Bike Sharing Demand Prediction Assignment

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?

Ans: From our analysis of the categorical variables from the dataset we can predict the formula for the best fit line equation:

Equation for the best fit line:

cnt = 0.260910 + (0.238304 * yr) - (0.058982 * holiday)-(0.125579 * windspeed)+(0.245523 * season_2)+(0.323927 * season_3)+(0.186103 * season_4) +(0.040413 * mnth_3)+(0.071054* mnth_5)-(0.009695 * mnth_7) +(0.071267 * mnth_9)+(0.119027 * mnth_10)-(0.080676* weathersit_2) -(0.272362* weathersit_3)

Interpretation of the equation:

Intercept (0.260910): This is the baseline value of the bike count (cnt) when all other variables are zero. This value is not very interpretable in this context because it doesn't make sense to have all variables at zero (e.g., no year, no season).

Year (yr, 0.238304): For each additional year, the bike count increases by 0.238304 units, holding all other variables constant. This indicates that bike usage has been increasing over time.

Holiday (-0.058982):If the day is a holiday, the bike count decreases by 0.058982 units compared to a non-holiday, holding all other variables constant. This suggests that fewer people use bikes on holidays.

Windspeed (-0.125579): For each unit increase in windspeed, the bike count decreases by 0.125579 units, holding all other variables constant. This implies that higher windspeed discourages bike usage.

Season (season_2, 0.245523; season_3, 0.323927; season_4, 0.186103): Compared to the baseline season (season_1, which is not included in the model), being in season_2 (spring) increases the bike count by 0.245523 units, being in season_3 (summer) increases it by 0.323927 units, and being in season_4 (fall) increases it by 0.186103 units. This indicates that bike usage is highest in summer, followed by spring and fall, compared to winter.

Month (mnth_3, 0.040413; mnth_5, 0.071054; mnth_7, -0.009695; mnth_9, 0.071267; mnth_10, 0.119027):Compared to the baseline month (mnth_1, January, which is not included in the model), being in March (mnth_3) increases the bike count by 0.040413 units, being in May (mnth_5) increases it by 0.071054 units, being in July (mnth_7) decreases it by 0.009695 units, being in September (mnth_9) increases it by 0.071267 units, and being in October (mnth_10) increases it by 0.119027 units. This shows that bike usage varies across the months, with the highest increase in October.

Weather Situation (weathersit_2, -0.080676; weathersit_3, -0.272362): Compared to the baseline weather situation (weathersit_1, clear or partly cloudy), being in weathersit_2 (mist + cloud) decreases the bike count by 0.080676 units, and being in weathersit_3 (light rain or snow) decreases it by 0.272362 units. This indicates that adverse weather conditions significantly reduce bike usage.

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

Ans:

Avoiding Multicollinearity:

Multicollinearity occurs when predictor variables in a regression model are highly correlated, which can make it difficult to determine the individual effect of each predictor. When all categories of a categorical variable are encoded into separate dummy variables without dropping one, they become linearly dependent (i.e., one dummy variable can be perfectly predicted from the others).

Simpler and More Interpretable Models:

Dropping the first dummy variable can lead to simpler models because it reduces the number of predictors. This can make models easier to interpret and understand. The coefficients of the remaining dummy variables will represent the difference in the outcome variable between each category and the dropped category, which serves as a baseline. This can make interpretation more straightforward.

Efficiency: Using fewer dummy variables can reduce computational cost and memory usage, which can be important when dealing with large datasets.

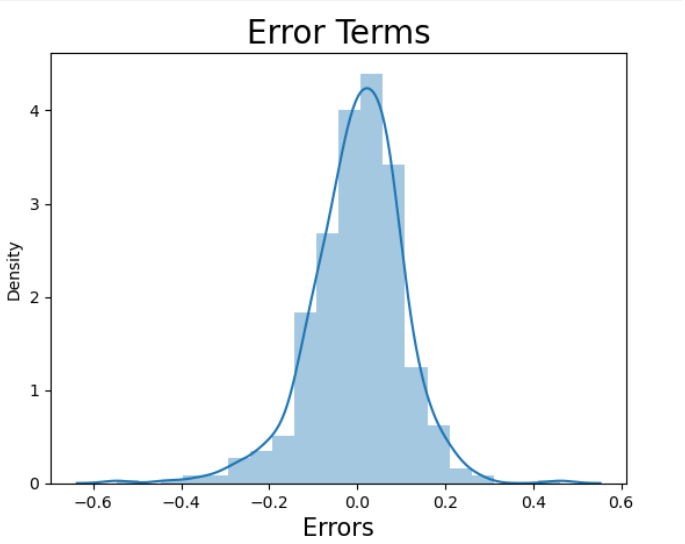
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: By looking at the pair-plot, 'TEMP' has the highest correlation among the other numerical Variables with the 'CNT' as the target variable.

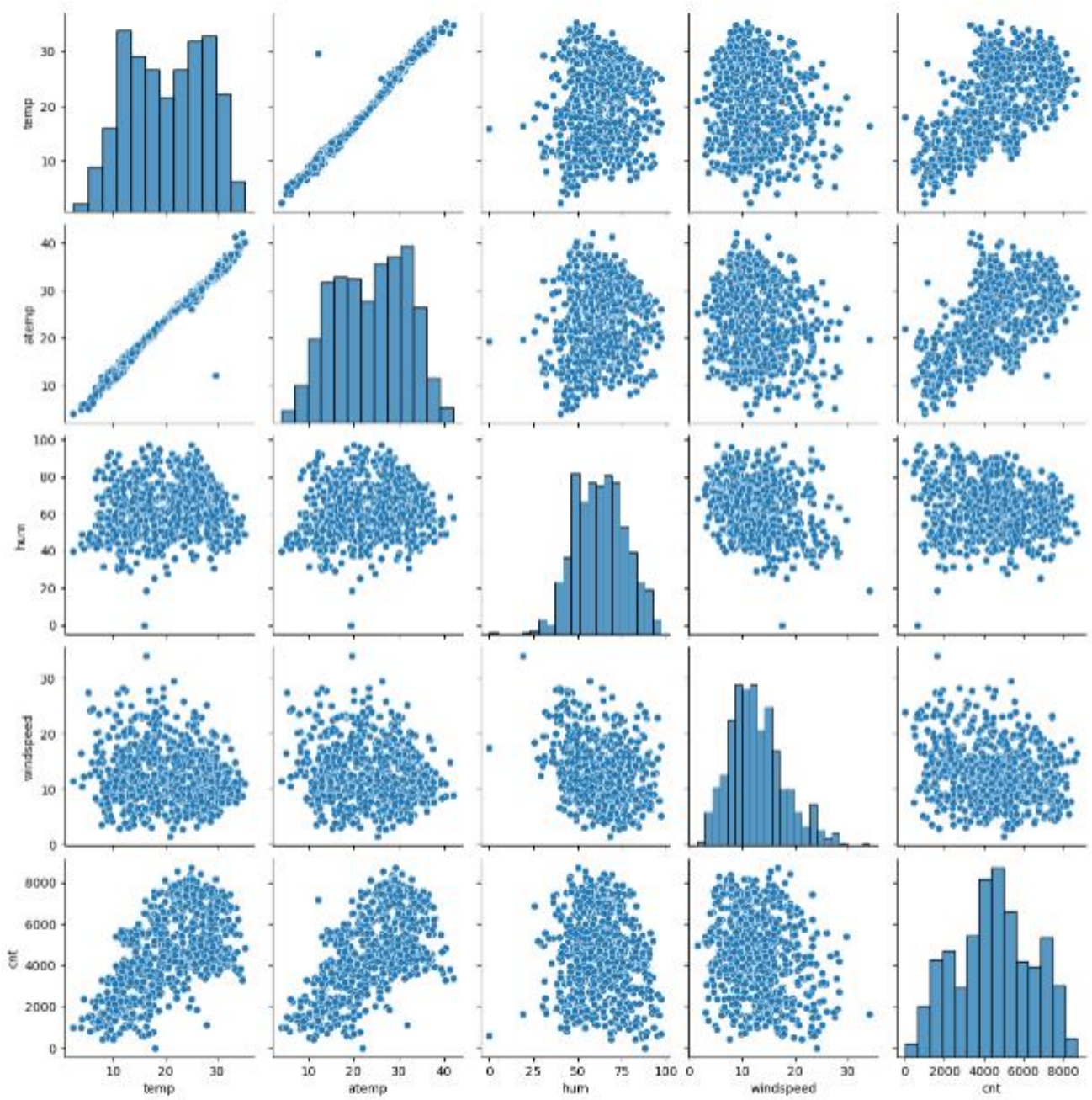
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: After building the model we can validate the assumptions of Linear Regression,

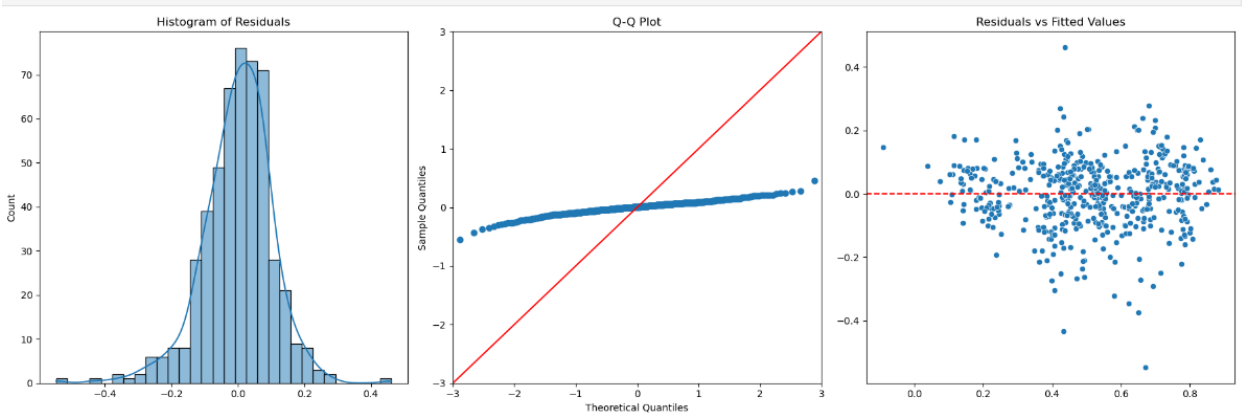
• Using histogram as we see the Residuals are normally distributed and maximum of the Error Terms are revolving around Zero.



Using the pair plot, we could see there is a linear relation between temp and atemp variable with the predictor 'cnt'.



Check for Residuals and Assumptions:



Histogram of Residuals:

The histogram of residuals appears to be roughly bell-shaped and symmetric around zero, which suggests that the residuals are approximately normally distributed. This is a good indication that the assumption of normality is reasonably met.

Q-Q Plot:

The Q-Q plot shows some deviation from the line, especially in the tails. This indicates that there might be some issues with normality, particularly in the tails. However, given the large sample size, slight deviations can be tolerated.

Residuals vs. Fitted Values:

The residuals vs. fitted values plot shows a random scatter around the horizontal line at zero, which is a good sign that there is no clear pattern. This indicates that the assumptions of linearity and homoscedasticity (constant variance) are reasonably met.

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

Ans: As per the Final Model, the top 3 Predictor Variables that are needed for the prediction

purposes are:

- Weather Situation:
- Season
- Year

weathersit_3 (light rain or snow) decreases it by 0.272362 units. This indicates that adverse weather conditions significantly reduce bike usage

Season (season_2, 0.245523; season_3, 0.323927; season_4, 0.186103):Compared to the baseline season (season_1, which is not included in the model), being in season_2 (spring) increases the bike count by 0.245523 units, being in season_3 (summer) increases it by 0.323927 units, and being in season_4 (fall) increases it by 0.186103 units. This indicates that bike usage is highest in summer, followed by spring and fall, compared to winter.

Year (yr, 0.238304):For each additional year, the bike count increases by 0.238304 units, holding all other variables constant. This indicates that bike usage has been increasing over time.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In its simplest form (simple linear regression), it uses a linear equation

$$y=\beta 0+\beta 1x+\epsilon$$

to predict the dependent variable y from the independent variable x. Multiple linear regression extends this to multiple predictors:

$$y=\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta pxp+\epsilon.$$

Key assumptions include linearity, independence of observations, homoscedasticity (constant variance of errors), normality of residuals, and absence of multicollinearity among predictors. The model is typically fitted using Ordinary Least Squares (OLS), which minimizes the sum of the squared differences between observed and predicted values.

For larger datasets, gradient descent can be used to iteratively adjust coefficients by minimizing the cost function. Model performance is evaluated using metrics such as R-squared (proportion of variance explained), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).Linear regression is valued for its simplicity and interpretability, where the coefficients indicate the strength and direction of the relationship between predictors and the outcome. However, it assumes a linear relationship, is sensitive to outliers, and can suffer from multicollinearity, which can affect coefficient estimates.

Despite its limitations, linear regression remains a foundational tool in predictive modeling, enabling both prediction and inference about relationships between variables.

Q2. Explain the Anscombe's quartet in detail.

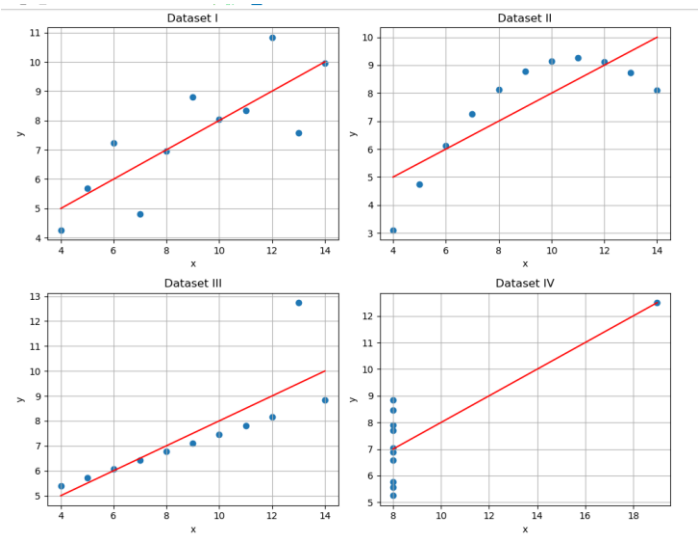
Ans: Anscombe's quartet consists of four datasets, each designed to have nearly identical simple statistical properties but very different distributions and visual characteristics. Created by statistician Francis Anscombe in 1973, the quartet illustrates the importance of graphical analysis of data.

The Four Datasets

Each dataset consists of 11 (x, y) points, with the following shared statistical properties:

- Mean of x is 9.
- Mean of y is 7.5.
- Variance of x is 11.
- Variance of y is approximately 4.122.
- Correlation between x and yy is 0.816.
- The linear regression line is y=3+0.5x

Graphical Illustrations



The provided graph visually demonstrates the differences between the datasets:

1. **Dataset I**:
    - Exhibits a typical linear relationship with points scattered around the regression line.
2. **Dataset II**:
    - Shows a clear non-linear, parabolic relationship. Despite the high correlation, the regression line poorly represents the data.

3. **Dataset III**:
    - Displays a linear relationship affected by a single outlier. This outlier influences the correlation and regression line, which would otherwise fit the majority of the points better.
4. **Dataset IV**:
    - Features a dataset with most points forming a vertical line, with one significant outlier that drives the correlation and regression line.

Importance of Anscombe's Quartet

Anscombe's quartet emphasizes several critical points in data analysis:

1. **Graph Data**: Visual inspections can reveal patterns and anomalies that summary statistics might miss.
2. **Outliers**: Outliers can heavily influence statistical analyses, including correlation and regression.
3. **Context Matters**: Statistical summaries alone do not provide a complete understanding of data. Contextual and visual exploration is essential.
4. **Misleading Statistics**: Identical statistical properties do not guarantee similar data distributions or relationships.

Anscombe's quartet is a powerful reminder of the importance of graphical analysis in data interpretation. While statistical summaries provide valuable information, visualizing data can uncover insights that purely numerical analysis might obscure. This approach ensures a more comprehensive understanding of data, leading to more accurate and reliable conclusions.

Q3. What is Pearson's R?

Ans: Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that evaluates the strength and direction of the linear relationship between two continuous variables. Named after Karl Pearson, this coefficient is widely used in statistics and various fields such as finance, psychology, and the social sciences to determine how closely two variables move in relation to each other.

Definition and Calculation

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. Mathematically, it is expressed as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where:

- Xi and Yi are the individual sample points,
- X^- and Y^- are the means of the X and Y variables, respectively.

Properties of Pearson's R

1. **Range**: Pearson's R ranges from -1 to 1.
    - r=1: Perfect positive linear relationship, meaning as one variable increases, the other variable also increases proportionally.

- r=−1: Perfect negative linear relationship, meaning as one variable increases, the other variable decreases proportionally.
- r=0: No linear relationship between the variables.

2. **Symmetry**: The coefficient is symmetric, meaning the correlation between X and Y is the same as between Y and X.
3. **Unitless**: Pearson's R is a dimensionless index, meaning it does not depend on the units of the variables being analyzed.

Interpretation

- **Strength**: The absolute value of Pearson's R indicates the strength of the linear relationship.
  - 0.0≤|r|<0.30: Weak correlation
  - 0.3≤|r|<0.7: Moderate correlation
  - 0.7≤|r|≤1.0: Strong correlation
- **Direction**: The sign of Pearson's R indicates the direction of the relationship.
  - Positive (r>0): As one variable increases, the other also increases.
  - Negative (r<0): As one variable increases, the other decreases.

Limitations

1. **Linearity**: Pearson's R only measures linear relationships. It may not accurately describe non-linear relationships.
2. **Outliers**: Sensitive to outliers, which can distort the correlation.
3. **Assumptions**: Assumes that the variables are normally distributed and the relationship is homoscedastic (equal variance of errors).

Application

Pearson's R is useful in various scenarios, such as:

- Determining the correlation between study hours and exam scores.
- Analyzing the relationship between income and expenditure.
- Investigating the connection between temperature and electricity usage.

In conclusion, Pearson's R is a fundamental statistical tool for assessing linear relationships between variables. However, it should be used with an understanding of its limitations and in conjunction with other analyses to ensure a comprehensive understanding of the data.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is a data preprocessing technique used to adjust the range and distribution of numerical data. In many machine learning algorithms, the range of input data can significantly impact performance, making scaling an essential step. By scaling, the goal is to ensure that each feature contributes equally to the model, avoiding dominance by features with larger ranges or magnitudes.

Why is Scaling Performed?

1. **Algorithm Efficiency**: Some algorithms, such as gradient descent-based methods, converge faster when features are scaled. Without scaling, the optimization process can become inefficient due to the varying scales of the features.
2. **Model Performance**: Algorithms like k-nearest neighbors (KNN) and support vector machines (SVM) are sensitive to the distances between data points. Features on different scales can distort these distances, leading to suboptimal model performance.
3. **Improved Interpretability**: Scaling makes it easier to compare the relative importance of different features. For instance, coefficients in linear regression models become more interpretable when features are on similar scales.
4. **Prevent Numerical Issues**: Large feature values can cause numerical instability in algorithms, leading to poor performance or even failure to converge.

Types of Scaling: Normalization and Standardization

Scaling methods can be broadly classified into normalization and standardization.

Normalization (Min-Max Scaling)

Normalization, also known as min-max scaling, rescales the data to a fixed range, typically [0, 1]. This method adjusts the values so that the minimum and maximum values of the feature correspond to 0 and 1, respectively.

Formula: $X'=X-Xmin/Xmax-Xmin$

Where:

- X is the original value.
- $Xmin$ and $Xmax$ are the minimum and maximum values of the feature.

Advantages:

- Simple to implement and interpret.
- Maintains the relationships between the original data values.

Disadvantages:

- Sensitive to outliers since they can skew the scaling range.
- Does not handle features with different distributions well.

When to Use:

- When the data needs to be bounded within a specific range.
- Suitable for algorithms that do not assume any particular distribution of the data, such as KNN and neural networks.

Standardization (Z-Score Scaling)

Standardization, also known as z-score scaling, transforms the data to have a mean of zero and a standard deviation of one. This method rescales the feature such that it has a mean ($\mu\mu$) of 0 and a standard deviation ($\sigma\sigma$) of 1.

Formula: $X'=X-\mu/\sigma$

Where:

- X is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

Advantages:

- Less sensitive to outliers compared to normalization.
- Suitable for data that follows a Gaussian distribution.

Disadvantages:

- Does not bound the values within a fixed range.
- The presence of outliers can still affect the mean and standard deviation.

When to Use:

- When the algorithm assumes a Gaussian distribution of the data, such as linear regression, logistic regression, and principal component analysis (PCA).
- When dealing with features that have different scales but need to be compared or combined.

Differences Between Normalization and Standardization

1. **Range**:
    - **Normalization**: Scales data to a fixed range, typically [0, 1].
    - **Standardization**: Scales data to have a mean of 0 and a standard deviation of 1, without bounding the range.
2. **Sensitivity to Outliers**:
    - **Normalization**: More sensitive to outliers, which can skew the min-max range.
    - **Standardization**: Less sensitive to outliers, although extreme values can still influence the mean and standard deviation.
3. **Usage Context**:
    - **Normalization**: Preferred when the algorithm is distance-based or when features need to be bounded.
    - **Standardization**: Preferred when the algorithm assumes normally distributed data or requires features to have comparable scales.

Scaling is a critical preprocessing step in machine learning that ensures fair contribution of features to the model, improves algorithm efficiency, and enhances model performance. Normalization and standardization are two primary scaling methods, each with its advantages, disadvantages, and appropriate use cases. Understanding when and how to apply these techniques is essential for developing robust and accurate machine learning models.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.For example, we would fit the following models to estimate the coefficient of determination R1

and use this value to estimate the VIF:

$X_1=C+ \alpha_2 X_2+\alpha_3 X_3+\cdots$

〖VIF〗$_1=1/(1-R_1^2)$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$X_2=C+ \alpha_1 X_1+\alpha_3 X_3+\cdots$

〖VIF〗$_2=1/(1-R_2^2)$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations

with same distributions.

Few advantages:

a) It can be used with sample sizes also.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

If two data sets —

• Come from populations with a common distribution.

• Have common location and scale.

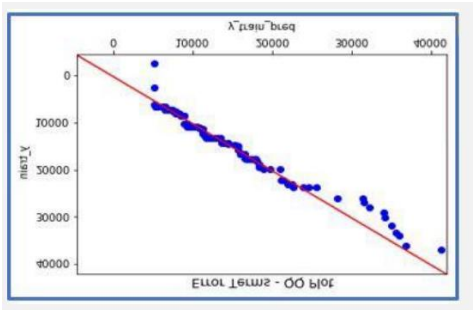
• Have similar distributional shapes.

• Have similar tail behavior.

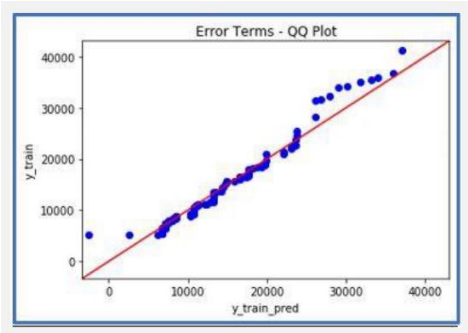
Below are the possible regressions for two data sets:

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis