Mixed Models in SEM

Overview

1. Fixed vs. Random
2. Pseudo-R^2s
3. SEM Example of mixed models
4. Causal Modeling with Random Effects
5. Fully hierarchical SEM

Fixed vs. Random. Comparison

Fixed	Random
Interested in drawing inferences / making predictions	Not particularly interested in any particular value or level
Represent values from the entire 'universe' of interest	A (random) sample from a larger pool of potential values
Levels not interchangeable	Levels interchangeable (could swap / relabel levels without any change in meaning)
Directly manipulated	Introduces incidental error (e.g., between subjects, blocks, sites, etc.)
Few levels / worth sacrificing d.f. to fit model	Many levels / cannot sacrifice d.f. to fit model

Fixed vs. Random. Why mixed models?

- More power than modeling the means of groups
- Reduces degrees of freedom necessary to fit model and estimate parameters (vs. modeling as a fixed effect)
- Accounts for uneven sampling within groups by using information across groups to inform the individual group means
- Can account for *non-independence* of observations by explicitly modeling their covariances (e.g., among sites, individuals, etc.)

Fixed vs. Random. Random structure

Different configurations of <u>random structure</u>:

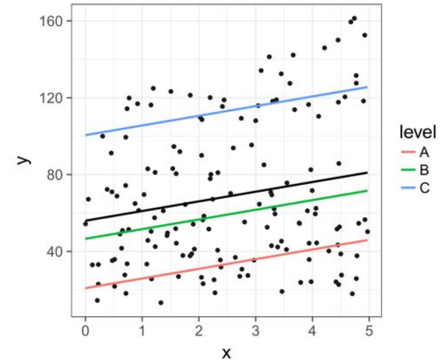
1. Varying intercept, fixed slope

2. Fixed intercept, varying slope

3. Varying intercept, varying slope

Fixed vs. Random. Varying intercept

- Estimates different intercept, same slope for all levels of the random effect

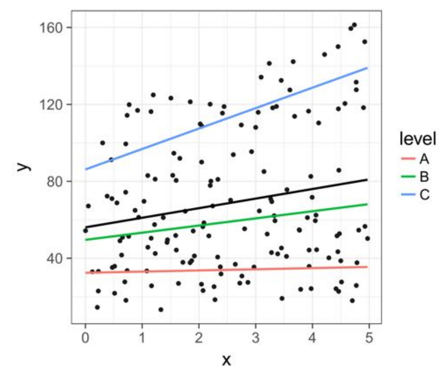


Fixed vs. Random. Varying intercept

- Good for block designs, repeated measures
- Can lead to overconfident estimates if levels are expected to respond differently (e.g., individuals in a drug trial)

Fixed vs. Random. Varying intercept AND slope

- Estimates different slope, different intercept for all levels



Fixed vs. Random. Varying intercept AND slope

- Addresses multiple sources of non-independence of within and between levels, leading to lower Type I *and* Type II error
- Random slopes can be extracted and used in other analyses (get error from lmerTools)
- Computationally intensive, may lead to non-convergence

Fixed vs. Random. Nesting

- Hierarchical models represent nested random terms (e.g., site within region)
- Nesting further addresses non-independence by modeling correlations within *and* between levels of the hierarchy
- Good for stratified sampling designs (varying intercept) and split-plot designs (varying slope, varying intercept)

Fixed vs. Random. Crossed effects

- Multiple random effects that are not nested but apply independently to the observation (e.g., space *and* time)

Fixed vs. Random. Random structures

(1|group)	random group intercept
(x|group) = (1+x|group)	random slope of x within group with correlated intercept
(0+x|group) = (-1+x|group)	random slope of x within group: no variation in intercept
(1|group) + (0+x|group)	uncorrelated random intercept and random slope within group
(1|site/block) = (1|site)+(1|site:block)	intercept varying among sites and among blocks within sites (nested random effects)
site+(1|site:block)	*fixed* effect of sites plus random variation in intercept among blocks within sites
(x|site/block) = (x|site)+(x|site:block) = (1 + x|site)+(1+x|site:block)	slope and intercept varying among sites and among blocks within sites
(x1|site)+(x2|block)	two different effects, varying at different levels
x*site+(x|site:block)	fixed effect variation of slope and intercept varying among sites and random variation of slope and intercept among blocks within sites
(1|group1)+(1|group2)	intercept varying among crossed random effects (e.g. site, year)

http://glmm.wikidot.com/faq

Fixed vs. Random. A warning

- Assumes fixed and random effects are *uncorrelated*
  - e.g., all of your warm data points don't come from a different site than your cool data points
- If possible, fit random effects as fixed effects and compare parameter estimates of other predictors
- Need to ensure appropriate replication at *lowest* level of nested factors (5-6 levels, *minimum*) – otherwise, fit as fixed effects

Fixed vs. Random. Different distributions

- *lme4* can fit many kinds of different distributions using `glmer`
- Does not provide *P*-values (d.d.f uncertain, see: https://stat.ethz.ch/pipermail/r-help/2006-May/094769.html)
  - Need to turn to *pbkrtest* package which estimates d.d.f. using the Kenward-Rogers approximation (less finicky than *lmerTest*)
  - *piecewiseSEM* does this for you automatically using `coefs`

Fixed vs. Random. Different distributions

- *nlme* can only handle normal distributions
  - Ives (2015): "For testing the significance of regression coefficients, go ahead and log-transform count data"
- `glmmPQL` in the *MASS* package uses penalized quasi-likelihood to fit models, can incorporate many different distributions and their quasi- equivalents (e.g., quasi-Poisson)
  - Quasi-distributions estimate a separate term for how the variance scales with the mean, so ideal for over/under-dispersed data
  - Quasi-likelihood means no likelihood based statistics (e.g., AIC, LRT, etc.) for any models fit with `glmmPQL`
  - Implementing R^2 for quasi-distributions right now

Fixed vs. Random. Troubleshooting

- R has the most infuriating error messages
- Can sometimes solve by switching to a different optimizer
  - `lmeControl(opt = "optim")` usually works
- Reduce tolerance for convergence
  - `lmeControl(tol = 1e-4)`
- Respecify random structure
  - Optimizer constrained to have cov > 0, can sometimes get stuck bouncing around when random components are very close to 0
- https://stackexchange.com/
  - Ben Bolker to the rescue! https://dynamicecology.wordpress.com/2013/10/04/wwbbd/

Overview

1. Fixed vs. Random
2. Pseudo-R^2s
3. SEM Example of mixed models
4. Causal Modeling with Random Effects
5. Fully hierarchical SEM

Pseudo-R^2s. Omnibus test

- Fisher's *C* is the global fit statistic for local estimation but has many shortcomings:
  - Sensitive to the number of d-sep tests and the complexity of the model (harder to reject as the complexity increases)
  - Sensitive to the size of the dataset (e.g., high *n* leads to low *P*)
  - Fails symmetricity when dealing with unlinked non-normal intermediate variables

Pseudo-R^2s. Local tests

- How do we infer the confidence in our SEM?
  - Examine standard errors of individual paths, qualitatively assess cumulative precision
  - Explore variance explained (i.e., R^2), qualitatively assess cumulative precision

Pseudo-R^2s. General linear regression

- Coefficient of determination (R^2) = proportion of variance in response explained by fixed effects
- For OLS regression, simply 1- the ratio of unexplained (error) variance (e.g., SS_{error}) over the total explained variance (e.g., SS_{total})
- Ranges (0, 1), independent of sample size
- Not good for model comparisons since R^2 monotonically increases with model complexity

Pseudo-R^2s. Generalized linear regression

- Likelihood estimation is not attempting to minimize variance but instead obtain parameters that maximize the likelihood of having observed the data
- In a likelihood framework, equivalent R^2 = 1- the ratio of the log-likelihood of the full model over the log-likelihood of the null (intercept-only) model
- Leads to identical R^2 as OLS for normal (Gaussian) distributions, not so for GLM – need to use likelihood-based pseudo-R^2 (e.g., McFadden, Nagelkerke)

Pseudo-R^2s. Generalized mixed models

- Becomes even worse for mixed models because variance is partitioned among levels of the random factor, so what is the error variance?
- Need a new formulation of R^2 :
  - Marginal R^2 = variance explained by fixed effects only

Fixed effects variance

$$R^2_{\mathrm{GLMM}(m)} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2}$$

Fixed effects variance

Random effects variance

Residual variance

Distribution-specific variance

Pseudo-R^2s. Generalized mixed models

- Conditional R^2 = variance explained by both the fixed and random effects

Fixed effects variance

Random effects variance

$$R^2_{\mathrm{GLMM}(c)} = \frac{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2}$$

Fixed effects variance

Random effects variance

Residual variance

Distribution-specific variance

Pseudo-R^2s. Generalized mixed models

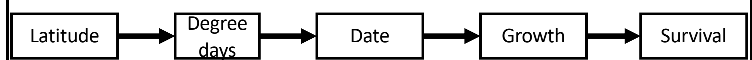
- Comparison of marginal and conditional R^2 can lead to roundabout assessment of ‘significance’ of the random effects (e.g., if conditional R^2 is larger relative to marginal R^2)
- Best to report both and allow readers to determine how their magnitude affects the inferences

Overview

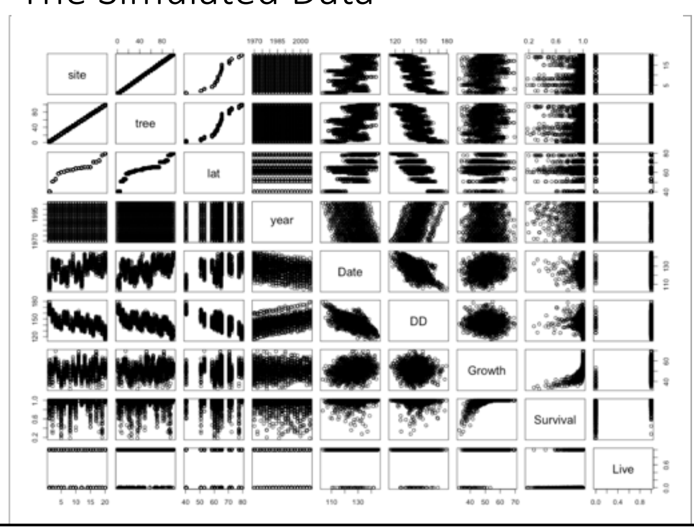
1. Fixed vs. Random
2. Pseudo-R^2s
3. SEM Example of mixed models
4. Causal Modeling with Random Effects
5. Fully hierarchical SEM

SEM Example. Shipley 2009

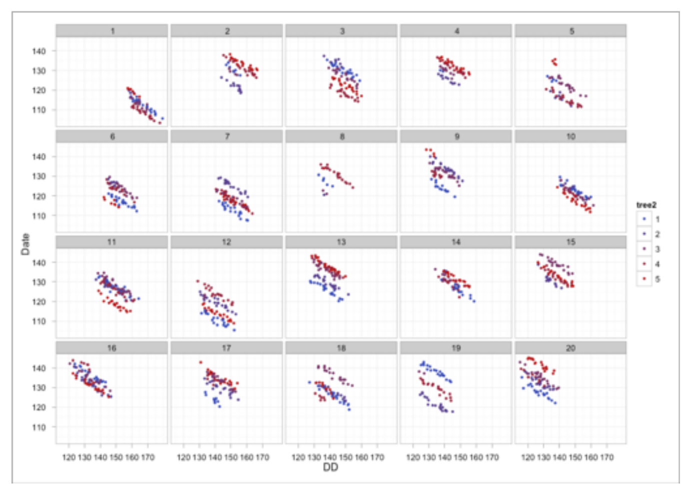
- Hypothetical dataset: predicting latitude effect on survival of a tree species
- Repeated measures on 5 subjects at 20 sites from 1970-2006
- Survival (0/1) influenced by phenology (degree days until bud break, Julian days until bud break), size (stem diameter growth)



The Simulated Data



Nested Structure in the Data



SEM Example. Shipley 2009

- Two distributions: normal, binary (survival)
- Random effects:
  - Site-only: latitude
  - Site and year: degree days, date
  - Site, year, and subject: diameter, survival

Latitude → Degree days → Date → Growth → Survival

SEM Example. What is the basis set?

Latitude → Degree days → Date → Growth → Survival

- Date $\perp$ Lat | (Degree days)
- Growth $\perp$ Lat | (Date)
- Survival $\perp$ Lat | (Growth)
- Growth $\perp$ Degree days | (Date, Lat)
- Survival $\perp$ Degree days | (Growth, Lat)
- Survival $\perp$ Date | (Growth, Degree days)

SEM Example. List of equations

Latitude → Degree days → Date → Growth → Survival

```
# Load data
shipley <- read.csv("shipley.csv")

# Create list of structural equations
shipley.sem <- psem(
  lme(DD ~ lat, random = ~1|site/tree, na.action = na.omit,
      data = shipley),
  lme(Date ~ DD, random = ~1|site/tree, na.action = na.omit,
      data = shipley),
  lme(Growth ~ Date, random = ~1|site/tree, na.action = na.omit,
      data = shipley),
  glmer(Live ~ Growth + (1|site) + (1|tree),
        family=binomial(link = "logit"), data = shipley)
)

# Get summary
summary(shipley.sem)
```

SEM Example. D-sep tests

Latitude → Degree days → Date → Growth → Survival

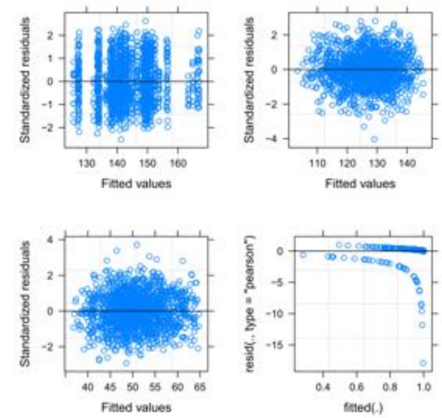
```
---
Tests of directed separation:

      Independ.Claim Estimate Std.Error   DF Crit.Value P.Value
    Date  ~  lat + ...  -0.0091    0.1135   18    -0.0798  0.9373
  Growth  ~  lat + ...  -0.0989    0.1107   18    -0.8929  0.3837
    Live  ~  lat + ...   0.0305    0.0297   NA     1.0280  0.3039
   Growth  ~  DD + ...  -0.0106    0.0358 1329    -0.2967  0.7667
     Live  ~  DD + ...   0.0272    0.0271   NA     1.0038  0.3155
   Live  ~  Date + ...  -0.0466    0.0298   NA    -1.5626  0.1181

Global goodness-of-fit:

  Fisher's C = 11.538 with P-value = 0.483 and on 12 degrees of freedom
```

SEM Example. Extract coefficients

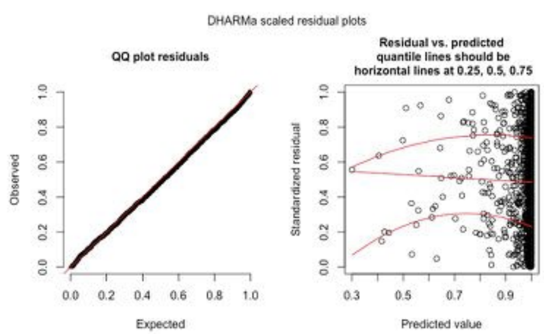
Latitude → Degree days → Date → Growth → Survival

```
Coefficients:

  Response Predictor Estimate Std.Error   DF Crit.Value P.Value Std.Estimate
        DD       lat  -0.8355    0.1194   18    -6.9960       0      -0.6877 ***
      Date        DD  -0.4976    0.0049 1330  -100.8757       0      -0.6281 ***
    Growth      Date   0.3007    0.0266 1330    11.2917       0       0.3824 ***
      Live    Growth   0.3479    0.0584 1431     9.9552       0           NA ***

  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05

Individual R-squared:

  Response method Marginal Conditional
        DD   <NA>     0.49        0.70
      Date   <NA>     0.41        0.98
    Growth   <NA>     0.11        0.84
      Live  delta     0.16        0.18
```

Evaluate residuals by lapplying plot



SEM Example. For GLMMs, use DHARMa




```
#residuals for a glmm
library(DHARMa)
sims <- simulateResiduals(shipley.sem[[4]])
plot(sims)
```

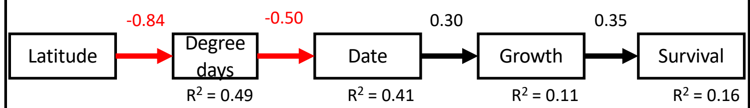
SEM Example. Populate final model

Latitude →(-0.84) Degree days →(-0.50) Date →(0.30) Growth →(0.35) Survival

Degree days: $R^2 = 0.49$; Date: $R^2 = 0.41$; Growth: $R^2 = 0.11$; Survival: $R^2 = 0.16$

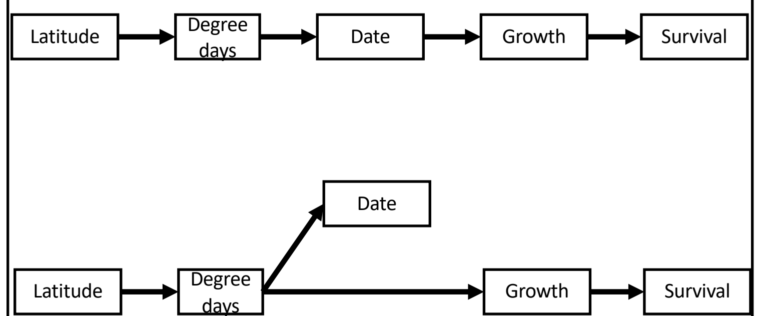
```
Coefficients:

  Response Predictor Estimate Std.Error   DF Crit.Value P.Value Std.Estimate
        DD       lat  -0.8355    0.1194   18    -6.9960       0      -0.6877 ***
      Date        DD  -0.4976    0.0049 1330  -100.8757       0      -0.6281 ***
    Growth      Date   0.3007    0.0266 1330    11.2917       0       0.3824 ***
      Live    Growth   0.3479    0.0584 1431     9.9552       0           NA ***

  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

SEM Example. Populate final model

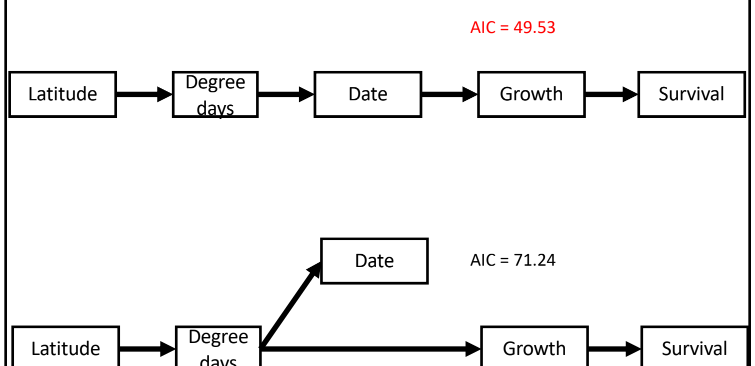



```
Individual R-squared:

  Response method Marginal Conditional
        DD    <NA>     0.49        0.70
      Date    <NA>     0.41        0.98
    Growth    <NA>     0.11        0.84
      Live   delta     0.16        0.18
```

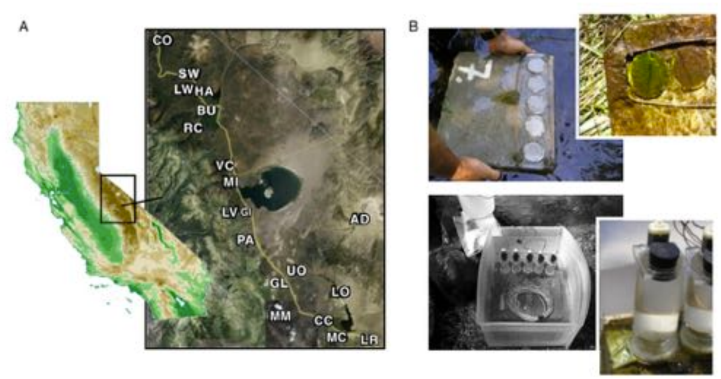
SEM Example. Compare these models



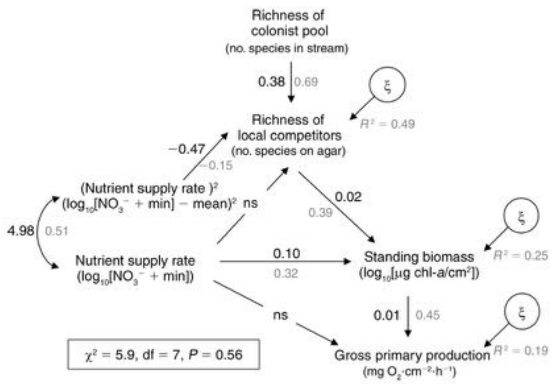
SEM Example. Compare these models



Cardinale et al. (2009) - Does diversity drive productivity or vice versa?

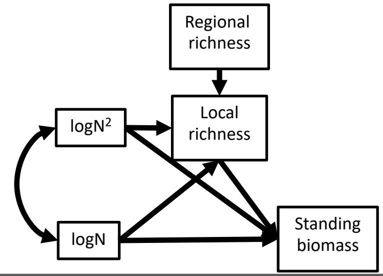


Multivariate Diversity-Productivity Model



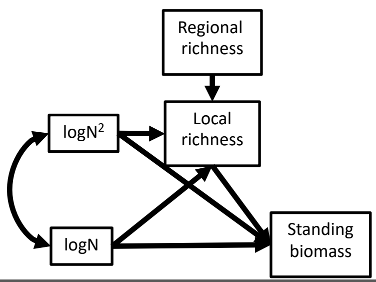
Cardinale et al. 2009

Remember to center!



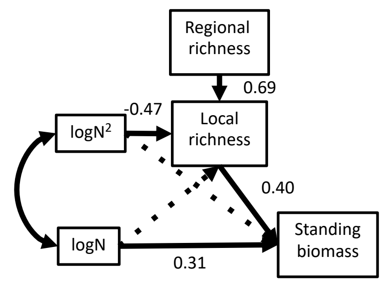

```
# Load data
cardinale <- read.csv("../data/cardinale.csv")

# Take log of N and N^2
cardinale$logN <- log10(cardinale$N + 1e-6)

cardinale$logN2 <- cardinale$logN ^ 2

# Take log of chl (standing biomass)
cardinale$logChl <- log10(cardinale$Chl)
```

Remember to center!



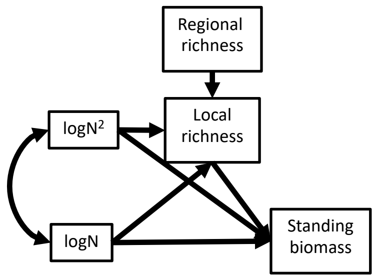

```
# Center polynomial to reduce collinearity
cardinale$logN.cen = scale(cardinale$logN, scale = F)

cardinale$logN2.cen = scale(cardinale$logN, scale = F) ^ 2
```

The Model without Mixed Effects



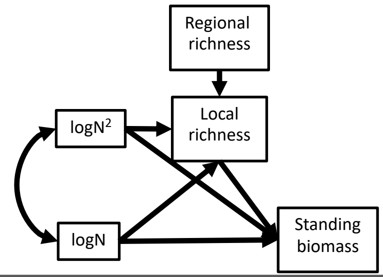

```
> coefs(cardinale.sem2)
  Response Predictor Estimate Std.Error  DF Crit.Value P.Value Std.Estimate
1       SA  logN.cen   0.3668    0.4460 123     0.8223  0.4125       0.0618
2       SA logN2.cen  -0.4742    0.2424 123    -1.9568  0.0526      -0.1470
3       SA        SR   0.3838    0.0359 123    10.6844  0.0000       0.6893 ***
4   logChl        SA   0.0201    0.0040 123     5.0327  0.0000       0.3946 ***
5   logChl  logN.cen   0.0944    0.0275 123     3.4320  0.0008       0.3116 ***
6   logChl logN2.cen   0.0032    0.0150 123     0.2108  0.8334       0.0193
```

Exercise: Fit with Stream as grouping variable



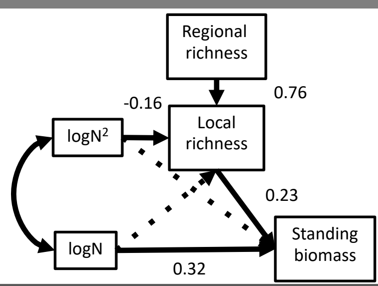
1. What does the model look like?
2. How does it differ from the fixed effects only model?

The Model without Mixed Effects




```
# Re-fit SEM using centered predictors
cardinale.mixed <- psem(
  lme(SA ~ logN.cen + logN2.cen + SR,
      random = ~1|Stream, data = cardinale),

  lme(logChl ~ SA + logN.cen + logN2.cen,
     random = ~1|Stream,  data = cardinale),

  data = cardinale
)
```

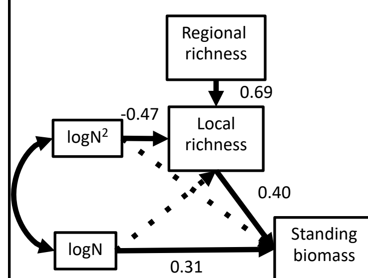
The Model without Mixed Effects



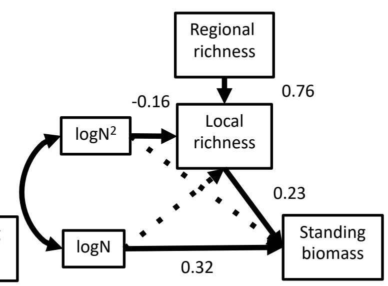

```
> coefs(cardinale.mixed)
  Response Predictor Estimate Std.Error  DF Crit.Value P.Value Std.Estimate
1       SA  logN.cen   0.4356    0.3279 105     1.3284  0.1869       0.0734
2       SA logN2.cen  -0.5136    0.1783 105    -2.8806  0.0048      -0.1592  **
3       SA        SR   0.4260    0.0663  18     6.4251  0.0000       0.7651 ***
4   logChl        SA   0.0117    0.0050 104     2.3525  0.0205       0.2285   *
5   logChl  logN.cen   0.0970    0.0223 104     4.3535  0.0000       0.3205 ***
6   logChl logN2.cen  -0.0022    0.0123 104    -0.1811  0.8566      -0.0136
```

Compare with v. Without Random Effects

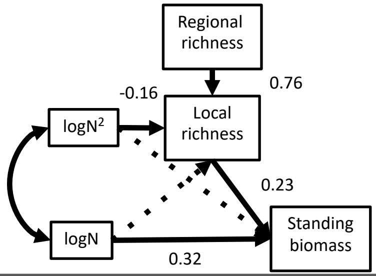
Fixed Effects Only Model



Mixed Model



Compare R^2



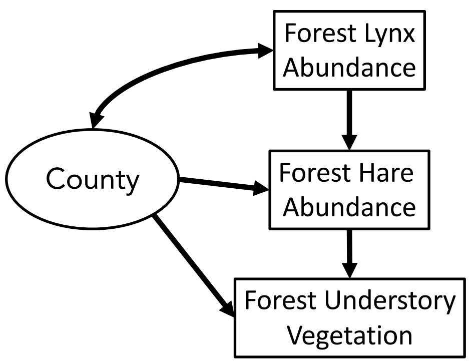

```
> rsquared(cardinale.sem2)
  Response   family     link method R.squared
1       SA gaussian identity   none 0.4882395
2  logChl gaussian identity   none 0.2538638

> rsquared(cardinale.mixed)
  Response   family     link method  Marginal Conditional
1       SA gaussian identity   none 0.5255357   0.7535010
2   logChl gaussian identity   none 0.1568633   0.4773681
```

Overview

1. Fixed vs. Random
2. Pseudo-R^2s
3. SEM Example of mixed models
4. Causal Modeling with Random Effects
5. Fully hierarchical SEM

Mixed Models and Graphical Models

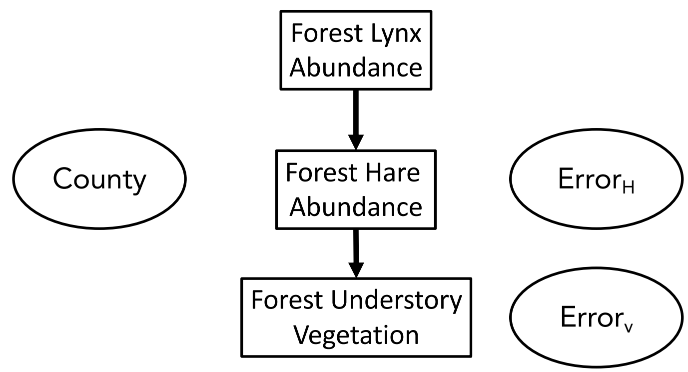
Random Effects are latents…



?

Mixed Models and Graphical Models

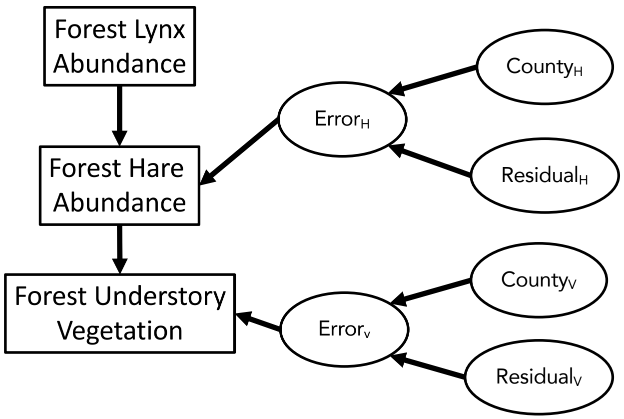
Error is also a latent with mean 0 and some SD



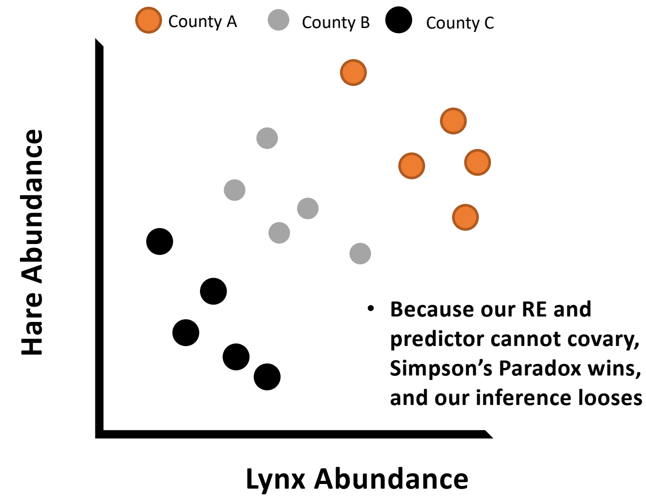
Really, mixed model error is RE variability + Residual variability

What's Really Going on in Mixed Models

Note that there is no way for our RE to covary with our exogenous variable



What if our RE and Predictors Covary?



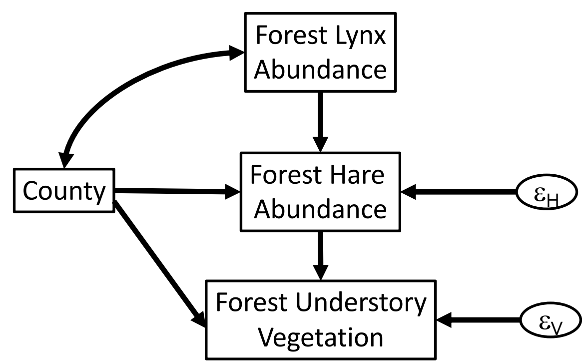
Solutions to our RE and Predictors Covarying

1. Have our RE as a fixed effect
   - Can have interaction effects for variable slopes
   - BUT – can cost DF, and open to critique of generalizability
   - BUT – that doesn't matter if you are interested in causal identifiability

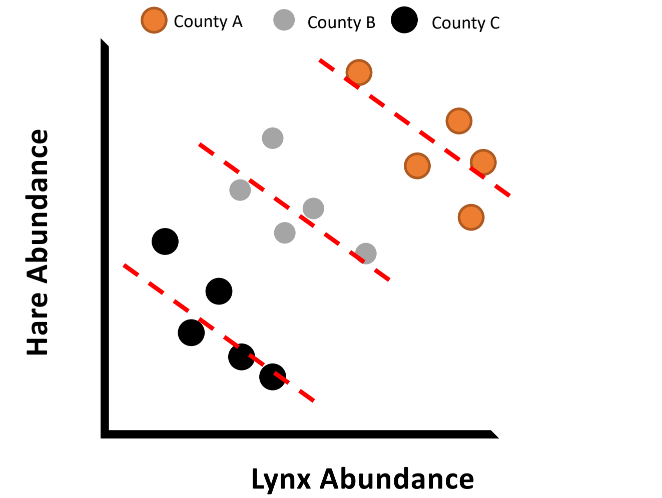
2. Include centered group-level predictor and RE
   - Covariate effect now estimated after controlling for correlation with group level mean
   - Understanding that correlation can be tricky
   - Interpretation of group-level covariate difficult

3. Include centered group-level predictor, deviation from group level predictor, and RE
   - Correlation broken, so both terms easier to interpret
   - Caution: group-level predictor contaminated by other site-level effects

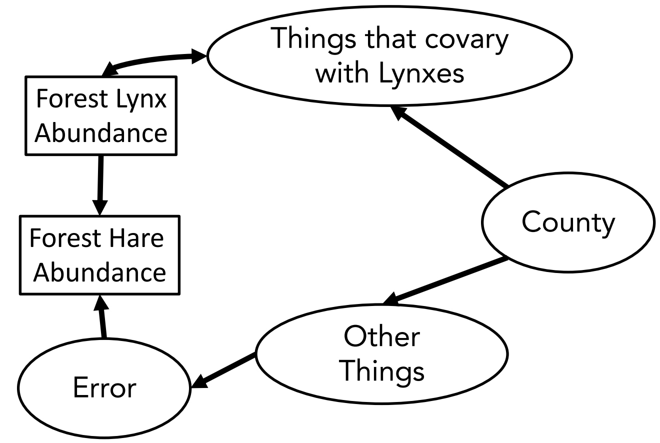
1. Fixed Effect Model



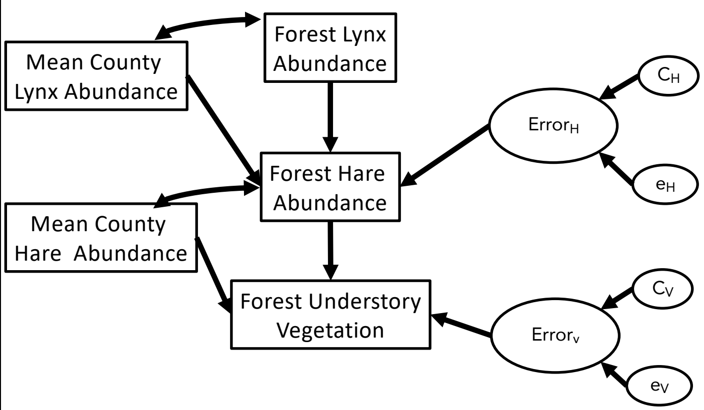
FE Model Handles Simpson's Paradox



2. Where does Group-Level Covariate Come From?

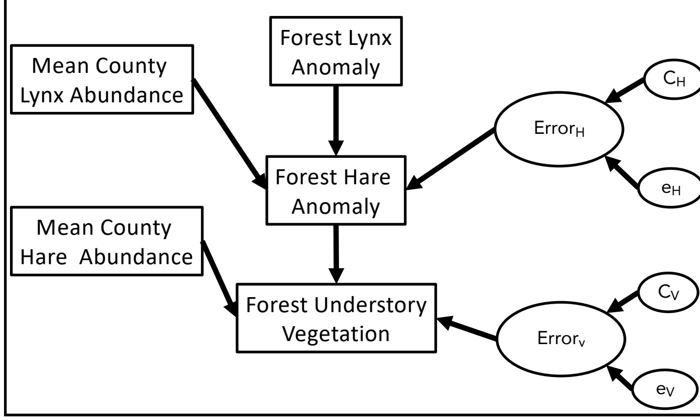


2. Incorporating Group Covariates



3. Incorporating Group Covariates & Centered Predictors

$Anomaly_{ij}$ = $Abundance_{ij}$ – Mean $Abundance_i$



Are Random Effects Always the Answer?

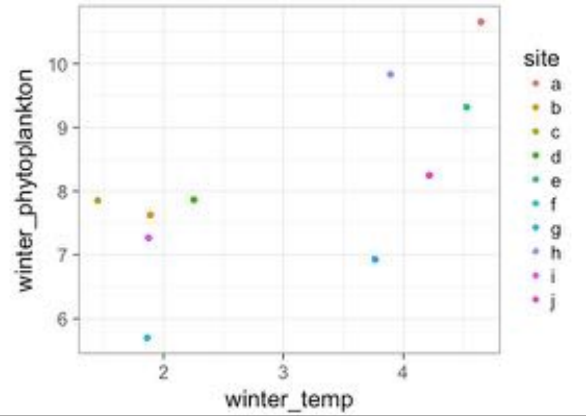
- No!
- We need to be careful that we are not opening a new back door by relying on random effects
- But, through careful consideration of model structure, we can hold that back door shut, and then some!

Overview

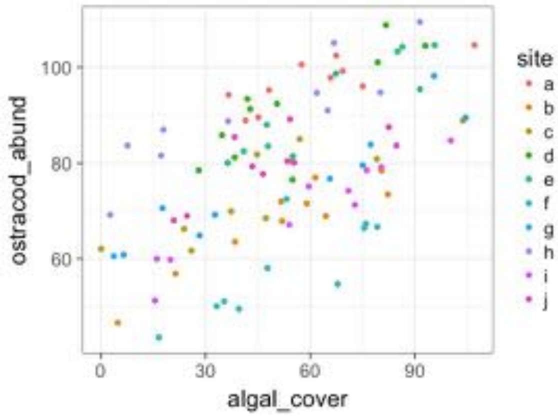
1. Fixed vs. Random
2. Pseudo-R^2s
3. SEM Example of mixed models
4. Causal Modeling with Random Effects
5. Fully hierarchical SEM

Filter Feeding Ostracods Living In Algae

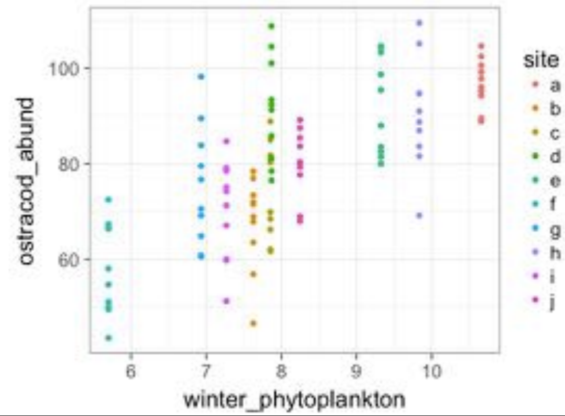
Site-Level Environmental Relationships



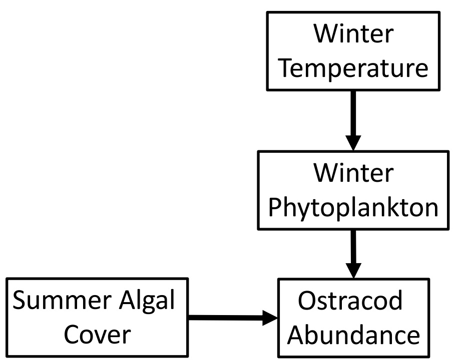
Plot-Level Biotic Relationships



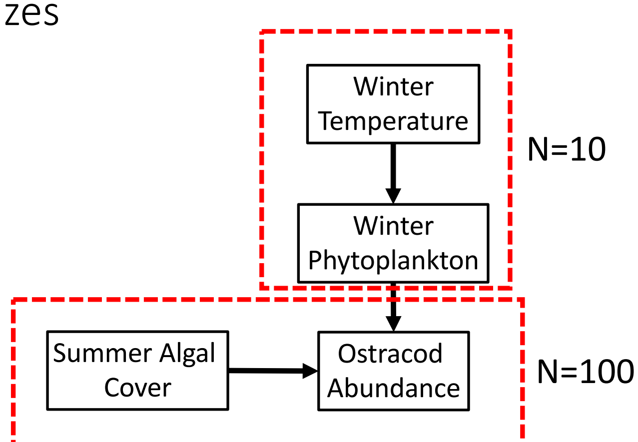
But…site-level drivers of local phenomena



Our Model: What is the basis set?



The problem: Variable Sample Sizes

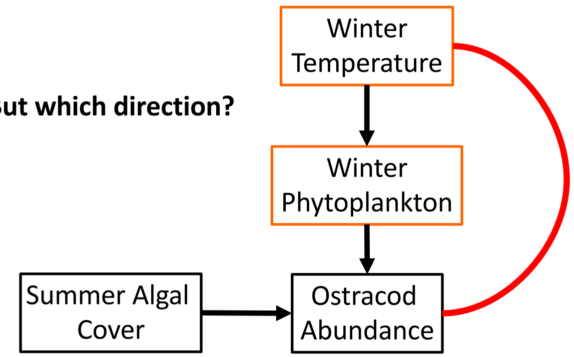


Quandaries with hierarchical SEM

- What *is* our sample size?
  - To some extent solved by hierarchical linear models
  - But, different model components will have different n – and hence different power
- How do we evaluate the basis set?
  - Trickier…but, we can manage!

Our Basis Set

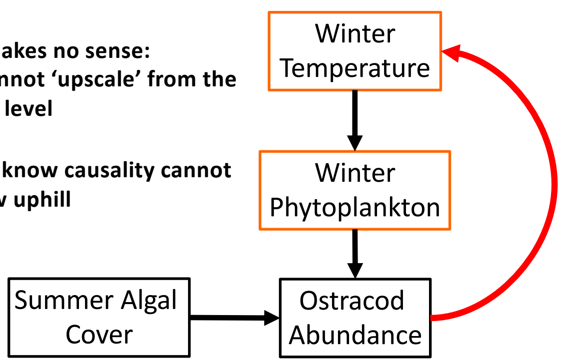
But which direction?



Basis Solutions

This makes no sense:

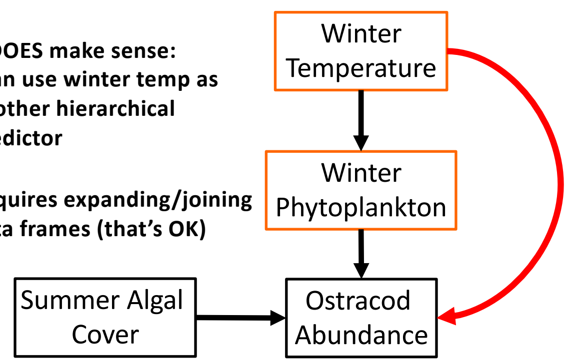
- **Cannot ‘upscale’ from the site level**
- **We know causality cannot flow uphill**



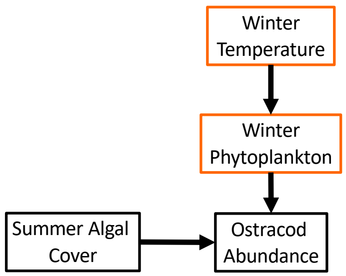
Basis Solutions

This DOES make sense:

- **Can use winter temp as another hierarchical predictor**
- **Requires expanding/joining data frames (that’s OK)**



Our Model: Two Pieces (for now)

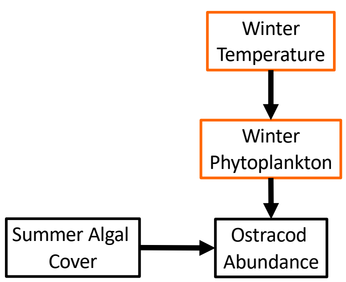



```
ostra_site <- read.csv("../data/ostracod_sitelevel.csv")
ostra_plot <- read.csv("../data/ostracod_plotlevel.csv")
```

Our Model: Two Pieces (for now)

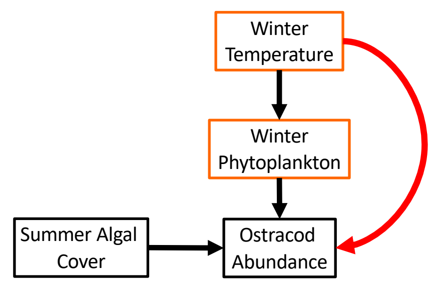
```
#site level
ostra_site_model <- psem(
  lm(winter_phytoplankton ~ winter_temp, data = ostra_site),

  data = ostra_site
)

#plot level
ostra_plot_model <- psem(
  lme(ostracod_abund ~ algal_cover + winter_phytoplankton,
      random = ~1|site, data = ostra_plot),

  data = ostra_plot
)
```

To get C, sum up C from submodels, and get hierarchical C



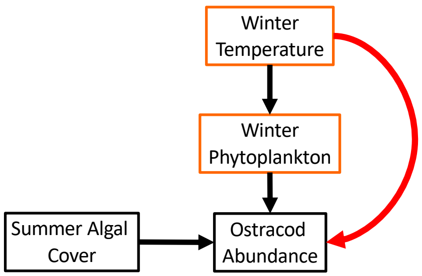

```
> fisherC(ostra_site_model)
  Fisher.C df P.Value
1  0  0       1
2  > fisherC(ostra_plot_model)
  Fisher.C df P.Value
1         0  0       1
```

To get C, sum up C from submodels, and get hierarchical C



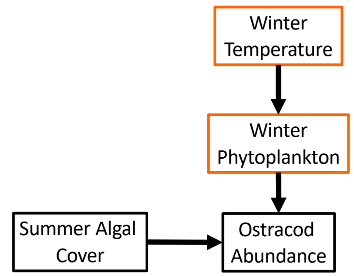

```
basis_mod <-  lme(ostracod_abund ~ algal_cover +
                    winter_phytoplankton + winter_temp,
                    random = ~1|site, data = ostra_plot)
```

To get C, sum up C from submodels, and get hierarchical C




```
Fixed effects: ostracod_abund ~ algal_cover + winter_phytoplankton +
winter_temp
                        Value Std.Error DF   t-value p-value
(Intercept)          5.677315 13.167862 89  0.431149  0.6674
algal_cover          0.324035  0.019371 89 16.728113  0.0000
winter_phytoplankton 6.508925  2.080789  7  3.128105  0.0167
winter_temp          1.210299  2.385633  7  0.507328  0.6275
```

To get C, sum up C from submodels, and get hierarchical C




```
> fish_c <- 0 + 0 + -2*log(0.6275)

> 1 - pchisq(fish_c, df = 1)
[1] 0.3343377
```

Hierarchical Models in SEM

- This is a new and fast developing area
  - Additional methods in next version of lavaan, too
- In essence, everything is the same…
- Except we need to think carefully about what is the correct test of conditional independence
- Otherwise, we use conventional HLMs, as in a univariate sense