ADVANCED STATISTICAL MODELS WITH *R* FOR BIOLOGICAL SCIENCES

COURSE PROGRAM

CIIMAR – 20 HOURS; 10-14 March 2025

Session 1. 10:00 – 14:00 h (break from 11:45 to 12:15h)
1.1 Reminder of linear regression
1.2 Exercise
1.3 Reminder of ANOVA
1.4 Exercise

Session 2. 10:00 – 14:00 h (break from 11:45 to 12:15h)
2.1 General linear model (GLM). Introduction and example 1
2.2 GLM, Example 2
2.3 Exercise
2.4 GLM with random factors (mixed effects models). Example

Session 3. 10:00 – 14:00 h (break from 11:45 to 12:15h)
3.1 GLM with random factors. Hierarchical (nested) designs
3.2 GLM with random factors. Random slope designs
3.3 Exercise
3.4 Exercise

Session 4. 10:00 – 14:00 h (break from 11:45 to 12:15h)
4.1 Generalized linear models (GLZ). Introduction and example binomial distribution
4.2 GLZ, example Poisson distribution
4.3 GLZ, example with over-dispersion (negative binomial distribution)
4.4 GLZ with random factors. Example

Session 5. 10:00 – 14:00 h (break from 11:45 to 12:15h)
5.1 Exercise
5.2 Generalized additive models (GAM). Introduction and example
5.3 GAM models with random factors. Example
5.4 Exercise

Instructor: Aldo Barreiro Felpeto. CIIMAR.
Price: 250 € (200 € for CIIMAR/UP members)
Registration: after announcement, up to fill 25 available positions.
Registration, together with the payment information, is available in the CIIMAR website, through the link that is sent with the e-mail announcing the course.
Steps to register:
- Ask via email (abarreiro@ciimar.up.pt) if there are spots available.
- Register through the course link sent in the information email (or found at CIIMAR website)
- Send proof of payment required to book the place (send proof to abarreiro@ciimar.up.pt).

After sending the proof of payment, a confirmatory e-mail for the registration will be sent.

Important information:

- All the course will be taught through a zoom platform.
- The course will be taught in English.
- At least a beginner's background in R and basic statistics is recommended.
- All the information and materials necessary for the development of the course (instructions to

install R and R packages, pdf with lessons content, scripts with examples and exercises, data for examples and exercises) will be made available for all the participants in the course through a link to the *Open Science Framework* platform.

Syllabus

Course description

The course's first session is a reminder of linear regression and analysis of variance, the two statistical techniques that constitute the basis of the most important frames of statistical models. These statistical models (GLM, GLZ and GAM) will be the content of the next four sessions.

GLM (General Linear Model) will be studied in the second and third sessions. Due to its robustness and relative simplicity, this is the most widely used statistical model frame. For this reason, in the course we spend more time with GLM than with the others modelling frameworks (GLZ, GAM). Through worked examples will be explained the following topics: I) GLM assumptions, how to test them and how to solve problems when certain of these assumptions are not met (data transformation for non-normality, alternative fitting by generalized least squares for variance heteroscedasticity and residual autocorrelation); II) contrast of hypothesis for the main effects and post-hoc tests for the simple effects; III) visualization of model predictions.

In the GLM will be explained with particular detail the inclusion of random factors (simple mixed effects models, hierarchical designs, and random slopes designs).

GLZ (Generalized Linear Models) are an extension of GLM for error distributions that could be different than Gaussian. Three worked examples will be shown for different distributions (binomial Poisson and negative binomial, for over-dispersed data), and a simple example including random factors.

GAM is a statistical model frame that could incorporate any features of GLM and GLZ. But, in addition, and considering additional restrictions, they could incorporate smoothed functions (non-linear but non-parametric) to describe the relationship between the response variable and one or several predictor variables. This statistical model frame has been gaining popularity in different fields of science due to the increase in computer power over the last 20 years.

At least a beginner's level in R language is recommended for this course, and also some familiarity with basic statistical techniques (regression, contrast of hypothesis, ANOVA).

The course is open to any level from undergraduate students to senior researchers. It is considered as the "second part" of another course that is regularly taught in CIIMAR entitled *Introductory Statistics for Biological Sciences*.

Course methodology

The course contents are short theoretical introductions to each topic followed by worked examples in *R* language, throughout which all features of the specific topic are shown. At the end of each section, the students will be provided data to perform exercises that will be revised and corrected during the lessons.

The theoretical explanations, as well as the worked examples are fully developed in a pdf, with R scripts with the examples as additional support. Access to all this material will be provided days before the course through a link to the *Open Science Framework* platform. Solutions to the exercises will also be delivered to the students by the end of the course.

General aim of the course

Overview of the main statistical model frames to understand the differences between them and consequently, their specific contexts of applicability.

Specific aims of the course

- Understand the principles of the statistical model frames in order to apply them according to the features of data.

- Test model assumptions with contrast of hypothesis and diagnosis graphs, and learn to solve problems associated with this process (data transformation, alternative fitting techniques, alternative error distributions, using non-linear or non-parametric fits).
- Understand the difference between random and fixed factors and the possible ways to incorporate random factors in the model design.
- Learn about the different contrast of hypothesis that can be performed in a statistical model (analysis of variance and analysis of deviance for the main effects, Tukey, Dunnett, least significant differences, etc. for the post-hoc).
- Learn the most standard methods to select the best model among a series of candidates fit to the same data.

Aldo Barreiro Felpeto is a researcher at Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR) associated to the University of Porto (Porto, Portugal). His research career has focused in plankton ecology. He defended his Ph.D. dissertation in 2007 in the Department of Ecology at the University of Vigo (Vigo, Spain) about interactions between zooplankton and toxic phytoplankton species from the Spanish NW Atlantic coast, southern Baltic sea and southern Tirreno coast. In 2008-2010, he performed a post-doctorate in the Department of Ecology and Evolutionary Biology at Cornell University (Ithaca, New York, USA). Since 2011 he is a researcher at CIIMAR.
He developed a strong background in statistics and dynamic modelling with R software, attending 10 courses in the period 2006-2018 and since 2013, organizing 14 editions of courses about different aspects of statistics and programming with R, mostly in CIIMAR, but also in the University of Vigo (Spain) and the University of Magallanes (Chile). He co-authored two books about statistics and programming: *Tratamiento de Datos* (Ed. Díaz de Santos, Madrid, 2006) and *Tratamiento de Datos con R, SPSS y ESTATISTICA* (Ed. Díaz de Santos, Madrid, 2010).
Due to his expertise in statistics and programming, he has developed collaborations in different fields of ecology, but also environmental sciences and molecular biology. He has published 60 articles, accounting for an *h* index of 27 and an *i10* index of 48.