DS203-2024-S1: Exercise – 1

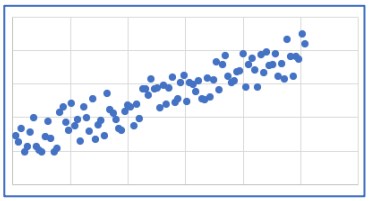
- **Submissions due by: Aug 16, 2024, 09:00 AM. No cribs will be entertained.**
- **Follow the Submission Guidelines given at the end of this document**
- **(-1) marks will be added to your account for late / non submissions.**
- **(-10) marks will be added to your account for copied / fraudulent submissions. Blank and woefully inadequate / irrelevant submissions will be considered fraudulent.**

Part - A

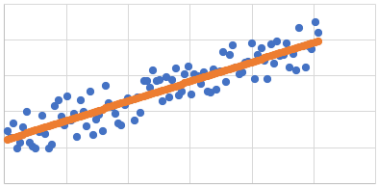
- Review **Simple Linear Regression Derivation.pdf** (uploaded to Moodle) to understand the closed form derivations of the Simple Linear Regression (SLR) coefficients **a** and **b**.

Part – B

Note: All steps in Part – B should be completed using a spreadsheet such as Excel, LibreOffice, etc.

1. Using a spreadsheet create a dataset comprising 100 pairs (x_i, y_i) as per the following guidelines:
   - Create 100 random values of x_i lying between 0 and 1 (both inclusive)
   - Corresponding to each x_i, create y_i such that the scatter plot of y_i v/s x_i looks like the image below:
   - 
2. Create the scatter plot resulting from *your* dataset (x_i, y_i)
3. Using the (x_i, y_i) data, calculate the regression coefficients **a** and **b** (all calculations should be entirely done using the spreadsheet). The equation of the resulting regression model (line) will be as shown below.

$$\hat{y}_i = a \cdot x_i + b$$

4. Using this regression line predict $\hat{y}i$ corresponding to every x_i
5. Superimpose the regression line over the scatter plot created in step 3, as shown below:
   - 
6. Calculate the prediction error e_i corresponding to every y_i, and calculate the error metrics SSE and MAE.
7. Create a scatter plot of e_i v/s x_i

8. As discussed in class, the simplest (and naïve) model is one that predicts the mean, as shown below. Using this model calculate $\mathbf{e}_i$, SSE and MAE and create the scatter plot of $\mathbf{e}_i$ v/s $\mathbf{x}_i$

$$\hat{y}_i = \bar{y}$$

9. Compare the error metrics and error scatter plots resulting from the above two distinct models and record your analysis and explain the differences between the two error scatter plots. (Note: Stating obvious facts is NOT analysis!)

Submission Guidelines

Create a **properly formatted report** covering all the above steps: Explain the steps and equations used to create the ($\mathbf{x}_i$, $\mathbf{y}_i$) dataset in step 2 and include all the subsequent plots, error metrics, your observations and analysis in this report. **List down your main learnings from this exercise**.

Upload the following files to the E1 submission point on Moodle: Note – the file names should start with **E1-YourRollNo**

1. The spreadsheet containing the data set (and all the calculations that you may have done in the spreadsheet).
2. PDF of your report.

oooOOOooo