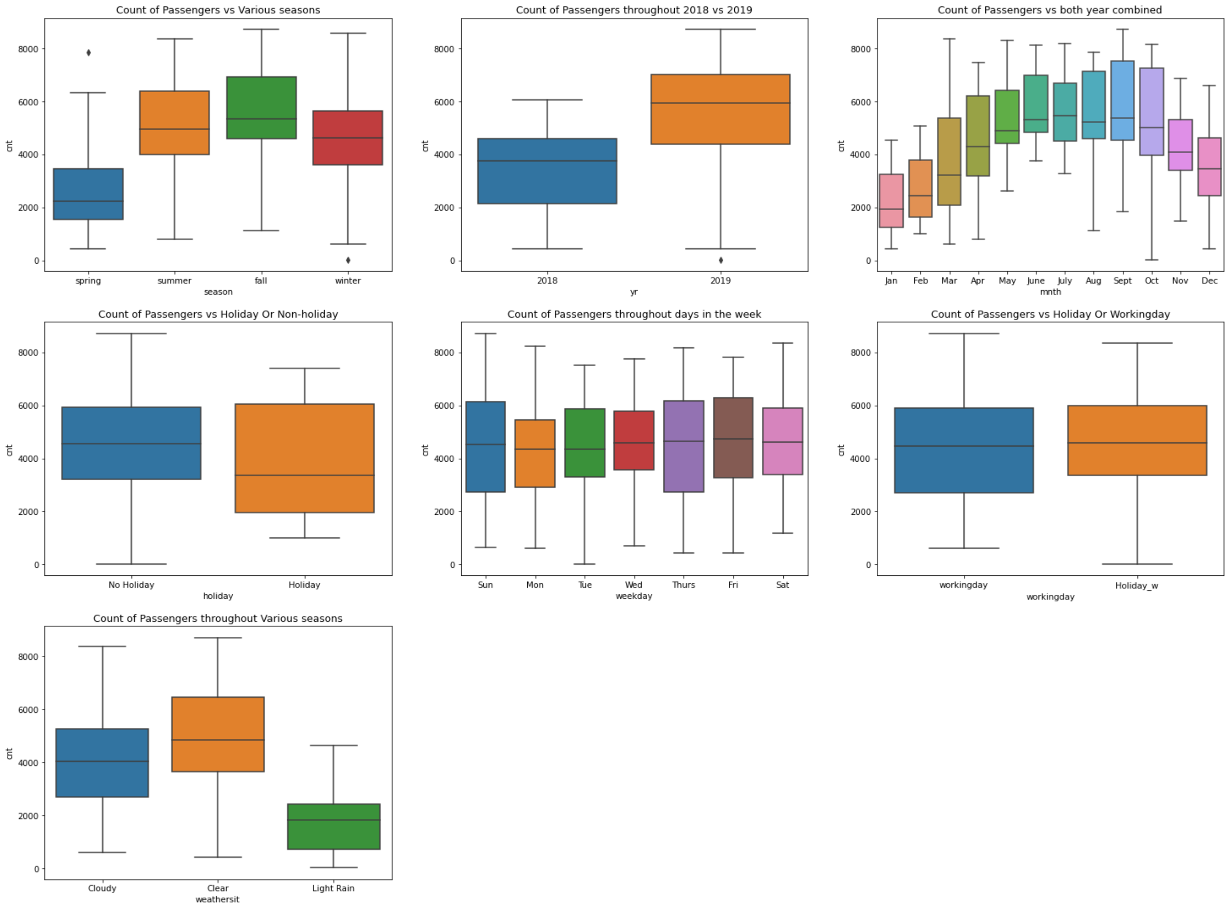
Assignment-based Subjective Questions

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

Answer: The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables.

When we see the boxplot of the data of each category, we got know that -:



•Bike demand in the fall is the highest.

•Bike demand takes a dip in spring.

•Bike demand in year 2019 is higher as compared to 2018.

•Bike demand is high in the months from May to October.

•Bike demand is high if weather is clear or with cloudy while it is low when there is light rain or light snow.

•The demand of bike is almost similar throughout the weekdays.

•Bike demand doesn’t change whether day is working day or not.

coef	std err	t	P>|t|	[0.025	0.975]	
const	0.1285	0.017	7.644	0.000	0.095	0.161
temp	0.5522	0.020	27.660	0.000	0.513	0.591
windspeed	-0.1552	0.025	-6.127	0.000	-0.205	-0.105
yr_2019	0.2331	0.008	28.030	0.000	0.217	0.249
season_summer	0.0892	0.010	8.565	0.000	0.069	0.110
season_winter	0.1297	0.010	12.364	0.000	0.109	0.150
mnth_Sept	0.0959	0.016	6.012	0.000	0.065	0.127
weekday_Mon	-0.0464	0.012	-3.925	0.000	-0.070	-0.023
weathersit_Cloudy	-0.0786	0.009	-8.877	0.000	-0.096	-0.061
weathersit_Light Rain	-0.2833	0.025	-11.326	0.000	-0.332	-0.234

Summer, winter falls under season category and have been dummy encoded. weathersit_Cloudy and weathersit_Light Rain falls under weathersit category and have been dummy encoded. Similarly, month variables fall under mnth category and have been dummy encoded. We can infer from above image that these variables are statistically significant and explain the variance in model very well.

Question: Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

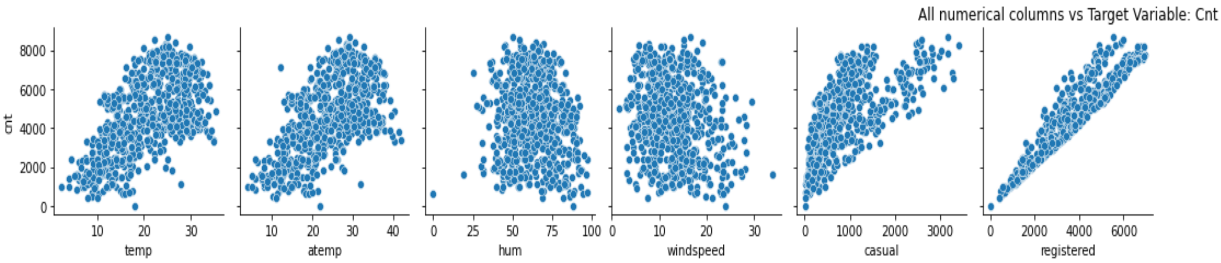
Answer: We can use the drop first = True during the dummy variable creation because

1. It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.
2. For Example: We have three variables: Cloudy, Rain and Clear. We can only take 2 variables as Cloudy will be 1-0, Rain will be 0-1, so we don’t need Clear weather as we know 0-0 will indicate Clear. So, we can remove it .

3. It is also used to reduce the collinearity between dummy variables.

Question: Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable? (1 mark)

Answer:



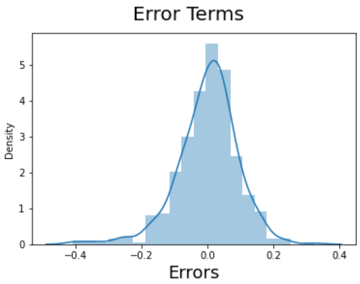
While looking to pair-plot among all numerical variables we got to know that the atemp and temp variable are highly correlated to each other with target variable “cnt” So, before model building and training, the pair plot shows highest correlation for registered variable having correlation 0.945. But we are not using casual and registered in our pre-processed training data for model training. casual + registered = cnt. This might leak out the crucial information and model might get overfit. So, excluding these two variables atemp is having highest correlation with target variable cnt which is followed by temp. As per the correlation heatmap, correlation coefficient between atemp and cnt is 0.631. And correlation coefficient between temp and cnt is 0.627.

Question: How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

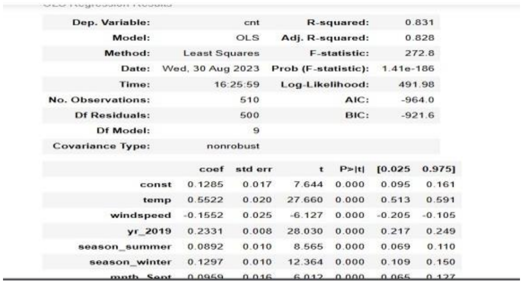
Answer: To validate the assumptions of Linear Regression after building the model on the training set, we see the following points -:

- **Residual Analysis -:** So, now to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression)

  Error terms are independent of each other, When we plotted the scatter plot of residuals, we did not find any pattern or clustering or any other continuous pattern which concludes that they are not dependent on each other. The residuals are following the normally distribution with a mean 0.



- **Linear relationship between predictor variables and target variable -:** This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-Squared value on training set is 0.831 and adjusted R-Squared value on training set is 0.828. This means that variance in data is being explained by all these predictor variables.



OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.831
Model:	OLS	Adj. R-squared:	0.828
Method:	Least Squares	F-statistic:	272.8
Date:	Wed, 30 Aug 2023	Prob (F-statistic):	1.41e-186
Time:	16:25:59	Log-Likelihood:	491.98
No. Observations:	510	AIC:	-964.0
Df Residuals:	500	BIC:	-921.6
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1285	0.017	7.644	0.000	0.095	0.161
temp	0.5522	0.020	27.660	0.000	0.513	0.591
windspeed	-0.1552	0.025	-6.127	0.000	-0.205	-0.105
yr_2019	0.2331	0.008	28.030	0.000	0.217	0.249
season_summer	0.0892	0.010	8.565	0.000	0.069	0.110
season_winter	0.1297	0.010	12.364	0.000	0.109	0.150
mnth_Sept	0.0959	0.016	6.012	0.000	0.065	0.127

- **Error terms are independent of each other**: Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

Question: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: These are top three features which is contributing significantly towards explaining the demand of shared bikes.

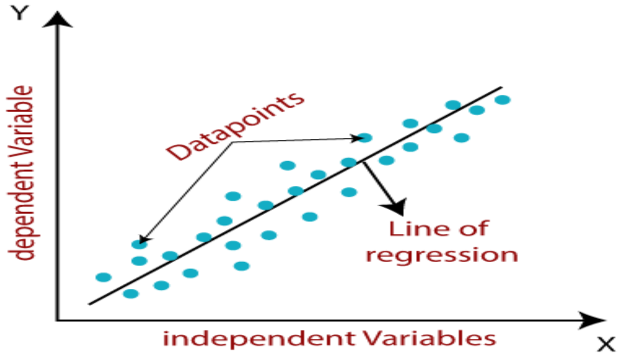
- **mnth_Sept (coef :0.0959)**
- **temp (coef :0.5522)**
- **yr_2019 (coef :0.231)**

General Subjective Questions

Question: Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear regression is a type of supervised machine learning algorithm.

Linear Regression computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variables.



Types of Linear Regression -:

- ### Simple Linear Regression

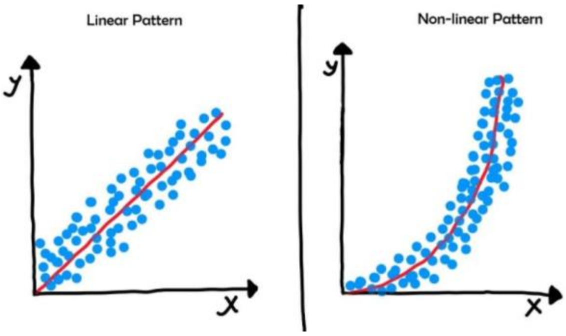
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- ### Multiple Linear regression:

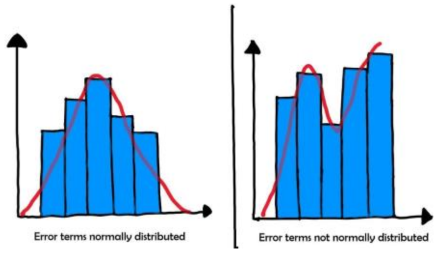
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

The assumptions of linear regression are -:

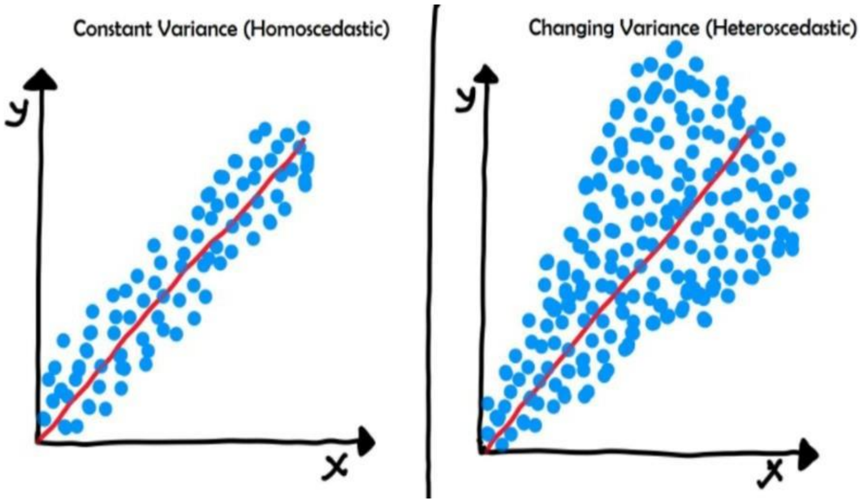
A. The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables



B. 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.

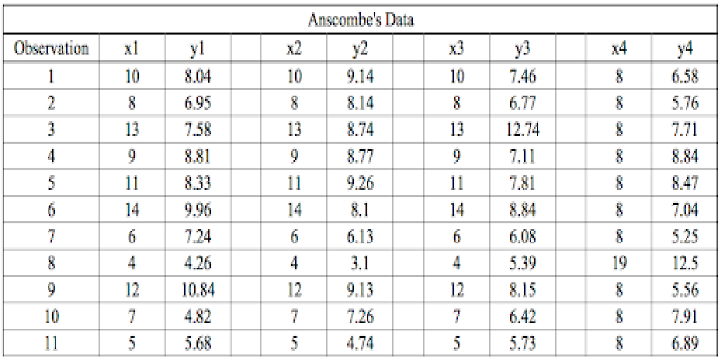


C. Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.



Question: Explain the Anscombe's quartet in detail. (3 marks)

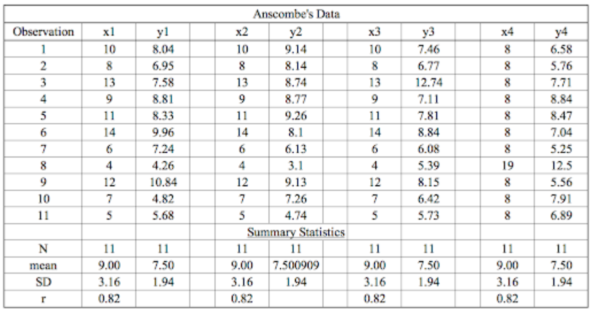
Answer: Anscombe's quartet is a group of four data sets thar are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set.



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

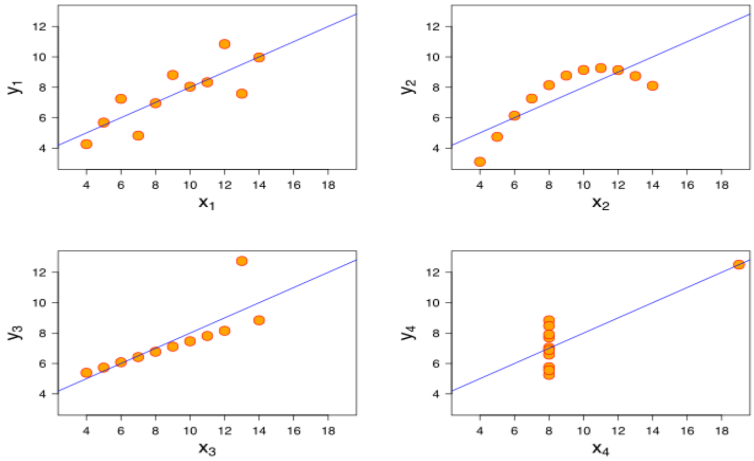
These four data sets have nearly the same statistical observations, which provide the same Information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.



Anscombe's Data

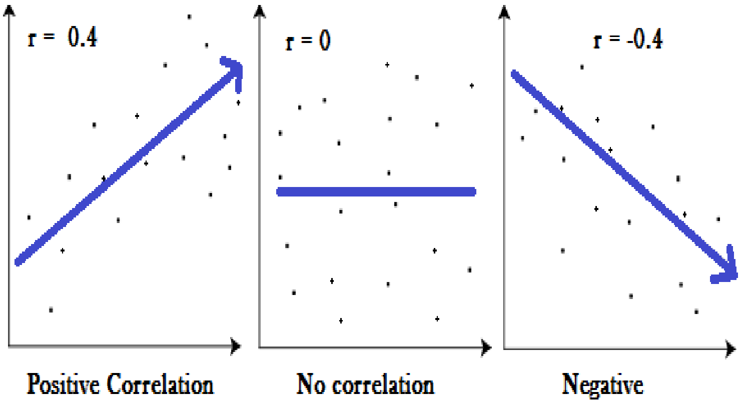
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	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	

The data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.



- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear
- Data Set 3: It shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: It shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Question: What is Pearson’s R? (3 marks)
Answer:



Pearson’s R or correlation coefficient is a measure of linear correlation between twosets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

Question: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

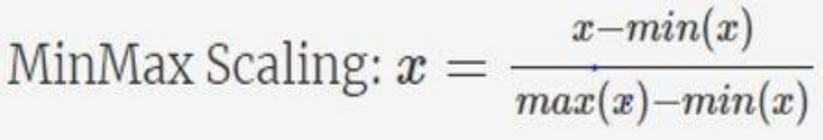
Answer: Scaling is a method used to normalize the range of independent variables or features of the data.

Since the range of values of raw data varies widely, in some machine learning algorithms, objectives functions will not work properly without normalization.

For example. Many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has to broad range of values, the distance will be normalized so that each features contributes approximately proportionately to the final distance.

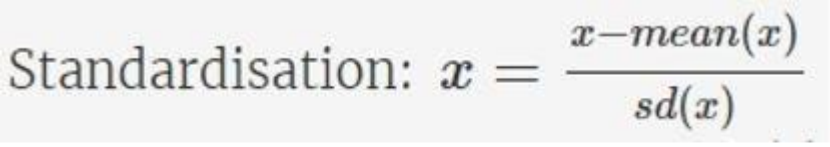
Another reason why features scaling is applied is that gradient descent converges much faster with features than without it.

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



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

- 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**).



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

Question. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: The Variance Inflation Factor (VIF) is a statistical measure used to assess multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, which can cause problems in interpreting the coefficients and can lead to unstable and unreliable estimates.

- The formula for calculating VIF for a particular independent variable in a regression model is:

  VIF = 1 / (1 - R^2)

- Where R^2 is the coefficient of determination when the variable in question is regressed against all the other independent variables in the model. If R^2 is equal to 1, the denominator in the formula becomes 0, leading to a division by zero, which results in an infinite VIF value.

- So, the VIF becomes infinite when one of the independent variables in your regression model can be perfectly predicted by a linear combination of the other independent variables. In other words, when there is perfect multicollinearity, meaning that one variable is a perfect linear function of one or more other variables in the model, the VIF becomes infinite.

- In practical terms, when you encounter infinite VIF values, it's a strong indicator that you have a severe multicollinearity problem in your regression model. To address this issue, you may need to consider removing one or more of the highly correlated variables from your model or finding alternative ways to handle the multicollinearity, such as combining the correlated variables into a composite variable or using regularization techniques like ridge regression.

Question. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks

Answer: Q-Q (quantile-quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution, such as the normal distribution. It compares the quantiles (ordered values) of the dataset to the quantiles of the chosen theoretical distribution, typically the normal distribution. This comparison helps you determine how closely the dataset matches the expected distribution.

Here's how a Q-Q plot works:

1. Data Sorting: First, you sort the data in ascending order, so you have the smallest data point at the beginning and the largest at the end.

2. Calculation of Expected Quantiles: For the chosen theoretical distribution (e.g., the normal distribution), you calculate the expected quantiles at various percentiles using a mathematical formula or a lookup table. These expected quantiles represent what the data should look like if it follows the theoretical distribution.

3. Plotting: You plot the observed quantiles (the sorted data points) on the vertical axis against the expected quantiles (from the theoretical distribution) on the horizontal axis. Each point on the plot corresponds to a specific data point's quantile.

4. Interpretation: By visually inspecting the Q-Q plot, you can assess whether the data points fall approximately along a straight line. If they do, it suggests that the data follows the theoretical distribution closely. If there are deviations or bends in the line, it indicates departures from the theoretical distribution.

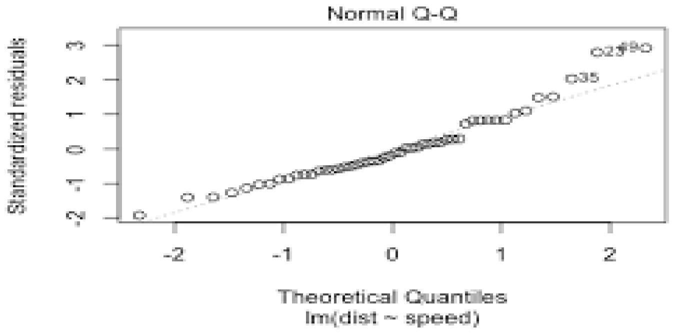
The use and importance of a Q-Q plot in linear regression:

1. Assumption Checking: In linear regression analysis, one of the key assumptions is that the residuals (the differences between observed values and predicted values) are normally distributed. Q-Q plots are valuable tools for checking this assumption. By plotting the residuals of a linear regression model against the expected quantiles of a normal distribution, you can assess whether the residuals follow a normal distribution pattern.

2. Detecting Departures from Normality: If the Q-Q plot shows a strong deviation from a straight line, it suggests that the residuals do not follow a normal distribution. This information is essential because violations of the normality assumption can affect the validity of regression results, including hypothesis tests and confidence intervals.

3. Model Improvements: If the Q-Q plot reveals that the residuals do not conform to the normal distribution, it may indicate the need for model adjustments or transformations. For example, you might consider applying a transformation to the response variable or adding additional predictor variables to address non-normality in the residuals.

In summary, a Q-Q plot is a graphical tool that allows you to visually assess whether a dataset or a set of residuals from a linear regression model follows a specific theoretical distribution, such as the normal distribution. It is important in linear regression because it helps you verify the normality assumption and make informed decisions about model adequacy and potential improvements.



Salil Chandan
(Executive PG Program in Machine Learning and AI)