FES 524: Natural Resources Data Analysis

Reading 4.2: Transformations

Contents

1 Nested versus crossed factors

Last week we saw examples of crossed factors with the factorial designs, this week we will see both *nested* and crossed factors.

1.1 Crossed factors

Recall from last week that when we have crossed factors, all combinations of the levels of the factors are present in the study. Figure 1 shows a diagram of crossed factors of "school" with two levels, and "teacher" with three levels. All teachers taught at each school in this example.

1.2 Nested factors

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

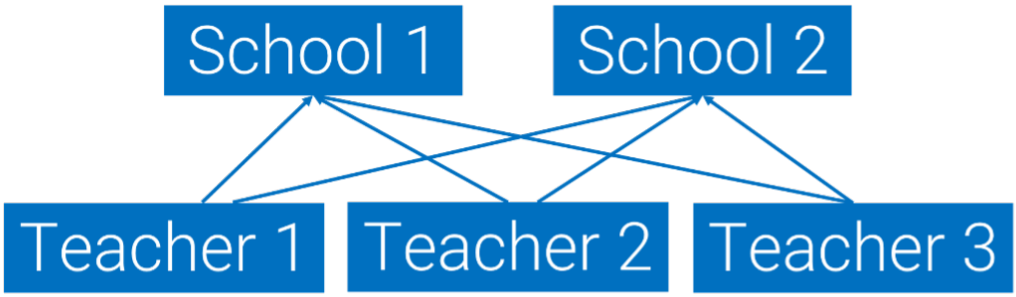

Figure 1: Diagram of two crossed factors, one with two levels and the other with three.

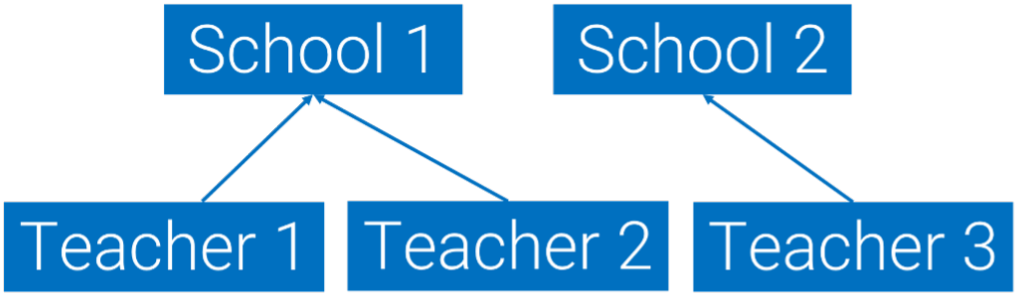

Figure 2: Diagram of a factor, teacher, nested in the school factor.

1.3 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.

1.3.1 Crossed random effects

Being able to recognize crossed versus nested factors can be important when those factors are to be treated as random. We will not see any crossed random effects in this class, but 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. 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.

1.4 Explicit versus implicit naming for nested factors

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 *implicitely named* factors. However, this can cause confusion. For example, in Table 1, is plant 3 at site 1 the same as plant 3 at site 2? It is difficult to tell by just looking at the dataset, and a R will not be able to tell that these are two unique plants. This kind of implicit naming

Table 1: Example dataset with implicite naming for plants within sites.

site	plant	treatment
1	1	hand pollen
1	1	open pollen
1	2	hand pollen
1	2	open pollen
1	3	hand pollen
1	3	open pollen
2	1	hand pollen
2	1	open pollen
2	2	hand pollen
2	2	open pollen
2	3	hand pollen
2	3	open pollen
3	1	hand pollen
3	1	open pollen
3	2	hand pollen
3	2	open pollen
3	3	hand pollen
3	3	open pollen

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.

1.5 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, plants could have been given unique identifiers 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 Table 1, there was no variable to represent the different plants. If we had a description of the study design, though, we would know that each plant within a site is unique. To make unique names for plants, we can combine the site and plant columns. In Table 2, it is much easier to see that plant 1 within site 1 is different from plant 1 in site 2 because observations from different sites are associated with unique plant identifiers.

2 Studies with different sizes of physical units

This week we will 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 sub-sampling, 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.

Table 2: Example dataset with a unique plant identifier.

site	plant	plant_id	treatment
1	1	1_1	hand pollen
1	1	1_1	open pollen
1	2	1_2	hand pollen
1	2	1_2	open pollen
1	3	1_3	hand pollen
1	3	1_3	open pollen
2	1	2_1	hand pollen
2	1	2_1	open pollen
2	2	2_2	hand pollen
2	2	2_2	open pollen
2	3	2_3	hand pollen
2	3	2_3	open pollen
3	1	3_1	hand pollen
3	1	3_1	open pollen
3	2	3_2	hand pollen
3	2	3_2	open pollen
3	3	3_3	hand pollen
3	3	3_3	open pollen

One term you will see for studies like the ones we will 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.

2.1 An example experiment with two sizes of units

Figure 3 shows an example of an experiment that has multiple sizes of experimental units. At the first level, the researchers have fields that they cut in half. Each half of the field is randomly assigned a treatment, either undisturbed or disturbed soil using tilling. Ignoring the next level of this experiment for a moment, we might recognize this experiment as a complete randomized block design with blocks as fields and halves of fields as our experimental unit to which we apply a soil disturbance treatment (either tilled or not). The statistical model for this first level of the experiment can be written as

$$y_{ijk} = \mu + \alpha_j + \gamma_k + \epsilon_{ijk}$$

where μ is the overall mean, α_j is the effect of the j^{th} level of the disturbance factor, γ_k is the random block factor for field, and ϵ_{ijk} is the error term for replicate i, where the half-field is the experimental unit (i will just be 1 in this example).

Let's turn our attention to the second level of this experiment. Each half field is split into four plots that are randomly assigned a fertilizer concentration, from no fertilizer to high with two concentrations in between. At this level, within a half-field, we have a completely randomized design for the fertilizer experiment. How do we combine these two experiments and fit one model that allows for multiple sizes of experimental units?

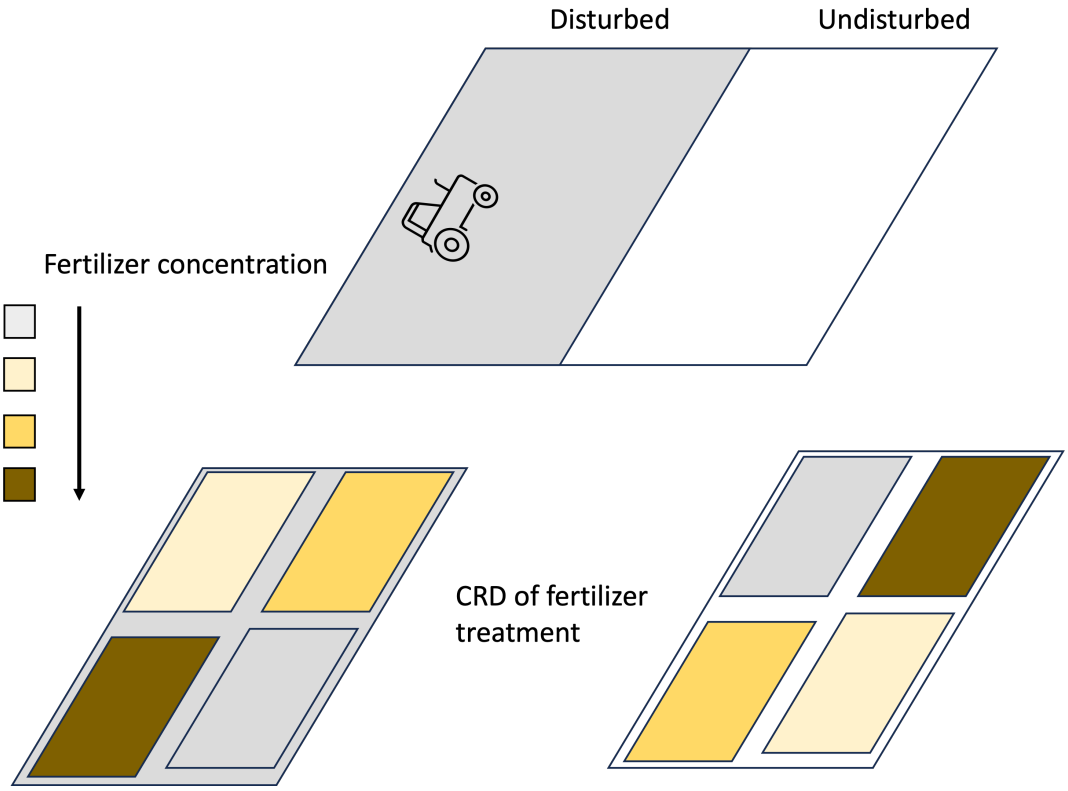

Figure 3: Example experiment with multiple sizes of experimental units.

3 Statistical model and analysis

You will see that the statistical model gets a bit complicated, but the focus of this week is on the additional random effects. The model for this experiment can be written as

$$y_{ijklm} = \mu + \alpha_i + \gamma_j + \epsilon_{ijk} + \beta_l + (\alpha\beta)_{il} + u_{ijklm}$$

where the first four terms are as described above (noticed I have changed the subscripts to be consecutive due to how many subscripts we need in this case), with β_l as the effect of fertilizer treatment l, averaged over disturbance, $(\alpha\beta)_{il}$ is the interaction between disturbance treatment i and fertilizer treatment l, and u_{ijklm} is the error term associated with plot replicate m. As usual, we have the sum-to-zero constraints on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $(\boldsymbol{\alpha\beta})$, and the iid assumptions for $\mathbf{u}$.

3.1 Nested random effects

Notice that we didn't just drop the error term from the block design model for the disturbance experiment. Instead, we model it as a random effect for the half-field, nested within the (random) field effect. This helps to highlight that the error terms *are* random effects, so you have been using mixed models since day one! Fitting this model using `lme4` is relatively easy once the data are structured properly with explicit naming columns. You will get some practice with this in Lab 5.