Q & A: Linear Regression Assignment

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)

- The demand of bikes is less in the month of ***spring*** as compared to other seasons.
- The demand of bikes was less in the year ***2018*** whereas it's increased in the year ***2019***
- The demand of bikes is less when ***weathersit*** **(3)** Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

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

By default ***get_dummies()*** create one dummy variable for every level of the input categorical variable. If ***drop_first*** = ***True***, then it will drop the first category. So, if you have K categories, it will only produce K – 1 dummy variables. This will avoid data redundancy because k-1 dummy variables are sufficient enough to store all info of a categorial variable even for the dropped one by storing zeros in all dummy variables of that categorial variable.

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

temp

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

a. The Pair plot showing linear relationship between predictor and the target variables.
b. Error terms are normally distributed with mean zero as shown in histogram.
c. Error terms are independent of each other and have constant variance.

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

1. temp,
2. yr,
3. weathersit/weathersit_3

General Subjective Questions

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

- Linear regression is a statistical algorithm used for modelling the relationship between a dependent variable (also known as the target variable) and one or more independent variables (also known as predictor variables or features). It assumes a linear relationship between the variables and aims to find the best-fitting line that represents this relationship.
- Linear regression is widely used in various domains, including finance, economics, social sciences, and machine learning, as it provides a simple yet powerful framework for modelling and predicting continuous variables based on input features.
- Linear regression models can be classified into two types depending upon the number of independent variables:
  Simple linear regression: When the number of independent variables is 1
  Multiple linear regression: When the number of independent variables is more than 1

  Simple Linear Regression: In simple linear regression, we have one dependent variable (Y) and one independent variable (X). The relationship between X and Y is represented as:
  $Y = \beta0 + \beta1 * X$
  Y: Dependent variable (target) we are trying to predict.
  X: Independent variable (feature).
  β0 and β1: Coefficients of the model. β0 represents the intercept, and β1 represents the slope of the line.

  Multiple Linear Regression: In multiple linear regression, we have one dependent variable (Y) and multiple independent variables (X1, X2, X3, ..., Xn). The relationship is represented as:
  $Y = \beta0 + \beta1 * X1 + \beta2 * X2 + ... + \beta n * Xn$
  Y: Dependent variable (target) we are trying to predict.
  X1, X2, ..., Xn: Independent variables (features).
  β0, β1, β2, ..., βn: Coefficients of the model.

  Cost Function: The linear regression algorithm aims to minimize the difference between the predicted values and the actual values in the training dataset. This is done using a cost function, often the Mean Squared Error (MSE) or Mean Absolute Error (MAE). The cost function calculates the average squared or absolute difference between the predicted values and the actual values.

  Gradient Descent: To find the best-fitting line (or hyperplane) that minimizes the cost function, linear regression uses an optimization algorithm called gradient descent. Gradient descent iteratively updates the model's coefficients to reach the minimum cost. It calculates the gradients of the cost function with respect to each coefficient and updates the coefficients in the opposite direction of the gradients.

Training the Model: The training process involves feeding the algorithm with a labelled dataset, where the dependent variable (Y) and independent variables (X) are known. The algorithm adjusts the coefficients using gradient descent until the cost function converges to a minimum.

Testing and Prediction: Once the model is trained, it can be used to make predictions on new, unseen data. By providing the values of the independent variables (features), the model calculates the predicted value of the dependent variable (target).

Evaluating the Model: The performance of the linear regression model can be evaluated using various metrics, such as the R-squared (coefficient of determination), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), etc. These metrics help assess how well the model fits the data and how accurate its predictions are.

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

Anscombe's quartet is a set of four small datasets that have nearly identical statistical properties, despite having very different graphical representations. Each dataset in Anscombe's quartet contains 11 data points and consists of two variables: X and Y. Despite having similar mean, variance, correlation, and regression line, they display significant differences when plotted graphically.

The significance of Anscombe's quartet lies in its demonstration that summary statistics alone may not be sufficient to understand the underlying relationships in the data. Although each dataset has nearly identical summary statistics (mean, variance, correlation, and regression line), they exhibit very different patterns when visualized.

This emphasizes the importance of data visualization in data analysis and decision-making. Graphical representations allow us to spot patterns, trends, and outliers that might not be apparent from summary statistics alone. Anscombe's quartet serves as a cautionary example against relying solely on numerical summaries and highlights the value of exploratory data analysis using graphs and plots.

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

Pearson's correlation coefficient, often denoted as "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Pearson's correlation coefficient is widely used in various fields, including statistics, social sciences, economics, and data analysis, to quantify and understand relationships between variables.

Pearson's correlation coefficient can take values between -1 and 1:

A positive value of r (closer to +1) indicates a positive linear relationship, meaning that as one variable increases, the other tends to increase as well.

A negative value of r (closer to -1) indicates a negative linear relationship, meaning that as one variable increases, the other tends to decrease.

A value of r close to 0 indicates a weak or no linear relationship between the two variables.

The formula to calculate Pearson's correlation coefficient for two variables X and Y is:

$$R = (\Sigma((X - \bar{X})(Y - \bar{Y}))) / (\text{sqrt}(\Sigma(X - \bar{X})^2) * \text{sqrt}(\Sigma(Y - \bar{Y})^2))$$

Where:
X and Y are the values of the two variables.
$\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.
Σ represents the sum across all the data points.

The important properties of Pearson's R are -

- It is symmetric, meaning that the correlation between X and Y is the same as the correlation between Y and X.
- It is sensitive to linear relationships but not to other types of relationships (nonlinear, monotonic, etc.).
- It is affected by outliers, as extreme values can heavily influence the correlation coefficient.
- It is not appropriate for categorical or ordinal variables.

Interpreting the magnitude of Pearson's R:

- The closer the value of R is to +1 or -1, the stronger the linear relationship between the variables.
- A value close to 0 indicates a weak or no linear relationship.
- In addition to the coefficient itself, a p-value is often calculated to determine the statistical significance of the correlation. The p-value helps assess whether the observed correlation is statistically significant or likely due to random chance.

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

Scaling is a preprocessing technique used in data analysis and machine learning to transform features (variables) to a common scale or range. It involves modifying the values of the features so that they all have similar magnitudes, which can be helpful for improving the performance of certain algorithms and making the data more interpretable.

Why is scaling performed?

Algorithm Performance: Many machine learning algorithms, such as gradient descent-based optimization and distance-based methods, work better when the features are on similar scales. Scaling prevents certain features from dominating others due to their larger magnitudes.

Convergence Speed: Scaling can help algorithms converge faster during training by reducing the number of iterations needed to find optimal solutions.

Interpretability: When features are on the same scale, it becomes easier to interpret the importance and effect of each feature on the model's predictions.

Distance-Based Metrics: In clustering and similarity-based methods, scaling ensures that distances between data points are not overly influenced by differences in feature magnitudes.

The difference between Normalized Scaling and Standardized Scaling is as follows:

Normalized Scaling (Min-Max Scaling):

Normalized scaling transforms the features to a specified range, typically between 0 and 1. It preserves the original distribution of the data.

The formula for normalized scaling is:

X_normalized = (X - X_min) / (X_max - X_min)

Here, X_normalized is the scaled value of feature X, X_min is the minimum value of X, and X_max is the maximum value of X.

Normalized scaling is suitable when the features have a bounded range and are not sensitive to outliers.

Standardized Scaling (Z-score Scaling or Standardization):

Standardized scaling transforms the features to have a mean of 0 and a standard deviation of 1. It centers the data around the mean and scales it relative to the standard deviation.

The formula for standardized scaling is:

X_standardized = (X - X_mean) / X_std

Here, X_standardized is the scaled value of feature X, X_mean is the mean of X, and X_std is the standard deviation of X.

Standardized scaling is appropriate when the features have different scales, and the data may contain outliers. It makes the features more robust to extreme values**.**

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

In statistics, the Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in regression analysis. It quantifies how much the variance of the estimated regression coefficients is inflated due to the presence of correlations between predictor variables. A high VIF indicates a high degree of multicollinearity.
Theoretically, the VIF can be calculated for each predictor variable using the formula:

$$VIF = 1 / (1 - R^2)$$

Where R^2 is the coefficient of determination for the regression model with the predictor variable as the dependent variable and the remaining predictor variables as the independent variables. However, there are cases where the VIF value is calculated to be infinite.

The VIF becomes infinite when the coefficient of determination (R^2) is equal to 1, indicating perfect multicollinearity. Perfect multicollinearity occurs when there is an exact linear relationship between two or more predictor variables. In this case, one or more variables can be expressed as a linear combination of the other variables, making the VIF calculation problematic. This can happen in scenarios such as:

- Including duplicate or highly correlated variables in the regression model.
- Including derived variables that are linear combinations of other variables.

When perfect multicollinearity is present, the regression model becomes unidentifiable, and it is not possible to estimate the coefficients accurately. Therefore, it is essential to identify and address multicollinearity issues in the dataset before performing regression analysis. This can be done by examining correlation matrices, calculating VIF values, and considering variable selection techniques to remove redundant or highly correlated variables.

If the VIF value is calculated to be infinite, it signifies the presence of perfect multicollinearity, indicating that the variables involved are linearly dependent. In such cases, it is necessary to investigate and resolve the multicollinearity issue before proceeding with the regression analysis.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution. It compares the quantiles of the sample data against the quantiles of the expected distribution. It is particularly useful for evaluating whether a dataset follows a specific distribution, such as the normal distribution.

The Q-Q plot works as follows:

- The data points are sorted in ascending order.
- The quantiles of the sorted data are computed.
- The expected quantiles are calculated based on the chosen theoretical distribution.
- The computed quantiles are plotted against the expected quantiles on a scatter plot.

The importance of a Q-Q plot in linear regression lies in its ability to assess the assumption of normality. In linear regression, one of the key assumptions is that the residuals (the differences between the observed and predicted values) should be normally distributed. By examining the Q-Q plot of the residuals, we can visually inspect whether they follow a normal distribution.

If the Q-Q plot of the residuals approximately follows a straight line, it indicates that the residuals are normally distributed. Deviations from the straight line suggest departures from normality. E.g.
- If the points on the Q-Q plot deviate from the straight line at the ends, it suggests heavy tails or outliers.
- If the points curve upward or downward in the middle, it indicates skewness.

A well-behaved Q-Q plot with points closely following the straight line suggests that the residuals meet the normality assumption, supporting the validity of the linear regression model. However, if the Q-Q plot exhibits substantial deviations from the straight line, it indicates a violation of the normality assumption, which can affect the reliability of the regression analysis.

By assessing the Q-Q plot, we can identify potential issues with the distributional assumption and take appropriate actions. This may involve transformations of variables, employing robust regression techniques, or considering alternative regression models.

In summary, a Q-Q plot is a valuable tool in linear regression as it provides a visual assessment of the distributional assumption, specifically the normality assumption. It helps identify departures from normality in the residuals, allowing for adjustments or considerations to ensure the validity of the regression model.