UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Analisi Statistica Multivariata

2526-2-E4101B042

Learning objectives

The aim of the course is to provide the statistical tools necessary for the joint analysis of multiple variables measured on the same set of statistical units. The skills acquired during the course enable students to:

- explore and summarize data;
- model data using regression techniques;
- produce and interpret the output of real data analyses carried out using the R programming language.

The course contributes to strengthening knowledge and understanding in the field of multivariate statistics, in line with the “Statistics” learning area of the Bachelor's Degree Programme in Statistical and Economic Sciences.

Contents

The course is composed of the parts: Statistical Models and R (first part, 9 CFU) and Exploratory data analysis (second part, 6 CFU).

- **Statistical Models and R (9 CFU)** provides “hands-on” training for learning how to analyse data in the R statistical software package and offers an introduction to linear regression models. It covers data input/output, data management and manipulation, and how to make useful and informative graphics, as well as how to handle a complete regression analysis.
- **Exploratory Analysis (6 CFU)** offers an introduction to the statistical analysis of multivariate observations with the goal of dimensionality reduction thereby facilitating the understanding of the data.

Detailed program

First part: Statistical Models and R (9 CFU)

- **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.
- **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.

Second part: Exploratory Analysis (6 CFU)

- Graphical representation of multivariate data
- Total and generalized variance
- Spectral decomposition theorem
- Principal components analysis
- Cluster analysis: K-means and hierarchical methods
- Factorial analysis

Prerequisites

Knowledge of the notions given in the courses "Statistics I", "Probability", "Matrix 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 (9 CFU)** section 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.
- The **Exploratory Analysis (6 ECTS)** part includes a total of **42 hours** of lectures conducted in person, with 7 of these hours taking place in the computer lab. Additionally, tutoring activities will be provided to support students, conducted remotely in synchronous mode.

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 in-term tests: the first concerning programming in R and the second concerning linear models.
- 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).

The exam for **Exploratory Analysis (6 CFU)**

- is divided into two parts: the *first written* part consists of 3 open-ended questions, including theoretical questions and numerical exercises to be solved without the use of a computer, and the *second part* consists of 2 data analysis exercises to be completed using R/RStudio on the online exam platform.
- Students and the instructor may request an optional oral exam covering the entire program.
- 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.
- The use of mobile phones or any digital support is not allowed during the exam.
- The evaluation of the two parts of the Exploratory Analysis exam is proportional to the credits allocated during the course to the theoretical and computational parts.

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.
- 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.

Second part: Exploratory Analysis (6 CFU) - Lecture notes from the instructor

- Johnson, Wichern (2014) Applied Multivariate Statistical Analysis (6th Edition), Pearson Prentice Hall
- Everitt, Hothorn (2011) An Introduction to Applied Multivariate Analysis with R, Springer

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