Assignment-based Subjective Questions/Answers:

Q:From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

As per our analysis of categorical columns using the boxplot and bar plots, below are a few points we can infer from the visualisation:

- The fall season attracted a higher number of bookings, and the booking count has increased significantly from the year 2018 to 2019.
- As per observation, bookings happen between Q2 (April–June) and Q3 (July–Sep).
- Sunny weather is getting more bookings.
- As per weekly basics, we are getting more bookings on Thursday, Friday, Saturday, and Sunday.
- There is no difference seen between a working day and a non-working day.
- 2019 attracted a higher number of bookings than the previous year.

Q:Why is it important to use drop_first=True during dummy variable creation?

Answer:

drop_first = true is used when we create dummies, which helps us reduce extra columns. Hence, it reduces the correlations created among dummy variables.

Syntax -

drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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

Answer:

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

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

Answer:

- Normality of error terms

Error terms should be normally distributed.

- Multicollinearity check

There should be insignificant multicollinearity among variables.

- Linear relationship validation

Linearity should be visible among variables.

- Homoscedasticity

There should be no visible pattern in residual values.

- Independence of residuals

No auto-correlation

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

Answer:

The following 3 features contribute significantly to increasing the demand for shared bikes:

- temp(numerical variable)
- winter(catagorical variable)
- sep(catagorical variable)

General Subjective Questions

Explain the linear regression algorithm in detail.

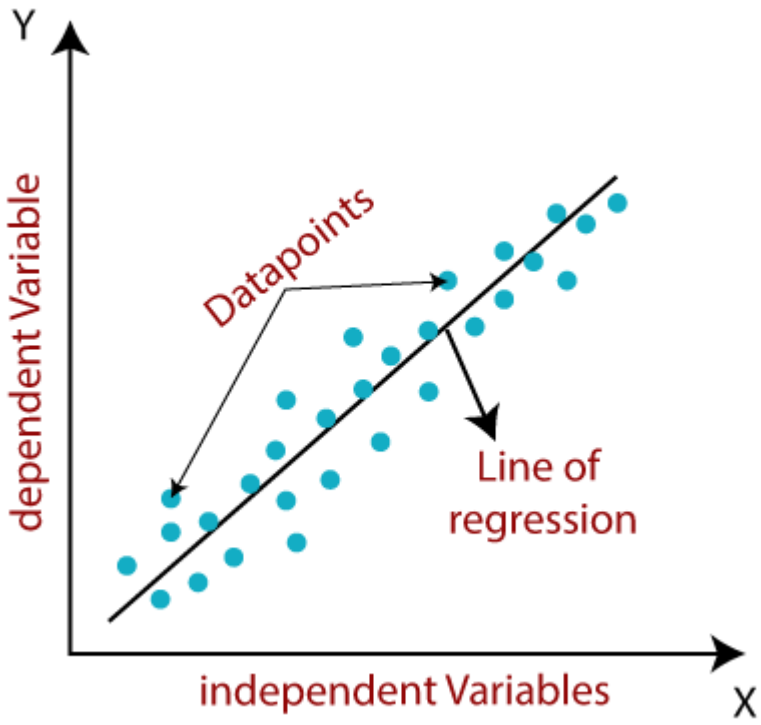
Answer:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable

is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



$y= a_0+a_1x+ \varepsilon$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

○ **Simple Linear Regression:**

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

○ **Multiple Linear regression:**

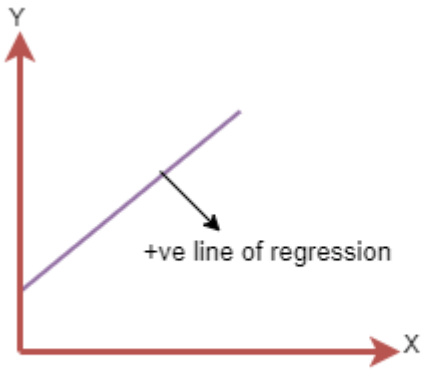
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**

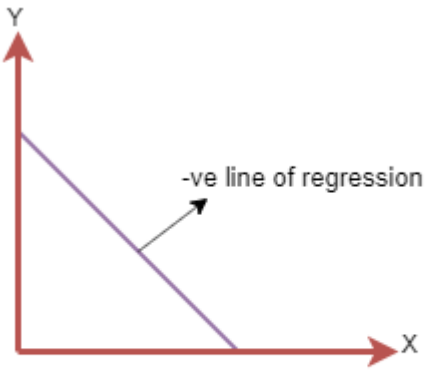
If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $\mathbf{Y= a_0+a_1x}$

- **Negative Linear Relationship:**

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line, so to calculate this we use cost function.

Cost function-

- The different values for weights or coefficient of lines (a_0, a_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

$$MSE = 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1x_i + a_0))^2$$

Where,

N=Total number of observation

Yi = Actual value

$(a1x_i+a_0)$= Predicted value.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent:

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

Linear relationship between the features and target:
Linear regression assumes the linear relationship between the dependent and independent variables.

Small or no multicollinearity between the features:
Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.

Homoscedasticity Assumption:
Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

Normal distribution of error terms:

Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
It can be checked using the **q-q plot**. If the plot shows a straight line without any deviation, which means the error is normally distributed

Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Purpose of Anscombe's Quartet

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's Quartet Dataset

The four datasets of **Anscombe's quartet.**

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	7.71
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	7.81	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	9.13	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

Find the Descriptive Statistical Properties for the all four Dataset

- Find mean for x and y for all four datasets.
- Find standard deviations for x and y for all four datasets.
- Find correlations with their corresponding pair of each datasets.
- Find slope and intercept for each datasets.
- Find R-square for each datasets.
- To find R-square first find residual sum of square error and Total sum of square error

While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

Q: What is Pearson's R?

Answer:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables

Correlation coefficients are used to measure how strong a relationship is between two variables. There are different types of formulas to get correlation coefficient, one of the most popular is Pearson's correlation (also known as Pearson's R) which is commonly used for linear regression. The Pearson's correlation coefficient is denoted with the symbol "R". The correlation coefficient formula returns a value between 1 and -1. Here,

- -1 indicates a strong negative relationship
- 1 indicates strong positive relationships
- And a result of zero indicates no relationship at all

Pearson's Correlation Coefficient Formula

The Pearson's correlation coefficient formula is the most commonly used and the most popular formula to get the correlation coefficient. It is denoted with the capital "R". The formula for Pearson's correlation coefficient is shown below,

$$R= n(\sum xy) – (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$$

The Pearson's correlation helps in measuring the strength(it's given by coefficient r-value between -1 and +1) and the existence (given by p-value)of a linear relationship between the two variables and if the outcome is significant we conclude that the correlation exists.

Cohen (1988) says that an absolute value of r of 0.5 is classified as large, an absolute value of 0.3 is classified as medium and an absolute value of 0.1 is classified as small.

The interpretation of the Pearson's correlation coefficient is as follows:-

1. A correlation coefficient of 1 means there is a positive increase of a fixed proportion

of others, for every positive increase in one variable. Like, the size of the shoe goes up in perfect correlation with foot length.

2. If the correlation coefficient is 0, it indicates that there is no relationship between the variables.

3. A correlation coefficient of -1 means there is a negative decrease of a fixed proportion, for every positive increase in one variable. Like, the amount of water in a tank will decrease in a perfect correlation with the flow of a water tap.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

*It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.*

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- ***sklearn.preprocessing.scale*** *helps to implement standardization in python.*
- *One disadvantage of normalization over standardization is that it* ***loses*** *some information in the data, especially about* ***outliers****.*

Example:

elow shows example of Standardized and Normalized scaling on original values

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

(3 marks)

Answer:

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.

The ***variance inflation factor*** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing *collinearity/multicollinearity*. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

How the VIF is computed

The *standard error* of an *estimate* in a *linear regression* is determined by four things:

The overall amount of noise (error). The more noise in the data, the higher the standard error.

The variance of the associated predictor variable. The greater the variance of a predictor, the smaller the standard error (this is a *scale* effect).

The sampling mechanism used to obtain the data. For example, the smaller the sample size with a simple random sample, the bigger the standard error.

The extent to which a predictor is correlated with the other predictors in a model.

The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the *R-squared* statistic of the regression where the predictor of interest is predicted by all the other predictor variables ( ). The *variance inflation* for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

Some statistical software use *tolerance* instead of VIF, where tolerance is:

$$1 - R^2 = \frac{1}{VIF}.$$

The VIF can be applied to any type of predictive model (e.g., CART, or deep learning). A generalized version of the VIF, called the *GVIF*, exists for testing sets of predictor variables and generalized linear models.

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

Answer:

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Quantiles And Percentiles

Quantiles are points in a dataset that divide the data into intervals containing equal probabilities or proportions of the total distribution. They are often used to describe the spread or distribution of a dataset. The most common quantiles are:

1. **Median (50th percentile)**: The median is the middle value of a dataset when it is ordered from smallest to largest. It divides the dataset into two equal halves.
2. **Quartiles (25th, 50th, and 75th percentiles)**: Quartiles divide the dataset into four equal parts. The first quartile (Q1) is the value below which 25% of the data falls, the second quartile (Q2) is the median, and the third quartile (Q3) is the value below which 75% of the data falls.
3. **Percentiles**: Percentiles are similar to quartiles but divide the dataset into 100 equal parts. For example, the 90th percentile is the value below which 90% of the data falls.

Note:

- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- For reference purposes, a 45% line is also plotted; **For** if the samples are from the same population then the points are along this line.