Exercices on session *Statistics with R session 2/3*

Load the packages of the *tidyverse* and *ggeffects*

Data on vascular plants in the British Islands

The Arrhenius relationship (1921) predicts that the number of species in an ecosystem increases with its area to the power z according to the equation

$$S = cA^z$$

where S(A) is the number of species observed on area A, c is a constant that depends on the ecosystems and the taxon considered, and z is the parameter of interest.

We are going to explore this relationship using data on the number of vascular species in the British Islands (Johnson et Simberloff, 1974), available here.

- Load the data as a tibble and look at their structure

The data contains the number of species of vascular plants (variable *species*) for different *island*. The other variables give some characteristics of these islands, including the *area* in km^2.

Simple linear regression

The Arrhenius relation can be linearised by taking the logarithm of S and A:

$$log(S) = \beta_0 + \beta_1 log(A)$$

- Explore graphically the relationship between the log-transformed area and the log-transformed number of species

- Fit the following log-log model, look at the model results and interpret them

$$log(S) \sim \mathcal{N}(\beta_0 + \beta_1 \times log(A), \sigma^2)$$

- Do the model validation

Multiple linear regression

The goodness of fit of the previous model could probably be improved. As the variables *elevation*, *latitude*, *dist_britain* and *soil_types* could also be influencing the number of species, we could add them as explanatory variables in our model.

- Let's first explore graphically the correlation between all pairs of numerical variables, *area* (log-transformed), *elevation*, *latitude*, *dist_britain*, using the package GGally. What do you conclude?
- Let's also make boxplots to explore graphically the relationship between these 4 variables and the type of soil (don't forget to transform the variable *soil_types* to a factor). What do you conclude?
- Fit the following model (without standardising the variables), look at the model results and interpret them

$$log(S) \sim \mathcal{N}(\beta_0 + \beta_1 \times log(A) + \beta_2 \times latitude, \sigma^2)$$

- Do the model validation
- Check the VIF (from library *car*) to check for the absence of collinarity between the explanatory variables

Presenting the result of the best model

- Compare the AIC of the two models. Which is the best one?
- Present the coefficients in a table
- Present the result graphically: first, represent the predicted relationship between A and S, setting the latitude to its mean
- then the predicted relationship between latitude and S, setting the area to its mean

For next time

- Make sure the following packages are already installed: *vegan*, *tidyverse*

Acknowledgments

- Marchand P. *Analyses et modélisation des données écologiques*
- Marcon E. *cours-R-Geeft*