UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models II

2324-2-F8203B042-F8203B013M

Learning objectives

The course aims are to provide analytical and inferential procedures using advanced statistical models and simulations. The content is presented to develop a critical understanding of the underlying theoretical underlying assumptions.

Knowledge and understanding
This teaching enables the student to:

- Analyze data with advanced statistical models developed for both categorical and continuous response variables; implement simulations independently;
- Use the semantics of the R software also through the RMarkdown environment to create reproducible documents containing the code, results, and comments of the analysis;
- Rigorously interpret the results of the empirical analyses and provide a description also for dissemination purposes.

Ability to apply knowledge and understanding

The course allows the student to:

- Develop statistical inference using modern bootstrap techniques;
- Estimate, select, and interpret the statistical model-based clustering techniques especially considering finite mixtures of distributions for heterogeneous populations;
- Apply theoretical knowledge to the analysis of data collected in various fields, including epidemiology, medicine, biology, genetics, and public health;
- Estimate models with latent variables
- Implement the code in the language of the R software.

The exercises that are assigned weekly are aimed at integrating the theory part with explanations of the procedures and results of empirical analysis through applications in an organic manner.

The teaching is fundamental for the subsequent university course as it provides the essential concepts for the development in both the theoretical and applied fields of statistics for the target job contexts (biostatistics/statistics/demography and related) of students on the Biostatistics degree course.

Contents

In the first part of the course, methods are introduced to generate pseudo-realizations from random variables. Next, the student is introduced to some resampling methods: bootstrap and jackknife, along with their inferential purposes.
The Expectation-Maximization (EM) algorithm is illustrated for incomplete-data problems through the estimated parameters of a generalized linear model. Then it is illustrated as an optimization method for the estimation model parameters of the finite mixture models of Gaussian distributions and latent variable models such as the latent class model. The course provides skills in the use of the semantic of the software R.

Detailed program

The first part of the course deals with simulation methods and linear congruential methods to generate pseudo-random numbers. Graphical tools for testing the series and statistical tests such as Kolmogorov-Smirnov and Chi-Squared tests are illustrated. Simulation of random numbers from specific distributions is considered. Some theoretical features of the exponential, binomial, and Gaussian distributions and convolution of random variables are exposed. The convolution of random variables and the generation of realizations from it are considered.
The main resampling methods, such as the jackknife and bootstrap, are introduced in the second part of the course. The bootstrap is applied for bias adjustment and the estimation of dispersion. Bootstrap confidence intervals based on the percentile method and the bias-corrected accelerated bootstrap method are explained.
The autoregressive Poisson model for count data and the similar model based on the negative binomial distribution to account for overdispersion are introduced. The models are applied to the analysis of COVID-19 counts based on daily national data series.
Among the optimization methods, the Expectation-Maximization Algorithm is considered and explained first as a tool to impute missing values through a generalized linear model and then as a tool to maximize the log-likelihood function for incomplete data problems. Finite mixture models are introduced both for continuous and categorical data, and a particular focus is given to the mixture of Gaussian distributions and latent variable models for categorical data.
Some time is devoted to explaining the theory by imparting the flavor of the empirical applications using data collected from different fields arising in epidemiology, pharmacoepidemiology, medicine and biology, and ecology and environmental sciences. They are developed within the statistical software R, RStudio with the RMarkdown interface. The main R packages used are bootstrap, dplyr, MASS, MultiLCIRT, tscount, mclust e skimr.
The student is encouraged to develop reproducible documents in which he/she critically comments on the code and the analysis results, also through cooperative learning.

Prerequisites

For an easier understanding of the course content, it is recommended to know Probability and Statistical Inference notions. The student should also know the basic semantics of the programming language in the R environment.

Teaching methods

Lectures are provided on the theoretical part concerning the basic concepts of statistical models. The theory lessons are complemented by practical exercises that allow the student to learn theory and data analysis techniques. Lessons take place in the computer lab. Weekly summarising exercises are assigned based on the application of the proposed models to real or simulated data related to the syllabus. During the course with the help of R in the RStudio environment and the RMarkdown interface, students learn to process reproducible documents. They are encouraged to tackle the application problem with the further aim of developing cooperative learning.

Assessment methods

The following methods of verifying learning apply to both students attending and non-attending lectures in presence. The examination is in written form with open questions and with optional oral; there are no intermediate tests. Students may receive a bonus if they provide selected excercises when required. The written exam has a maximum total duration of two hours and takes place in the computer lab. During the examination, open theory questions must be answered, and exercises must be solved in the light of the theoretical topics developed during the course. The theory questions allow verifying the learning of the theoretical concepts taught during the course. The empirical analyses are conducted using the R environment, Rstudio, and RMarkdown and allow verifying the ability to understand the problem and its resolution by applying advanced statistical models to real or simulated data and the elaboration of reports in which the procedure is described, and the results are illustrated. The examination is open book and students can consult the R code used during the course. The student passes the test with a mark of at least 18/30.

Textbooks and Reading Materials

The teaching material consists mainly of handouts prepared by the teacher. They cover both the theory topics and the applications developed with R or SAS software. All the files are available on the page of the e-learning platform of the university dedicated to the course. In addition, the teacher publishes at the end of each lesson: the slides, the calculation programs, the exercises, the datasets, and the solutions of the exercises. On the same page are also published some previous exam texts.

The primary reference texts are listed in the bibliography of the handouts; among others, the following are noted. Some of these also available in ebook the following:

Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). Latent Markov Models for longitudinal data, Chapman and Hall/CRC, Boca Raton.

Bishop, Y. M., Fienberg, S. E., Holland, P. W. (2007). Discrete multivariate analysis: theory and practice. Springer Science & Business Media, New York.

Blitzstein, J. K., Hwang, J. (2014). Introduction to probability, Chapman & Hall/CRC.

Gentle, J. E., Hardle W., Mori Y. (2004). Handbook of computational statistics. Springer-Berlin.

Lange, K. (2010). Numerical analysis for statisticians, 2nd Edition, Springer, New York.

Pennoni, F. (2023). Dispensa di Modelli Statistici II, parte di teoria e applicazioni con R. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Semester

Semester I, cycle I, October-November 2023

Teaching language

The course is provided in Italian. Erasmus students can use the handouts material in English and ask the teacher to carry out the exam in English.

Sustainable Development Goals

GOOD HEALTH AND WELL-BEING | REDUCED INEQUALITIES | CLIMATE ACTION