Assignment

In this assignment, we will implement a number of linear models (including linear regression) to model relationships between different clinical features for heart failure in patients.

For the dataset, we use "heart failure clinical records data set at UCI:
`https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records`

Dataset Description: From the website: "This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features."

These 13 features are:

1. age: age of the patient (years)
2. anaemia: decrease of red blood cells or hemoglobin (boolean)
3. high blood pressure: if the patient has hypertension (boolean)
4. creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
5. diabetes: if the patient has diabetes (boolean)

6. ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
7. platelets: platelets in the blood (kiloplatelets/mL)
8. sex: woman or man (binary)
9. serum creatinine: level of serum creatinine in the blood (mg/dL)
10. serum sodium: level of serum sodium in the blood (mEq/L)
11. smoking: if the patient smokes or not (boolean)
12. time: follow-up period (days)

target death event: if the patient deceased (DEATH_EVENT = 1) during the follow-up period (boolean)

We will focus on the following subset of four features:

1. creatinine phosphokinase
2. serum creatinine
3. serum sodium
4. platelets

and try establish a relationship between some of them using various linear models and their variants.

Question 1:

1. load the data into Pandas dataframe. Extract two dataframes with the above 4 features: df_0 for surviving patients (DEATH_EVENT = 0) and df_1 for deceased patients (DEATH_EVENT = 1)
2. for each dataset, construct the visual representations of correponding correlation matrices M_0 (from df_0) and M_1 (from df_1) and save the plots into two separate files
3. examine your correlation matrix plots visually and answer the following:
   (a) which features have the highest correlation for surviving patients?
   (b) which features have the lowest correlation for surviving patients?
   (c) which features have the highest correlation for deceased patients?
   (d) which features have the lowest correlation for deceased patients?
   (e) are results the same for both cases?

Question 2: In this question you will compare a number of different models using linear systems (including linear regression). You choose one feature X as independent variable X

and another feature Y as dependent. Your choice of X and Y will depend on your facilitator group as follows:

1. Group 1: X: creatinine phosphokinase (CPK), Y: platelets
2. Group 2: X: platelets, Y: serum sodium
3. Group 3: X: serum sodium, Y: serum creatinine
4. Group 4: X: platelets, Y: serum creatinine

We will now look for the best model (from the list below) that best explains the relationship for surviving and deceased patients. Consider surviving patients (DEATH_EVENT = 0). Extract the corresponding columns for X and Y. For each of the models below, we will take 50/50 split, fit model with X_{train} and predict Y_{test} using X_{test}. From the predicted values $\text{Pred}(y_i)$ we compute the residuals $r_i = y_i - \text{Pred}(y_i)$. We can then estimate the loss function (SSE sum of the squared of residuals)

$$L = \sum_{x_i \in X_{test}} e_i^2$$

You do the same analysis for deceased patients. You will consider the following models for both deceased and surviving patients:

1. $y = ax + b$ (simple linear regression)

2. $y = ax^2 + bx + c$ (quadratic)

3. $y = ax^3 + bx^2 + cx + d$ (cubic spline)

4. $y = a \log x + b$ (GLM - generalized linear model)

5. $\log y = a \log x + b$ (GLM - generalized linear model)

For each of the model below, you will do the following (for both deceased and surviving patients)

(a) fit the model on X_{train}

(b) print the weights $(a, b, \ldots)$

(c) compute predicted values using X_{test}

(d) plot (if possible) predicted and actual values in X_{test}

(e) compute (and print) the corresponding loss function

Question 3: Summarize your results from question 2 in a table like shown below:

Model	SSE (death_event=0)	(death_event=1)
$y = ax + b$ $y = ax^2 + bx + c$ $y = ax^3 + bx^2 + cx + d$ $y = a \log x + b$ $\log y = a \log x + b$		

1. which model was the best (smallest SSE) for surviving patients? for deceased patients?
2. which model was the worst (largest SSE) for surving patients? for deceased patients?