STAT 301 - Statistical Modelling for Data Science

When and where?

- The lectures will be on Tuesdays and Thursdays from 12:30 to 14:00
- The lectures will be held in Henry Angus building, Room 234
- Office hours will be held ONLINE on Wednesdays from 15:30 to 16:00

Instructor: Gabriela Cohen Freue and Rodolfo Lourenzutti

Webpage: https://canvas.ubc.ca/courses/89390/modules/items/3539310 Check regularly for updates on course policy and content.

Calendar Description

Explanatory and predictive data analysis with multiple explanatory variables. Choosing the right methods to apply based on the statistical question and data at hand. Trade-offs between model-based and non-model based approaches. Emphasis placed on case studies and real data sets, as well as reproducible and transparent workflows when writing computer scripts for analysis and reports.

Prerequisites: *note that these are hard pre-requisites.*

- STAT 201: Statistical Inference for Data Science
- One of MATH 100, MATH 102, MATH 104, MATH 110, MATH 120, MATH 180, MATH 184, SCIE 001
- Access to a computer. If a student does not have their own laptop or chromebook, students may be able to loan a laptop from the UBC library (https://services.library.ubc.ca/computers-technology/technology-borrowing/)

Learning Outcomes

By the end of the course, students are expected to be able to:

- Describe real-world examples of explanatory modelling (e.g. A/B testing optimization & regression with variable selection) and predictive modelling problems.

- Explain the trade-offs between model-based and non-model based approaches, and describe situations where each might be the preferred approach.

- Explain the difference between creating models for explanation vs prediction, in the context of both how you choose and evaluate models as well as how you interpret the results.

- Choose & apply a suitable method (e.g., regression, GLM's, sample size estimation, controlling for multiple testing, peeking, bandit algorithms, variable selection, model diagnostics) based on the statistical question and data at hand. Discuss the advantages and disadvantages of different methods that may be suitable for a given problem.

- Correctly interpret computer output when performing the statistical analyses presented in this course, in the context of the statistical question being asked and the audience being reported to.

- Identify the assumptions / conditions required for each method to produce reliable results. Choose techniques to check (or at least be able to falsify) those assumptions. Discuss the consequence(s) of mapping the wrong methods to the question and/or data type.

Software Platforms

- Students will learn to perform their analysis using the R programming language (https://cran.r-project.org/)
- Worksheets and tutorial problem sets as well as the final project analysis, development, and reports will be done using Jupyter Notebooks (http://jupyter.org/)
- Students will access the worksheets and tutorials in Jupyter Notebooks through Canvas

Recommended textbooks

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Available ONLINE at UBC Library: http://resolve.library.ubc.ca/cgi-bin/catsearch?bid=6667014
- Rafael Irizarry. *Introduction to Data Science*. Available ONLINE at https://rafalab.github.io/dsbook/
- Ismay, C. and Kim, A. Y. (2021) *ModernDive: Statistical Inference via Data Science.* Available ONLINE at https://moderndive.com