Multiple Linear Regression Assignment

Bike Sharing Assignment

By, Annette Benny

Assignment-based Subjective Questions

Q1. 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, holiday, working day, and weather situation significantly influence the total bike demand (`cnt`):

- Season: Demand fluctuates with seasons, peaking in summer and fall due to better weather conditions.
- Holiday: Bike usage tends to drop on holidays compared to regular working days.
- Working Day: Bike demand is higher on working days, likely driven by commuting needs.
- Weather Situation: Unfavorable weather, such as heavy rain or snow, reduces bike demand.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Using drop_first=True during dummy variable creation helps avoid the dummy variable trap, where multicollinearity arises due to redundant information. By dropping one category, the model can infer its presence from the other categories, ensuring a more stable and interpretable regression model.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The pair plot indicates that the variable `registered` has the strongest positive correlation with the target variable (`cnt`), as the majority of bike demand is driven by registered users.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The assumptions of linear regression were validated as follows:

- Linearity: Assessed through scatterplots of residuals versus predicted values to confirm the absence of patterns.
- Homoscedasticity: Ensured by checking for consistent variance of residuals across predicted values.
- Normality of Residuals: Verified using a Q-Q plot or a histogram to confirm a normal distribution.
- Multicollinearity: Evaluated by calculating Variance Inflation Factors (VIF) to ensure predictors are not strongly correlated.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top three features influencing demand are:

- Registered Users (`registered`): Exhibits a strong positive correlation with bike demand.
- Temperature (`temp`): Demand increases with higher temperatures.
- Season (`season`): Summer and fall experience significantly higher demand compared to winter and spring.

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 (target) and one or more independent variables (predictors). The core idea is to find the best-fitting line that minimizes the sum of squared differences (residuals) between the observed data points and the predicted values. The model is represented by the equation:
$Y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon$

Where:

- Y is the dependent variable.
- X1,X2,…,Xn are the independent variables.
- β0 is the intercept.
- β1,β2,…,βn are the coefficients for each predictor.
- ϵ is the error term.

The algorithm aims to find the values of β0,β1,…,βn that minimize the residual sum of squares using techniques like ordinary least squares (OLS). Once the coefficients are estimated, the model can be used for prediction.

2. Explain the Anscombe’s quartet in detail

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but very different distributions and relationships when visualized. The purpose of Anscombe's Quartet is to demonstrate the importance of data visualization and show that summary statistics alone can be misleading. Each dataset in the quartet has 11 data points, but:

The relationship between the variables in each dataset is different (linear, quadratic, etc.).Visualizing the data reveals patterns and outliers not captured by the statistics. Anscombe’s Quartet emphasizes the need to use plots, such as scatterplots, to uncover underlying data structures that summary statistics cannot.

3. What is Pearson’s R?

Pearson’s correlation coefficient (denoted as r) measures the strength and direction of the linear relationship between two continuous variables. The value of r ranges from -1 to 1, where:

- r = 1: Perfect positive correlation,
- r = -1: Perfect negative correlation
- and r = 0: No linear relationship.

Pearson’s r is calculated as: $r = \text{cov}(X,Y) / \sigma X \sigma Y$

Where:

- cov(X,Y) is the covariance between XX and YY,
- σX and σY are the standard deviations of XX and YY, respectively.

Pearson’s r is sensitive to outliers and only measures linear relationships.

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

Scaling refers to the process of transforming the features of a dataset so they have similar ranges or distributions, which improves the performance of machine learning algorithms. Scaling ensures that no variable dominates others due to differing magnitudes.

- **Normalized scaling** (Min-Max scaling): Transforms the features so they are within a specified range, typically [0, 1]. This is done by subtracting the minimum value and dividing by the range of the data. Xscaled = $X - \min(X) / \max(X) - \min(X)$
- **Standardized scaling** (Z-score scaling): Transforms the features to have a mean of 0 and a standard deviation of 1. It is done by subtracting the mean and dividing by the standard deviation. Xscaled = $X - \mu / \sigma$

Why Scaling is Performed: Scaling is important because many machine learning algorithms (e.g., gradient descent) are sensitive to the scale of input features. Features with larger ranges may overpower those with smaller ranges, leading to inefficient learning.

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

The Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors. A high VIF indicates that a predictor is highly correlated with other predictors, leading to multicollinearity. When the VIF value is infinite, it typically happens due to perfect multicollinearity, where one predictor is a linear combination of others. This results in a determinant of the design matrix being zero, causing instability in the regression model and making it impossible to estimate the coefficients.

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** is a graphical tool used to assess if a dataset follows a particular theoretical distribution, commonly a normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the data follows the normal distribution, the points will lie on a straight line. In the context of linear regression, a Q-Q plot is used to check the **normality of residuals**, which is one of the assumptions of linear regression. Normal residuals indicate that the model is appropriate for the data, while deviations from the straight line suggest that the residuals are not normally distributed, potentially invalidating the regression results.

THE END