Advanced Statistical Methods

Course's introduction

Leonardo Egidi, Francesco Pauli, and Matide Trevisani

March 2024

Università di Trieste

General information

Instructors

- Leonardo Egidi (24 h)
- Francesco Pauli (12 h)
- Matilde Trevisani (12 h)

Lecture schedule

- The lectures will be given:
    - monday (10.00-12.00), room 4b (H2bis)
    - tuesday (9.00-11.00), room 0b (H3)
    - friday (9.00-11.00), room 5b (H2bis)

Lecture days

There won't be lectures in the following days:

- 29 march
- 1,2 april
- 15 april

Any further variation will be communicated in advance. The course is given exclusively in presence.

Intructors’ appointments slots

- Leonardo Egidi (at DEAMS, room 2.19)
  - tuesday: 15.00 - 17.00
  - thursday: 11.00 - 13.00
- Francesco Pauli (at DEAMS, room 2.14)
  - check here: https://www.units.it/persone/index.php/from/abook/persona/8755
- Matilde Trevisani (on meeting, to be fixed by email)

The offices are located at the second floor, building D (DEAMS).

- It is possible to fix an appointment by email.
- Email:
  - legidi@units.it
  - francesco.pauli@deams.units.it
  - matilde.trevisani@deams.units.it

Course aims

- **Knowledge and understanding**: students will have to show that they have understood the essential ideas that motivate the use of advanced statistical techniques and the functionalities that limit their use.
- **Applied knowledge and understanding**: the student will have to show that he knows how to use the techniques learned for the analysis of real data, even using appropriate software tools. Specifically, the student is required to have a very good use of the Stan software and of the 'rstan' library available in R.
- **Making judgements**: the student must be able to navigate in the context of the analysis of real data, with always priority and vigilant attention to the sampling scheme of the data, to their possible hierarchical/multilevel structure and to their granularity.
- **Communication skills**: the student will be able to effectively communicate the results of data analysis using appropriate tools (including modern techniques for compiling dynamic documents, such as RMarkdown). In addition, the student is also required to make a 'visualization' effort, suitable for the production of graphical tools that summarize complex trends (above all, for example, the use of R libraries such as 'ggplot2').
- **Ability to learn**: at the end of the course the student will be able to consult theoretical and applied scientific works that use advanced statistical techniques, critically analyze the application of the models and algorithms explained in class, and illustrate case studies through the use of probabilistic scientific programming.

Texbooks

- Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2006.
- Gelman, Andrew, Jennifer Hill, and Aki Vehtari. Regression and other stories. Cambridge University Press, 2020.
- Wood, Simon N. Generalized additive models: an introduction with R. CRC press, 2017.
- Ruppert D., Wand M.P., Carroll R.J. Semiparametric regression. Cambridge University Press, 2010
- Blei D.M. et al., Latent dirichlet allocation, J. Mach. Learn. Res. (2003).
- Griffiths T.L. et al., Finding scientific topics, Proc. Natl Acad. Sci. (2004)
- Handbook of Mixed Membership Models and Their Applications, Edited By Edoardo M. Airoldi, David Blei, Elena A. Erosheva, Stephen E. Fienberg (2015)
- Notes from lectures

Course structure

Prerequisites

- Basic knowledge of statistics (equivalent to two courses in a three-year bachelor, or in any case equivalent to having taken and passed the Statistical Methods exam).
- Ability to program and use the R software.

The course will make use of teaching tools available on the moodle2, MS/Teams and wooclap platforms. In addition, all students are expected to use R software, so they must own or have access to a computer.

Syllabus

1. Hierarchical/multilevel statistical models, with use of the Stan software (in this case the R library ‘rstan’).

a) Definition of multilevel data and general structure.
b) Linear and generalized linear
multilevel models with variable slope
and intercepts, estimated with Frequentist and Bayesian methods.
c) Extensions of canonical models: models for grouped data; non-nested models; models for repeated measurements.

2. Causal inference

a) Randomized experiments.
b) Treatment interactions and post-stratification.
c) Observational studies
d) Sub-classification: effects and estimates for different sub-populations.
e) Use of instrumental variables to estimate the causal effect.

Syllabus (continues)

3. Semi-parametric/non-parametric regression

a) Introduction to local regression methods
b) Spline functions
c) Penalized likelihood: classical estimation and Bayesian estimation d) Splines and hierarchical models.

4. Mixed-membership models

a) Understanding multiple membership data structures
b) Examples in multilevel models and in other types of analysis (Text, Social network, Survey, Population genetics, Ecology, Marketing and Clustering analyses)
c) Latent Dirichlet Allocation model and extensions
d) Bayesian and Variational inference
e) Lab with R pkgs (lda, stm, ldatuning, ldavis, rlda)

Exam information

- **Oral** final exam consisting of both theoretical and practical questions. The student will be asked to do a preliminary work assignment, send it to the professors at most two days before the exam takes place, and then discuss the results at the exam's day.