UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Statistical Modeling

2223-1-FDS01Q004

Learning objectives

The course aims to provide students with methodological and applied background on advanced statistical models: multiple linear regression and generalized linear models used to analyze data arising in many different fields. The course also introduces the student to some model-based approaches to cluster analysis. Along with modeling, the focus of the course is also tailored to introduce some computational methods.

Knowledge and understanding
The student is introduced to the advanced concepts of statistical models and the related assumptions underlying each model formulation for discrete and continuous response variables. Maximum likelihood estimation of the model parameters is considered. Then, he/she learns how to apply the models through the open-source software R to perform solid statistical analyses in many different contexts using simulated data and data from economics, finance, business, astronomy, environmental and social sciences. The student is encouraged to interpret the parameter estimates, evaluate the models, and verify their sustainability.

Ability to apply knowledge and understanding
The course provides skills in using the semantics of the software R for descriptive multivariate data analysis and for estimating univariate and multivariate model parameters. The student learns through R and RStudio to make statistical thinking and data analysis processes visible and reproducible running the code and showing the results and comments through a simple plain text file. Theory and practical applications are jointly explained to support students with deep practical knowledge. 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 course of student studies.

Contents

In the first part of the course, after a brief introduction to the statistical learning, the course presents the multiple

linear regression model introduced with its assumptions, ordinary least squares, and maximum likelihood estimation, statistical properties of the least square estimators, estimation of the variance, measures of fit, regression diagnostics, and prediction. Next, the student is introduced to the bootstrap resampling method and its inferential purposes. Generalized linear models are also covered. The expectation-maximization algorithm is introduced as a maximum likelihood estimation tool for new techniques such as model-based clustering. The course provides skills in using the semantics of the software R also using the RMarkdown interface through the library knitr to integrate code, results and comments on the same file.

Detailed program

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

- The first part presents the linear regression model with multiple explanatory variables. The deviance decomposition and the method of the ordinal least squares are recalled, and comparisons are made with the maximum likelihood approach. The properties of the ordinal least square estimators are discussed according to the model assumptions. Inference for the regression coefficients is illustrated.
- 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.
- Diagnostic tools to evaluate the model through residuals are proposed, and some criteria for model selection, such as the Akaike Information Criterion, and Bayesian Information Crierion are exaplained.
- The student also learn how to make out of sample prediction and associate a measure of uncertainty to this prediction.
- The bootstrap procedure is also introduced as a basic tool for inference and applications are illustrated.
- Generalized linear models are introduced for binary and categorical data.
- Among the optimization methods, the expectation-maximization algorithm is considered as a tool to maximize the log-likelihood function for incomplete data problems.
- Finite mixture models are introduced as model-based clustering approaches. A particular focus is given on the mixture of Gaussian distributions for continuous data.

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 methods

Lectures are provided on the theoretical aspects, and they 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.

Assessment methods

The following methods of verifying learning apply to both students attending and non-attending lectures in presence. 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 the 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 verifying 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. The examination is carried out with an open book, and students can consult the R code implemented during the course and the exercises. 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 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 at the end of each lesson: the slides, the calculation programs, and the datasets. 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 ebook:

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 (2022). 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 2023

Teaching language

The course is delivered in English, and the examination will be in English.

Sustainable Development Goals

QUALITY EDUCATION | REDUCED INEQUALITIES | PARTNERSHIPS FOR THE GOALS