Applied Machine Learning and Predictive Modelling 1

Dr. Luisa Barbanti and Dr. Matteo Tanadini
luisa.barbanti@hslu.ch

Fall Semester 2025 (HSLU)

Testing: Multiple linear regression

- A simple linear regression model contains a single predictor
- A multiple linear regression model contains several predictors[1]
- The effects estimated in a multiple linear regression model are said to be conditional. This means that the effect of any predictor is estimated after accounting all the other predictors present in the model

[1] Multiple linear regression is not to be confounded with multivariate linear regression which is a regression model, where several response variables are modelled simultaneously.

Testing: Categorical variables

- In **R** categorical variables are called "factors"
- Factors can be tested via F-tests by using the `drop1()` function or by comparing two models via the `anova()` function
- Comparisons among levels of a factor (i.e. "contrasts") can be performed by using the `glht()` function
- Contrasts (usually) imply a p-value correction to avoid the "multiple testing" problem
- Sequential Sums of Squares Anova (i.e. calling the `anova()` function fed with one model) should be avoided

Testing: Continuous or discrete variables

- Continuous (and discrete) variables can be tested via F-tests (with `drop1()`) or by t-tests (with `summary()`). The results are fully equivalent
- Sometimes the inferential results for continuous variables are best displayed and communicated with confidence intervals

Testing: Interactions

- If an interaction term is shown to be significant, then all terms involved in this interaction play a relevant role
- An interaction involving two predictors is called "two-fold interaction" (e.g. age `*` `species`)
- An interaction can involve more than two predictors (three-fold, four-fold, ...)
- Parameter interpretation for interaction terms higher than two-fold can be hard

Testing: Model assumptions

- All the testing procedure discussed here rely on assumptions about the errors that will be discussed later in the course
- In particular, errors are assumed to follow a normal distribution
- p-values and confidence intervals, as discussed here, are fully equivalent (duality)[2]

[2] For example, a continuous predictor significant at the 5 % level will have a CI that does not contain zero.