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 relevant additional explanatory variables.

Here we revisit the previous loyn data analysis, and ask if a 'better' model for these data could be achieved by including additional explanatory variables and by performing model selection. 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). This will hopefully avoid 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 and why. 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 try to 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 end up with a model that is implausible or not very useful (and all models need to be useful!).

1. Import the 'loyn.txt' data file into RStudio and assign it to a variable called `loyn`. Here we will be using all the explanatory variables to explain the variation in bird density. If needed, remind yourself of your data exploration you conducted previously. Do any of the remaining variables need transforming (i.e. `AREA`, `DIST`, `LDIST`) or converting to a factor type variable (i.e. `GRAZE`)? Add the transformed variables to the `loyn` dataframe.

2. Let's start with a very quick graphical exploration of any potential relationships between each explanatory variable (collinearity) and also between our response and explanatory variables (what we're interested in). Create a pairs plot using the function `pairs()`of your variables of interest. 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]`. If you feel like it, 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).

3. Now, let's fit our maximal model. 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!). Perhaps, call this model `M1`.

4. Have a look at the summary table of the model using the `summary()` function. You'll probably find this summary is quite complicated with lots of parameter estimates (14) and P values testing lots of hypotheses. Are all the P values less than our cut-off 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 an F test based model selection approach. This will allow us to decide which explanatory variables 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?

6. Update and refit your model and remove the least significant explanatory variable (from above). Repeat single term deletions with `drop1()` again using this updated model. You can update the model by just fitting a new model without the appropriate explanatory variable and assign it to a new name (`M2`). Alternatively you can use the `update()` function instead. I'll show you both ways in the solutions below.

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

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

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

10. 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?

11. 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. This time when we remove each term we are looking for the model with the lowest AIC (remember that lower AIC values are better). If you like, you can repeat the model selection you did above, starting with the same most-complex model (M1) as before, but this time use the `drop1()` function and perform model selection using AIC instead (omitting the `test = "F"` argument), each time removing the term that gives a model with a lower AIC.

A2. However, the "superpower" of AIC is the ability to *simultaneously* compare multiple competing models, something we are not taking advantage of when we perform a stepwise model selection process. So, another approach to selecting our best model is to decide which set of models we are going to compare **before** we fit any of them, then fit them all, extract the AIC values, and see which model(s) have the lowest.

Fit 5 different models with the following terms, giving each model appropriate names to distinguish from the other models we have fitted so far: 1) "LOGLDIST + YR.ISOL + ALT + LOGAREA + FGRAZE + LOGAREA:FGRAZE", 2) "LOGLDIST + YR.ISOL + ALT + LOGAREA + FGRAZE + LOGLDIST:YR.ISOL + LOGAREA:FGRAZE", 3) "YR.ISOL + LOGAREA + FGRAZE", 4) "LOGAREA + FGRAZE + LOGAREA:FGRAZE", 5) "LOGAREA + FGRAZE". If you like, you can add further models using combinations of terms you think might give the best model. Which model do you think will have the lowest AIC?

A3. Now we have fitted our set of models, we can calculate the AIC of each using the `AIC()` function. Which has the lowest? Which has the highest? Are there any within 2 AIC of the lowest AIC? Are there any models that differ in a single term and have an AIC difference of about 2? What does this tell us about the additional term?

If all goes well, the best 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 in a stepwise manner. This 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 we have already done this previously.

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 here 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 fitting all of our models and then use the `AIC()` function to calculate the AIC values for each model rather than `drop1()`. Note that the values differ slightly between the two approaches; using either is fine but best not to mix AIC values from the drop1() and AIC() functions.

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

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

418.54

4.83

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

420.34

6.63

LOGAREA + FGRAZE

422.61

8.90

YR.ISOL + LOGAREA + FGRAZE

424.60

10.89

Which model selection approach do you prefer? Can you think of pros and cons of either approach?

Whichever model selection approach you chose, what is key to remember is to perform a fully manual and thought-through model selection process, informed by the understanding of theory in the research area, and of the research questions, with rigorous model validation throughout.

End of the model selection exercise