Statistical modelling

Syllabus

Dr. Léo Belzile
HEC Montréal

Organization

- Weekly meetings in room Hélène-Desmarais (Wednesday, 15:30-18:30).
- Two instructors! Each teaching one half of the course
  - Léo Belzile (CSC 4.850, `leo.belzile@hec.ca`)
  - Juliana Schulz (CSC 4.809, `juliana.schulz`)

Course content

- All of the course material can be downloaded from the course website: `https://lbelzile.github.io/statmod/` and from ZoneCours
  - course notes
  - video recordings
  - **R** package with datasets
  - exercises and solutions
  - **R** demos

Communication policies

- Submit your assignments/projects via *ZoneCours*
- Ask course-related questions on **Teams** Class Channel: `O365-MATH 60604A - A2024`
  - enter the code `r2e6mjw` to join.
- For other matters, *email* both of us: `leo.belzile@hec.ca` and `juliana.schulz@hec.ca`

Course content

> *All models are wrong, but some are useful.*
>
> — Georges Box

1. Basic principles in inference and statistical modelling
2. Linear models
3. Generalized linear models
4. Models for longitudinal data and correlated data
5. Linear mixed models
6. Introduction to survival analysis

Evaluations

- Three part-project worth 30% (teamwork) using Montreal Bixi data with due dates
  - October 11th,
  - November 8th,
  - December 6th.
- Midterm exam (30%) on Tuesday, October 29th 2024, 12:00-15:00.
- Final exam (40%) on Thursday, December 12th, from 13:30-16:30.

One sided letter paper for crib-sheet is allowed for midterm, two-sided for the final.

What is the format of the course?

At home:

- reading course notes
- watch videos (when applicable)
- exercises (with solutions)

In class:

- lectures with weekly summary
- **R** demonstrations
- question period

What is the workload for this course?

- 3 credits = 135 hours of work
- an average of 9 hours per week
- do not underestimate the initial time investment:
  - installing required software
  - learning programming basics
  - getting up to speed with prerequisites

What is the target audience?

Students enrolled in the Data Science and Business Analytics M.Sc program.

Students admitted normally have a bachelor in

- engineering
- physics or
- mathematics.

Basic knowledge of calculus and linear algebra is assumed.

What are the prerequisites?

A first course in probability/statistic covering the following notions:

- probability axioms and combinatorics
- random variables
  - moments (expectation, variance, correlation)
  - discrete distributions: Bernoulli, binomial, Poisson
  - continuous distributions: uniform, exponential, normal
- descriptive statistics
- hypothesis tests
- comparison of means and proportions (one and two samples)
- simple linear regression

What software will we use in class?

I am avid support of open-access software and of **R**, a programming language written by the community

- its free!
- multi-platform support
- download from cran.r-project.org
  - I recommand the free RStudio Desktop IDE by Posit.

Will there be programming?

Yes. We will cover the basics of **R** to fit models and visualized data.

- I will provide code only for exercises and slides.

You must provide your code for the group project

- I should be able to reproduce *exactly* your analyses.
- submit as a `.txt` file (otherwise, you won't be able to submit your work on *Zonecours*)
- use UTF8 encoding
- follow the instructions for naming scripts/files ($\neq$ `mycode.txt`)

Learning R

The CAMS offers free tutorials (mandatory registration, limited space).

- An introduction to R: Part 1 on Thursday, August 29th 2024, from 15:30-17:30
- An introduction to R: Part 2 on Thursday, September 5th 2024, from 15:30-17:30

The objective of this course is not to become expert **R** programmers, rather to use R to carry out data analyses.

- Additional resources are listed on the course website

What are the professors expectating of you?

- Active participation in class: students are expected to be in class
  - ask questions! there is no silly question
- Autonomy: you are sole responsible for your learning.
  - stay up to date and do your readings
  - don't stay in the dark: ask questions (to instructors or your peers)!
- Feedback: problems or unclear explanations? let us know asap

Inclusive and respectful environment

Harassment, discriminatory views, etc. are not tolerated.

Let us know if

- We can do something to improve the course experience for you or other students
- a statement or attitude makes you uncomfortable
- your name/prefered pronoun/gender differs from the information provided on *HEC en ligne*
- your performance is affected by external factors: we will do our best to help you or direct you to external resources.

Plagiarism

Please don't. There are consequences and its an insult to your intelligence. Ask for help if necessary!

- if you take and adapt code from elsewhere (e.g., `StackOverflow`), cite your sources!
- you must program yourself your code for individual assignments (discussion with peers is okay, but code sharing, copy paste or similar wording is punishable)

Generative AI

Midterm and final are closed-book exam (no computer).

Projects: students can use generative artificial intelligence (AI) or large language models (LLM) to improve and edit the deliverable (either code or text), but only as a proof-reading tool. That is, sudents must create the original draft of all components of the assignment (including all statistical analyses, code, and accompanying text).

My policy reflects that of Andrew Heiss

Using genAI and large language models (LLM)

Any use of generative AI or LLM must be cited adequately and students should provide their work with

1. details about the tool used (name and version),
2. the list of prompts,
3. a copy of the original draft or code,
4. a description of any modification to the proposed output,
5. a short reflection on the usage of generative AI or LLM.