FES 524: Natural Resources Data Analysis

Reading 5.1

**Contents**

[Other readings to do before class 1](#_Toc62631676)

[Studies with different sizes of physical units 1](#_Toc62631677)

[Nested vs crossed factors 1](#_Toc62631678)

[Crossed factors 1](#_Toc62631679)

[Nested factors 2](#_Toc62631680)

[Partial crossing 2](#_Toc62631681)

[Crossed random effects 3](#_Toc62631682)

[Explicit naming 3](#_Toc62631683)

[Missing column for physical units 3](#_Toc62631684)

[Making explicit names 4](#_Toc62631685)

[Statistical model and analysis 4](#_Toc62631686)

[Random effects 4](#_Toc62631687)

[Overall tests 6](#_Toc62631688)

# Other readings to do before class

Read Week 5 Handout 1: “Study on total microbial biomass under different overstory species” and come prepared to use this example to talk about the topics covered over the past few weeks as well as in reading 5.1. Make sure you’ve had time to think about the answers to the questions in that handout.

Week 5 Handout 3 is a companion to the section of this reading that touches on statistical models.

# Studies with different sizes of physical units

This week we’ll be focusing on study designs where there are different sizes of physical units within the same study. Sometimes having different sizes of physical units is due to subsampling, such as the designs we saw in week 2, and we can average over the smallest physical units to simplify the analysis. However, this week our examples will have different factors of interest measured on or applied to physical units of different sizes.

One term you will see for studies like we’ll see this week, coming from experimental design language, is *split plot*. The examples this week are actually examples of *blocked split plot* designs. However, it is unnecessary to use this sort of jargon to describe a study. It is better practice to clearly describe the study design so anyone can understand what was done without using this statistical jargon. Even if you do include this kind of language, which is common in some fields, make sure you still thoroughly describe the design rather than relying on this terminology.

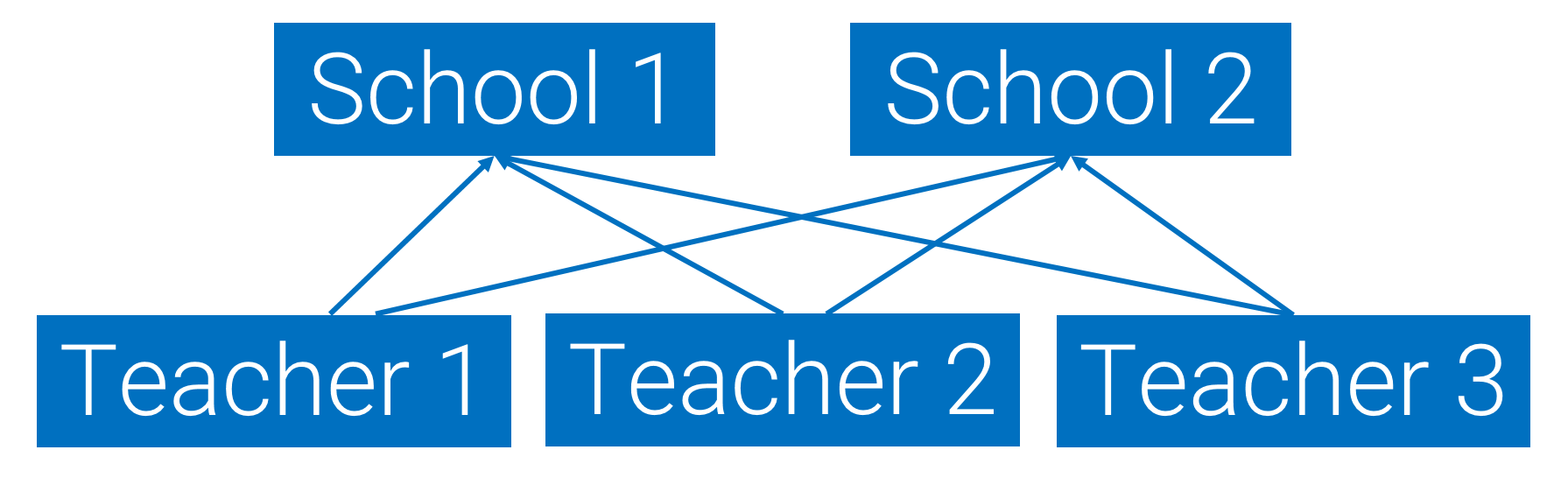
# Nested vs crossed factors

This week we will see both nested and crossed factors.

## Crossed factors

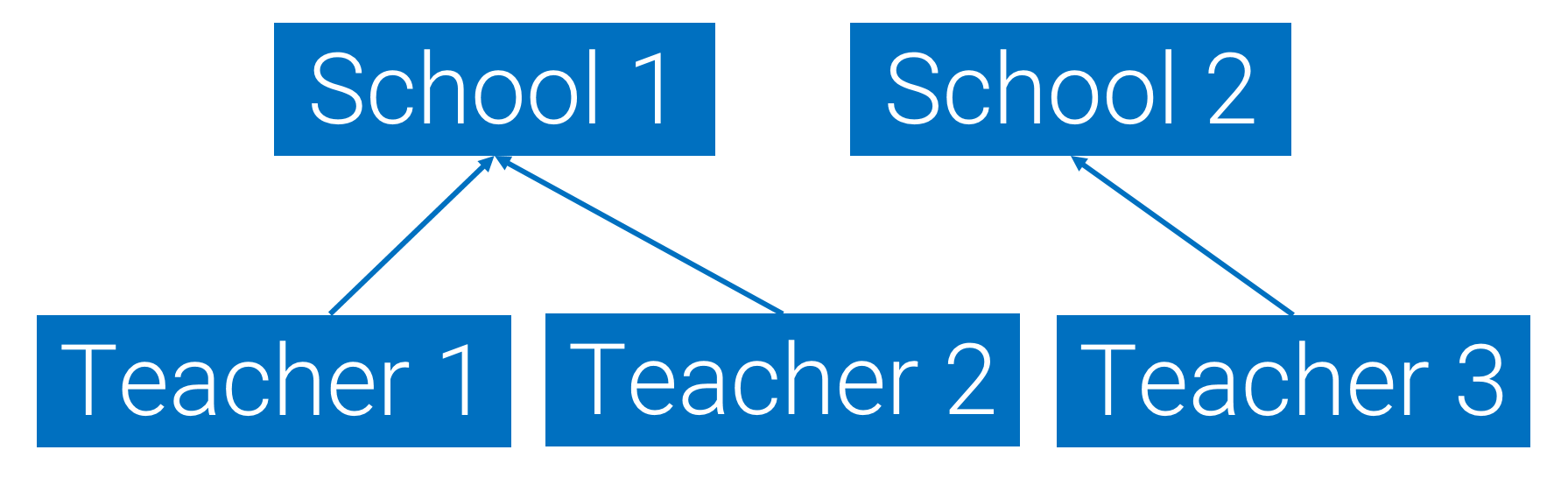
We formally discussed crossed factors last week. Remember that in crossing all combinations of factor levels are present.

Here is an example, showing that all three teachers taught in the two schools.



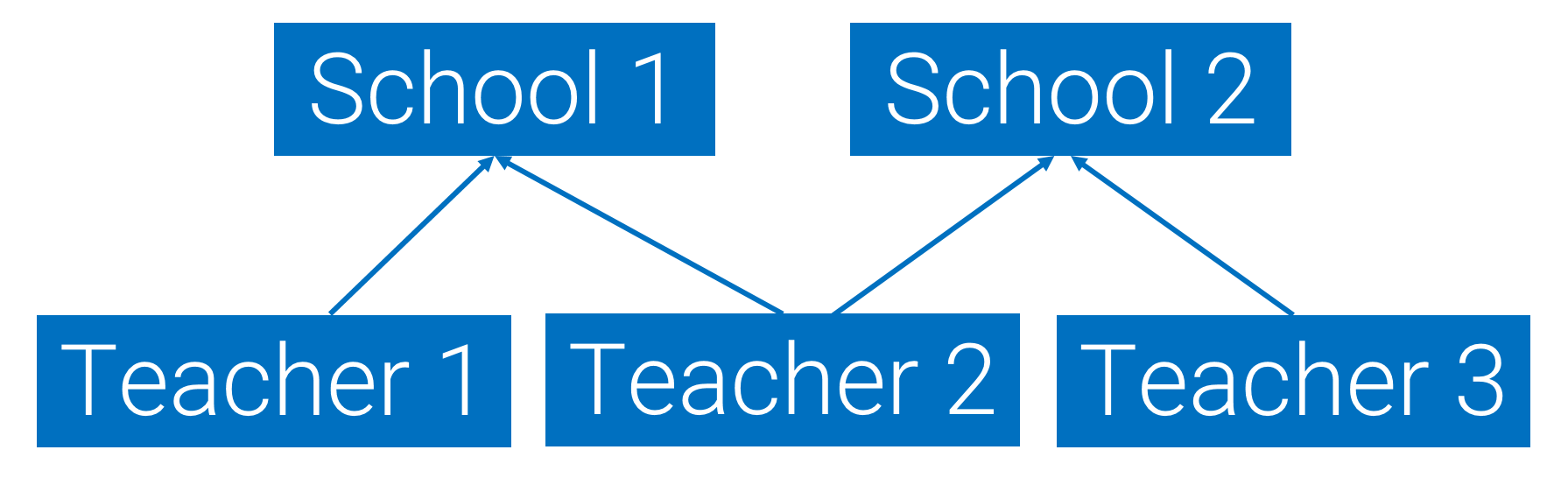
## Nested factors

When one factor is *nested* in another, levels of one factor occur only within one specific level of another factor. In the example below, showing teachers nested in schools, you can see that teachers 1 and 2 only taught at school 1 and teacher 3 only taught at school 2.



## Partial crossing

In partial crossing, levels of one factor occur with more than one level of another factor. However, not all levels of one factor occur with every level of another factor like in full crossing. This can happen when, for example, a teacher switches schools during the time frame of a study. In the example schematic below, you can see that teacher 2 taught in both school 1 and school 2 while the other two teachers only taught in one school. These factors are partially crossed.



## Crossed random effects

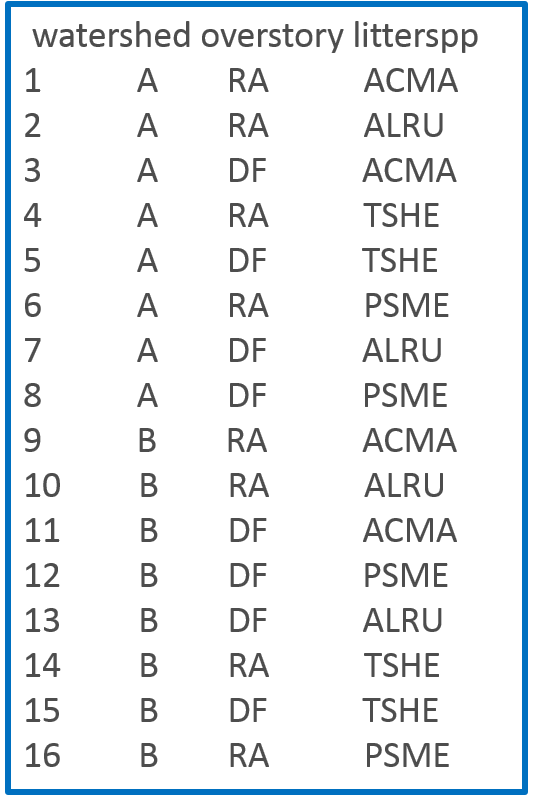
The reason I bring up nesting vs crossing (full or partial) is that being able to recognize nested vs crossed factors comes into play when doing an analysis. When factors that are to be treated as random are *crossed*, we will need a different model fitting approach than what we are learning in this class. All the models you learn in this class will have nested random effects.

A common example of crossed random effects is when working with space-time studies, where multiple study units were measured across time. Having both time (e.g., year) and space (e.g., plot) factors that you want to treat as random effects often indicates you need a model that allows for crossed random effects. The optimization needed for crossed random effects is more complicated, and not all software packages can fit such models. While crossed random effects can technically be done in package nlme, the syntax is difficult and models can be very slow to fit. If you are a situation with crossed random effects in R you will likely find yourself needing to use a different package, such as lme4 or glmmTMB.

All of the issues of crossed random effects come up with partial crossing, which is why you need to be able to recognize if the factors you are using as random effects are partially crossed.

# Explicit naming

## Missing column for physical units

A common issue when working with factors from different sized physical units, where smaller-sized units are nested in larger-sized units, is to have variables that represent these units missing from the dataset.

You can see an example of this on the left, showing a “watershed” variable plus the two factors of interest in the motivating example for two of the study watersheds. Stands are the replicate physical unit for the overstory factor, but there is no stand variable in the dataset.

Is the stand that represents RA in watershed A the same as the one that represents RA in watershed B? It is difficult to tell by just looking at the dataset, and a computer won’t be able to tell that these are two unique stands. This kind of *implicit* naming is relatively common, but it makes it more difficult to figure out all the sources of variation that need to be included in a model.

Implicit naming can lead to confusion.

## Making explicit names

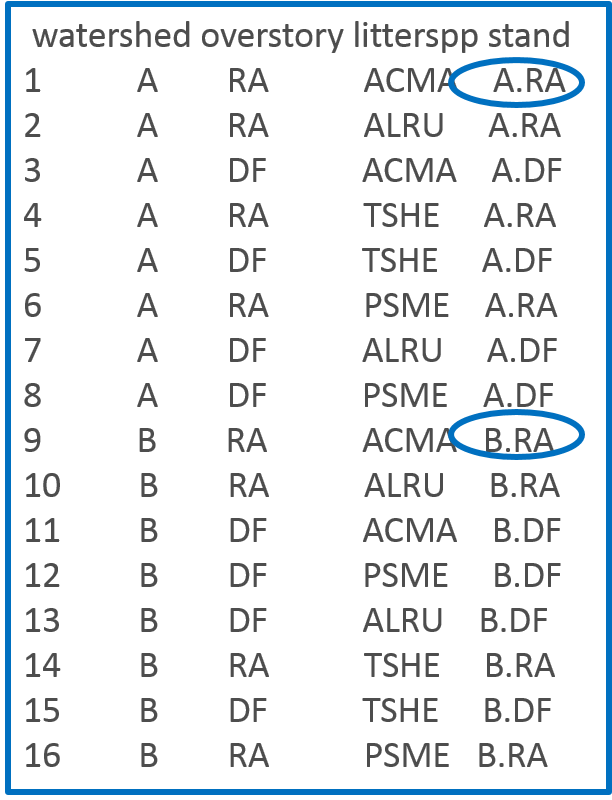
I recommend keeping the physical units as distinct variables from the other factors in the dataset. This will make models (and the dataset) easier to understand. To do this we need to create *explicit* names for the physical units in the study.

Making explicit names is often most straightforward to do during data collection. For example, stands could have been given unique, numeric names that were entered in the dataset along with the other collected data. If we didn’t do that, though, we can still create unique names using the other factors in the dataset.

In the first example I showed above, there was no variable to represent the different stands. Since we have a description of the study design, though, we know that each overstory species represents one stand within a watershed. To make unique names to represent stands we can combine the watershed and overstory factor levels. Now it is much easier to see that stand A.RA is different than stand B.RA because observations from different stands are associated with unique stand identifiers.

Creating explicit names from other factors in the dataset can be trickier than may seem in this example, so we will practice this more in class. For example, sometimes folks get confused why, when stands now have unique names, there are four A.RA rows in the dataset.

We might also want to make a variable to represent the litter bag location. This would make it very clear that the locations are unique physical units and helps us recognize this as a source of variation. However, since the litter bag locations are the level of the measurement of the response variable and we are fitting a linear mixed model, explicit naming for the observation-level variable is less important. As you know, we don’t need a variable to represent the observation-level random effect to be explicitly put into the model. Explicit naming is still good practice even with the observation-level physical units, though, since when working with more advanced models like generalized linear mixed models we may need to put an observation-level effect to address that source of variation explicitly into the model.



# Statistical model and analysis

You’ll see the statistical model in Week 5 Handout 3, based on the motivating example, is starting to get pretty complicated. The focus of this week is on the additional random effects.

## Random effects

When we have nested physical units in a study we have more sources of variation that affect the response variable that we (usually) are not specifically interested in. These factors are part of the design and must be accounted for in any model. If we leave any of these out of the model used for analysis after designing a study with different sizes of physical units we won’t meet the assumption of independence of errors.

Watershed variation

Here’s the part of the statistical model that is for the largest physical unit in the motivating example, the watersheds:

br is the random effect of *rth* watershed on log total microbial biomass, br ~ N(0, σb2) and Cov(br, br’)=0

Note that I’m using a new shorthand notation to indicate that watersheds are assumed to be independent of each other.

Stand variation

The second largest physical unit in the example study is the stand:

cs is the random effect of the *sth* stand on log total microbial biomass, cs ~ N(0, σc2) and Cov(cs, cs’)=0

We don’t expect stands *within* watersheds to be independent of each other. The assumed independence listed here is only after we’ve accounted for the correlation caused by watersheds (since “watershed” is in the model).

Litter bag location variation

The observation-level random effect is represented by the litter bag locations, since there as many unique litter bag locations as there are observations in the dataset. Always make sure you can figure out what physical unit is at the level of the observation.

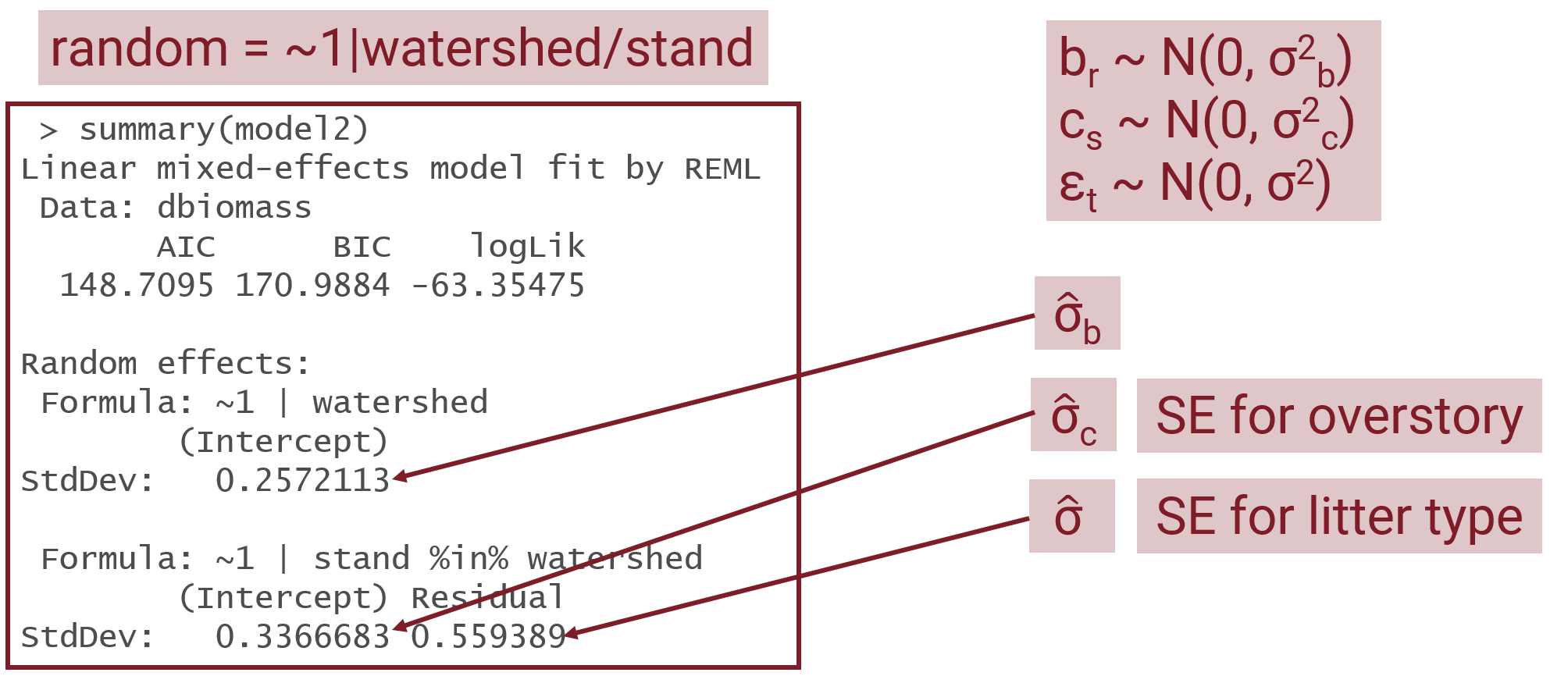
Here is a description of the observation-level random effect:

εt is the residual error term, which is the random effect of the *tth* litter bag location on log total microbial biomass, εt ~ N(0, σ2), and Cov(εt, εt’)=0, and br, cs, εt are all independent

We still assume independence of the errors, which is valid only if all other variables that cause non-independence (watersheds, stands, factors of interest) are in the model.

Output in R

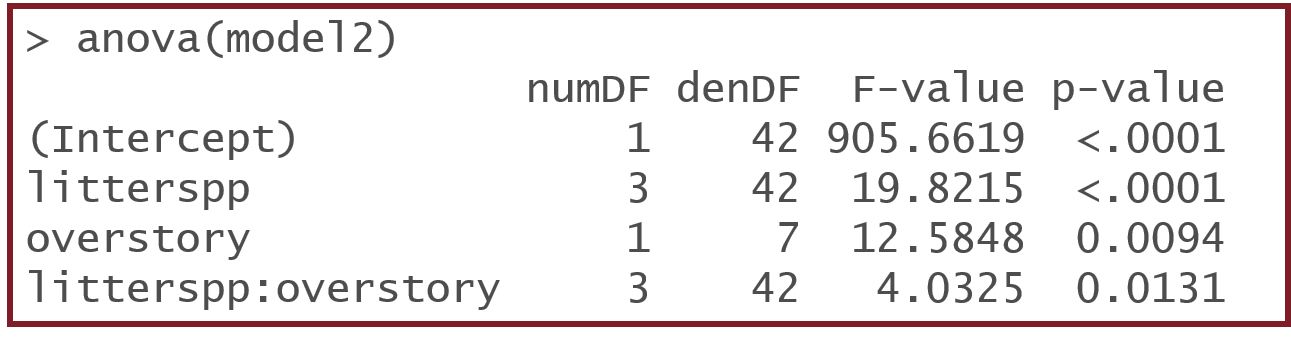
Random effects are about estimating variances. You can see there are now three standard deviations reported in the output after fitting a linear mixed model in R in function lme().



One of the reasons getting these random effects correct when fitting the model is that all tests and confidence intervals for a factor are based on the variance from the replicates of that factor. Since stands are the replicates of overstory species, the variance for stands is used when making inference about overstory species (such as hypothesis tests and confidence intervals). Since litter bag locations are the replicates of litter type, the litter bag location variance (observation-level variance) is used for all inference about litter type. Modern software is powerful, and R can automatically tell which variance to use for which fixed effects for a classic design like this one as long as we write the random effects correctly.

## Overall tests

When you look at overall F tests when using different size physical units, you will see differences in the degrees of freedom for different tests. The smallest physical unit, litter bag location, has the largest number of replicates. You can see that all factors at that level have more denominator degrees of freedom. Tests for factors at the level of the larger physical unit, in this case anything measured at the stand level, have fewer denominator degrees of freedom since there are fewer replicates of the larger physical units.



Remember that you are only going to report tests that are of interest to the investigators even though the anova.lme function automatically reports many tests.