Topics in Advanced Statistics
Final individual assignment

Andreas Alfons
Erasmus School of Economics, Erasmus Universiteit Rotterdam

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

(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 6 pages** (excluding references). Please submit your report in pdf format **via Canvas**, together with the R file containing your implementation and the R file with your simulations. The deadline for submission is Sunday, **March 1, 2019, 23:59**.

Assignment

In this assignment, you will study missing data and robustness against cellwise outliers in a regression setting. The problem of missing data will be addressed with multiple imputation and the bootstrap, while the DetectDeviatingCells algorithm will be used to filter cellwise outliers.

1. Implement the procedures for obtaining point estimates and valid standard errors for MM-regression via multiple imputation and the bootstrap.[1] Include the possibility to run the DetectDeviatingCells (DDC) algorithm before imputation.[2] 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**.[3] Submit the R file with your implementations together with your report.

[1] You can use functions such as `irmi()` and `kNN()` from package `VIM` for the individual imputations, but you need to implement the multiple imputation and bootstrap algorithms yourself. That is, you cannot use functionality from packages `mi`, `mice`, `boot`, etc.

[2] You can use function `DDC()` from package `cellWise` for this step.

[3] 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. Design a small simulation study to investigate the behavior of the following methods under missing data and cellwise contamination: (i) MM-regression with multiple imputation, (ii) MM-regression after DDC with multiple imputation, (iii) MM-regression with the bootstrap, (iv) MM-regression after DDC with the bootstrap. Are the point estimates accurate after imputation? Are the standard errors accurately estimated (and how do you suggest to measure this)? How do cellwise outliers affect the point estimates and standard errors?

   Make suitable choices regarding the missing data generating mechanisms and cellwise outlier generating processess.[4] Ensure that those choices in the simulation design illustrate any potential advantages/disadvantages of the different methods.

   Submit the R file with your simulation code together with your report.

[4] You are **not** allowed to use ready-made functions such as `ampute()` from package `mice` to generate missing values. You need to show that you understand the missing data mechanisms by programming them yourself. The same holds for cellwise outlier processes.