23637 - Modern Statistical Computing in R

Syllabus Information

Academic Course: 2021/22
Academic Center: 304 - Faculty of Law and Economics
332 - Faculty of Economic and Business Sciences

Study: 3327 - Bachelor's degree in International Business Economics

Subject: 23637 - Modern Statistical Computing in R
Credits: 5.0
Course: 3 and 4
Teaching languages:

Theory: Group 1: English

            Group 2: English

Seminar: Group 101: English

            Group 102: English

Teachers: Ferran Carrascosa Mallafre, Pere Lopez Brosa
Teaching Period: Second Quarter o Fourth Quarter

Presentation

Over the recent years, **R (**https://www.r-project.org**)** has become the leading tool for statistical computing and graphics. The basic language of **R** is greatly enhanced by numerous contributed packages submitted by users. The majority of computing in the leading applied statistical journals is done in **R**, and it is used almost exclusively in some of the leading-edge applications, such as in genetics and data sciences. This software permits data analysts to interact with their data and to design personalized protocols for statistical analysis. **R** is *free software* that can run in most of the computer platform systems (Windows, OS of Mac, Unix, etc.). The purpose of this course is to set a foundation for full exploitation and creative use of this statistical modern language for computing and graphics.

Much of the statistical methodology implemented in software packages is used in the form of a black box. This is advantageous for a user who is not interested in the details of the methods, but the result is often a second-rate application, because the implementation, even if of high quality, is often meant for a different context, small details in the setting of options are ignored or misunderstood, and the orientation in the output, formatted for general interest, is difficult.

The course will introduce students to the syntax and inner workings of **R**, to become proficient in everyday computational tasks with datasets of all kinds, skilled in applications of elementary statistical methods, with an emphasis on (initial) data exploration and simple graphics. Focus will also be placed on opportunities to enhance the learning experience in other statistical courses by illustrating and applying basic statistical concepts using **R.**

The course will also include a gentle introduction to Python, a platform that has been essential in data processing in the recent years.

Learning outcomes

At the end of the course, students will have learned

- to use fundamental tools for computing in the practice of quantitative analytical methods (the 'paper-and-pencil' tool of the 21st century), that can work for the small jobs (like a pocket calculator) as well as for the big jobs (complex statistical data analysis).

- programming, data handling, transformations, subsetting, exploratory data analysis, probability distributions and simulations, regression and linear models, summarizing data, how to handle large data sets, effective graphics.

- modern concepts of statistics based on simulations and writing a report of a quantitative analysis.

Prerequisites

An introductory course on probability and statistics is basic for enrolment to this course. For the students of UPF, the compulsory requirement is the Probability and Statistics of the second year in the studies of ECO/ADE/IBE.

Contents

1. Introduction to R

2. Variables and data structures.

3. Data input and output

4. Basic and advanced graphics

5. Flow control and programming structures

6. Exploratory data analysis

7. Linear models, Matrix algebra

8. Simulation, permutation tests and bootstrap

9. Optimization

10. Introduction to Python

Teaching Methods

The course will combine theory and seminar sessions in small groups of students.

The theory will develop the general concepts and methods of the subject.

The seminar classes are 6 sessions throughout the term for each of the two groups into which the class is divided. The seminars will be taught in the computer room and will use R and Python software. In the first 4 of them, the exercises previously posted in the Global Classroom will be resolved as well as any doubts that may arise about their resolution. The last two seminars will be devoted entirely to topics of discussion and preparation of the final project.

Both the exercises and project should be elaborated with RMarkdown and/or Python notebooks.

Evaluation

Method of Assessment

Assessment is composed of the following inputs:

1. Continual evaluation: contribution to class + homework (15%)
2. Main Project (35%)
3. Final exam (50%).

A minimum of 3 points, out of 10, is required in each of the three items to pass the course

The main project will involve some computing in R/Python and submission of a report of up to 6 typed pages (not counting appendices).

Students will select their projects from topics of their own interest (upon the acceptance of the instructors) and will make a brief oral presentation at the end of the course.

The Retake Final Exam (RFE): Those students that have attend the course complying with all the required activities (including the final exam) and that have fail the subject, are entitled to a RFE and revise their failed final grade. The revised grade will be a weighted average (weights: .3 and .7 respectively) of (1) the CEG attained during the course and (2) the grade of the RFE. A necessary condition for the revision of the grade based on the RFE, is the score of RFE to be grater or equal than 4 (on a scale 0 to 10).

Bibliography and information resources

Required Readings: Handout material will be posted on the web as the course evolves.

Recommended bibliography:

Students are encouraged to consult the following sources on their own.

Dalgaard, P. (2002), *Introductory Statistics with R*, Springer

Dennis, B. (2013). *The R Student Companion,* Taylor & Francis Group

Matloff, N. (2011). *The Art of R Programming: A Tour of Statistical Software Design*, William Pollock

Chihara, L. and Hesterberg, T. (2011), *Mathematical statistics with resampling and* R, Wiley

Lander, J. P. (2014) R *for Everyone: Advanced Analytics and Graphics*, Addison-Wesley Data & Analytics Series

J. VanderPlas (2016). *Python Data Science Handbook*, O'Reilly Media, Inc.