UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models

2324-2-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 concepts underlying statistical models and related assumptions. Learn about models through their use with real and simulated data. Learn to interpret the results and verify the sustainability of the model. Aspects of graphical analysis, and computational analysis using matrix notation are covered.
Ability to apply knowledge and understanding
The course develops the skills for analyzing data of a multivariate nature and coming from various information sources: business, economic, biological, physical, medical, astronomical, environmental, social and sports contexts. The student deepens the skills in the semantics of the R software for multivariate descriptive statistical analyzes and for the application of multiple linear regression, classical linear, logistic regression models. The student learns to create reports where he illustrates the analyzes carried out and comments on the results obtained.
The course allows the student to acquire the basic elements of theory and application of statistical models and qualifies as indispensable both for the subsequent university professional training course in data science.

Contents

The course is divided into four parts:

1. STATISTICAL MODELS. The introduction explains the concept of statistical analysis and statistical model. Some fundamental concepts necessary for carrying out the course are then introduced
2. THE MULTIPLE LINEAR REGRESSION MODEL. Methodologies for solving a descriptive multiple regression are explored. In particular, the least squares estimation method, the goodness-of-fit criteria and

the choice of one regression model over another are analysed. The presence of multicollinearities and outliers that can affect the quality of an analysis are evaluated.

3. THE CLASSICAL LINEAR REGRESSION MODEL In an inferential key, the classic linear model is proposed with its six assumptions and the hypothesis of normality of errors. It occurs, as the least squares sample estimators possess optimal properties to study the parameters of the population. Furthermore, under the assumption of normality of the errors, hypothesis tests are constructed on the significance of the individual parameters, and related confidence intervals. At this point, the criteria for choosing the models among the various possible ones and the use of the models for explanatory purposes are described.
4. EXTENSIONS The model with qualitative and mixed explanatory variables is proposed, the main transformations of the variables and the logistic regression model is analyzed in particular

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.

The multiple linear regression model is introduced first considering three variables with the extended notation and then through the matrix notation. The deviance decomposition and the method of the ordinal least squares are recalled. 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, 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

The lectures take place in a computer lab.In this context of theoretical lectures, theory part is supported by the development of applications concerning multivariate data referring to both real and simulated case studies and to different application fields: there are very numerous outputs on real and simulated data presented in R environments with the help by RMarkdown. In the practical exercises you will learn the necessary procedures and the codes necessary to carry out the exercises independently. In fact, with the help of R in the RStudio environment and the RMarkdown interface, the student learns the related programming language and creates reproducible documents. During the exercises the student is encouraged to recognize the problem of the exercise, and to identify the most suitable methodology, as well as to apply the analyzes and comment on the results.
Before each lesson, the slides and parts of the handout relating to the topics presented will be available on the students' e-learning page.

Assessment methods

The test is in written form. There are no intermediate tests. The following ways of verifying learning are valid both for students attending face-to-face and non-attending lessons.

The test takes place in the laboratory. The student must answer two theoretical questions from a set of predetermined questions that he/she will already know at the beginning of the course. It is necessary to argue the answer in understandable and comprehensive terms by reporting the required demonstrations. The reference point for answers is the slides and the handout: of course, knowledge gained from the recommended books can be reported. Formulas and graphs should be reported: if it is difficult, they can be written on paper with pen and then scanned. The required length of the answers will depend on the question: answers not exceeding four typed sheets in 12-gauge spacing 1.5 (12000 characters including spaces) are suggested.
The second part of the test will consist of a practical exercise on real or simulated data provided by the lecturer through the use of statistical packages. The statistical tools to be used will be those learned in the course. In the paper all graphs and outputs should be appropriately commented on, both from a theoretical point of view and with respect to the application under examination. The development is done through the R environment. The student may use the exercise codes during the exam. These codes will be provided on the day of the test

The second part of the exam will consist of a practical exercise on real or simulated data provided by the teacher through the use of statistical packages. The statistical tools that he will have to use will be those learned in the course. The paper must include detailed comments on the codes used and the results obtained. The development takes place through the R environment. The student will be able to use the codes of the exercises during the exam. These codes will be provided on the exam day.

Textbooks and Reading Materials

I principali testi di riferimento sono
• Spinelli, D. , Vittadini G.(2023) course slides
• Pennoni, F. Spinelli D, Vittadini G. (2023). 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.
• - Baltagi B. H. (2008), Econometrics, fourth Edition, Springer Berlin
• Faraway, J. J. (2014). Linear models in R, Second Edition, Chapman & Hall, CRC Press.
• - Freund, R. J., Wilson, W. J., and Sa, P. (2006), Regression Analysis: Statistical Modeling of a Response Variable, 2nd edition, Academic 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.
• Littell, R. C., Freund, R. J., and Spector, P. C. (2002), SAS for Linear Models, 4th Edition, Cary, NC: SAS Institute Inc.
• Manual SAS/STAT 15.1
• Nolan, D., & Lang, D. T. (2015). Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving. Chapman & Hall, CRC Press.
• R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
• Wooldridge, J. M. (2015). Introductory econometrics: A modern approach. Cengage learning.

Semester

II Semester, III Cycle

Teaching language

Italian

Sustainable Development Goals

QUALITY EDUCATION