

Week 5: Interaction effects

Exercise 1: Water \times shade

Real biological systems are full of interactions, at least if you push the variables controlling the system far enough. For example, if you take away all CO₂ available to a plant, then it doesn't matter how much sunlight it receives; these variables interact, because the effect of sunlight depends upon the amount of CO₂ available. Likewise, the effect of CO₂ depends upon the amount of sunlight.

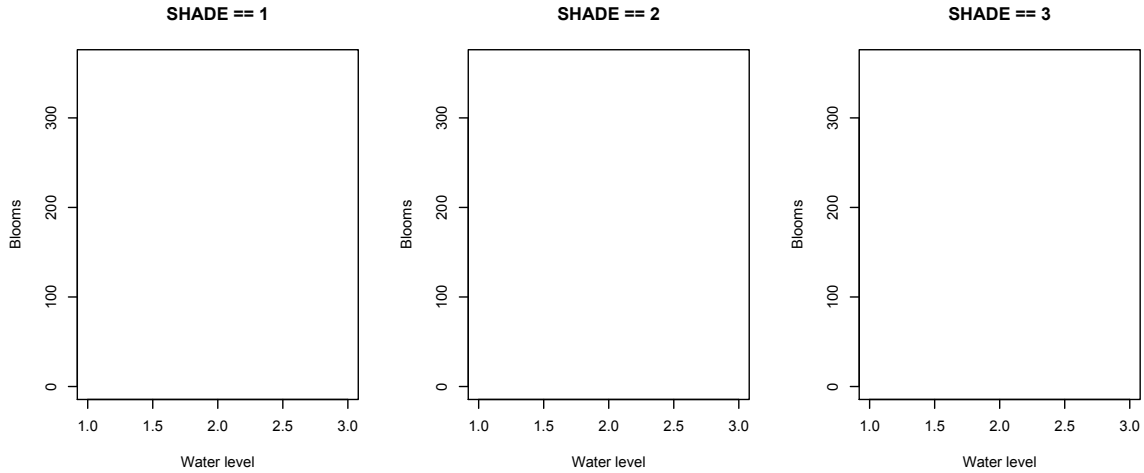
You'll analyze this kind of relationship, with the goal of getting a grasp on interactions between two continuous variables. The data contained in "tulips.csv" are the number of blooms produced by 27 different tulip plants, planted at three levels of water (higher numbers mean more water) and three levels of shade (higher numbers mean more shade). The BED variable is just the location in the greenhouse—you may find BED useful as a control variable.

(a) Analyze the main and interaction effects of water and shade, possibly controlling for BED, using model comparison (AIC and multiple models). Does model comparison support the hypothesis that water level and shade level interact? If so, how strongly?

(b) Using the best model (by AIC) that interacts water and shade, plot the predictions of the fit model. There are a number of ways to plot interactions, but we'll use perhaps the simplest, "tritych," method. (trip·tych |'triptik| noun. A set of three associated artistic, literary, or musical works intended to be appreciated together.)

1. Make three scatter plots with blooms on the vertical axis and water on the horizontal axis. Call these #1, #2, and #3.
2. In #1, plot those data points where the shade level is 1. Make sure the vertical axis displays the entire observed range of bloom counts, using `ylim=range(d$BLOOMS)` in your call to `plot`. (Remember that you can extract part of a list of numbers using indexing. e.g. `d$x[d$z==1]` gives those elements of `d$x` where `d$z` is equal to 1.)
3. Superimpose on the scatter plot the predicted relationship between blooms and water, given that the shade value is 1. (The easiest way to do this is to use the `curve` function.)
4. Repeat steps 2 and 3 for plots #2 and #3, where shade level should be set to 2 and 3, respectively.

You should have three plots like those below, but with the appropriate data points and regression lines drawn within each.



(c) Repeat (b), but using the highest ranking model that DOES NOT contain the interaction between WATER and SHADE.

(d) Compare the two triptych plots from (b) and (c). What is the model without the interaction failing to do correctly? That is, why does the interaction model fit the data so much better?

(e) What do you think is happening here, biologically? How would you explain the interaction in biological terms, rather than statistical terms?

Exercise 2: Sickle cell anemia

Sickle cell anemia is a condition caused by a rare hemoglobin variant, Hb^S . Individuals with two copies of this allele suffer from severe anemia—their blood doesn't carry oxygen very well, especially when oxygen is needed most.

Individuals with only one copy of this allele, however, can show resistance to malaria infection—the fancy malaria animal has a hard time parasitizing Hb^S blood cells, and an $Hb^A Hb^S$ individual produces both S cells and A (normal) cells. Thus in the presence of some kinds of malaria (*falciparum* esp.), heterozygote individuals have an advantage.

The data contained in the file “sickle.csv” are genotypes and ages at death for 1000 Tanzanian births. You will use the columns pA and pS—the counts of normal A and sickle S alleles in each individual—to analyze the effects of each allele on mortality rates.

(a) Plot a histogram of the ages at death. Do you need to transform these data, before using `lm()` to analyze them? If so, construct your transformed variable and add it to the data frame, as a new column.

(b) Analyze the model that contains only the main effects of pA and pS. Why do you fail to get an estimate for pS? (Hint: Think about the nature of information in each variable.) What biological conclusions does this model suggest?

If you transformed your data, be sure to convert the estimates back to the natural scale. Remember, if you used a regression equation:

$$\log(y_i) = \alpha + \beta x_i,$$

then if you convert the measures $\log(y_i)$ back to natural scale by exponentiating, then you just exponentiate the right-hand side, as well. This implies:

$$\begin{aligned}\exp(\log(y_i)) &= \exp(\alpha + \beta x_i), \\ y_i &= \exp(\alpha) \exp(\beta x_i)\end{aligned}$$

So a unit change in x_i will result in some unknown proportional change, Δy , in y_i that we can solve for. Let $x_1 = x$ and $x_2 = x + 1$. Then:

$$\begin{aligned}\text{proportional change in } y &= \Delta y = \frac{y_i|x_2}{y_i|x_1} \\ \Delta y &= \frac{\exp(\alpha) \exp(\beta(x + 1))}{\exp(\alpha) \exp(\beta x)} = \frac{\exp(\beta x) \exp(\beta)}{\exp(\beta x)}, \\ \Delta y &= \exp(\beta).\end{aligned}$$

So if x increases by 1 unit, we expect a proportional change in the size of the outcome of $\exp(\beta)$. For example, if $\exp(\beta) = 2$, we expect a unit increase in x to double y . If $\exp(\beta) = 0.5$, we expect a unit increase in x to halve y . This multiplicative relationship on the natural scale results from the nature of logs: things that are additive on the log scale are multiplicative on the normal scale.

(c) Analyze the model that contains the interaction of pA and pS. What biological conclusions does this model suggest? If you transformed your data, be sure to convert the estimates back to the natural scale. What work is the interaction effect doing in this context? What biological phenomenon is it managing to capture?

(d) Compare the models from (b) and (c), using AIC. Interpret.