Datamining & Statistics

Assignment 1: Regression

The assignment for Regression is relatively straight-forward:

Find a data set of your own choice (different from the ones used in class), perform a multivariate regression analysis on it, including test/validation, and report the results.

Use the ***DMS - S4 - Assignment Template*** notebook that is available on Moodle to complete this assignment.

To get a 6:

Complete chapters 1, 2 and 3 in the DMS - S4 - Assignment Template. It should be possible for a student next year if given your notebook, that he/she can follow and understand what you've done. This includes therefor a brief explanation of a term, argumentation of some choices you make, how well the model performs (using some of the metrics), and what the results mean.

The file should be able to run from start to finish. Note that you have to use a dataset different from the ones used in class.

To get a 7, 8 or 9 add one or more of the following:

> **+1**
> Complete chapter 4. This is about adding a section on choosing the independent variables to include/exclude from your equation. Perhaps based on the VIF's but perhaps another technique.
>
> **+1**
> Complete chapter 5. This is about adding another method to split the data using a cross-validation method.
>
> **+1**
> Complete chapter 6. This is about adding a section where you apply at least one of the mentioned regression techniques at the end of the lecture. This would also require some more explanation on how this technique works.

To get a 10:

Get to a 9 but add something substantial, related to regression that would surprise me ☺. A 10 is given to those who have done something beyond expectation.

To submit:

- Jupyter notebook file
- Dataset file(s) (or link to this in case it's too large)

Method to submit:

- One zip or rar file containing the files. Include your student number in the name of the zip/rar file.
- Upload on Moodle.

FAQ's

Q1: Can we work together?
A1: No and yes. No, it is an individual assignment, yes you can of course always help each other. But no direct copy-pasting. Copy-pasting, or copy-pasting and changing just one or two words, either from fellow students, or the internet, books, etc. is considered fraud. This will be reported to the examination board.

But…: But some lines of code will be standard, and only a few parameters or variable names will be different…..
A-but: The copy-pasting is more on the text surrounding the code, and indeed some parts of the code will be copy-pasted. This is acceptable and probably normal in IT that you copy some code from the internet and make a few adjustments to make it work for your data.

Q2: Can we use the same dataset?
A2: Yes, as long as it isn't one of the two datasets used during the session (the Soccer and UFC dataset). If you do plan on using the same dataset as perhaps your project group, I'd recommend not to share your code with the others to prevent the suspicious of plagiarism.

Q3: I'm not clear about the '+1', can you explain?
A3: Sure. There are some criteria to get a 6. To get to a 7 add one of the three things described at a '+1', your choice which one. To get to an 8 add two, and to get to a 9 add all three.