Dipartimento di Economia e Finanza

DEGREE COURSE IN STATISTICS
ACADEMIC YEAR 2023 - 2024
MULTIVARIATE STATISTICS

General information	
Year of the course	III
Academic calendar (starting and ending date)	First term (11/09/2023 – 15/12/2023)
Credits (CFU/ETCS):	10 CFU
SSD	Statistica, SECS-S/01
Language	Italian
Mode of attendance	Class attendance is strongly suggested

Professor	
Name and Surname	Alessio Pollice
E-mail	alessio.pollice@uniba.it
Telephone	080 504 9267
Department and address	Room n. 3, 5-th floor
Virtual room	MS Teams channel “Prof. A. Pollice - Ricevimento studenti”, code: y7zenm7
Office Hours (and modalities: e.g., by appointment, on line, etc.)	Tuesday 9.30 – 11.30, Friday 9.30 – 11.30, or by appointment.

Work schedule			
Hours			
Total	Lectures	Hands-on (laboratory, workshops, working groups, seminars, field trips)	Out-of-class study hours/Self-study hours
250	46	24	180
CFU/ETCS			
10 CFU			

Learning Objectives	Understanding and knowledge of the theoretical foundations and methodological developments of multivariate inference, linear models and multidimensional data analysis. Familiarity and autonomy in the application of the aforementioned methods with the aid of the R software.
Course prerequisites	• Notions of mathematical analysis; • Notions of matrix algebra; • Notions of probability theory; • Notions of statistical inference.

Teaching strategies	• The Multivariate Statistics course provides an introduction to statistical techniques that allow the study of multiple variables, in order to make inferences on their joint distribution, study the relationship between variables, the similarity between statistical units and represent the statistical units and/or variables in a space of reduced dimensionality. The lessons are accompanied by exercises in which the theoretical topics covered are applied, using the statistical software R, to real examples and case studies. • Frontal lessons on theoretical topics (about 46 hours) and exercises on the same topics using the R software (about 24 hours). If necessary, lessons and exercises can be provided in distance learning mode. • Course materials, self-assessment tests and exercises on the e-learning platform.

	• Self-assessment tests (multiple choice questions for each chapter of the course) are provided for the purpose of familiarizing with the procedures for carrying out the (partial) exams and are to be taken individually within the pre-established time limits. The outcome of the self-assessment tests helps to improve the overall evaluation of the commitment of the students enrolled in the course.

Expected learning outcomes in terms of	
DD1 Knowledge and understanding	The expected learning outcomes coincide with the acquisition of skills on the various chapters of the detailed program: understanding and knowledge of the theoretical foundations and methodological developments related to multivariate inference, linear models and multidimensional data analysis.
DD2 Applying knowledge and understanding	Familiarity in applying multivariate statistics methods to the analysis of economic data or data from other application contexts with the aid of R software.
DD3-5 Soft skills	• *DD3 - Making informed judgments and choices* Autonomy in the choice of multivariate statistics methodologies and in the evaluation of the results of their application with the R software to exercises and case studies referring to the economic context or to other application fields; • *DD4 - Communicating knowledge and understanding* Ability to communicate to specialists and non-specialists the theoretical contents of the discipline, the reasons for the choices to be made for carrying out exercises and examples and the interpretation of the results of the analyses conducted with the R software; • *DD5 - Capacities to continue learning* Autonomy in learning the topics of multivariate statistics and in the practice of using the R software in order to enhance skills and competences in subsequent studies and in work activity.

Content knowledge (Syllabus)	First part: Multivariate inference and linear models. • Discrete and continuous multidimensional random variables. Stochastic independence. Expected values. Variance/covariance matrices. Covariance matrices. Moment generating functions. • Multivariate Normal distribution and its parameters. Bivariate Normal distribution. Standardization. Moment generating function. Properties of the multivariate Normal. Wishart's and Hotelling's Distributions. • Inference on the parameters of the multivariate Normal. Maximum likelihood estimation. Sampling distributions of maximum likelihood estimators. Multivariate central limit theorem. Multivariate tests: union-intersection principle, generalized likelihood ratio. Hotelling's test. Hypothesis testing on the variance/covariance matrix. • General linear model. Multiple linear regression. Parameter estimation with the least squares method. Properties of estimates. Parameter estimation with the maximum likelihood method. Sum of squares and R^2. Hypothesis testing and confidence intervals for regression coefficients. Forecasts using the linear model. Removal of assumptions, analysis of residuals, intrinsically linear models, choice of explanatory variables, generalized least squares (heteroskedasticity and 1st order autocorrelation), multicollinearity, ridge estimators. Analysis of variance. Analysis of covariance.. • Generalized linear models. Exponential family, score and total score functions, canonical exponential family. Definition of GLM and generalities. Estimation of GLM parameters (Newton-Raphson and scoring methods), sampling distribution of the estimators. Model assessment. Quasi likelihood (outline).

	Second part: Multivariate data analysis. • Discriminant analysis. Fisher's linear discriminant function. Maximum likelihood discriminant function. Bayesian discriminant analysis. Least cost of misclassification. Estimation of the misclassification probability (outline). • Principal component analysis. Definition of principal components, sampling properties. Application problems and interpretation of the principal components. Choosing the number of principal components.. • Canonical correlation analysis. Definition of canonical components, sampling properties. Hypothesis testing. • Factor Analysis. Factor model. Model estimation: principal axis factorization, maximum likelihood method. Factor rotation. Factor scores: Bartlett and Thompson estimators. • Cluster analysis. Dissimilarity matrix. Hierarchical techniques. Non-hierarchical techniques. Finding the number of groups. • Correspondence analysis. Graphical representation. Introduction to multiple correspondence analysis.
Texts and readings	A. Pollice, Course handouts, available in e-learning mode. G. James, D. Witten, T. Hastie, R. Tibshirani (2021) An Introduction to Statistical Learning: with Applications in R Second edition. Springer Editor.
Notes, additional materials	The use of slides or personal notes is STRONGLY NOT RECOMMENDED and considered insufficient for the preparation.
Repository	In e-learning mode it is possible to carry out the self-assessment tests and download the course handouts, additional teaching materials, traces and data useful for carrying out the exercises with the R software. The address and password of the Multivariate Statistics course in e-learning mode are shared at the beginning of the course.

Assessment	
Assessment methods	For attending students, the evaluation of training activities is distributed over the semester and concludes with the end of the course. The outcomes of the self-assessment tests and laboratory activities contribute to integrating the overall evaluation of the profit. An intermediate and a final exam referring to distinct parts of the program and both based on a test with multiple choice questions and a laboratory exercise with the R software contribute to the evaluation. At the end of each exam, a short oral interview follows. Non-attending students must take a test referring to the entire course program and based on multiple choice questions and a laboratory exercise with the R software. An oral interview follows.
Assessment criteria	• *Knowledge and understanding* The first part of the exams referring to the two parts of the course consist of a test with 20 multiple choice questions to be carried out in 30 minutes. For non-attending students, the test covers the entire course program and includes 40 multiple choice questions to be completed in 50 minutes. • *Applying knowledge and understanding* The second part of the exams involves carrying out an exercise in which students are required to analyze of a dataset with the R software in 2 hours. Similarly, non-attending students are required to carry out the analysis of a dataset with R software with reference to the methodologies of the entire course program. • *Making informed judgments and choices* In order to evaluate the independence of judgment of the candidates, the comments contained in the execution of the exercises will be evaluated referring to: • the reasons for the choices made during the analyses;

	• the interpretation of the results obtained applying the methodologies of the course program with the R software. • *Communicating knowledge and understanding* After completing the exercise with R, the candidates will be called individually for a short oral public discussion based on the answers to the multiple choice questions. In order to evaluate the communication skills of the candidates, the comments contained in the execution of the exercises will also be considered referring to: • the reasons for the choices made during the analyses; • the interpretation of the results obtained applying the methodologies of the course program with the R software. • *Capacities to continue learning* In order to evaluate whether the candidates have developed the learning skills necessary to undertake further studies with a high degree of autonomy, they are questioned individually for a short oral public discussion based on the answers to the multiple choice questions.
Final exam and grading criteria	The evaluation of the exercises with R are carried out in the days following the exam, paying close attention to the authenticity of the contents. The exams showing identical sentences or expressions in the interpretation of the results or the same errors in the R commands are automatically cancelled. The cancellation of the exam entails taking the exam on the entire course program. Each multiple choice test gives rise to an evaluation out of 100. Even the exercises with R give rise to evaluations out of 100. The result of the oral interview improves or worsens the score obtained in both. The assessments of the intermediate and final exams give rise to a vote proposal out of thirty obtained as a synthesis of the two assessments. For non-attending students, the grade proposal is formulated on the basis of the evaluations of the test, the exercise with R and the oral interview referring to the entire program of the course.

Further information	