Universidad Zaragoza

Academic Year/course: 2022/23

61343 - Multivariate Analysis Techniques

Syllabus Information

Academic Year: 2022/23
Subject: 61343 - Multivariate Analysis Techniques
Faculty / School: 109 - Facultad de Economía y Empresa
Degree: 525 - Master's in Economics
ECTS: 3.0
Year: 1
Semester: Second semester
Subject Type: Optional
Module:

1. General information

1.1. Aims of the course

The course and its expected results respond to the following approaches and objectives:

The approach of the subject is markedly instrumental since its objective is to provide the student with a set of statistical tools widely used in carrying out a multidimensional exploratory analysis. All topics will be approached from a practical point of view, using different data sets to illustrate the techniques explained. For this, the programming language R will be used, which integrates a multitude of packages that increase its capacity and versatility. R has basic functions related to descriptive data analysis, and the most complex and current models related to the latest advances in the field of statistics or data analysis in areas such as Economics.

The course begins with a topic dedicated to the **Introduction to R** in which the most common commands are presented, very useful to acquire some basic notions and concepts. Next, different types of data are detailed and, later, basic commands to create, delete, define and operate with the data are explained.

The following topic entitled **Initial Data Analysis** shows the need to examine the data prior to the application of any statistical technique, in this way a basic understanding of the data and the relationships between variables is achieved. This analysis provides simple methods for organizing and preparing data, handling and evaluating missing data, identifying outliers, and testing underlying assumptions in most multivariate techniques. We begin with the one-dimensional statistical analysis and we provide a guide with the most advisable graphical representations and numerical summaries to describe the most important characteristics in each case. The next step is to analyze the existence of possible relationships between variables. This study is carried out from a two-dimensional perspective and the three general situations that may arise are analyzed. Thus, the guidelines to follow are provided when both variables are qualitative, both are quantitative and when one is quantitative and the other qualitative.

The rest of the topics already address multivariate techniques, therefore, first of all, they are presented in an intuitive way, and an adequate classification of them is made so that the student is able to select the technique that should be applied at each problem. Once the interest in these techniques has been motivated through different examples, next topic entitled **Factor Analysis** presents the two most used techniques in data reduction processes: Principal Component Analysis and Factor Analysis, indicating their advantages and disadvantages. Both techniques are used to analyze interrelationships between a large number of metric variables, explaining interrelationships in terms of a smaller number of variables called principal components or factors.

The next topic entitled **Cluster Analysis** seeks to classify individuals into internally homogeneous and externally heterogeneous groups, analyzing the differences between them. It is an example of an unsupervised classification technique since the groups are unknown a priori they are derived from data.

The last topic entitled **Discriminant Analysis** presents a supervised classification technique since the observations are already previously classified into two or more groups, looking for the reasons that explain that classification.

These approaches and objectives are directly aligned with the following Sustainable Development Goals (SDGs) of the United Nations 2030 Agenda:

Goal 7: Affordable and clean energy.

Goal 11: Sustainable cities and communities

Goal 12: Responsible consumption and production

Goal 13: Climate action

since the examples that are shown in class analyze databases that address these issues. Although it is true that all the training provided by this subject (theoretical and practical) contributes transversally to the 2030 AGENDA and SDGs, since their training enables the student to contribute to the analysis and management of the 245 SDG indicators. In short, the acquisition of the learning outcomes of the subject provides training and competence to contribute to some extent to the achievement of the SDGs.

1.2. Context and importance of this course in the degree

The course has a markedly instrumental nature and aims to train students in the handling of multivariate quantitative techniques commonly used in the implementation of exploratory analysis of multivariate data, which constitutes the first phase of any statistical-econometric study of an applied nature. The techniques studied are fundamentally of two types. The first group (principal components analysis and factor analysis) seek to carry out a data reduction process by identifying and calculating latent factors whose purpose is to express the information contained in a data set in terms of a smaller number of variables that represent the relationships between the analyzed variables. The second group (cluster analysis) seeks to classify objects/individuals into internally homogeneous and externally heterogeneous groups, analyzing the differences between them. All this results in a better understanding of the information contained in a data set that may be relevant when proposing a more formal analysis.

1.3. Recommendations to take this course

To have completed a course of introduction to both descriptive and inferential statistics and a course of Introduction to Econometrics

2. Learning goals

2.1. Competences

To pass the course, the student will be competent to use statistical tools to extract relevant information to develop and defend projects applied of economic character.

2.2. Learning goals

At the end of the course, the student should be able to:
- Perform an initial analysis of a multivariate data set
- Carry out a principal component analysis and interpret the results obtained
- Carry out a factorial analysis and interpret the results obtained
- Relate the factorial analysis and the principal components, highlighting the similarities and differences between them
- Carry out a data classification process using agglomerative hierarchical procedures
- Carry out a data classification process using partitioning procedures
- Design classification procedures that allow discriminating between groups of observations in the most efficient way possible, as well as analyzing the differences between them
- Validate and interpret the results obtained in a classification procedure

2.3. Importance of learning goals

Statistical techniques studied in the course will equip students with a set of powerful tools to make an initial exploratory analysis of the information contained in economic databases. In this way the student will learn to provide scientific rigor to solving economic problems by analyzing empirical data, which is an important step in implementing the scientific approach to problem solving.

3. Assessment (1st and 2nd call)

3.1. Assessment tasks (description of tasks, marking system and assessment criteria)

The student must demonstrate that it has achieved the intended learning outcomes by performing:

- 6 multiple choice questionnaires for each of the topics, except for topic 1 where there will be a questionnaire for the one-dimensional analysis and another for the two-dimensional analysis. These questionnaires will account for 40% of the mark if their average exceeds 3.

- 1 report corresponding to the resolution of exercises related both one-dimensional and two-dimensional exploratory analysis applied to a database (topics 0 and 1) that will account for 30% of the final mark if is higher than 3.

- 1 report consisting of the resolution of exercises that apply the multivariate techniques (topics 2, 3 and 4), that will account for 30% of the final mark if is higher than 3.

(40%) 6 questionnaires

(30%) Report on topics 0 and 1

(30%) Report on topics 2, 3 and 4

Students who do not pass this continuous evaluation system or those who opt for the global evaluation, will take a final evaluation exam that assesses the theoretical and practical knowledge of the subject through exercises applied to different databases. The answers to the questions will be made through scripts using the R program, and writing a report with the conclusions obtained.

Course assessment will be onsite. In the case of a new pandemic wave assessment will become partly online or fully online. It should be noted that in any online assessment task the student performance may be recorded, following the regulations described in: "
https://protecciondatos.unizar.es/sites/protecciondatos.unizar.es/files/users/lopd/gdocencia_reducida.pdf"_

4. Methodology, learning tasks, syllabus and resources

4.1. Methodological overview

The learning process is based on the combination of exposure, by the teacher, of the underlying theoretical basis to each of the techniques explained, with application to case studies conducted in the classroom using the R program. This will be done in a participatory environment in which both teacher and students discuss among themselves the interpretation of the results, which will increase the degree of applicability of the explained techniques.

All lectures and seminars will be imparted on site. In the case of a new health emergency caused by the current pandemic all teaching will be moved online.

4.2. Learning tasks

Theoretical and practical sessions (50%-50%)	30	100%
Work preparation and independent study	60	--------

4.3. Syllabus

Topic 0: Introduction to R **(4 hours)**
Topic 1: Initial Data Analysis **(8 hours)**
Topic 2: Factor Anaysis (Principal Component Analysis) **(7 hours)**
Topic 3: Cluster Analysis **(7 hours)**
Topic 4: Discriminant Analysis **(4 hours)**

The course will be taught in sessions of two hours in the computer room.

4.4. Course planning and calendar

Presentation: In the first session of the course, the objectives and contents of the subject are explained in detail, the teaching methodology used in the classes is presented and the evaluation criteria are clearly exposed. Likewise, an overview of multivariate statistical techniques will be given.

Theoretical-practical sessions: Throughout the course, 15 theoretical-practical sessions will be held in a computer room in which the theoretical bases of the statistical techniques will be explained and then these techniques will be illustrated by means of the analysis of real cases related to the world of economics. For this, the R program will be used as tool.

Autonomous work: To pass the course, the student must submit varied exercises for each of the topics.

4.5. Bibliography and recommended resources

The updated bibliography is incorporated through the Library Center and can be accessed by the web