**Data analysis for insurance**

Prof. Diego Zappa; Prof. Gabriele Cantaluppi

Module I: *Prof. Diego Zappa*; Module II: *Prof. Gabriele Cantaluppi*

***COURSE AIMS AND INTENDED LEARNING OUTCOMES***

We aim at introducing basic methods of data exploratory analysis using both Excel and R mostly on dataset with a special origin from the insurance market. We will first analyse datasets just using Excel but next we will move into the knowledge of the R environment to learn both the basic of this language and the capabilities offered to start basic methods of statistical learning and categorical data analysis with a special aim at generalized linear models. Most of lectures will be provided in the computer lab. To speed up the knowledge of R along with a recap of some fundamental results from mathematical statistics, some exercises will be based on the application of asymptotic results (e.g. Central Limit Theorem or Likelihood Ratio Test). Only if strictly necessary other ad hoc statistical software will be presented.

On successful completion of the course participants are expected to possess:

1. knowledge of concepts, terms and methods of the most used statistical learning techniques and grasp of their strengths and weaknesses (DD1- Knowledge and understanding);
2. ability to correctly apply statistical learning methods to real insurance, economics and management problems (DD2- Applying knowledge and understanding);
3. quantitative thinking addressed to make independent judgements, driven by application of statistical learning methods (DD3- Making judgements);
4. ability to present statistical learning arguments and the conclusions from them, by means of the extraction of qualitative information from quantitative data, with clarity and accuracy and in forms that are suitable for the audiences being addressed, both orally and in writing (DD4-Communication);
5. mastery of statistical learning methods, rigorous reasoning and data-driven decision-making, useful for quantitative analyses in other courses of the curriculum, as well as for analyses required in careers in insurance and in all fields involving management of data (DD5- Lifelong learning skills).

***COURSE CONTENT***

Module I: *R programming and data analysis.*

1. Some remarks from mathematical statistics useful for the master degree in Actuarial sciences.

2. Excel and the open source software R.

3. Data types. R objects: data frame and “friends”. Functions. Libraries. Data import. Graphics tools

4. Likelihood and inference: a recap using R

5. Linear and Nonlinear models. The Gauss-Newton algorithm. Examples with R.

7. From linear Gaussian models to Generalized Linear Models. The Exponential Dispersion Family. Inference.

8. GLMs for count, continuous and categorical responses.

9. Logistic regression for binary response data

Module II: *Statistical learning*

1. Introduction to Statistical Learning. Supervised and unsupervised learning. Assessing model accuracy.

2. Resampling methods. Cross-validation and bootstrap.

3. Variable selection in regression problems. Best subset selection and stepwise selection.

4. Shrinkage methods. The ridge regression and the lasso.

5. Cluster analysis. Hierarchical and non-hierarchical methods. The K-Nearest Neighbors algorithm.

6. Tree based methods. Regression trees and classification trees.

***READING LIST***

Lecturer’s notes / slides

I module

L. Fahrmeir-Th. Kneib-S. Lang-B. Marx, *Regression. Models, Methods and Applications,* Springer, New York, 2013.

A. Agresti, *An Introduction to Categorical Data Analysis,* John Wiley, New York, 2018.

R.A. Irizarry, *Introduction to Data Science,* Chapman and Hall/CRC, New York, 2019, https://rafalab.github.io/dsbook/.

II module

G. James-D. Witten-T. Hastie-R. Tibshirani, *An Introduction to Statistical Learning,* Springer, New York, 2017, https://www.statlearning.com/.

L. Fahrmeir-Th. Kneib-S. Lang-B. Marx, *Regression. Models, Methods and Applications,* Springer, New York, 2013.

(Details about which part of the books are suggested for reading will be given at the beginning of each module)

***TEACHING METHOD***

Lectures with examples from actuarial, business, economic, financial, health case studies.

***ASSESSMENT METHOD AND CRITERIA***

Module I: for those attending regularly the lectures, the exam will consist of a mix of problems and a case study to be solved with R. For those that cannot attend regularly the lectures, written exam with questions on the methods presented during the course (Time: 45 minutes).

Module II: A written exam consisting of two open-ended questions on the statistical learning subjects. An optional question concerning the interpretation of an R code can also be answered. Time: 45 minutes.

The final mark will be the weighted average of the marks in the 2 modules: I module=55%, II module=45%. If the final score does not correspont to an integer, the mark will be rounded up to the nearest larger integer.

Aim of the exam is to assess reasoning analytic abilities on the course subjects. Language properties and communication abilities are also assessed.

***NOTES AND PREREQUISITES***

Students enrolling in this course should have a basic understanding of mathematical and statistical techniques at the level of the bachelor degree (undergraduate programme) in economic studies (see for an example the profile “Quantitative methods for finance and insurance” taught at the faculty of “Banking, Finance and Insurance Sciences”).