UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Modelli Statistici e R

2526-2-E4101B042-E4101B042-1

Learning objectives

The course has a twofold objective: to introduce students to programming in R and to provide the fundamental statistical tools needed to jointly analyze multiple variables measured on the same set of statistical units. By the end of the course, students will be able to:

- use the R programming language to perform data analysis and address probabilistic or inferential problems, including through simulations;
- read, understand, write, and execute R scripts, both their own and those written by others;
- analyze relationships between variables using linear regression, both from a theoretical perspective and through practical application in R.

Contents

The course provides an introduction to the R programming language, covering topics in programming, descriptive statistics, probability, and statistical inference. The second part of the course is devoted to the specification, estimation, validation, and diagnostics of statistical models.

Detailed program

First part:

- **Introduction to R language:** using R as a scientific calculator; introduction to objects and their classes (vectors, booleans, matrices, data.frames, lists); conditional structures and loops; functions.

- **Descriptive statistics in R:** review of the main topics in univariate and bivariate descriptive statistics applied to datasets, including graphical representations and their customization.
- **Probability calculations in R:** key functions for working with random variables; Monte Carlo methods to approximate integrals and probabilities.
- **Statistical inference in R:** study of the properties of estimators through simulations; numerical methods for likelihood analysis.

Second part:

- **Multidimensional random variables:** joint density and distribution functions; marginalization; moments; mean vector and variance-covariance matrix; multivariate normal random variables and their properties.
- **Model specification:** steps for specifying a statistical model; model classification.
- **Simple linear regression model:** assumptions; parameter interpretation; parameter estimation (least squares and maximum likelihood); properties of estimators; Gauss-Markov theorem; coefficient of determination.
- **Model validation and usage:** hypothesis testing on the value of a single coefficient; hypothesis testing for model goodness-of-fit; using the model for point and interval predictions.
- **Model diagnostics:** methods to evaluate assumptions related to model structure, errors, and absence of unusual observations.
- **Multiple linear regression model:** model specification in matrix form and its assumptions; parameter interpretation; parameter estimation (least squares and maximum likelihood); properties of estimators; Gauss-Markov theorem; multiple coefficient of determination.
- **Qualitative variables:** incorporating qualitative variables into the model using dummy variables; interactions.
- **Testing a system of linear hypotheses:** general theory and specific cases.
- **Model selection:** absolute and relative contribution of an explanatory variable; partial determination index (PDI); criterion-based selection of explanatory variables using backward, forward, and stepwise approaches; AIC and BIC.

Prerequisites

Knowledge of the notions given in the courses "Statistics I", "Probability", "Linear Algebra", and "Statistical inference (Statistics II)" is required.

Teaching methods

The course is delivered in Italian and includes both classroom lectures and computer lab sessions.

The classroom lectures aim to deepen the student's theoretical knowledge on the course topics and their formalization. The computer lab sessions focus on the implementation aspects of the models on real and simulated data using the R software.

In particular, ** the **Statistical Models and R** course includes a total of **73 hours** of lectures conducted in person, each consisting of 2 or 3-hour blocks, many of which will be held in a computer lab. Additionally, tutoring activities will be provided to support the students.

Assessment methods

To pass the Multivariate Statistical Analysis course, it is necessary to obtain a grade of 18 or higher in both parts that make up the course (Statistical Models and R (9 CFU) and Exploratory Analysis (6 CFU)). The final grade is determined by the weighted average (with the respective CFU) of the grades obtained in the partial exams.

For the **Statistical Models and R (9 CFU)** part:

- the exam is written and consists of 3-4 exercises. These exercises include theoretical questions, programming exercises, classic written exercises and/or real data analysis, and also involve the use of R.
- There are two midterm assessments: the first (winter session) focuses on R programming, and the second (spring session) on linear models. In all other sessions, the full written exam must be taken.
- The use of texts or any other materials is not permitted during the exam, except for the codes provided by the instructor at the beginning of the exam.
- Students, as well as the instructor, can request an optional oral exam (covering the entire 9 CFU program).

Textbooks and Reading Materials

First part: Statistical Models and R (9 CFU)

- Lecture notes from the instructor
- Albert, J. & M. Rizzo (2012). *R by Example*. Springer.
- Stefano Leonardi. Impariamo R - Un'introduzione facile al linguaggio. Città Studi edizioni.
- Venables, W. N., Smith D. M. & the R Core Team (2021). *An Introduction to R*.
- M. Grigoletto, F. Pauli, L. Ventura, Modello lineare, teoria e applicazioni con R. Giappichelli, 2017
- J. Fox. Applied regression analysis and generalized linear models, third edition. Sage.
- Piccolo, D. (2010), Statistica, Terza edizione, Il Mulino.

Semester

The course is scheduled in the first semester and in the second part of the second semester.

Teaching language

Italian

Sustainable Development Goals

QUALITY EDUCATION