BERNARD BARUCH COLLEGE

# MTH 4330

## Instruction

- You should be able to produce a report in the form of a single pdf file that should be no longer than 8 pages.
- You should use Python 3 and the sklearn, numpy libraries for your statistical, numerical analysis.
- Any figure you include in the report should be mentioned in the text and the figure should be discussed in detail trying to answer the following questions: Why am I including this figure in this document? Which information is contained in the figure?
- Every figure should have labels on its axis and a caption summarizing the data represented in the figure.
- Please report the analysis of your statistical tests in the form of tables or graphs without forgetting to include a description of these tables and graphs and the information you can conclude from them.
- Please include snippet of python code you are using for your statistical analysis. Statements like, "Table 1 reports the output of the python function xxx (specify which function you are using)" or "the code to compute the coefficient of determination is ...", are encouraged.
- Feel free to email me if you have any question.

## Ex3

The attached file called train.txt contains two columns of data based on which you will build your model. The first column represents the independent variable that we interpret as time and the second column is relative to the price of some good, say ice cream. Please answer the following questions:

(1) Upload and plot the data.

(2) Use the sklearn LinearRegression function to fit a linear model $y = \beta_0 + \beta_1 x$ to the data contained in train.txt .
   (a) Which estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ do you get?
   (b) Can you conclude that there is a linear trend in the data? This is equivalent to ask whether you can conclude that $\beta_1$ is equal to zero or not. Justify your answer using statistical inference.
   (c) What about $\beta_0$, can you conclude whether this parameter is zero or not?

(3) Improve your model by adding non-linear features (still using only the data contained in train.txt). Suppose your colleague is an economist and she suggests you that, in addition to the linear model you built in the previous question, the following non linear features can help you improve your model: $\cos(x)$, $\log(x)$, $\cos(4x)$, $\sin(3x)$, $\sin(5x)$ and $\sin(2x) \times \cos(2x)$. Which procedure are you going to use to choose among them? You can come up with your own procedure for *feature selection* (in this case please don't forget to describe it and justify it) but please always use the the adjusted $R^2$ coefficient (see the textbook or the wikipedia page `https://en.wikipedia.org/wiki/Coefficient_of_determination`) to quantify the quality of your model. Summarize the procedure you are using and conclude which features you should choose to improve your model and why you are choosing these features.

(4) Use Lasso and Ridge regression with all the features suggested by the your frined economist. Do you obtain a comparable result than the one you got at the previous point using feature selection?

(5) The file test.txt contains additional data you can use to evaluate your model on prediction. Compute the output of your model using as input the times in the first column of test.txt and plot the result. Compare the output of the model with the recorded value contained in the second column of text.csv. Do you observe a good agreement between the model and the data in the test.txt file? How do you quantify this agreement? Explain.