General Subjective Questions

Q.1 Explain the linear regression algorithm in detail.

Below are the important pointers about linear regression algorithm -

1. Linear regression is used for predictive analysis. It is a statistical method.
2. It helps to make predictions for **continuous variables** such as price of a product, salary of a person, price of a house etc.
3. The reason it is called linear regression because it shows a relationship between dependent (y) variable and one or more independent variable and hence is used to find out how the dependent variable values change according to the value of independent value.
4. Below equation represents a linear regression

   $y= a_0+a_1x+ \varepsilon$

   y-> Dependent variable
   x-> Independent variable
   a0 -> Intercept of the line
   a1 -> Coefficient of linear regression
   ε -> Random error

5. Linear regression has below two types

   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.

6. **Best Fit Line**
   The different values for weights or the coefficient of lines (a_0, a_1) gives a line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line.

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

   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.

Our aim is to find out the best model out of various models.

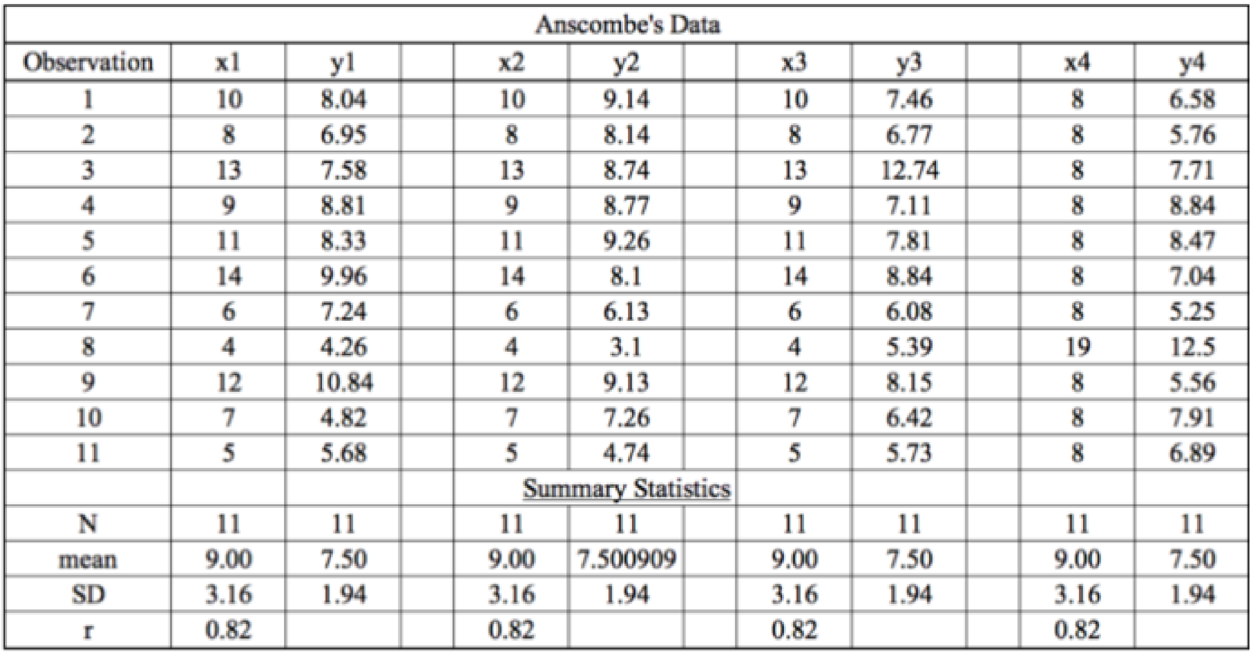
R-Squared

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.

Q.2 Explain the Anscombe's quartet in detail.

Anscombe's Quartet

1. It can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
2. We can see below that the summary statistics are almost the same for all data sets.



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Q.3 What is Pearson's R?

1. Pearson's r is a numerical summary of the strength of the linear association between the variables.
2. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
3.

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association

4. A correlation can be calculated between two numerical values (e.g.price of house and area) or between two category values (e.g., type of product and profession). However, a company may also want to calculate correlations between variables of different types. One method to calculate the correlation of a numerical variable with a categorical one is to convert the numerical variable into categories.

Q.4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling

This means **transforming your data so that it fits within a specific scale**, like 0-100 or 0-1.

There are two techniques of Scaling

1. **Normalisation-**

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

X' = (X-Xmin)/(Xmax-Xmin)

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1
- If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

2. **Standardisation**

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

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

1. If there is perfect correlation, then VIF = infinity.

2. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity

3. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let’s discuss some of the probability distributions.

The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

Assignment based questions on the next page

Assignment Based Questions

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

If we consider the final model using RFE –

#cnt = -0.9058 + 0.5461*temp -0.3611*weathersit_2 -1.3408*weathersit_3 + 0.4308*season_2 + 0.6508*season_4 + 1.0327*yr_1 + 0.2483*mnth_8 + 0.5315*mnth_9 + 0.2813*weekday_6 + 0.2484*workingday_1

Categorical variables have effect on the dependent variable

1. Weathersit_2 – If the weather is Mist , count of bikes is less
2. Weathersit_3 – If there is light snow, light rain then the count of bikes is less
3. Season_2 – Summer, bike count increases
4. Season_4 – winter, bike count increases
5. Yr_1 – 2019, bike count is higher compared to 2018
6. Mnth_8 – Aug bike count increases
7. Mnth_9 – September bike count increases
8. Workingday_1 – count is more during working days

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

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Example:

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

Looking at the pair-plot among the numerical variables, the one which has the highest correlation with the target variable is- temp

The correlation is 0.65

And the pairplot also shows a linear relationship between count and temperature

Lower the temperature , the count of bike will get impacted negatively.

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

Below are the assumptions of the Linear regression.

The relationship between X and the mean of Y is linear – This is clear from the graphs where we can see a linear graph formation.

Observations are independent of each other – There is no relationship between independent observations. Only relationship can be seen between independent and dependent variables.

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

Top 3 features which are contributing significantly towards the demand of shared bikes are.

Temp – higher temp, higher count, lower temp lower count

Year – 2019 has a higher bike count compared to 2018

Season4 - Winter