Regression Subjective Questions - Answers

Assignment-based Subjective Questions

Question 1:

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

Answer: From the analysis, categorical variables like `season` and `weathersit` significantly affect bike demand:
- Season: Winter shows the highest demand, followed by Summer. Spring and Fall have comparatively lower demand.
- Weathersit: Clear weather leads to higher demand, whereas adverse conditions like snow, rain, or mist significantly reduce demand.

Question 2:

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

Answer: Using `drop_first=True` avoids the dummy variable trap, which occurs when one category can be perfectly predicted from the others, leading to multicollinearity. Dropping the first category ensures the model remains interpretable without redundancy.

Question 3:

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

Answer: The variable `registered` has the highest correlation with the target variable `cnt`, as most bike rentals are driven by registered users. However, `temp` also has a strong positive correlation with `cnt`.

Question 4:

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

Answer: Residual analysis was performed to validate linear regression assumptions:
1. Linearity: Residuals were randomly scattered around zero.
2. Homoscedasticity: Residual spread was consistent across predicted values.
3. Normality: Histogram and Q-Q plots (if performed) confirmed that residuals follow a normal distribution.
4. Independence: Residuals did not show patterns indicating independence of errors.

Question 5:

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

Answer: The top three features based on their coefficients are:
1. Year (`yr`): Demand was significantly higher in 2019 compared to 2018.
2. Season (`Winter`): Higher bike demand during winter.
3. Season (`Summer`): Moderate bike demand during summer.

General Subjective Questions

Question 6:

Explain the linear regression algorithm in detail.

Answer: Linear regression is a supervised learning algorithm used to model the relationship between one or more independent variables (X) and a dependent variable (Y). It assumes a linear relationship:
$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betanXn + \varepsilon$
The goal is to minimize the sum of squared residuals (SSR) using techniques like Ordinary Least Squares (OLS). Key assumptions include:
- Linearity of relationships.
- Independence of errors.
- Homoscedasticity (constant error variance).
- Normality of error terms.

Question 7:

Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet consists of four datasets with nearly identical statistical properties (mean, variance, correlation, and regression line) but drastically different visual patterns. It emphasizes the importance of visualizing data before relying solely on statistical metrics. Each dataset demonstrates unique behaviors like outliers or non-linear trends that would be missed without visualization.

Question 8:

What is Pearson's R?

Answer: Pearson's R is a measure of linear correlation between two variables. It ranges from -1 to 1:
- R = 1: Perfect positive correlation.
- R = -1: Perfect negative correlation.
- R = 0: No linear correlation.
It quantifies the strength and direction of the linear relationship.

Question 9:

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

Answer: Scaling adjusts the range of features to improve model performance, especially for algorithms sensitive to magnitude differences.
- Normalization scales data to a specific range, typically [0, 1].
- Standardization centers data to a mean of 0 and standard deviation of 1.
Normalization is useful when the range is bounded, while standardization is better for Gaussian-distributed data.

Question 10:

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

Answer: An infinite Variance Inflation Factor (VIF) occurs when a predictor is perfectly correlated with one or more other predictors, leading to multicollinearity. To resolve this, redundant variables should be removed or transformed.

Question 11:

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

Answer: A Q-Q (quantile-quantile) plot compares the quantiles of residuals against a theoretical normal distribution. If the points align along a 45-degree reference line, it indicates normality. This is critical in linear regression to validate the assumption that residuals are normally distributed, ensuring reliable hypothesis tests and confidence intervals.