Linear Regression Assignment

Name: Saurabh Dugdhe

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? (3 marks)**

Ans: From the bike sharing database, we studied following categorical variables and their effect on the `cnt` variable

   i. `yr` : We can see that with every passing year, we have more rentals. The rentals increased in the year 2019 over 2018
   ii. `season`: There are higher chances of renting a bike on in Summer, followed by fall and then winter. Spring has the lowest bike rentals.
   iii. `temp`: More bikes are rented when temperature is higher
   iv. `weathersit`: There are high rentals when the weather is clear, followed by misty and cloudy. But the rentals are very low in case of light rains or light snow. There is no data for heavy rains and snow.
   v. `woorkingday`: The rentals are marginally higher on working days than that on weekends.
   vi. `mnth`: Bike rentals are the highest during the months from May to October. The rentals generally increase from the start of the year, but then sharply fall after October.

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

Ans: If we want to create dummy variables for a categorical column, we will get one column for each unique value available for that variable. These columns contain 0 or 1 based on if the category is true. Out of these columns, only one can have a value `1`, while others will be `0`. Hence, these are correlated.

If we were to drop say one column out of them, we are still able to represent the dropped category i.e., when all other columns are `0`. This also helps in reducing the multicollinearity between the columns.

Pandas provides an attribute drop_first=True which gives us (n-1) columns for n features. This helps in reducing multicollinearity and also simplifies the data.

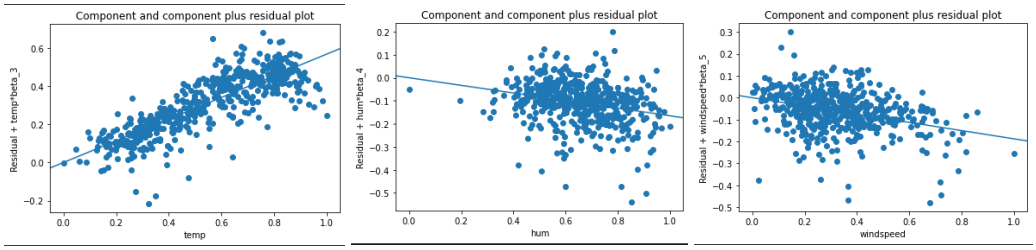
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Ans: After looking at the pair plot of the numerical variables, we can see that `temp` and `atemp` are highly correlated with the target variable `cnt`.

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

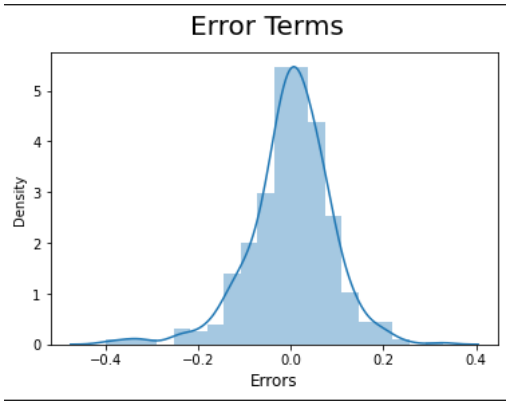
Ans: After building the model on training set, I performed following to validate the model:

   i. **Linear Relationship:** The X and Y must have linear relationship. The expected value of dependent variable is a straight-line function of each independent variable. Following figures show that the independent variables have a linear relationship with the dependent variable.



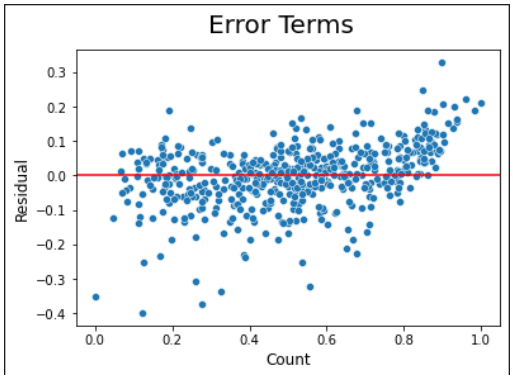
ii. **Normality in distribution of Errors:** The error terms must be normally distributed.

I plotted the distribution of residuals using the actual values and the predicted values of the test data. The distribution was normally centered around 0.



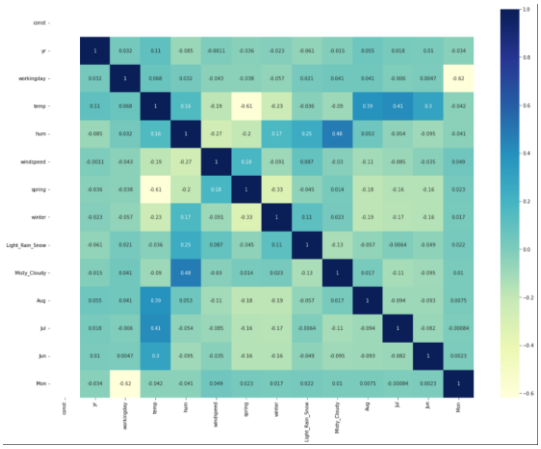
iii. **Homoscedasticity:** Error terms must have a constant variance.

The plot of residuals against the count variable shows not perceptible pattern to the error terms and hence there is no constant variance. Therefore, Homoscedasticity is preserved.



iv. **Multicollinearity:** There must be no or very little correlation between the variables.

As you can see there is no or very little correlation between variables.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Ans: The top 3 features contributing significantly towards explaining the demand of shared bikes are:

i. temp
ii. Light_Rain_Snow
iii. yr

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Ans:

Linear Regression is a type of supervised machine learning algorithm.

It is a method of modelling a target variable based on independent predictor variables.

It is a technique in which the dependent variable is of continuous nature. The relationship between the dependent and the independent variables is of linear nature.

Equation of a best fitted line is used to represent a linear regression model.

There are two types of Linear regressions:

1. **Simple Linear regression**:

Linear regression in which the number of independent variables is one and there is a linear relationship between it and the dependent variable. This c

Equation: $y = \beta_0 + \beta_1 x_1$

Here y is the dependent variable,

x is the independent variable,

β_0 is the intercept, and

β_1 is the slope

2. **Multiple Linear Regression**:

Linear regression in which there are multiple number of independent variables and there is a linear relationship between it and the dependent variable.

Equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_n + ... + \beta_n x_n$

Here y is the dependent variable,

$x_1, x_2, ..., x_n$ are the independent variables,

β_0 is the intercept, and

$\beta_1, \beta_2, ..., \beta_n$ are the slopes

Strength of Linear Regression:

The strength of a Linear regression model is explained by R^2,

$R^2 = 1 - (RSS/TSS)$

where,

RSS - Residual Sum of Squares

TSS - Total Sum of Squares

The R^2 value always falls between 0 and 1, and 1 being the perfect fit.

Assumptions of Linear Regression Model.

There are four principal assumptions to justify the use of linear regression model for prediction and inference.

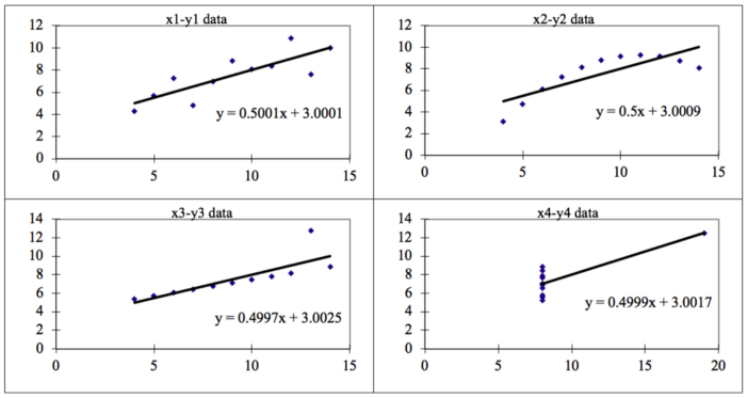
i. Linearity relationship of the dependent variable with the independent variables: The expected value of the dependent variable is a straight line for each independent variable. The slope does not depend on other variables.
ii. Statistical independence of errors (Absence of Multicollinearity): There must be no correlation between the variables.
iii. Homoscedasticity: There must be constant variance of errors verses time, predictions or any independent variable.
iv. Normality of the distribution of errors: Error terms must be normally distributed

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:

Anscombe's Quartet is a group of four datasets which are nearly identical in summary statistics. But in reality, they have very different distribution and distinct plots when plotted as a scatter plot. This fools the regression algorithm.

Let's take an example:
There are 4 datasets, whose summary statistics are the same, but when plotted they look different as shown in the following figure:



Here,

i. The first plot shows a linear relationship between x and y. this can fit the regression model.
ii. The second plot is non linear and has a curvature, hence it could not fit the linear regression model.
iii. The third plot has outliers, and hence cannot be handled well by the linear regression model
iv. The fourth plot has outliers, and hence cannot be handled well by the linear regression model

3. What is Pearson's R? (3 marks)

Ans:

Pearson's r is used to measure the strength of linear relationship between two variables. It is the ratio between covariance of two variables and the product of standard deviations. It can take values between [-1, 1].
Formula: $\mathbf{P_{x,y} = Cov(x,y)/(\sigma_x * \sigma_y)}$

i. Positive Value (>0): If the value of Pearson's r is greater than 0, then the relationship between variables is positive, hence as x increases, y increases.

ii. Negative Value (<0): If the value of Pearson's r is less than 0, then the relationship between variables is negative, hence as x increases, y decreases.
iii. Zero (0): If the value is 0, then there is no correlation between the variables x and y.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

Machine learning algorithm just sees raw numbers. It assumes that higher ranging numbers are always higher and superior to lower ranging numbers.

Scaling is the method that is used to normalize the range of all the independent variables. It brings all the variables to a similar scale, so that it is easy for the machine learning algorithm to compare them.

There are two types of scaling:

i. **Normalization**: In normalization, the values of an independent variable are transformed into a specific range of [0, 1]. This type of scaling is also known as Min-Max scaling.
Formula: $x_{new} = (x - x_{min}) / (x_{max} – x_{min})$
Where,
x_{new} is Scaled value
x is Actual value
x_{min} is minimum value of x
x_{max} is the maximum value of x

ii. **Standardization**: In standardization, the values of independent variables are transformed such that the resulting distribution has a mean of 0 and a standard deviation of 1. It is also known as z-score normalization.
Formula: $x_{new} = (x – \mu) / \sigma$
Where,
x_{new} is the Scaled value
x is the actual value
μ is the mean of the feature x
σ is the standard deviation of the feature x

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

VIF is the variance inflation factor. It is used to measure the multicollinearity in a set of multiple regression variables. When the VIF value is infinite, it means that there is perfect correlation between the two variables.

The formula for VIF is: $VIF = 1/(1-R^2)$

In case of perfect correlation, the R^2 is 1. Hence by putting it in above formula, we get the value of VIF as infinity. This means that there is a perfect correlation between the two variables. One variable can be expressed perfectly by a linear combination of the other variable.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Ans:

Q-Q plots or the Quantile-Quantile plot is a scatter plot that is plotting data quantiles and normal/standard quantiles against each other. They are used to determine if two samples of data are actually a subset of the same population.

Use:

i. Used to find if the sample data are from same population
ii. Used to find if two samples have same distribution shape
iii. Used to find if the data samples have similar tail behaviour

Importance:

i. It is used to find many distributional aspects such as shift in location, shifts in scale, changes in symmetry and presence of outliers.