Case Study: Bike Sharing

Nimish Mohan S

Assignment-based Subjective Questions

1) **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   Ans: Categorical variables such as season, year, month, holiday, weekday, working day, and weather situation have a significant effect on the dependent variable (i.e., the demand for shared bikes). These variables are important in understanding the demand for shared bikes across different seasons, years, months, and weather situations. For example, we can infer that the demand for shared bikes might be higher during the summer season than in winter, or during weekdays than on weekends. The effect of these categorical variables on the dependent variable can be seen through visualization techniques such as box plots, bar plots, or heatmaps.

2) **Why is it important to use drop_first=True during dummy variable creation?**

   Ans: When creating dummy variables from categorical variables, we need to drop one category to avoid the dummy variable trap, which is a scenario where one variable can be predicted from the others. Therefore, we need to use the drop_first=True parameter to drop one category from each categorical variable. By dropping one category, we avoid perfect multicollinearity and ensure that the regression model works correctly.

3) **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   Ans: From the pair-plot among the numerical variables, we can see that the variable 'temp' has the highest correlation with the target variable 'cnt', which is the count of shared bikes. This means that the temperature has a strong influence on the demand for shared bikes. We can use this information to build a more accurate model for predicting the demand for shared bikes.

4) **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Ans: After building the model on the training set, we validated the assumptions of linear regression by checking for the following:

   Linearity: We checked if the relationship between the independent variables and dependent variable is linear by plotting a scatter plot of the actual values against the predicted values.

   Normality: We checked if the residuals are normally distributed by plotting a histogram or a density plot of the residuals.

   Homoscedasticity: We checked if the variance of the residuals is constant across all values of the dependent variable by plotting a scatter plot of residuals against the predicted values.

   Independence: We checked if the residuals are independent of each other by plotting a scatter plot of residuals against the predicted values or time.

5) **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Ans: Based on the final model, the top three features contributing significantly towards explaining the demand for shared bikes are:

Temperature (temp), Season (season), Weather situation (weathersit)

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Ans: Linear regression is a supervised learning algorithm used for predicting continuous numerical values. The goal of linear regression is to find a linear relationship between the independent variable(s) and the dependent variable. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are multiple independent variables. The algorithm works by finding the best-fit line that minimizes the sum of the squared errors between the predicted and actual values. This is done by estimating the slope and intercept of the line using the least squares method. Once the line is fitted, it can be used to make predictions for new data.

2) Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four datasets that have the same statistical properties, such as mean, variance, correlation, and regression line, but look very different when plotted. The datasets were created by statistician Francis Anscombe to demonstrate the importance of data visualization and to show that numerical summaries can be misleading. The quartet consists of four datasets that have different patterns of data points, ranging from linear to non-linear, and from a strong relationship to no relationship at all. The quartet highlights the importance of visualizing data and checking assumptions before drawing conclusions.

3) What is Pearson's R?

Ans: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables. It measures the strength and direction of the linear association between the two variables, with values ranging from -1 to 1. A value of -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. The coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations. Pearson's R is commonly used in data analysis, including in linear regression, to determine the relationship between variables.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is the process of transforming variables to have a similar scale. Scaling is performed to ensure that variables with larger ranges do not dominate variables with smaller ranges in a model. Normalized scaling and standardized scaling are two types of scaling. Normalized scaling transforms variables so that they have a range of 0 to 1. Standardized scaling transforms variables so that they have a mean of 0 and a standard deviation of 1. The difference between the two is that normalized scaling preserves the shape of the distribution, while standardized scaling transforms the distribution to have a mean of 0 and a standard deviation of 1.

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF, or variance inflation factor, is a measure of the multicollinearity of variables in a linear regression model. It measures how much the variance of the estimated regression coefficient is increased due to the correlation between the independent variables. Sometimes, the value of VIF can be infinite, which means that there is perfect multicollinearity between the independent variables. This happens when one or more of the independent variables can be expressed as a linear combination of the other independent variables. Perfect multicollinearity makes it impossible to estimate the regression coefficients and makes the model invalid.

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A Q-Q plot is a graphical tool used in statistics to assess whether or not a set of data can be assumed to be normally distributed. It compares the data distribution to the expected normal distribution by plotting the quantiles of the data against the quantiles of the normal distribution. In linear regression, Q-Q plots are used to check the normality assumption of the residuals, which is a necessary condition for valid inference using linear regression models.