UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Modelli Statistici

2021-2-E4102B084-E4102B085M

Learning objectives

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

Knowledge and understanding

The student is introduced to the basic concepts of statistical models and to the related assumptions. He/she learns on how to apply the models to perform solid statistical analysis in many different applied contexts: such as 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. H/she learns how to verify the tenability of the model. The course provides skills in use of the semantic of the software R and SAS for descriptive multivariate data analysis and for the application of the multiple linear and logistic regression. H/she also learn 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 the student with a 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 to the

multivariate graphical examination of the data as well as to the use of linear total and partial correlation coefficients to inspect the linear associations among variables.

During the course the following main issues are raised. The multiple linear regression function is introduced with the assumptions. The ordinary least square estimation method is explained and the main basic properties of the estimators are illustrated. The bivariate and multivariate Gauss distributions are illustrated with their properties also 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 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 on 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 to the concepts of causal inference. The following concepts are 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 with the model assumptions. Inference for the regression coefficients is illustrated.

During the course the knowledge of the student 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 Gauss distribution are depicted and described.

Many diagnostic tools are proposed to evaluate 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 inflated factor is used to provide a measure of relative importance of each covariate. The way to forecast a new observation and the average value of the response are illustrated. The ideas of training e testing sets is 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 amount of time is devoted to explain the theory by imparting 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. It is recommended to know the content of the course of Statistical Inference.

Teaching methods

Theoretical lectures as well as exercises are held in the lab. Theory is explained and during the lectures, many practical examples based on real and simulated multivariate data referred to different contexts of application are proposed to the students to be analyzed with R, RStudio with RMarkdown to make reproducible documents. The same analyses are carried out also by using the SAS software.

The student is encouraged to develop the cooperative learning in order 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. A tutor is available to help students with the weekly assignments.

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 are valid also for students not attending lectures. The exam is written and it is held in the lab. It is carried out by answering open questions related to the theoretical part and an applied context. The latter involves real data analysis to be performed with R or SAS. The student by making a reproducible document 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 learning material is available.

The exam allows to evaluate the understanding of the theoretical parts, the analytical skills as well as the ability of writing a report. The oral examination is not compulsory and can be required if the student earns a score of at least 18/30 in the written part. Intermediate assessments are not planned. The scores are published in the e-learning page. A scheduled meeting is planned to see the spreadsheet of the exam.

During the Covid-19 emergency the exam will be carried out according to the guidelines of the University.

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. At the end of each lecture the following teaching material is downloadable from the webpage of the course: lectures ‘notes, slides, R scripts, SAS code, exercises, solutions and datasets.

During the Covid-19 emergency period the videotaped lessons will be also published.

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. (2020). *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 (2020). 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 2020.

Teaching language

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