Exercise

Model selection with the Loyn data

The exercise from the previous practical assumed a pre-conceived model with the area of the patch `LOGAREA`, and the grazing intensity `FGRAZE` as interactive effects. This is useful as a training exercise, and might be the way to approach the analysis of these data if the experiment had been designed to test these effects only. However, if other predictors are presumed to be important, not including them in the model could bias our results. Alternatively, if the goal of the analysis is just to explore what model form(s) explain the data in a parcimonious way (as opposed to formally testing hypotheses), we would also want to include these extra predictors.

Here we revisit the Loyn data analysis, asking if a better model for the data could be achieved by including additional predictors, and applying a model selection procedure. Because we would like to test the significance of the interaction between `LOGAREA`, and `FGRAZE`, while accounting for the potential confounding effects of other predictors, we will want to force `LOGAREA`, `FGRAZE` and their interaction to remain in the model until the very last step of the model selection exercise.

1. Load the loyn data and repeat the data preparations done previously (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 required variables to the data set.

2. We assume that all the predictors have been collected by the authors *because* they were believed to be biologically relevant for explaining bird abundance. However, it is a good idea to pause and think what might be relevant or not-so relevant or partly redundant and why, before even exploring the relationships with bird abundance (yes, even graphically). You could do this in a table format, and include a hypothetical ranking of importance. Is there anything that limits your ability to fill such a table?

3. It's useful to start with a graphical exploration of the relationships between predictors and between predictors and response. A pair-plot with `pairs()` is a very effective way of doing this when the number of variables is not too large.

- Hints:
  - 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]`

4. Start with a model of ABUND containing all predictors. Don't include any interactions other than `LOGAREA * FGRAZE` at this point: I suggest you simplify this exercise by including only the main effects (unless you have identified some interactions that you expect to be biologically important and you really want to include them).

5. Have a look at the summary of the model, and compare the estimates of the coefficients with their standard errors. Are there any which have large SE relative to the coefficient estimate? Do any of these correspond with predictors that you had identified as being potentially collinear?

6. Is every term needed (everything significant?) in this model? To find out, perform a model selection step using `drop1()` for choosing which single term might be candidate for deletion (remember to use the test = "F" argument to perform F tests). What is that term? What hypothesis is being tested when we do this model selection step?

7. Update the model and repeat single term deletions with `drop1()`, until there are no longer any non-significant terms, ignoring `LOGAREA` or `FGRAZE` (we want to leave them in for now, irrespective of what `drop1` suggests).

8. If all goes well, you should end up the previous question with the interactive model again `lm(ABUND ~ LOGAREA * FGRAZE)`. Do you need to simplify this model? Do you need to use `drop1()` for that?

9. Let's simplify the model anyway, considering the additive-only model `lm(ABUND ~ LOGAREA + FGRAZE)`. Although we could have validated models at each step the model selection procedure, this can become impractical. However, you really should validate your models at least in the final stages of model selection, by creating plots of the residuals for the candidate final model (I say "candidate" because should the model fail the validation, it may need revisiting irrespective of what the model selection procedure suggested). Remember that you can split your plotting device into 2 rows and 2 columns using the `par()` function before you create the plots. Check each of the assumptions of the model using these plots and report if these assumptions are acceptable.

10. Obtain summaries of the model output using the `anova()` and `summary()` functions. Make sure you understand the difference between these two summaries (e.g. what specific hypotheses are being tested

for each of them), and the interpretation of the coefficients in the summary table: a good test of your understanding is to reconstruct the model formula in writing (on paper or in your script), to be able to make predictions by hand (see optional questions 'A2' and 'A3' at the end). If in doubt, try it and seek assistance!

11. What inference can you make from this model? What are the biological interpretations, and the statistical lessons you take away from this analysis of the Loyn data?

12. Had we not been aiming to test the effect of the `LOGAREA * FGRAZE` interaction statistically, we could also have used AIC to perform the model selection. Let's try this (taking the AIC values returned by `drop1` or using the function `step`), and summarize the performance of the alternative models in a table.

End of the model selection exercise

Optional additional questions, if you're fast or want to take it further

A1. Taking the final additive model from this practical: Since the `anova` function does sequential tests of the effects, the results could be different if we put `FGRAZE` first. Run the corresponding model and its analysis of variance. What null hypotheses are being tested? Do you reject or fail to reject the null hypotheses? What percentage of variation does the model explain overall? Hint: (SST-SSE)/SST. How much variation do `LOGAREA` and `FGRAZE` explain respectively?

A2. Looking at the summary table of the additive model, interpret all the coefficients, in terms of what they measure and how they affect the predictions of the model. Let's then check that it all fits together: write down the equation of the model with the appropriate parameter estimates from the summary. By hand, calculate the predicted bird abundance (A) for a patch with `LOGAREA`= -0.5 and `GRAZE`= 1, and (B) for a patch with `LOGAREA`= -0.5 and `GRAZE`= 3. Can you predict the difference in expected abundance between (A) and (B) before doing the calculation? Hint: what measures the difference between `GRAZE3` and `GRAZE1` for a given patch area? Now, predict (C) for `LOGAREA`= 0.5 and `GRAZE`= 3. What does the difference between (C) and (B) correspond to?

A3. Check if you can make sense of the interactive model structure (Model `birds.inter.1` in the previous practical), by writing down the equation of the model with the appropriate parameter estimates from the summary. Then, calculate again the predicted bird abundance (A) for a patch with LOGAREA= 2.5 and GRAZE= 1, and (B) for a patch with LOGAREA= -0.5 and GRAZE= 5.

A4. Let's compare the residuals diagnostics of the interactive model structure (Model `birds.inter.1` in the previous practical) with those of the additive model (final model from this practical). Remember, that you can split your plotting device into 2 rows and 2 columns using the `par()` function before you create the plots. Does the additional complexity of the interaction make a big difference?