PROBLEM SET 1

SUBMIT BY 11 OCTOBER 2024

INSTRUCTIONS

Each homework should be submitted on Canvas by **5:00pm on 11 October 2024** as an `ipynb` file (i.e. a Jupyter notebook).

The 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 ipynb file should contain your answers to the questions raised below in markdown (or "text") cells and not Python cells.

The use of generative AI is prohibited and will be filed, in line with the university's policy, as academic misconduct. Read more here:

https://www.universityofgalway.ie/academicintegrity/.

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%.

Late submissions will be accepted at a 10% deduction every 24 hours past the deadline, starting with 10%. For example, submitting two hours after the deadline on Canvas will incur a 10% deduction, whereas 26 hours after the deadline will incur 20%.

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.)

Date: September 27, 2024.