DS203-2024-S1: Exercise – 1

- **Submissions due by: Jan 28, 2025, 23:55 Hrs. 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

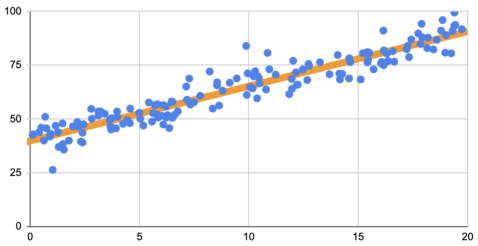
- Review **Simple Linear Regression Derivation.pdf** (uploaded to Moodle) to understand the closed form derivations of the Simple Linear Regression (SLR) coefficients **a** and **b**.

Part – B

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

1. Download and use the dataset *E1.csv*. This dataset contains 150 pairs (x_i, y_i) suitable for simple linear regression.
2. Create the scatter plot resulting from the above 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, MSE, RMSE, and MAE. Research and find out the context in which these error metrics are used.
7. Create a scatter plot of e_i v/s x_i.
8. Create a histogram of the errors (e_i), adjust the bin size, and comment on the distribution of the values. Is it a good regression from the error analysis point of view?
9. How to find out if the distribution is normal? Deduce it based on an analysis of the skewness and kurtosis values of e_i.
10. Compute **R-squared** and comment on the goodness of fit based on the value of R-squared.

11. Using the model $\hat{y}_i = \mathbf{a} \cdot \mathbf{x}_i$, calculate $\mathbf{e}_i$, SSE, MSE, RMSE, and MAE for this model and create the scatter plot of $\mathbf{e}_i$ versus $\mathbf{x}_i$.

12. 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. **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