Exercise

Model selection with the Loyn data

In the previous exercise you fitted a pre-conceived model which included the main effects of the area of the forest patch (`LOGAREA`), the grazing intensity (`FGRAZE`) and the interaction between these two explanatory variables (`FGRAZE:LOGAREA`). This was useful as a training exercise, and might be a viable approach when analysing these data if the experiment had been designed to test these effects only. However, if other potentially important variables are not included in the model this may lead to biased inferences (interpretation). Additionally, if the goal of the analysis is to explore what models explain the data in a parsimonious way (as opposed to formally testing hypotheses), we would also want to include these extra explanatory variables.

Here we revisit the loyn data analysis, asking if a 'better' model for the data could be achieved by including additional explanatory variables, and applying a model selection procedure. Because we would like to test the significance of the interaction between `LOGAREA`, and `FGRAZE`, whilst accounting for the potential effects of other explanatory variables, we will also include `LOGAREA`, `FGRAZE` and their interaction in the model as before. Including other interaction terms between other variables may be reasonable but we will focus only on the `FGRAZE:LOGAREA` interaction as we have relatively little information in this data set (67 observations). What we want to avoid is fitting an overly complex model which will estimate many parameters for which we have very little data. This is a balance you will all have to maintain with your own data and analyses (or better still, perform a power analysis before you even collect your data). No 4-way interaction terms in your models please!

It's also important to note that we will assume that all the explanatory variables were collected by the researchers *because* they believed them to be biologically relevant for explaining bird abundance (i.e. data were collected for a reason). Of course, this is probably not your area of expertise but it is nevertheless a good idea to pause and think what might be relevant or not-so relevant or partly redundant and why. Do we need to include all variables? This highlights the importance of knowing your study organism / study area and discussing research designs with colleagues and other experts in the field before you collect your data. What you should avoid is collecting heaps of data across many variables (just because you can) and then expecting your statistical models to make sense of it for you. As mentioned in the lecture, model selection is a relatively controversial topic and should not be treated as a purely mechanical process. Your expertise needs to be woven into this process otherwise you may come unstuck quite quickly.

1. Import the 'loyn.txt' data file and repeat the data preparations done previously (i.e. data transformations - `FGRAZE`, `LOGAREA`). Here we will be using all the explanatory variables to explain variation in bird density. If needed, remind yourself of your data exploration you conducted previously. Do any of the remaining variables need transforming? If so, what transformation did you apply? Add the transformed variables to the `loyn` dataframe.

2. It's useful to start with a quick graphical exploration of any potential relationships between each explanatory variable (collinearity) and also between our response and explanatory variables. A pairs plot using the function `pairs()` is a very effective way of doing this when the number of variables is not too large. You can also add the correlations to the lower triangle of the plot as you did previously (don't forget to define the function first).

Hint: restrict the plot to the variables you actually need. An effective way of doing this is to store the names of the variables of interest in a vector `VOI <- c("Var1", "Var2", ...)` and then use the naming method for subsetting the data set `Mydata[, VOI]`

3. Start with a model of `ABUND` and include all explanatory variables as main effects. Also include the interaction `LOGAREA:FGRAZE` but no other interaction terms as justified in the preamble above. Don't forget to include the transformed versions of the variables where appropriate (but not the untransformed variables as well otherwise you will have very strong collinearity between these variables!).

4. Have a look at the summary of the model using the `summary()` function. This is now quite a complicated model with lots of parameter estimates (14) and P values testing lots of hypotheses. Are all the P values less than our cutoff of 0.05? If not, then this suggests that some form of model selection is warranted to simplify our model.

5. Let's perform a first step in model selection using the `drop1()` function and use a *F* test based model selection approach. This will allow us to decide which explanatory variable may be suitable for removal from the model. Remember to use the `test = "F"` argument to perform *F* tests when using `drop1()`. Which explanatory variable is the best candidate for removal and why? What hypothesis is being tested when we do this model selection step?

7. Update the model and remove the least significant explanatory variable (from above) and repeat single term deletions with `drop1()`. You can update the model by just fitting a new model without the appropriate explanatory variable and assign it to a new name. Alternatively you can use the `update()` function. I'll show you both ways in the solutions below.

8. Again, update the model to remove the least significant explanatory variable (from above) and repeat single term deletions with `drop1()`.

9. Once again, update the model to remove the least significant explanatory variable (from above) and repeat single term deletions with `drop1()`.

10. And finally, update the model to remove the least significant explanatory variable (from above) and repeat single term deletions with `drop1()`.

11. If all goes well, your final model should be `lm(ABUND ~ LOGAREA + FGRAZE + LOGAREA:FGRAZE)` which you encountered in the previous exercise. Also, you may have noticed that the output from the `drop1()` function does not include the main effects of `LOGAREA` or `FRGRAZE`. Can you think why this might be the case?

12. Now that you have your final model, you should go through your model validation and model interpretation as usual. As we have already completed this in the previous exercise I'll leave it up to you to decide whether you include it here (you should be able to just copy and paste the code). Please make sure you understand the biological interpretation of each of the parameter estimates and the interpretation of the hypotheses you are testing.

OPTIONAL questions if you have time / energy / inclination!

A1. If we weren't aiming to directly test the effect of the `LOGAREA:FGRAZE` interaction statistically (i.e. test this specific hypothesis), we could also have used AIC to perform model selection. Repeat the model selection as in the previous *F* test questions above but this time use the `drop1()` function and perform model selection using AIC instead. NOTE: you can also do this using the `AIC()` function. Don't forget, if we want to perform model selection based on AIC with the `drop1()` function we need to omit the `test = "F"` argument)

A2. Refit your model with the variable associated with the lowest AIC value removed. Run 'drop1() again.

A3. Refit your model with the variable associated with the lowest AIC value removed and run `drop1()` again on your new model.

A4. Repeat your model selection by removing the variable indicated by the model with the lowest AIC.

A5. Rinse and repeat.

If all goes well, your final model should be `lm(ABUND ~ LOGAREA + FGRAZE + LOGAREA:FGRAZE)`. This is the same model you ended up with when using the *F* test based model selection above. This, however, might not always be the case and generally speaking AIC based model selection approaches tend to favour more complicated minimum adequate models compared to *F* test based approaches.

We don't need to re-validate or re-interpret the model, since the minimum adequate model is the one we came up with using the `drop1()` function (this is not always the case!).

However, don't forget that there's no replacement for a well thought-through model selection approach, informed by the understanding of theory in the research area and of the research questions. Remember to use your expertise and brain when performing model selection.

I guess the next question is how to present your results from the model selection process (using either *F* tests or AIC) in your paper and/or thesis chapter. One approach which I quite like is to construct a table which includes a description of all of our models and associated summary statistics. Let's do this for the AIC based model selection but the same principles apply when using *F* tests (although you will be presenting *F* statistics and P values rather than AIC values). Although you can use the output from the `drop1()` (and do a bit more wrangling) let's make it a little simpler by just using the `AIC()` function to calculate the AIC values for each model.

Or if you prefer a prettier version to include directly in your paper / thesis. You will need to install the `knitr` package before you can do this. See the 'Installing R Markdown' in Appendix A in the Introduction to R book for more details.

Model

AIC

deltaAIC

LOGAREA + FGRAZE + LOGAREA:FGRAZE

413.71

0.00

ALT + LOGAREA + FGRAZE + LOGAREA:FGRAZE

414.57

0.86

ALT + LOGDIST + LOGAREA + FGRAZE + LOGAREA:FGRAZE

416.41

2.70

LOGDIST + YR.ISOL + ALT + LOGAREA + FGRAZE + LOGAREA:FGRAZE

418.37

4.66

LOGLDIST + LOGDIST + YR.ISOL + ALT + LOGAREA + FGRAZE + LOGAREA:FGRAZE

420.34

6.63

End of the model selection exercise