Topics in Advanced Statistics
Intermediate group assignment

Andreas Alfons
Erasmus School of Economics, Erasmus Universiteit Rotterdam

For this assignment you will work in **groups of 4 students**. It is ok to work in a group of 3 (for example, to simplify planning or logistics if you cannot be at university too often). However, note that a group of 3 will of course increase the individual workload because the total work is split among fewer people. It is not allowed to do this assignment individually or in a group of 2, though, as this goes against the spirit of working in a group.

It is up to you to find your partners. You can sign up for a group on Canvas.

Perform the tasks below and prepare a report. Do not answer to questions individually point-by-point, but instead tell the whole story in a properly structured report that contains

- an introduction to the problem,
- a brief description of of the methodology with a focus on the aspects or properties that are relevant for answering the questions below,
- a discussion of your simulations, including a complete and clear description the data generating process,
- a discussion of any other results

(not necessarily in that order). Furthermore, motivate any choices that you make in your analyses. **Be concise** and use as few pages as necessary, your report may be **no longer than 8–10 pages** (excluding references). Please submit your report in pdf format **via Canvas**, together with the R file containing your implementation. The deadline for submission is Tuesday, **February 11, 2019, 11:59** (noon).

Assignment

1. Implement the deterministic MCD algorithm (including the reweighting step). In addition, construct and implement a robust linear regression estimator based on the deterministic MCD covariance matrix, which in the following will be referred to as the plug-in robust regression estimator.

   On Canvas, you can find an R file containing a code skeleton and some explanations. Please use this file, **do not change any function names and make sure that you use the correct input and output as specified for each function**.[1] Submit the R file with your implementations together with your report.

[1] This will simplify grading because each groups's code will be similarly structured. **Not using the specified input or output format could lower your grade if your code cannot be tested properly**.

2. How is this plug-in robust regression estimator related to the least trimmed squares (LTS) estimator of regression? Discuss what its advantages and disadvantages are (in terms of statistical properties).

3. Consider the `Eredivisie28` data to explain the market value of football players by their age. Are there any clear outliers? Why is this relevant when computing the empirical influence function (EIF)? Compute the EIF for the intercept and the slope coefficient of the plug-in robust regression estimator, the LTS estimator[2], and the OLS estimator. Include plots in your report. What are your findings?

4. Study the behavior of the estimators by means of a simulation study. Also include the LTS estimator and the OLS estimator in the simulation study for a more complete comparison.

   How do you propose to evaluate coefficient estimates and prediction performance in such a simulation study? How well do the robust methods perform if there are no deviations from the model distribution? How do outliers affect the estimators? How do the robust methods perform with respect to outlier detection?

   Make suitable choices regarding the numbers of observations/variables, outlier configurations and contamination levels. Make sure that those choices in the simulation design illustrate any potential advantages/disadvantages of the different methods. Include only a selection of interesting results in the main text of your report.

`Eredivisie28` data

The Eredivisie, the highest football league in the Netherlands, is known for producing young, talented football players. This data set from the 2013/14 season contains the market value (in Euros) and age (in years) of all 351 players of age 28 years or younger.

[2] Use function `ltsReg()` in package `robustbase`.