Linear Regression Subjective Questions

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? (3 marks)

Ans 1. The analysis of categorical variables in the dataset reveals significant effects on the dependent variable, bike rental demand.

- **Season:** The season variable has a noticeable impact on bike rental demand. Spring, summer, and winter seasons exhibit higher average daily bike rentals compared to the overall average. Specifically, fall and summer seasons outperform winter and spring in terms of bike rental demand. This suggests that the season significantly affects bike rental patterns, with warmer and more pleasant seasons driving increased demand.
- **Month (Mnth):** Certain months such as 6, 9, 8, 7, 5, and 10 have higher average daily bike rentals compared to the overall average, while months 4, 11, 3, 12, 2, and 1 exhibit lower demand. This suggests a strong seasonal pattern, with the warmer months attracting more bike rentals. September, stands out as a month with high demand.
- **Weekday:** Weekdays 2, 3, 4, 5, and 6 (corresponding to Tuesday through Saturday) have higher average daily bike rentals compared to the overall average. In contrast, weekdays 1 (Sunday) and 7 (Monday) have lower demand. This reflects a weekday/weekend pattern, with higher bike rentals during the workweek.
- **Working Day (Workingday):** Working days exhibit higher average daily bike rentals compared to non-working days. This implies that individuals might use bike-sharing services more for their daily commutes or work-related trips on working days.
- **Weather Situation (Weathersit):** Weather conditions categorized as "Clear," "Few clouds," "Partly cloudy," and "Partly cloudy" have a positive impact on bike rental demand, with higher average daily rentals compared to the overall average. In contrast, weather conditions such as "Mist + Cloudy" and "Light Snow" are associated with lower demand.

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

Ans 2. Using **drop_first=True** during dummy variable creation is important because it helps prevent multicollinearity in the regression model. When all dummy variables are included without dropping one as a reference, they become perfectly correlated, leading to multicollinearity. This can destabilize the model and make it challenging to interpret individual category effects.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans 3. Looking at the pair-plot among the numerical variables, variable **temp** and **atemp** have the highest correlation with the target variable **cnt.**

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans 4. The assumptions of Linear Regression were validated after building the model on the training set using several techniques:

- **Residual Analysis:** Residuals, which are the differences between the observed and predicted values, were analyzed. This involved:
    - **Residual Plot:** We visualized the residuals with a residual plot to check for patterns. The plot showed that Residual terms are randomly distributed and there is no pattern which means the output is explained well by the model and there are no other parameters that can explain the model better.
    - **Normality of Residuals:** Checking if the residuals followed a normal distribution. This was done by plotting a histogram of residuals and it was observed that the Residuals were normally distribured with mean 0.
- **Linearity:** The linearity assumption was validated by examining scatter plots of the dependent variable against the predictor variables. Through this plot it was observed that there was a linear relationship between the predictors and the response variable.
- **Homoscedasticity:** Homoscedasticity, which assumes that the variance of the residuals is constant across all levels of the independent variables, was assessed through residual plots. Specifically, a plot of residuals versus predicted values was examined to check for any cone-shaped or funnel-shaped patterns, which would indicate heteroscedasticity (varying variance).Through this plot, we observed that variance of the residuals (error terms) is constant across predictions. i.e error term does not vary much as the value of the predictor variable changes.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans 5. Based on the final model, the top three features contributing significantly towards explaining the demand for shared bikes are:

- **Temperature (temp):** Temperature is the most influential feature, with a coefficient value of 0.412069. This suggests that for every unit increase in temperature, the demand for shared bikes is expected to increase significantly.
- **Weather Situation - Light Snow (weathersit_light_snow):** The presence of light snow in the weather situation is the second most significant feature, with a coefficient value of 0.296761. This indicates that when there is light snow, bike rental demand decreases considerably.
- **Year (yr):** The year variable, particularly the year 2019, is the third most important feature, with a coefficient value of 0.235197. This suggests that in 2019, there was a substantial increase in bike rental demand compared to the previous year (2018).

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans 1. The Linear Regression Algorithm can be explained as :-

- **Linear Relationship :**
  - Linear regression explores how one variable (e.g., study time) is related to another (e.g., test scores) using a simple equation: $Y = \beta_0 + \beta_1 X$.
  - Y is what we're trying to predict (e.g., test scores), X is the influencing factor (e.g., study time), β_0 is a starting point, and β_1 measures the impact of X on Y.
- **Model Fitting :**
  - We adjust β_0 and β_1 to create a line that closely fits our data points. The goal is to minimize the gap between the line and the actual data.
  - However, some randomness (ε) always exists, representing unexplained variations in real-world data.
- **Assumptions :**
  - Linear regression assumes that our data behaves in specific ways: it's linear, errors are independent, and they follow a bell-shaped curve (normal distribution).
  - If these assumptions don't hold, we might need to consider alternative regression methods.
- **Predictive Tool :**
  - Linear regression helps us understand and predict how changes in one variable affect another.
  - It's a valuable tool in data analysis, allowing us to make predictions based on data patterns and relationships.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Ans 2. Anscombe's quartet can be explained as follows :-

- Anscombe's Quartet is a famous statistical phenomenon discovered by the statistician Francis Anscombe in 1973. It comprises four datasets, each with 11 data points, and was designed to highlight the importance of data visualization and the limitations of relying solely on summary statistics.
- What makes Anscombe's Quartet fascinating is that despite having vastly different distributions and relationships between variables, they share nearly identical summary statistics. For example, all four datasets have the same means, variances, correlations, and regression lines. This similarity can mislead analysts who rely solely on these summary statistics.
- Anscombe's Quartet underscores the critical role of data visualization in understanding data. While the datasets may appear statistically similar on paper, plotting them reveals starkly different patterns. It serves as a reminder that exploratory data analysis through graphs and visualizations is essential for gaining a deeper insight into data and uncovering hidden relationships or patterns that summary statistics might miss.

We can say that Anscombe's Quartet is a powerful illustration of the need for both quantitative and visual analysis in statistics. It cautions against drawing conclusions solely based on summary statistics without considering the actual data distribution and relationships.

3. What is Pearson's R? (3 marks)

Ans 3. Pearson's R can be explained as follows :-

- Pearson's correlation coefficient, often denoted as "r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It assesses how closely the data points in a scatterplot cluster around a straight line.
- The value of Pearson's R ranges from -1 to 1. A positive R indicates a positive linear relationship, meaning as one variable increases, the other tends to increase as well. A negative R suggests a negative linear relationship, where one variable typically decreases as the other increases. An R value of 0 signifies no linear relationship between the variables. The closer the R value is to -1 or 1, the stronger the linear relationship. A value of -1 or 1 indicates a perfect linear relationship.
- Pearson's R is widely used in statistics, research, and data analysis to:
    - Assess the association between variables.
    - Determine the strength and direction of the relationship.
    - Identify outliers or influential data points that may affect the correlation.
    - Make predictions based on the observed relationship.

It is a valuable tool for understanding connections between variables in various fields, including economics, social sciences, and natural sciences.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans 4. Scaling can be defined as the process of transforming numerical variables to a specific range or distribution to ensure that they have a comparable scale. It is performed to avoid issues related to different units or scales of variables when working with machine learning or statistical models. Scaling ensures that all variables contribute equally to the analysis.

The benefits of performing Scaling are as follows :-

- Scaling is crucial in various machine learning algorithms and statistical techniques like k-means clustering, principal component analysis (PCA), and gradient descent optimization.
- It ensures that no single variable dominates the analysis due to its scale.
- Choosing between normalized scaling and standardized scaling depends on the specific requirements of the analysis and the characteristics of the data.

The differences between normalized scaling and standardized scaling are as follows :-

- **Normalized Scaling (Min-Max Scaling):**
    - Normalization scales data to a specific range, typically between 0 and 1.
    - It maintains the relative relationships between data points.
    - It is useful when we want to preserve the original distribution of the data.

- **Standardized Scaling (Z-Score Scaling):**
    - Standardization transforms data to have a mean (average) of 0 and a standard deviation of 1.
    - It centers the data around 0 and scales it proportionally to the standard deviation.
    - It is useful when we want to compare variables with different units on the same scale.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans 5. Sometimes the value of VIF (Variance Inflation Factor) is infinite because :

- **Perfect Multicollinearity :**
    - The primary reason for VIF to become infinite is the presence of perfect multicollinearity among predictor variables.
    - Perfect multicollinearity means that two or more variables in a regression model are linearly dependent on each other, to the point where one or more variables can be expressed as exact linear combinations of others.
    - When perfect multicollinearity exists, it becomes impossible for the regression model to estimate the unique contribution of each predictor. This is because the model cannot distinguish the individual effects of these perfectly correlated variables.
- **Mathematical Implication :**
    - VIF is calculated based on the variance of the coefficient estimate of each predictor. In the case of perfect multicollinearity, the model encounters a mathematical problem where the true variance of the coefficient estimate becomes zero.
    - When dividing by zero in the VIF formula, it results in an undefined or infinite VIF value.

To summarise, infinite VIF values occur due to perfect multicollinearity, which essentially renders some predictors as linearly dependent on others. This multicollinearity issue creates a mathematical problem when calculating VIF, leading to the phenomenon of infinite VIF values. Resolving multicollinearity through variable selection or transformation is necessary to address this issue and obtain finite and interpretable VIF values.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans 6. A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles expected from the theoretical distribution.

Use and Importance :

- In Linear Regression:
    - A Q-Q plot is used to validate one of the key assumptions of linear regression: the normality of the residuals (error terms).

    - It helps assess whether the residuals from the regression model are normally distributed, which is crucial for the validity of statistical inference, hypothesis testing, and confidence intervals.
    - By plotting the residuals on the Q-Q plot, you can visually inspect whether they follow a straight line, which would indicate normality. Deviations from a straight line may suggest departures from normality.
- In Data Analysis:
    - Beyond linear regression, Q-Q plots are used in various statistical analyses to check the goodness of fit of data to a chosen distribution, making them essential in hypothesis testing and model validation.
    - Q-Q plots are particularly useful when working with small sample sizes or when assumptions about data distribution need to be confirmed.

Interpretation :

- In a Q-Q plot, if the points closely follow a straight diagonal line (the line of equality), it indicates that the data is well-approximated by the chosen theoretical distribution (e.g., normal distribution).
- Deviations from the diagonal line suggest departures from the assumed distribution, indicating potential issues that need further investigation.
- Q-Q plots are a visual and powerful tool for identifying outliers, skewness, and non-normality in data, helping analysts make informed decisions about data transformations or model adjustments.

To summarise, we can say that a Q-Q plot is used in linear regression and data analysis to assess the normality of residuals and validate statistical assumptions. It provides a visual way to confirm whether data follows a specified distribution, helping ensure the reliability of statistical inferences and the quality of regression models.