Exercises

Exercise: Linear model with single categorical explanatory variable

1. As in previous exercises, either create a new R script (perhaps call it 'linear_model_2') 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. Once again import the data file 'loyn.txt' into R and take a look at the structure of this dataframe using the `str()` function. In this exercise you will investigate whether the abundance of birds (`ABUND`) is different in areas with different grazing intensities (`GRAZE`). Remember, the `GRAZE` variable is an index of livestock grazing intensity. Level 1 = low grazing intensity and level 5 = high grazing intensity.

3. As we discussed in the graphical data exploration exercise the `GRAZE` variable was originally coded as a numeric (i.e. 1, 2, 3, 4, 5). In this exercise we actually want to treat `GRAZE` as a categorical variable with five levels (aka a factor). So the first thing we need to do is create a new variable in the `loyn` dataframe called `FGRAZE` in which we store the `GRAZE` variable coerced to be a categorical variable with the `factor()` function (you can also use the `as.factor()` function if you prefer).

4. Explore any potential differences in bird abundance between each level of `FGRAZE` graphically using an appropriate plot (hint: a boxplot might be useful here). How would you interpret this plot? What might you expect to see in your analysis? Write your predictions in your R script as a comment. What is the mean number of birds for each level of `FGRAZE`?

5. Fit an appropriate linear model in R to explain the variation in the response variable, `ABUND`, with the explanatory variable `FGRAZE`. Remember to use the `data =` argument. Assign this linear model to an appropriately named object (`birds_lm` if your imagination fails you!).

6. Produce the ANOVA table using the `anova()` function on the model object. What null hypothesis is being tested? Do you reject or fail to reject the null hypothesis? What summary statistics would you report? Summarise in your R script as a comment.

7. Use the `summary()` function on the model object to produce the table of parameter estimates (remember these are called coefficients in R). Using this output what is the estimate of the intercept and what does this represent? What is the null hypothesis associated with the intercept? do you reject or fail to reject this hypothesis?

Next we move onto the the `FGRAZE2` parameter, how do you interpret this parameter? (remember they are contrasts). Again, what is the null hypothesis associated with the `FGRAZE2` parameter? do you reject or fail to reject this hypothesis?

Repeat this interpretation for the `FGRAZE3`, `FGRAZE4` and `FGRAZE5` parameters. Summarise this as a comment in your R script.

8. Now that you have interpreted all the contrasts with `FGRAZE` level 1 as the intercept, set the intercept to `FGRAZE` level 2 using the `relevel()` function, refit the model, produce the new table of parameter estimates using the `summary()` function again and interpret.

Repeat this for `FGRAZE` levels 3, 4 and 5. Can you summarise which levels of `FGRAZE` are different from each other?

9. Staying with the summary table of parameter estimates, how much of the variation in bird abundance does the explanatory variable `FGRAZE` explain?

10. Now onto a really important part of the model fitting process. Let's check the assumptions of your linear model by creating plots of the residuals from the model. Remember, you can easily create these plots by using the `plot()` function on your model object. Also remember that if you want to see all plots at once then you should 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 in your R script.

11. This is an optional question and really just for information. I'll give you the code in the solutions so don't overly stress about this! Use Google (yep, this is OK!) to figure out how to plot your fitted values and 95% confidence intervals. Try Googling the `gplots` package or the `effects` package.

Alternatively, have a go at using our old trusty `predict()` function to calculate the fitted values and standard errors. Add the fitted values and 95% confidence intervals to a plot of bird abundance and graze level (to add your upper and lower confidence intervals will need to use either the `segments()` or `arrows()` function).

Or we can even use the `ggplot2` package. Check out the solutions code if you're thoroughly confused!

End of the linear model with single categorical explanatory variable exercise