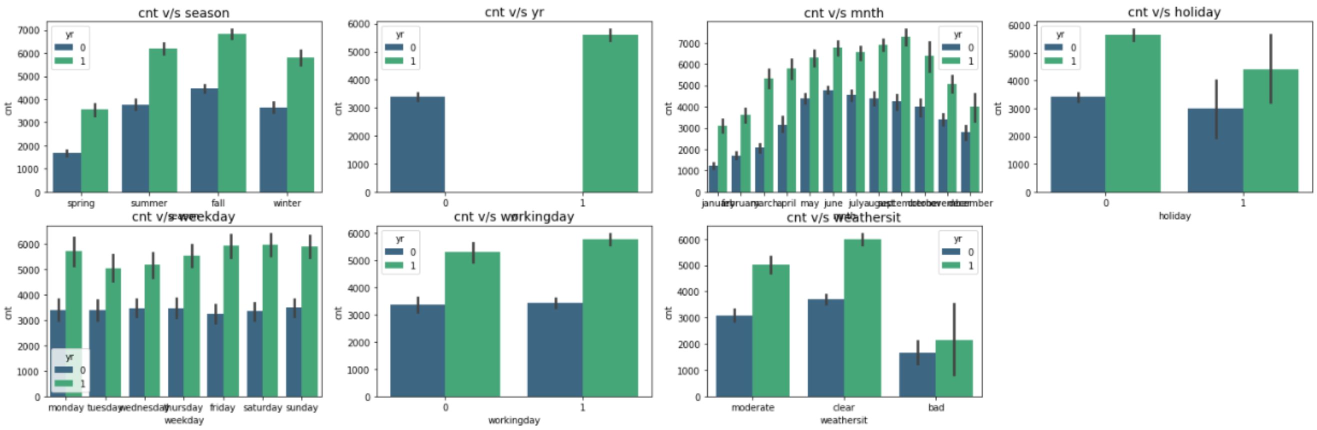
Assignment-based Subjective Questions

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

A:

- display a bar plot for each categorical variable with the variable 'cnt' as the target variable, displaying the trend year by year



The influence of categorical variables on the target variable :

- The hiring rate is higher for 2019 than for 2018.
- Fall shows a higher demand followed by summer season.
- The May to October period also shows a higher demand than other parts of the year.
- Working days are in higher demand than holidays and weekends.
- Clear weather days attract more hiring demand than other days.
- Temperature has a positive correlation (0.63) with demand. And higher temparature attract more bike hiring.

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

A:

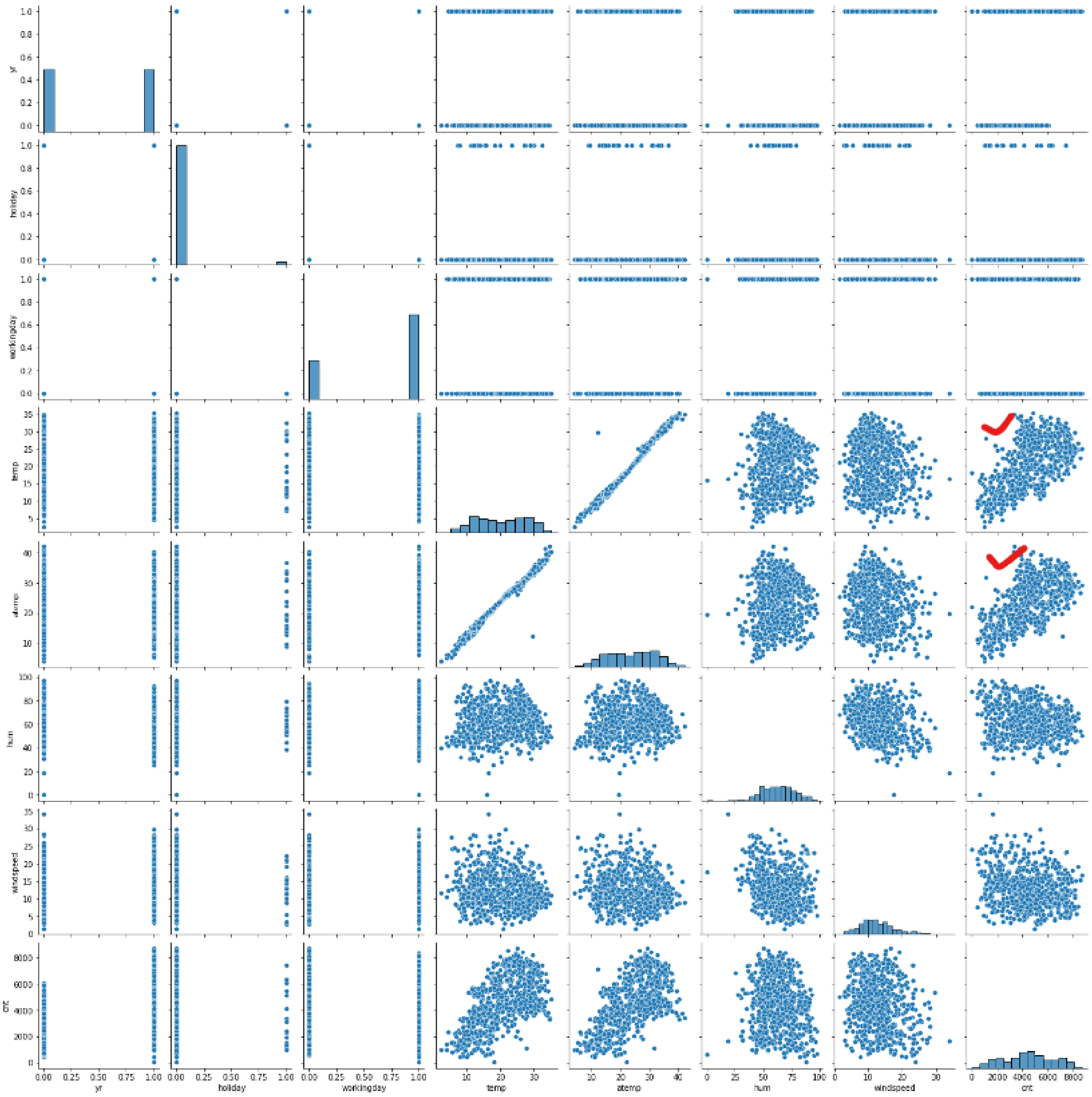
- N-1 columns required to represent n categorical variables, in order to reduce the correlation created among dummy variables, we need to remove any one dummy variable such that all dummy variables will be independent.

- Pandas has the option to remove one dummy variable by passing the drop_first=True argument during dummy variables creation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

A: Temperature has the highest correlation (0.63) with the target variable.

```
<Figure size 1080x1080 with 0 Axes>
```



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

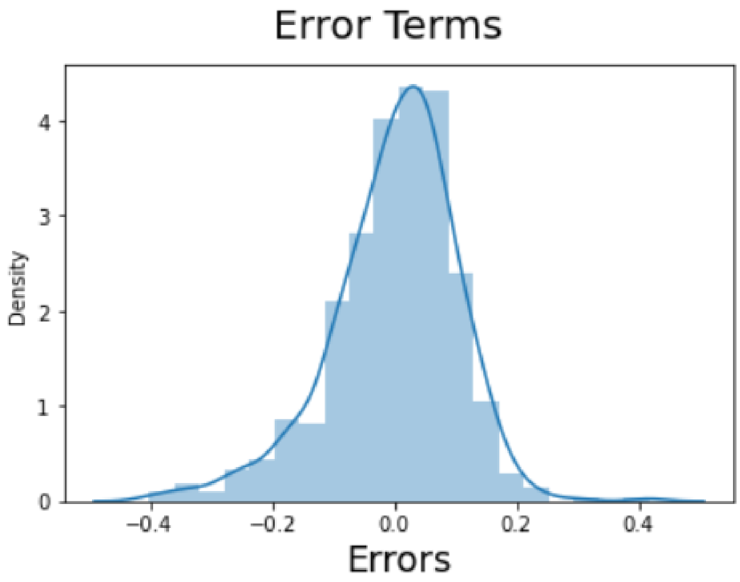
A:

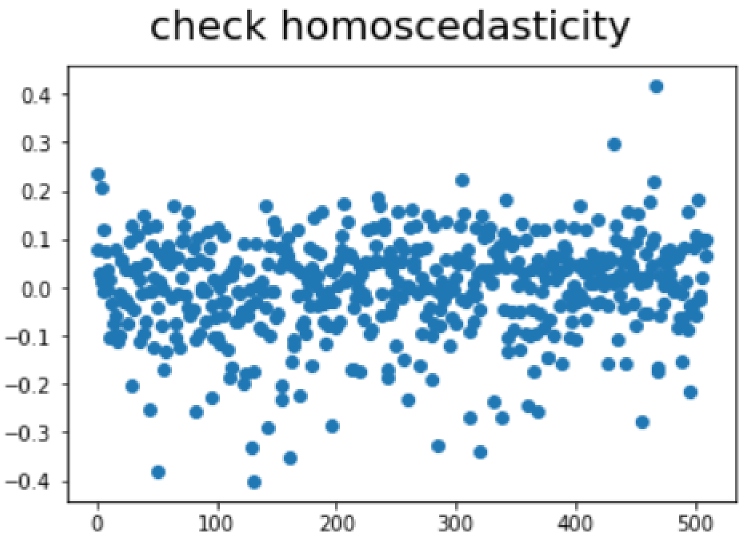
The assumptions of linear regression are:

- X and Y have a linear relationship.
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other.
- Error terms have constant variance (homoscedasticity).

There is a linear relationship between X and Y.

From the distribution plot for error terms, it clearly shows that the mean is centered around zero and normally distributed.





From the above scatter plot, it clearly shows that error has constant variance and no visible patterns are there. It satisfies the condition of homoscedasticity.

All the assumptions of linear regression are satisfied.

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

A:

According to my analysis and the trained model, the top three factors contributing to shared bike demand are the year, the weather situation, and the month of September.

General Subjective Questions

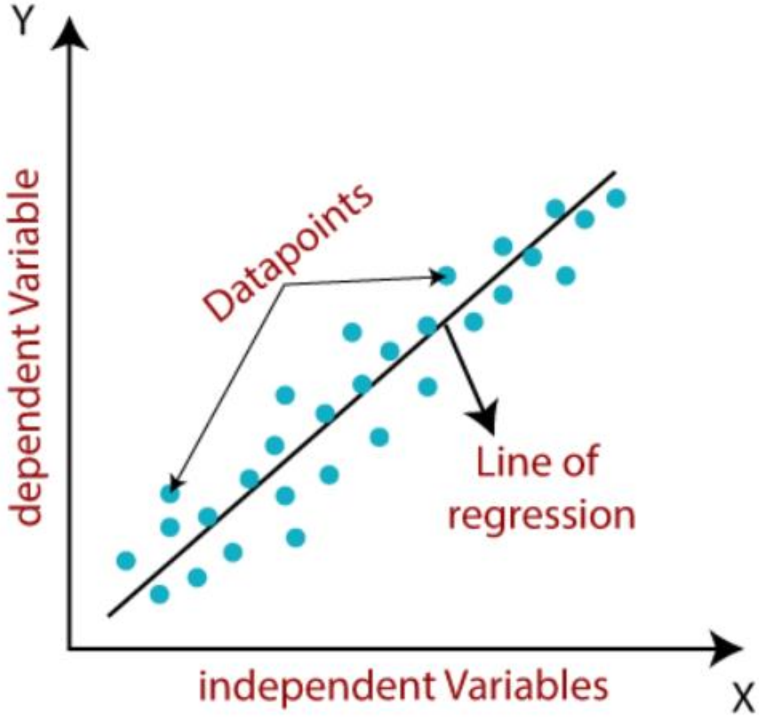
1. Explain the linear regression algorithm in detail. (4 marks)

A:

Linear Regression is a supervised machine learning algorithm. It performs a regression task, which means predicting a numerical value based on independent variables. The linear regression algorithm has a linear relationship between a dependent variable (y) and one or more independent (x) variables. Since it is a linear relationship, the dependent variable changes according to the value of the independent variables.

The linear regression model fits a straight line to represent the relationship between the dependent and independent variables.

Please check the below image for a better understanding:



Equation of Simple Linear Regression:

$$y = b_o + b_1x$$

The independent variable is x, and the dependent variable is y, bo is the intercept, b1 is coefficient or slope

Equation of Multiple Linear Regression:

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots. + b_nx_n$$

bo is the intercept, b1,b2,b3 ... ,bn are coefficients or slopes of the independent variables x1,x2,x3 ... ,xn and y is the dependent variable.

The assumptions of linear regression are:

- X and Y have a linear relationship.
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other.
- Error terms have constant variance (homoscedasticity).

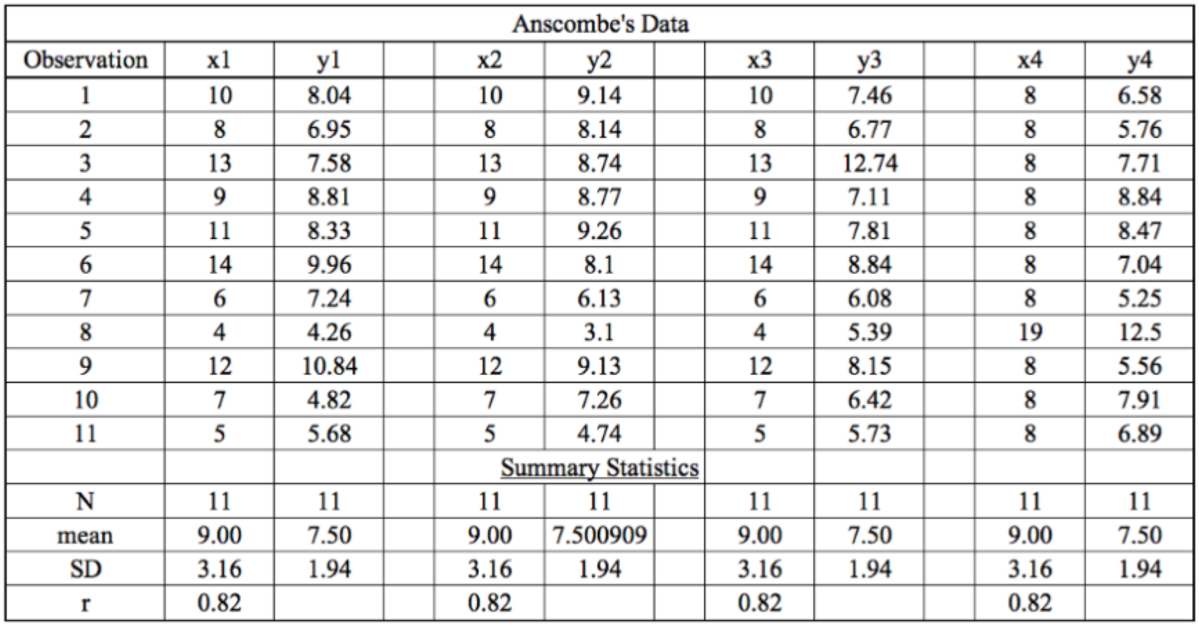
2. Explain the Anscombe’s quartet in detail. (3 marks)

A:

Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of data visualization before constructing the machine learning model. There are four datasets which have nearly the same statistical information like mean and variance. When we try to fit a linear regression model, all four datasets have almost the same linear regression model parameters but failed to generalize the distribution of the data, so this dataset can easily fool the linear regression model.

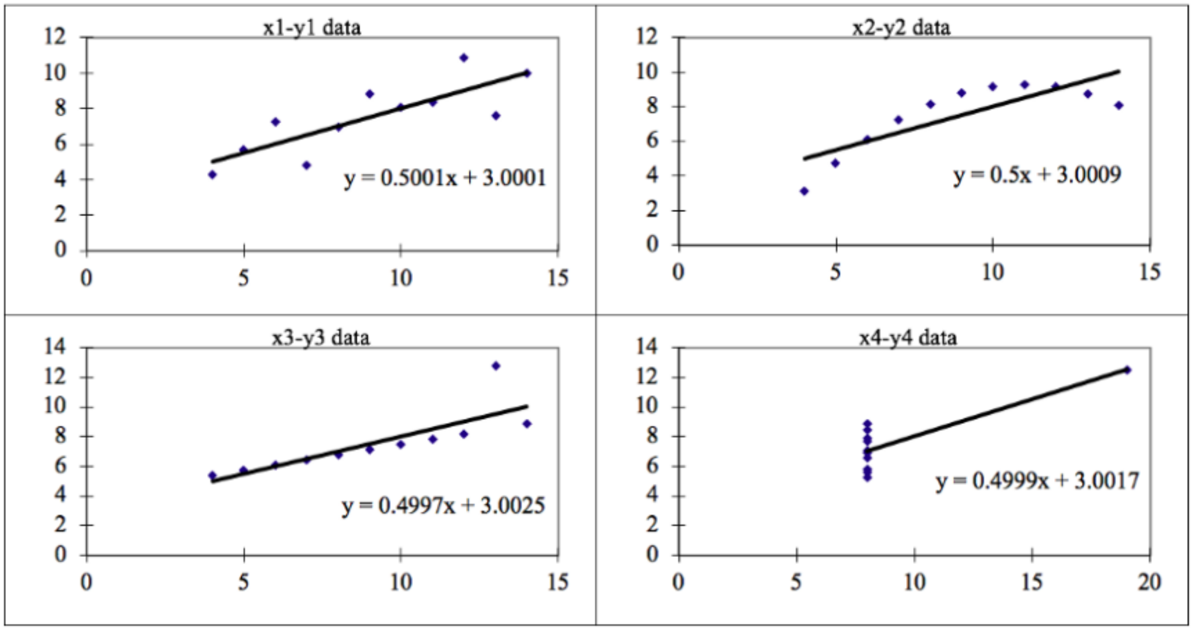
So it is very important to analyze the data by visualization before building any model on it. Data visualization helps us to understand the distribution of the data, outliers in the data, and linear separability of the data.

Anscombe’s Data and Statistics shown below:



					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 we check the below plots, all the data sets show the same regression line but have different distribution data.



The four datasets can be described as:

Dataset 1: Linear regression fits pretty well.

Dataset 2: non-linear data; linear regression doesn’t fit well.

Dataset 3: Outliers present in the data, linear regression doesn’t fit well.

model

Dataset 4: outliers present in the data, linear regression doesn’t fit well.

model

The summary is that Anscombe’s quartet helps us to understand the importance of data visualization and how someone can fool a regression model with the data. So before building any model, first try to visualize the data and understand the different properties, and then build the model.

3. What is Pearson’s R? (3 marks)

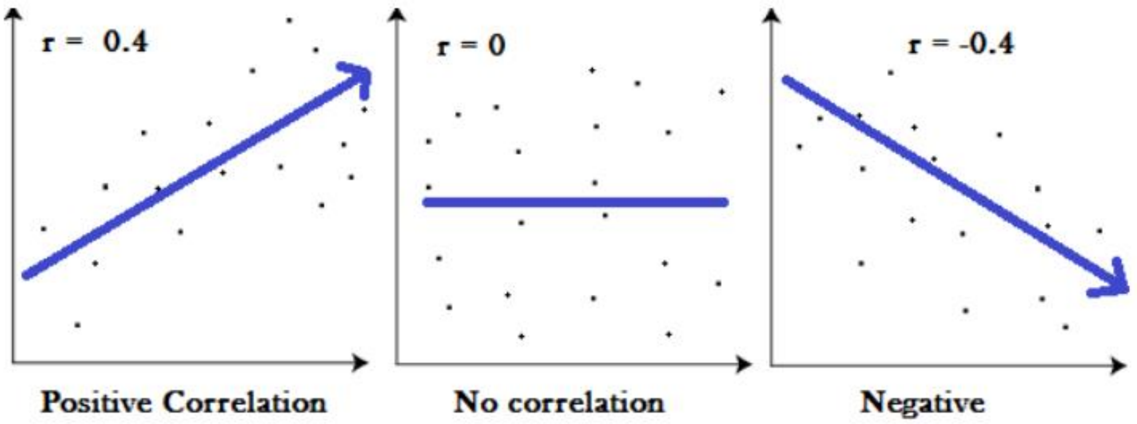
A:

Pearson’s R, also called Pearson’s correlation, is a correlation coefficient used in linear regression. Basically, correlation measures the relationship between two variables.

Pearson’s correlation coefficient formula.

$$\mathbf{r} = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

Correlation coefficient values range between -1 and 1, where:



- A correlation coefficient of 1 means positive correlation; an increase in one variable results in the same proportion of increase in another variable.
- A correlation coefficient of -1 means negative correlation; an increase in one variable results in the same proportion of decrease in another variable.
- "Zero correlation" means these two variables are not related to each other.

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

A:

If the data is not scaled, machine learning algorithms are biased towards large numerical data values. As a result, we must scale all of the data values into a common range so that the machine learning model can fit on it.Also, it helps to converge machine learning algorithms faster.

There are a variety of scaling techniques in machine learning. The primarily used methods are normalized scaling and standardized scaling.

Normalized Scaling

This method is used to transform features into scales that are similar. The new value is calculated by using the below equation.

X_new = (X - X_min)/(X_max - X_min)

The output data range is [0, 1] or [-1, 1].

This method is useful when there are no outliers and the user doesn't have information about the distribution of the data. This transformation squishes the n-dimensional data into an n-dimensional hypercube.

standardized scaling:

This method is used to transform the data into a standard normal distribution. All the features are subtracted by the mean and divided by the standard deviation. This is often called the Z-score.

X_new = (X - mean)/Std

This method is useful when the user knows about data distribution. It will also affect dummy variables (Ex: 0 &1). Standardized scaling does not get affected by outliers.

Differences between normalized scaling and standardized scaling:

Normalization	**Standardization**
Scaling uses the minimum and maximum value of features.	Standardization uses the mean and standard deviation of features.
This method is used when the features are of different scales.	This method is used to ensure zero mean and unit standard deviation.
Output values range between [0, 1] or [-1, 1].	There is no output range criteria.
This method is affected by outliers.	This method is less affected by outliers.
It is useful when we don't know about the distribution.	It is useful when the feature distribution is normal or Gaussian.

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

A:

- VIF gives a basic idea about how many feature variables are correlated to each other. That means how well a predictor variable is correlated with all other variables, excluding the target variable.
- The formula for VIF is 1/(1-R2).
- If there is perfect correlation, we get R2 = 1, which leads to 1/(1-R2) infinity, then VIF = infinity, the corresponding variable may be expressed exactly by a linear combination of other variables.
- We need to drop the variable from the dataset that is causing this perfect multicollinearity.

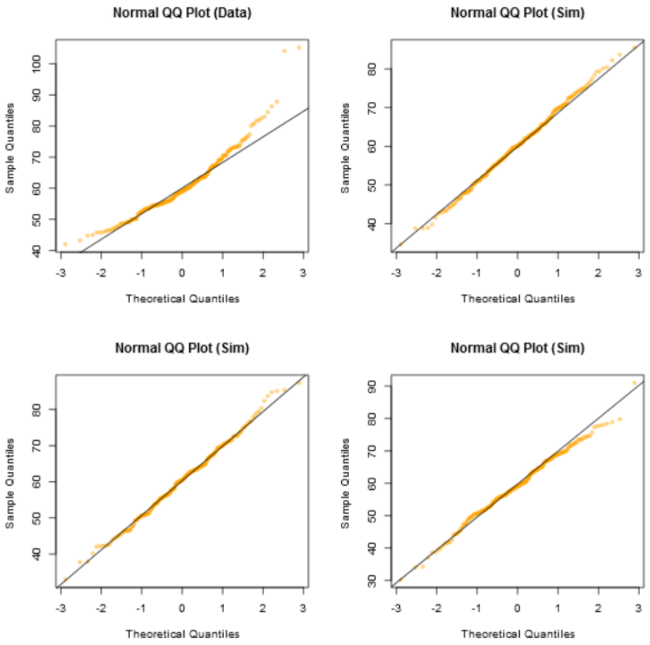
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

A:

A Quantile-Quantile (Q-Q) plot is used to compare two data sets from a population having the same distribution or not. For example, it can compare the uniform, normal, and exponential distributions of the two datasets. It compares properties like location, scale, and skewness.

A 45-degree reference line on the Q-Q plot



The median is a quantile where 50% of the data falls below that point and 50% lies above it. A 45-degree line is plotted on the Q-Q plot.

If two datasets have the same distribution, then the points lie exactly or approximately on the 45-degree line. If the two distributions are different, then points may lie above or below the 45-degree line.