Course	DS 203: Programming for Data Science
Activity	Type: Exercise Title: E02 – Simple Linear Regression
Background	• Simple Linear Regression (SLR) is the term used when there is only one independent variable (x) in the regression • Data analysis tools provide functions and interactive methods to perform regression using observations (xi,yi). • Several metrics are generated as part of the regression. These metrics collectively determine the quality of regression. • It is important to know what these metrics are, understand how to interpret and use them.
Expected outcomes of this exercise	• Acquire some of the skills required to visualize, explore and analyze data sets. • Perform SLR using tools like Excel, Google Sheets, SmartOffice Sheets, etc. and understand the outputs generated • Understand the terms: Independent (x) and dependent variable (y), coefficients of regression (a, b or β_1, β_0), F-test, t-test, p-value • Understand metrics such as variance, standard deviation, R^2, SST, SSR, SSE, F-value, t-value, p-value, and learn how to interpret and apply them • Acquire the skill to compare the metrics across multiple data sets / multiple regressions and interpret the trends. • Most important in Data Science: Learn how to concisely collate and present results, use them for analyses, draw conclusions, and take decisions.
Tools	• At least one of the following ○ Microsoft Excel ○ Libreoffice Calc • Recommended: Use both these tools !
Effort estimate	• **6 hours**
Submission type	• **Mandatory submission**
Due date and time	• **17th August, 23:55 Hrs** • (This assignment will be discussed in the class of 18th Aug)
Submission instruction	• Submit to the appropriate Moodle submission point • Your final submission should be in the form of a single PDF file. • Your solutions and answers should include explanations / graphs / charts / Tables that are necessary to fully explain the solution(s). • **Complete this exercise and submit well within time. No extensions will be granted, no email submissions will be accepted.**
Marks for the exercise	• Credit will be given for timely submission to Moodle • The exercise itself will not carry marks. • **Your understanding and skill – expected to be gained by completing this exercise – will be gauged in a quiz, test, or viva that will be conducted subsequently.**
References	• Lecture notes • Help documentation for tools like Excel, etc. • Articles, blogs and other sources

Note:

- Data file to be used in this exercise: **E2-data-sets.xlsx**
- This data file contains 5 sheets, each with 2 columns of data.

Part 1: (Estimated effort: 2 hours)

1. Thoroughly review the uploaded document “**DS203-2023-08-11-board.pdf**” to understand the Linear Regression concepts and steps covered therein.
2. Replicate the steps in the above document using data set ‘**Data1**’
3. Submit the outputs and your interpretation of the results.

Part 2: (Estimated effort: 4 hours)

- Plots, statistics, and outputs generated as part of this exercise must be captured in a **Table**, the format of which is shown later, with an example.
- Complete the following steps for each data set (Data1 through Data5)
- Create scatter plots (**y_i v/s x_i**) to visualize the relationship between y and x.
- Fit Simple Linear Regression (SLR) models on each data set and capture the metrics.
- Using the regression coefficients generate the predicted values $\hat{y}_I$ and also the corresponding error values **e_i** (y_i - $\hat{y}_I$)
- Create combined scatter plots of **y_i and $\hat{y}_I$ v/s x_i**
- Create ‘error plot’, **e_i v/s x_i**
- Analyze the Table to answer the questions listed below.
  - Analysis should typically be based on the following statistics and metrics: Variance(y_i), correlation coefficient, SSE, MSE, R^2, F-value, Significance F, regression coefficients and their p-values.
  - **Please do not ‘state the obvious’ in your analysis and conclusions**. Eg. – statements like ‘the variance of data set Data1 is 10.3 and that of Data2 is 11.4’ is not analysis! Your analysis should explain what the numbers and / or the differences mean, along with the trend observed, if any, and your conclusions based on such analysis.

  Note:

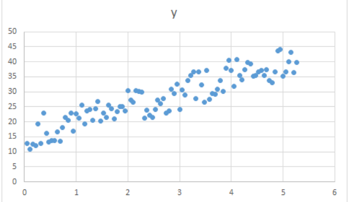
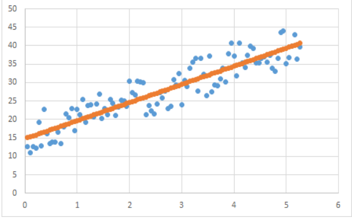
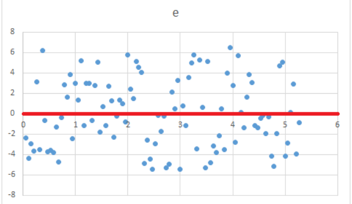
  - MSE (Mean Squared Error) is defined as **SSE/n**, and it is often used as a metric to compare the quality of regression across multiple data sets or regressions using multiple strategies. One of it’s use, along with other metrics like R^2, is to select the most acceptable regression model from many candidate models.
  - Look up the term “Pearson’s Correlation Coefficient”. What does it indicate, how is it calculated, and how to obtain its value in Excel / Libreoffice Calc.

Questions:

1. Data sets Data1 and Data2 are **samples** of different sizes randomly drawn from the same population. Care has been taken to ensure that the samples are true representative of the population. What can you say about the impact of sample size on regression quality? Explain by comparing the following: regression coefficients and their p-values, R2, MSE, F-value, Significance F. Which of these is majorly impacted, and what can you conclude from this? Can you explain why?
2. Compare the data sets Data2 and Data3 by focusing on the plots, variances, and correlation coefficients. What differences do you observe? What do they indicate?
3. Compare the regression outputs from Data2 and Data3. Focus on the following: correlation coefficients, regression coefficients and their p-values, R2, MSE, F-value, Significance F. What major differences do you observe? What do they indicate about the relative quality of these regressions?
4. Compare the data sets Data1 and Data4, and their regression outputs. What are your observations about the relative quality of the data sets themselves? How do the two regression outcomes compare? What conclusions can you make from this analysis – particularly related to data quality and regression quality?
5. Consider Data5. In this case you have fitted a linear model over non-linear data. What is the impact on the regression metrics and how do they correlate with the data quality – variances, correlation coefficient, regression coefficients and their p-values, R^2, MSE, F-value, and Significance F? In your analysis also compare these metrics with those observed in the previous 4 data sets – and state your observations and conclusions.
6. Across the data sets, what relationship do you observe between the correlation coefficients and the quality of regressions? Can you detect any mathematical relationship between the correlation coefficient and R^2?
7. Error plots and error metrics are an important consideration while assessing regression quality. Observe all the error plots. Qualitatively, in what way does the error plot for Data5 differ from the other error plots? Why do such error plots indicate an incorrect regression?
8. Based on all the observations, analyses, and conclusions you have made so far, now list down the specific criteria and steps you will take to decide upon the quality of a Linear Regression model, and to decide whether you will use the regression model for prediction!

The next page shows the Table format for capturing the data set characteristics and the regression outcomes.

Table format for data capture

Data Set	Data1	Data2			
Scatter plot • y_i v/s x_i					
Scatter plot • y_i v/s x_i • $\hat{y}_i$ v/s x_i					
Error plot ● e_i v/s x_i					
Data size (n)		100			
Variance(y)		67.504			
stdev(y)		8.216			
Corr(x,y)		0.911			
Coeff ‘a’		4.903			
Coeff ‘b’		14.810			
p-value ‘a’		1.5E-39			
p-value ‘b’		4.6E-39			
R^2		0.830			
$1-R^2$		0.17			
F-value		497.833			
Significance F		1.54E-39			
SSE		1133.415			
MSE (SSE/n)		11.334			

*****