Introduction to Regression Analysis

DPP Assignment

Prepared by: **Sarthak Sahu**

Day 1: Understanding Basics and Assumptions

Question 1: Define simple linear regression in your own words.

Answer: Simple linear regression is a statistical technique used to model the relationship between one independent variable and one dependent variable by fitting a straight line that best represents the data.

Question 2: List and explain the assumptions of simple linear regression.

Answer: The assumptions are: (1) Linearity – relationship is linear; (2) Independence – observations are independent; (3) Homoscedasticity – constant variance of errors; (4) Normality – residuals are normally distributed; (5) No influential outliers.

Question 3: Using the given dataset, plot a scatter plot of Hours_Studied vs Exam_Score and identify whether the relationship appears linear.

Answer: The scatter plot shows an upward trend, meaning exam scores generally increase with study hours, indicating a linear relationship.

Question 4: Discuss what a linear relationship implies in this context.

Answer: A linear relationship implies that each additional hour of study leads to an approximately constant improvement in exam score.

Day 2: Fitting the Model

Question 1: Write the mathematical equation of a simple linear regression model.

Answer: The regression equation is: $Y = b0 + b1X$.

Question 2: Fit a simple linear regression model to the dataset.

Answer: Using Excel regression formulas, the fitted model gives slope ≈ 0.7 and intercept ≈ 2.1.

Question 3: Identify the slope and intercept and interpret their meanings.

Answer: Slope means exam score increases about 0.7 marks per hour studied. Intercept means expected score is about 2.1 when hours studied is zero.

Question 4: Plot the regression line on the scatter plot of the data.

Answer: The regression line is the best-fit straight line drawn through the scatter plot.

Question 5: Explain how the model minimizes the sum of squared residuals.

Answer: The model chooses parameters b0 and b1 such that the sum of squared differences between actual and predicted values is minimized.

Day 3: Evaluating the Model

Question 1: Define R^2, Adjusted R^2, and Mean Squared Error (MSE).

Answer: R^2 measures explained variance, Adjusted R^2 adjusts for number of predictors, and MSE measures average squared prediction error.

Question 2: Calculate R^2 and MSE for the fitted model.

Answer: Excel provides $R^2 \approx 0.61$. MSE is computed using the squared residuals average.

Question 3: Interpret the R^2 value.

Answer: About 61% of the variation in exam scores is explained by hours studied.

Question 4: Discuss the importance of Adjusted R^2.

Answer: Adjusted R^2 becomes important when multiple predictors are added to prevent misleading increases in R^2.

Day 4: Outliers

Question 1: Plot scatter plot of the extended dataset to identify outliers.

Answer: The point (50,100) is far away from the rest and is an outlier.

Question 2: Calculate residuals and identify outliers.

Answer: The outlier has an extremely large residual compared to other points.

Question 3: Explain how the outlier affects regression line and metrics.

Answer: It pulls the regression line upward, increases error (MSE), and distorts R^2.

Question 4: Discuss strategies to handle outliers.

Answer: Strategies include removing the outlier, transforming data, or using robust regression methods.

Day 5: Real-World Problem: Predicting Housing Prices

Question 1: Plot House_Size vs Price scatter plot.

Answer: The scatter plot shows that larger houses tend to have higher prices.

Question 2: Fit a regression model to predict Price based on House_Size.

Answer: Excel calculates the slope and intercept for the housing dataset.

Question 3: Write the regression equation derived from the model.

Answer: Price = b0 + b1(House_Size).

Question 4: Predict the price of a house with size 1000 sq ft.

Answer: Substituting X=1000 into the regression equation gives the predicted price.

Question 5: Evaluate the model using R^2 and MSE.

Answer: R^2 shows the proportion of price variation explained by size, and MSE gives prediction error.

Question 6: Discuss limitations of using simple linear regression here.

Answer: House price depends on many factors such as location, rooms, and market trends, so size alone is not sufficient.