Assignment: Numerical Based Problem on Ridge Regression

1 Objective

The purpose of this assignment is to understand the application of Ridge Regression (L2 regularization) for solving regression problems in the presence of multicollinearity or overfitting. You will apply Ridge Regression to a numerical dataset and compare the results with Ordinary Least Squares (OLS) regression.

2 Instructions

2.1 1. Introduction to Ridge Regression

- Briefly explain the difference between Ridge Regression and Ordinary Least Squares (OLS) regression.
- Define the Ridge Regression cost function and explain the role of the regularization parameter λ.

2.2 2. Dataset

Below is a dataset containing three independent variables (X1, X2, X3) and a dependent variable (Y). Your task is to use Ridge Regression to predict the dependent variable.

X1	**X2**	**X3**	**Y**
100	25	10	20
200	35	12	30
300	45	15	40
400	55	17	50
500	65	19	60
600	75	22	70
700	85	25	80

2.3 3. Tasks

1. Perform Ordinary Least Squares (OLS) regression on the dataset. Calculate the coefficients β_0, β_1, β_2, and β_3 using the OLS method.
2. Use Ridge Regression to fit the same dataset with $\lambda = 5$. Calculate the coefficients β_0, β_1, β_2, and β_3 using the Ridge Regression formula:
$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$
3. Compare the coefficients obtained from OLS regression and Ridge regression. Discuss how the regularization term λ affects the magnitude of the coefficients.
4. Use the Ridge Regression model to predict the dependent variable Y for a new set of values: $X1 = 350$, $X2 = 50$, $X3 = 16$. Compare the predicted value from the Ridge Regression model with the prediction from the OLS model.
5. Evaluate the model performance for both OLS and Ridge Regression using the following metrics:
   - Mean Squared Error (MSE)
   - R-squared

   Compare the performance of both models based on these metrics and discuss which model performs better on the given dataset.

3 Bonus Task

- Experiment with different values of λ (e.g., $\lambda = 0.1$, $\lambda = 10$, $\lambda = 50$) and observe the changes in the model coefficients. Discuss the effect of different values of λ on model performance and overfitting.

4 Guidelines for Solution

- Use any programming language or statistical tool (Python, R, Excel) to perform the calculations and solve the regression problems.
- Provide a report with the following:
  - Step-by-step solution to each task, including all calculations for coefficients.
  - Visualizations such as scatter plots, residual plots, or coefficient comparison charts (optional but encouraged).
  - A discussion of your findings, including the impact of regularization on model performance and coefficient shrinkage.

5 Submission Guidelines

- Submit a PDF or Word report that includes your detailed solutions, plots, and interpretations.
- Ensure all your steps are clearly documented, and any assumptions or explanations are provided for each part of the assignment.

6 Dataset (for easy reference)

X1	**X2**	**X3**	**Y**
100	25	10	20
200	35	12	30
300	45	15	40
400	55	17	50
500	65	19	60
600	75	22	70
700	85	25	80

7 Example of Step 2 Ridge Regression Formula

For Ridge Regression, you will calculate the coefficients using the following formula:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

Where:

- X is the design matrix (with the independent variables X1, X2, X3).
- λ is the regularization parameter.
- I is the identity matrix.