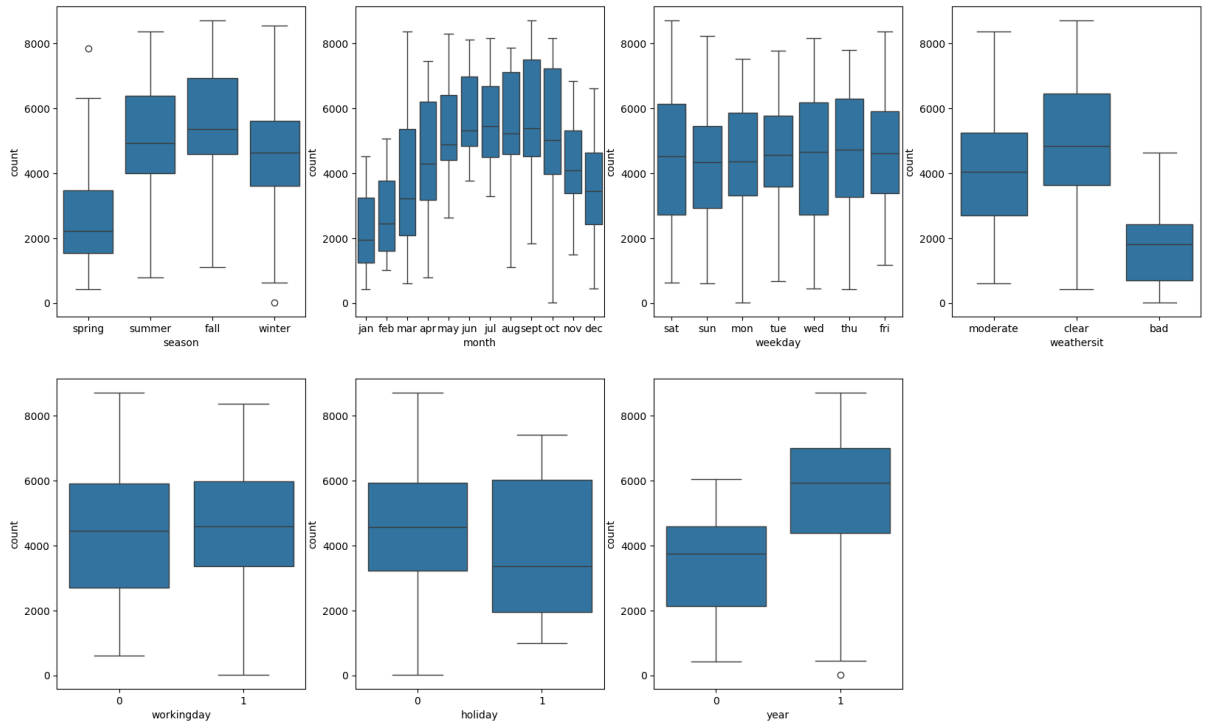
By : Maimana Kowatly

Linear regression assignment

Assignment-based Subjective Answers:

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

Answer:
Analysis of Categorical Columns:
Seasonal Trends: The fall season saw the highest number of bookings, with a significant increase in bookings from 2018 to 2019 across all seasons.
Monthly Trends: The majority of bookings occurred from May to October, with a trend of increasing bookings from the start of the year until mid-year, followed by a decline towards the year's end.
Weather Influence: Clear weather led to more bookings.
Day of the Week: Bookings were higher on Thursdays, Fridays, Saturdays, and Sundays compared to the beginning of the week.
Holiday Impact: Fewer bookings were made when it was a holiday, people prefer to stay home and spend time with family on holidays.
Workdays vs. Non-Workdays: Bookings were almost equal on working and non-working days.
Yearly Comparison: There were more bookings in 2019 than in the previous year, indicating business growth.



2) Why is it important to use drop_first=True during dummy variable creation?
Answer:

Importance of Using drop_first=True in Dummy Variable Creation
Using drop_first=True when creating dummy variables is essential because it helps prevent the creation of redundant columns, thereby reducing multicollinearity among the dummy variables. Here's a detailed explanation:

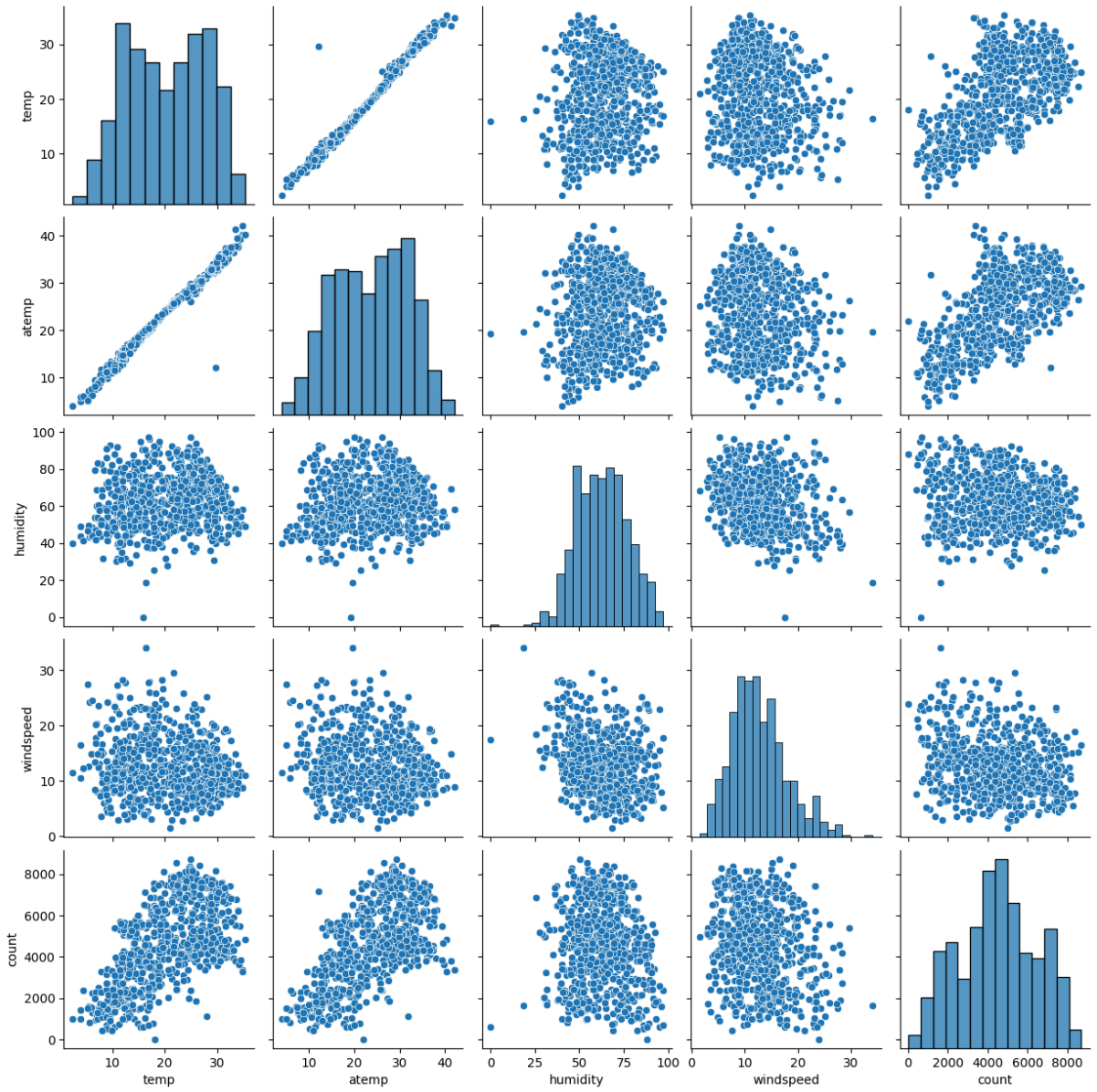
Explanation
When converting categorical variables into dummy variables (one-hot encoding), it's common to generate a separate binary column for each category. However, including a dummy variable for every category introduces redundancy into the dataset. This redundancy can lead to issues with certain statistical models, especially those that are sensitive to multicollinearity, such as linear regression.

Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a significant degree of accuracy. This can cause unstable estimates and make it difficult to assess the impact of each predictor.

In our case study: (bike sharing system)
Dummies created for:

- Season
- Month
- Weekdays
- Weather

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

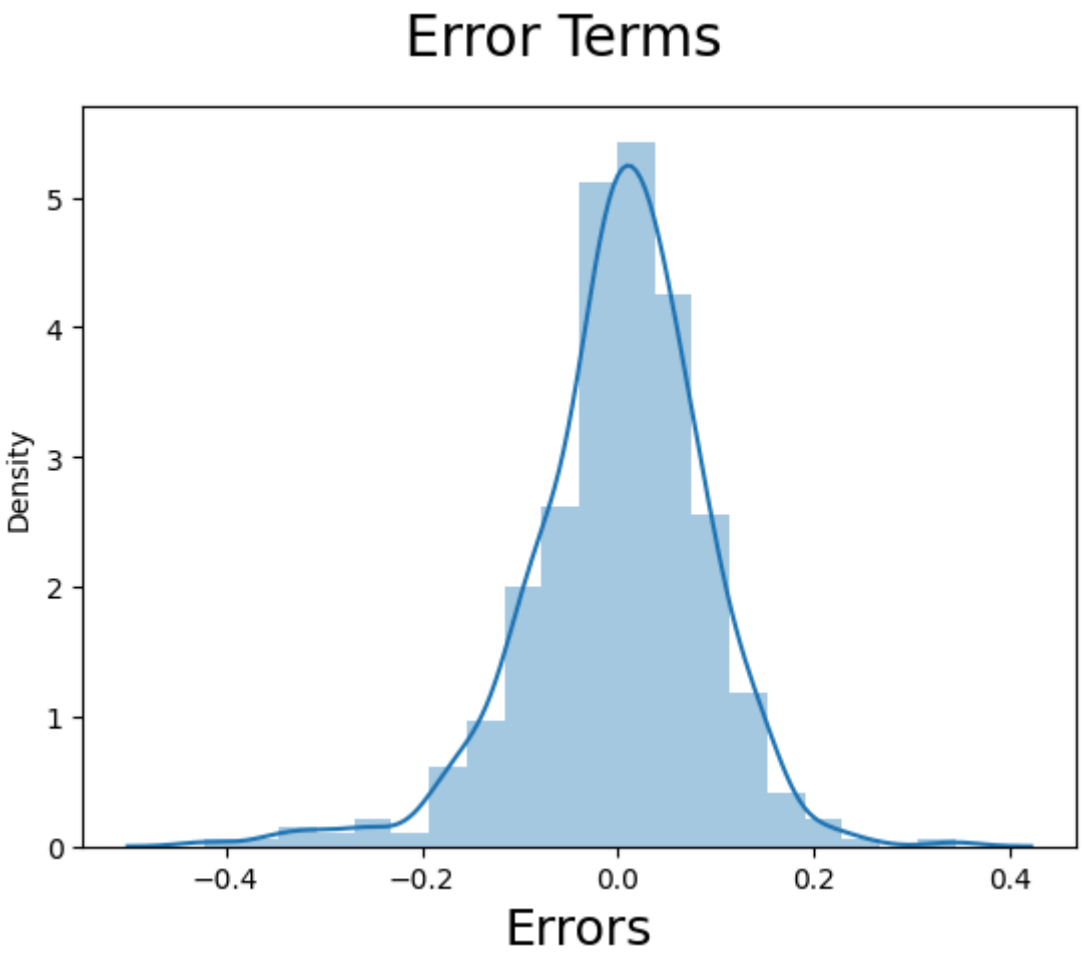


The temp and ‘atemp’ column variables have the highest correlation when compared to the rest with the target variable as “cnt”’.

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

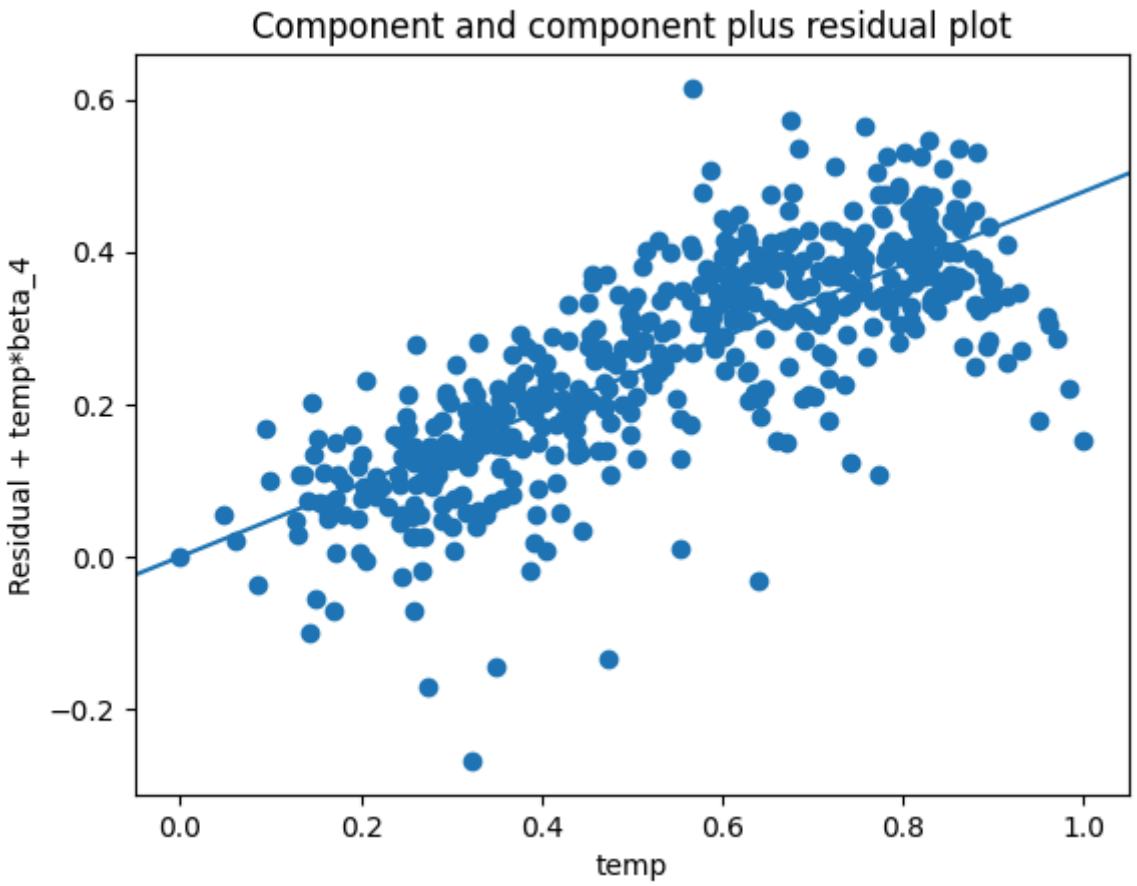
- Normality of Error Terms:

  Error terms should follow a normal distribution.



- Multicollinearity Check:
  There should be minimal multicollinearity among variables.

- Linear Relationship Validation:
  There should be a clear linear relationship among the variables.



- Homoscedasticity:
  Residual values should not show any discernible pattern.
- Independence of Residuals:
  There should be no autocorrelation among residuals.

Durbin-Watson: 2.101 which signifies there is no autocorrelation.

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

Answer:
Temperature
Season
Month

General Subjective Questions

1) Explain the linear regression algorithm in detail.
Answer:
Detailed Explanation of the Linear Regression Algorithm
Linear Regression is one of the most fundamental and widely used algorithms in statistics and machine learning. It aims to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to observed data.

1. The Linear Regression Model

In its simplest form, the linear regression model can be represented as:

$$y = \beta0 + \beta1x1 + \beta2x2 + \ldots + \beta nxn + \epsilon$$

Where:

- $y\ is\ the\ dependent\ variable\ (target)$.
- $x1, x2, \ldots, xnare\ the\ independent\ variables\ (features)$.
- $\beta0 is\ the\ intercept\ of\ the\ regression\ line\ (the\ value\ of yy\ when\ allxx's\ are\ 0)$.
- $\beta1, \beta2, \ldots, \beta nare\ the\ coefficients\ (slopes)\ for\ the\ respective\ features$.
- $\epsilon\ is\ the\ error\ term\ (the\ difference\ between\ the\ predicted\ and\ actual\ values)$.

2. Assumptions of Linear Regression

- Linearity: The relationship between the independent and dependent variables should be linear.
- Independence: Observations should be independent of each other.
- Homoscedasticity: The residuals (errors) should have constant variance.
- Normality: The residuals should be normally distributed.
- No Multicollinearity: Independent variables should not be highly correlated with each other.

3. Ordinary Least Squares (OLS) Method

The most common method to fit a linear regression model is the Ordinary Least Squares (OLS) method. OLS estimates the parameters (coefficients) by minimising the sum of the squared differences between the observed and predicted values:

Minimise $\sum i = 1n(yi - y\hat{}i)2$
Where yi is the actual value. $y\hat{}\ i\ .\ is\ the\ predicted\ value$

is the predicted value.

4. Steps in Linear Regression Analysis

- Data Preparation:
  - Collect and clean the data.
  - Handle missing values.

    - Encode categorical variables (if any).
    - Split the data into training and test sets.
- Model Training:
    - Fit the linear regression model using the training data.
    - Use the OLS method to estimate the coefficients.
- Model Evaluation:
    - Evaluate the model using various metrics like R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc.
    - Check the assumptions of the model (linearity, independence, homoscedasticity, normality, multicollinearity).
- Model Interpretation:
    - Interpret the coefficients to understand the relationship between the independent and dependent variables.
    - Assess the significance of each predictor variable.

- Prediction:
    - Use the fitted model to make predictions on new data.

5. Metrics for Model Evaluation:

- R-squared (Coefficient of Determination):

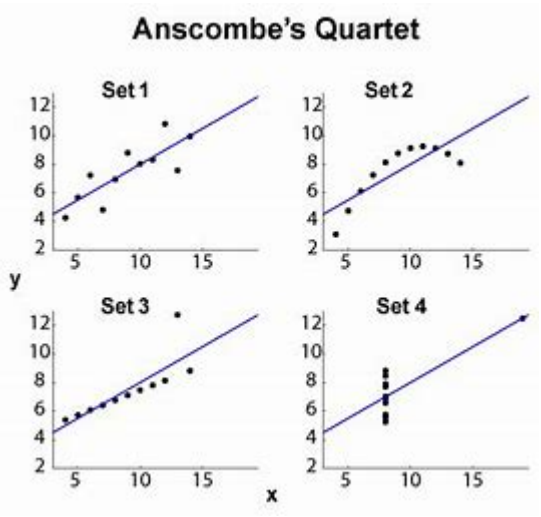
    1. Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.
    2. Ranges from 0 to 1, with higher values indicating better model fit.
- Mean Squared Error (MSE):

    Measures the average of the squares of the errors (the average squared difference between the observed actual outcomes and the outcomes predicted by the model).
- Root Mean Squared Error (RMSE):
  The square root of the MSE. Provides a measure of the average magnitude of the errors in the same units as the dependent variable.
- Mean Absolute Error (MAE):
  Measures the average magnitude of the errors in a set of predictions, without considering their direction.

Linear regression is a powerful tool for understanding and predicting the relationship between variables. By meeting its assumptions and carefully interpreting its results, it provides valuable insights that can drive data-driven decision-making.

2. Explain the Anscombe's quartet in detail.
Answer:

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. These datasets were constructed by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analysing it and to demonstrate the effect of outliers and the limitations of simple statistical measures.



Components of Anscombe's Quartet:

Each dataset in Anscombe's Quartet consists of eleven data points (x, y pairs) and shares the following properties:

- Identical mean of x: The mean of the x-values in each dataset is the same.
- Identical mean of y: The mean of the y-values in each dataset is the same.
- Identical variance of x: The variance of the x-values in each dataset is the same.
- Identical variance of y: The variance of the y-values in each dataset is the same.

- Identical correlation between x and y: The correlation coefficient between the x and y values is the same.
- Identical linear regression line: The linear regression line (y = mx + b) is nearly identical in each case.

Despite these similarities, the datasets differ significantly when visualised graphically.

The Four Datasets

Here are the four datasets (x, y pairs) in Anscombe's Quartet:

- ❖ Dataset I:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

- ❖ Dataset II:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

- ❖ Dataset III:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

- ❖ Dataset IV:

x: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 19
y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 5.56, 7.91, 6.89, 12.50

Common Statistical Properties:

- Mean of x: 9
- Mean of y: 7.5
- Variance of x: 11
- Variance of y: 4.125
- Correlation between x and y: ~0.816
- Linear regression line: y = 3 + 0.5x

Visualisation and Interpretation:
When visualised, the differences between the datasets become clear:

- ❖ Dataset I:
  Appears as a roughly linear relationship with some scatter around the regression line.
- ❖ Dataset II:
  Displays a clear nonlinear relationship. A curve might fit the data better than a straight line.
- ❖ Dataset III:

Has a linear relationship but with an obvious outlier that greatly influences the regression line.

- ❖ Dataset IV:

Consists mostly of identical x-values with one significant outlier, which skews the regression line.

Importance and Lessons:

- Graphical Analysis: Always visualise your data. Simple statistical summaries do not capture all aspects of the data distribution and relationships.
- Outliers: Be aware of the influence of outliers. They can drastically affect statistical measures like the mean and regression coefficients.
- Nonlinearity: Check for non-linear relationships. Linear models may not always be appropriate.
- Data Distribution: Descriptive statistics (mean, variance, correlation) may be identical, yet the underlying data can be fundamentally different.

3) what is Pearson R:
Answer:

Pearson's R: An Overview

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the degree to which a relationship between two variables can be described using a straight line.

Definition and Formula

The Pearson correlation coefficient, denoted as r, is defined as the covariance of the two variables divided by the product of their standard deviations. Mathematically, it is represented as:

$$r = \sigma X \sigma Y cov(X, Y)$$

Where:

$cov(X, Y)\ is\ the\ covariance\ between\ variables XX\ and YY.$

$\sigma X is\ the\ standard\ deviation\ of\ variable XX.$

$\sigma Y is\ the\ standard\ deviation\ of\ variable YY.$

The formula for calculating r from a sample of data is:

$$r = \sum i = 1n(xi - x^{-})2\sum i = 1n(yi - y^{-})2\sum i = 1n(xi - x^{-})(yi - y^{-})$$

Where:

$n\ is\ the\ number\ of\ data\ points.$

$xi\ and\ yi\ are\ the\ individual\ sample\ points.$

$x^{-} and y^{-}\ are\ the\ means\ of\ the\ x\ and\ y\ values,\ respectively.$

Interpretation of Pearson's R

The value of r ranges from -1 to 1:

- r=1: Perfect positive linear correlation. As one variable increases, the other variable also increases proportionally.

- $r=-1$: Perfect negative linear correlation. As one variable increases, the other variable decreases proportionally.
- $r=0$: No linear correlation. There is no linear relationship between the variables.

Strength of the Correlation

The strength of the correlation can be interpreted as follows:

- 0.0 to 0.1: No correlation.
- 0.1 to 0.3: Small (weak) correlation.
- 0.3 to 0.5: Medium (moderate) correlation.
- 0.5 to 1.0: Large (strong) correlation.

Assumptions of Pearson's R

For Pearson's R to be a valid measure of correlation, certain assumptions need to be met:

1. Linearity: The relationship between the two variables should be linear.
2. Continuous Data: Both variables should be continuous.
3. Normality: The variables should be approximately normally distributed (especially in small samples).
4. Homoscedasticity: The variability of one variable should be constant across the range of the other variable.
5. Independence: Observations should be independent of each other.

---

4) What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Answer:

Scaling is a data preprocessing technique used to adjust the values of numerical features to a common scale without distorting differences in the ranges of values. It is essential in many machine learning algorithms to ensure that all features

contribute equally to the result and to improve the performance and training stability of the model.

Why Scaling is Performed

1. Equal Contribution: In many algorithms, like gradient descent-based methods, features with larger ranges can dominate the learning process. Scaling ensures all features contribute equally.
2. Improved Convergence: Algorithms like gradient descent converge faster with scaled features because the optimization landscape becomes more symmetric.
3. Improved Accuracy: Distance-based algorithms (e.g., k-nearest neighbors, SVMs) are sensitive to the range of features. Scaling improves their accuracy by ensuring that all features are treated equally.
4. Preventing Bias: Models can become biased towards features with larger ranges if data is not scaled.

Types of Scaling: Normalised Scaling vs. Standardised Scaling

1. Normalised Scaling

Normalisation rescales the values of features to a range of [0, 1] or [-1, 1]. This technique is particularly useful when you know the data distribution does not follow a Gaussian distribution and when you want to preserve the relationships between data points.

The formula for normalisation is:

$$x' = max(x) - min(x)x - min(x)$$

Where:

- x is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.
- x' is the standardised value.

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

Answer:

Infinite VIF: Understanding the Causes

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, meaning they contain redundant information.

Definition of VIF

For a given predictor

Xi, the VIF is defined as:

$VIF(Xi) = 1 - Ri21$

Where

$Ri2\ is\ the\ coefficient\ of\ determination\ of\ the\ regression\ of XiXion\ all\ other\ predictors.$

Causes of Infinite VIF

VIF values can become infinite when $Ri2$ equals 1. This happens when there is perfect multicollinearity in the dataset, meaning one predictor is an exact linear combination of one or more other predictors. Here are the common causes:

1. Exact Linear Dependence: If a predictor variable can be perfectly predicted from a linear combination of other predictor variables

   $Ri2\ will\ be\ 1.$

   $For\ example,\ if X3 = 2X1 + 3X2X3 = 2X1 + 3X2,\ then\ the\ VIF\ for X3\ will\ be\ infinite.$

   2. Duplicate Columns: If the dataset contains duplicate columns (or nearly columns with extremely high correlation), the VIF for these variables will approach infinity.

3. Dummy Variable Trap: In categorical variables encoded using one-hot encoding, including all dummy variables without dropping one (the reference category) will lead to perfect multicollinearity. This is known as the dummy variable trap. For instance, if a categorical variable with three levels is encoded into three dummy variables $D1$,$D2$, and $D3$ their sum will always be 1. Including all three in the model leads to perfect multicollinearity.

4. Numerical Precision Issues: Sometimes, numerical precision issues can cause a situation where the regression algorithm perceives perfect multicollinearity, especially when dealing with large datasets or very small/large values.

Solutions to Avoid Infinite VIF

1. Remove Redundant Variables: If predictors are exact linear combinations of others, remove one of the redundant predictors.
2. Drop One Dummy Variable: In one-hot encoding, always drop one dummy variable to avoid the dummy variable trap.
3. Principal Component Analysis (PCA): Use PCA to transform the predictors into a set of linearly uncorrelated components.
4. Regularisation: Techniques like Ridge Regression can help mitigate the effects of multicollinearity by adding a penalty to the regression coefficients.

Conclusion

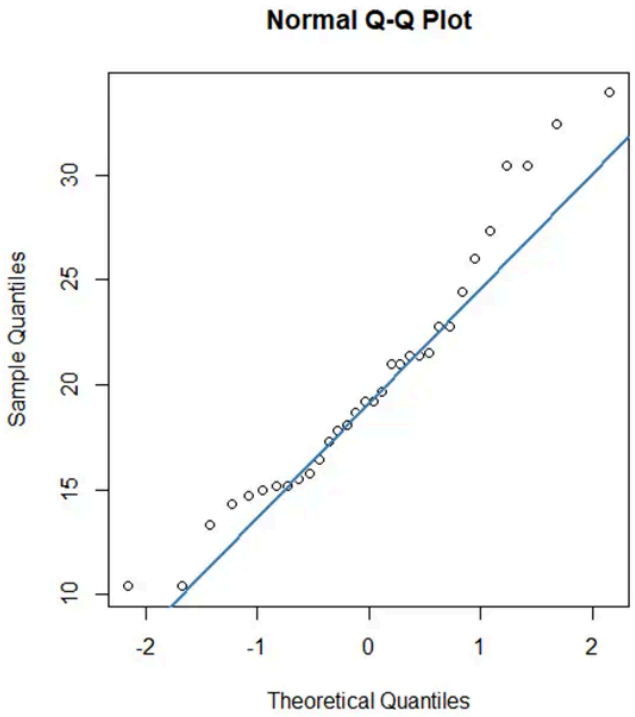
Infinite VIF indicates perfect multicollinearity, where one predictor is an exact linear combination of other predictors. Understanding and addressing the causes of multicollinearity is crucial for building reliable regression models. By carefully examining the relationships between predictors and applying appropriate solutions, one can ensure the robustness of the regression analysis.

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

Answer:

Q-Q Plot: An Overview

Q-Q Plot, or Quantile-Quantile Plot, is a graphical tool used to assess whether a set of data follows a particular theoretical distribution, such as the normal distribution. It is a plot of the quantiles of the data against the quantiles of the theoretical distribution. If the data follows the specified distribution, the points on the Q-Q plot will lie approximately along a straight line.



Components of a Q-Q Plot

1. Quantiles of the Data: These are the values that divide the data into intervals with equal probabilities.
2. Quantiles of the Theoretical Distribution: These are the values from the theoretical distribution at the same cumulative probabilities as the data quantiles.

Steps to Create a Q-Q Plot

1. Sort the Data: Arrange the observed data in ascending order.
2. Calculate Quantiles: Compute the quantiles of the observed data.
3. Determine Theoretical Quantiles: Obtain the corresponding quantiles from the theoretical distribution.
4. Plot the Quantiles: Plot the observed data quantiles on the x-axis and the theoretical quantiles on the y-axis.

Importance of Q-Q Plot in Linear Regression

In the context of linear regression, Q-Q plots are primarily used to check the assumption of normality of the residuals (error terms). The key assumptions in linear regression include:

1. Linearity: The relationship between the predictors and the response variable should be linear.
2. Independence: The residuals should be independent.
3. Homoscedasticity: The residuals should have constant variance.
4. Normality: The residuals should be approximately normally distributed.

Checking Normality of Residuals

- Normal Distribution Assumption: Many inferential statistics and diagnostic tests in linear regression, such as hypothesis tests on coefficients, rely on the assumption that the residuals are normally distributed. Non-normal residuals can indicate that the model is not capturing all the patterns in the data.
- Identifying Deviations: Q-Q plots help identify deviations from normality, such as skewness (asymmetry) or kurtosis (heavy tails or outliers). Deviations from the straight line suggest that the residuals do not follow a normal distribution.

Interpretation of a Q-Q Plot

1. Straight Line: If the points fall roughly along a 45-degree line (y=x), the data follows a normal distribution.
2. S-shaped Curve: Indicates skewness. Right skewed data will bend upwards, and left skewed data will bend downwards.

3. Upward/Downward Deviation at Ends: Indicates heavy tails or light tails in the distribution.