Advanced Data Analysis, third assignment, 2020-10-31

The file `children.txt` contains a matrix with two columns. The first column is the age of each child in months, and the second the weight in Kg. The data is from the National Health and Nutrition Examination Survey of 2017-2018 and represents a sample of children up to 24 months old.

For answering the following questions, you can use or adapt any code from the lectures or exercises if you want.

When required, you must justify your answer. Please provide brief, clear and direct justifications. Invoking irrelevant factors, giving vague justifications or writing more than necessary will be penalized.

Read all questions before starting to implement your code, since the first question depends on a proper understanding of the assignment as a whole.

Question 1 [4 points out of 20]

In the following questions you will be asked to fit a polynomial curve to the data after finding the best degree for the polynomial regression, and then to estimate the expected error your curve will have when predicting the weight of a child based on their age in months.

In this part, start by organizing your data in the different sets you will need and plot the different sets. Explain how you divided your data and justify this division.

Question 2 [6 points out of 20]

Trying polynomial curves of degrees 1 through 12, find the best polynomial degree and the coefficients of the best curve. Explain your approach, explaining how you use your data for this task and justifying your conclusions.

Question 3 [3 points out of 20]

Indicate what is the expected error of your curve when predicting the weight of children from their age in months. Note that the error should be in Kg, so take careful note of the units you are computing. Justify your answer, both the value obtained and the data used to obtain it.

Question 4 [7 points out of 20]

Plot the best curve you found along with the data used to fit the curve coefficients and represent the uncertainty of this curve with different shades at the levels the confidence intervals of 95% and 99%. Explain how your code works.