PROBLEM SET 1

SUBMIT BY 9 OCTOBER 2023

INSTRUCTIONS

Three problem sets count for 40% of the module assessment, and the exam counts for the other 60%. The lowest of the three sets will be ignored.

Each homework should be submitted on Canvas as an archive file (tar or zip) consisting of

(1) your python code (a py or a ipynb file),
(2) a pdf file of your written responses (can be embedded into your ipynb file instead),
(3) any data file used as input for your python program. You do not need to include the supplementary files I provide (like csv files).

The py file (or ipynb) should be machine readable, and it must reproduce your answers when I run it. Develop your own Python code rather than simply using existing Python modules for computing the relevant mathematics (e.g. `statsmodels`). If in doubt, send me an email. The pdf document should contain your answers to the questions in which you provide a description of the mathematical methods used. There should be no Python code, but you may refer to specific Python functions if you need.

The homework will be graded according to a scheme in which *content* (i.e. correctness of your answers, choice of methods, python code) is weighted at 80% and *presentation* (i.e. manner in which you present your answers, methods, and code) is weighted at 20%.

Please submit on Canvas by **09.10.2023** as a single archive file (tar or zip).

PROBLEMS

Problem 1. Write a Python function called `csv_to_linreg` that takes as input a string to a csv file and outputs a `tuple` where the first entry is a list (or tuple, numpy array, etc.) of the coefficients for the (affine) hyperplane of best fit and the second entry of the tuple is its r^2 value. Assume the last column is corresponds to the (unique) dependent variable.

Run your function on the following supplementary csv files:

- `small_sample.csv`,
- `medium_sample.csv`,
- `large_sample.csv`,

and describe the meaning of the output as well as the r^2 value.
(Hint: Consider writing two functions `csv_to_linreg` and `dataframe_to_linreg`—could be useful later.)

Problem 2. Write a Python function called `dataframe_to_plot` that takes as input a pandas `dataframe` (you may assume only two columns: independent and dependent variable resp.) and outputs a matplotlib figure plotting the following items:

(1) the data set,
(2) the line of best fit,
(3) the parabola of best fit, and
(4) the cubic of best fit.

The data set and each of the three curves should all be different colours. The plot should also display the r^2 values of the three curves; you can do this with the legend, but you might have a different solution.
(Hint: Save time and space by using the function `dataframe_to_linreg` from the hint in Problem 1.)

Problem 3. Run your function from Problem 2 on the following csv files:

- `small_sample.csv`,
- `four_pts.csv`,
- `logging_logs.csv`.

What, if anything, can you conclude from seeing the output and why?

Date: October 18, 2023.