Boom-Bikes

LINEAR REGRESSION ASSIGNMENT

Daniella Brito| Linear Regression| 24-12-2020

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?

Answer: The plots above shows the relationship between categorical variables and a Target variable.

• Bike Rentals are more during the Fall season and in Summer .

• Bike Rentals are more in the year 2019 compared to 2018

• Bike Rentals are more in clear weather condition and in times when it is partly cloudy .

• Bikes are rented more on Working days and more during the months- September and August and the least during January and possible reason could be people are out on holidays to other places .

• Bike rentals are observed at higher "feel-like" temperatures and temperatures in general.

• More bike rentals seem to be made when the humidity is higher.

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

Answer: Here drop_first=True is important to use as it helps in reducing the extra column created during the dummy variable creation. Hence it reduces the correlations created among the dummy variables.

More dummy features make it harder for the Algorithm to fit or even worse make it easier to overfit. To avoid over fitting of the model and to avoid the redundant values.

In this assignment we converted all the categorical values to dummy variables and then applied the drop_first to them using get _dummies function. Using drop_first=True is more common in statistics and is often referred to as "dummy encoding" while drop_first=False refers to one hot encoding in ML.

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

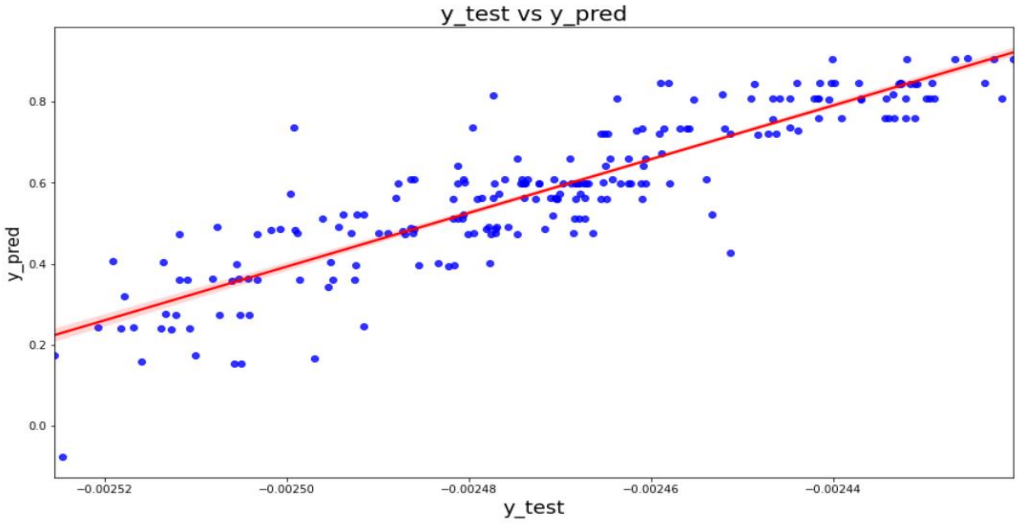
Answer: Bike rentals are more correlated to registered as 0.95, casual as 0.67 and temp as 0.63 when we consider count to be the target variable , highest being registered which is count of registered users .

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

Answer: We arrived at a very decent model for the the demand for shared bikes with the significant variables. We can see that year variable is having the highest coefficient 0.2459, which means if the temperature increases by one unit the number of bike rentals increases by 0.2459 units. Similarly we can see coefficients of other variables in the equation for best fitted line. We also see there are some variables with negative coefficients. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease. We have Spring, mist cloudy , light snow and summer variables with negative coefficient. The coefficient value signifies how much the mean of the dependent variable changes given one unit shift in the independent variable while holding other variables in the model constant.

The steps to validate can be –

- The predicted values have linear relationship with the actual values.



- The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.
- No Autocorrelation in residuals . Using the Durbin-Watson Test.

```
==============================================================================
Omnibus:                       60.252   Durbin-Watson:                   1.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              145.320
Skew:                          -0.619   Prob(JB):                     2.78e-32
Kurtosis:                       5.303   Cond. No.                         8.79
==============================================================================
```

When DW=2 would be ideal where there is no autocorrelation . In our case the value is 1.945 which indicates a positive autocorrelation.

0<DW<2 - indicates a positive autocorrelation

0<DW<4 – indicates a negative autocorrelation

- No perfect multicollinearity . Another common way to check is by calculating the VIF (Variance Inflation Factor) values. Where we know

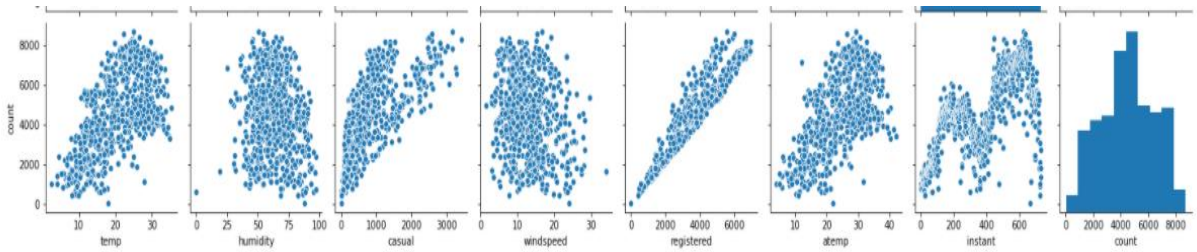
  If VIF=1 , Very less Multicollinearity

  If VIF<5 Moderate Multicollinearity

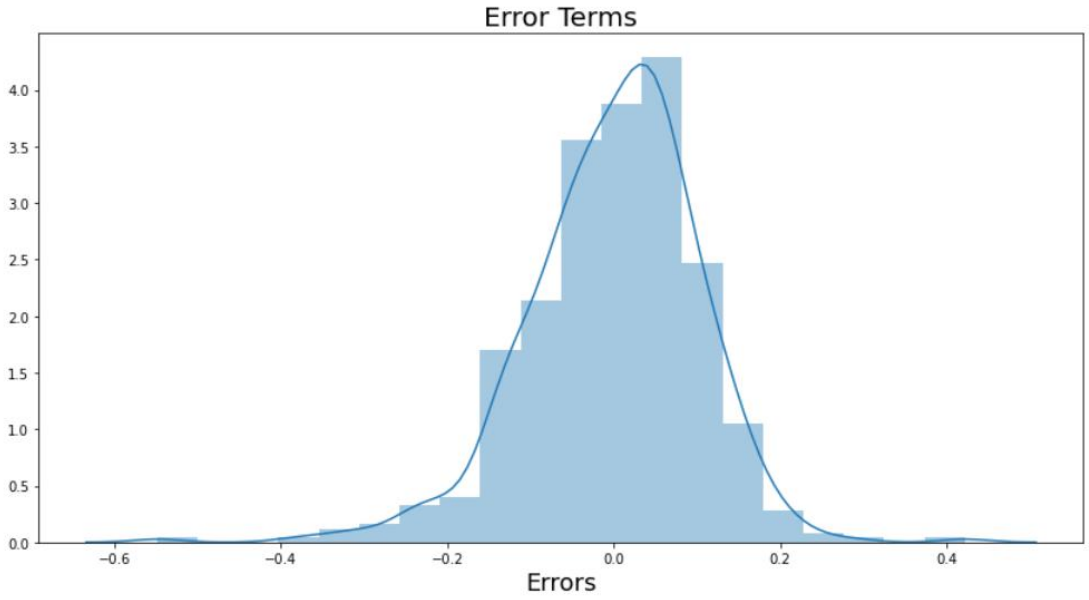
  If VIF is very high we removed a few of them in our case. To obtain a good model.

	Features	VIF
2	windspeed	3.23
3	season_spring	2.37
0	Year	1.74
4	season_summer	1.70
6	month_Jan	1.62
10	weathersit_Mist + Cloudy	1.48
7	month_Nov	1.23
8	month_Sep	1.16
5	month_Dec	1.13
9	weathersit_Light Snow	1.07
1	holiday	1.06

- A pair-plot can help us know if the independent variable has a relationship with the dependent variable . As we can observe in this case if we take the target variable as count in this case .



- The error terms are normally distributed as observed from the assignment .



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

Answer:

- Year : 0.2459
- Sep : 0.0607
- Spring season : -0.2367
- Jan : -0.1216
- Mist :-0.3647

We can see that bike-sharing provider Boom Bikes can focus more on Year and time for the business to pickup its pace .We can see demand for bikes was more in 2019 than 2018, so just focus on time as there is an increase in 2019 . We can focus more on Summer & Fall season, September months, weekends, working days as they have good influence on bike rentals. We also can see that there are some variables with negative coefficients which suggests that as the independent variable increases , the dependent variable tends to decrease. We have spring , mist and cloudy and light snow variables with the negative coefficient.

GENERAL SUBJECTIVE QUESTIONS

1 .Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being x and y as dependent variable . Here the idea is to estimate the future value based on the historical data by learning the behavior or patterns from the historical data.

Regression analysis is used for three types of applications:

1. Finding out the effect of Input variables on Target variable.
2. Finding out the change in Target variable with respect to one or more input variable.
3. To find out upcoming trends.

Here are the types of regressions:

1. Linear Regression
2. Multiple Linear Regression
3. Logistic Regression
4. Polynomial Regression
5. Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Where x and y are two variables on a regression line -

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

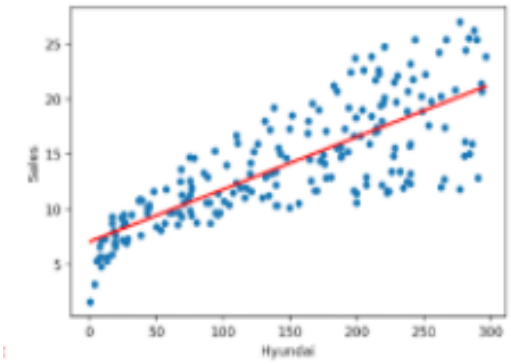
y = Dependent variable from dataset

Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

- Let's visualize the relationship between the features and the sales response using scatterplots.
- Now let's Estimate the model coefficients for Linear Regression by using single feature to predict quantitative response.
- To calculate coefficients, we will use the least square criterion, which means we will find a line that will decrease the sum of squared errors.
- In this step we will load stats models to estimate the model coefficients for the advertising data. Stats models allows users to fit statistical models by importing OLS. As shown below we are going to fit the model using stats models OLS.
- From step 4- we got the value of A and B. we will use the model to predict the future sales . Let's say in the new market is spending 50 thousand dollars in advertising. That means the new value of X will be 50. Now using Y = A + BX to predict the new value.

- Now let's plot the least square line by creating a data frame with the minimum and maximum values and predict for x value and store that value in preds variable. Let's plot the observed data graph and the least square line using preds value and new x value



In linear regression, the observation blue dots are assumed to be the result of random deviation from an underlying relationship (red line) between a dependent variable (y) and an independent variable (x). Here the goal is to decrease the distance between the red line and the blue dots. If all the blue dots are on red lines that means Root means square error will low and better.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

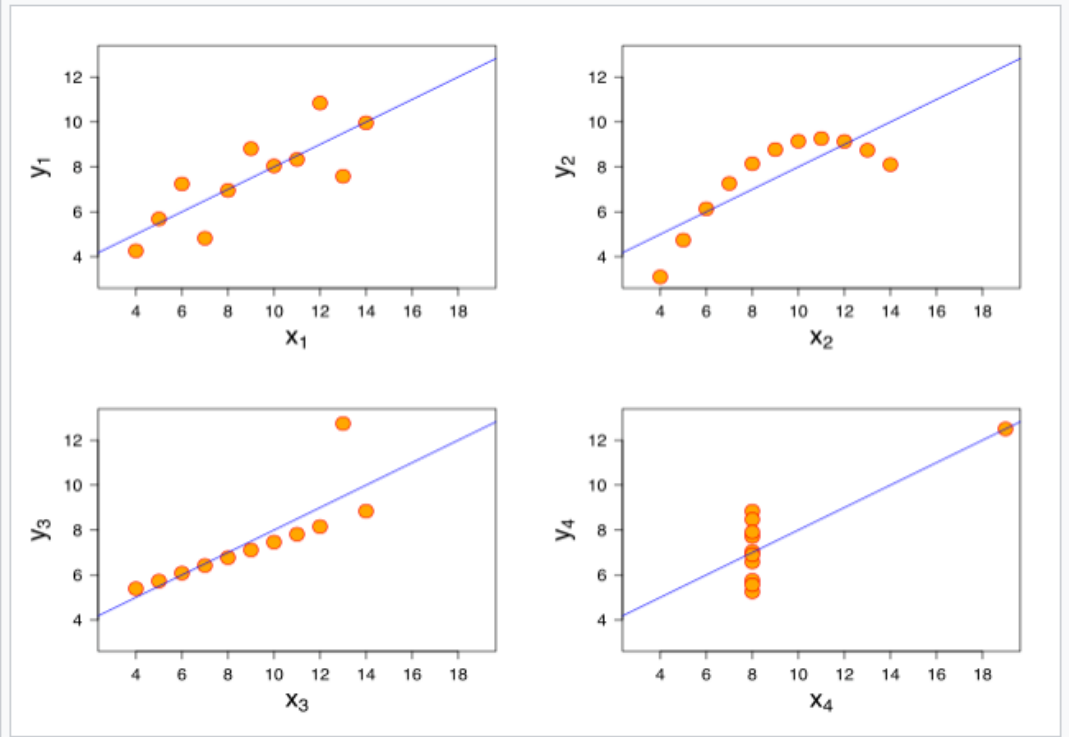
Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

2.Explain the Anscombe's quartet in detail.

Answer:

Statistics have long been used to describe data in general terms. For example, things like variance and standard deviation allow us to understand how much variation there was in some data without having to look at every data point individually. They give us a rough idea as to how consistent data is. However, knowing variance alone does not give you the full picture as to what the data truly is in its native form. Statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set. Francis Anscombe realized this in 1973 and created several data sets, all with several identical statistical properties, to illustrate it. These data sets, collectively known as "Anscombe's Quartet.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

The summary thus obtained was -

Summary

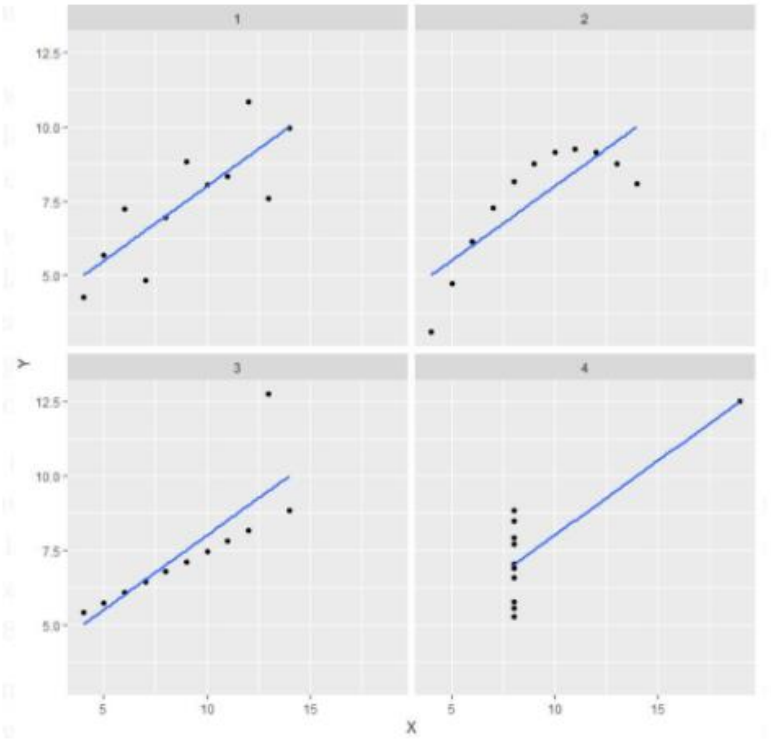
Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

On writing a code in python to plot the same we get this –

Python program to plot scatter plot

```
# Import the required libraries
from matplotlib import pyplot as plt
import pandas as pd

# Import the csv file
df = pd.read_csv("anscombe.csv")

# Convert pandas dataframe into pandas series
list1 = df['x1']
list2 = df['y1']

# Function to plot scatter
plt.scatter(list1, list2)

# Function to show the plot
plt.show()
```

Output:



Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

3.What is Pearson's R?

Answer: Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the yaxis (vertically).The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

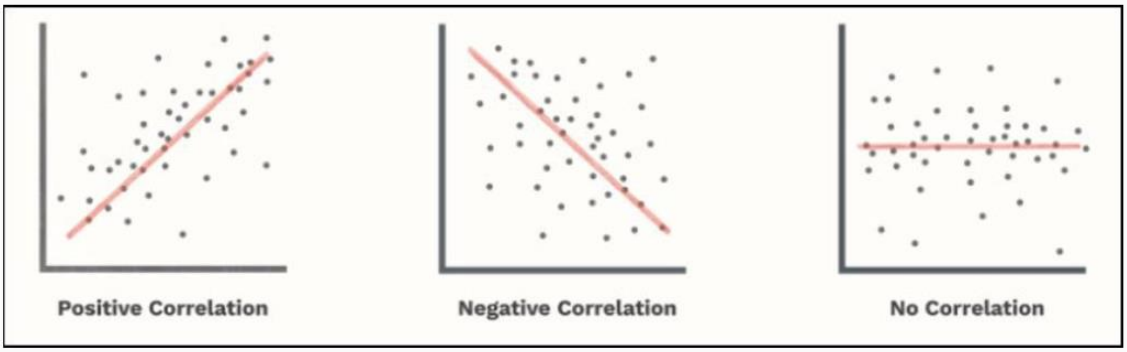
Coefficient of Correlation (r):

1. It measures the strength and the direction of a linear relationship between two variables (x and y) with possible values between -1 and 1.

2. **P*ositive Correlation:*** It indicates that two variables are in perfect harmony. They rise and fall together. +1 is perfect +ve correlation

3. ***Negative Correlation:*** It indicates that two variables are perfect opposites. One goes up and other goes down. -1 is perfect -ve correlation

4. ***No correlation***: If there is no linear correlation or a weak linear correlation, r is close to 0.



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

Answer: It is a step of data Pre- Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range.
. Sometimes, it also helps in speeding up the calculations in an algorithm. When data set contains features highly varying in magnitudes , units and range . If scaling is not done then the algorithm only takes magnitude into account and not units hence there is incorrect modelling . To solve this issue we have to do scaling to bring all the variables to the same level of magnitude .Scaling only affects coefficients and none of the parameters such as t-statistic, f-statistic , p-values and r-squared values .

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

Why do we perform scaling? We perform scaling to achieve gradient descent faster.
What do we do in scaling? We convert the data points of a column in a specific range to another range(0-1) based on the method(normalization/standardization). So we are not changing the data point but we are changing it's scale. This change in scale is reflected in coefficients because we use OLS method to find the optimal coefficients. So the predictability power is same but the scale and coefficient are changed.

The terms normalization and standardization are used interchangeably, but they usually refer to Normalization – means to scale a variable to have values between 0 and 1. While Standardization transforms data to have a mean of zero and a standard deviation of 1. This standardization is called as z-score and the data points can be standardized with the formula.

$$z_i = \frac{x_i - \bar{x}}{s}$$

A z-score standardizes variables.

We take a example of dataset where we have salary variable and age variable. Age can take range from 0 to 90 where salary can be from 25thousand to 2.5lakh.
We compare difference for 2 person then age difference will be in range of below 100 where salary difference will in range of thousands.
So if we don't want one variable to dominate other then we use either Normalisation or Standardization. Now both age and salary will be in same scale but when we use standardiztion or normalisation, we lose original values and it is transformed to some values. So loss of interpretation but extremely important when we want to draw inference from our data.Normalization rescales the values into a range of [0,1]. also called min-max scaled.
Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1.So it gives a normal graph.

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

Answer:

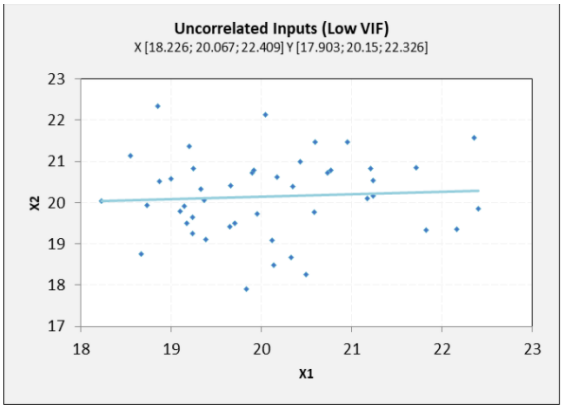
VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

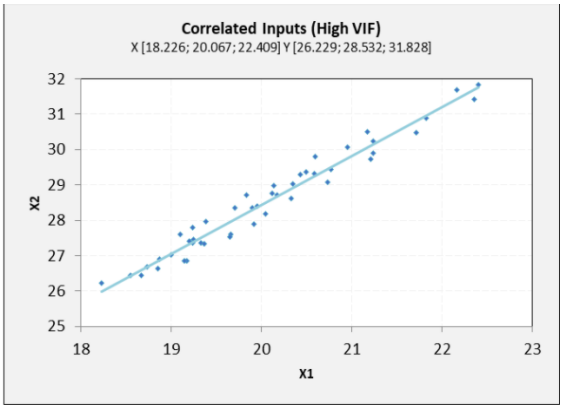
VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.

〖VIF〗_1=1/(1-R_1^2)

In order to determine VIF, we fit a regression model between the independent variables. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).

The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.





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

Answer:

Quantile-Quantile(Q-Q)plot , is a graphical tool to help us assess if a set of data possibly came from theoretical distribution such as Normal ,exponential or uniform distribution. It also helps to determine if two data sets come from population with a common distribution . This helps in a scenario of linear regression when we have a training and test data set received separately and then we can confirm using the Q-Q plot that both the data sets are from populations with the same distributions.

FEW ADVANTAGES:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios if 2 data sets -

i. Come from populations with a common distribution

ii. Have common location and scale

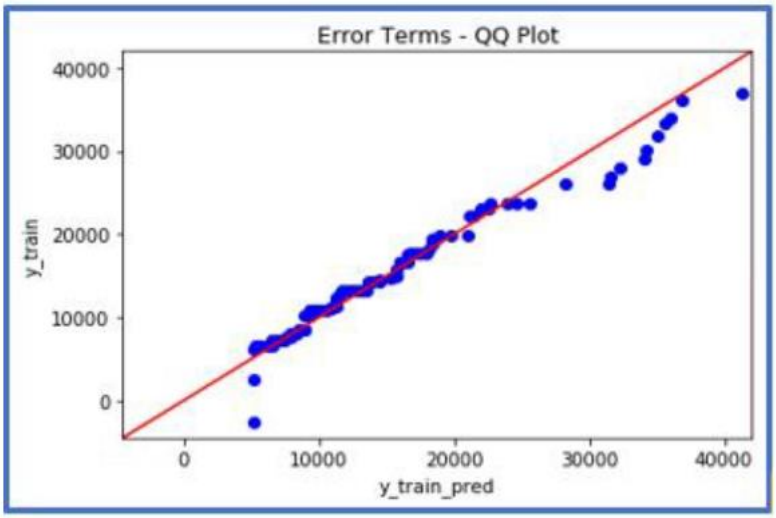
iii. Have similar distributional shapes

INTERPRETATION:

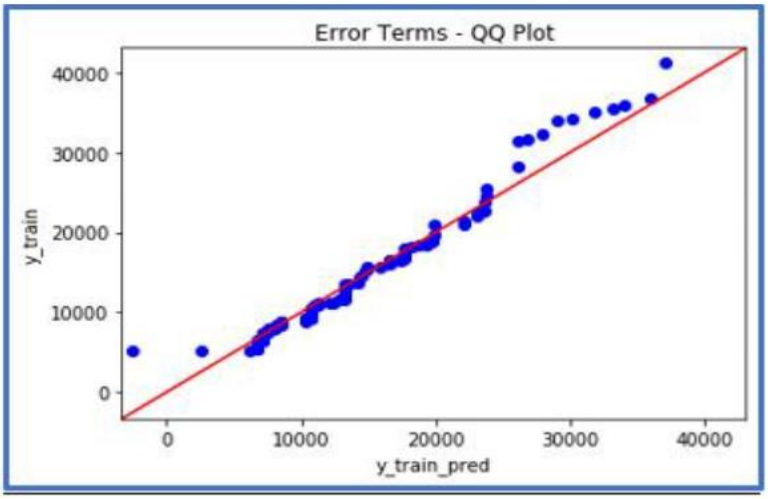
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an

angle of 45 degree from x -axis

As the name suggests, the Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words, you plot quantiles against quantiles. Whenever you interpret a Q-Q plot, you should concentrate on the 'y = x' line.

You also call it the 45-degree line in statistics. It entails that each of your distributions has the same quantiles. In case you witness a deviation from this line, one of the distributions could be skewed when compared to the other.