Exercise

This practical has two exercises, as well as optional supplementary questions. Note that exercise 1 omits important steps of the normal modelling workflow, in particular the model validation and interpretation. This is addressed in Exercise 2.

Exercise 1: Linear model with interaction between continuous and categorical predictors (= explanatory variables)

This exercise builds on the linear model with one continuous predictor, and the linear model with one categorical predictor, by adding these two sources of variation in the same model and allowing their effects to interact (i.e., the effect of one predictor changes with the value of the other predictor).

1. As in previous exercises, either create a new R script (perhaps call it linear_model_3) or continue with your previous R script in your RStudio Project. Again, make sure you include any metadata you feel is appropriate (title, description of task, date of creation etc) and don't forget to comment out your metadata with a `#` at the beginning of the line.

2. Import the data file 'loyn.txt' into R and take a look at the structure of this dataframe using the `str()` function. We know that the abundance of birds `ABUND` increases quickly with the area of the patch `LOGAREA`, and more slowly for the larger patches (a saturating "log-linear relationship"). We now also know that bird abundance changes in a non-linear way with the grazing intensity `FGRAZE`. But how do these effects combine together? Would a small patch with low grazing intensity have more birds than a larger patch with high grazing intensity? Could the (poor) fit of the `ABUND ~ LOGAREA` model for the large patches be improved, if we accounted for grazing intensity in the patches?

3. As previously we want to treat `AREA` as a log-transformed area to limit the influence of the few disproportionately large patches, and `GRAZE` as a categorical variable with five levels. So the first thing we need to do is create the corresponding variables in the loyn dataframe, called `LOGAREA` and `FGRAZE`.

4. Explore the relationship between grazing and patch area, using a scatterplot. You could explore the joint effect of `FGRAZE` and `LOGAREA` on `ABUND`, using panel plots. Hint: See the function `coplot` in the Data exploration lecture slide 24, and/or the help page for `coplot`. Factor levels increase from the bottom-left panel to the top-right panel. What pattern do you see? Is it okay to assume the effect of `LOGAREA` to be the same for all grazing levels? This is effectively asking if we should let the slope of `LOGAREA` vary across `FGRAZE` levels, which is the definition of an interaction.

5. Fit an appropriate linear model in R to explain the variation in the response variable, `ABUND` with the explanatory variables `LOGAREA` and `FGRAZE` acting interactively. Hint: `*` is the interaction symbol! Remember to use the `data =` argument. Assign this linear model to an appropriately named object, like `birds.inter.1`.

6. Let's first check the assumptions of your linear model by creating plots of the residuals from the model. 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 using these plots and report whether your model meets these assumptions.

7. Use the `summary()` function on the first model object to produce the table of parameter estimates. Using this output, take each line in turn and answer the following questions: (A) what does this parameter measure, specifically? (B) What is the biological interpretation of the corresponding estimate? (C) What is the null hypothesis associated with it? (D) Do you reject or fail to reject this hypothesis? I encourage you to get someone to discuss your answers with you.

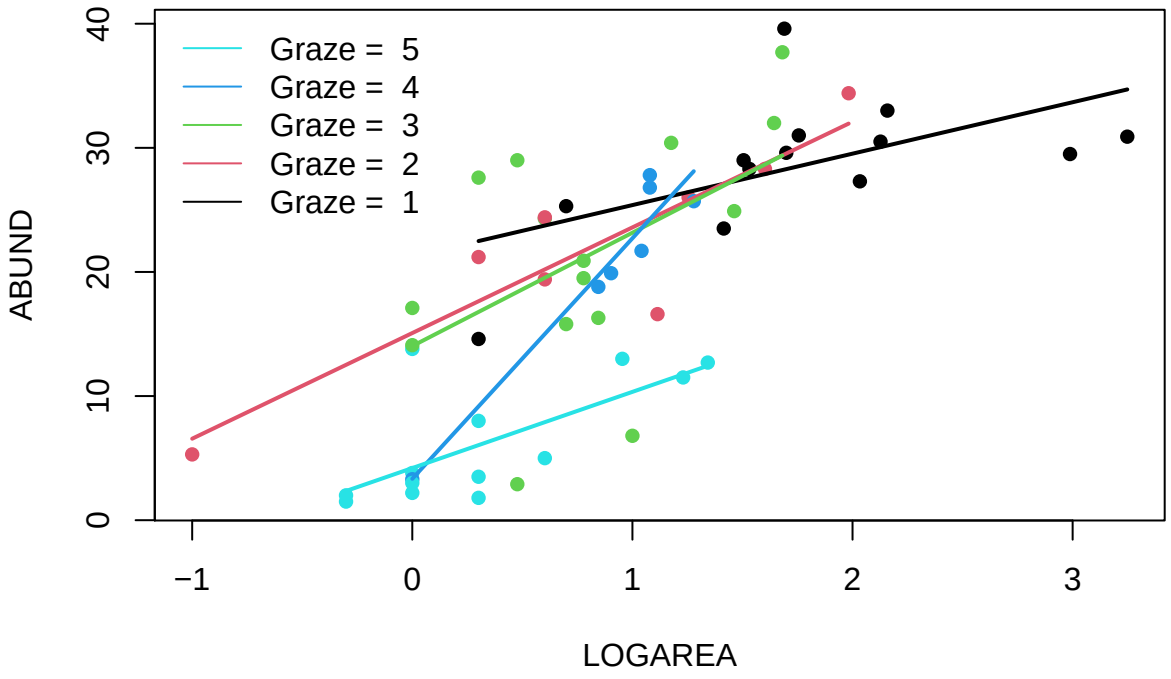
8. Let's now plot the predictions of your initial model to figure out how it really fits the data. Here's a recipe, using the `predict()` function.

- plot the raw data, using a different colour per `FGRAZE` level
- for each `FGRAZE` level in turn,
- create a sequence of `LOGAREA` from the minimum value to the maximum within the grazing level (unless you wish to predict outside the range of observed values)
- store it in a data frame (e.g. `dat4pred`) containing the variables `FGRAZE` and `LOGAREA.` Remember that `FGRAZE` is a factor, so it requires double quotes.
- add a predicted column containing the predictions of the model for the new data frame, using `predict()`
- plot the predictions with the appropriate colours

See the script below, for one of many ways of doing this.

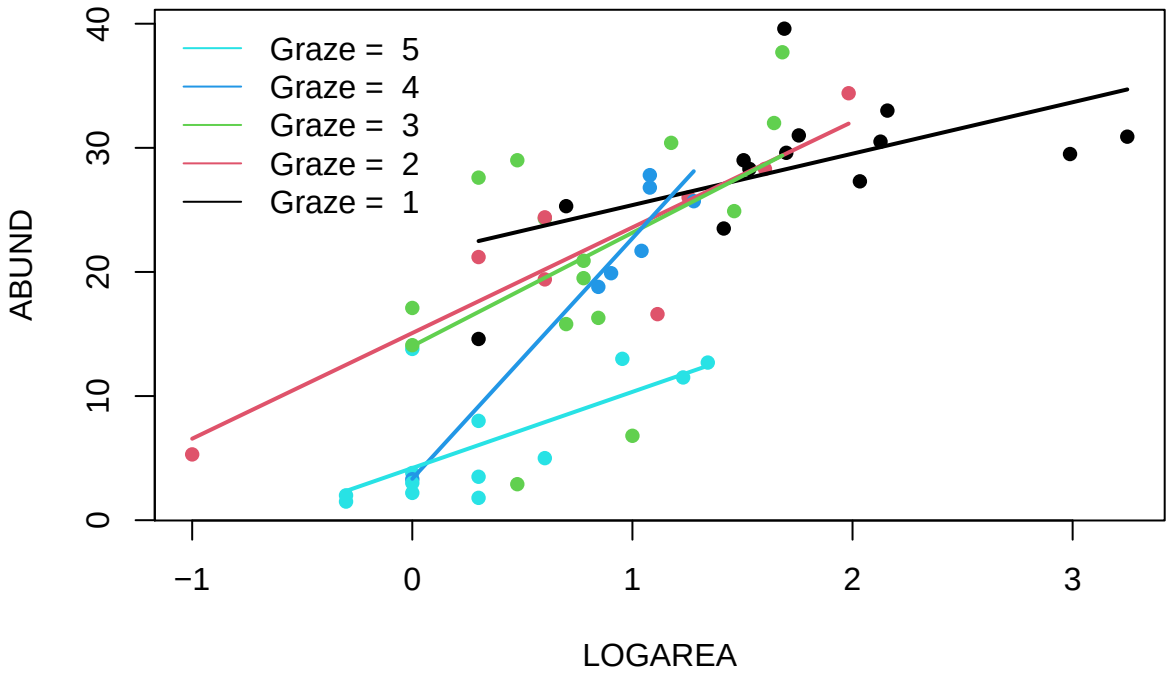
```
par(mfrow= c(1, 1))
plot(ABUND ~ LOGAREA, data= loyn, col= GRAZE, pch= 16)
# Note: # color 1 means black in R
# color 2 means red in R
# color 3 means green in R
# color 4 means blue in R
# color 5 means cyan in R

# FGRAZE1
# create a sequence of increasing Biomass within the observed range
LOGAREA.seq<- seq(from= min(loyn$LOGAREA[loyn$FGRAZE == 1]),
```

```
                    to= max(loyn$LOGAREA[loyn$FGRAZE == 1]),
                    length= 20)
# create data frame for prediction
dat4pred<- data.frame(FGRAZE= "1", LOGAREA= LOGAREA.seq)
# predict for new data
dat4pred$predicted<- predict(birds.inter.1, newdata= dat4pred)
# add the predictions to the plot of the data
lines(predicted ~ LOGAREA, data= dat4pred, col= 1, lwd= 2)

# FGRAZE2
LOGAREA.seq<- seq(from= min(loyn$LOGAREA[loyn$FGRAZE == 2]),
                    to= max(loyn$LOGAREA[loyn$FGRAZE == 2]),
                    length= 20)
dat4pred<- data.frame(FGRAZE= "2", LOGAREA= LOGAREA.seq)
dat4pred$predicted<- predict(birds.inter.1, newdata= dat4pred)
lines(predicted ~ LOGAREA, data= dat4pred, col= 2, lwd= 2)

# FGRAZE3
LOGAREA.seq<- seq(from= min(loyn$LOGAREA[loyn$FGRAZE == 3]),
                    to= max(loyn$LOGAREA[loyn$FGRAZE == 3]),
                    length= 20)
dat4pred<- data.frame(FGRAZE= "3", LOGAREA= LOGAREA.seq)
dat4pred$predicted<- predict(birds.inter.1, newdata= dat4pred)
lines(predicted ~ LOGAREA, data= dat4pred, col= 3, lwd= 2)

# FGRAZE4
LOGAREA.seq<- seq(from= min(loyn$LOGAREA[loyn$FGRAZE == 4]),
                    to= max(loyn$LOGAREA[loyn$FGRAZE == 4]),
                    length= 20)
dat4pred<- data.frame(FGRAZE= "4", LOGAREA= LOGAREA.seq)
dat4pred$predicted<- predict(birds.inter.1, newdata= dat4pred)
lines(predicted ~ LOGAREA, data= dat4pred, col= 4, lwd= 2)

# FGRAZE5
LOGAREA.seq<- seq(from= min(loyn$LOGAREA[loyn$FGRAZE == 5]),
                    to= max(loyn$LOGAREA[loyn$FGRAZE == 5]),
                    length= 20)
dat4pred<- data.frame(FGRAZE= "5", LOGAREA= LOGAREA.seq)
dat4pred$predicted<- predict(birds.inter.1, newdata= dat4pred)
lines(predicted ~ LOGAREA, data= dat4pred, col= 5, lwd= 2)

legend("topleft",
 legend= paste("Graze = ", 5:1),
 col= c(5:1), bty= "n",
 lty= c(1, 1, 1),
 lwd= c(1, 1, 1))
```



[Optional] Alternative method, using a loop:

```
# Okay, that was a long-winded way of doing this.
# If, like me, you prefer more compact code and less risks of errors,
# you can use a loop, to save repeating the sequence 5 times:
par(mfrow= c(1, 1))
plot(ABUND ~ LOGAREA, data= loyn, col= GRAZE, pch= 16)

for(g in levels(loyn$FGRAZE)){# `g` will take the values "1", "2",..., "5" in turn
    LOGAREA.seq<- seq(from= min(loyn$LOGAREA[loyn$FGRAZE == g]),
                                        to= max(loyn$LOGAREA[loyn$FGRAZE == g]),
                                                length= 20)
    dat4pred<- data.frame(FGRAZE= g, LOGAREA= LOGAREA.seq)
    dat4pred$predicted<- predict(birds.inter.1, newdata= dat4pred)
    lines(predicted ~ LOGAREA, data= dat4pred, col= as.numeric(g), lwd= 2)
}
legend("topleft",
 legend= paste("Graze = ", 5:1),
 col= c(5:1), bty= "n",
 lty= c(1, 1, 1),
 lwd= c(1, 1, 1))
```



Take some time to observe the predictions from the model, and how the lines have different intercepts but the same slope (as assumed by the model with additive effects only)

9. From a biological point of view, what have we learned so far from the interactive model? (Assume that the assumptions are adequately met, for now). Do you think the model is biologically plausible? Is it supported statistically?

End of the Linear model with interactive continuous and categorical predictors exercise

Exercise 2: Model selection

Exercise 1 above assumes 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.

Exercise 2 revisits 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.

10. 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.

11. 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?

12. 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]`

13. 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).

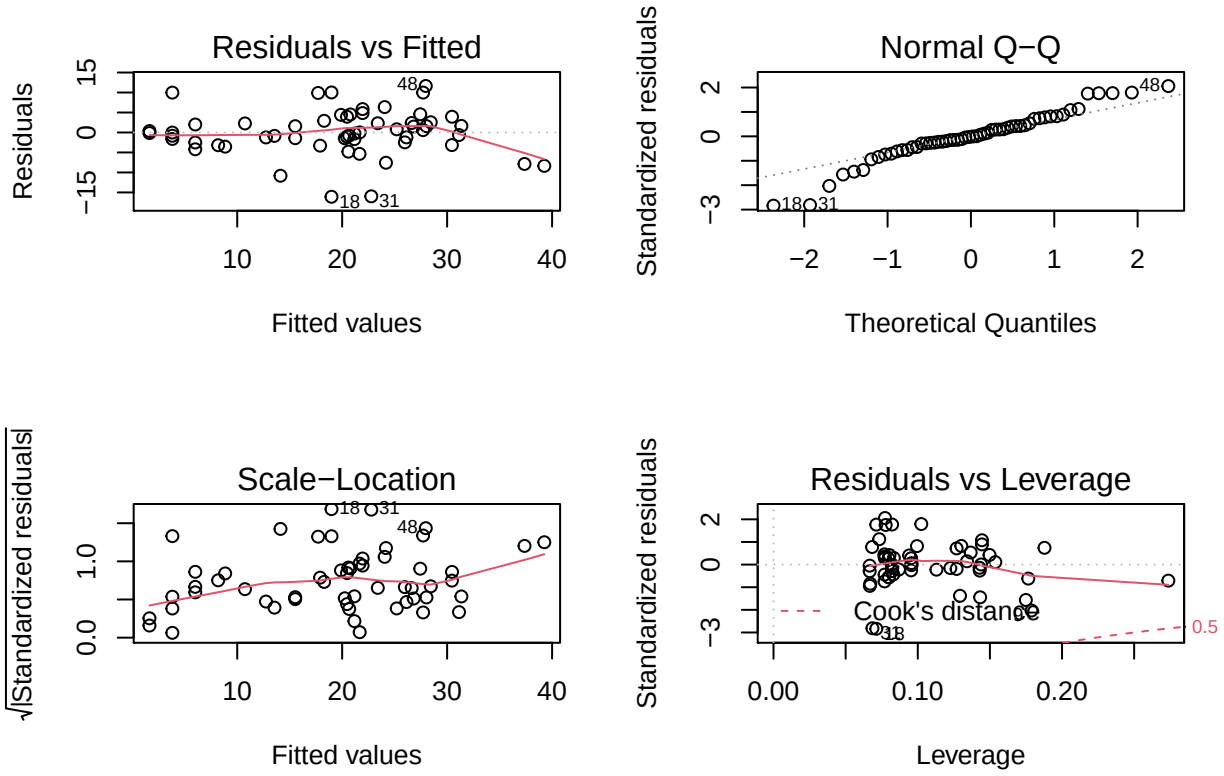
14. Check for collinearity using the `vif()` function in the `car` package.

15. 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?

16. 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).

17. 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?

18. 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.



19. 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). In doubt, try it and seek assistance!

20. 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?

21. 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 questions, if you're fast or want to take it further

A1. Taking the final additive model from Exercise 2: 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: 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 exercise 1), 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 exercise 1) with those of the additive model (final model from exercise 2). 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?