Introduction to Regression Analysis DPP Assignment

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 – the relationship between variables is linear; (2) Independence – observations are independent; (3) Homoscedasticity – constant variance of errors; (4) Normality – residuals are normally distributed.

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 where exam scores generally increase with study hours, indicating a linear relationship.

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

Answer: It implies that for every additional hour spent studying, the exam score is expected to increase by a consistent amount on average.

Day 2: Fitting the Model

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

Answer: The equation is $Y = \beta_0 + \beta_1 X + \varepsilon$, where Y is the dependent variable, X is the independent variable, β_0 is the intercept, β_1 is the slope, and ε is the error term.

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

Answer: The model is fitted in Excel using the SLOPE and INTERCEPT functions applied to the dataset.

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

Answer: The slope represents the average increase in exam score for each additional hour studied. The intercept represents the predicted exam score when study hours are zero.

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

Excel Mapping: Right-click a data point in your scatter plot -> Add Trendline -> Select Linear.

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

Answer: The least squares method adjusts the regression line so that the sum of squared differences between actual values 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 how much variance in the dependent variable is explained by the model. Adjusted R^2 adjusts R^2 based on the number of predictors. MSE measures the average squared difference between actual and predicted values.

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

Answer: R^2 is calculated using the RSQ function in Excel, and MSE is calculated as the average of squared residuals.

Question 3: Interpret the R^2 value.

Answer: The R^2 value indicates the proportion of variation in Exam_Score that is explained by Hours_Studied.

Question 4: Discuss the importance of adjusted R2R^2 when multiple predictors are added (for context, as there is only one predictor here)

Answer: R^2 will always stay the same or increase as you add more predictors (even if they are useless), Adjusted R^2 will decrease if the new predictor does not improve the model significantly. It prevents "overfitting."

Day 4: Outliers

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

Answer: The scatter plot shows the point (50, 100) far away from the main cluster, clearly identifying it as an outlier.

Question 2: Calculate residuals for each data point and identify any outliers.

Answer: Residuals are calculated as Actual Value minus Predicted Value. The data point with an exceptionally large residual is identified as an outlier.

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

Answer: The outlier pulls the regression line toward itself, significantly altering the slope and reducing model accuracy.

Question 4: Discuss strategies to handle outliers in regression analysis.

Answer: Outliers can be handled by removal (if erroneous), data transformation, or using robust regression methods.

Day 5: Real-World Problem – Predicting Housing Prices

Question 1: Plot a scatter plot of House_Size vs Price to examine the relationship.

Answer: The scatter plot shows a strong positive relationship between house size and price.

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

Answer: A simple linear regression model is fitted in Excel using the housing dataset.

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

Answer: The regression equation is written using the calculated slope and intercept values from Excel.

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

Answer: Using the regression model, the predicted price for a 1000 sq ft house is obtained using Excel's FORECAST.LINEAR function.

Question 5: Evaluate the model using R2R^2 and MSE. Interpret the results

Excel Result: R^2 is approximately **0.99**. **Interpretation:** House size explains 99% of the variation in price, indicating an excellent model fit for this specific data.

Question 6: Discuss the limitations of using a simple linear regression model in this context.

Answer: The model ignores other important factors such as location, number of rooms, and property age, which limits prediction accuracy.