BIKE SHARING & LINEAR REGRESSION ASSIGNMENT

Submitted By: SOUMAYADEEP MANNA

Assignment-based Subjective Questions

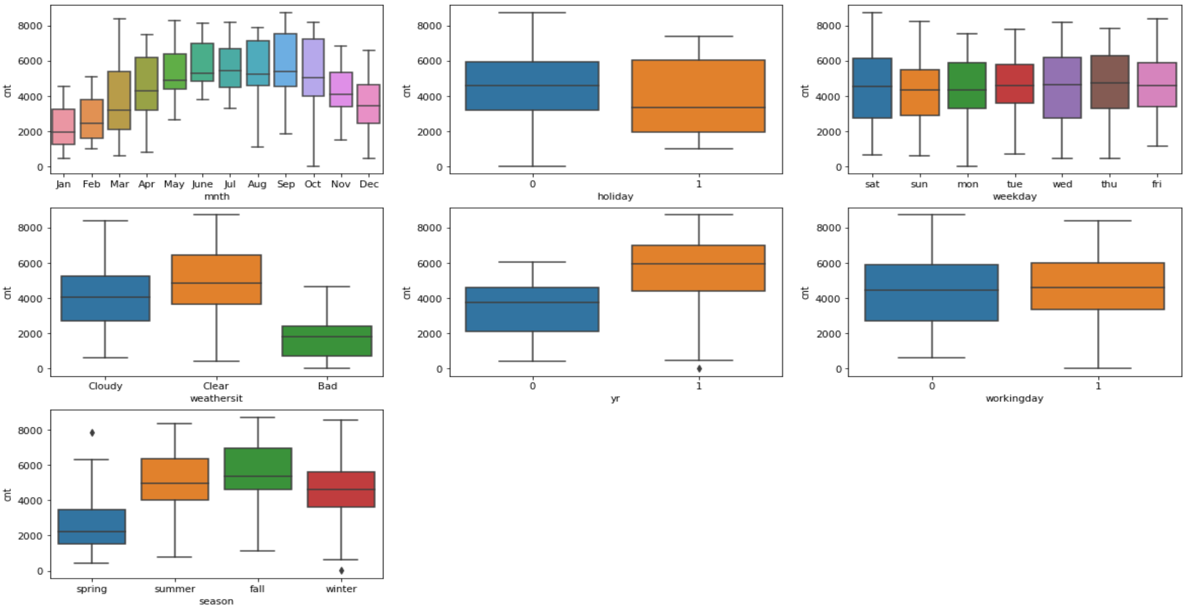
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans:

Following insights and points can be inferred from the analysis of the categorial variables

1. Month wise during September bike sharing is more, compare to year ending and starting it is the lowest.
2. On holidays there is less demand of bike sharing.
3. Days of the week not clearly distinguish any clear data.
4. Bike sharing demand increases in Clear, Few clouds, partly cloudy, partly cloudy weather conditions.
5. Season wise Fall has the highest bike sharing demand.
6. Month-wise bike sharing demand is gradually increasing till September and it starts to decrease gradually.
7. Bike demand for the next year is high.

Plot attached:



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

Ans:

drop_first=True is important to use during dummy variable creation, because it helps to reduce extra number of columns that will be get created. Which can cause high correlation between dummy variables, so if we do drop_first=True it will not create extra column and hence correlation will be reduced. ***So, if there are n categories of a feature it can be explained by n-1 dummy variables.***

Example:

Let "Grade" be a feature with 3 categories A/B/C, if we create dummies for grade:

Var1 -→ grade_A

Var2 -→grade_B

Now, machine already known that third or last one will be grade_C, so if we don't do drop_first=True, it will create Var3 →grade_C in spite of already knowing the 3rd dummy value.

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

Ans:

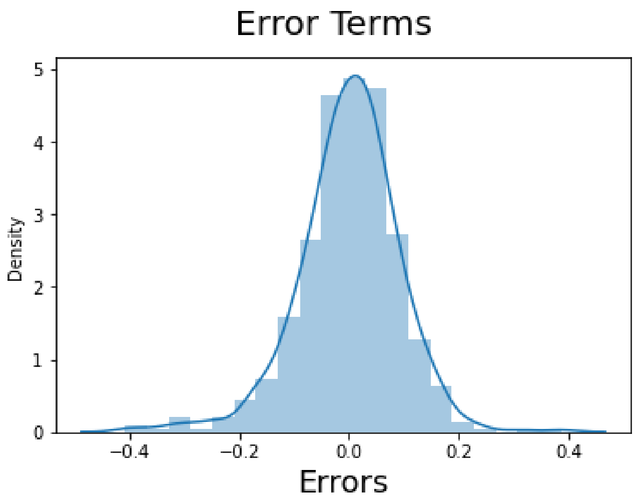
Feeling temperature (atemp) has the highest correlation with the target variable with the value of **0.630685**.

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

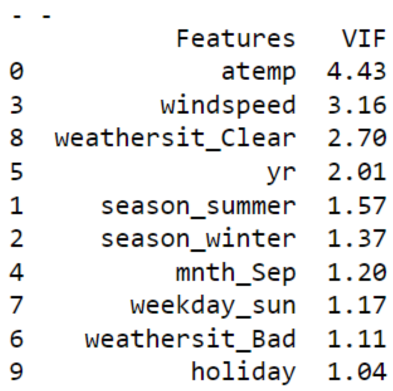
Ans:

Assumptions of linear Regression after building the model on training set is validated in following ways:

- ***Error terms are normally distributed with mean zero.*** We have plotted error terms in a distribution plot, plot result attached below which clearly explains that the error terms are normally distributed with mean near to the zero.

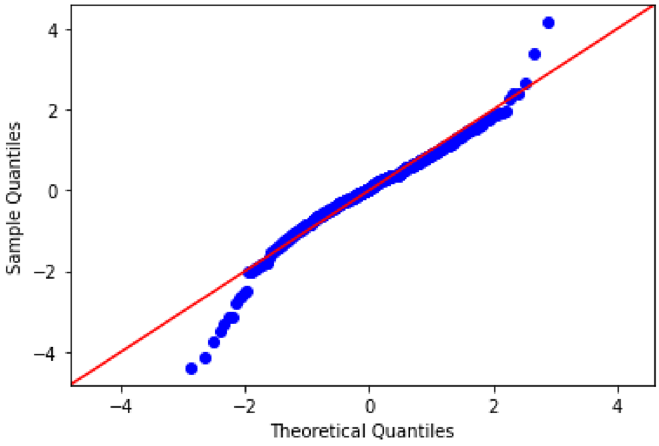


- **There should be no high multi collinearity among the features of the models.** To verify the condition, we have calculated Variance Inflation Factors (VIF) for the final model features. VIF values are as follows, *where all the VIF values are less than standard threshold value of 5.0.*



```
 - -
            Features   VIF
0              atemp  4.43
3          windspeed  3.16
8  weathersit_Clear  2.70
5                 yr  2.01
1     season_summer  1.57
2     season_winter  1.37
4           mnth_Sep  1.20
7       weekday_sun  1.17
6    weathersit_Bad  1.11
9            holiday  1.04
```

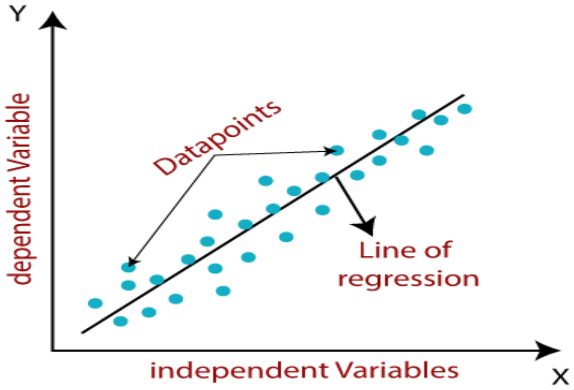
- Normal distribution on the residuals by plotting Q-Q plot.
  Plot shows the normal distribution of the residual's values with slight fat tails.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Ans:
   The top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows:
   - *Actual Temperature (atemp) with a highest positive coefficient value of 0.5751.*
   - *Year 2019 (yr: 1) with second highest positive coefficient value of 0.2343.*
   - *Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (weathersit_Bad) with highest negative coefficient value of -0.2016.*

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Ans:
   Linear Regression is a ML algorithm which based from the statistical model Linear Regression. Linear Regression is used to find the Linear Relationship between 1 or more independent variables and a dependent variable.



In the above image, if we consider the variable X as an independent and the dependent variable y, we could see that there is a linear trend between data points those are plotted based on X and y. Once this linear trend is spotted, try fitting a line through the data points in a way that the error between the data points and the predicted values on the line, for a given value of x in minimum.

Mathematically the relationship can be represented with the help of following equation –

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept.
If X = 0, Y would be equal to c

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases

- Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases.

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions –

The following are some assumptions about dataset that is made by Linear Regression model –

- Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Linear regression model assumes that the relationship between response and feature variables must be linear.
- Error terms should be normally distributed
- There should be no visible pattern in residual values.

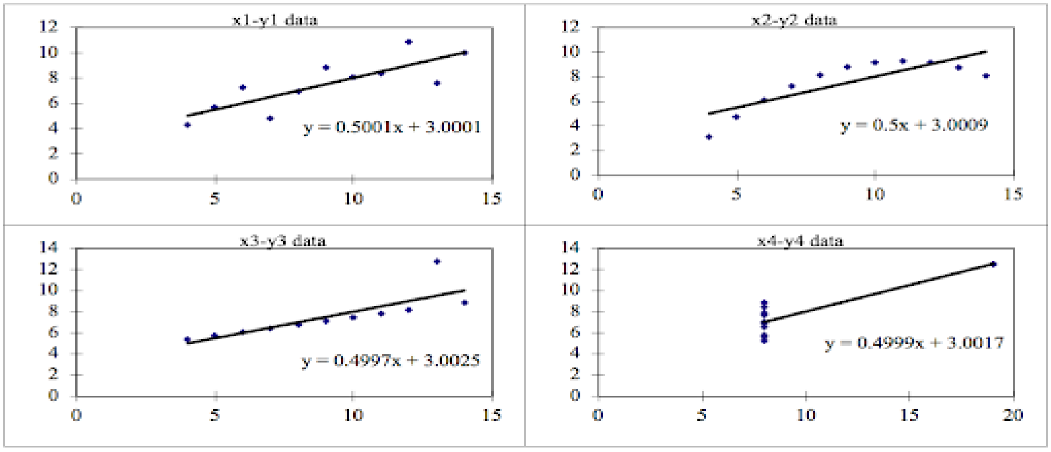
2. **Explain the Anscombe's quartet in detail.**

Ans:

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to signify the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Summary Statistics consists of calculating values like the mean, median, mode etc. While it does give us a good idea about the data, it doesn't really help us look at the shape or the distribution of data.

When these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



So, despite of having similar four datasets the following visualization of data shows that,

- Model 1 has linear relationship
- Model 2 is not linear
- Model 3 is linear but it has outliers present in the dataset
- Model 4 also shows outliers value present in the dataset

3. **What is Pearson's R?**

<u>Ans:</u>

Pearson's R is a statistical measure that is used to determine the measure of the strength of association between two numerical and Linear Variables. Pearson's correlation coefficient is usually calculated by plotting the values of the independent variable of a sample on the x-axis and the corresponding
values of the dependent variable of the sample on the y-axis.

Note that, the variables strictly don't have to be dependent on one and another.

After plotting the values on the graph, the covariance values are calculated by
the formula:

$$Cov(m, n) = \Sigma(m_i - \overline{m})(n_i - \overline{n})/N - 1$$

Where,

- m_i – m value of an individual data point
- n_i - n value of an individual data point
- $\overline{m}$- Mean of m
- $\overline{n}$ - Mean of n

Once the covariance value is calculated, the Correlation coefficient is calculated by dividing the value of covariance with the standard deviation of m and n.

$$R = Cov(m, n)/\sigma m \sigma n$$

- σm - Standard Deviation of m
- σn- Standard Deviation of n

The Correlation coefficient tells us if there is a strong positive relationship, strong negative relationship, weak positive relationship, weak negative relationship or no relationship between the 2 variables. The Correlation Coefficient value would only range in between -1 to 1.

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

<u>Ans:</u>

Scaling is a process of converting a feature variable or multiple feature variables into a standard scale or a common scale.

The reason this is done is because as multiple feature variables come with their own scales of data and their distribution curves, the ML model which is using the features to predict the response variable would implicitly take the feature variable with a higher scale to be more important than a feature variable with a lower scale. Another reason why scaling is recommended is it makes the interpretation of the model really hard since the coefficients for different variables would have extremely high and low values, due to various scales.

Scaling prevents some of the above listed issues and makes the data much more interpret-friendly and it also expedites the process of finding the coefficients since gradient descent algorithm responds better to data in a constricted scale than in a large/ very small variable scale combination.

<u>The two common types of scalers used are:</u>

1. MinMax Scaler
2. Standard Scaler

<u>MinMax Scaler or Normalized Scaler</u> scales the data to fit it within the range of 0 to 1, no matter however big the range of the data is.

<u>MinMax Scaler</u> achieves this by learning the values of the minimum and maximum values initially, subtracting the value of each data point from the minimum value and dividing it by the range of the data.

MinMax Scaler = (x – xmin)/ (xmax – xmin)

Standard Scaler converts the values of the variable to a format where it sets the average value of the data close to zero and the standard deviation of the data to 1. This method of scaling is quite beneficial if the variable which is about to be scaled follows a normal distribution.

Standard Scaler = (x – mean(x)) / standard deviation of x

Both the scalers are affected by outliers since the parameters used to scale the values are affected by outliers.

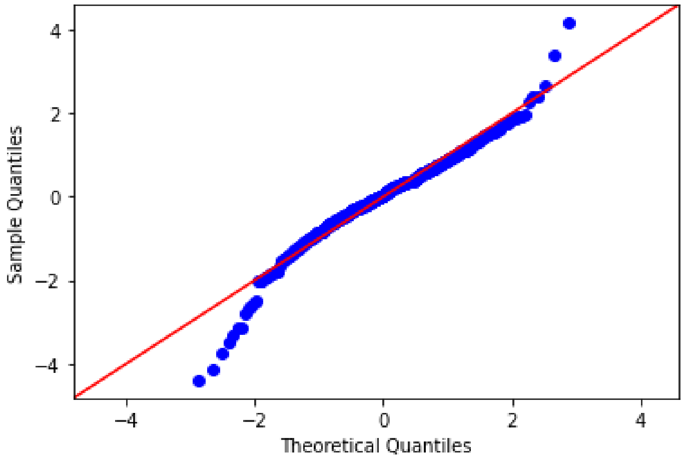
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
Ans:
If there is perfect or high correlation, then VIF value equals to infinity. This shows a perfect correlation between two independent variables.

In the case of perfect correlation, we get R2 =1, As per formula

VIF = 1 / 1-R2

Where R2 is the R-squared value of the fit between a feature variable as the dependent variable and the other feature variables as independent variables.
To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
Ans:
Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
The purpose of this plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



In this above plot a linear regression model residual values are plotted on a theoretical quantile to check distribution and linearity of the data.
If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.
In Linear Regression Q-Q can be used to validate the linearity that is normal distribution of error terms that is the assumptions make on a model based on test data.