Assignment Number 1

General notes:

Solutions are to be written in a jupyter notebook
Time to submit the assignment: 4 weeks.
Reading a parquet file:
Import pandas as pd
pd.read_parquet(path/to/data1.parquet)

Question 1

Objective: understand an alternative view of considering the covariates x as constant without error.

Along with this assignment, you have received a parquet file “data1.parquet” containing samples of measured quantities x and y (x appears on the left column).
y is a response variable which can be written as an explicit function of x.
You know that the technique used for measuring x is twice as better than that for measuring y in the sense of **error variance**, i.e. the variance of the error in x is twice as small as the variance of error in y. (Note that this is something new, you haven’t encountered before - in class we have assumed x to be constant and not a random variable)

Your task is to model y as a function of x respecting the above information regarding the way the data was generated. Compare your solution against a simple linear regression (plot the two regression results one against the other).

You may impose further assumptions (additional to the ones given above) to your model if you find it necessary. Write a short description (within the notebook) of the assumptions you’ve made, if any, the model you’ve found, and most importantly: explain and justify your solution.

Question 2

Objective: get familiar with quantile regression.

Along with this assignment, you have received a parquet file “data2.parquet” containing samples of measured quantities x and y.

Model y as a function of x using a linear regression.

Look at the residuals, do they admit the linear regression assumptions?

Model the data using **quantile regression** with quantiles **0.05, 0.5, 0.95** using the package **statsmodels** (0.12.2) or **scikit-learn**

Add a section in the notebook describing quantile regression - explaining how it works, what it is aiming to maximize/minimize and when it is adequate to use it.

note: import it before importing any other package, otherwise you might encounter import issues).

Question 3

Objective: get familiar with scikit learn and ridge regression.

Along with this assignment, you have received a parquet file “data3.parquet” containing samples of some data. The response (label) column is named “label”.

Use scikit-learn to fit a ridge regression where the coefficient of the regularizer (α) can vary from 0 to 1 in intervals of 0.01.

Your task it to search over all possible values of α. For each to perform cross validation Using the **RepeatedKFold** function with 10 splits and 3 repeats. Then find the α that minimizes the mean absolute error using **grid search** (look for a function that incorporates the cross validation in the search).

Print the best MAE score and its corresponding coefficient α