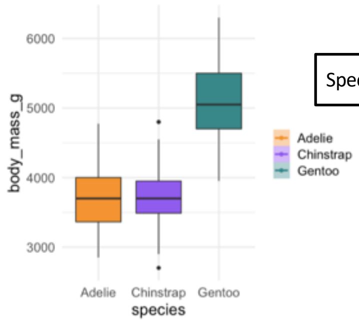
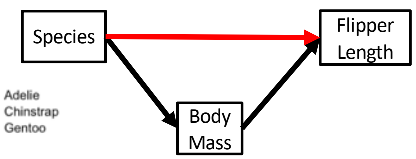
Categorical Exogenous Variables and SEM

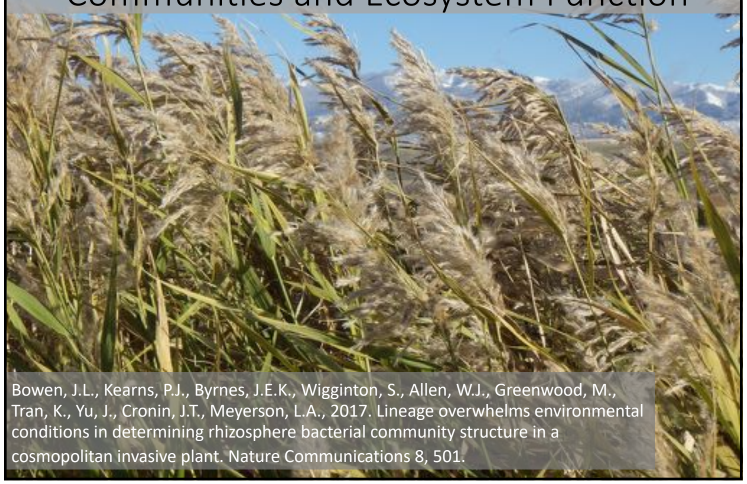




Categorizing our SEMs

1. Categorical Predictors
2. Categorical Predictors and Model Comparison
3. Categorical Variables and Shutting the Backdoor

Effect of *Phragmites* Genotype on Microbial Communities and Ecosystem Function

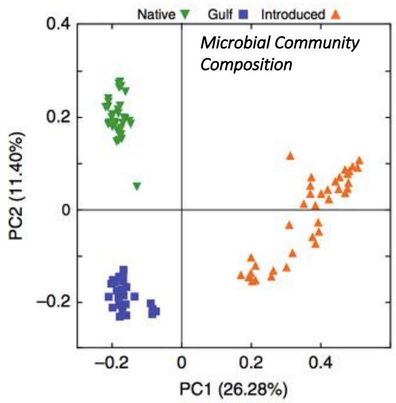


Bowen, J.L., Kearns, P.J., Byrnes, J.E.K., Wigginton, S., Allen, W.J., Greenwood, M., Tran, K., Yu, J., Cronin, J.T., Meyerson, L.A., 2017. Lineage overwhelms environmental conditions in determining rhizosphere bacterial community structure in a cosmopolitan invasive plant. Nature Communications 8, 501.

Multiple *Phragmites* Genotypes Across the US

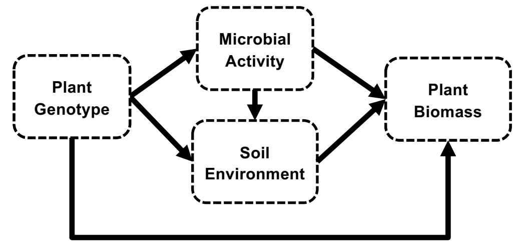
- Local environment should shape soil microbial communities
- These communities should regulate ecosystem function

But…genotype rules all!



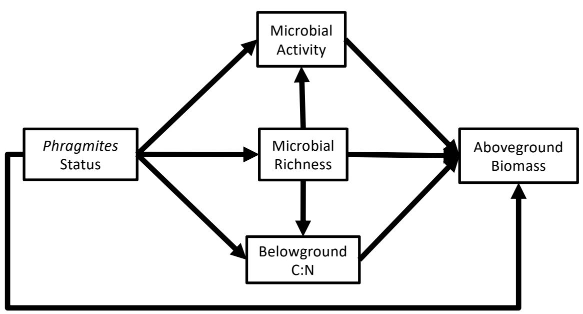
- How does this genotype effect translate to ecosystem function?
- Are we looking at direct or indirect effects on productivity?

What Drives Plant Performance?



- Three types of Phragmites, with multiple Genotypes: Native, Gulf Coast, Invasive: Fixed Effect
- Piecewise SEM with Mixed Models using Genotype as Random Effect

A Model from a Common Garden Experiment



Multiple genotypes per *Phragmites* status group – random effect!

The Four Submodels

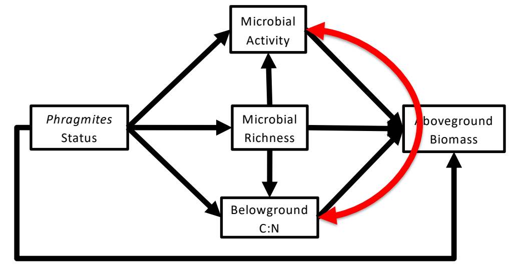
```
bowen <- read.csv("../data/bowen.csv")

###
# A categorical model
###
div_mod <- lme(observed_otus ~  status,
                      random =~ 1|Genotype,
                      data = bowen, method = "ML")

activity_mod <- lme(RNA.DNA ~ status + observed_otus,
                         random =~ 1|Genotype,
                         data = bowen, method = "ML")

c_mod <- lme(below.C ~ observed_otus +  status,
                    random =~ 1|Genotype,
                    data=bowen, method="ML")

biomass_mod <- lme(abovebiomass_g ~ RNA.DNA + observed_otus + belowCN + status,
                         random =~ 1|Genotype,
                          data = bowen, method="ML")
```

`method = "ML"` for accurate estimates of fixed effects

Build a pSEM

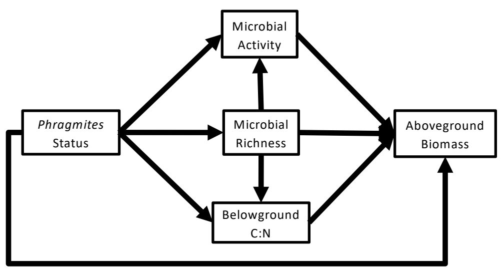
```
bowen_mod <- psem(
  div_mod,
  activity_mod,
  c_mod,
  biomass_mod,
  data = bowen
)
```

Assessing D-Separation



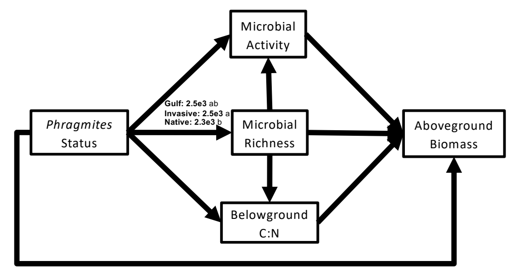

```
> dSep(bowen_mod)

          Independ.Claim Test.Type DF Crit.Value   P.Value
1 below.C ~ RNA.DNA + ...       coef 57 -0.1203216 0.9046515
```

Coefficients




```
> coefs(bowen_mod)

Warning message:
Categorical or non-linear variables detected. Please refer to documentation for
interpretation of Estimates!
```

ANOVA and Means Estimates

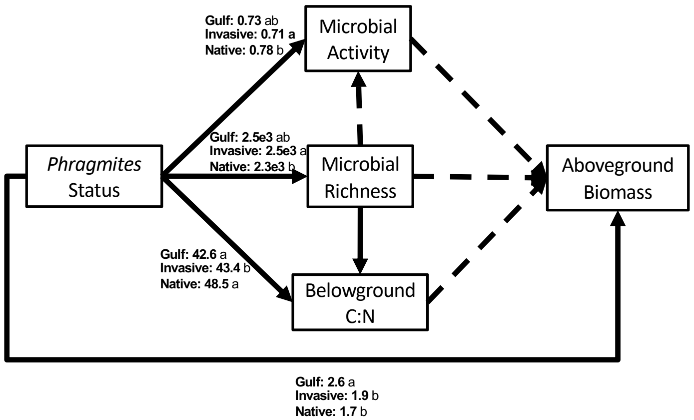


From `car::Anova`

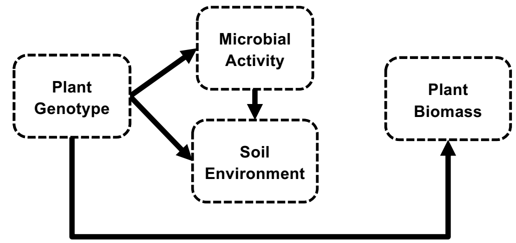
```
> coefs(bowen_mod)
       Response              Predictor  Estimate Std.Error DF Crit.Value P.Value
1   observed_otus                 status         -         -  2     5.9969  0.0499
2   observed_otus      status = native 2259.9211  101.4606 12    22.2739  0.0000
3   observed_otus    status = invasive 2530.0265   53.7327 12    47.0854  0.0000
4   observed_otus status = introduced 2533.8853  126.0395 14    20.1039  0.0000
```

From emmeans

Displaying Results of Posthocs on an SEM



Conceptual Answers!



- It's all about that genotype!
- With the exception of biomass, invasive are different from Gulf and East Coast genotypes

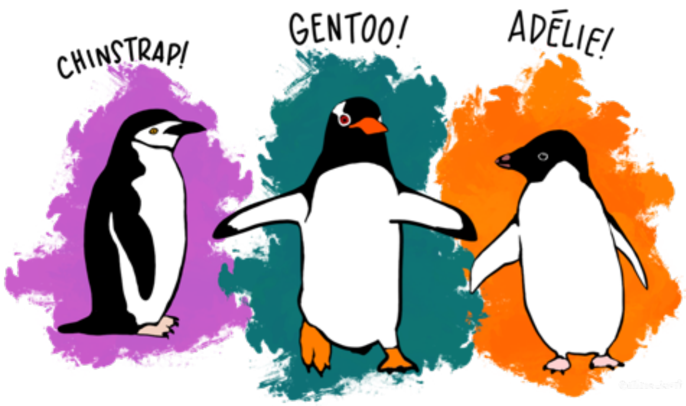
Categorizing our SEMs

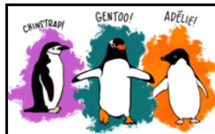
1. Categorical Predictors
2. Categorical Predictors and Model Comparison
3. Categorical Variables and Shutting the Backdoor

Let's Look at Some Penguins!

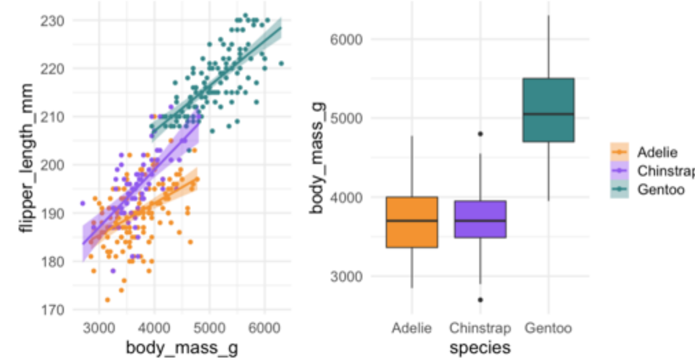
From data science superstar Alison Horst. All Illustrations are hers.

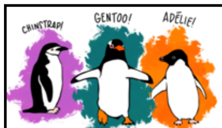
What Determines Flipper Length?



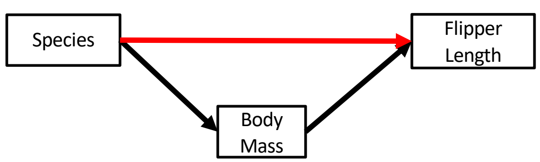


What Determines Flipper Length?

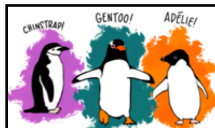




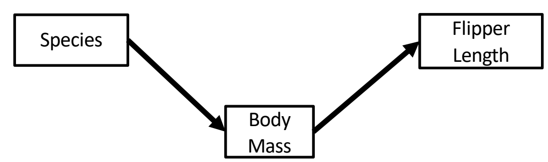
What Determines Flipper Length?




```
#fit a model with no direct species effect
size_mod <- lm(body_mass_g ~ species ,
               data = penguins)


flipper_mod <- lm(flipper_length_mm ~ body_mass_g,
               data = penguins)

penguin_fit <- psem(
  size_mod,
  flipper_mod)
```

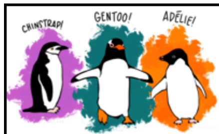


This Model…. Is not Good

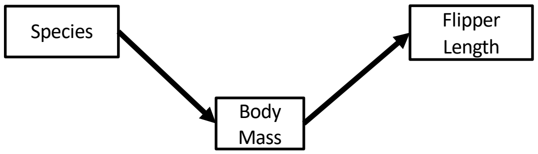



```
> fisherC(penguin_fit)
  Fisher.C df P.Value
1  157.519  2       0

> LLchisq(penguin_fit)
    Chisq df P.Value
1 159.434  2       0
```

Wait, df = 2?

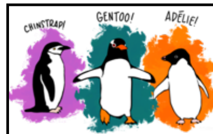


Only One Test for Categorical Variables

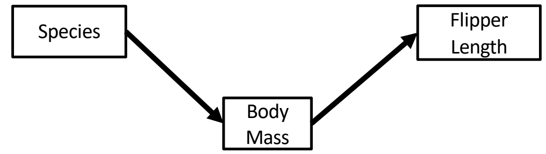



```
  Independ.Claim                                Test.Type DF Crit.Value       P.Value
1 flipper_length_mm ~ species + ...                 anova  2   101.0202 6.24099e-35 ***
```

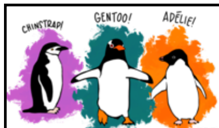
- How many DF? Do we use all factor levels coded as 1/0?
- We evaluate categorical variables with LRTs or F-Tests
- Thus, each category only contributes 2 DF (2*1 link) to our tests



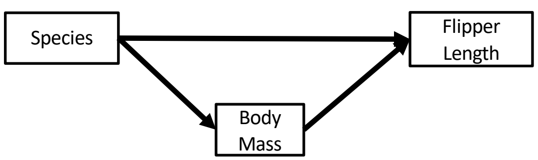
So, This Model.... Is not Good



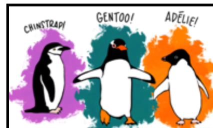

```
> fisherC(penguin_fit)
  Fisher.C df P.Value
1  157.519  2       0

> LLchisq(penguin_fit)
    Chisq df P.Value
1 159.434  2       0
```



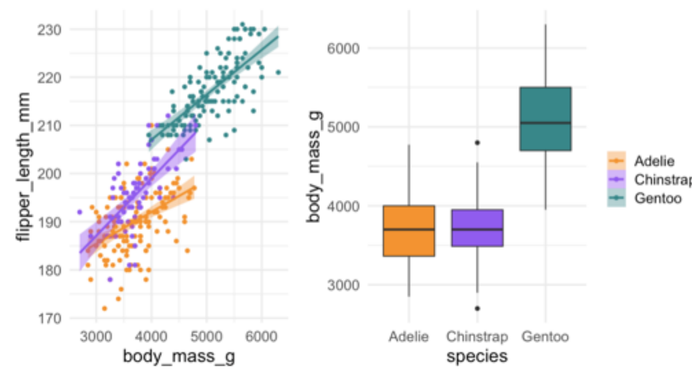
A Partial Mediation Model (Saturated)

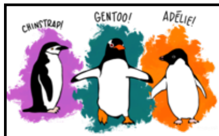



```
flipper_mod_sp <- lm(flipper_length_mm ~
                       body_mass_g + species,
                     data = penguins)

penguin_fit_species <- psem(
  size_mod,
  flipper_mod_sp)
```

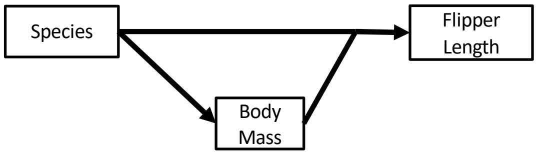


We might still ask, as that slope constant?

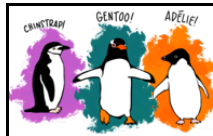




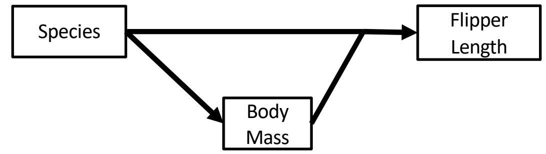
An Interaction Model




```
flipper_mod_int <- lm(flipper_length_mm ~
                        body_mass_g * species,
                      data = penguins)

penguin_fit_int <- psem(
  size_mod,
  flipper_mod_int)
```

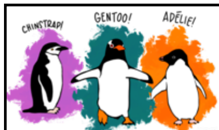


Comparing Models with AIC

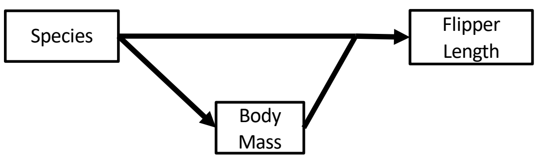



```
> anova(penguin_fit, penguin_fit_species, penguin_fit_int)
Chi-square Difference Test

         Df  AIC.AIC AIC.K AIC.n   Chisq Chisq.diff Df.diff P.value
 Model 1  2 7264.774     7   333 159.434
 Model 2  0 7109.340     9   333   0.000    159.434       2       0 ***
 Model 3  0 7104.757    11   333   0.000    159.434       2       0 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

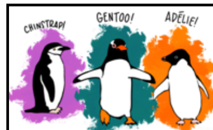
Best AIC by 5, and Interaction is Different from 0



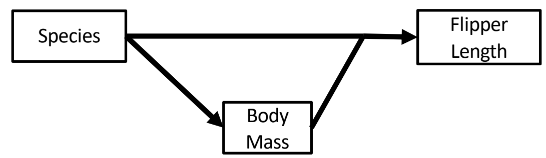
We are Fixing Coefs….




```
> coefs(penguin_fit_int)

Error in `[.data.frame`(data, , vars, drop = FALSE) :
  undefined columns selected
```

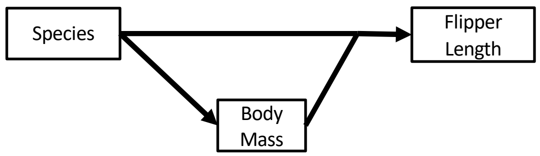


You Can Get Slopes and Effects with `emmeans`



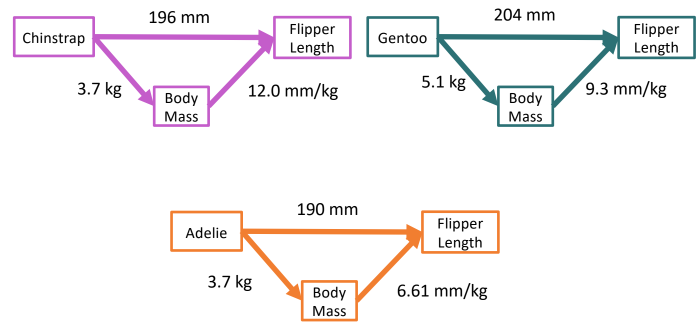

```
library(emmeans)


emtrends(flipper_mod_int,
         ~species,
         var = "body_mass_g")
```

Then Get Effect of Species on Mass and Flipper Length at Average Mass




```
#mass
emmeans(size_mod, ~species)

#avg mass
weight_list <- split(penguins$body_mass_g, penguins$species)
weight_vec <- sapply(weight_list, mean)

emmeans(flipper_mod_int, ~ species|body_mass_g,
        at = list(body_mass_g = weight_vec) )
```

Presentation: Separate Diagrams

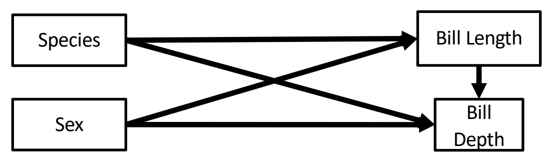


Species effects on flipper at average species mass

Future Directions

- Shoring up Interaction Effects and coefs
- Plotting categorical variables and interactions
- Categorical endogenous variables will require implementation of multinomial logistic regression within piecewiseSEM (but you can implement by hand?)

Exercise

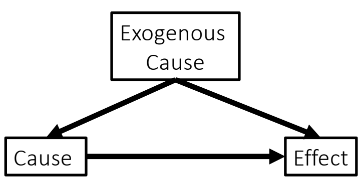


1. Is the effect of species and sex on bill depth partially or fully mediated by bill length?
2. Do species and sex interact?
3. Extra: Play with structure and interactions more.

Categorizing our SEMs

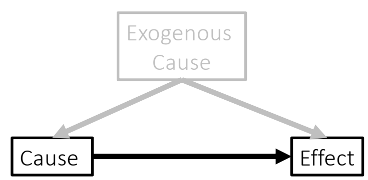
1. Categorical Predictors
2. Categorical Predictors and Model Comparison
3. Categorical Variables and Shutting the Backdoor

The Omitted Variable Bias Problem



- We assume that sampling means that omitted variables average to 0
  - Omission produces downward bias in SE of coefficients
- But, if omitted variables are correlated causally with a predictor, they likely are not averaged out
- This *will* bias your estimates
  - You will not know in what direction

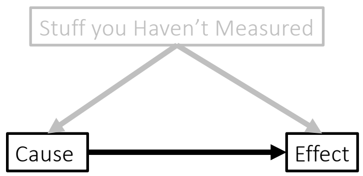
OVB and Causal Identification



This path is not causally identified

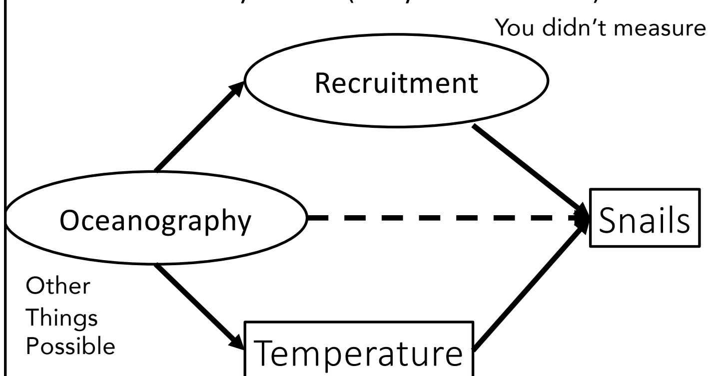
AGH! I Forgot to Measure That

- This happens to everyone
- Sometimes, you know what you didn't measure – or COULDN'T measure
- Sometimes, you don't!

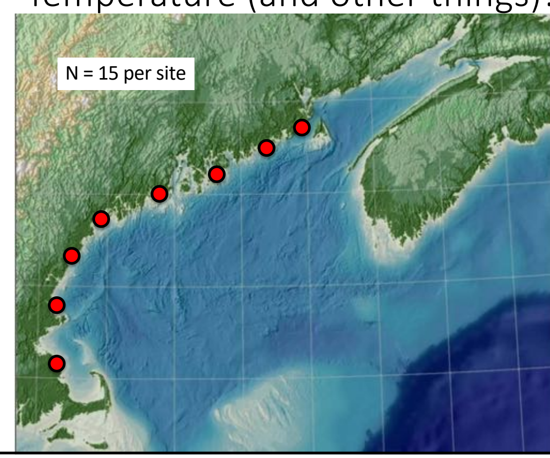


Let's Talk Snails

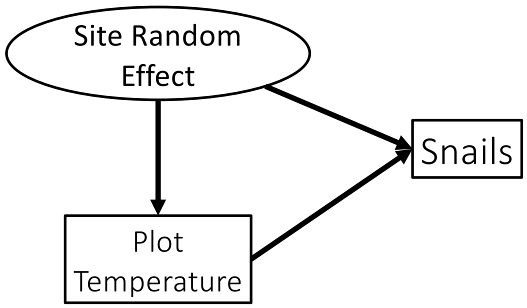
An Ecological Example: The System (as you know it)



Snail Sampling at Sites Varying in Temperature (and other things)!

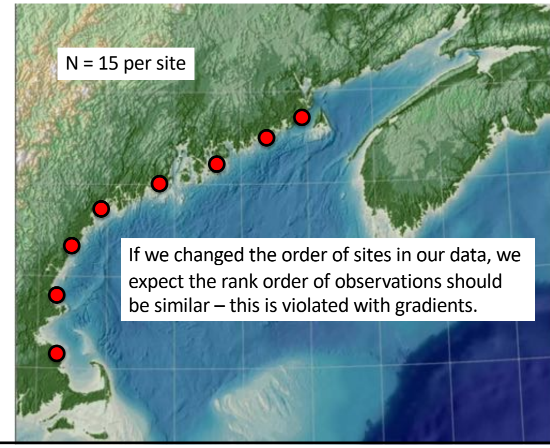


Your Naïve Random Effects Model

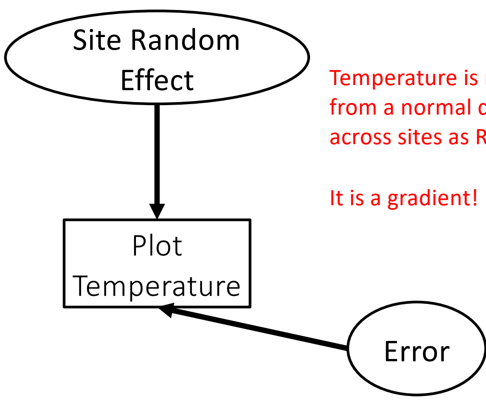



```
library(lme4)

temp_mod <- lmer(temperature ~ (1|Site), data = dat)

snail_mod <- lmer(snails ~ temperature + (1|Site),
                  data = dat)
```

Problem #1: Violation of Exchangeability



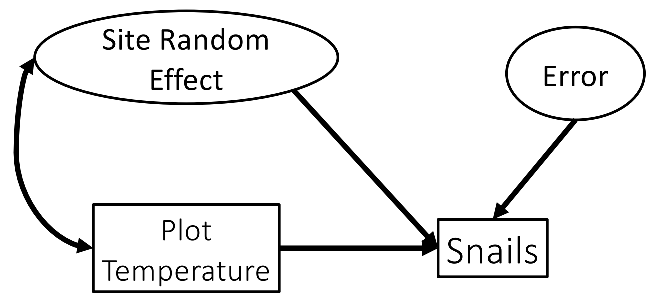
Violation of Exchangeability Make this Model Invalid



Temperature is not drawn from a normal distribution across sites as RE implies!

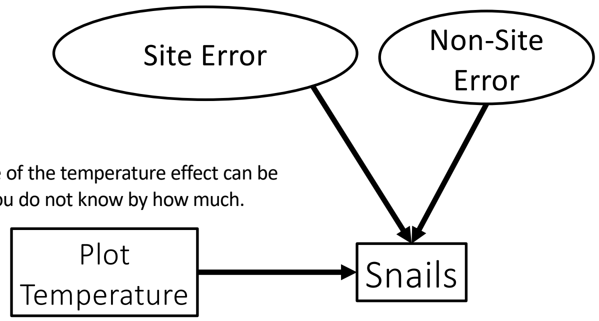
It is a gradient!

Problem #2: Correlation Between Site and Temperature Violates Random Effects Assumption for Snail Piece



This is not how mixed models work correct

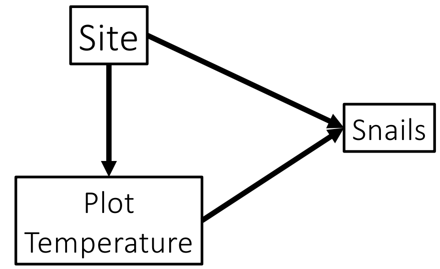
An SEM/DAG Way of Thinking About Mixed Models: An Open Backdoor!



Your estimate of the temperature effect can be **biased** and you do not know by how much.

Note, we are no longer adjusting for correlation!

Solution #1: Fixed Effects Shut the Back Door!

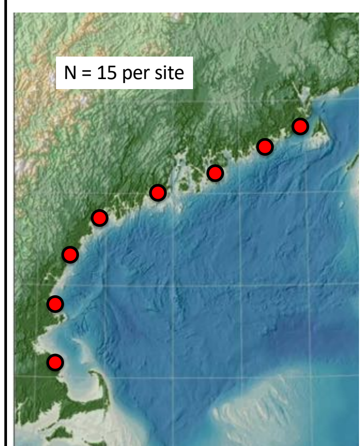


- Site now embodies all non-temperature site-level variation
- We adjust for site-temp relationship in snail piece of the model
- Temperature is now causal!
- Site coefs don't contain causal meaning

Problems with Fixed Effects Approach

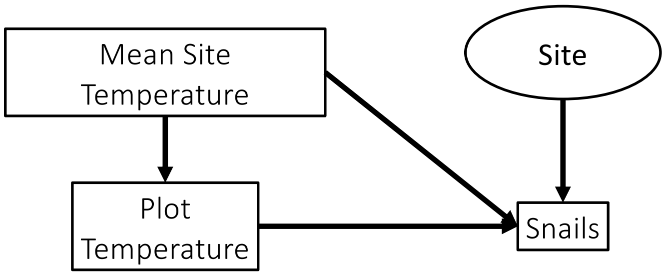
- ***Ineffecient***
  - Eats up DF for model pieces
  - Lots of sites and small n per site leads to trouble
- What if we want to partition out site variability not due to things correlated with site?
  - Is there a role for a site RE somehow

Group Mean Models



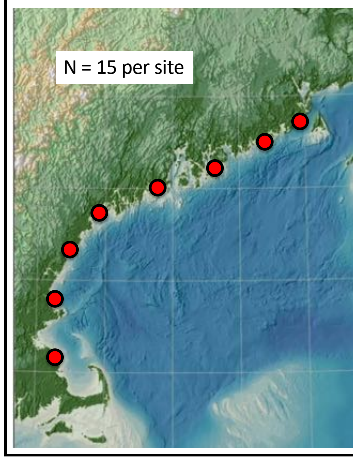
- Get the mean value of predictors at each site
- This encompasses variability due to site-level processes correlated with predictors
- Site RE is now uncorrelated site-level variability

Solution #2 Group Mean Model (Mundlak Device)



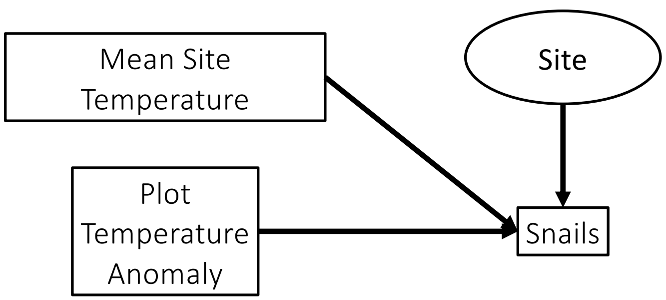
- Site now embodies all non-temperature correlated site-level variation
- We adjust for correlated site-level effects
- Temperature is now causal!
- Mean temp is everything correlated with temp at a site level
- It is not causal

Group Mean Centering



- Get the mean value of predictors at each site
- For each plot, subtract the site mean
- What remains is the plot temperature anomaly, or, **group mean centered** value

Solution #3 Group Mean Centered Model

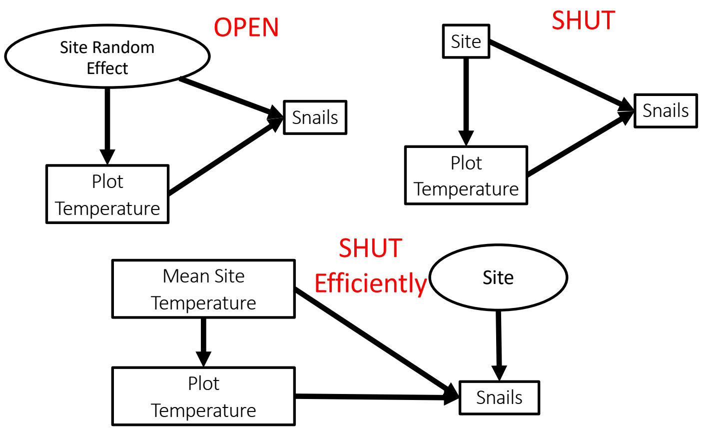


- Not really an SEM anymore…
- Site embodies all non-temperature correlated site-level variation
- We adjust for correlated site-level effects
- Temperature is now causal!

Does it Work: Results from a Simulation with 10 sites, n = 10/site

Model	Temperature Effect	SE Effect
Naive RE Model	-0.1485831	0.1087651
Fixed Effects Model	-0.2401346	0.1183225
Mundlak Device	-0.2401346	0.1176633
Group Mean Centered Model	-0.2401346	0.1176633

So, To Shut the Back Door…



Common Uses of Categorical Fixed Effects, Group Means, and Group Mean Centering

- Great for nested data
- Great for longitudinal data
- Can have multiple groups (spatio-temporal, etc), and can extend conceptually to any structure
- Rich literature on this in Econometrics

Categorical Variables: Conclusions

- They are very useful!
- No reason not to include in piecewiseSEM
- Can help shut the back door to ensure causal identification
- But, as always, be careful of interpretation and how quickly they eat DF

Get Yourself CATegorically Centered!

Yogakatz.com