


MULTIPLE LINEAR REGRESSION ASSIGNMENT SUBMISSION

Name: Mounica Yelchuri

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable

- **Spring season** has less count of bookings compared to other seasons.
- Number of bookings **got increased in 2019**.
- Bikes are used **more on working days** compared to holiday
- No pattern seen among the spread around weekdays.
- No pattern seen for working day.
- **Weathersit light snow has very less cnt** compared to weathersit mist and weathersit clear.
- **cnt has increased with raise in temperature till 30 but there is reduce in cnt after 30**(indicating curvilinear relationship) .As there is no linear relationship, considered the variable as categorical variables and created 6 bins out of the temp. **Clear pattern is visible among temperature bins for cnt.**
- **Cnt has reduced when humidity** is greater 80 . As there is no pattern visible for humidity range 30-80 ,created bins for humidity as well.
- Atemp is highly correlated with temp . Hence dropped the attribute to avoid multi collinearity.
- Windspeed has no impact on cnt.

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

Consider a categorical feature having k distinct values.

If we create dummy variables without drop_first=True, then k new columns will be created.

If we create a model using these k attributes, We might end up having multi collinearity issue as one of the columns can be considered as linear dependent of others.

Example :

Weathersit has three categories : clear , mist ,light snow.

If we create two variables Weathersit_clear, Weathersit_mist then

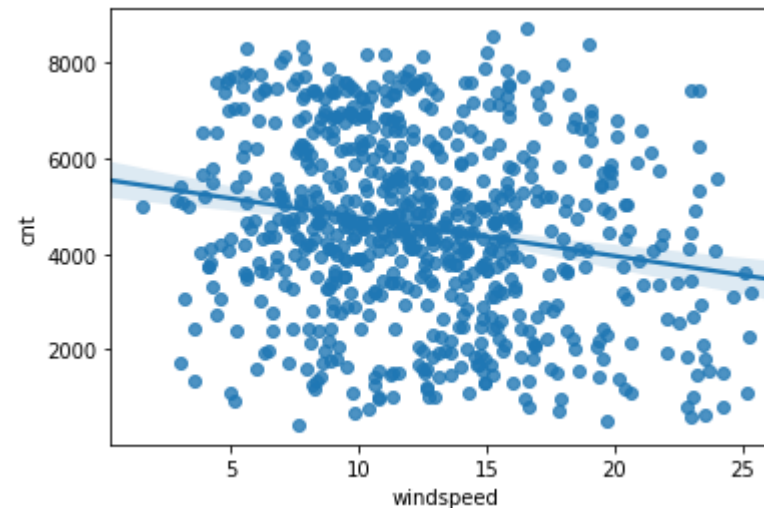
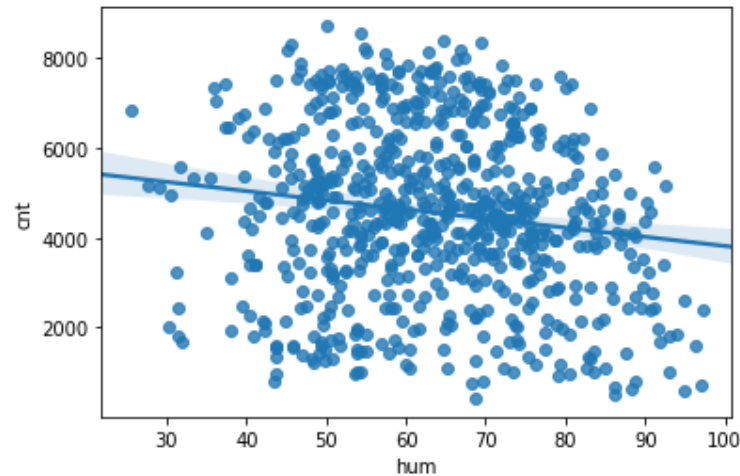
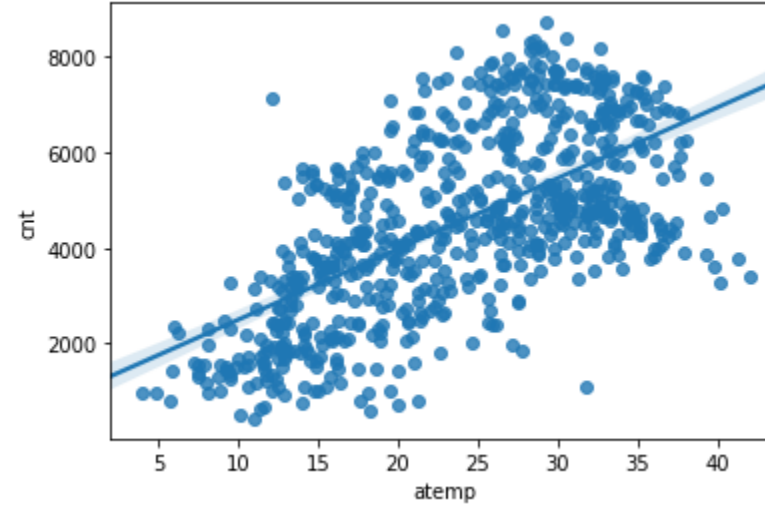
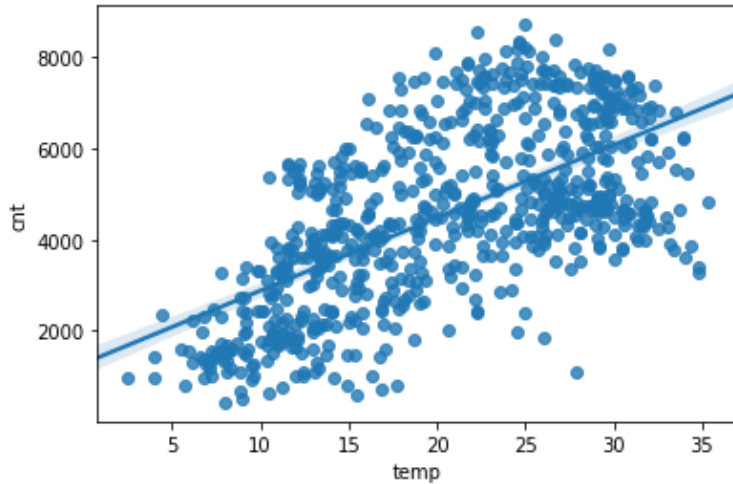
Clear can be represented by 10

Mist can be represented by 01

Light snow can be represented by 00.

We can interpret the third variable with help of two variables only.

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

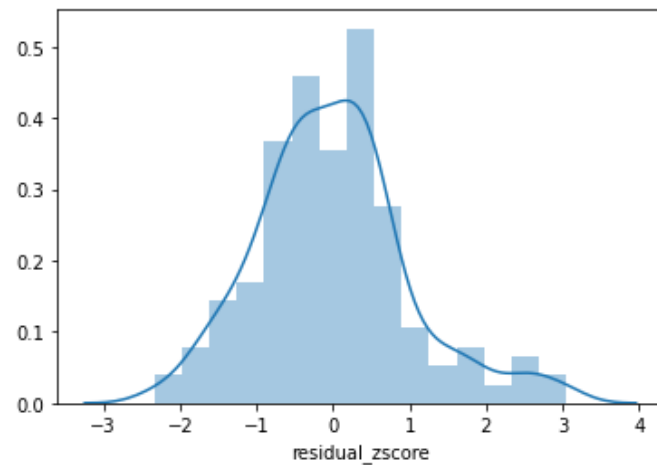


- Temp and atemp has correlation with cnt.
- As temp and atemp are highly correlated, we considering only temp.

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

Assumptions of linear regression model :

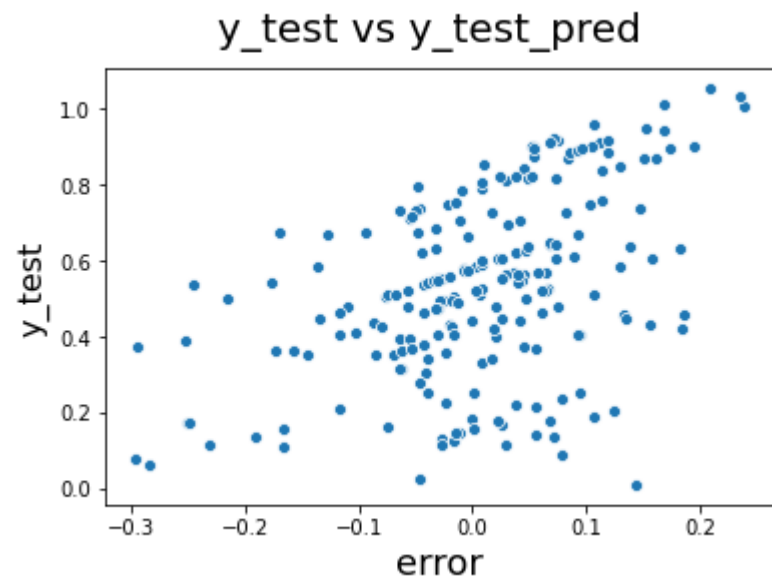
- Error terms(residuals) are normally distributed with mean zero.



Residuals are within -3 to 3 standard deviations . Hence they are normally distributed.
Also the mean of the distribution is around 0

Validating the assumptions of Linear Regression

- Error terms are independent of each other. Error terms have constant variance that is variance should not follow any pattern as the error terms change.



Error terms are randomly spread. Slightly high variance is seen only in extremes.

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

const	0.325929
yr2019	0.264099
holiday	-0.075604
season_spring	-0.152674
weathersit_Light Snow	-0.164421
weathersit_clear	0.076160
mnth_Oct	0.053643
temp_bin_2.42-7.91	-0.108214
temp_bin_7.91-13.39	-0.090328
temp_bin_18.88-24.36	0.131690
temp_bin_24.36-29.84	0.178877
temp_bin_29.84-35.33	0.118629
hum_bin_79.29-97.25	-0.052828

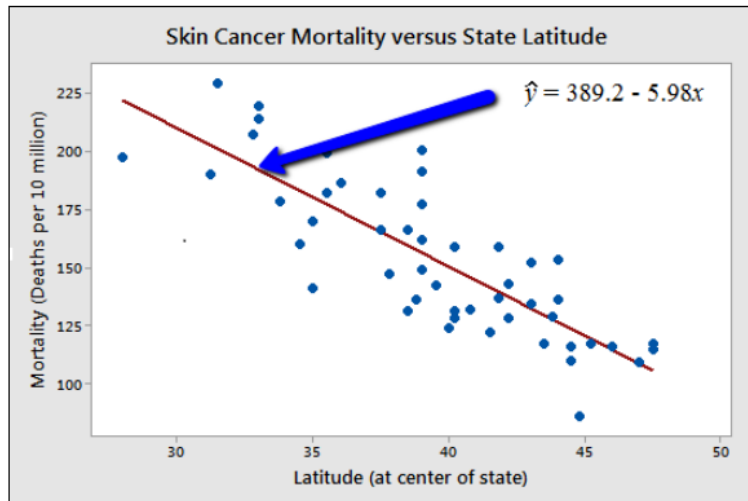
Year 2019 , temperature between 18 to 35 , clear weathersit ,not a holiday signify the demand of the shared bikes.

Linear regression Algorithm

Linear regression is a one of the supervised learning methods that allows us to summarize and study relationships between a dependent variable and independent variables (two or more continuous variables)

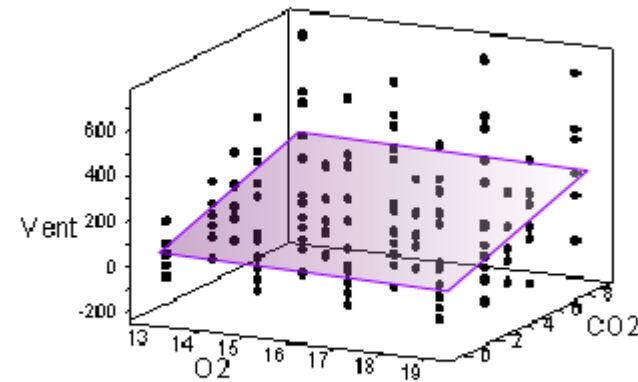
Linear regression consists of two algorithms :

Simple Linear regression



Simple Linear regression explains the relationship between a dependent variable and **one** independent variable using a straight line

Multiple linear regression



Multiple linear regression explains the relationship between a dependent variable and **more than one** independent variables using a plane/line.



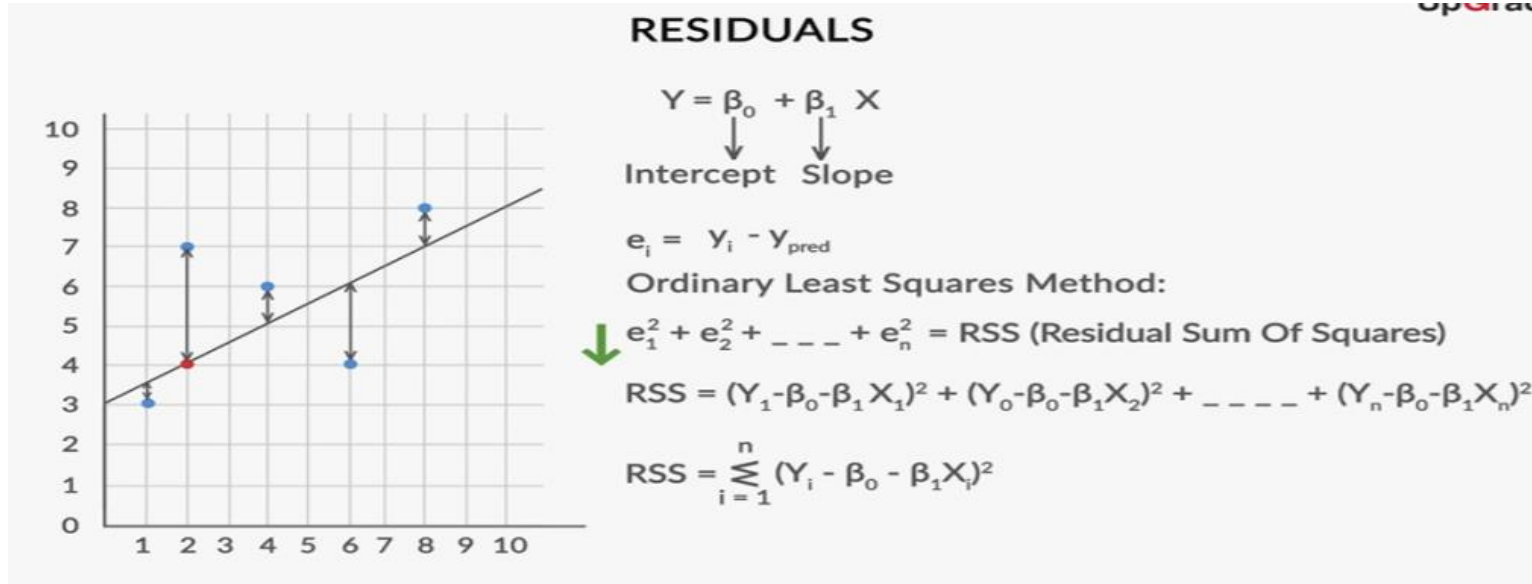
Simple Linear regression Algorithm

The standard equation of the regression line is given by the following expression

$$\hat{y}_i = b_0 + b_1 x_i$$

We can draw as many lines joining these points, but we need find the **best line which summarizes the trend between two variables.**

The best-fit line can be found by minimizing the sum of squares of residual(difference of y value predicted using the linear equation and actual line).



Measure of linear regression model

Efficiency of model can be explained using following metrics:

Coefficient of determination

F –test

Coefficient of determination (r- square)

RSS (Residual Sum of Squares): sum of squares of the difference between the expected and the actual output.

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable.

$$r^2 = 1 - \left(\frac{RSS}{TSS} \right)$$

Higher the r^2 , the variables are more statistically significant.

But we cannot completely rely on r-square for predicting the model.

Hypothesis testing using F-test

Null hypothesis : slope parameter beta1 is zero

Alternate hypothesis : beta1 is not equal to zero.

We calculate f –statistic from the below equation

Source of Variation	DF	SS	MS	F
Regression	1	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$
Residual error	$n-2$	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	$SSTO = \sum_{i=1}^n (y_i - \bar{y})^2$		

- P -value is determined by comparing f -value to an F distribution with 1 numerator degree of freedom and $n-2$ denominator degrees of freedom.
- If the p value is less than significant level then we reject the null hypothesis that beta1 is zero.
- If the p value is greater than significant level then we accept the null hypothesis that there is no significance of the independent variable.

Linear regression limitations

- There is a linear relationship between the independent and dependent variables.
- Error terms(residuals) are normally distributed with mean zero.
- Error terms are independent of each other.
- Error terms have constant variance that is variance should not follow any pattern as the error terms change.

Multiple linear regression

- Multiple linear regression is an extension of simple regression where we have multiple independent variables instead of one.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times x_1 + \hat{\beta}_2 \times x_2 + \hat{\beta}_3 \times x_n \dots \hat{\beta}_n \times x_n$$

- Along with assumptions of simple linear regression, multiple regression defines that these independent variables should be linearly independent.

Measure the efficiency of multiple linear regression model

- We use the same hypothesis testing used earlier to eliminate the not so significant independent variables.
- To identify the linear dependency between the variables, we use variance inflation factor(VIF) we need to make sure that VIF of the variables in the final model is less than 5.
- Adjusted R-squared is a better metric than R-squared** to assess how good the model fits the data. R-squared always increases if additional variables are added into the model even if is insignificant. Adjusted R-squared, on the other hand, penalises R-squared for unnecessary addition of variables. So, if the variable added does not increase the accuracy adequately, adjusted R-squared decreases although R-squared might increase.

Anscombe's quartet

Anscombe's quartet was constructed by statistician Francis Anscombe which contains four datasets that have nearly identical statistical properties having very different distributions which can be understood better when graphed.

Interpreting the data from statistical properties

Anscombe's quartet

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

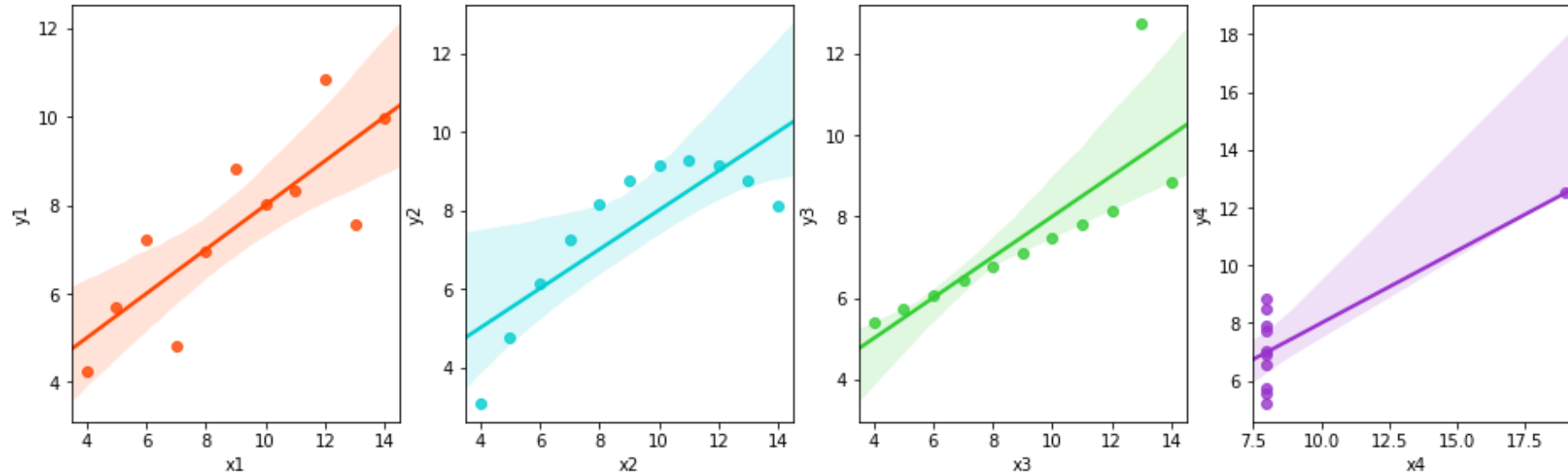
Summary

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

Statistical properties(mean , standard deviation, correlation between attributions) are almost same for these datasets.

Anscombe's quartet

Interpreting data using visualization



- In the first plot there seems to be a linear relationship between x_1 and y_1 .
- In the second plot a non-linear relationship between x and y can be easily misinterpreted as linear relationship.
- In the third plot there is a good linear relationship for all the data points except one which seems to be an outlier (indicated far away from that line) which will mislead to building an inaccurate linear regression model.
- In the fourth plot there is one point which is far away from the other which produces a high correlation coefficient, even though the other data points do not indicate any relationship between the variables

We understand the importance of visualizing data graphically before starting to analyze according to a particular type of relationship and also understand that of basic statistic properties are inadequate for describing realistic datasets

- Pearson r is directly related to the coefficient of determination (r-square).

$$r = \pm \sqrt{r^2} \quad r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad r = \frac{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \times b_1$$

- The sign of r depends on the sign of the estimated slope coefficient b_1 .
- If b_1 is negative, then r takes a negative sign.
- If b_1 is positive, then r takes a positive sign

Interpreting r value :

- If $r = -1$, then there is a perfect negative linear relationship between x and y .
- If $r = 1$, then there is a perfect positive linear relationship between x and y .
- If $r = 0$, then there is no linear relationship between x and y

Caution of r / r^2

- The coefficient of determination r^2 and the correlation coefficient r **quantify the strength of a linear relationship**. It is possible that $r^2 = 0\%$ and $r = 0$, means there is no linear relation between x and y , and a perfect curved (or "curvilinear" relationship) exists.
- A large r^2 value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data.
- The coefficient of determination r^2 and the correlation coefficient r can both be **greatly affected by just one data point (or a few data points)**.

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

- Scaling is performed to normalize the range of the independent variable
- Linear regression uses gradient descent as optimization technique which would perform better if data is scaled across all independent variables.

Normalized scaling

Values are shifted and rescaled so that they end up ranging between 0 and 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

X_{min} is the minimum of the feature x

X_{max} is the maximum value of feature X

If we apply the above formula, new value for X_{min} would be zero and scaled value for X_{max} would be 1.

Standardization scaling:

Values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

- Normalization is good to use when the distribution of data does not follow a Gaussian distribution. 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
- One **disadvantage of normalization** over standardization is that it **loses** some information in the data, especially about **outliers**

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

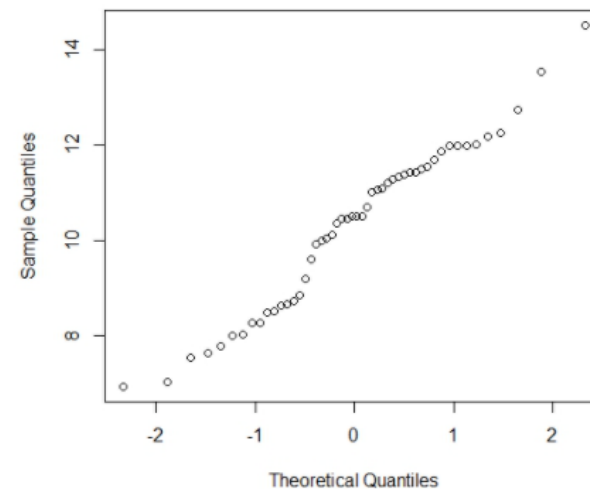
Usually we observe the VIF as infinite only when there is multi-collinearity in the model.

$$VIF = \frac{1}{1 - R^2(x_1)}$$

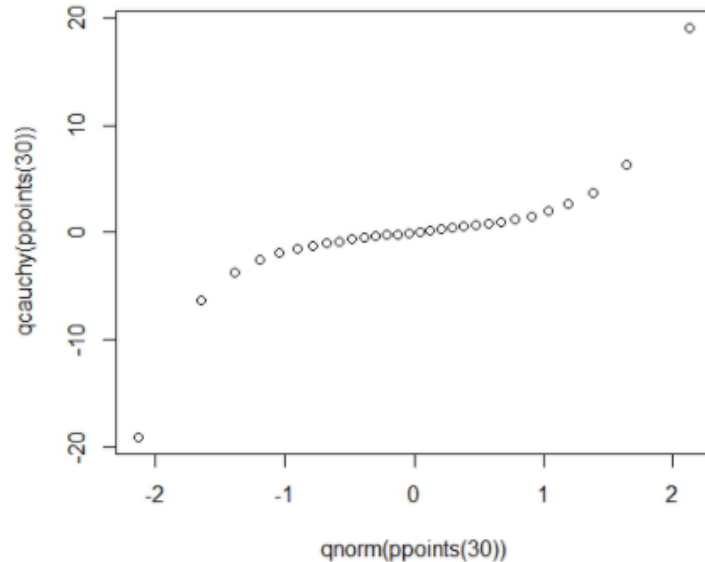
VIF is infinite when denominator is zero (i.e. r-square is 1).

r-square being 1 indicates that the variable is highly correlated (linearly dependent) with another variable. We need to drop one of the variables to improve the model.

- What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
- Quantile-quantile plots helps us in understanding if the distribution is normal or exponential.
- A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



Q-Q plot



The points fall along a line in the middle of the graph, but curve off in the extremities. Q-Q plots that exhibit this behavior usually mean data has more extreme values than expected if they truly came from a Normal distribution.

We can use the Q-Q plot to validate the assumption of linear regression that error residuals are normally distributed or not.

If q-q plot depicts a line then we can consider that error are normally distributed.