UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models

2122-2-E4102B084-E4102B085M

Learning objectives

The course aims to provide students with methodological and applied background on the multiple linear regression model and the multiple logistic regression models.

Knowledge and understanding

The student is introduced to the basic concepts of statistical models and the related assumptions. Then, he/she learns how to apply the models to perform solid statistical analysis in many different applied contexts: economics, business, biology, physics, astronomy, environmental and social sciences.

Ability to apply knowledge and understanding

Some theory related to computations using the matrix algebra is illustrated. He/she learns how to verify the tenability of the model. The course provides skills in using the semantic of the software R and SAS for descriptive multivariate data analysis and multiple linear and logistic regression. He/she also learns to draft reports with the illustration of the analyses and comments on the results. Theory and practical applications on real and simulated data 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

At the beginning of the course, the student is introduced to the big picture of statistical inference and the multivariate graphical examination of the data, and the use of linear total and partial correlation coefficients to inspect the linear associations among continuous variables.

During the course, the following main issues are raised. The multiple linear regression function is introduced with its assumptions. The ordinary least square estimation method is explained, and the main basic properties of the estimators are illustrated. The bivariate and multivariate Gaussian distributions are illustrated with their properties which are also explicated through applicative examples and simulations.

The model is evaluated by considering the following aspects: fit indices, information criteria, selection of explicative variables. Model diagnostics tools for checking model assumptions and unusual observations are taken into account, along with the multicollinearity issue. Prediction and linearization methods are introduced. Odds and odds ratios are introduced, and the multiple logistic regression is explained along with its estimation methods and uncertainty associated with the parameter estimates through the standard error and the interpretation of the resulting coefficients.

The R environment within the Rstudio and RMarkdown interface is employed to develop live code and output in the same interface and to make reproducible documents. SAS is employed to develop students’ skills in multivariate data analysis and multiple linear and logistic regression.

Detailed program

The course starts with an introduction to the big picture of statistical inference and causal inference concepts. The following features are also recalled: type of variables, the variance and covariance matrix, the correlation and partial correlation matrices.

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, and they are illustrated by means of the scatterplots in two and three dimensions. The contours of the Bivariate Gaussian distribution are depicted and described.

Many diagnostic tools are proposed to evaluate the model’s residuals, and some criteria for the variable selection, such as the Bayesian Information Criterion, the Mallow Cp index, are introduced. The multicollinearity is explained, and the variance inflation factor is used to provide a measure of the relative importance of each covariate. The way to forecast the response value for a new observation and the average value of the response is illustrated. The ideas of training e testing sets are also illustrated.

Other arguments raised during the course are i) maximum likelihood estimation method for the model parameters; ii) transformation of the variables; iii) categorical covariates; iv) models with some orders of interactions between covariates; v) odds and odds ratios; *vi*) categorical response variables and the general logistic model.

Some time is devoted to explaining the theory by imparting the flavor of the applications on real data collected from different fields. They are developed within the statistical environment R, RStudio with RMarkdown to make reproducible documents. The student is introduced to the semantic of the SAS software to carry out multivariate analysis and multiple linear and logistic regression.

Prerequisites

Positive examinations are required on the following courses: Statistics I, Mathematics, Linear Algebra, and Probability. For an easier understanding of the course content, it is strongly recommended to be familiar with the concepts of statistical inference taught in the Statistics II course.

Teaching methods

All the lessons take place in the computer lab: the theory part is flanked by the development of applications involving multivariate data referring to both real and simulated case studies and to different application areas.

Tutoring classes are also provided so that the student can be assisted in learning the theory and carrying out the exercises assigned weekly.

During the exercises, with the help of R in the RStudio environment and the RMarkdown interface, the student learns the relevant programming language and creates reproducible documents. The same analyses are carried out also by using the SAS software.

The student is encouraged to develop cooperative learning to interact with other students and finalize the required steps of the analysis. Exercises are carried out in a written form, and the results are reported with comments.

During the Covid-19 emergency period, the lessons will take place in the online asynchronous mode (videotaped lessons) with scheduled videoconferences meetings and some real meetings according to the availability suggested by the University.

Assessment methods

The following assessment methods of learning apply to both in-class and non-attending students. The exam is written, and it has a total duration of one hour and thirty minutes with optional oral, and it is held in the lab. It is carried out by answering open questions related to the theoretical and applied contexts using the computer. Real data analysis is carried out with R or SAS. Making a reproducible document, the student carries descriptive analysis on real and simulated data and applies the multiple linear regression model or logistic model. The student has to provide explanations concerning the code employed for the analyses and the results.

The exam allows evaluating the understanding of the theoretical parts, the analytical skills, and the ability to write a reproducible report. The oral test is optional and covers both theory and applications. It can be requested by those who have a result of at least 18/30 in the written test at the moment of publication. Results are published on the e-learning page dedicated to the course.

During the Covid-19 emergency, the exam will be the same, but it will be carried out in the lab or videoconference through Webex according to the university's guidelines.

Textbooks and Reading Materials

The professor’s lecture notes are available from the webpage of the course of the e-learning website of the university. In addition, at the end of each lecture, the following teaching material is downloadable from the course's web page: slides, R scripts, SAS code, exercises, solutions, and datasets.

On the same page are published some exam texts from previous years. In addition, the teacher is available weekly for interviews with students according to the times published weekly on the e-learning page of the course.

The teaching language is Italian. However, Erasmus students can meet the professor to define proper English textbooks and they can require to carry out the exam in English.

During the Covid-19 emergency period, video recordings of classes are also posted on the course page.

The reference texts are listed in the handout bibliography. Some of these include the following:

Faraway, J. J. (2014). Linear models in R, Second Edition, Chapman & Hall, CRC Press.

Johnson, R. A., and Wichern, D. W. (2002). Applied multivariate statistical analysis, Pearson Education International, Prentice-Hall.

Hastie, T., D. & Tibshirani, R. (2013). An introduction to statistical learning, New York, Springer.

Nolan, D., & Lang, D. T. (2015). Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving. Chapman & Hall, CRC Press.

Pennoni, F. (2021). Dispensa di Analisi Statistica Multivariata –Modulo Modelli Statistici- parte di teoria e applicazioni con R e SAS. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

SAS/STAT 9.4. PROC SGSCATTER, PROC CORR, PROC REG, PROC GLM, PROC GLMSELECT, User's guide, SAS Institute, 2012.

Semester

II Semester, III cycle: from February to April 2021

Teaching language

The course is taught in Italian. Erasmus students can use the didactic material in English and can ask the teacher for the exam to be held in English.