Data Science for Medicine

Nikolaos Demiris, Assistant Professor, AUEB, nikos@aueb.gr

Overview

Data Science is changing the way that treatments are decided and administered in medical science. This course will introduce the basic concepts of Biostatistics and Epidemiology and will illustrate how big data are used within survival analysis and infectious disease Epidemiology. The methods will be applied using R and assignments will give students experience with analyses and reporting through real data examples.

Key Outcomes

By completing the course, the students will be able to:

- understand the basic concepts of Survival Analysis and Epidemiology of communicable and non-communicable Diseases
- learn how big data is used in medical practice, including web queries informing infectious disease prediction and genetic data utilised for disease classification and prediction.
- Analyse medical data using R
- deliver appropriate presentations of the various analyses

Requirements and Prerequisites

This course combines theory and practice, including multiple individual exercises and culminating in a hands-on group project.

The course does not assume any prior experience in R or other software. Basic knowledge of programming will be an advantage.

Required Course Materials

There is no required textbook. All course materials will be provided in class and available for downloading.

Books

There are many books on the subject but no single book covers the entire syllabus; the following selection contain in detail the aspects of the theory discussed in class:

- David W. Hosmer, Jr., Stanley Lemeshow, Susanne May, 2008 Applied Survival Analysis: Regression Modeling of Time to Event Data, 2nd Edition. Wiley Series in Probability and Statistics
- Diekmann O., Heesterbeek, J.A.P. and Britton, T. (2013). Mathematical tools for understanding infectious disease dynamics. Princeton UP.
- Kenneth J. Rothman, Sander Greenland, Timothy L. Lash, 2012 Modern Epidemiology Third Edition, Lippincott Williams & Wilkins

Software/Computing requirements

- R special packages will be used like glmnet, ahaz and surveillance. An up to date version of R is required.

Grading

Students will be graded as follows:

- Participation - 10%

- Mini assignments - 25% they refer to small exercises after each lecture that have to be returned at the next one and they mainly refer to points related to the comprehension of the lecture. They can be simple exercises, small essays, work with data, search for cases etc

- Group Project - 40% They will one projects working as groups (randomly selected). They have a 2-week duration and you have to submit a detailed report

- Final presentation – 25% It refers to the presentation for the group projects above. Here the ifocus is upon the oral communication of the methodology and the findings as well as the overall message in a fun and scientifically accurate way

The course does not have written exams.

Participation

In-class contribution is a significant part of your grade and an important part of our shared learning experience. Your active participation helps us to evaluate your overall performance. You can excel in this area if you come to class on time and contribute to the course by:

- Providing strong evidence of having thought through the material.
- Advancing the discussion by contributing insightful comments and questions.

Turn off all electronic devices prior to the start of class. Cell phones tablets and other electronic devices are a distraction to everyone. In lectures you need to use laptop you will be informed to do so.

Assignments

Late assignments will not be accepted.

Attendance Requirements

Class attendance is essential to succeed in this course and is part of your grade. An excused absence can only be granted in cases of serious illness or grave family emergencies and must be documented. Job interviews and incompatible travel plans are considered unexcused absences. Where possible, please notify the instructor in advance of an excused absence.

Students are responsible for keeping up with the course material, including lectures, from the first day of this class, forward. It is the student's obligation to bring oneself up to date on any missed coursework.

Code of Ethics

Students may not work together on individual graded assignments unless the instructor gives express permission.

Exercise integrity in all aspects of one's academic work including, but not limited to, the preparation and completion of all other course requirements by not engaging in any method or means that provides an unfair advantage. In any case of doubt, students must be able to prove that they are the sole authors of their work by demonstrating their knowledge to the instructor.

Clearly acknowledge the work and efforts of others when submitting written work as one's own. Ideas, data, direct quotations (which should be designated with quotation marks), paraphrasing, creative expression, or any other incorporation of the work of others should be fully referenced. No plagiarism of any sort will be tolerated. This includes any material found on the internet. Reuse of material found in question and answer forums, code repositories, other lecture sites, etc., is unacceptable. You may use online material to deepen your understanding of a concept, not for finding answers.

Please report observed violations of this policy. Any violations will incur a fail grade at the course and reporting to the senate for further disciplinary action.

Course Syllabus

The course comprises **six** units of three hours each.

Unit 1: Introduction to Survival Analysis

The survival and hazard functions. Time-to-event data.

The likelihood function.

Inference for parametric models using R.

Assignment (individual): ***exercise.***

Unit 2: Advanced Survival Analysis

Proportional and additive hazard models.

Fitting using R.

Checking model adequacy: martingale and Shoenfeld residuals.

Applications in medicine and reliability theory

Assignment (individual): ***exercise.***

Unit 3: Big Data in Survival Analysis

The LASSO and its use in the Cox model. The Poisson interpretation of the semi-parametric approach.

Examples using the glmnet R package

Applications in personalized medicine

Assignment (individual): ***exercise.***

Unit 4: Infectious Disease Modelling

Challenges and pitfalls in Modelling Infectious Diseases, including Google Flu Trends

The main epidemic Models: Deterministic and Stochastic

Disease Control

Analysis of real data using the Surveillance R package

Assignment (individual): ***exercise.***

Unit 5: Epidemiology

Analysis of contingency tables

Risk assessment in retrospective and prospective studies

Deconfounding

Project Development: Question time regarding each project.

Unit 6: Final presentation of projects

Each team presents their finished project to peers and instructors. The final project will involve:

- Perform exploratory data analyses from a real-world dataset using R
- choose the model(s) that best support the proposed project, taking into account model fit and parsimony considerations
- design a presentation narrating the medical insight gained from the project

Written documentation due.