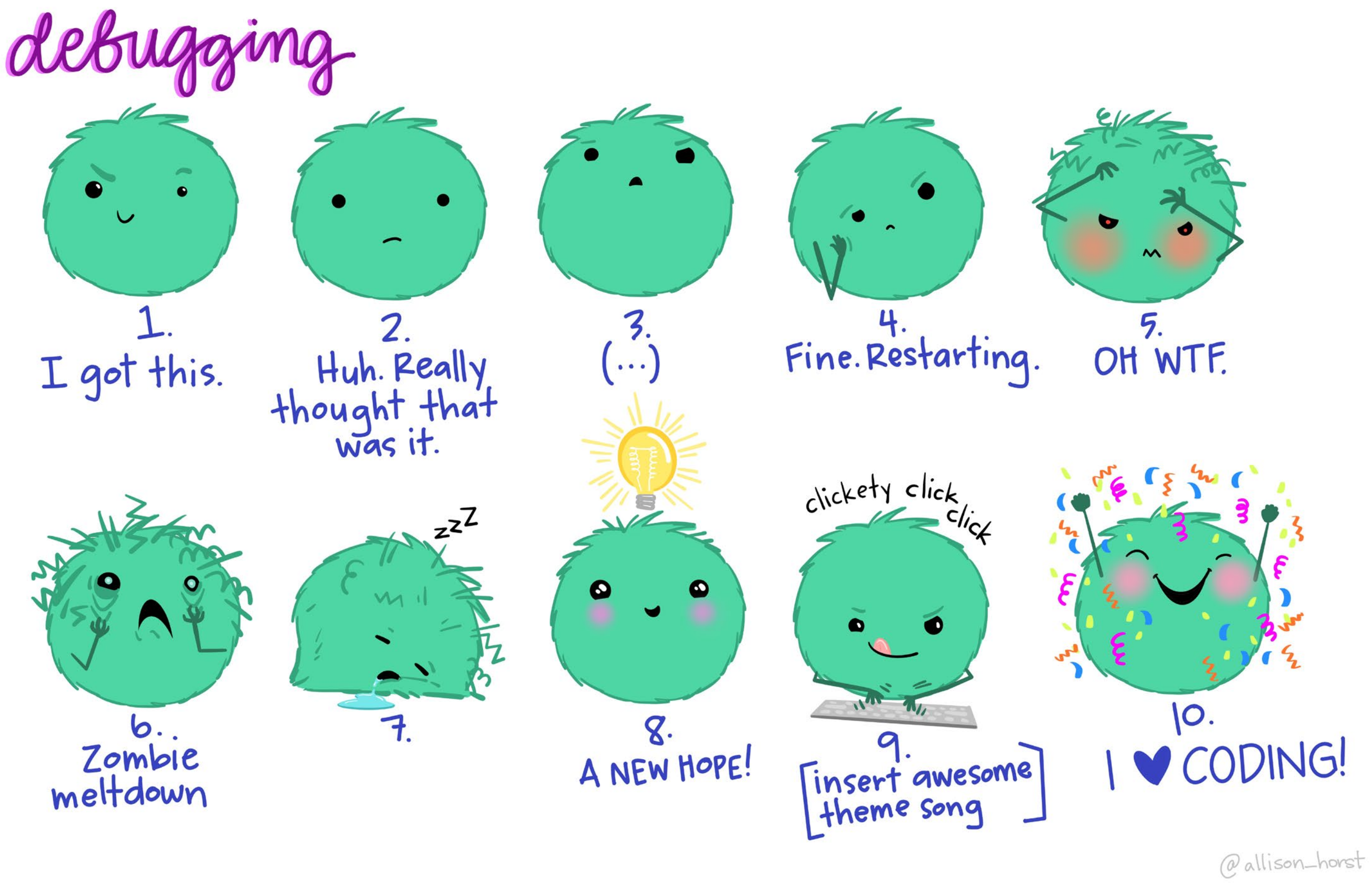
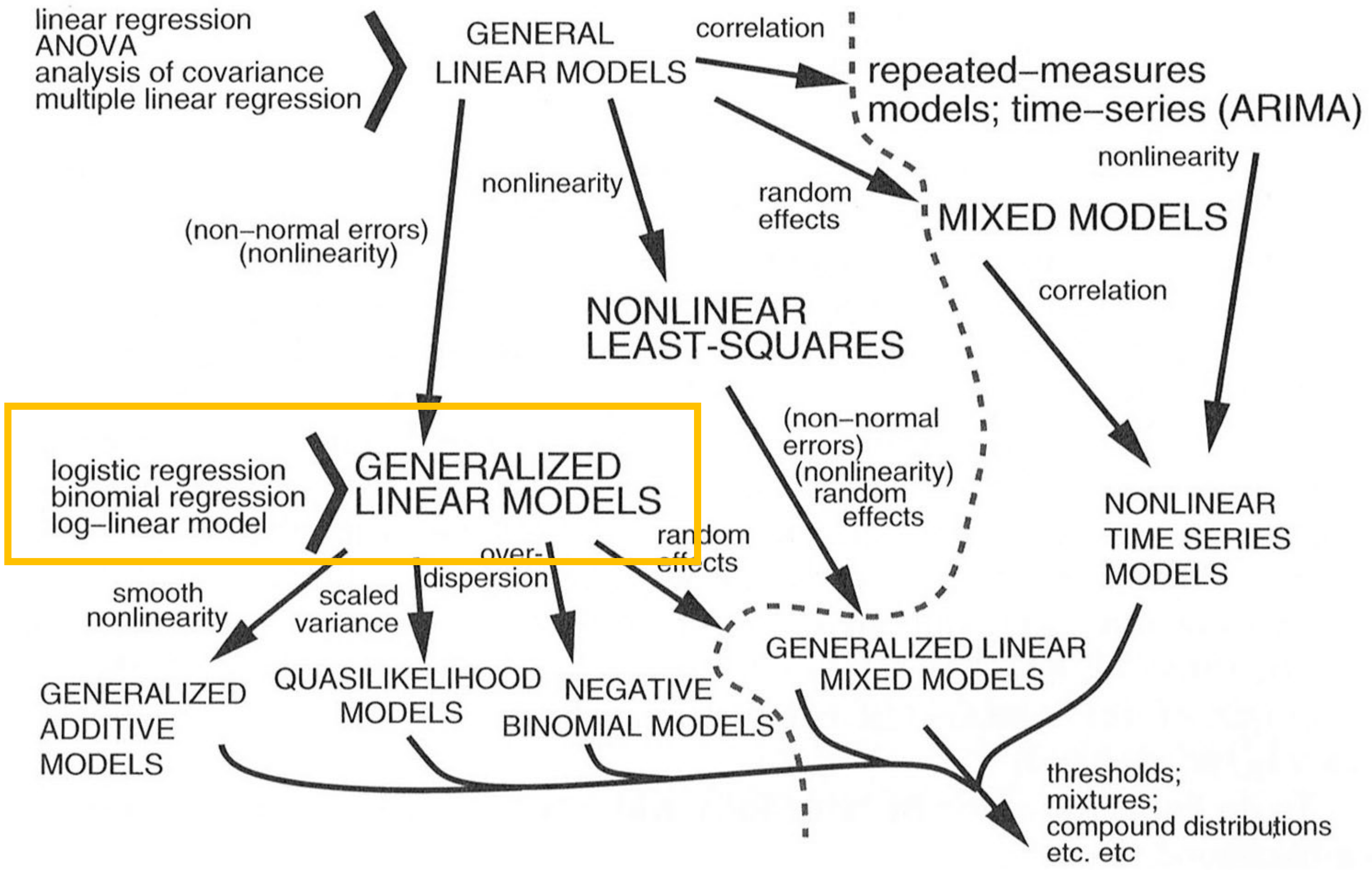
Questions from last week? How did lab go?

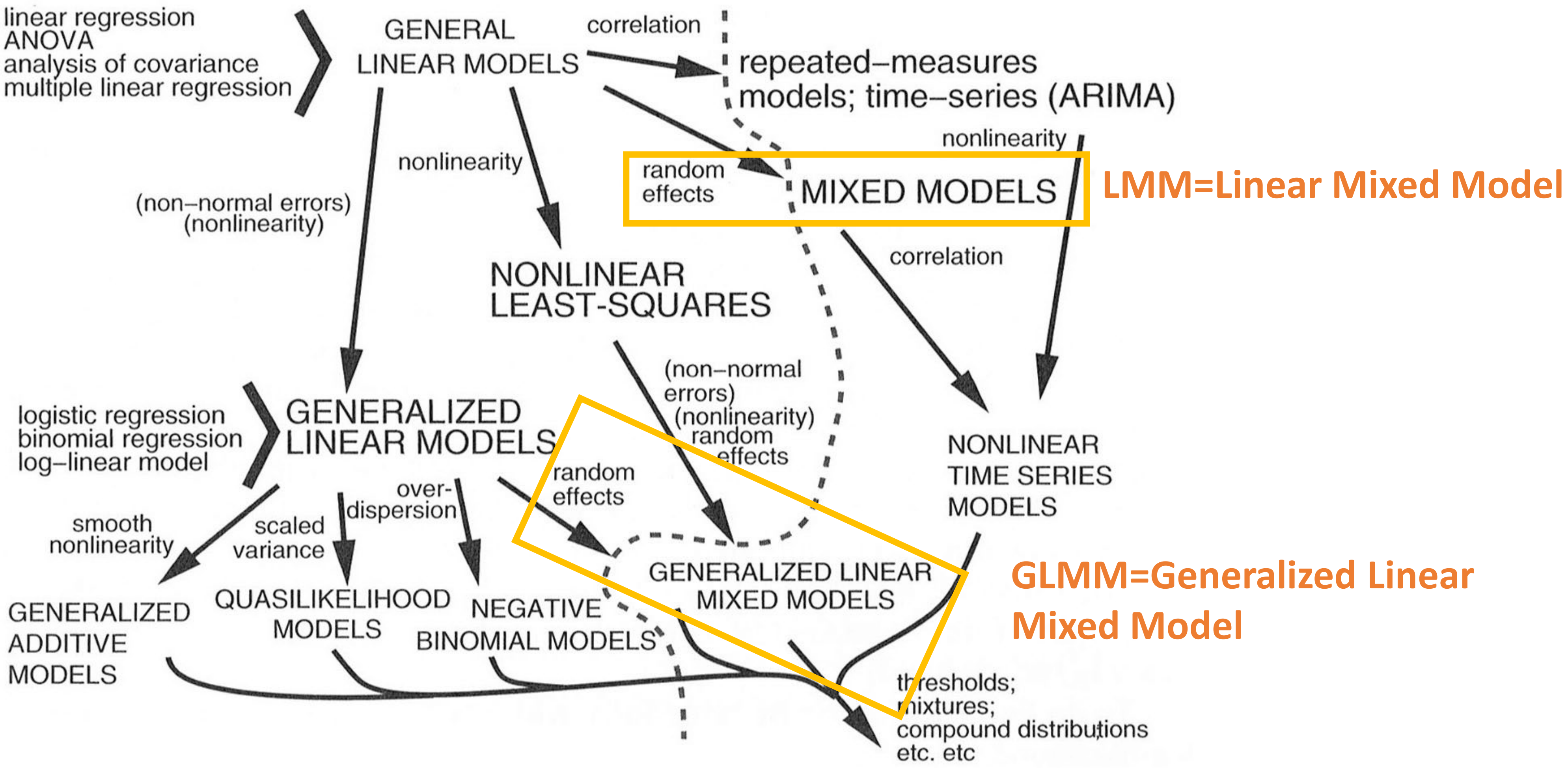
- More glm practice:
  - Interpreting coefficients
  - hypothesis testing
  - interactions
- Hurdle models to deal with zero-inflated data



A road map for the next ~10 weeks

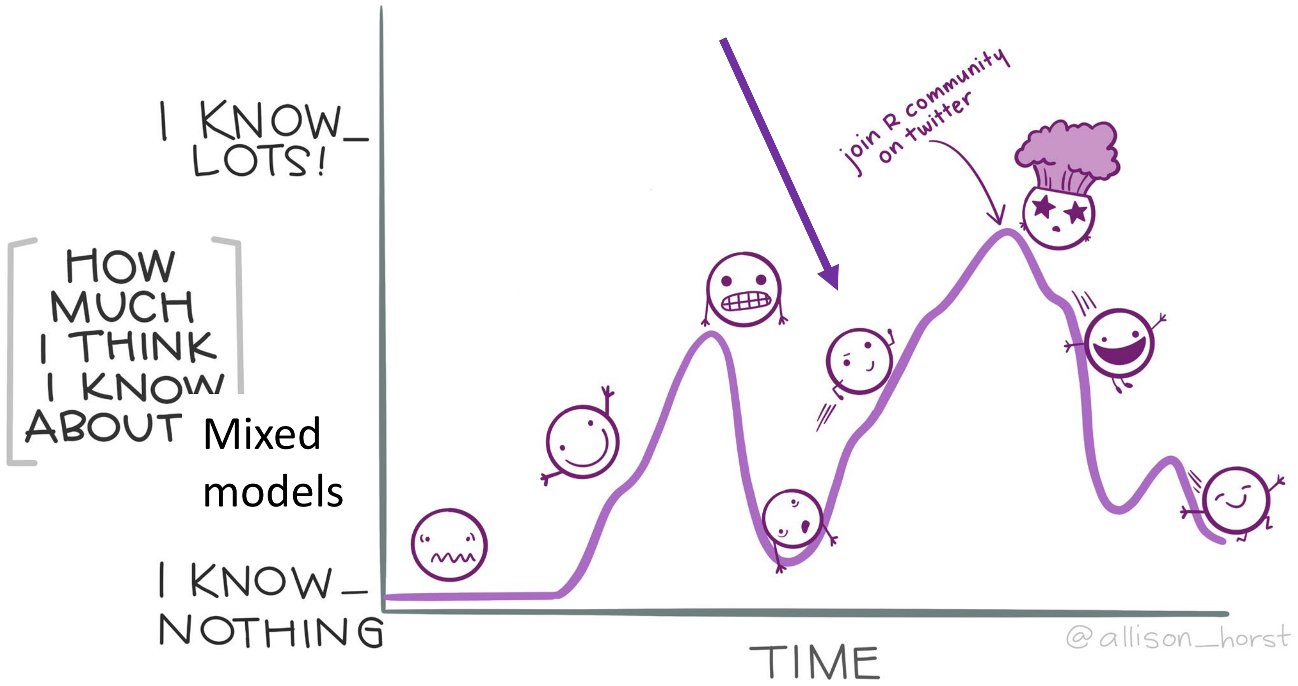


A road map for the next ~10 weeks



Caveat lector

- I'm not a statistician
- We are now wading into waters where methods are under active development and research in the field of statistics
- When appropriate, I'll let you know if you've hit the edge of my knowledge, and point you to more authoritative resources



This week

- Why do we need mixed models?
- What are fixed vs. random effects?
- Think about how fixed/random effects relate to experimental design
- Practice identifying fixed vs. random effects
- Understand the architecture of multilevel/mixed models

Recall the (general) linear model

$$One\ response\ variable = one\ or\ more\ linear\ combinations\ of\ predictor\ variables + error$$

Assumed to be *independent* observations

Resulting in *independent* errors

(Murray)

Mathematical way to understand independence: variance-covariance matrix

$$\text{One response variable} = \text{one or more linear combinations of predictor variables} + \text{error}$$

Homogeneity of variance

$$y_i = \underbrace{\beta_0 + \beta_1 \times x_i}_{\text{Linearity}} + \varepsilon_i \qquad \varepsilon_i \sim \underbrace{\mathcal{N}(0, \sigma^2)}_{\text{Normality}} \qquad \mathbf{V} = cov = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \cdots & \sigma^2 & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{pmatrix}$$

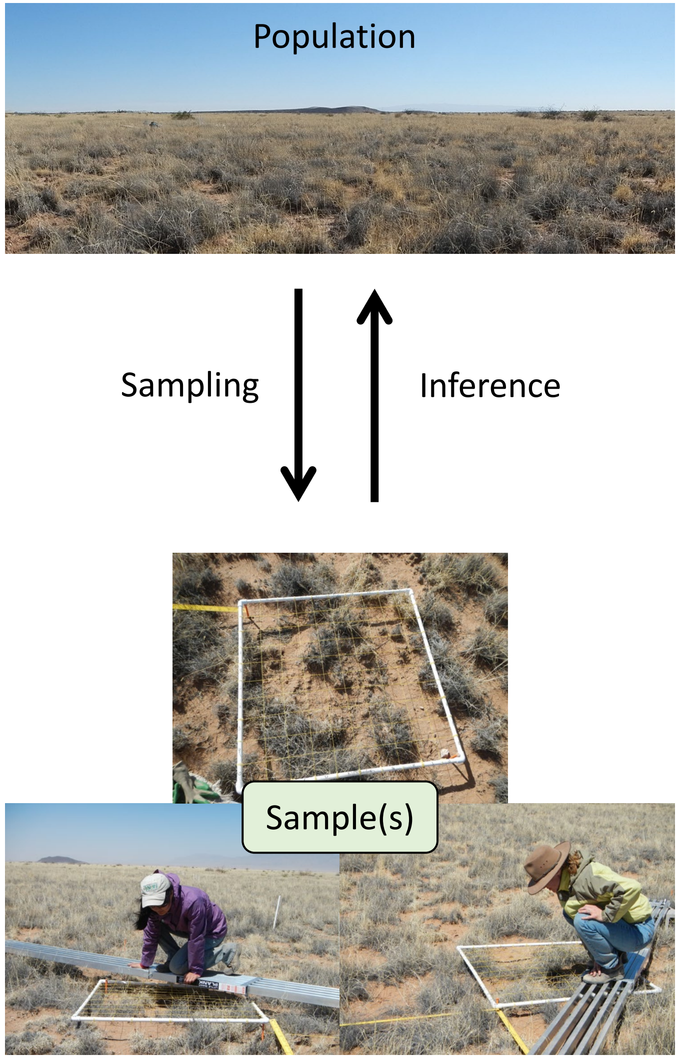
Zero covariance (=independence)

Covariance: A measure of association between two variables, x and y, where *n* = sample size
i is the *i* th observation in your dataset, and the bar indicates the mean value of x or y.

$$COV_{(x,y)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{n-1}$$

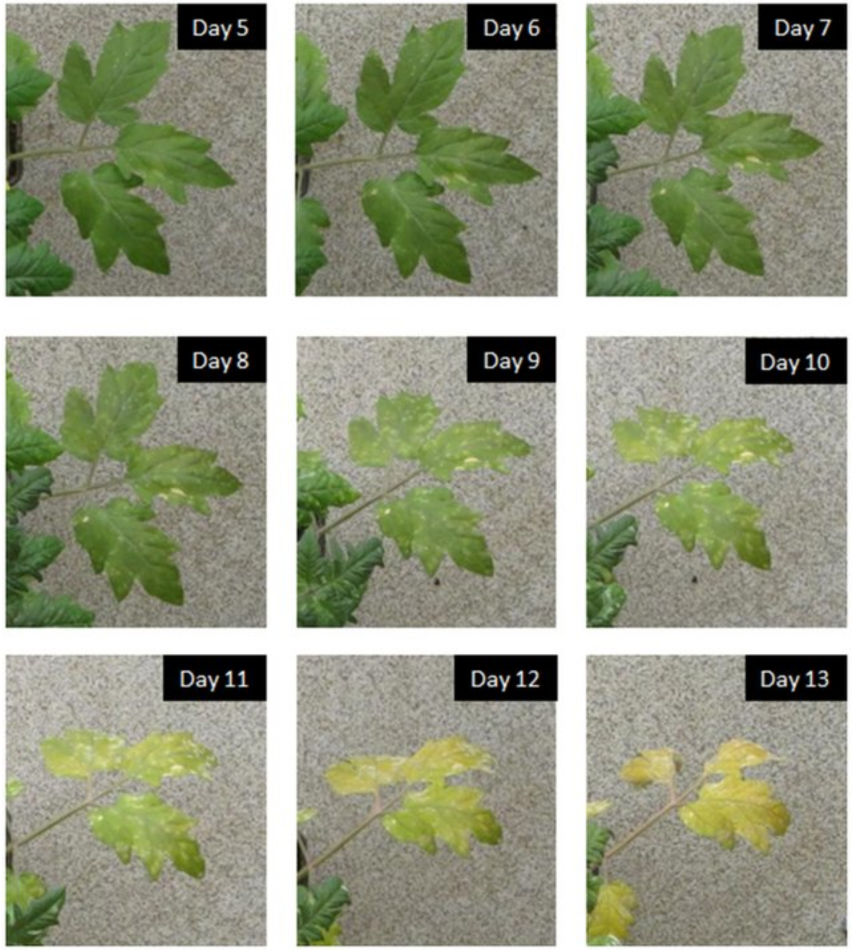
What are some examples where we might violate the “independence” assumption?

Examples of bad samples: non-independence



- In this example, what is the sampling unit?
- The quadrats are independent of each other
- However, the plants within each quadrat are not independent

Repeated measures



Powdery mildew on tomato (Raza et al. 2015)

Often violation of independence comes from some sort of hierarchy in the data

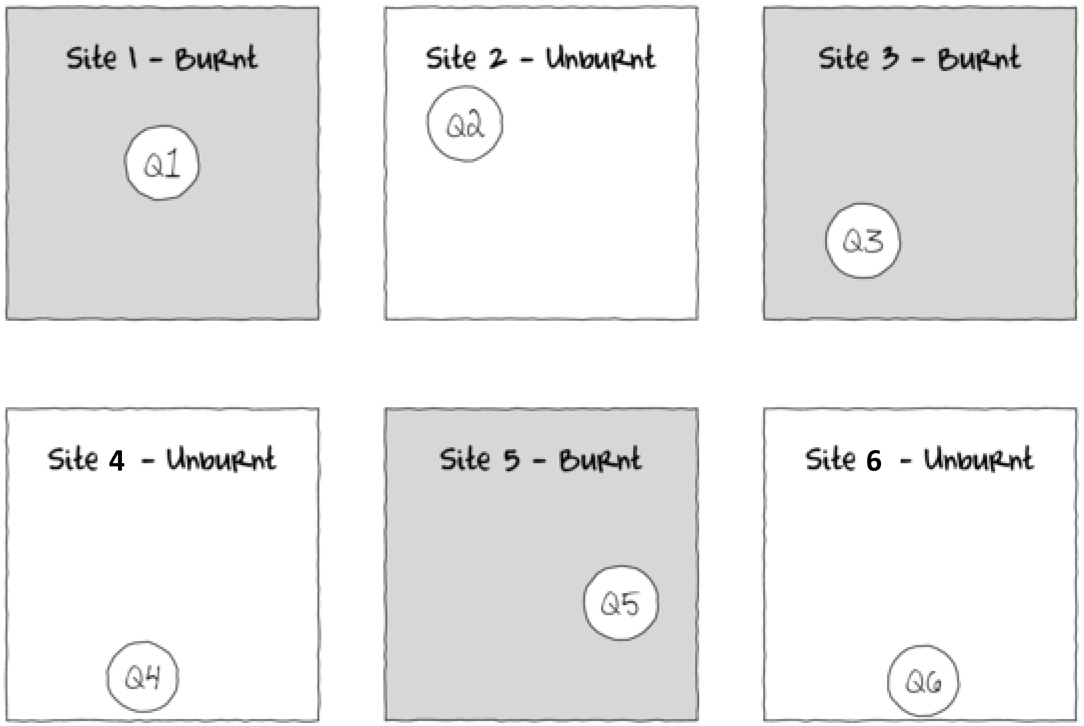
- Hierarchical/multilevel designs help us reduce unexplained variation across space, time, or experimental subjects

Some examples of commonly-found hierarchies in familiar data types

- Quadrats are nested within sites
- Sites are nested within ecosystem/biome
- Individual caves are nested in cave type
- Observations are nested within a quadrat
- Time Series data
- Plots are nested within block
- Individuals are nested within genetic families
- Species are nested within functional groups or phylogenetic clades (e.g., family)
- Plots are nested within one treatment, but not another

Examples of multilevel experimental designs

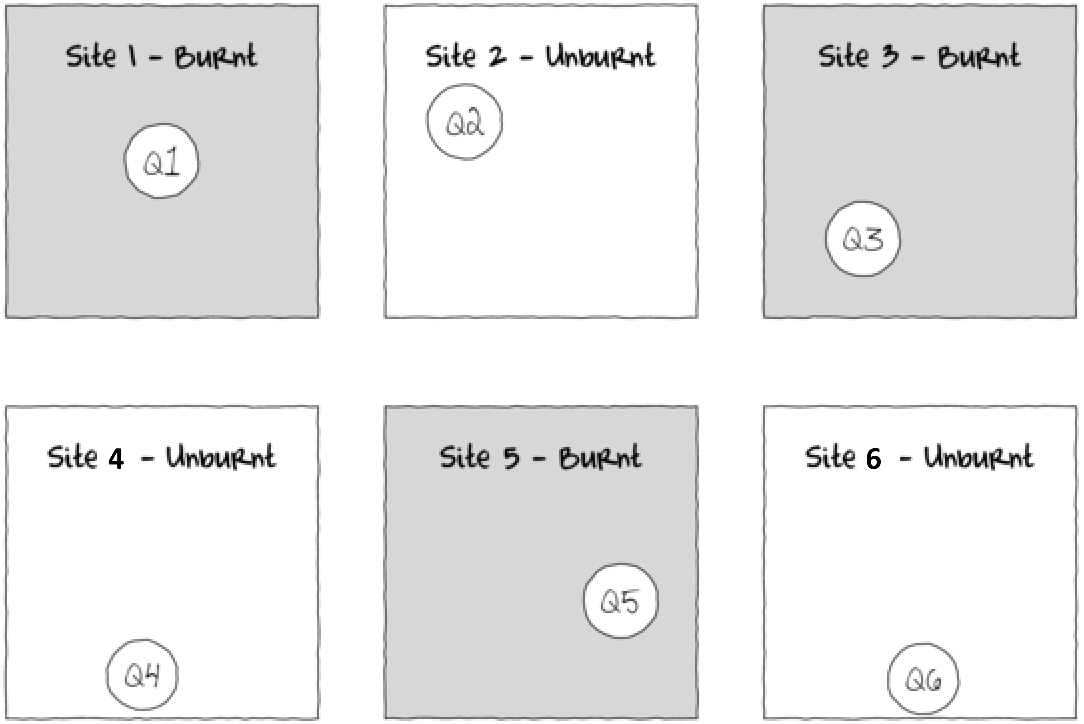
Q: What is the effect of fire on vegetation?



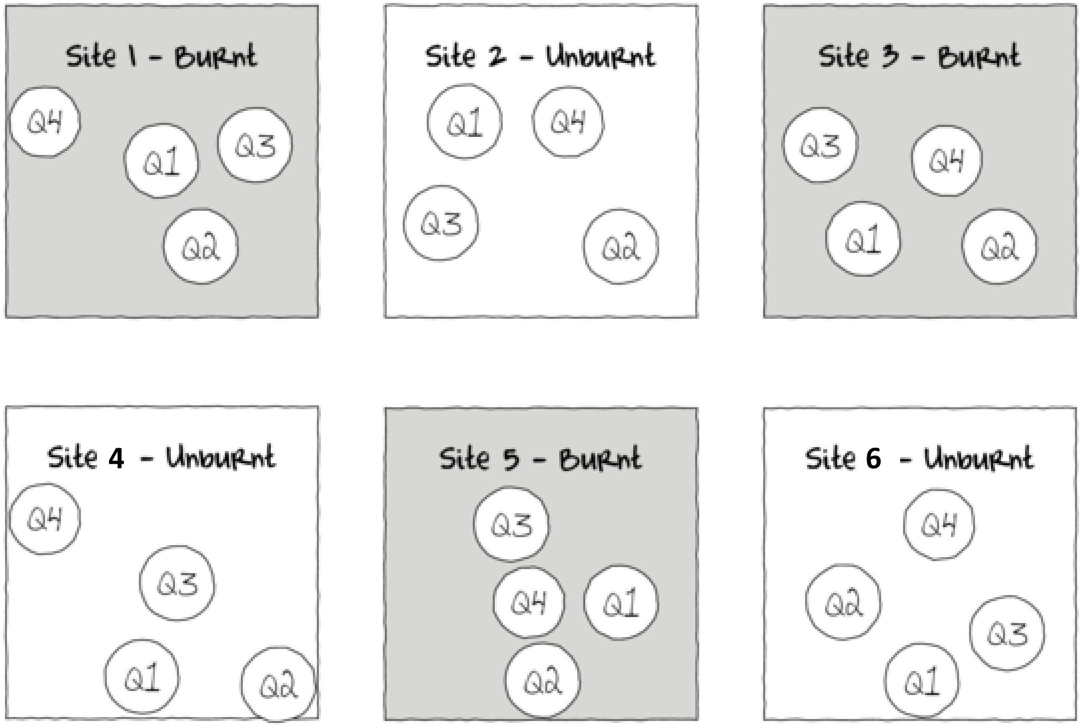
Experimental unit?
Level of replication for the fire treatment?

Examples of multilevel experimental designs

Q: What is the effect of fire on vegetation?



4 quadrats *nested* within each site



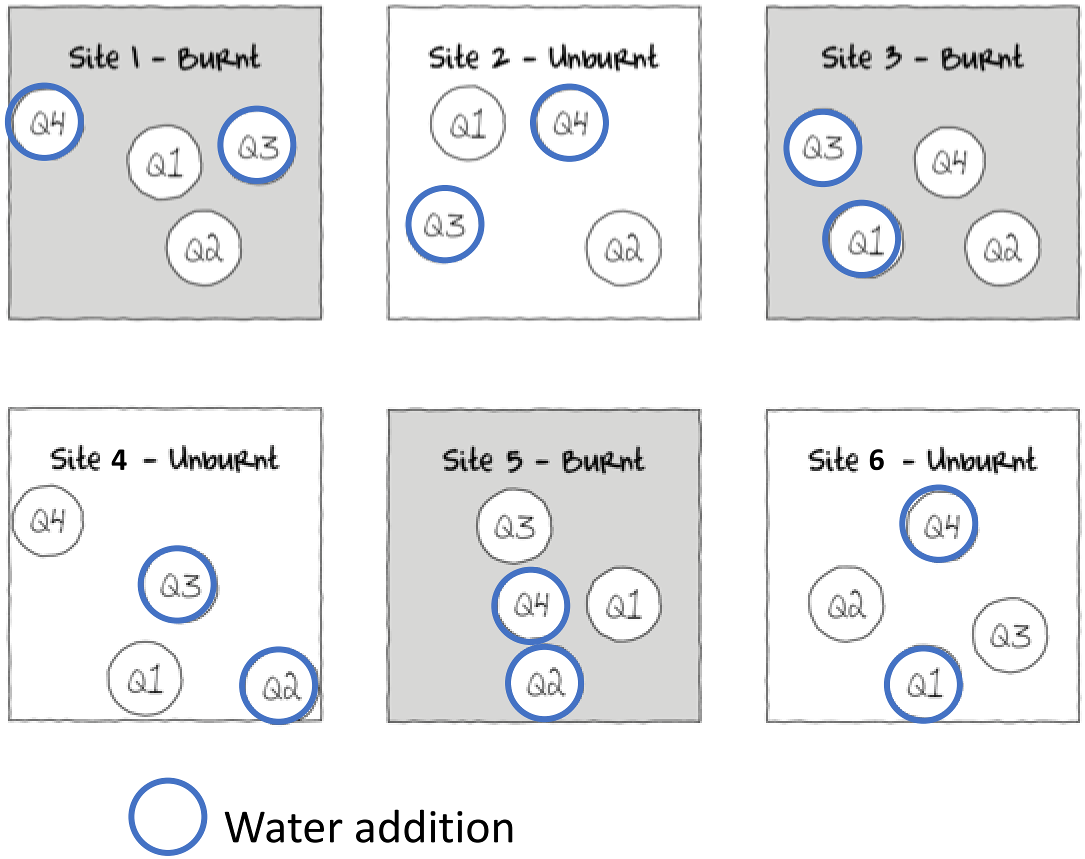
Experimental unit?

Level of replication for the fire treatment?

Unit of observation?

Examples of multilevel experimental designs

2 x 2 factorial experiment
(split plot design)

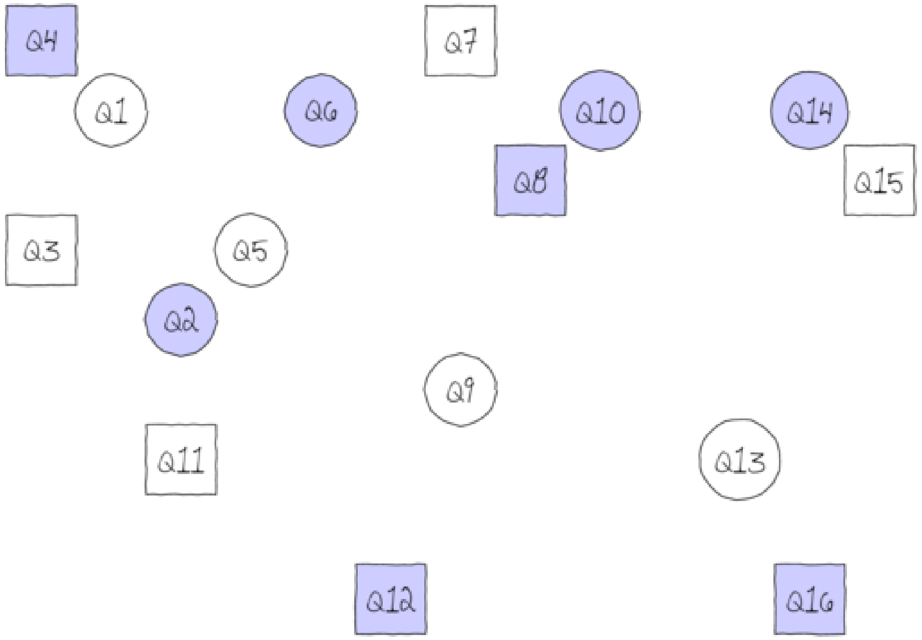


Experimental unit?
Level of replication?
Nesting?

Why do this?

Examples of multilevel experimental designs

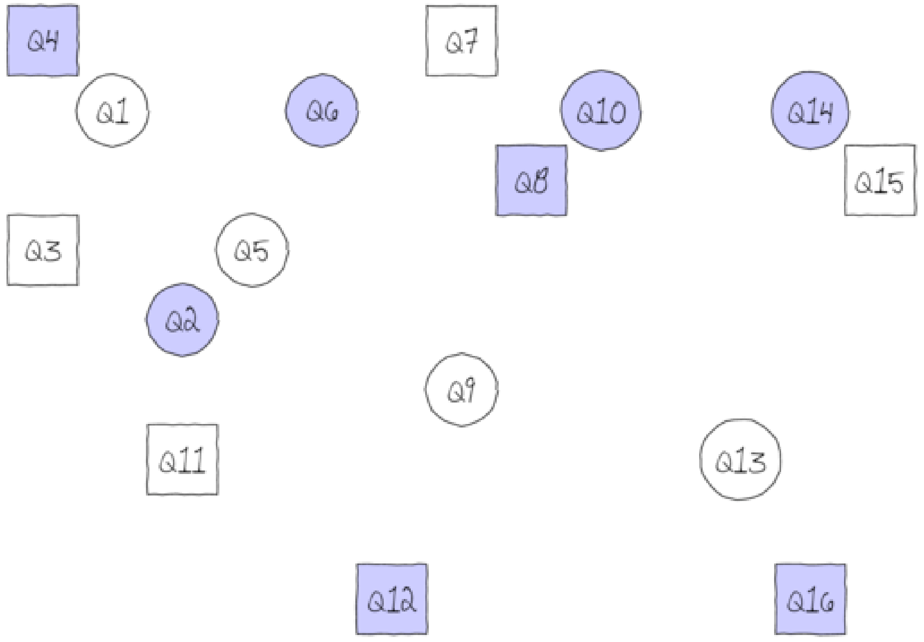
2 x 2 factorial experiment
(shape is one factor, color is another factor)



Experimental unit?
Level of replication?

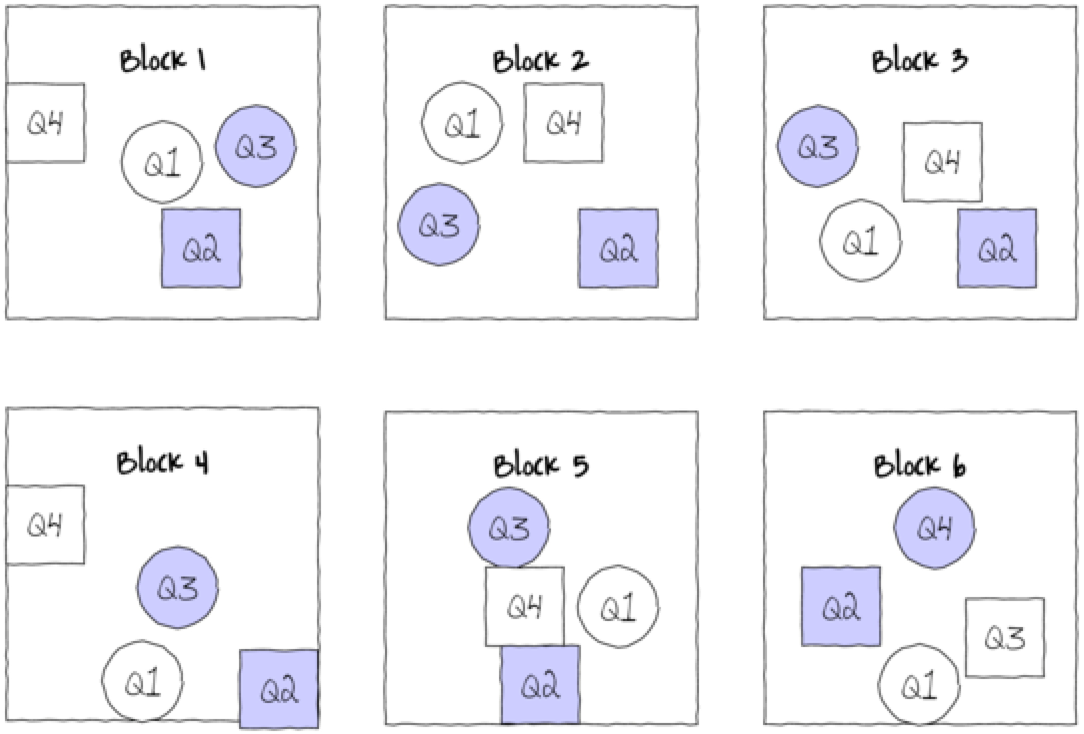
Examples of multilevel experimental designs

2 x 2 factorial experiment
(shape is one factor, color is another factor)



Experimental unit?
Level of replication?

2 x 2 factorial experiment
Randomized complete block design



Break

Often violation of independence comes from some sort of hierarchy in the data

- Hierarchical/multilevel designs help us reduce unexplained variation across space, time, or experimental subjects
- However, need to be treated with care in statistical analysis

Homogeneity of variance

$$y_i = \underbrace{\beta_0 + \beta_1 \times x_i}_{\text{Linearity}} + \varepsilon_i \qquad \varepsilon_i \sim \underbrace{\mathcal{N}(0, \sigma^2)}_{\text{Normality}} \qquad \mathbf{V} = cov = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \cdots & \sigma^2 & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{pmatrix}$$

Zero covariance (=independence)

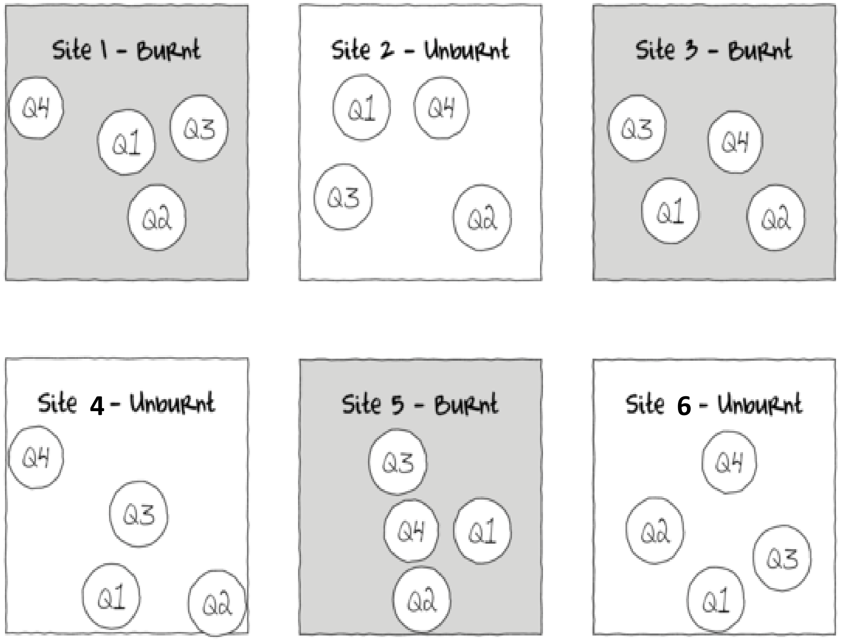
We are usually interested in the variation due to specific factors *in spite of* underlying patterns in noise.

We can model the dependence between variance due underlying patterns in noise and our factors of interest by incorporating them as random effects.

Models that include a *mix of fixed and random effects* are called *mixed models*

Benefits of mixed models

- Accurately represent the original design with your statistics
- Get statistical tests for comparing complex designs across… species, experiments, biomes, etc.
- Don’t lose information (e.g., variance) by averaging
- Have more degrees of freedom
- Ability to make predictions for unmeasured groups



What are fixed vs. random effects?

1. Fixed effects are constant across individuals, and random effects vary. For example, in a growth study, a model with random intercepts α_i and fixed slope β corresponds to parallel lines for different individuals i, or the model $y_{it} = \alpha_i + \beta t$ (Kreft and de Leeuw 1998).
2. Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population (Searle, Casella and McCulloch 1992).
3. "When a sample exhausts the population, the corresponding variable is fixed; when the sample is a small (i.e., negligible) part of the population the corresponding variable is random" (Green and Tukey 1960).
4. "If an effect is assumed to be a realized value of a random variable, it is called a random effect" (LaMotte 1983).
5. Fixed effects are estimated using least squares (or, more generally, maximum likelihood) and random effects are estimated with shrinkage (Snijders and Bosker 1999).

Gelman et al. proposes thinking about random effects as "grouping variables"

Gelman 2005

What is the practical difference?

$$\mathbf{y} = \beta_0 + \sum \beta_i \mathbf{x}_i + \gamma + \varepsilon$$

Response variable

Global intercept

Fixed effect parameters

Fixed effect variables

Random effect variance

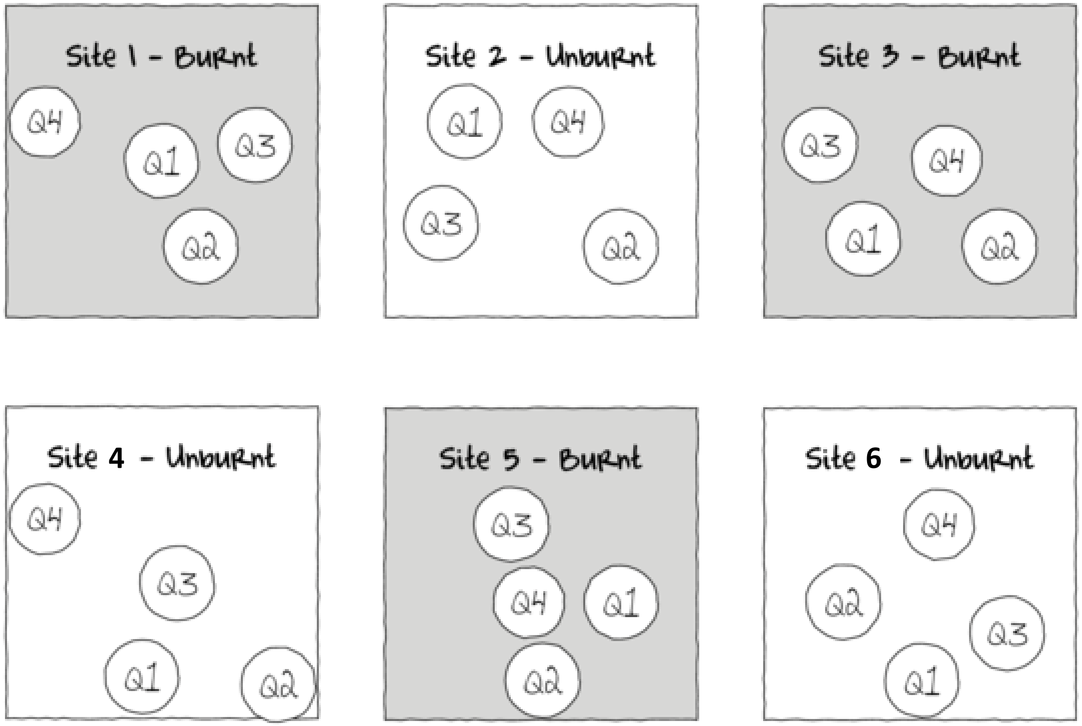
Residual variance

Figure 1: A mathematical and verbal representation of a simple mixed effects model. y is the response variable, β_0 is the global intercept (the expectation of y when all fixed effects are zero, and for members of an average group in the random effect), x_i is the measured value of the ith fixed explanatory variable, β_i is the additive expected change caused by the value of each of the fixed explanatory variables, γ is a draw from the distribution of category means for a normally distributed random effect (with mean of zero and variance equal to the random effect variance), and ε is a draw from the normal distribution of residuals (with mean of zero and variance equal to the residual variance).

(Silk et al. 2020)

Back to the experimental designs earlier: which are the fixed vs. random effects?

4 quadrats *nested* within each site

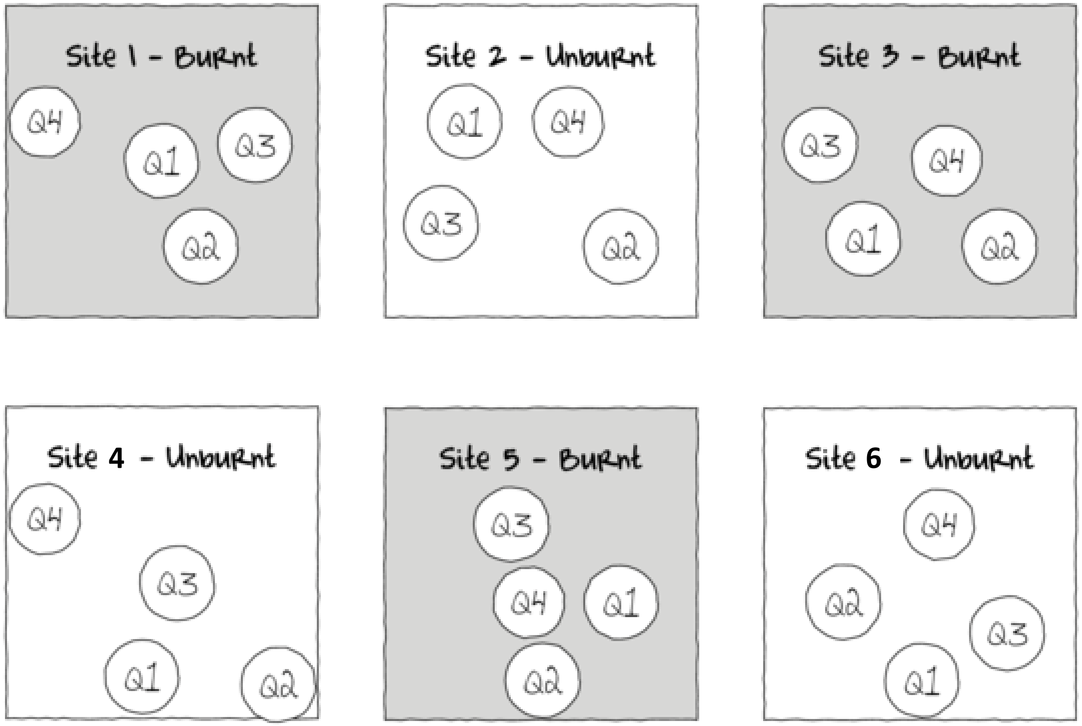


Q: What is the effect of burning on plant composition?

Effect	Fixed	Random	Not sure	Not applicable
Burning				
Site				
Quadrat				

Back to the experimental designs earlier: which are the fixed vs. random effects?

4 quadrats *nested* within each site



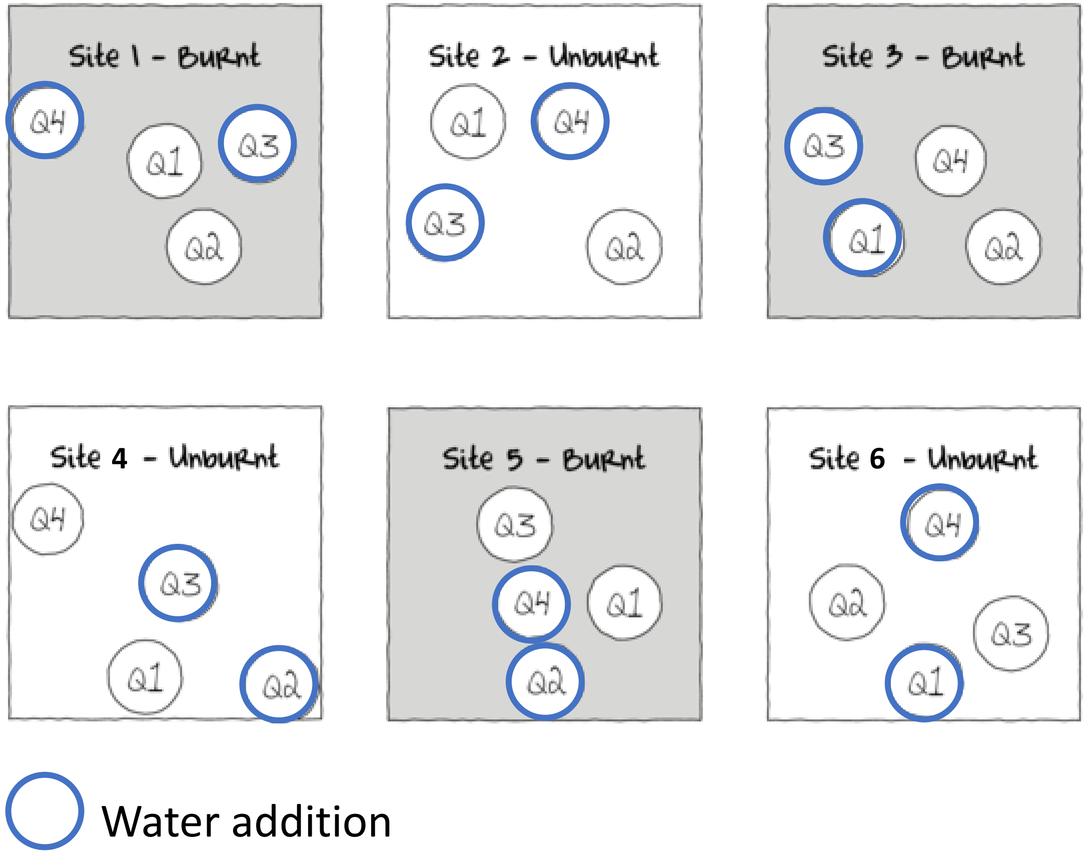
Q: What is the effect of burning on plant composition?

Fixed: burning
Random: site

Quadrat is not involved in model architecture because it is the unit of observation/data point and not a grouping variable

Back to the experimental designs earlier: which are the fixed vs. random effects?

2 x 2 factorial experiment
(split plot design)

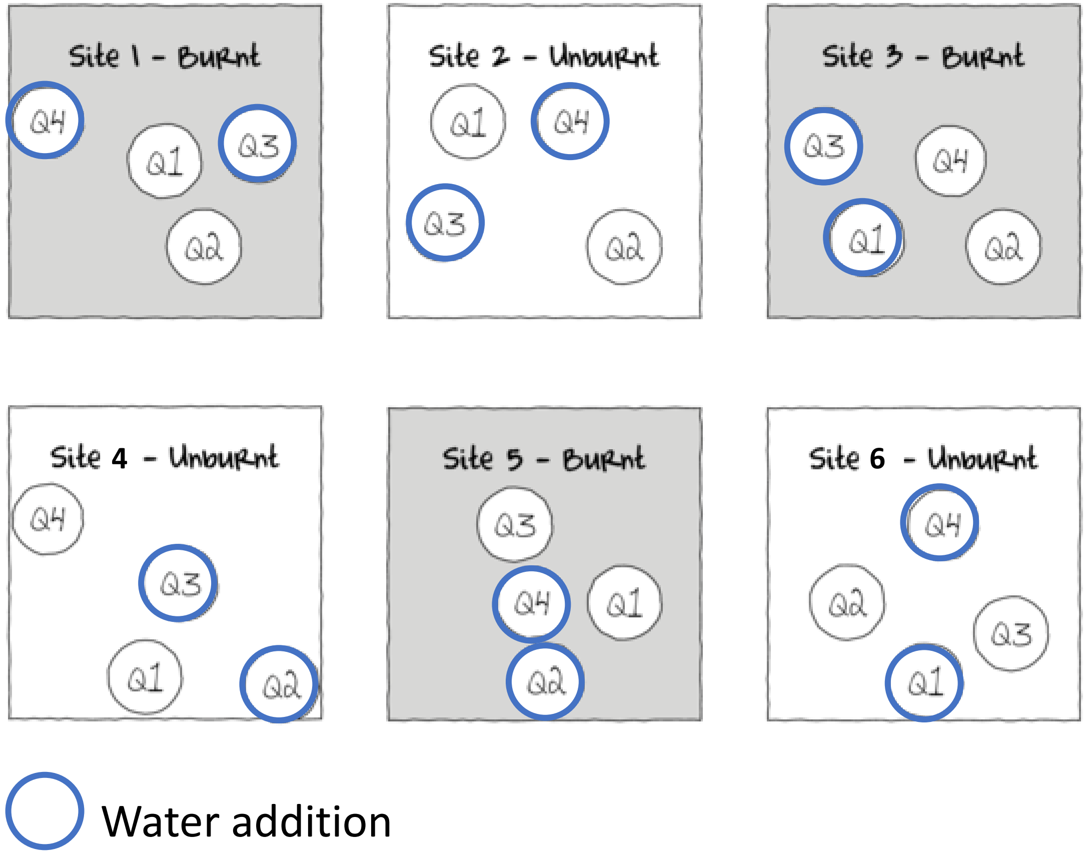


Q: What are the effects of burning and water addition on plant composition?

Effect	Fixed	Random	Not sure	Not applicable
Burning				
Watering				
Burn:Water				
Site				
Quadrat				

Back to the experimental designs earlier: which are the fixed vs. random effects?

2 x 2 factorial experiment
(split plot design)



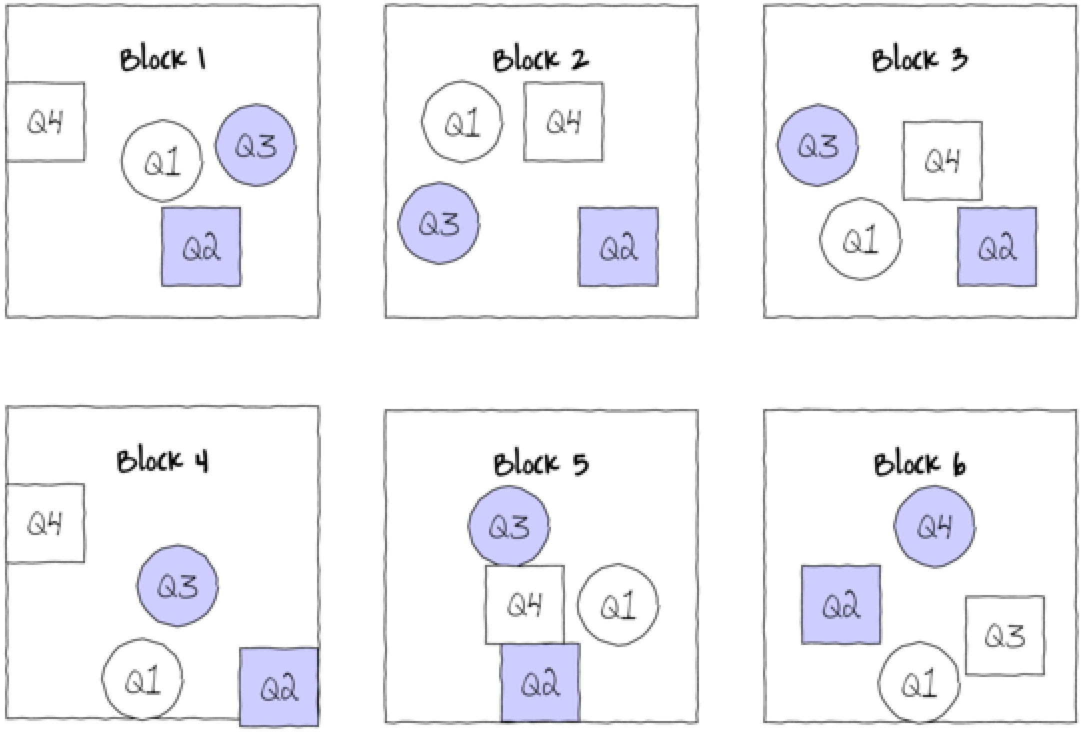
Q: What are the effects of burning and water addition on plant composition?

Fixed:
Burning+ Watering+ Burning:Watering
Or
Burning*Watering

Random: Site

Back to the experimental designs earlier: which are the fixed vs. random effects?

2 x 2 factorial experiment
Randomized complete block design
(shape is factor 1, color is factor 2)

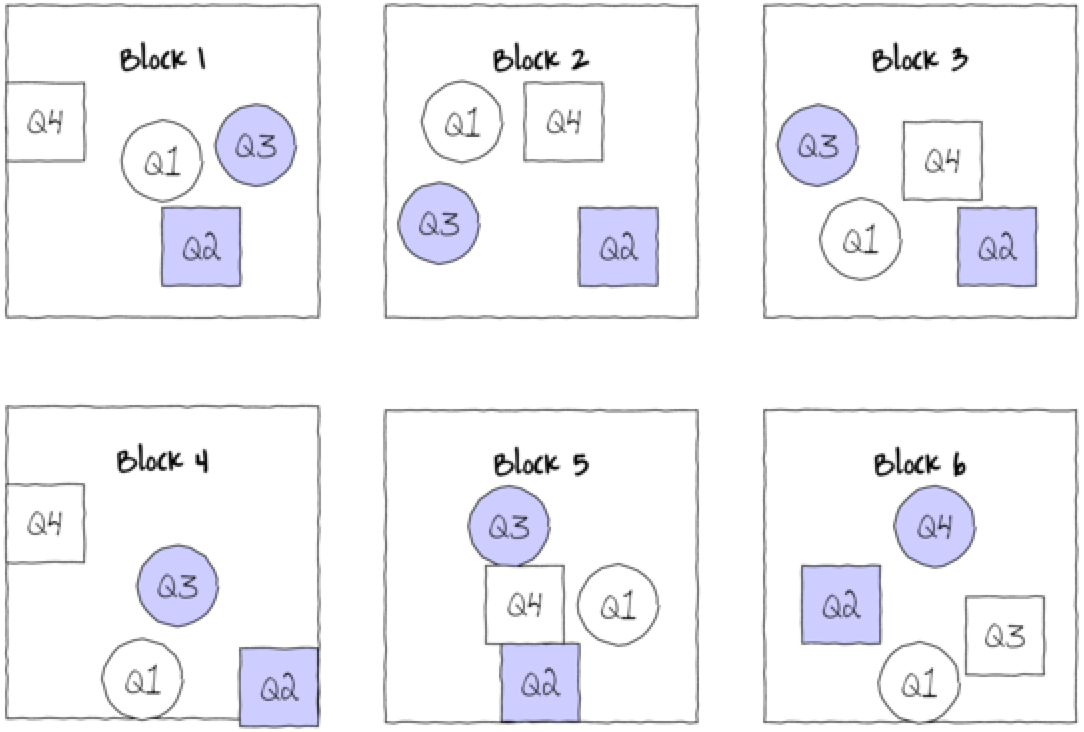


Q: What are the effects of Factor1 and Factor2 on a quadrat-level response?

Effect	Fixed	Random	Not sure	NA
Factor 1				
Factor 2				
Factor1: Factor2				
Block				
Quadrat				

Back to the experimental designs earlier: which are the fixed vs. random effects?

2 x 2 factorial experiment
Randomized complete block design
(shape is factor 1, color is factor 2)

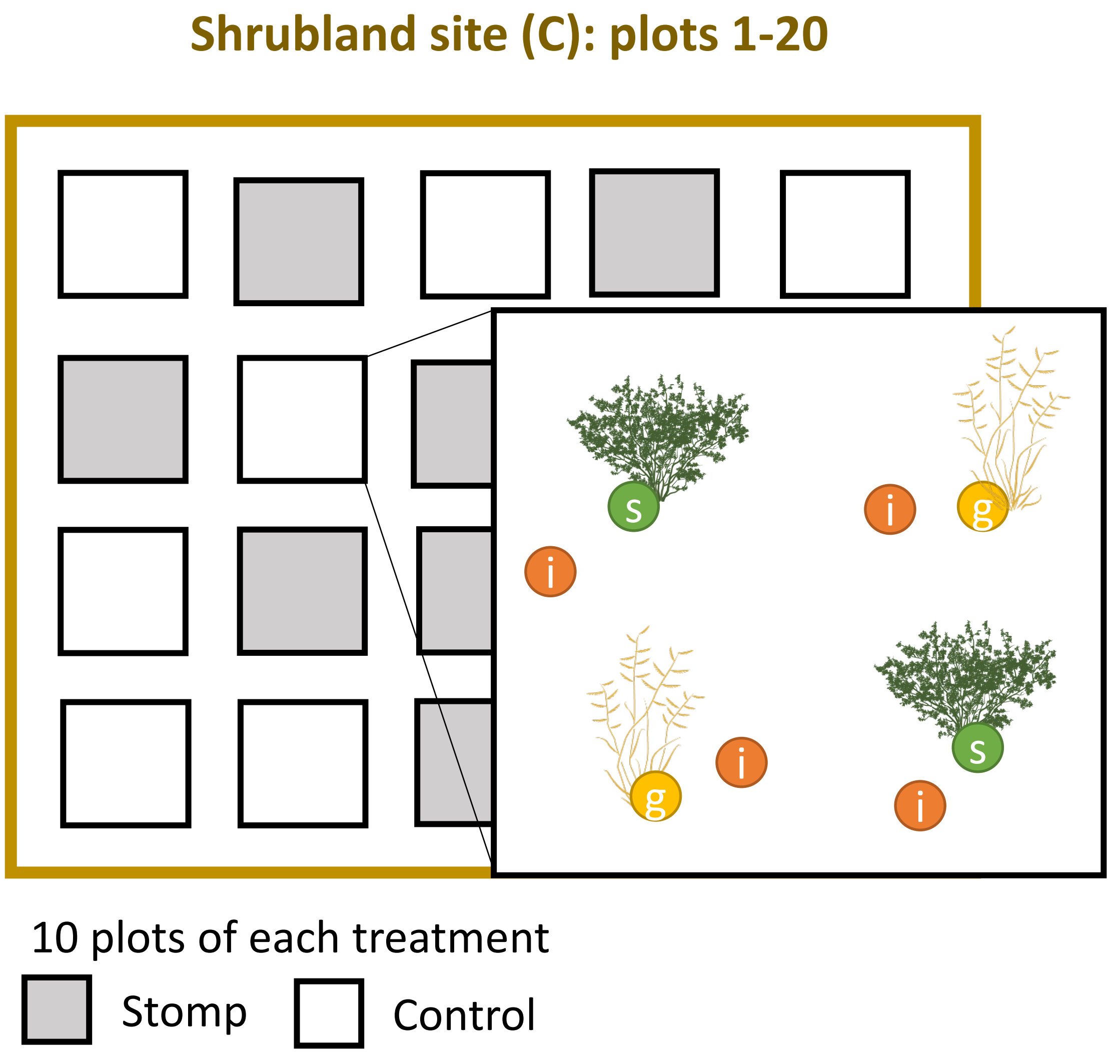


Q: What are the effects of Factor1 and Factor2 on a quadrat-level response?

Fixed:
Factor1+ Factor2+ Factor1:Factor2
Or
Factor1*Factor2

Random: Block

Another example: revisiting the biocrust disturbance experiment



Q: What are the effects of stomping on biocrust cyanobacteria activity among different microsite types in a shrubland?

Effect	Fixed	Random	Not sure	NA
Site				
Stomping				
Plot				
Microsite				
Stomp: Microsite				

Break

Additional things to think about

- Random effect restrictions
- Multiple random effects (crossed vs. nested random effects)
- Random intercepts vs. random slopes
- The same variable can be a fixed or random effect depending on your question

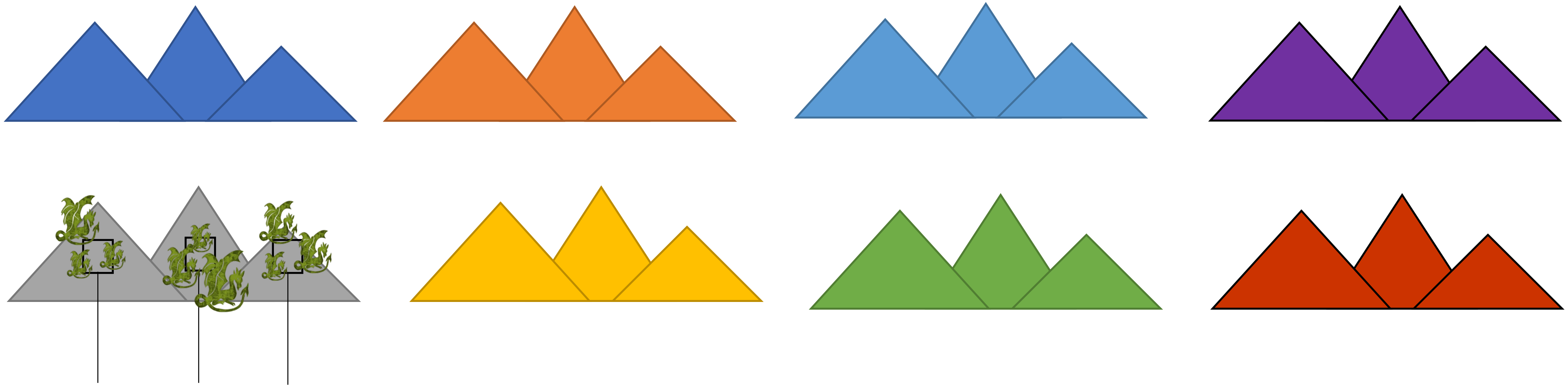
Random effect restrictions

- Remember, we are estimating “distributions” for random effects
- That means random effects should have enough levels to estimate variance
- In the biological sciences people generally say *at least 5 levels*
- In the social sciences folks aim for 30 levels
- Random effects can only be categorical
- Some unbalanced data is okay, but not too unbalanced
- The residual distribution and homoscedasticity assumptions are same as would be applied to the “fixed” part of the model

Multiple random effects: example from reading

8 mountain ranges

Q: Is dragon intelligence dependent on dragon body length?



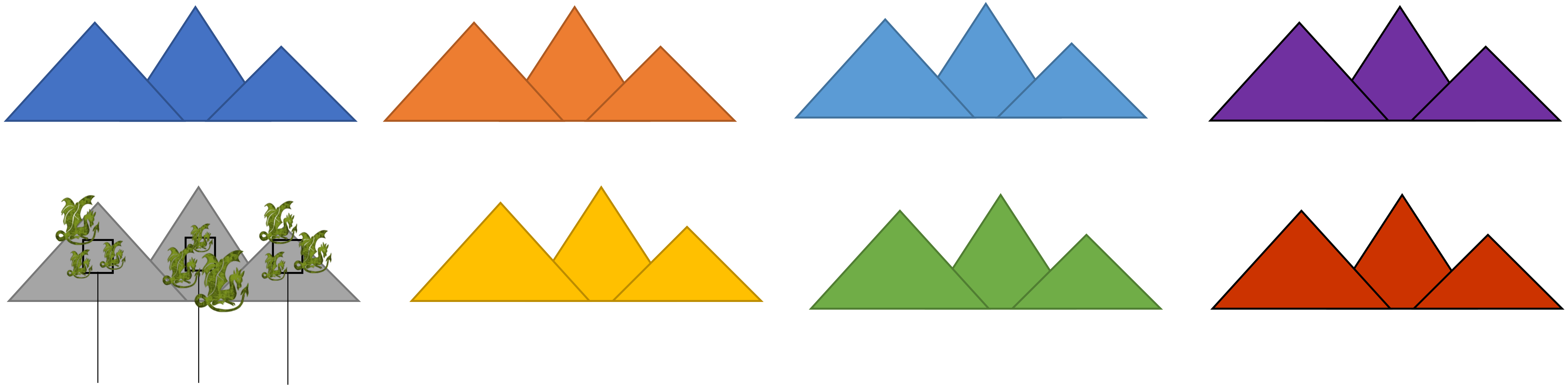
3 sites per mountain range

At each site, sample many dragons of different body lengths and assess their intelligence

Multiple random effects: example from reading

8 mountain ranges

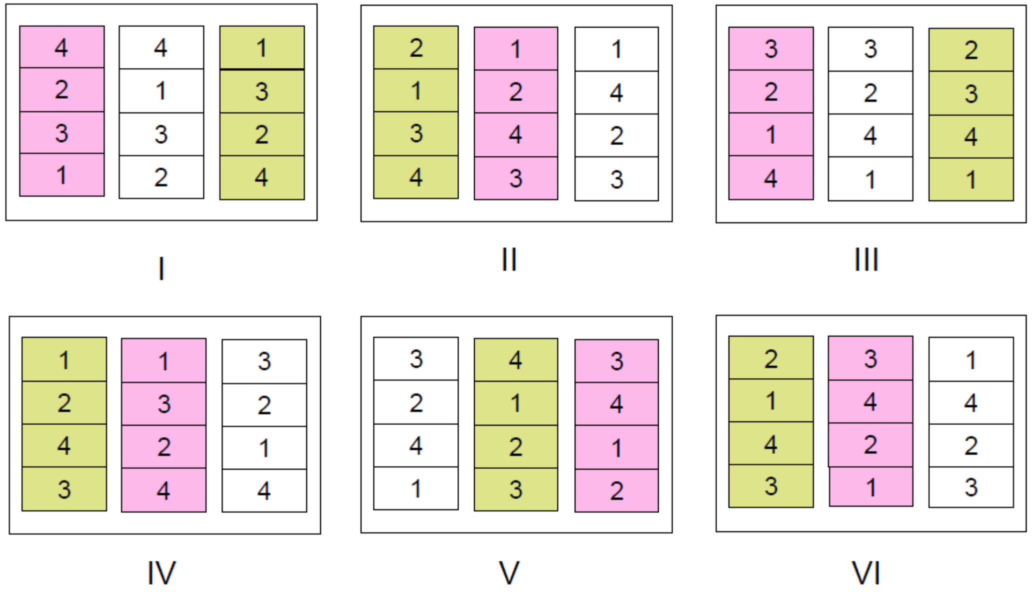
Q: Is dragon intelligence dependent on dragon body length?



3 sites per mountain range

- Mountain Range and Site are both random effects
- Site is nested within Mountain Range
- Total number of random levels to be estimated: 3X8=24
- The way to express this in code depends on how you named your sites

Practice: Blocked split-plot design



I-VI: Six spatial blocks
Colors: 3 different oat varieties
Plots: each rectangular strip containing 1 variety
1-4: 4 nitrogen fertilization levels
Subplot: each small rectangle (unit of yield measurement)

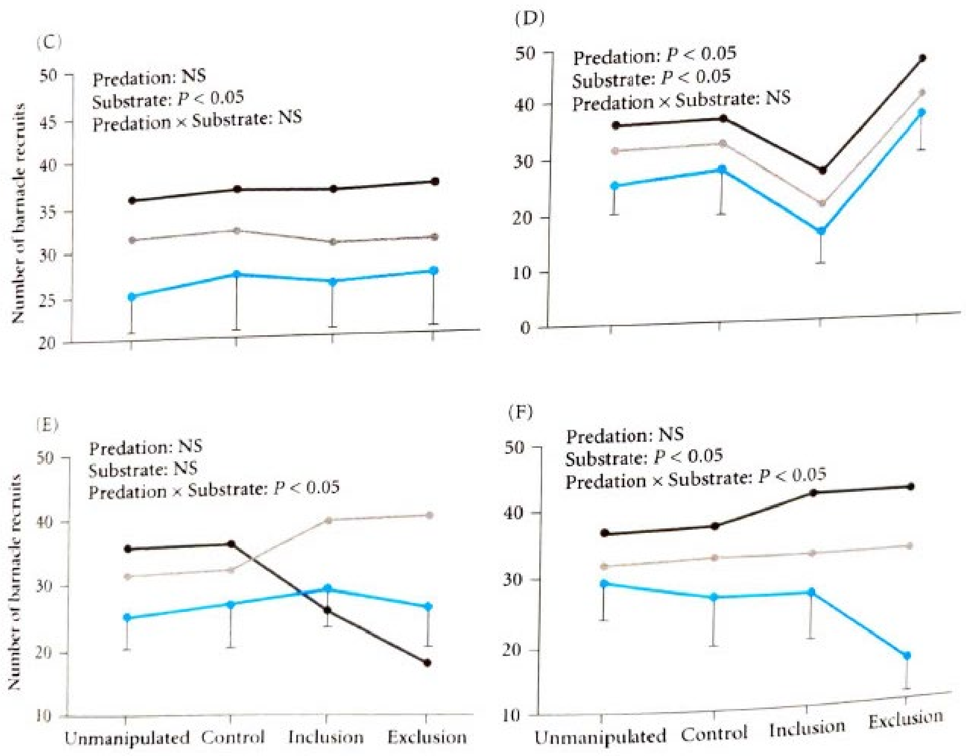
Q: What are the effects of oat variety and N fertilization on yield?

Effect	Fixed	Random	Not sure	NA
Block				
Variety				
Plot				
Nitrogen				
Variety: Nitrogen				
Subplot				

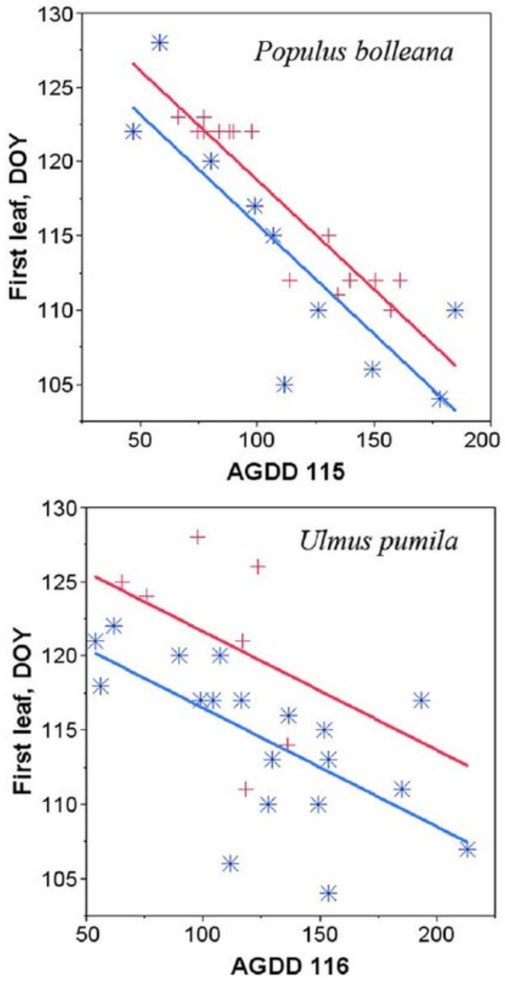
What is the random effects structure?

Random intercepts vs. random slopes

Recall:
Main effects vs. interactions

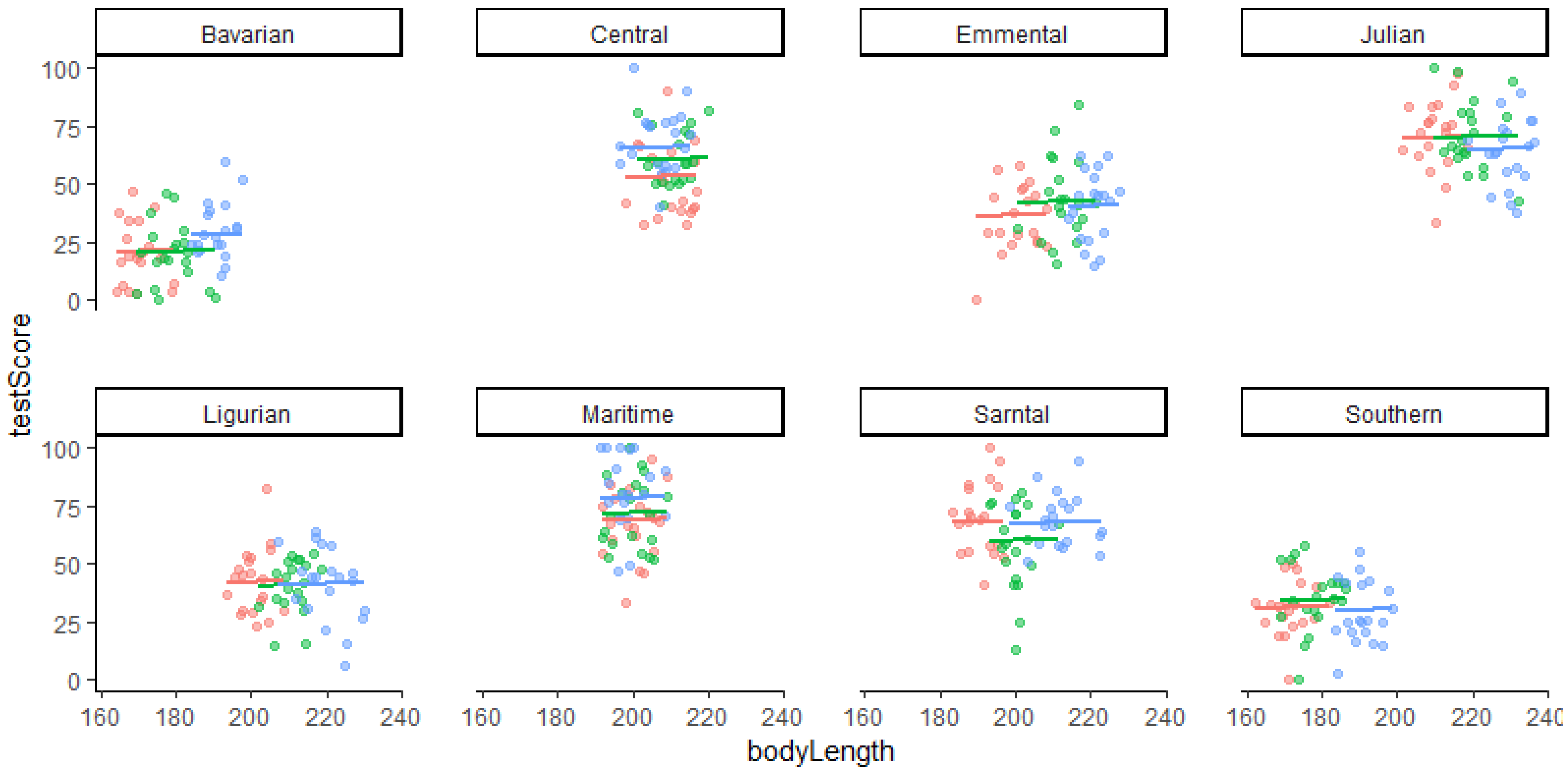


Species:AGDD interaction

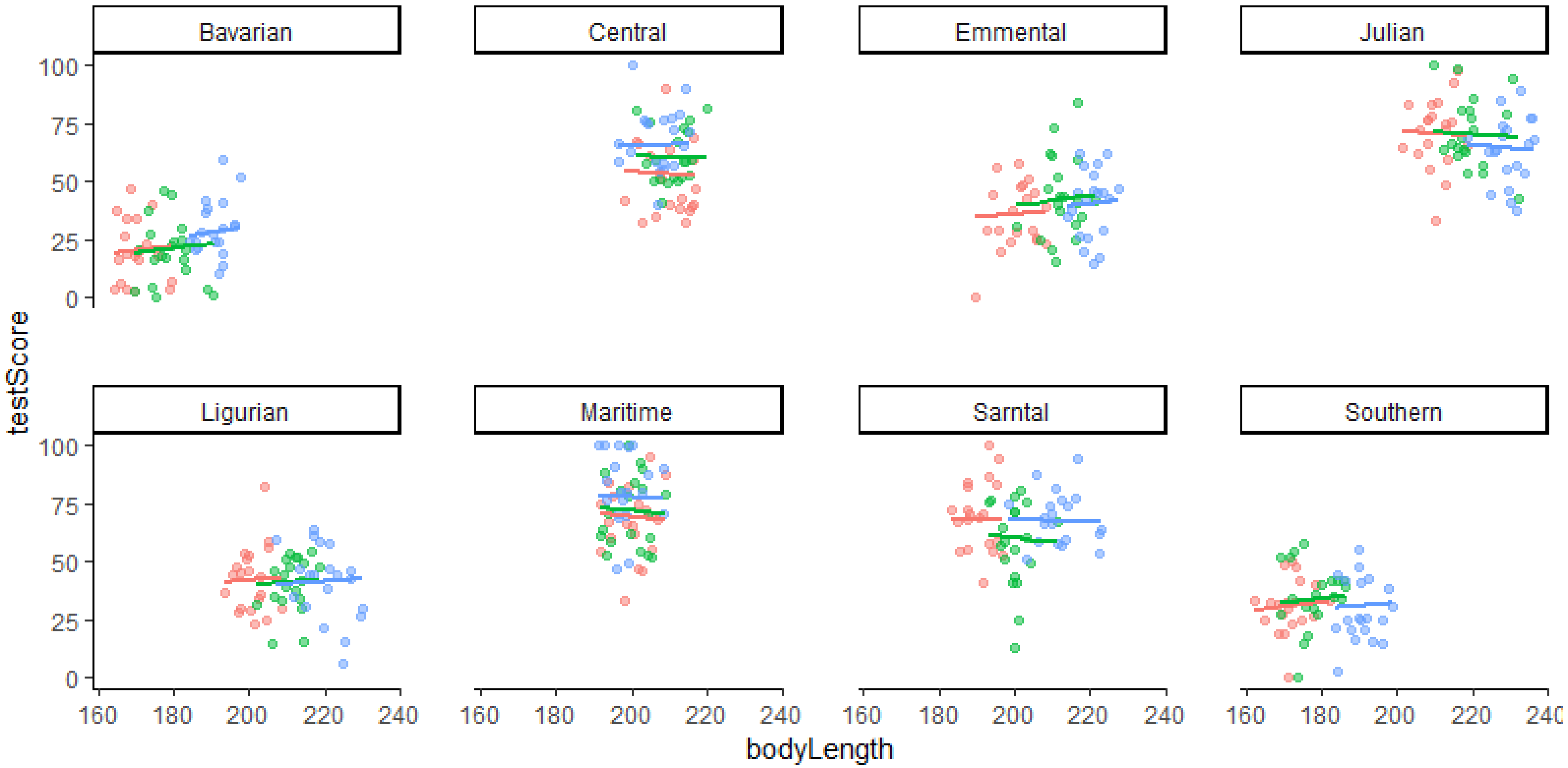


For categorical predictors, main effects can also be thought of as affecting the *intercept*, and interactions as affecting the *slope*.

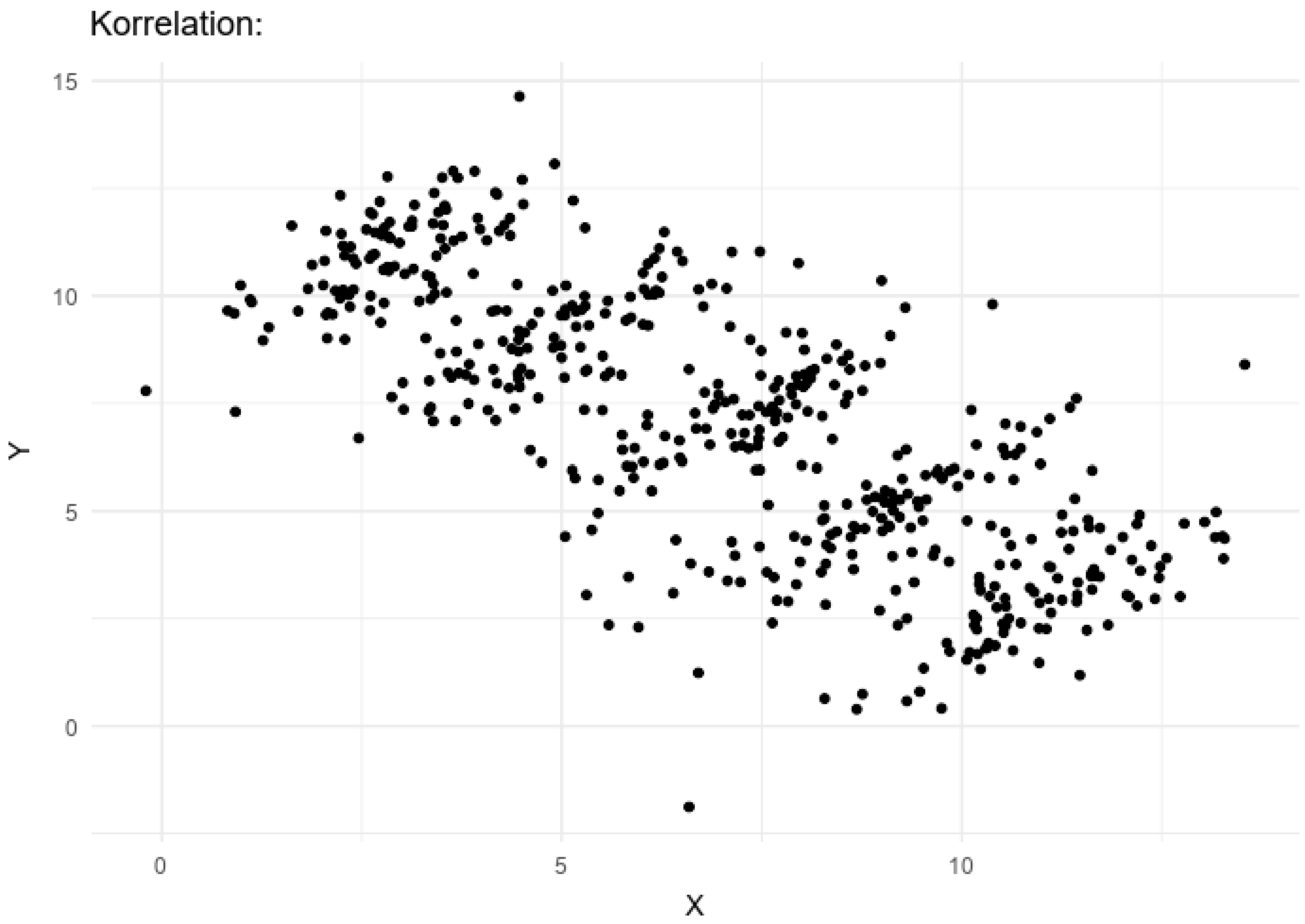
Example of random intercept model



Example of random intercepts + slopes model



“Simpson’s Paradox”



Many variables can fit within both “fixed” and “random” classes

Example: Time as fixed vs. random, depending on your question

Data: A time series of plant biomass at the same site throughout the years

Q: Does plant biomass vary with precipitation?

→year as *random effect* to account for non-independence of observations from the same year

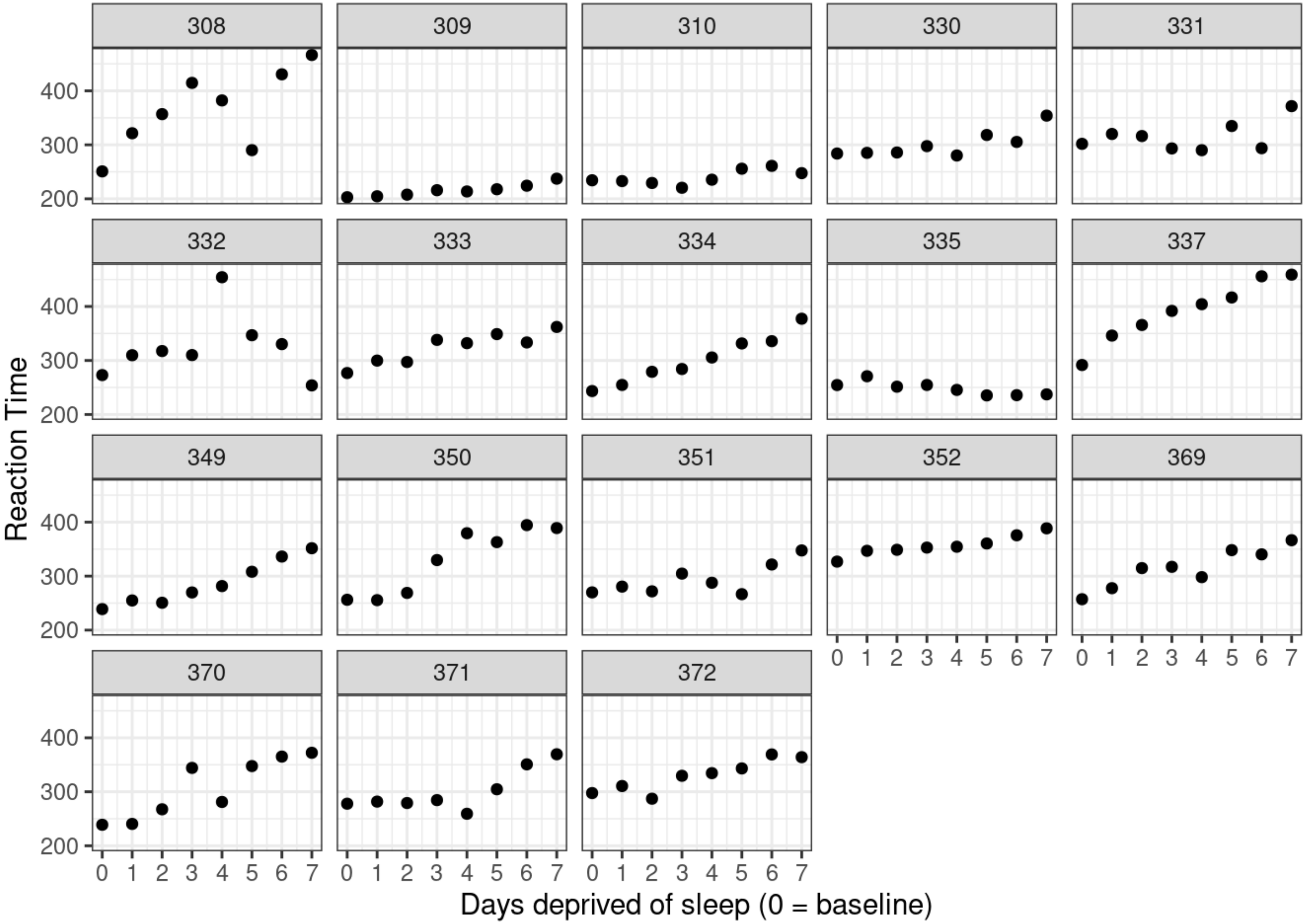
BIOMASS ~ Precipitation + (1|YEAR)

Q: Does plant biomass increase/decrease over time?

→year as *fixed effect* to test influence of specific years

BIOMASS ~ YEAR   (fits a linear regression, is slope non-zero?)

Brief overview of mechanics



Response variable: Reaction time (Y_{sd})
Fixed effect: Days of sleep deprivation (*d*)
Random effect: Subject (*s*)

If we ignored the random effect and just fit one big model (also called complete pooling), the lm model would be expressed like this:

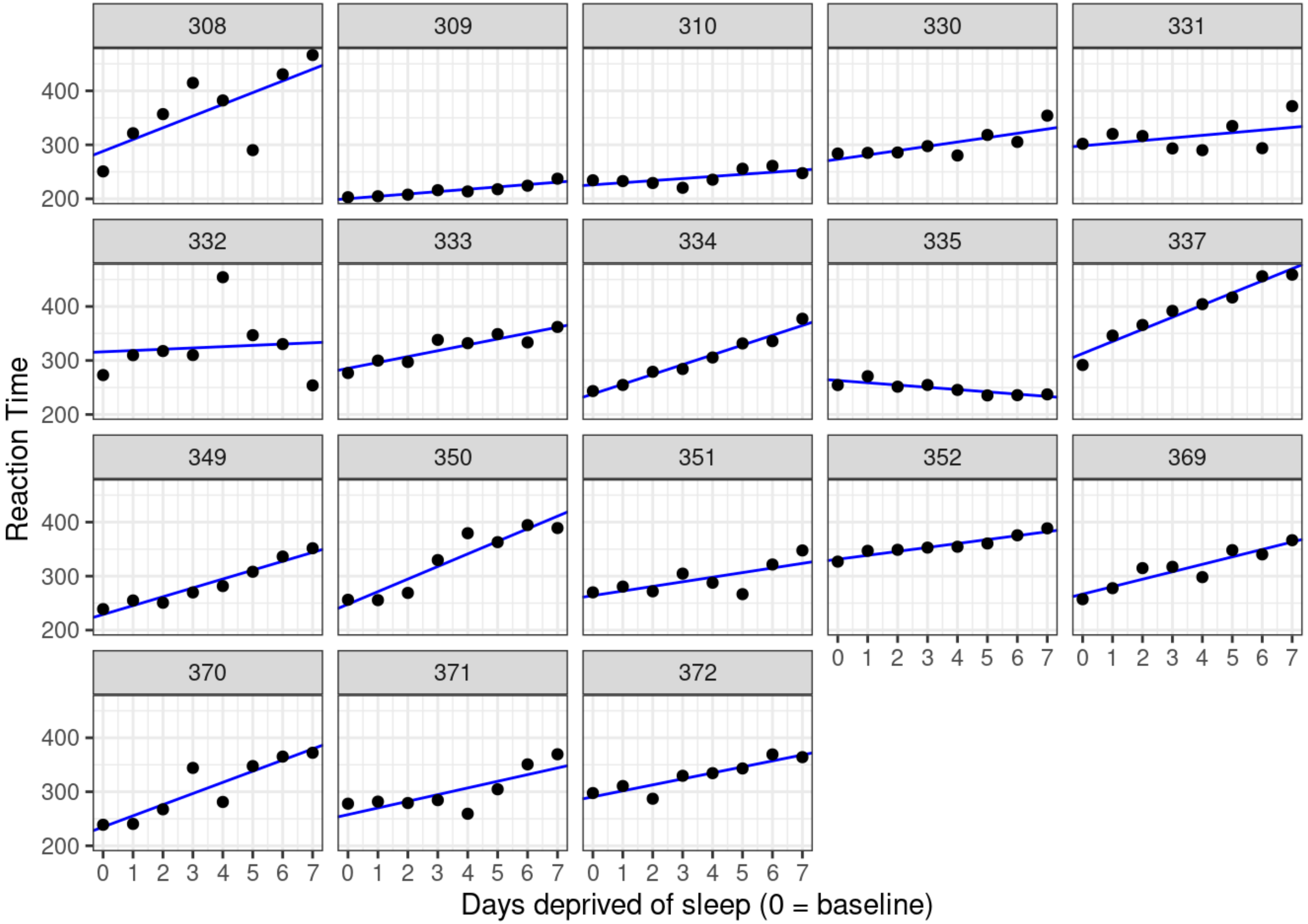
$$Y_{sd} = \beta_0 + \beta_1 X_{sd} + e_{sd}$$

$$e_{sd} \sim N\left(0, \sigma^2\right)$$

Obviously not ideal, since it violates independence

(Example from Dunn 2020 Ch 5, data from Belenky et al. 2003)

Brief overview of mechanics



Response variable: Reaction time (Y_{sd})
Fixed effect: Days of sleep deprivation (*d*)
Random effect: Subject (*s*)

Alternatively, we could include Subject (*s*) as a predictor in our model (no pooling):

Reaction Time ~ Days*Subject

However, we this means we are fitting 18 parameters, which takes away a lot of degrees of freedom, and we really don't care about each *specific* subject's result

We want to know what the relationship between reaction time and sleep deprivation days is, *controlling for the variation among subjects*.

Brief overview of mechanics

To include the random effect of subject (*s*), we recognize that the variance should be modeled at 2 levels (partial pooling):

Level 1:

$$Y_{sd} = \beta_{0s} + \beta_{1s} X_{sd} + e_{sd}$$

Brief overview of mechanics

To include the random effect of subject (*s*), we recognize that the variance should be modeled at 2 levels (partial pooling):

Level 1:

$$Y_{sd} = \beta_{0s} + \beta_{1s} X_{sd} + e_{sd}$$

Fixed intercept

Level 2:

$$\beta_{0s} = \gamma_0 + S_{0s}$$

Random intercept offset for every level of subject *s*

$$\beta_{1s} = \gamma_1 + S_{1s}$$

Random slope offset for every level of subject *s*

Fixed slope

Brief overview of mechanics

To include the random effect of subject (*s*), we recognize that the variance should be modeled at 2 levels (partial pooling):

Level 1:

$$Y_{sd} = \beta_{0s} + \beta_{1s} X_{sd} + e_{sd}$$

Level 2:

$$\beta_{0s} = \gamma_0 + S_{0s}$$

$$\beta_{1s} = \gamma_1 + S_{1s}$$

Variance Components:

$$\langle S_{0s}, S_{1s} \rangle \sim N\left(\langle 0, 0 \rangle, \mathbf{\Sigma}\right)$$

$$\mathbf{\Sigma} = \begin{pmatrix} {\tau_{00}}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & {\tau_{11}}^2 \end{pmatrix}$$

$$e_{sd} \sim N\left(0, \sigma^2\right)$$

Random intercept / random slope pairs ⟨S0s,S1s⟩ are drawn from a bivariate normal distribution centered at the origin ⟨0,0⟩ with variance-covariance matrix Σ

- Assuming all group means are drawn from a common distribution causes their estimates to drift towards the global mean.
- Also known as *shrinkage*
- Can also lead to smaller and more precise standard errors around means.

Brief overview of mechanics

To include the random effect of subject (*s*), we recognize that the variance should be modeled at 2 levels (partial pooling):

Level 1:

$$Y_{sd} = \beta_{0s} + \beta_{1s} X_{sd} + e_{sd}$$

Level 2:

$$\beta_{0s} = \gamma_0 + S_{0s}$$

$$\beta_{1s} = \gamma_1 + S_{1s}$$

Random intercept variance, which captures how much subjects vary in their mean response time on Day0

Variance Components:

$$\langle S_{0s}, S_{1s} \rangle \sim N(\langle 0, 0 \rangle, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} = \begin{pmatrix} {\tau_{00}}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & {\tau_{11}}^2 \end{pmatrix}$$

Covariance between random intercepts and slopes (correlation times the square root of variances)

$$e_{sd} \sim N(0, \sigma^2)$$

Random slope variance, which captures how much subjects vary in their susceptibility to the effects of sleep deprivation

Brief overview of mechanics

To include the random effect of subject (s), we recognize that the variance should be modeled at 2 levels (partial pooling):

Level 1:

$$Y_{sd} = \beta_{0s} + \beta_{1s} X_{sd} + e_{sd}$$

Level 2:

$$\beta_{0s} = \gamma_0 + S_{0s}$$

$$\beta_{1s} = \gamma_1 + S_{1s}$$

Variance Components:

$$\langle S_{0s}, S_{1s} \rangle \sim N(\langle 0, 0 \rangle, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} = \begin{pmatrix} {\tau_{00}}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & {\tau_{11}}^2 \end{pmatrix}$$

$$e_{sd} \sim N\left(0, \sigma^2\right)$$

Random intercept variance, which captures how much subjects vary in their mean response time on Day0

Covariance between random intercepts and slopes (correlation times the square root of variances)

Random slope variance, which captures how much subjects vary in their susceptibility to the effects of sleep deprivation

$$\mathbf{\Sigma} = \begin{pmatrix} {\tau_{00}}^2 & 0 \\ 0 & 0 \end{pmatrix}$$ Random intercepts only

$$\mathbf{\Sigma} = \begin{pmatrix} {\tau_{00}}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & {\tau_{11}}^2 \end{pmatrix}$$ Random intercepts and slopes

$$\mathbf{\Sigma} = \begin{pmatrix} 0 & 0 \\ 0 & {\tau_{11}}^2 \end{pmatrix}$$ Random slopes only

Brief overview of mechanics

To include the random effect of subject (*s*), we recognize that the variance should be modeled at 2 levels (partial pooling):

Level 1:

$$Y_{sd} = \beta_{0s} + \beta_{1s} X_{sd} + e_{sd}$$

Level 2:

$$\beta_{0s} = \gamma_0 + S_{0s}$$

$$\beta_{1s} = \gamma_1 + S_{1s}$$

Variance Components:

$$\langle S_{0s}, S_{1s} \rangle \sim N\left(\langle 0, 0 \rangle, \mathbf{\Sigma}\right)$$

$$\mathbf{\Sigma} = \begin{pmatrix} {\tau_{00}}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & {\tau_{11}}^2 \end{pmatrix}$$

$$e_{sd} \sim N\left(0, \sigma^2\right)$$

Variable	Type	Description
Y_{sd}	observed	Value of `Reaction` for subject s on day d
X_{sd}	observed	Value of `Day` (0-9) for subject s on day d
β_{0s}	derived	level 1 intercept parameter
β_{1s}	derived	level 1 slope parameter
e_{sd}	derived	Residual (Y_{sd} - $\hat{Y}_{sd}$) for subject s, day d
γ_0	fixed	Grand intercept (“gamma”)
γ_1	fixed	Grand slope (“gamma”)
S_{0s}	derived	Random intercept (offset) for subject s
S_{1s}	derived	Random slope (offset) for subject s
$\mathbf{\Sigma}$	random	Variance-covariance matrix
${\tau_{00}}^2$	random	Variance of random intercepts
ρ	random	Random correlation between intercepts and slopes
${\tau_{11}}^2$	random	Variance of random slopes
σ^2	random	Error variance

Brief overview of mechanics

To include the random effect of subject (*s*), we recognize that the variance should be modeled at 2 levels (partial pooling):

Level 1:

$$Y_{sd} = \beta_{0s} + \beta_{1s} X_{sd} + e_{sd}$$

Level 2:

$$\beta_{0s} = \gamma_0 + S_{0s}$$

$$\beta_{1s} = \gamma_1 + S_{1s}$$

$$Y_{sd} = \underbrace{\gamma_0}_{\text{Fixed intercept}} + \underbrace{S_{0s}}_{\text{Random intercept offset}} + \left(\underbrace{\gamma_1}_{\text{Fixed slope}} + \underbrace{S_{1s}}_{\text{Random slope offset}}\right) X_{sd} + \underbrace{e_{sd}}_{\text{Residual}}$$

Variance Components:

$$\langle S_{0s}, S_{1s} \rangle \sim N\left(\langle 0, 0 \rangle, \mathbf{\Sigma}\right)$$

$$\mathbf{\Sigma} = \begin{pmatrix} {\tau_{00}}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & {\tau_{11}}^2 \end{pmatrix}$$

$$e_{sd} \sim N\left(0, \sigma^2\right)$$

This week

- Why do we need mixed models?
- What are fixed vs. random effects?
- Think about how fixed/random effects relate to experimental design
- Practice identifying fixed vs. random effects
- Understand the architecture of multilevel/mixed models