MACHINE LEARNING ASSIGNMENT (SUBJECTIVE)

Submitted by : **SOURAV KARMAKAR**

Email Id : **skrmkr1992@gmail.com**

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: The effect of the categorical variables on the dependent variable (cnt) are as below:

Season: When we look into the boxplot of season vs cnt, we see that the demand of bike increases in the seasons 2 and 3. Here season 2 and 3 are summer and fall respectively. The count is lesser in the other two seasons of the year i.e. in season spring and winter. The highest count of bike shares are recorded during fall.

Yr (year): If we look into the plot yr vs cnt, we see that the demand of bike is much higher in the year 2019 than in 2018.

Mnth (Month): Looking into the plot we can say that the demand stays higher during the months of June to October. And the highest demand may fly around September.

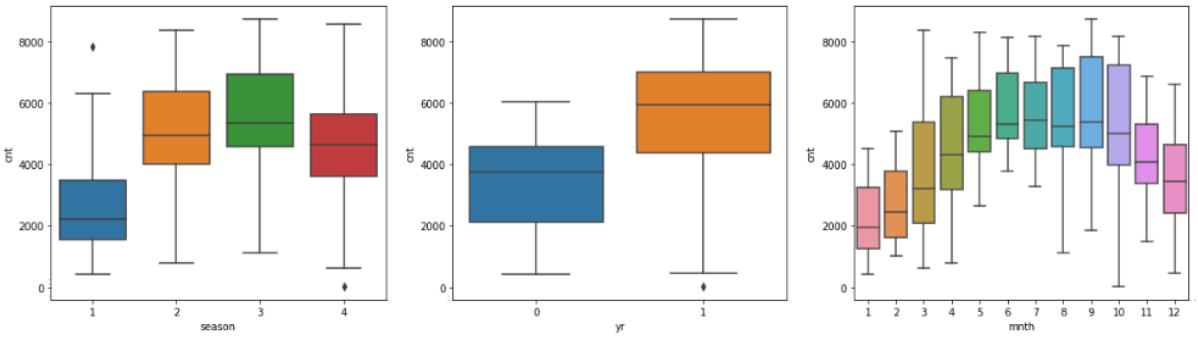
Holiday: If the day is not a holiday, the demand is higher than the day if it's a holiday. We can say that by looking into the plot since the median for a day which is not a holiday is much higher than that of a holiday.

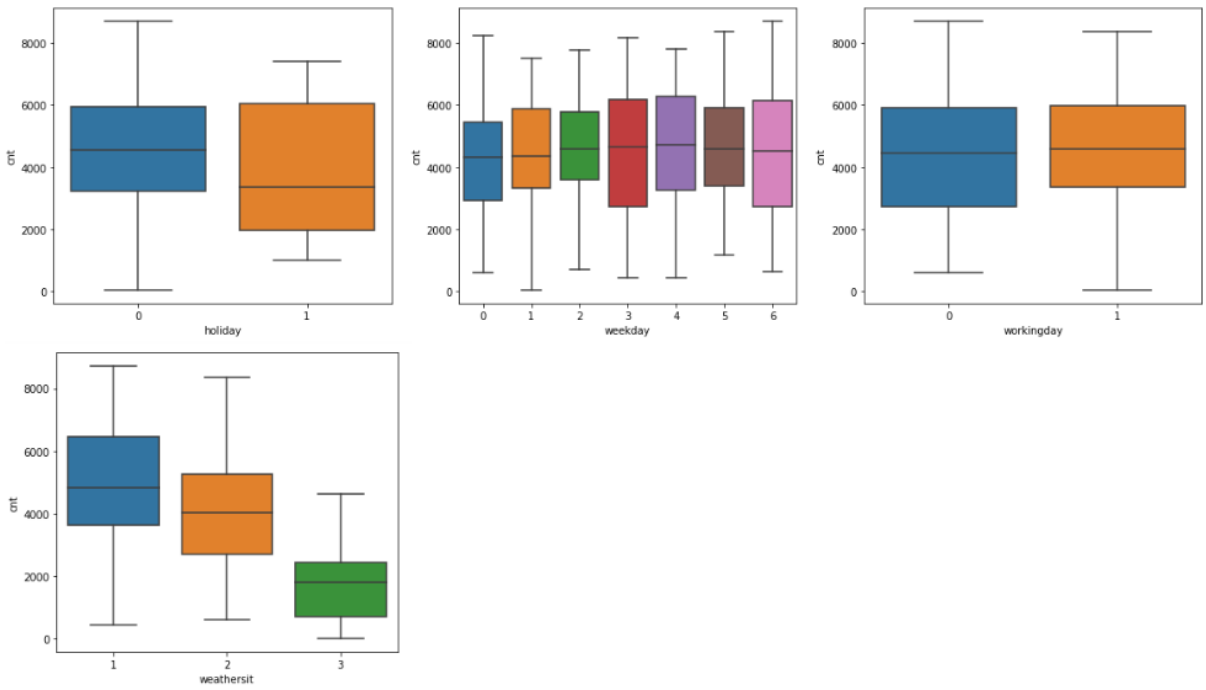
Weekday: By looking into the plots of weekdays vs cnt we can say that there is not much difference amongst the days of a week since the median for each days are almost similar.

Workingday: On a working day of the week, the demand is lesser that that of a non-working day. Even though the median is almost same, the distribution of the plot is bigger in case of a non-working day which helps us to infer this idea.

Weathersit (Weather situations): Bu looking into this plot, we can easily understand that if the weather condition is Clear, Few clouds, Partly cloudy or Partly cloudy, the demand of bikes are the highest. The demand decreases as the weather gets extreme.

Below are the respective plots to support the above explanation:





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

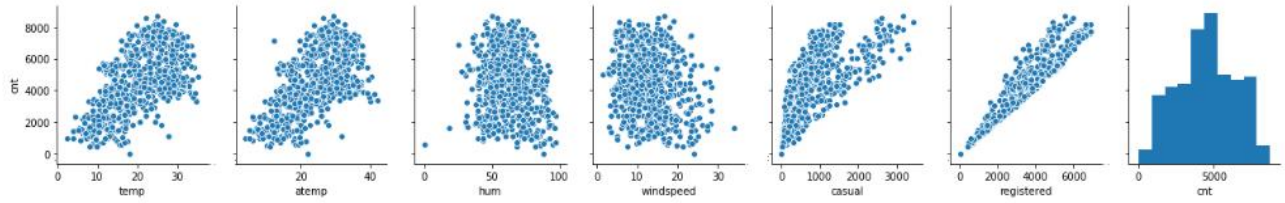
Ans: While dummy variable creation for any categorical variables, the number of dummies created is one less than the number of categories the actual variable has. If the categorical variable has K number of categories, the number of dummy variables required to explain all the categories is (K-1). That is the reason to use drop_first=True while creating dummy. Let's look into the below example to understand better.
The categorical variable Blood_group has 4 different categories i.e. A, B, AB and O. So if we create 4 dummies and represent the categorical variable in tabular format, it will look like below:

A	B	AB	O
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

So, group A means 1 in A and all others are 0, group B means 1 in B and all others 0 and so on.
So we can assume A=1000, B=0100, AB=0010 and O=0001. Now if we drop the first group, we can say that B=100, AB=010, O=001 and also A=000. This means, we can understand each of the categories with one less dummy variables. This is why the first dummy variable is dropped as it reduces one extra variable without impacting any efficiency of the model.

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

Ans: The pair plot of the numerical variables looks as below (only the target variable row is picked for explaining the answer):

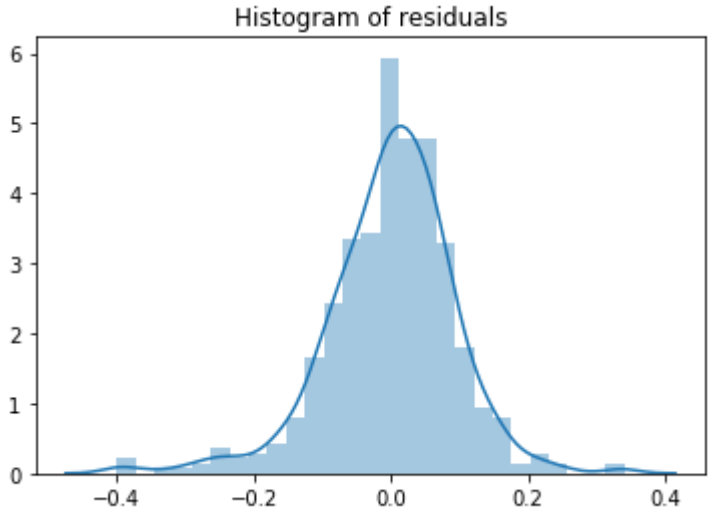


By looking into the above plots, we can say that the variable "registered" has the highest correlation with the target variable "cnt". However, as the variables "casual" and "registered" combined creates the target variable "cnt", the next variable that has highest correlation with "cnt" variable is "temp".

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

Ans: The major assumptions of Multiple Linear Regression that are checked after building the model were as below:

a) Error terms are normally distributed: After building the model, I calculated the residuals for all the predicted points by calculating the difference between the actual target value and the predicted target value. Then I plotted the histogram for the residuals and got a normally distributed histogram as below:



This diagram also proves one more assumption to be true and that is the mean of the residuals are lying around zero (0).

b) The 2nd assumption i.e. there is no multicollinearity was proved by calculating the VIF for all the variables from the final model. The calculated VIF for all the variables are below 5 which is the acceptable range for MLR. Below is the table for VIF of all the variables from final model:

	Features	VIF
1	atemp	4.92
2	windspeed	3.02
0	yr	2.00
3	summer	1.83
5	aug	1.54
4	winter	1.50
7	cloudy	1.49
6	sep	1.30
8	light rain	1.08

c) The final model was including all the variables from the given dataset. It only contains few of the features which are the most affecting ones. With these particular features, the r-squared value is the optimum. If we add more features in the model, the value of r-squared does not increase much. Also, if we remove any of the variables, the r-squared value is reduced drastically. This also proves that feature selection is also a very important aspect of model building.

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

Ans: The final model equation is as below:

cnt = 0.1226 + (0.2337 * yr) + (0.5412 * atemp) - (0.1334 * windspeed) + (0.1005 * summer) + (0.1302 * winter) + (0.0694 * aug) + (0.1180 * sep) - (0.0807 * cloudy) - (0.2718 * light rain)

From the above equation, we can see that the top three features contributing significantly towards explaining the demand of shared bikes are:

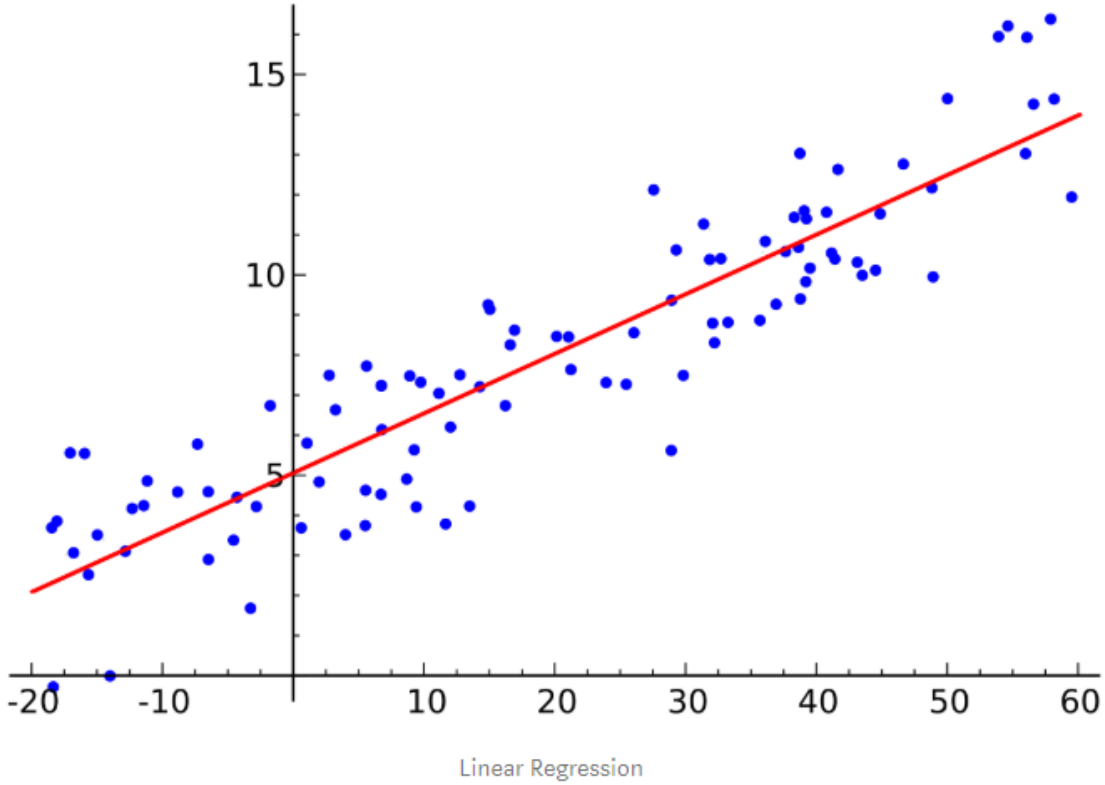
1) atemp (feeling temperature) – Coefficient = 0.5412
2) light rain – Coefficient = -0.2718
3) yr (year) – Coefficient = 0.2337

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Regression is a method to build a model to describe the target variable based on one or more independent variables. When the relationship between the independent variable(s) and the dependent variable or the target variable is linear in nature, the type of regression is called linear regression.
Based on the number of independent variables or also called as predictors, linear regression is divided into two categories i.e. a) Simple Linear Regression (SLR) – one independent variable and b) Multiple Linear Regression (MLR) – more than one independent variables.

In Simple Linear Regression, the model can be expressed in the form of a simple linear equation of a straight line. The diagram of SLR is as below:



The model equation for SLR is: $Y = \beta0 + \beta1*X$
Where Y is the target variable (dependent)
β0 is the intersection from the X-plane
β1 is the increase in value of Y for unit change in X
X is the independent variable or the predictor

The red line in the above diagram is called the best fit line (for SLR). The significance of this line is that the model designed predicts the best target variable on the line. For MLR, this becomes a best fit on a 'hyperplane' since there are multiple predictor variables.

The difference between actual Y and the predicted Y is known as residual.
Residual(Ɛ i) = Yi – Y pred [Y pred = predicted target var]

Residual Sum of Squares (RSS) = $Ɛ\ 1^2 + Ɛ\ 2^2 + Ɛ\ 3^2 + \ldots\ldots\ldots + Ɛ\ n^2$
RSS is also known as Cost Function or Objective Function.

$$RSS = \sum_{i=1}^{n} (Yi - \beta0 - \beta iXi)^2$$

The best fit line is determined by minimizing the quantity called Residual Sum of Squares or Cost Function.

There are two different methods for reduction Cost Function:

1. Differentiation
2. Gradient Descent

Gradient Descent is an optimization algorithm that optimizes the Cost Function or Objective Function to reach the optimal solution.

Total Sum of Squares (TSS) = $(Y1 - \bar{Y})^2 + (Y2 - \bar{Y})^2 + (Y3 - \bar{Y})^2 + \ldots.. + (Yn - \bar{Y})^2$

$= \sum_{i=0}^{n} (Yi - \bar{Y})^2$ [$\bar{Y}$ - average of Y]

$R^2 = 1 - \frac{RSS}{TSS}$ [RSS: Residual Sum of Squares, TSS: Total Sum of Squares]

The higher the value of R^2 for any model, the better the model is.

Assumptions of Single Linear Regression:

1. The target variable and the independent (input) variable are linearly dependent and there is no other relationship is there for them.
2. Error terms or residuals are normally distributed
3. Error terms are independent of each other
4. The variance or standard deviation of the error terms are constant

MLR or Multiple Linear Regression: the best fit for MLR is a hyperplane unlike a line for SLR. The coefficients for MLR are also obtained by the same method as SLR which is minimizing the sum of squared errors (RSS).

For MLR, the equation of model looks like: $Y = \beta0 + \beta1*X1 + \beta2*X2 + \beta3*X3 + \ldots..$
Interpretation: Change in the mean response E(Y) is the per unit increase in the variable (X) when other predictors or independent variables are held constant.

Assumptions:

1. The target variable and the independent (input) variable are linearly dependent and there is no other relationship is there for them.
2. Error terms or residuals are normally distributed
3. Error terms are independent of each other
4. The variance or standard deviation of the error terms are constant
5. Adding more independent variables in the model is not always helpful

Adding more variable to the model could create two problems:

a) Overfitting: When due to large number of predictor variables, the model starts remembering each and every points fitting too well to the train set and cannot generalize. Also the model becomes very much complex.
b) Multicollinearity: Due to large number of predictor variables, associations between predictor variables start showing.

Due to multicollinearity, the interpretation of MLR does not hold true. Also, there are large swing in the coefficients, signs can get changed to negative and also p-value becomes unreliable.

VIF = $\frac{1}{1-R^2}$: is one method to identify multicollinearity between variables. If the multicollinearity of any variable is more than 5 and the variable is statistically insignificant to the target variable, we can drop that variable and rebuild the model once again.

Scaling of numerical variables are done to bring all the numerical variables in a common range so that the model predicts the target variable properly. Scaling only affects the coefficients of the variables and not any other parameters such as t-statistics, f-statistics, p-value, R-squared etc.
Two different process of scaling numerical variables are:-

1. Standardization
2. MinMax Scaling (Normalization)

Feature Selection: This is one of the most important aspect for building the best possible model for a set of data. As described earlier, not all features participate in the best model and also adding more features not always help in improvement of the model. Hence, feature selection is very important to build a good model. There are few methods to select most significant features. Those methods can be broadly divided into two categories:

1. Manual Feature Elimination
   - Build model
   - Drop features that are least helpful in prediction (high p-value)
   - Drop features that are redundant (using VIF)
   - Rebuild model and repeat
2. Automated approach
   - Top ‘n’ features: Recursive Feature Elimination (RFE)
   - Forward Selection
   - Backward Selection
   - Stepwise selection (Based on AIC)
   - Regularization
3. Mixed Approach: Combination of Automated approach (RFE) to select top ‘n’ features and then fine tune the model by using manually eliminating least significant features one by one to finally build a model with highest possible R-squared value.

Q2. Explain the Anscombe's quartet in detail.

Ans: When we have a spreadsheet of business data we can easily calculate the statistical features of the data such as mean, median, variance, correlation between x/y variables and even the best fit line of the model built from the data. However, even though we can get a fare idea of the model built, only relying on the summery is very much dangerous since it draws no conclusion on the distribution of the data.

The most elegant demonstration of the danger of relying only on the summery statistics is Anscombe's Quartet. It's a group of four datasets each containing eleven pair (x, y) of data which seem similar when only looked into the summary statistics but are extremely different when plotted.

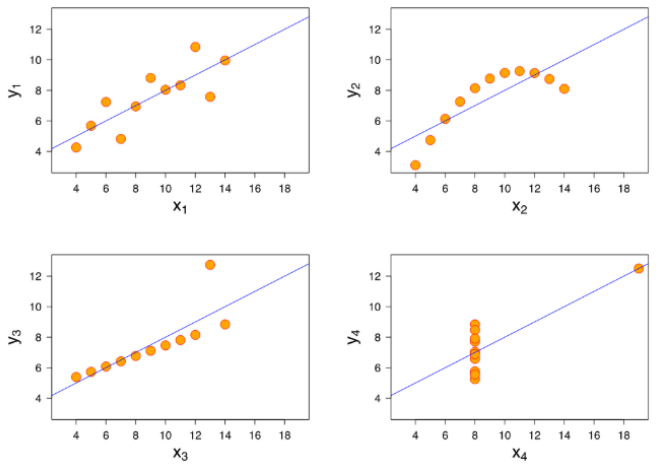
Let's take the below four data sets for example:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The statistics that we can think of are as below:

- The average value of x variables are same for 4 data sets (9)
- The average value of y variables are same for 4 data sets (7.50)
- The variance or SD for x and y for all 4 sets are also same (11.0 and 4.12 respectively)
- The correlation between x and y for four sets are also same (0.816)
- The best fit line (LR) equation for four models are also represented with same equation i.e. y = 0.5x + 3

By looking inti the above statistics, it seems that the four datasets are very similar. But when the data sets are plotted, we see the actual difference is in the distribution. Below is the diagrams when they are plotted:



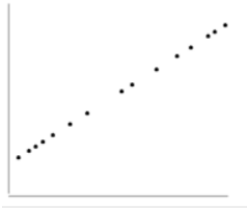
From the plot 1, we can say that there is a rough linear relationship between x and y.
From plot 2, the relation between x and y is not linear.
Plot 3 indicates a strong and tight linear relationship between x and y variable and plot 4 shows that the value of x remains constant.
So, by using Anscombe's Quartet, we can understand the real story of the data. Also, it means that only looking into the summary does not let us understand what is going on, we need to visualize the data to understand it completely.

Q3. What is Pearson's R?

Ans: Pearson's Correlation Coefficient (aka Pearson's R) is a measure of the strength of association between two quantitative, continuous variables. To understand if there is a linear relationship between two continuous variables or not, a scatter plot is drawn while putting the dependent variable in y-axis and the independent variable in x-axis. If the scatter is nearer a straight line, it means there is a strong association between the two variables.

The value of Pearson's Correlation Coefficient ranges from +1 to -1.

When R (Pearson's Correlation Coefficient) = +1, it means the two variables are positively correlated and the association can be shown using a perfectly straight line with a positive slope. Below is the scatter plot for R = +1



Positive correlation coefficient means, when one variable increases, the other variable also increases and vice-versa.

When R = -1, it means there is a strong negative association between the two variables. The relation can be shown by a perfectly straight line with a negative slope. Below is the scatter plot for R = -1

Negative Correlation Coefficient means, when one variable increases the other variable decreases and vice-versa.

When R tends to 0 (zero), it means there is no linear association between the two variables. A scatter plot to understand this scenario better looks like below:

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

Ans: Scaling: Scaling is a method to standardize the continuous variables when the variables hold data of different ranges. For example, in a data set there are many variables say velocity, mass etc. If the values of velocity ranges from 20 km/h to 60 km/h and the values of mass ranges from 10000 mg to 80000 mg and so on for other variables, the inference or prediction of the dependent variable will not be correct since the difference between minimum and maximum values will be high.

Scaling is performed during pre-processing to handle highly varying values or units. Without scaling, the model understands higher values, higher and lower values, lower irrespective of their units. For example, the model will consider 5000m higher than 10km if scaling is not done even though that is wrong in real life.

Difference between Normalized Scaling and Standardized Scaling:

Normalized Scaling: Normalized Scaling is also known as MinMax Scaling. In this method, data is normalized in such a way that the range of data becomes 0 to 1. The formula used for Normalized Scaling is as below:

$$X = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Standardized Scaling: In this method of scaling, the data are scaled in a way that the mean becomes 0 and the standard deviance becomes 1. The formula used in case of standardization is as below:

$$X = \frac{X - X\ \text{mean}}{Standard\ Devation}$$

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

Ans: VIF is an index to understand how much a variable is correlated with other variables i.e. multicollinearity. If there is multicollinearity between variables, we assume the value of VIF to be less than 5 for a good linear regression model. However, VIF can be more than 5 or even infinity. The formula to determine VIF is as below:

$$VIF = \frac{1}{1-R^2}$$

The value of R^2 ranges from 0 to 1. So the less the value of denominator, the more the value of VIF. Which means, $1 - R^2$ tends to zero

Or, R^2 tends to 1.

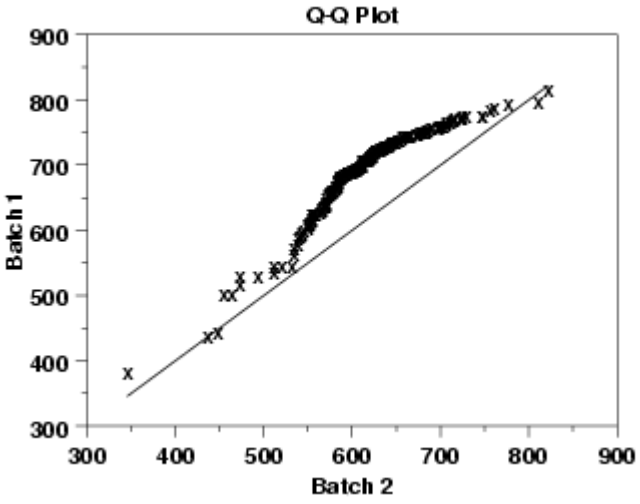
We know that more the value of R^2, better is the linear regression model. This means, the variable affects the target variable in a huge extent or the variable is in strong association either directly to some other independent variable or a combination of variables.

For VIF to be infinity, we can see from the above equation, $1 - R^2 = 0$
Or, $R^2 = 1$

This means, the model exactly defines the predicted variables without any error or the predictor variable is in absolute association with 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 or Quantile-Quantile plot is a graphical representation tool to understand if two given datasets came from populations with same distribution. This is a plot where in one axis, the quantiles of probability distribution are put and in the other axis, the quantiles of probability distribution for the other data set are put. A sample quantile plot is shown below:



After plotting the points (x, y), a 45* reference line is plotted. If the points are exactly falls around the reference line, it means that the two datasets came from population of same distribution.

However, if the plot (points) are departed from the reference line with considerably high value, it signifies that the data sets are coming from population with different distributions.