

DS203-2024-S1: Exercise – 2

- **Submissions due by: Aug 23, 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.**
-

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

1. In E1 you have already created a dataset (y, x) . Calculate the **Pearson Correlation Coefficient** (r) for this dataset and comment on the value.
2. Using the E1 dataset, fit a Regression line and generate detailed regression output **using the built-in Regression functionality of the spreadsheet**. Show the Regression output in your report. It should include the regression coefficients, their associated standard errors, p-values, confidence intervals, the F-statistic, and R^2 values. (All spreadsheets create detailed Regression output – some provide a dialog based UI to do it, while in some others appropriate function(s) have to be used.)
3. Answer the following questions:
 - a) Comment on the value of R^2 (the Coefficient of Determination). How good is the regression? What does this value represent?
 - b) Independently calculate the value of R^2 using its basic definition. Compare it with the value created by the Regression functionality.
 - c) Compare R^2 with r^2 (the square of the Correlation coefficient, calculated above. What is your observation?
 - d) What do you understand by the Standard Error values associated with the coefficients?
 - e) Are the coefficients of the Regression statistically significant? Justify your answer.
 - f) What do you understand by the 95% confidence interval associated with each of the coefficients. What happens if ZERO is a part of this interval?
 - g) Comment on the F-value / F-Statistic. What does it represent? Why is it important?
4. Create 5 variants of the E1 dataset – by changing the variance of the data. In each case fit a regression line using the built-in regression functionality.
 - a) For each variant, note down the regression outcomes and other statistics such as R^2 , p-value, F-value, SSE, MSE, variance of y (use a Table to record all these values for each variant).
 - b) Create a plot of R^2 v/s variance (of y). What do you observe? Why?
 - c) Analyze the effect of variance on the regression parameters and prediction errors and state your observations and conclusions.

Submission Guidelines

Create a **properly formatted report** covering all the above questions. Include the required Tables plots, error metrics, your observations and analysis in this report. **List down your main learnings from this exercise.**

Upload the following files to the E2 submission point on Moodle: Note – the file names should start with **E2-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.

ooo000ooo