UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

# SYLLABUS DEL CORSO

# Statistical Modeling

**2324-1-FDS01Q004**

## Aims

The course aims to provide students with methodological and applied background on advanced statistical models: multiple linear regression and some extensions, some generalized linear models and some model-based approaches to cluster analysis concerning finite mixture models of Gaussian distributions.

**Knowledge and understanding**

The student is introduced to advanced statistical models for analysing data with different types of response variables. The relevant assumptions underlying the theory are also illustrated by considering the maximum likelihood and least squares estimation methods for model parameters. Data analysis is conducted using R software and the RMarkdown environment, which allows reproducible documents containing code, results and comments to be created. Applications cover real and simulated data from various fields such as economics, finance, and social sciences. The student is also encouraged to provide a critical evaluation of the results obtained from the empirical analyses. The course allows the students to acquire solid elements of theory and applications. It concerns data science, and this knowledge is essential nowadays in each working environment, and it is compulsory for the next courses of student studies.

**Ability to apply knowledge and understanding**

The course provides skills in using the semantics of the open-source software R for the descriptive analysis of multivariate data and parameter estimation of univariate and multivariate models. Through R and RStudio, students learn how to organically set up statistical reasoning by analysing data and writing reports that illustrate code, analysis and results. Theory is complemented by practical applications. The course enables students to acquire a solid theoretical foundation and the ability to apply the proposed statistical models to real data.

## Contents

In the first part of the course, after a brief introduction to the conceptual framework of statistical inference, the resampling procedure known as bootstrap is illustrated in order to obtain measures of accuracy with respect to estimators of interest. Next, the multiple linear regression model is presented with its assumptions. Ordinary least squares and maximum likelihood estimation, as well as statistical properties of least squares estimators are introduced. Measures of fit, regression diagnostics and prediction are also covered. Generalized linear models are considered, with reference to the multiple logistic regression model and multinomial logit model. The expectation-maximisation algorithm is introduced as a tool for maximum likelihood estimation of classification model parameters. The course provides skills in the use of R software semantics, also using the RMarkdown library via the knitr library to integrate code, analysis results and comments.

## Detailed program

The course starts with an introduction to the picture of statistical inference and some related concepts of causal inference.

- The first part of the course is an introduction to the resampling method known as bootstrap for determining the standard error as a measure of accuracy. The method is applied to various estimators using data of interest.

- The course covers the multiple linear regression model, least-squares and maximum likelihood estimation methods. The properties of the least-squares estimators are discussed on the basis of the model assumptions.

- During the course, the student's knowledge based on univariate distributions is extended to include the bivariate and multivariate Gaussian distributions. Random realizations are drawn from these distributions. The distribution is also represented graphically with contour lines.

- Various diagnostic tools for model evaluation based on regression residuals are considered with particular emphasis related to the outliers, influential and leverage points. The problem of selecting the most relevant explanatory variables using informative criteria such as Akaike's criterion is addressed. One also learns how to evaluate the model in relation to its predictive ability.

- Generalised linear models for the analysis of categorical response variables with two or more categories are introduced. The multiple logistic regression model and the multinomial model are illustrated, with particular emphasis on the interpretation of regression coefficients.

- Among the optimization methods, the expectation-maximization algorithm is explained as a computational tool to maximize the likelihood function.
  Mixture models are introduced as model-based clustering methods. Features of mixtures of Gaussian distributions are presented in detail.

Some time is devoted to explaining the theory by imparting the flavor of the empirical applications using data from different fields such as economics, finance, biology, ecology, and environmental sciences. They are developed within the statistical software R, RStudio using many different libraries with the RMarkdown interface and the library knitr in order to introduce the student to principles of reproducible research. The student is to write reproducible reports in which he/she comments on the code and the analysis results critically, and through cooperative learning through the assigned homework.

## Prerequisites

For an easier understanding of the course content, it is recommended to know the contents of the course Foundations of Probability and Statistics. The course assumes prior knowledge of the following topics: probability of an event, probability distribution function, density, cumulative distribution functions, the law of total probability, independence of events, Bayes theorem, expectation and variance of a random variable, standardization and percentiles of a random variable, continuous and discrete random variables such as Bernoulli, binomial, Poisson, geometric, uniform, exponential, Gaussian, Student-t, chi-squared, graphs and numerical measures to describe data. Statistical inference, and maximum likelihood inference and basic knowledge of multivariate data analysis and linear algebra. Students should also know the basic semantics of the programming language in the R environment.

## Teaching form

Lectures are provided on the theoretical aspects, and are complemented by practical exercises that allow the student to learn theory and apply the models for analyzing real and simulated data. Lessons take place in the computer lab. Weekly summarizing exercises are assigned as homework to favor the learning of the theory and its applications. During the course, with the help of R in the RStudio environment and the RMarkdown interface, students also learn to process reproducible documents. They are encouraged to tackle the application problem with the further aim of developing cooperative learning. Tutoring lectures are also scheduled to help students develop exercises and compare solutions.

## Textbook and teaching resource

The teaching material consists mainly of handouts prepared by the teacher. They cover the theory topics and the applications developed with the R 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 the slides, the calculation programs, and the datasets at the end of each lesson. Weekly exercises are assigned, and some solutions are provided and discussed. On the same page are also published some examples of the examination text. The primary references will be listed in the bibliography of the handouts; among others, the following are noted. Some of these are also available at the library and in ebooks:

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.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). Model-based clustering and classification for data science: With applications in R. Cambridge University Press.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. D. (2021). Regression: Models, methods and applications. Springer Berlin, Heidelberg.

Faraway, J. J. (2014). Extending the Linear models with R, 2nd Edition, Chapman & Hall, CRC Press. Hastie, T., D. and Tibshirani, R. (2013). An introduction to statistical learning, New York, Springer.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, 2nd Edition. Chapman and Hall/CRC, London.

R Core Team (2023). R: A Language and Environment for Statistical

Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Xie, Y., Dervieux, C. and Riederer E. (2020). R Markdown Cookbook. Chapman & Hall, CRC

## Semester

Semester II, March-May 2024

## Assessment method

The following methods of verifying learning apply to students attending and non-attending lectures held in the lab. The examination is written with open questions and optional oral; there are intermediate tests. The written exam has a maximum total duration of an hour and a half in the computer lab. During the examination, open theory questions must be answered, and exercises must be solved in light of the theoretical topics covered during the course and the practical exercises assigned weekly during the course. The theory questions verify the learning of the theoretical concepts taught during the course. The empirical analyses are conducted using the R environment, Rstudio, and RMarkdown and allow demonstrating the ability to understand the problem and its resolution by applying advanced statistical models to real or simulated data and the elaboration of reproducible reports in which the procedure is described, and the results are illustrated. During the examination, the use of the study material and R code implemented during the course is permitted. Each question will be evaluated approximately 3 or 4 points. The student passes the exam with a score of at least 18 out of 30.

## Office hours

Weekly, annunced on the elearning page during the course.

## Sustainable Development Goals

GOOD HEALTH AND WELL-BEING | QUALITY EDUCATION | REDUCED INEQUALITIES | PARTNERSHIPS FOR THE GOALS