Linear –Regression Assignment

- Nithin John Jacob

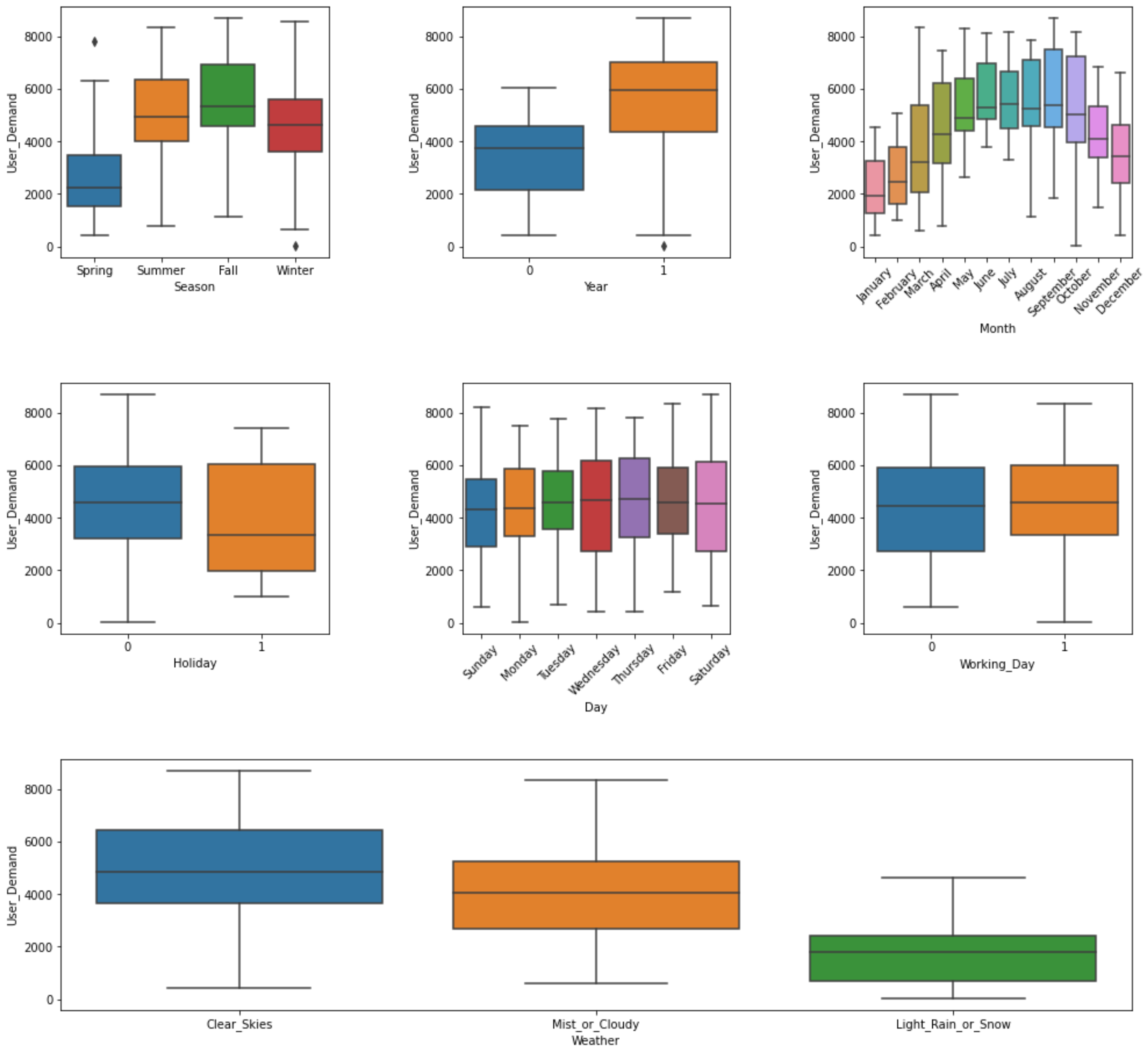
Assignment-based Subjective Questions

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

The following are the inference from the data:

Season: The average demand in fall is more than other seasons. It is the lowest in spring
Year: The average demand in 2019 is much higher to 2018.
Month: The average demand goes up during July – September while the lowest is around January.
Weather Situation: The demand goes down as weather becomes bad. Clear skies has the maximum demand.
Working Day: The average value does not seem to be different for working and non-working days
Day: The average demand is more or less same throughout the week.
Holidays: The average demand drops during non-weekend holidays.

Distribution of user Demand by Categories.



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

n-levels of a categorical column can be represented by n-1 variables. It is better to have less number of columns for easy modelling. The extra features are often cause multicollinearity and often lead to overfitting. Also the increased number of variables takes more memory and processing power and makes the modelling slower.

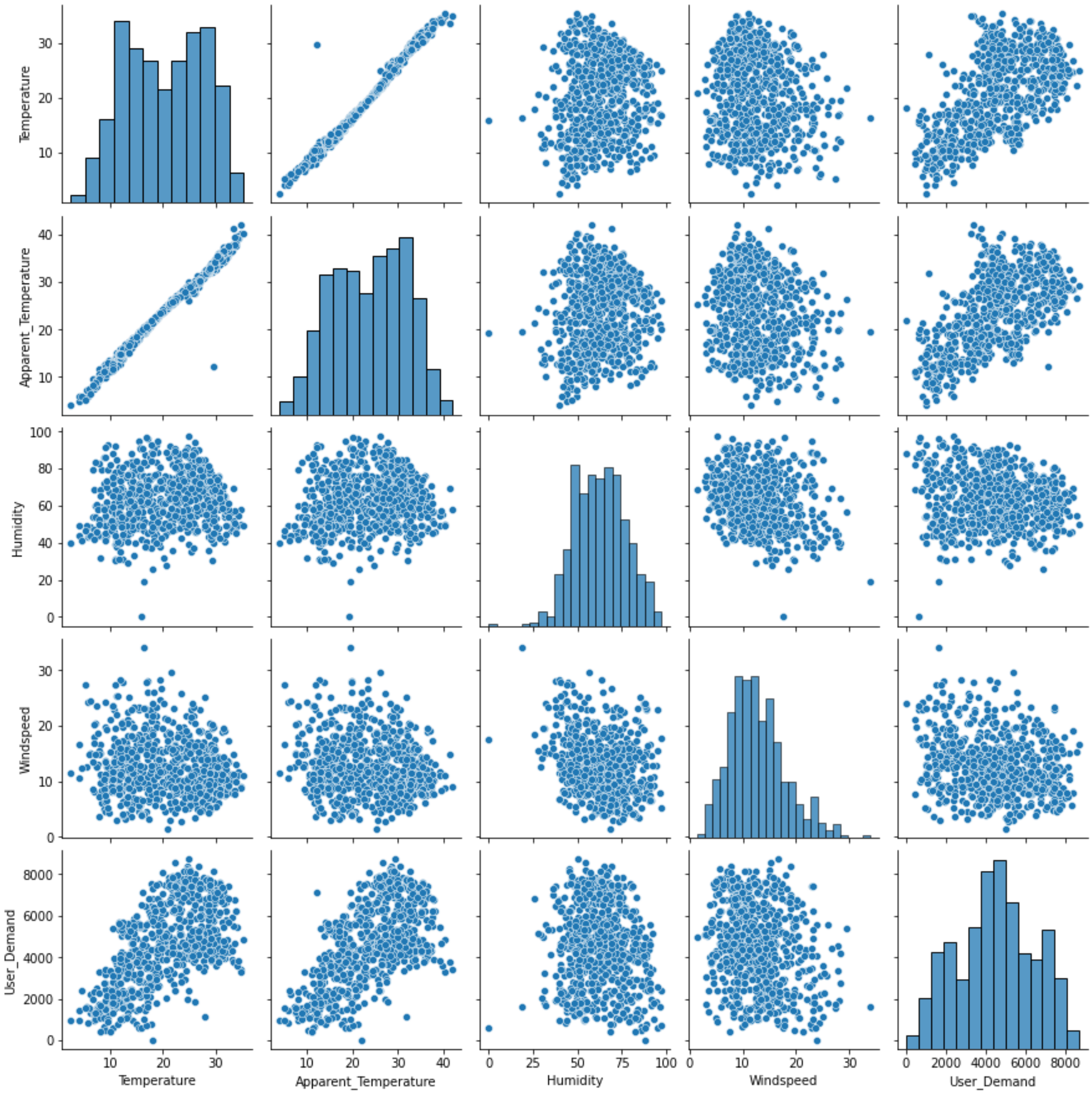
Eg: The first column given in the example below can be removed without losing any information as it is redundant. 00 will represent Maybe 10 Yes and 01 NO.

Option
Maybe
Yes
No

Option_Maybe	**Option_Yes**	**Opinion_No**
1	0	0
0	1	0
0	0	1

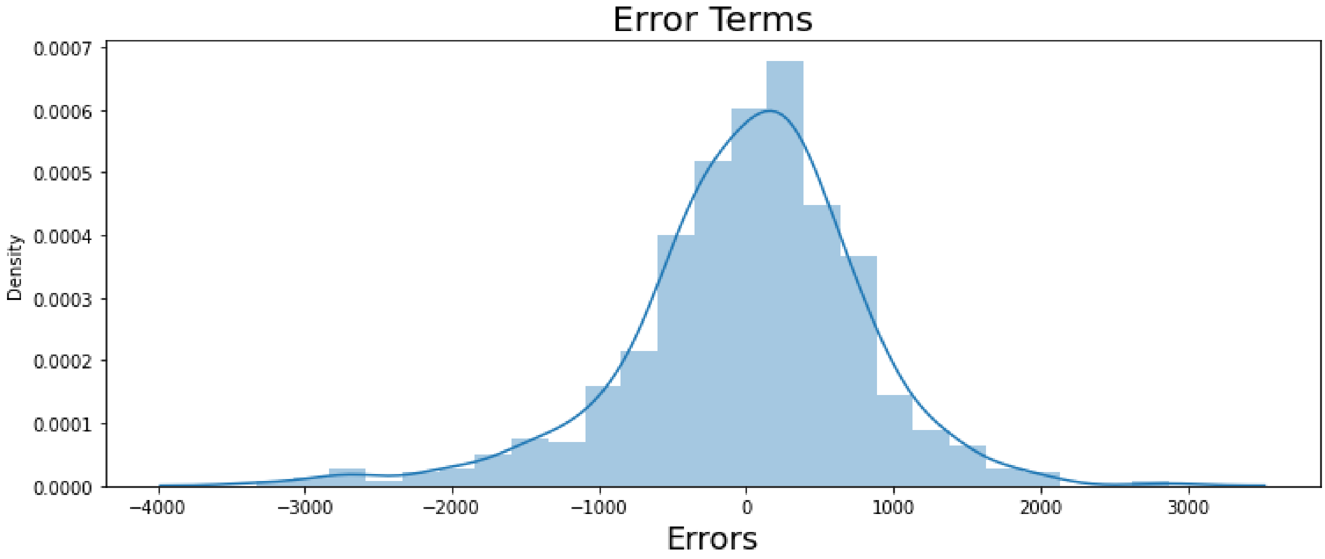
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

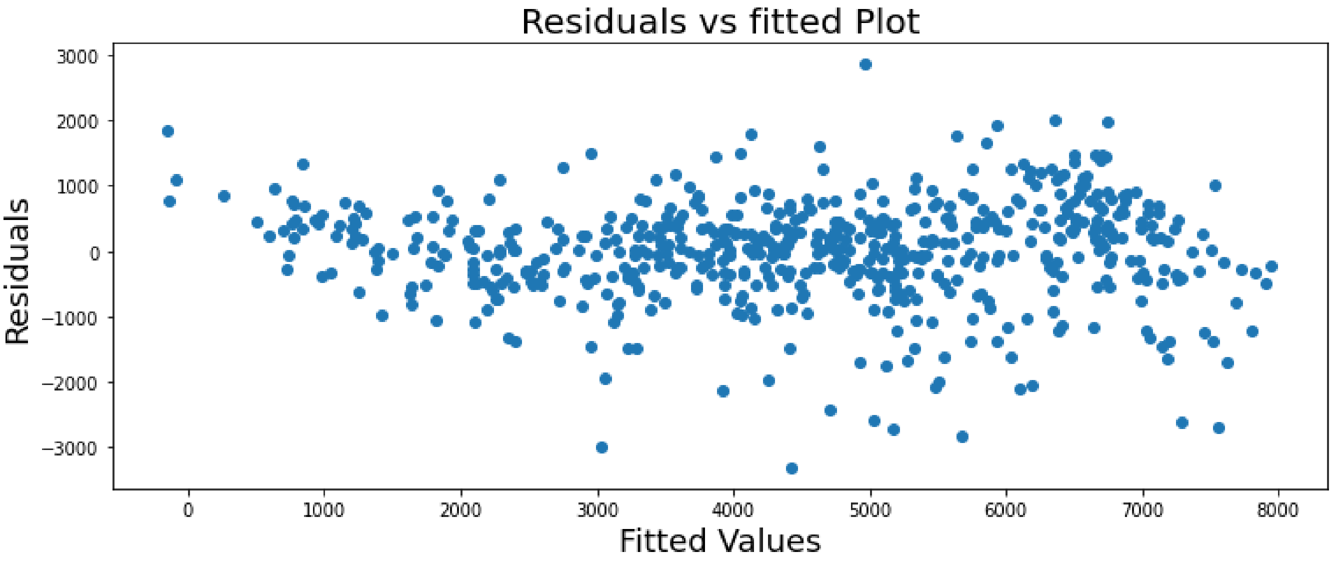
The Temperature and Apparent temperature has highest correlation with the target variable.



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

The residual analysis was done to see that the error terms are also normally distributed.





Multicollinearity was checked by verifying the VIFs of the model.

The Residuals vs fitted model was checked for constant variance.

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

Top Three Features are :

5347*Temperature -2054*Weather_Light_Rain_or_Snow + 1951*Year

Linear –Regression Assignment

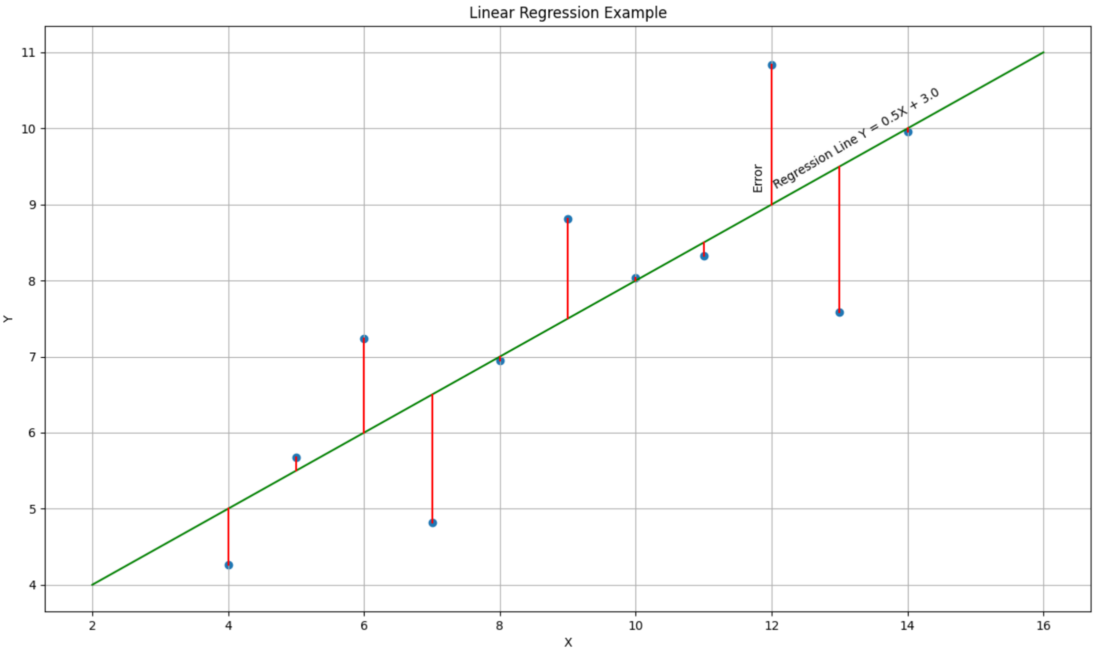
- Nithin John Jacob

General Subjective Questions

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

Regression is a type of Supervised Machine learning model, where the output variable is a continuous variable. A simple linear regression links a dependent (**output variables**) (Y)and an independent (**predictor variable**) (X) variable using a straight line. The basic idea is to use a set of data to calculate the relationship, and then use that and the new X to predict Y.

In regression, there is a notion of a best-fit line — the line which fits the given scatter-plot in the best way.The best-fit line is obtained by minimising the **cost function**. The cost function used here is a quantity called Residual Sum of Squares (RSS).



Linear regression are of two types:

Simple Linear Regression is used when dependent variable is predicted using one independent variable. The equation is of the form Y = $\beta_0 + \beta_1 X$

Multiple Linear Regression is used when dependent variable is predicted using n independent variable. The equation is of the form Y = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$

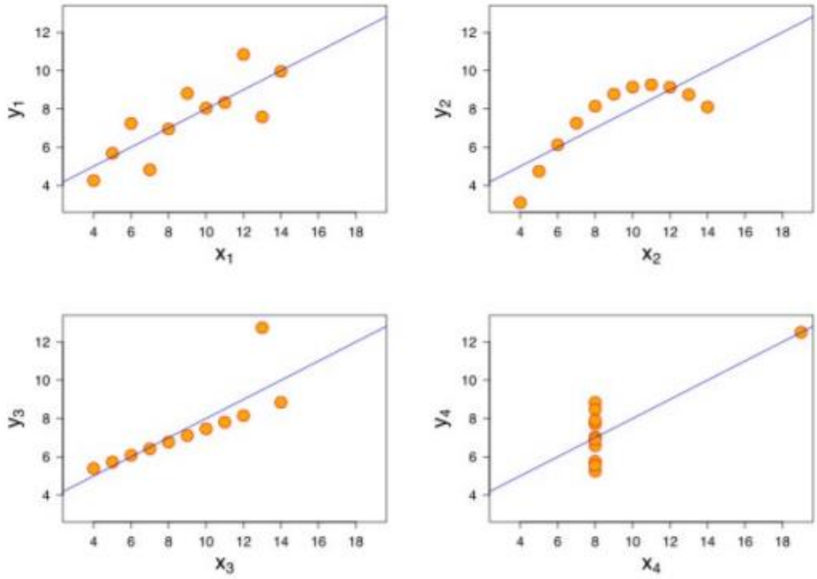
Cost function, $\text{RSS} = \sum_{i=0}^{n}(\hat{y}_i - y_i)$

We find the β_0 , β_1, …β_n for a minimum RSS.

Gradient Descent algorithm is one of the methods used to reach an optimal solution

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

Anscombe’s quartet is a group of dataset made by statistician **Francis Anscombe in** 1973. This data set was carefully created with identical statistical parameters to *emphasis the importance of data visualization* over just looking at statistical parameter. These four dataset although statistically equivalent render completely different graphs.



Anscombe's quartet

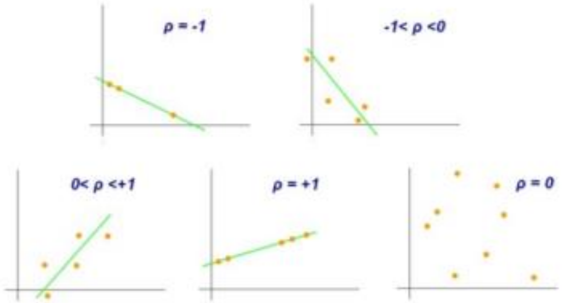
	I		II		III		IV	
	x	y	x	y	x	y	x	y
Data	10.00	8.04	10.00	9.14	10.00	7.46	8	6.58
	8.00	6.95	8.00	8.14	8.00	6.77	8	5.76
	13.00	7.58	13.00	8.74	13.00	12.74	8	7.71
	9.00	8.81	9.00	8.77	9.00	7.11	8	8.84
	11.00	8.33	11.00	9.26	11.00	7.81	8	8.47
	14.00	9.96	14.00	8.10	14.00	8.84	8	7.04
	6.00	7.24	6.00	6.13	6.00	6.08	8	5.25
	4.00	4.26	4.00	3.10	4.00	5.39	19	12.50
	12.00	10.84	12.00	9.13	12.00	8.15	8	5.56
	7.00	4.82	7.00	7.26	7.00	6.42	8	7.91
	5.00	5.68	5.00	4.74	5.00	5.73	8	6.89
Mean	**9.00**	**7.50**	**9.00**	**7.50**	**9.00**	**7.50**	**9.00**	**7.50**
Standard Deviation	**3.32**	**2.03**	**3.32**	**2.03**	**3.32**	**2.03**	**3.32**	**2.03**
Correlation between x and y	**0.816**		**0.816**		**0.816**		**0.816**	
Linear regression line	**y = 3.00 + 0.500x**		**y = 3.00 + 0.500x**		**y = 3.00 + 0.500x**		**y = 3.00 + 0.500x**	
Coefficient of determination of the linear regression, R^2	**0.67**		**0.67**		**0.67**		**0.67**	

3. What is Pearson’s R? (3 marks)

Pearson's r also known as Pearson correlation coefficient (PCC), Pearson product-moment correlation coefficient (PPMCC) is a numerical index that reflects the relationship between two variables. It was developed by Karl Pearson. It is given by:

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left[n\Sigma X^2 - (\Sigma X)^2\right]\left[n\Sigma Y^2 - (\Sigma Y)^2\right]}},$$

- r_{xy} is the correlation coefficient between X and Y;
- n is the size of the sample;
- X is the individual’s score on the X variable;
- Y is the individual’s score on the Y variable;
- XY is the product of each X score times its corresponding Y score;
- X^2 is the individual’s X score, squared;
- and Y^2 is the individual’s Y score, squared.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

A correlation can range in value from −1.00 to +1.00. The correlation is called a direct correlation or a positive correlation if variables change in the same direction. The correlation is called an indirect correlation or a negative correlation if variables change in opposite directions.

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

Feature Scaling is used to transform the features to a uniform range. If the features are in different scales, the LR algorithm gives more weightage to higher ranged values and less weightage to low range values. Model may not capture the impact of low range values. Normalization and Standardization are the most commonly used scaling Techniques.

Normalization is a scaling technique in which values are shifted and rescaled from 0 to 1. It is also known as Min-Max scaler.

It is calculated as X' = (X - Xmin)/(Xmax - Xmin)

Standardization is another scaling technique in which values are centred around mean with a unit standard deviation. The new distribution will have 0 as mean and 1 as standard deviation.
It is calculated as X' = (X - μ)/σ

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

Variance Inflation Factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity. It is given by (VIF) =1/(1 - R_i^2).

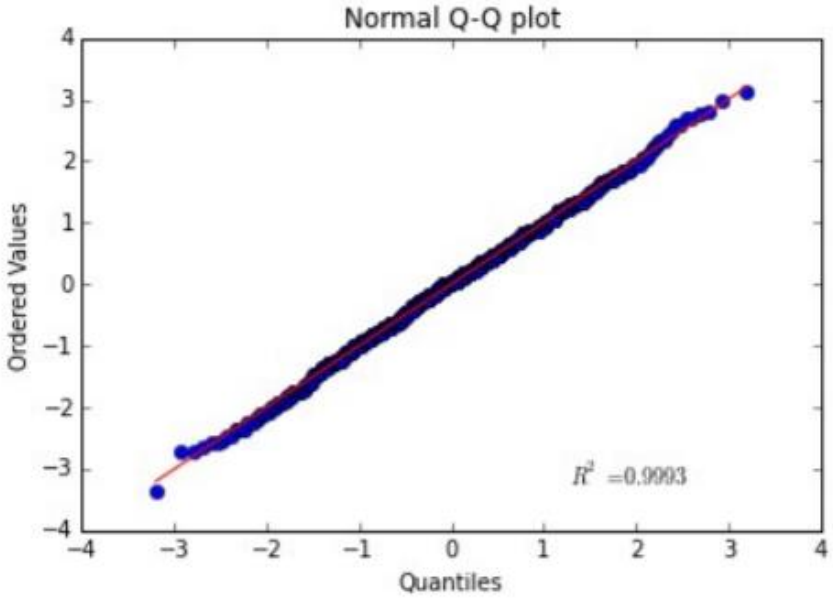
This VIF becomes very large when R_i^2 becomes close to 1. At R_i^2 =1, the equation becomes 1/0. The Infinite value hints a perfect correlation between the independent variables.

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

A Q–Q (quantile-quantile) plot is a probability plot used in statistics. It is used mainly to compare two distributions. We plot the different quantiles of a distribution on y-axis and similar quantiles for second distribution on x-axis to compare the two distributions.

eg: we can compare our test dataset with any standard Normal plot to see if test dataset is normal or not. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions.



Source: StackExchange Output Q-Q Plot