Linear Regression Subjective Questions - Answers

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?

Categorical variables like season, yr, mnth, holiday, weekday, workingday, and weathersit have distinct effects on bike demand (cnt). For example:
- Season: Bike demand is typically higher during summer and fall and lower during winter, indicating a strong seasonal effect.
- Year (yr): The demand for bikes has increased from 2018 to 2019, reflecting growing popularity.
- Month (mnth): Certain months, particularly during warmer weather, show higher demand.
- Weather Situation (weathersit): Clear weather conditions show higher bike demand compared to misty or rainy conditions.
These variables help capture the patterns and trends in bike usage based on different time periods and weather conditions.

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

Using drop_first=True in one-hot encoding helps avoid multicollinearity by dropping one level of the categorical variable. This ensures that the dummy variables are independent and prevents perfect multicollinearity, which can distort the results of regression analysis.

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

Among the numerical variables, atemp (feeling temperature) typically shows the highest correlation with the target variable cnt. This indicates that as the perceived temperature increases, bike demand also tends to increase.

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

Linearity: Checked using scatter plots of residuals vs. predicted values to ensure no patterns.
Homoscedasticity: Assessed using the spread of residuals to ensure constant variance.
Normality: Validated using Q-Q plots of residuals to see if they follow a normal distribution.
Independence: Ensured by checking that residuals are not correlated.

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

The top 3 features identified by the model are:
- yr_1 (Year 2019): Indicates higher demand in 2019 compared to 2018.
- temp (Temperature): Directly related to increased bike usage with warmer weather.
- atemp (Feeling Temperature): Closely related to actual temperature, also positively affecting bike demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The basic form is $y = \beta0 + \beta1x1 + \beta2x2 + ... + \betanxn + \varepsilon$, where:
- y is the dependent variable.
- x1, x2, ... , xn are the independent variables.
- β0 is the intercept.
- β1, β2, ... , βn are the coefficients.
- ε is the error term.

The goal is to find the best-fitting line by minimizing the sum of the squared differences between observed and predicted values (least squares method).

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets with nearly identical simple descriptive statistics, yet they appear very different when graphed. It demonstrates the importance of visualizing data before analyzing it, as statistical measures alone can be misleading. Each dataset in the quartet has the same mean, variance, correlation, and linear regression line but different distributions and patterns.

3. What is Pearson's R?

Pearson's R, or Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to 1, where:
- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

It quantifies the strength and direction of the linear relationship between variables.

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

Scaling: The process of transforming data to fit within a specific range or scale. It's performed to ensure that all features contribute equally to the model, particularly important for algorithms sensitive to the scale of data.

- Normalization (Min-Max Scaling): Rescales data to a range of [0, 1] using the formula $x' = (x - \min(x)) / (\max(x) - \min(x))$.
- Standardization (Z-score Scaling): Rescales data to have a mean of 0 and a standard deviation of 1 using the formula $x' = (x - \mu) / \sigma$, where μ is the mean and σ is the standard deviation.

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

Variance Inflation Factor (VIF) measures multicollinearity among predictor variables. An infinite VIF occurs when there is perfect multicollinearity, meaning one predictor variable is a perfect linear combination of other predictors. This makes the regression model unstable and the coefficients unreliable.

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

A Q-Q (Quantile-Quantile) plot compares the distribution of a dataset with a theoretical distribution (usually normal). Points should lie on a straight line if the data follows the theoretical distribution. In linear regression, Q-Q plots are used to check the normality assumption of residuals, ensuring that residuals are normally distributed.