EDDA: final assignment

- *Submit in* **canvas** *a well readable report, preferably as one pdf-file (resulting from your Rmd-file according to the template). Submitting an R-script file is also allowed.*
- *The submitted file should contain the number of your group (with the names), the R-code you used, and answers to the questions posed (if needed, including plots).*
- *Throughout this assignment tests should be performed using a level of 0.05.*

Exercise Galapagos

The dataset `gala.txt` contains measurements on the numbers of plant species and background variables on 30 Galapagos islands. The dataset contains the following variables: `Species` – the number of plant species found on the island, `Area` – the area of the island (km^2), `Elevation` – the highest elevation of the island (m), `Nearest` – the distance from the nearest island (km), `Scruz` – the distance from Santa Cruz island (km), `Adjacent` – the area of the adjacent island (square km).

Source: M.P. Johnson and P.H. Raven (1973). Species number and endemism: The Galapagos Archipelago revisited, *Science* 179, 893–895.

1. Find a suitable linear model for the response variable `Species` by the *step-down* method. (Do not consider interactions or transformations at this stage.)
2. Do the same for the transformed response variable `sqrt(Species)`.
3. Use standard graphical diagnostic tools to decide which of the two resulting models (from 1. and 2.) is better.
4. For the model for `sqrt(Species)` (from 2.), plot the values of Cook's distance. Which island appears to be an influence point? Does the model change a lot if this island is removed?
5. Consider transforming the explanatory variables. Start by making some summary plots.

   ```
   > par(mfrow=c(2,3))
   > for (i in 1:6) hist(gala[,i],main=colnames(gala)[i],xlab="",ylab="")
   > pairs(gala)
   > for (i in 1:6) hist(log(gala[,i]),main=colnames(gala)[i],xlab="",ylab="")
   > pairs(log(gala))
   ```

   What did we achieve by applying this transformation?
6. Fit `log(Species)` on the logarithmic transforms of the other variables:

   ```
   > modlog=lm(log(Species)~log(Area)+log(Elevation)+log(Nearest)+log(Scruz+1)
      +log(Adjacent),data=gala)
   ```

   The `R`-function `step` (applied to a model) selects a submodel by using the AIC-criterion. Use this function to reduce the model `modlog`:

```
> modlog1=step(modlog)
```

Are all variables in the model `modlog1` significant?

7. Can you think of an explanation in terms of the meaning of the variables why it might make more sense to model `log(Species)` as a linear function of `log(Area)` than `Species` as a linear function of `Area`?

8. For the model `modlog1`, make a plot of the values of Cook's distance, a `qqnorm` plot of the residuals and a plot of residuals versus fitted values. Comments?

9. The variable `Elevation` seems a natural explanatory variable. Consider the model

   ```
   > gala$logElevation=log(gala$Elevation)
   > gala$logArea=log(gala$Area)
   > modlog2=lm(log(Species)~logArea+logElevation+logScruz,data=gala)
   ```

   Investigate possible collinearities between the explanatory variables in this model. Investigate whether it is useful to include interactions between `logArea` and `logElevation`.

10. Which of the fitted models so far do you prefer? What do you think of the realism of the independent Gaussian errors with zero mean and constant variance in all the considered models? Can you propose an alternative?