Computational Statistics:
Data Science for Social Science

Lena Janys
Assistant Professor Institute for Finance and Statistics
ljanys@uni-bonn.de

July 12, 2022

Final project

The project should include

1. Description of the method
2. A simulation study using a realistic set-up from a research paper in economics as a motivation and benchmarking.
3. (An empirical application)

The discussion should focus on the theoretical properties of the method, the properties of the data most often encountered in empirical settings and the simulation study should reflect these two components. You may collaborate with other students who are working within the same general method, i.e. lasso, but everyone has to hand in a unique project.

You may add an introduction and/or a conclusion, but the take-away should be clear from the discussion of the results.

Formatting and Uploading

- The compiled PDF should be about 10 pages (from a Jupyter Notebook or an RMD file). Hand in both files.
- Name your files: `your_name.rmd`, `your_name.pdf` etc.
- Upload your .pdf and code file to the folder `Project Upload` on E-campus.

General Remarks

The final project should center one of the methods we discussed in class, specifically:

1. **Regularization**
   (a) Principal Components
   (b) Ridge regression
   (c) Lasso
2. **Tree based methods**
   (a) Regression Trees
   (b) Random forests

(c) Bagging
(d) Boosting
(e) Stacked methods/ensemble methods
(f) Causal forests

3. **Neural networks**

Some journals that you can consult for inspiration on empirical work:

- AEJ:Applied
- Journal of Labor Economics
- Journal of Health Economics
- Journal of Development Economics
- Journal of Public Economics
- Journal of Applied Econometrics
- ... AER, ReStud, Quantitative Economics

And a nice paper as an inspiration for using data science methods for economics applications: https://www.aeaweb.org/articles?id=10.1257/aer.p20171040

Your Homework

1. Look through the journals above
2. Pick **three** papers that interest you and that have either a prediction component (e.g. instrumental variables) or something else where you think the methods of this course might be useful (classification exercise etc.)
3. Write in a table: which variables are used, are the data available and if not: which descriptive statistics are reported etc.
4. How could you simulate the dgp?