DS203-2025-S1: E1

- **Submissions due by:** Aug 12, 2025; 23:55 Hrs.
- **Follow the submission guidelines** given at the end of this document.
- **Late / non-submissions:** (-1) mark will be added to your account.
- **Copied / fraudulent submissions:** (-10) marks will be added to your account. Blank and woefully inadequate / irrelevant submissions will be considered fraudulent.

Part A

- Review `Simple Linear Regression Derivation.pdf` (uploaded to Moodle)

Part B

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

1. Download the dataset *E1.csv* uploaded to Moodle. Review and create a description of the data in terms of its size (rows and columns), the type of data it contains - including the *level of measurement* that can be associated with each column.

2. Create a scatter plot to visualize the data and comment on the suitability of a *Simple Linear Regression* model for the data.

3. Using the data, calculate the regression coefficients β_0 and β_1. **All calculations should be entirely done using a spreadsheet**.

4. The equation of the resulting regression model (line) will be: $\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$; using this regression line, predict $\hat{y}_i$ corresponding to every x_i.

5. Create a plot by superimposing the predicted points $(x_i, \hat{y}_i)$ over the scatter plot of (x_i, y_i) created earlier.

6. For every y_i calculate the prediction error $e_i = (y_i - \hat{y}_i)$, and follow it up by calculating the following error metrics:

    a. SSE (Sum of Squared Errors)

    b. MSE (Mean Squared Error)

    c. RMSE (Root Mean Squared Error)

    d. MAE (Mean Absolute Error)

(*Find out the contexts and applications in which these error metrics are used*)

5. Create a scatter plot of e_i v/s x_i and record your observations.

6. Create a histogram of the errors e_i, and comment on the shape of the histogram. *Is it a good regression from the error analysis point of view?*

7. How to find out if the distribution is normal? Deduce it based on an analysis of the **skewness** and **kurtosis** values of e_i.

8. Compute R^2 for this regression and comment on the goodness of fit based on its value.

9. As discussed in class, sometimes (*when?*) we are only interested in the *slope* of the regression line and we do not need the intercept . In such cases we are interested in a simple linear regression model which can be expressed as: $\hat{y}_i = \beta_1 \cdot x_i$;

    - a. Derive the expression for β_1
    - b. Using the resulting model create the following plots:
    - c. $\hat{y}_i$ v/s x_i
    - d. e_i v/s x_i
    - e. Calculate the following error merics: e_i, SSE, MSE, RMSE, MAE for this model.
    - f. Compare these plots and metrics with the corresponding plots and metrics of the earlier model and record your conclusions based on this analysis.

      (Note: Stating obvious facts is NOT analysis!)

Submission Guidelines

1. Create a properly formatted **report** covering all the above steps.
2. List down your **main learnings** from this exercise.
3. Upload the following files to the E1 submission point on Moodle (**Note:** The file names should start with ***E1-YourRollNo***).
    - The **spreadsheet** containing the dataset and all calculations done in the spreadsheet.
    - **PDF of your report.**

---

oooOOOooo