Computational statistics:
Data Science for Social Science

Marina Khismatullina
Institute for Finance and Statistics, University of Bonn
marina.k@uni-bonn.de

About the course

The goal of this course is **to introduce basic methods from data science and machine learning** and **to apply them to problems in empirical social science**.
This is a very applied course which consists of (a) revisiting a particular method, (b) implementing it, and (c) interpreting the results afterwards. However, you will need an understanding of the theory behind all of these methods.

The course will be divided into three blocks:

1. Introduction to R
2. Basic Concepts of Data Science:
   (a) Classification
   (b) Validation
   (c) Structural estimation vs. prediction
   (d) Resampling methods
3. Methods in Data Science:
   (a) Regularization: principal components, ridge regression, lasso
   (b) Regression trees, random forests, bagging, boosting stacked methods/ensemble methods
   (c) Causal forests

I will interweave applications from social science, as well as common pitfalls when implementing these methods and particular problems that occur in social science applications.

Schedule

- All lectures and presentations will be conducted **live** via Zoom. Please turn your video on during the sessions. We will be meeting on Monday (from 10.15 am to 11.45 am) and on Wednesday (from 8.30 am to 10.00 am). Days, when the lectures will not be taking place, will be announced separately. Alternatively, you can find the dates of the holidays for the current semester on the website of Prüfungsamt.

- The link for the Zoom meetings is `https://uni-bonn.zoom.us/j/97950407912?pwd=d1lpQ2R5b3M5T29KaU42SENzeUJndz09`. Alternatively, you can log into Zoom using the **Meeting ID:** 979 5040 7912 and **password:** 998048.

- All materials will be available online on the GitHub repo `https://github.com/marina-khi/CompStat_2021` (hopefully very soon after the lectures).

- If you cannot be present during the lecture, please tell me about that.

- Starting from May, we will have weekly office hours for which **you need to sign in advance!** The office hours are currently planned for Wednesdays, from 11.00 to 12.30. Each of you who signed up for a meeting will get a 10 minute slot to discuss the issue, so please **prepare your questions**. You can also send me your question (or a piece of code) beforehand, ideally, no later than Tuesday afternoon. You will find the slots for the next meeting at (link tba). The slots are distributed on a first come, first served basis. You can use the same Zoom link for the office hours as for the lectures.

- Wednesdays are reserved for the problem sets. Once we finish the R introduction and revision of some econometric concepts, **everyone** is supposed to present their own solutions to the problem sets at some point. You will do it by sharing your screen and we will discuss the results of the solution together in class. You will also be expected to participate in the discussion even if you are not presenting. This means two things:

    1. Be prepared.
    2. Before the meeting, think of one or two questions you would like to ask the presenters or that you had while working on the problem set. Be prepared to be called to ask those questions at the end of the presentation.

Literature and languages

- Programming: we will use `R` for all implementations. You will need a distribution of `R` (can be downloaded at `https://cran.r-project.org`) and you may also download an editor for `R` (I recommend `Rstudio`: `https://rstudio.com/products/rstudio/`).
- The main reference will be James et al. (2013), which is available for free as a PDF. I will subsequently add more literature as we discuss specific topics. Further reading is Friedman et al. (2001)

Grade

You will need to register for the course with the Prüfungsamt until April 26th.

- Instead of an exam, there will be a final project (Hausarbeit). You will need to pick a topic for your project as soon as you decide to stay in the course. We will have a final discussion of the chosen topics in the beginning of July, where everyone will introduce their topics and we will discuss the issues that may arise. This project will contribute 90 % to your final grade.
- You will also be asked to present (parts) of a problem set, possibly in groups. Once everyone has registered for the course, you will be assigned a problem set to present. The problem sets will be directly connected with implementation of different methods in R and interpretation of the results. This presentation will contribute 10 % to your final grade.

Final project

The deadline for the **individual** project is August 15th, as the deadline has to be least 4 and at most 6 weeks after the topics were approved/assigned (this is the requirement of the examination office). This means that the topics will be finally approved (and discussed) in the beginning of July, when we will have a "mini-workshop" with you introducing your topics. The project will involve a simulation study with a realistic empirical set-up, and (ideally) an empirical application of one of the methods discussed in class in a social science setting.

Bibliography

Friedman, J., T. Hastie, and R. Tibshirani (2001): *The elements of statistical learning*, vol. 1, Springer series in statistics New York.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An introduction to statistical learning*, vol. 112, Springer.