

General Subjective Questions

Q1. Explain the linear regression in algorithm?

Ans- Linear regression is a machine learning algorithm. It is based on supervised learning method where the output variable to be predicted is continuous variable. It is only applicable whenever you have target variable available where target variable should always have a value 1. It is a part of regression analysis. Linear regression is of two types-Simple Linear Regression and Multiple Linear Regression. If there is single input variable(x) present, then such type of variable called Simple Linear Regression whereas if there are more than one input variable present then such type of linear regression called Multiple Linear Regression. Linear Regression shows the linear relationship between the independent variable(x-axis) and dependent variable(y-axis).

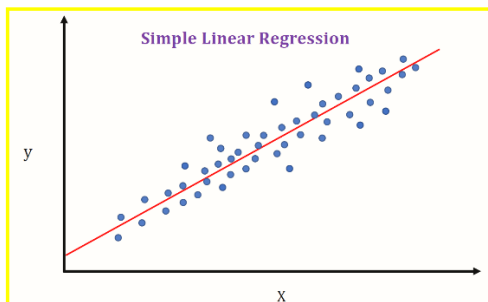
Simple Linear Regression: - Simple Linear Regression explains the relationship between a dependent variable and an independent one using a straight line. It uses simple slope- intercept formula,

$$Y = \beta_0 + \beta_1 X_1. \text{ Where,}$$

β_0 =intercept of the line,

X=independent variable.

Y = dependent variable



When the value of x increases, the value of y also increases. Regression method tries to find the best fit line which shows the relationship between dependent and independent variable with least error value.

Multiple Linear Regression: -Multiple Linear Regression is a regression model that shows the relationship between a quantitative dependent variable and two or more independent variable using straight line with formula - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$. Where,

Y = dependent variable,

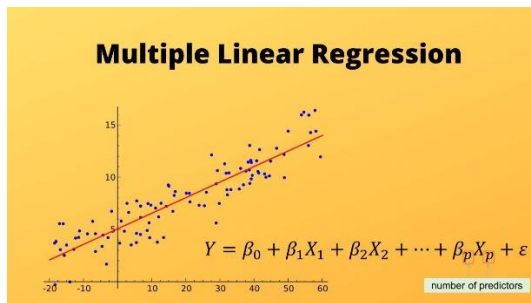
X=independent variable.

β_0 =intercept of the line 1,

β_1 = intercept of the line 2,

β_p = intercept of the nth line.

ϵ = Error value/Residual line.



Q2: Explain the Anscombe's quartet in detail.

Ans: - Anscombe's quartet is a group of four data sets which are nearly identical in simple descriptive statistics but have different distribution and appearance when graphed means it appeared differently on scatter plot. It was constructed in 1973 by statistician Francis Anscombe's to illustrate the importance of plotting the graphs before analysing and model building. The four data set plots which have nearly same statistical observations provides same statistical information that involves variance and mean of all x,y points in all four datasets.

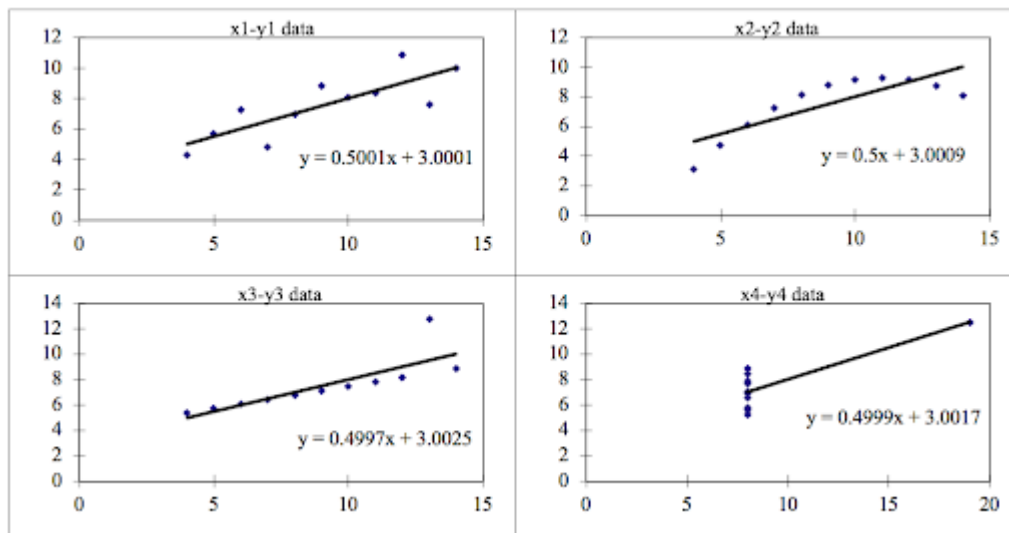
The four data plots can be defined as follows:

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

The statistical information for these four datasets are approximately same. We can compute them as follows:

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	

When these models are plotted on a scalar plot ,each dataset generates a different kind of plot that is not interpretable by any regression algorithm which we can see below:



WE can understand easily by seeing the four datasets given above.




- 1.Dataset1: -Fits the linear regression model well.
- 2.Dataset2: - Cannot fit the linear regression model because the data is non-linear.
- 3.Dataset3: -Shows the involvement of outliers in the dataset which cannot be handled by linear regression model.
- 4.Dataset4: Shows the outliers involved in the dataset which also cannot be handled by linear

Q3: What is Pearson's R?

Ans: -The Pearson correlation coefficient(r) is the most common way of measuring a linear correlation. It assigns a value between -1 and 1, where 0 is no correlation,1 is total positive correlation and -1 is total negative correlation. Pearson correlation coefficient (PCC), also referred to as Pearson's r , the Pearson product moment correlation coefficient (PMCC), or the bivariate

correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations.

Pearson's R Formula is as follows:


$$\text{Pearson Correlation Coefficient} = \rho(x,y) = \frac{\sum [(x_i - \bar{x}) * (y_i - \bar{y})]}{\sigma_x * \sigma_y}$$


$$\text{Pearson Correlation Coefficient} = \rho(x,y) = \frac{\sum [(x_i - \bar{x}) * (y_i - \bar{y})]}{(\sigma_x * \sigma_y)}$$

Where,

\bar{x} = Mean of x variable

\bar{y} = Mean of y variable

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

Ans: -Scaling means that you are transforming your data so that it fits in specific scale. It is a step of data pre-processing which is applied to independent variables to normalize the data within a range. It helps in speeding up the calculation in an algorithm. Most of the times, collected data contains features highly varying in magnitudes, units and range. If scaling will not be done then algorithm takes only magnitude in, consider not units which results in incorrect modelling. To solve this problem, we must do scaling to bring all the variables to the same level.

Difference between normalized scaling and standardized scaling:

1. In normalized scaling, minimum and maximum value of features being used whereas in standardized scaling mean and standard deviation are used.
2. Normalized scaling scales values between (0,1) or (-1,1), whereas standardized scaling is not having or is not bounded in a certain range.
3. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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

Ans: VIF stands for Variance Inflation Factor. It basically explains the relationship with all other independent variables. If there is perfect correlation, then $VIF = \infty$. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing the perfect multicollinearity. This shows a perfect correlation between two independent variables. The formulation of VIF is given below: a VIF value of greater than 10 is high, a VIF of greater than 5 should also not be ignored and inspected appropriately. An infinite VIF value indicates that the corresponding variable may be expressed by a linear combination of other variables.

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

Ans: - Q-Q plot are also known as Quantile-Quantile plot. It is a probability plot represents graphical method for comparing two probability distributions by plotting their quantiles against each other. Q-Q plot is a graphical tool which help us to analyse if set of data possibly came from some theoretical distribution such as Normal, exponential or uniform distributions. It also helps us to determine whether two distributions are similar or not. If the two distributions are similar, the points in the Q-Q plot will approximately lie on the line $y=x$. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line $y=x$. It also helps in estimating parameters in a location-scale family of distributions.

A Q-Q plot is used to compare the shapes of distributions, provides a graphical view of how properties such as locations, scale and skewness are similar or different in the two distributions.

Importance of Q-Q plot in linear regression:

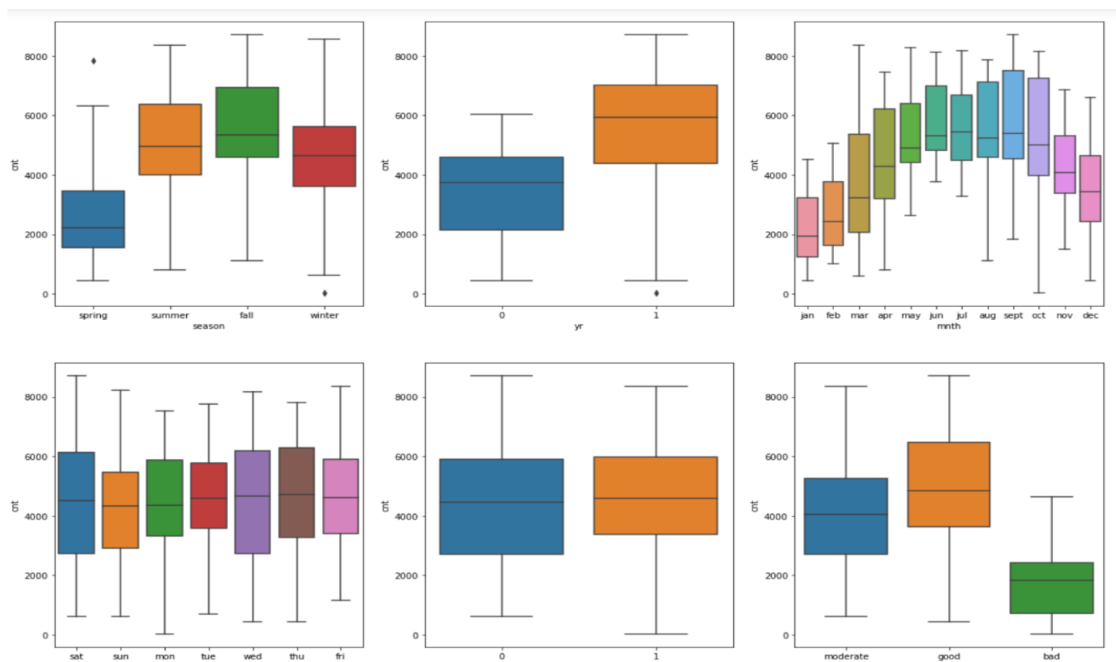
Q-Q plot are used when we have train-test dataset in a linear regression by which we can confirm that both the train and test dataset are from the same distribution or not. It's advantages are:

- 1.It can be used with sample size also.
- 2.If both datasets have similar type of distributions shape.
- 3.If both datasets have common location and common scale.
- 4.Many distributional aspects like shift in locations, shifts in scale, changes in symmetry and the presence of outliers can also be detected from this plot.

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: As per the analysis, we found that there are couple of variables like season, mnth, yr, weekday, workingday and weathersit have a major effect on the 'cnt' variable. Let's understand through the below figure:



These variables are visualized using bar plot and Box plot both.

Following are the effects for categorical variables on the dependent variables:

1. Company should focus on expanding business during Fall, Summer and Winter.
2. September month has shown great demand.

3. There is not much demand during the holidays.

4. There would be less booking during Bad and no demand in severe weather conditions.

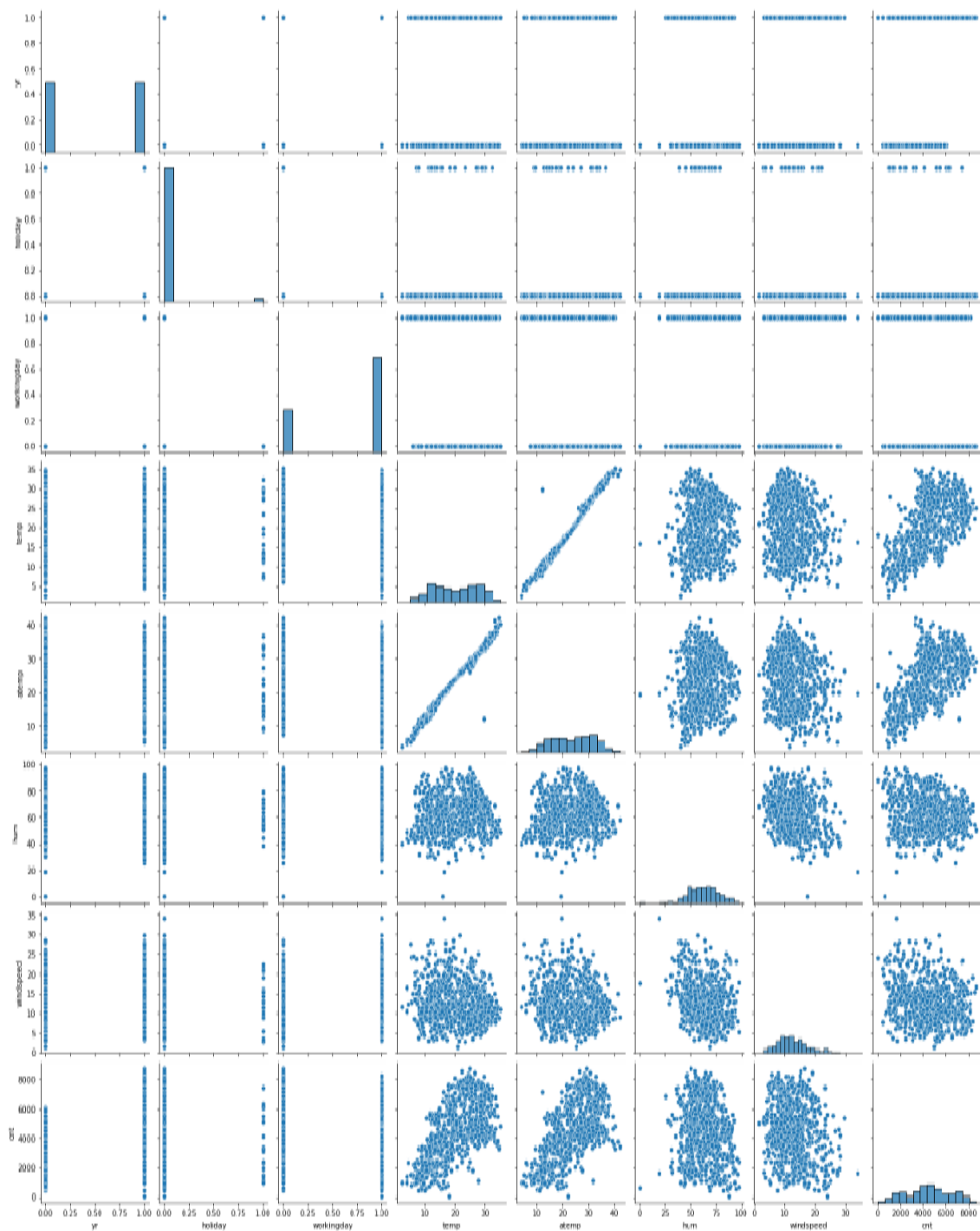
5. It has been seen that demands for bike rentals had gone up from 2018 to 2019.

Q2. Why is it important to use `drop_first=True` during dummy variable creation?

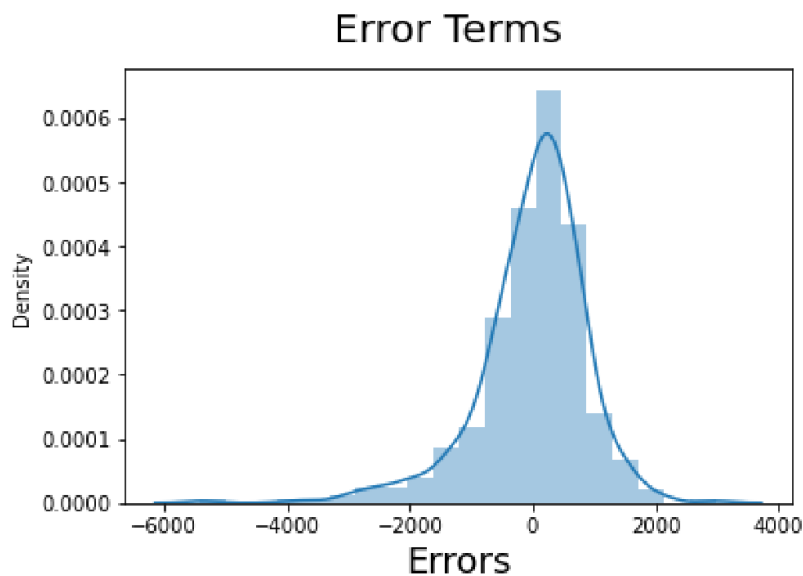
Ans: The use of `drop_first=True` is important as it helps in reducing unnecessary columns which are created during the creation of dummy variables. Hence, it reduces the correlation created among dummy variables.

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

Ans: After looking at the pair-plot among the numerical variables, 'temp' variable has the highest correlation with the target variables. Let's analyse through the below given figure:



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



Ans: Residual distribution should follow normal distributions and centred around 0. After validating the assumptions about residuals by plotting a distplot of residuals, we will see whether the residuals are following normal distributions or not. The above diagram shows that the residuals are distributed about mean=0.

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

Ans: Based on the final model, yr, temp and weather are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

