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?

Answer:

Below are the inferences drawn by analyzing the categorical variables:

a) Fall season seems to have attracted a greater number of bookings, while there has been a substantial increase in bookings for each season from 2018 to 2019.
b) The months of May, June, July, August, September, and October have seen the highest booking activity. There is a rising trend in bookings from the beginning of the year until the middle, followed by a gradual decline towards the end of the year.
c) Bookings are more prevalent during clear weather conditions, which is expected.
d) Thursdays, Fridays, Saturdays, and Sundays have a higher booking count compared to the earlier days of the week.
e) Non-holiday periods generally have fewer bookings, as people tend to prefer spending time at home and enjoying with their families during holidays.
f) Booking frequencies are relatively similar between working days and non-working days.
g) The year 2019 has experienced a higher number of bookings compared to the previous year, indicating positive business growth.

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

Answer:

When creating dummy variables, the parameter drop_first=True is used to exclude one of the categorical levels from the resulting dummy variable representation.

Avoiding multicollinearity: Including all the dummy variables without dropping one can lead to multicollinearity, which is a situation where two or more variables are highly correlated with each other. This can cause problems in statistical models, particularly in regression analysis, where it can affect the accuracy and interpretability of the results. By dropping one level, we ensure that the remaining dummy variables are linearly independent, reducing multicollinearity.

Uniqueness of baseline category: When creating dummy variables, one category is chosen as the baseline or reference category, against which the other categories are compared. By dropping the first category, we establish a unique baseline for comparison. This is important for correctly interpreting the coefficients of the dummy variables. The dropped category represents the omitted level, and the coefficients of the remaining dummy variables represent the difference from that baseline category.

Efficiency and simplicity: Including all dummy variables without dropping one can result in redundancy and unnecessarily increase the dimensionality of the dataset. By dropping one level, we reduce the number of variables and make the representation more efficient and easier to interpret.

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

Answer:

The 'temp' variable has the highest correlation with the target variable.

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

Answer:

I have validated the assumptions of linear regression based on the below points:

1. Normality of error terms
    - Error terms should be normally distributed
2. Multicollinearity
    - Inter feature dependency should be insignificant
3. Independence of residuals
    - There should be no visible pattern in residuals
4. Homoscedasicity
    - variability of the dependent variable is constant across different levels of the independent variable(s)

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

Answer:

Temperature, weather situation and year are the top features contributing significantly towards the demand of shared bikes.

General Subjective Questions

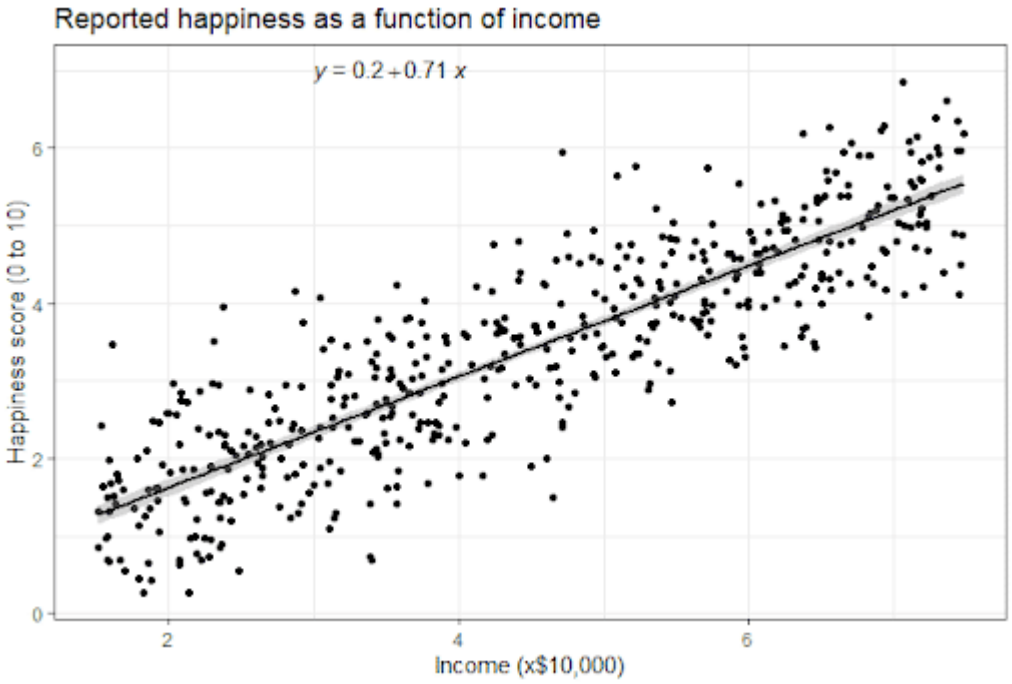
6. Explain the linear regression algorithm in detail.

Answer:
Linear regression is a widely used supervised learning algorithm for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the input variables and the target variable. The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted values and the actual values of the target variable.

In linear regression, the relationship between the input features (denoted as X) and the target variable (denoted as y) is represented by a linear equation of the form: y = b0 + b1x1 + b2x2 + ... + bnxn, where b0, b1, b2, ..., bn are the coefficients or weights associated with each input feature. The coefficient b0 represents the intercept or the value of the target variable when all input features are zero.

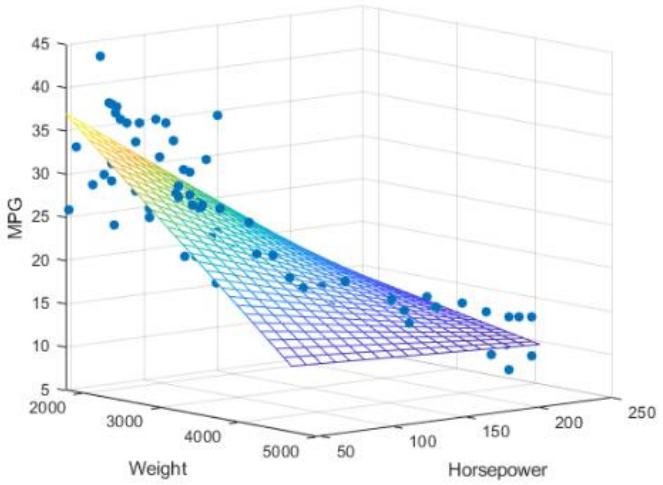
There are two types of linear regression based on the number of features:

1. Simple Linear Regression: Simple linear regression is a type of linear regression that involves predicting a continuous target variable based on a single input feature.



(Image reference: https://www.scribbr.com/wp-content/uploads//2020/02/simple-linear-regression-graph.png)

2. Multiple linear regression: Multiple linear regression is an extension of simple linear regression that involves predicting a continuous target variable based on multiple input features.



(Image reference: https://vitalflux.com/wp-content/uploads/2019/07/multilinear-regression-model.png)

Assumptions:

Below are the assumptions for a linear regression model:

1. Low or no Multicollinearity: Multicollinearity refers to a situation in multiple linear regression where two or more predictor variables are highly correlated with each other. It occurs when there is a strong linear relationship between the independent variables, making it difficult to determine the individual effects of each variable on the dependent variable. This should be very low between the predictors.

2. Linear relationship between variables: The relation between the feature variable and the output should be linear.

3. Normality of error terms: Error terms must be normally distributed.

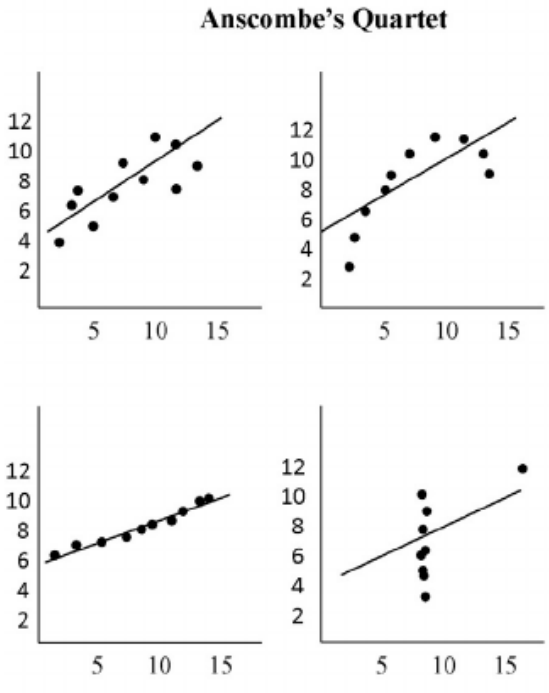
4. Homoscedasticity: Homoscedasticity, also known as homogeneity of variance, is a statistical concept that relates to the assumption of equal variance of errors or residuals in a regression model. It means that the variability of the residuals is consistent across different levels or ranges of the independent variables.

7. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a set of four datasets, each consisting of 11 x-y data points, created by the statistician Francis Anscombe in 1973. The quartet is designed to demonstrate the limitations of relying solely on summary statistics such as mean, variance, and correlation to

understand and analyze data. Despite having similar summary statistics, the four datasets exhibit distinct patterns and relationships when graphed, highlighting the importance of visualizing data.



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

(Image reference: https://www.researchgate.net/publication/285672900/figure/fig4/AS:305089983074309@1449750528742/Anscombes-quartet-of-different-XY-plots-of-four-data-sets-having-identical-averages.png)

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 I represents a linear relationship between x and y with a positive slope. When plotted, the data points closely align along a straight line, indicating a strong linear relationship.

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 II also represents a linear relationship, but with a slight curvature. It highlights the effect of an outlier, where a single data point significantly deviates from the general trend.

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 III shows a non-linear relationship, resembling a quadratic curve. The pattern is not apparent when relying solely on summary statistics.

Dataset IV:

x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
Dataset IV highlights the impact of an outlier, as a single data point drastically influences the summary statistics. The outlier pushes the regression line upward, indicating a relationship that doesn't accurately represent the majority of the data.

Anscombe's quartet serves as a cautionary example, emphasizing the importance of visually examining data. Although the summary statistics of the four datasets are nearly identical, their graphical representations reveal distinct patterns. It illustrates that relying solely on numerical summaries can overlook important details and lead to incorrect interpretations.

By showcasing different relationships, linearity, outliers, and curvature, Anscombe's quartet underscores the significance of data visualization in understanding and analyzing data. It reminds statisticians and data analysts to complement summary statistics with visualizations to gain a comprehensive understanding of the data at hand.

8. What is Pearson’s R?

Answer:

Pearson's R, also known as Pearson correlation coefficient or Pearson's product-moment correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It measures the degree of association between the variables on a scale ranging from -1 to 1.

Few key points about Pearson's R:

Range and interpretation: Pearson's R ranges from -1 to 1. A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases in a consistent manner. A value of 1 indicates a perfect positive linear relationship, where both variables increase or decrease together. A value of 0 indicates no linear relationship between the variables.

Linear relationship: Pearson's R specifically measures the linear association between variables. It assumes that a straight line can approximate the relationship between the variables. If the relationship is nonlinear, Pearson's R may not accurately capture the association.

Symmetry: Pearson's R is symmetric, meaning that the correlation between variable X and variable Y is the same as the correlation between variable Y and variable X. The order

of the variables does not affect the magnitude or interpretation of the correlation coefficient.

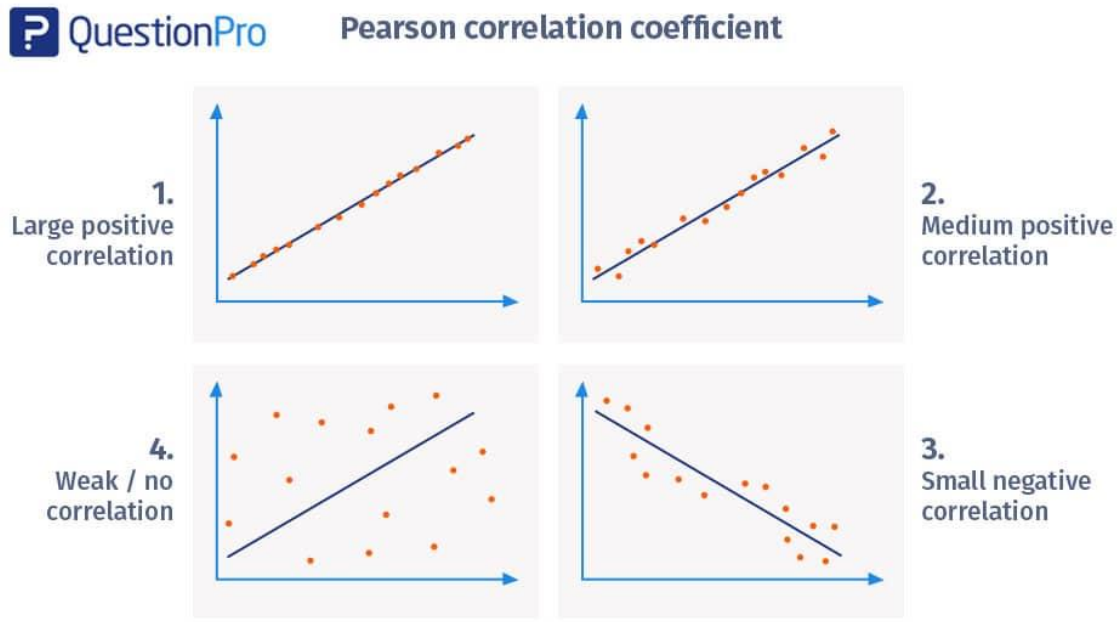
Strength of association: The magnitude of Pearson's R indicates the strength of the association between the variables. A value close to -1 or 1 suggests a strong linear relationship, while a value closer to 0 indicates a weaker association.

Interpretation of magnitude: There is no definitive threshold for determining what constitutes a "strong" or "weak" correlation, as it can depend on the context and field of study. However, commonly used guidelines consider correlations around ±0.3 to ±0.5 as moderate, ±0.5 to ±0.7 as strong, and above ±0.7 as very strong.

Statistical significance: In addition to the magnitude, the statistical significance of Pearson's R is important. Hypothesis tests can determine if the observed correlation coefficient is significantly different from zero, indicating a meaningful relationship between the variables.

Assumptions: Pearson's R assumes that the relationship between the variables is approximately linear, the variables follow a bivariate normal distribution, and there are no outliers or influential observations. Violations of these assumptions can impact the accuracy and validity of the correlation coefficient.

Limitations: Pearson's R measures only the linear relationship between variables and does not capture other types of relationships (e.g., nonlinear, curvilinear). It is sensitive to outliers and can be influenced by extreme observations.



(Image reference: https://www.questionpro.com/blog/wp content/uploads/2020/04/Pearson-correlation-coefficient-1.jpg)

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

Answer:

Scaling, also known as feature scaling or data normalization, is a preprocessing technique used to transform the values of different variables in a dataset to a standardized range. It ensures that all variables are on a similar scale, allowing for fair comparisons and avoiding the dominance of certain variables over others in data analysis and modeling.

Few of the key points about scaling:

Purpose of scaling: Scaling is performed to bring all variables to a similar scale, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. It is necessary when variables have different measurement units, scales, or ranges. Scaling helps in comparing variables with different units and prevents one variable from dominating or biasing the analysis based on its larger value range.

Normalized scaling: Normalization scales the values of a variable to a range between 0 and 1. It is achieved by subtracting the minimum value of the variable from each data point and dividing it by the range (maximum value minus minimum value). Normalization preserves the relative relationships and proportions between data points, but the distribution shape may be affected.

Standardized scaling: Standardization scales the values of a variable to have a mean of 0 and a standard deviation of 1. It involves subtracting the mean of the variable from each data point and dividing it by the standard deviation. Standardization transforms the data to have a standard normal distribution with a mean of 0 and a standard deviation of 1. It preserves the shape of the distribution and the relative distances between data points.

Differences between normalization and standardization:

Range: Normalization scales the values to a range between 0 and 1, while standardization rescales the values to have a mean of 0 and a standard deviation of 1.
Distribution shape: Normalization may alter the shape of the distribution, while standardization maintains the shape of the original distribution.
Outliers: Normalization is sensitive to outliers, as it relies on the minimum and maximum values. Standardization is more robust to outliers, as it uses the mean and standard deviation.
Interpretation: Normalization preserves the original units and can be useful when maintaining the interpretability of the data is important. Standardization transforms the data to z-scores, making it easier to compare variables with different units.

The choice between normalization and standardization depends on the specific requirements of the analysis or modeling task. Normalization is commonly used in algorithms that require input values within a specific range, such as neural networks. Standardization is often preferred when maintaining the distribution shape and handling outliers is important, such as in many statistical analyses and machine learning algorithms.

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

Answer:

In certain cases, the VIF value can be infinite, indicating a perfect linear relationship between the predictor variables. This occurs when one or more variables can be expressed as a perfect linear combination of the other variables in the model. As a result, the coefficient estimates for these variables become indeterminate, leading to infinite VIF values.

Here are some key reasons why the VIF can be infinite:

Perfect multicollinearity: Perfect multicollinearity arises when there is an exact linear relationship between two or more predictor variables. This means that one variable can be perfectly predicted using a linear combination of the other variables. In this scenario, the regression model cannot estimate the unique contribution of each variable, and the VIF becomes infinite.

Duplicated or redundant variables: If there are duplicated or redundant variables in the dataset, where two or more variables contain identical or nearly identical information, it can result in infinite VIF values. These redundant variables introduce perfect multicollinearity, making it impossible for the model to estimate their coefficients.

It is important to note that infinite VIF values are not desirable and indicate a problem in the regression analysis. They can lead to unreliable coefficient estimates and affect the interpretation and inference of the model.

To address infinite VIF values, it is necessary to identify and resolve the perfect multicollinearity issue. This typically involves examining the variable relationships, identifying redundant variables, and either removing one of the correlated variables or transforming the variables to remove the linear dependence.

Overall, the occurrence of infinite VIF values highlights the importance of identifying and addressing multicollinearity issues in regression analysis to ensure the accuracy and reliability of the results.

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

Answer:

A Q-Q plot, short for 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 observed data against the quantiles expected under a particular distribution, such as the normal distribution. The Q-Q plot allows visual inspection of whether the data deviates from the expected distribution and provides insights into the goodness-of-fit between the data and the assumed distribution.

Beow are few of the uses and importance of a Q-Q plot in linear regression:

Distributional assessment: In linear regression, it is often assumed that the residuals (the differences between the observed and predicted values) follow a normal distribution. The Q-Q plot helps assess the validity of this assumption by visually comparing the quantiles of the residuals against the quantiles of a normal distribution. If the residuals approximate a straight line on the Q-Q plot, it suggests that the residuals follow a normal distribution.

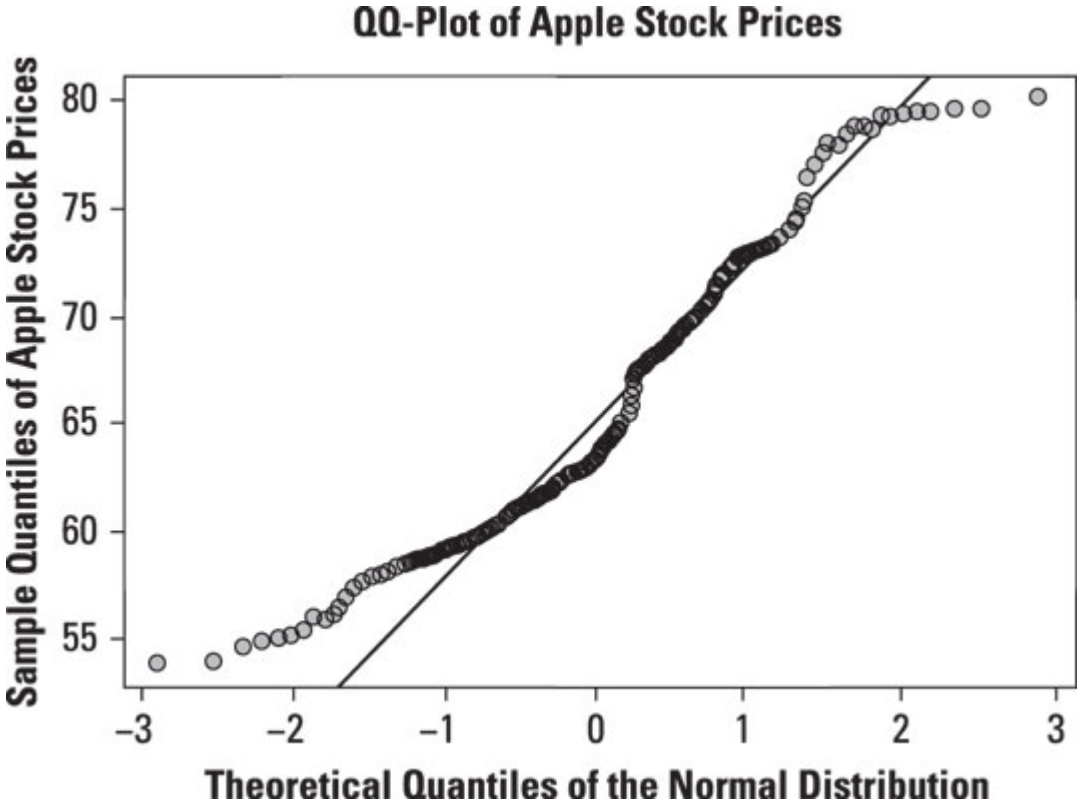
Detecting departures from normality: The Q-Q plot can reveal departures from normality in the residuals. If the plotted points deviate from a straight line, it indicates a departure from the assumed normal distribution. Departures may include skewness (asymmetric tails) or heavy-tailedness (excess kurtosis) in the residuals, which can affect the reliability of the linear regression model.

Assessing model assumptions: Linear regression relies on several assumptions, including linearity, independence, and homoscedasticity. Violations of these assumptions can lead to biased or inefficient coefficient estimates. The Q-Q plot helps diagnose departures from the normality assumption, which is a crucial assumption in linear regression analysis. If significant deviations from normality are observed, it may indicate violations of other assumptions as well.

Model refinement and diagnostics: The Q-Q plot provides insights for model refinement and diagnostics. If the Q-Q plot shows deviations from the expected straight line, it suggests that the model assumptions need further examination. It can guide the selection of appropriate transformations or suggest the need for robust regression techniques to account for non-normality or outliers.

Comparing alternative distributions: Besides assessing normality, the Q-Q plot can be used to compare the observed data against other theoretical distributions. This allows researchers to explore whether a different distribution may provide a better fit to the data, potentially leading to more accurate and reliable regression modeling.

By visually examining the patterns in the Q-Q plot, we can evaluate the distributional assumptions in linear regression and make informed decisions about the model's validity and reliability. It helps identify potential issues, guide model diagnostics, and support the selection of appropriate modeling techniques to improve the accuracy and interpretation of regression results.



(Image reference: https://www.dummies.com/wp-content/uploads/490217.image0.jpg)