12-Dec-2022

LINEAR REGRESSION SOLUTION ASSIGNMENT

ASSIGNMENT BASED SUBJECTIVE QUESTIONS

1. Seasons is linked with the Demand for bikes as we can see in the box plots as well. People tend to use more in the warmer seasons than the colder ones

   Weekdays have little significance on the predictability of data

   The monthly data follows a certain trend and it shows that it peaks during the summer months

   Weather situation has a clear bearing on the result. We can see that People do not tend to use Bikes in the snow

2. drop_first=True removes the column after hot encoding to adjust for high correlation values . It knows that the last column value can we represented as NOT of other variables , hence it is necessary to do so as it reduces the multicollinearity part of the analysis. If we include it we might get VIF = infinity when we have built the model
3. temp has the highest correlation with cnt
4. After we build the model , we try to check its r2 score and plot the scatter plot with the test set . if we see a linear line as an emergent line we can be rest assured that the model is correct
5. The top 3 variables are :
   i. Yr
   ii. Holiday
   iii. Working_day

2. GENERAL SUBJECTIVE QUESTIONS

1. Linear Regression Model is a statiscal analysis that is used for predictive analysis based on the formula

   $Y = m_1x_1 + m_2x_2 + ……. + k$

   Y is the dependent variable and x_1 x_2 …x_n are independent variables

   This involves that we try to estimate the test set and then we try to get meaningful association

2. Anscombe Quartet

   i. This signifies that we can represent 4 different lines, considering all other parameters to be same, yet have very different distributions and appear very different when graphed.

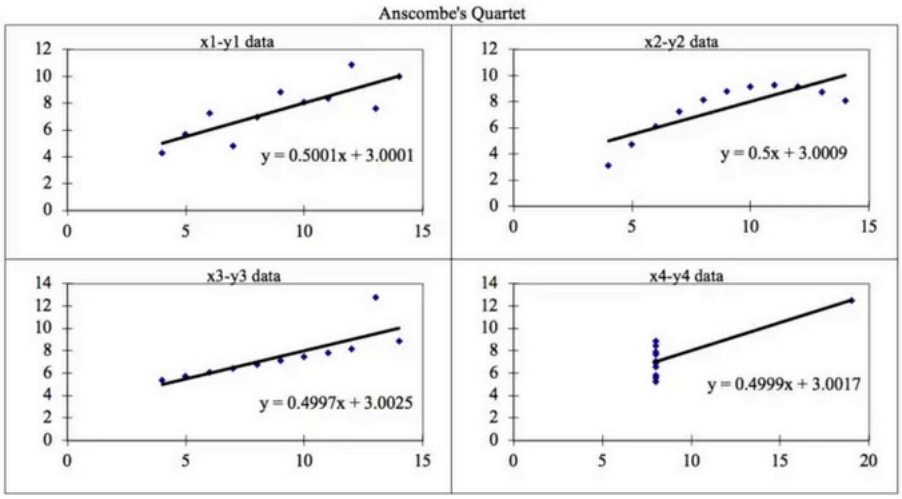
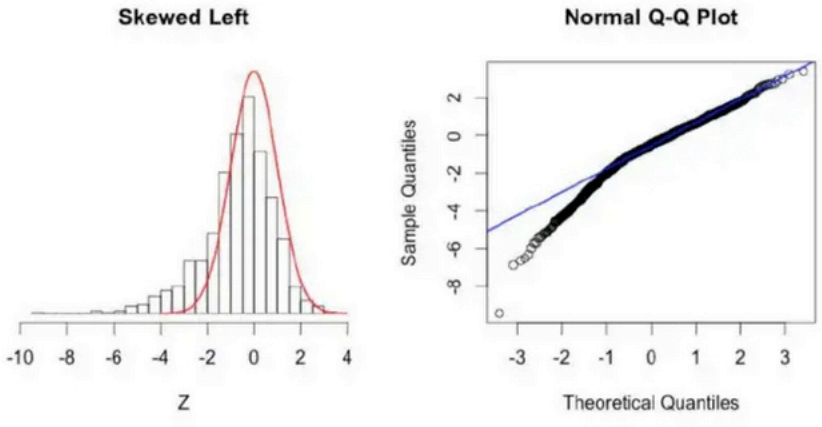


Image by Author

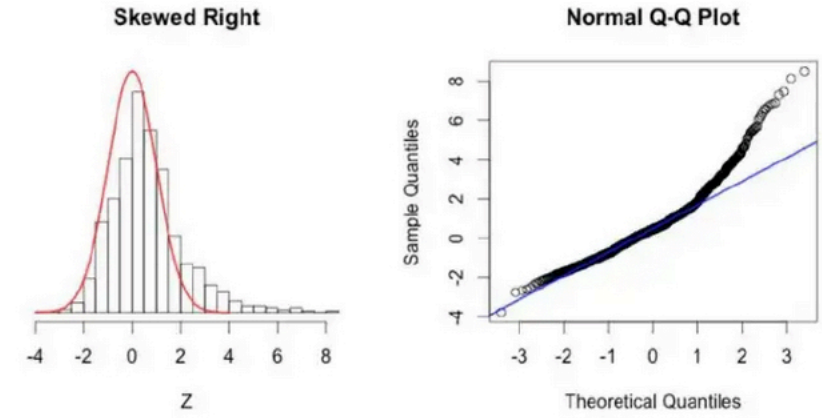
3. Pearson R – it is a measure of linear correlation between 2 variables. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationships or correlations.

4. Scaling is trimming the data to levels where it can be visualized and inferred quickly.
   Scaling of data may be useful and/or necessary under certain circumstances (e.g., when variables span different ranges). There are several different versions of scaling, the most important of which are listed below. Scaling procedures may be applied to the full data matrix, or to parts of the matrix only (e.g.. column wise).
5. VIF is infinite if two columns have very high collinearity. We observed that when we ran the model post hot encoding and not using drop -first functionality in hot encoding. The VIF values came to be infinite

   As is evident with the formula $1/(1-R^2)$ Higher the correlation much larger is the value of VIF
6. The Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. We can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.



Left Skewed Q-Q plot for Normal DIstribution



If the bottom end of the Q-Q plot deviates from the straight 45 degree line but the upper end is not, then we can clearly say that the distribution has a longer tail to its left or simply it is **left-skewed** (or ***negatively skewed***) but when we see the upper end of the Q-Q plot to deviate from the straight line and the lower and follows a straight line then the curve has a longer till to its right and it is **right-skewed**