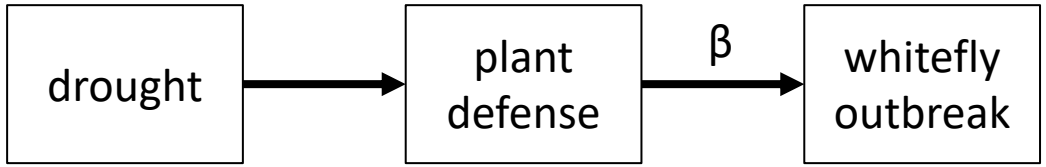
What is the regression model syntax for estimating drought effects on plant defense (β)?

plant defense ~ drought

Why?

Theory: directed acyclic graphs (DAGs)

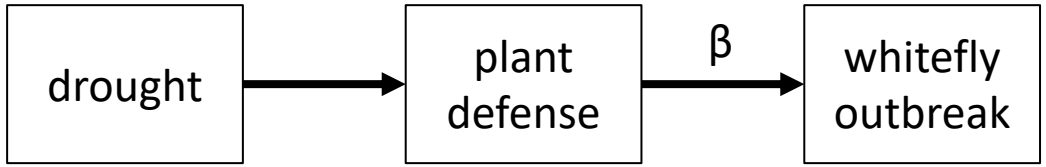
let's illustrate why DAGs are useful in general for multivariate regression models



What is the correct regression model for estimating the direct effect of plant defenses on whitefly outbreaks (β)?

Theory: directed acyclic graphs (DAGs)

let's illustrate why DAGs are useful in general for multivariate regression models

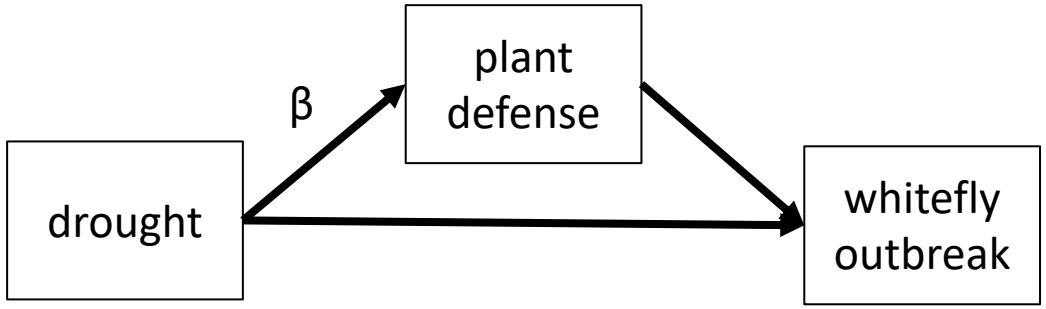


What is the correct regression model for estimating the direct effect of plant defenses on whitefly outbreaks (β)?

whiteflies ~ plant defense

Theory: directed acyclic graphs (DAGs)

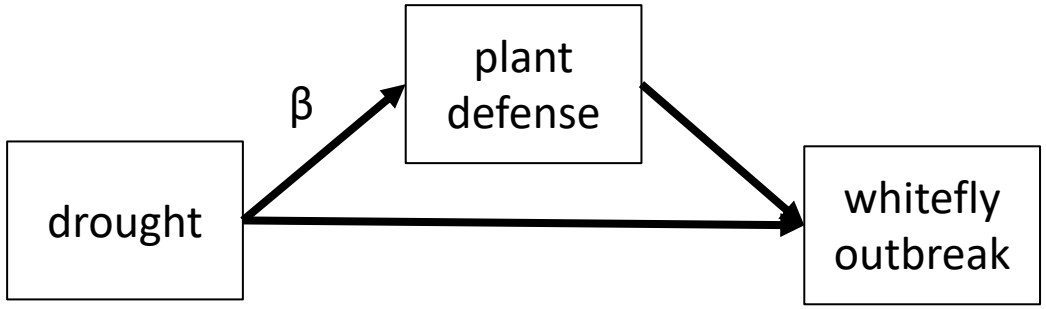
let's illustrate why DAGs are useful in general for multivariate regression models



How about now? What is the regression model syntax for estimating drought effects on plant defense (β)?

Theory: directed acyclic graphs (DAGs)

let's illustrate why DAGs are useful in general for multivariate regression models

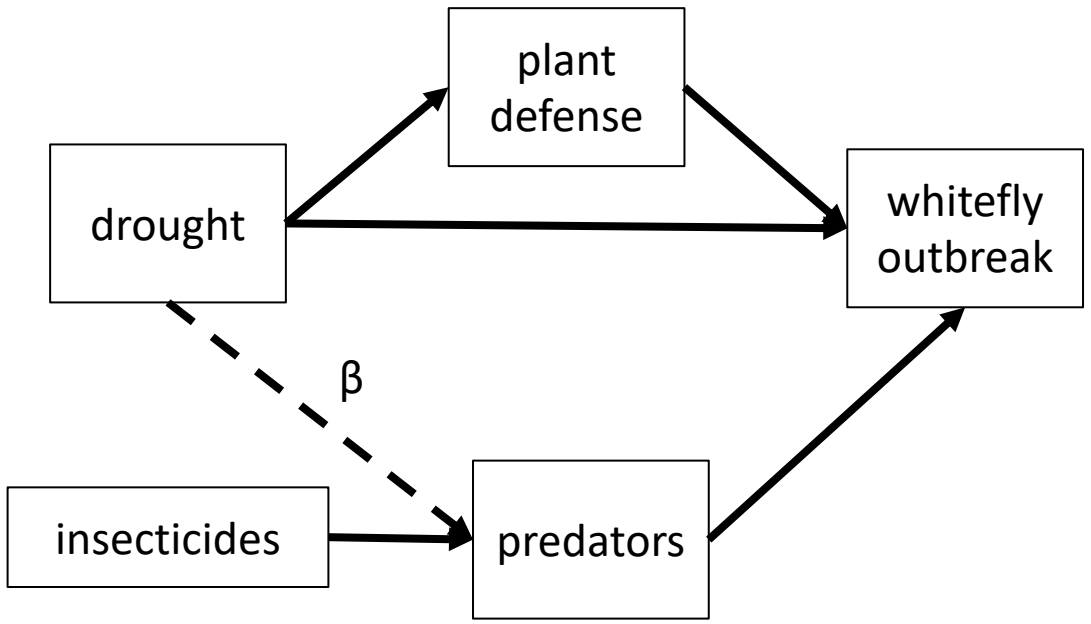


How about now? What is the regression model syntax for estimating drought effects on plant defense (β)?

plant defense ~ drought

Theory: directed acyclic graphs (DAGs)

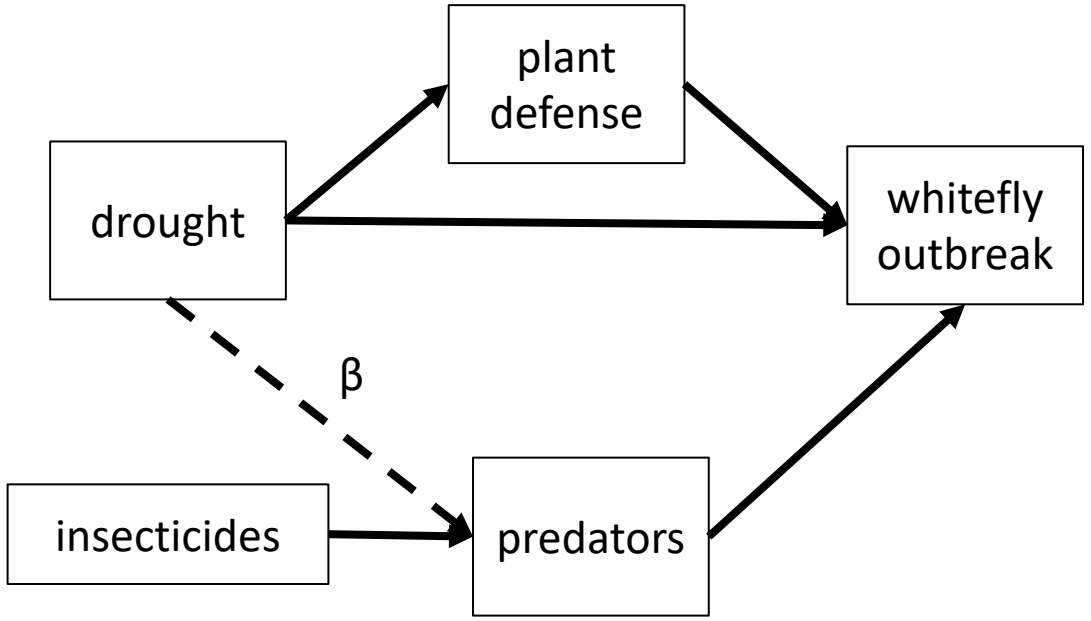
let's illustrate why DAGs are useful in general for multivariate regression models



Let's make it a little harder. Define the model that provides an estimate of the *direct* effect of drought on predators.

Theory: directed acyclic graphs (DAGs)

let's illustrate why DAGs are useful in general for multivariate regression models



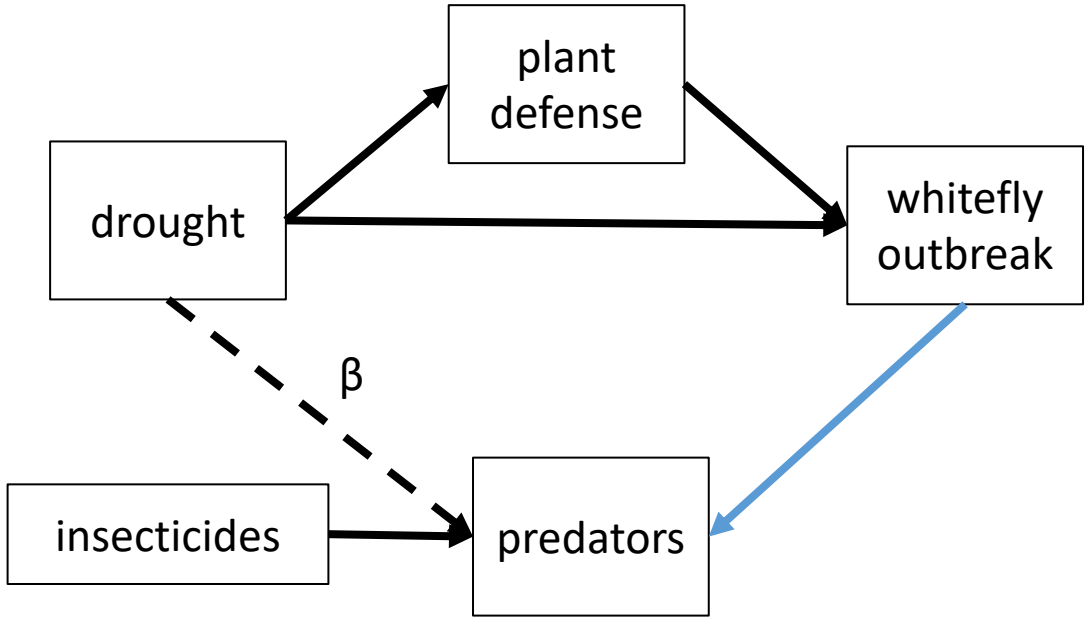
predators ~ drought

whitefly outbreak is a *collider variable*, meaning that several causal arrows enter into it, breaking the causal flow between antecedents (that is, there is no causal link between drought and predators mediated by plant defense and whitefly outbreak).

The analogy here is that collider variables are like a light switch in the OFF position. There is no current between drought and predators besides the direct linkage. However, by including whitefly outbreak in the model, we flip the switch to the ON position, leading to a drought-predator relationship with multiple causes, biasing our estimate of direct drought effects of predators.

Theory: directed acyclic graphs (DAGs)

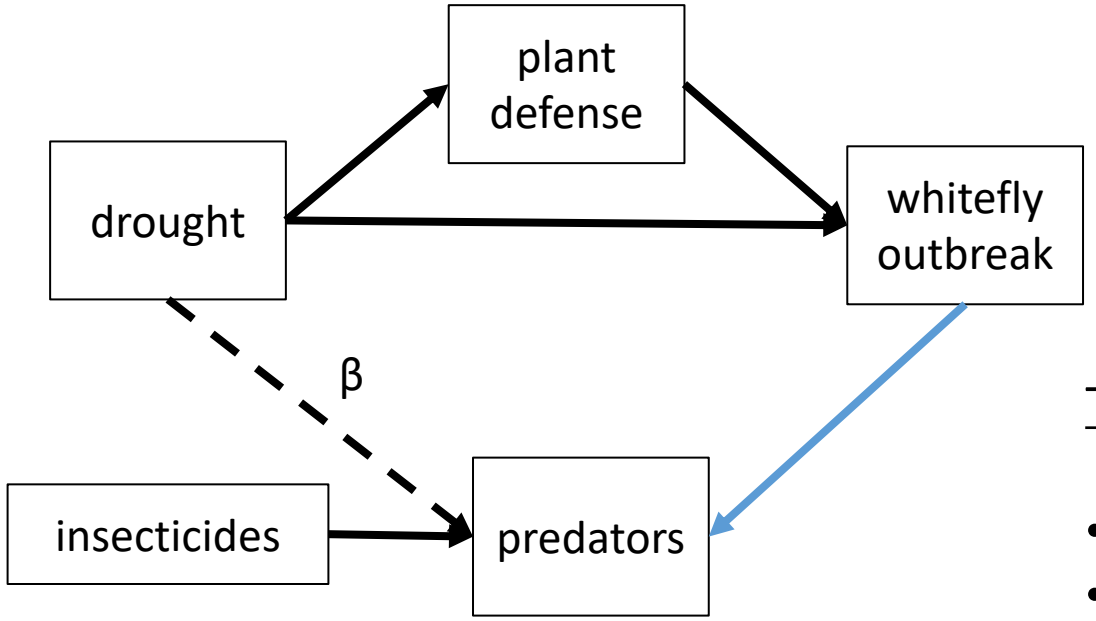
let's illustrate why DAGs are useful in general for multivariate regression models



How about now?

Theory: directed acyclic graphs (DAGs)

let's illustrate why DAGs are useful in general for multivariate regression models

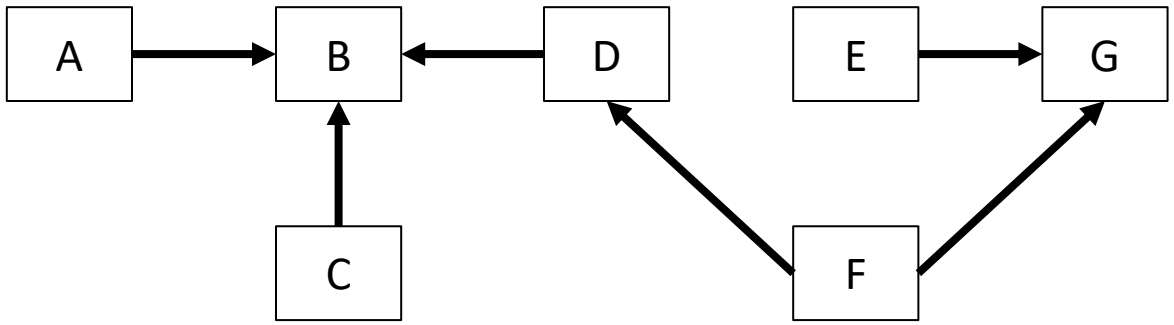


predators ~ drought + whiteflies

<u>Things to consider:</u>

- What would happen if we included insecticides?
- What would happen if we added plant defense to the model?
- What would the model look like if wanted to measure the total effect of drought on predators?
- Table 2 Fallacy and model selection

Theory: directional connection and separation



D⫫G|F –> D is d-separated from G *given (or conditioned on)* F
A⫫C|ϕ –> A is d-separated from C *unconditionally*
B⫫F|D –> B is d-separated from F *given D*
C⫫D|ϕ –> C is d-separated from D *unconditionally*
etc…

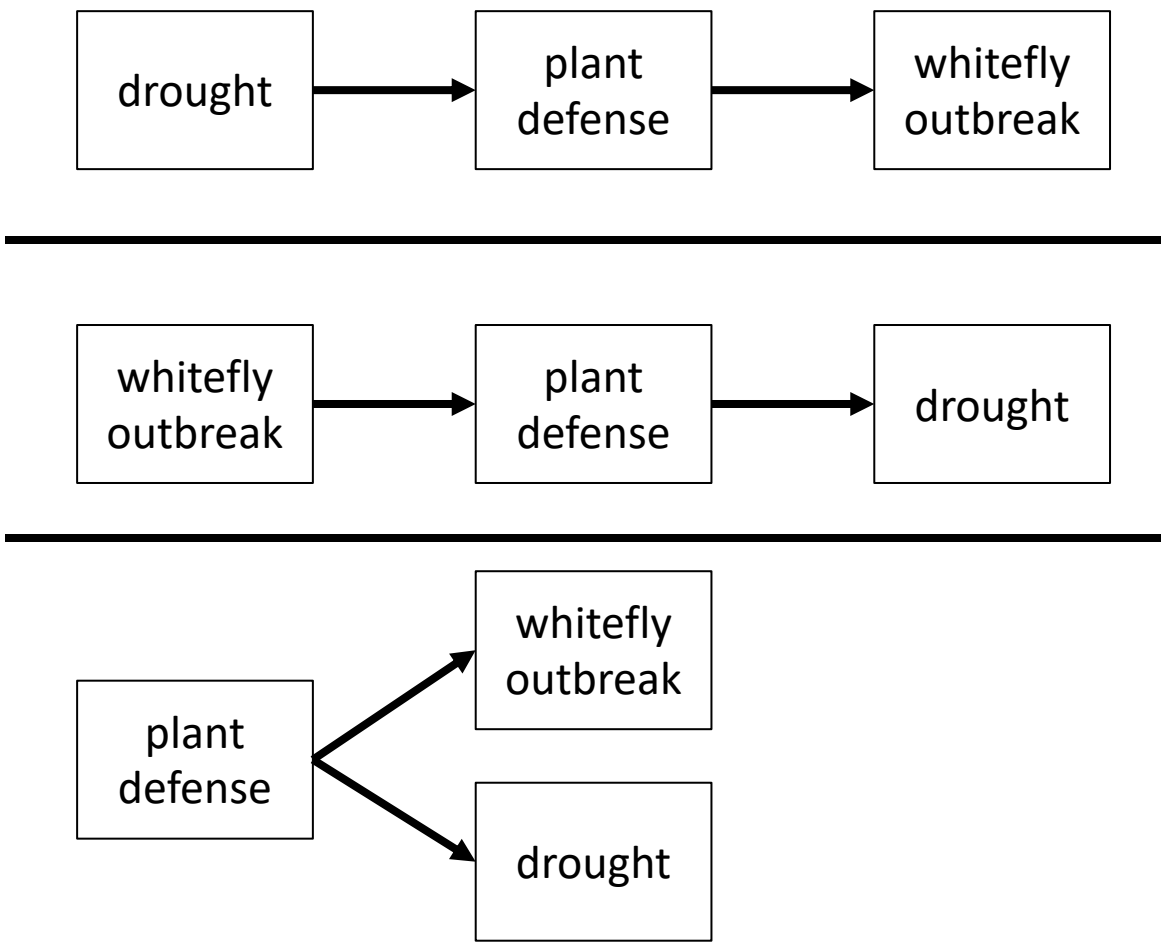
Up to this point, we have been practicing how to read DAGs, with an emphasis on describing relationships that are *directionally connected*.

We learned that DAGs contain a wealth of information about relationships among variables, and this information can be translated into regression models to quantify path coefficients accurately (i.e., without bias)

This process of creating independence claims is called ***directional separation (d-sep)*** and produces an exhaustive list of tests of independence. So, from the DAG, we get many claims of dependence and independence that we can test

Remember: the DAG is the puppeteer and the d-sep claims are the shadows on the cave wall—tests of these claims provide relative support for the hypothesized puppeteer

Theory: equivalent models



Let's revisit our simple DAG from earlier. We can construct different DAGs to compare ***alternative*** models of a given phenomena. However, multiple causal structures can produce the *same* correlation structures, and these are called ***equivalent models***.

The drought → plant defense → whitefly outbreak model has two additional correlation structures that we cannot distinguish among empirically, no matter how much data we collect. They each say whitefly outbreaks ⫫ drought | plant defense.

However, there is only one plausible option here, given our understanding of the biology. As your DAG grows, so does the number of equivalent models you will need to consider. And, it won't always be clear which direction the arrows should go

Theory: wrap-up

- DAGs describe a causal structure
- the causal structures cast correlational shadows
- correlational shadows can be converted into independence claims
- independence claims can then be converted into statements about probability distributions
- because causation and correlation are not 1:1, you must consider the equivalent models associated with your independence claims

Introduction to Structural Equation Modeling: Practice

May 17th, 2024

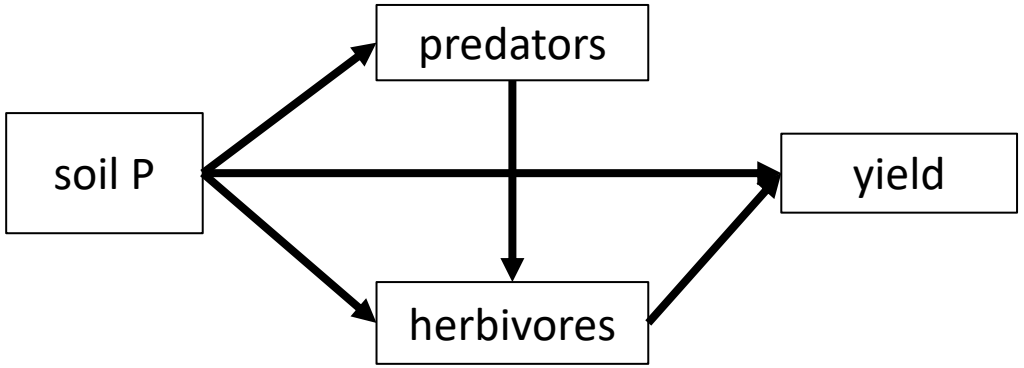
Jordan Croy

Theory: recap

- DAGs describe a causal structure
- the causal structures cast correlational shadows
- correlational shadows can be converted into independence claims
- independence claims can then be converted into statements about probability distributions
- because causation and correlation are not 1:1, you must consider the equivalent models associated with your independence claims

Practice: macro

1. create your DAG
2. generate a list of independence claims (this is called the *basis set*, and, thankfully, this process is automated in piecewiseSEM)
3. test the claims – significant p-values indicate violations of independence (Shipley's d-sep test)
4. Fisher's C often used a measure of goodness of fit of overall model (original equation shown to the right), there are better approaches now (log likelihood approach also shown right). Differences in log likelihood of hypothesized vs saturated model follows a chi-squared distribution
   - log likelihood for each sub model is summed and compared against the sum of log likelihood for each sub model of the saturated model—the difference between the two values follows a chi-squared distribution, producing P-values for your DAG
5. use AIC to compare competing hypotheses (optional)



1. yield ⫫ soil P | herbivores

p_i

$$C = -2\sum_{i=1}^{k} ln(p_i)$$

$$log(L_M(\theta|X)) = \sum_{i=1}^{k} log(L_i(\theta_i|X_i)$$

$$\chi^2_M L = -2(log(L(M_1)) - log(L(M_2))$$

$$AIC_{M1} = \sum_{i=1}^{v} AIC_i$$

Practice: less macro

1. create your DAG
2. specify/define each of the individual models (i.e., sub-models)
3. check/validate each sub-model
   - evaluate dependencies (nested, spatial, temporal)
   - inspect the distribution of the residuals
4. produce basis set via piecewiseSEM (does automatically when you assemble sub-models)
5. check the independence claims
   - include relationships that are deemed significant (Grace 2022) – improves model fit
   - Do NOT include the omitted links (Lefcheck) – consistent with theory-driven approach, but some cases require it
6. non-significant links
   - prune (Grace 2022)
   - I'd leave them in, and inspect the individual model again to make sure we have the power to test this relationship—if it is an unequivocal yes (to the power question), then I'd remove
7. summarize results
   - Fisher's *C* (or likelihood-based alternatives)
   - AIC (if comparing different causal structures)
   - standardized path coefficients
   - plot
8. organize results – strategies?

Practice: let’s go to R

Extra considerations

statistical power only limited by ability to run each individual model

generally, 15 observations / parameter you are estimating is a good

Keep in mind the equation to assess goodness of fit (note, Fisher's C can be used as a goodness of fit metric, but there are alternatives. But, I want to make a point about the limitations of Fisher's C):

What do you notice?

$$C = -2\sum_{i=1}^{k} ln(p_i)$$

high sample size can lead to greater probability of detecting a dependency that doesn't exist—leading to poor model fit

low sample size and overfitting can hinder ability to detect relationships that do exist—overshadowing significant dependencies and leading to good model fit

there is a likelihood-based alternative to *C* that avoids some of these issues

asymmetry in non-Gaussian models – need to specify directionality / exclude from d-sep test any claims involving non-normal error distributions (i.e., y1 | y2 (x1) != y2|y1 (x1) if error of y1 or y2 is non-Gaussian)