Exercise 2: Density of lampreys in North American streams

Background:

File Exercise_2_Lamprey_data.csv (available from our Github repository) contains a collection of data from a survey of lampreys (*Petromyzon marinus*) in different streams of North America, along three different sampling seasons. The dataset includes the following data:

- Depth.m - depth of sample unit in meters
- Velocity.ms - current velocity in m/s
- Season - season sample was collected
- Sample.area.m2 - sample unit area in meters squared
- No.caught- the number of lampreys caught

Overall objectives:

We are interested in studying the effects of sampling season, water velocity, and stream depth on the density of lampreys. We also want to develop a predictive model to predict the counts of lampreys caught as a function of those measured variables that have a significant effect on this response variable.

Step-by-step analyses:

1. First, we will check that our results are not biased due to using different sampling efforts (unit surveyed areas) in the three seasons. Use a suitable parametric method to test for differences in the average values of Sample.area.m2 as a function of Season (as a categorical factor). Check if the assumptions of the parametric test are met. In case they are not, use a non-parametric test. Can we conclude whether is there a significant effect of the season in the sampled areas? Specify the test(s) you used and give the p-value(s):

```
# Write your results as comment line(s) in your R script
```

2. Now we are going to produce some exploratory plots. Create a new column in the dataset for the density of lampreys (number of lampreys by sampled area). Generate two groups of faceted plots (for each season) to show the density of lampreys vs water velocity and the density of lampreys vs water depth. Add a linear trend line to each facet using geom_smooth().

3. Calculate a complete linear model for lamprey density as a function of season, water depth and water velocity. Which variables have a significant effect on the density of lampreys? Specify the p-values for all the tested variables:

```
# Write your results as comment line(s) in your R script
```

4. Check the normality of the residuals from the previous complete linear model and the homoscedasticity of lamprey densities sampled at different seasons. Do you think that the results of the model are reliable?

```
# Write your results as comment line(s) in your R script
```

5. Now, produce a linear model including only the variable(s) that have a significant effect on the lamprey density. Check the assumptions. Use the function anova() with the option test="LRT" (likelihood ratio test) to check if the complete model including all variables is significantly better than the reduced model including only the variable(s) that have a significant effect. Which model is the best to explain the experimental results without adding any unnecessary complexity?

```
# Write your results as comment line(s) in your R script
```

6. Now, instead of lamprey densities, we will work with lamprey counts. Model the lamprey counts as a function of all variables used in the previous complete model, adding the sampled area as another independent variable. Also check the assumptions of the model. Do you get similar results regarding which variables have significant effects? Has the sample area a significant effect on the counts, according to this linear model?

```
# Write your results as comment line(s) in your R script
```

7. Now model the lamprey counts using only the variable(s) with significant effects. Use function anova() with test="LRT" to check if there are significant differences between the complete and the reduced models.

```
# Write your results as comment line(s) in your R script
```

8. Re-do the calculations of Step 6 with Lamprey counts, but now using the function glm() and modeling the residuals using a Poisson distribution (instead of a gaussian distribution). Use the anova() function with the option test="LRT" to test the significancy of the effects of every variable). Which differences do you observe to the previous model from Step 6?

```
# Write your results as comment line(s) in your R script
```

9. Generate two additional glm() models. One using the velocity and the sampled area as factors, and the other using only the velocity, both fitting the residuals to a Poisson distribution. Now compare the three models of step 8 and 9 using anova() with test="LRT". Which one do you think is the best to fit the lamprey count data?

```
# Write your results as comment line(s) in your R script
```

10. Use the function predict(model, type = "response") to estimate two vectors with the predicted values of Lamprey counts for the best possible linear model from Step 7 and for the best possible model based on a Poisson distribution from Step 9. Bind these two vectors to the data frame usind bind_cols() and then calculate two new columns with the squared residuals for each model (i.e. the differences between the data and the predicted values for the lamprey counts from each model, squared). Print the sum of squared residuals for both models. Which one is smaller?

```
# Write your results as comment line(s) in your R script
```

11. Generate a scatter plot showing the experimental data points of Lamprey counts as a function of water velocity. Add two lines (using different colours) to show the values predicted by the two models (plot the predicted values, not the squared residuals), which you calculated in Step 10.