LINEAR REGRESSION ASSIGNMENT ON BIKE SHARING

- Vinoda Varshini Rajagopal

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)

Answer:

Categorical Variable that were analysed with dependent variable (cnt) are Season, Weathersit, Weekday, Month and Year

Season:

- **3 : Fall** has **highest** number of bike rentals.
- **1 : Spring** has **least** number of bike rentals.

Weathersit:

- When the weather is "**1: Clear, Few clouds, Partly cloudy, Partly cloudy**" it is mentioned as **Clear** in our code. There is **high** number of bike rentals.
- When the weathersit is "**3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds**", mentioned as **Light Snow** in our code, During Light Snow **bike rental has decreased**.

Weekday:

- Number of bike rentals are **high on Friday** where 487790 bikes were rented.
- **Second highest** is on **Thursdays** has total count of 485395.
- **Third highest** is on **Saturday**, which has total of 477807 bike rentals.
- Number of bike rentals are **least on Sunday** where 444027 bikes were rented.

Month:

- In the year 2018
    - **June** has **highest** count of bike rentals where **143512 bikes were rented**.
    - **July** has **next highest** count of bike rentals **141341.**
    - **February** has **lease number** where **48215** bikes were rented.

- In the year 2019
    - **September** has highest count of bike rentals **218573.**
    - Followed by **August** where **214503** were rented.

Year:

- Compared to 2018, 2019 has more number of bike rentals.
- In the year 2018, total number of bikes rented was 1243103.
- In the year 2019, total number of bikes rented was 2047742.

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

Answer:
It is important to use **drop_first=True** when we create dummy variables by pd.get_dummies().

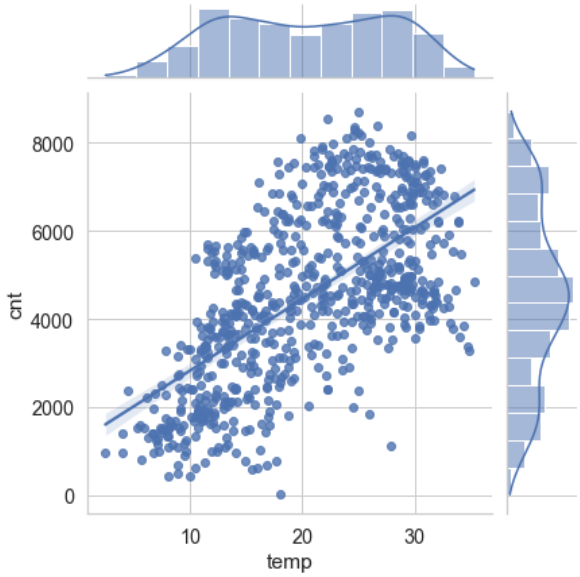
When we fail to do drop_first=True, it causes multicollinearity in our data.
This causes high correlation among our independent variable

Example: We have an independent variable "Income" which has three categories "High" ,"Medium" and "Low". When we use pd.get_dummies on this column without using drop_first=True, it will create three separate columns, where one column will have '1' and remaining columns have '0'.
So, this leads to multicollinearity.

When we use **drop_first=True,** we will get only n-1 column out of n columns and chances of multicollinearity is reduced.

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

Answer:
From the pair-plot among the numeric variables **temp(Temperature)** has highest correlation with target variable.

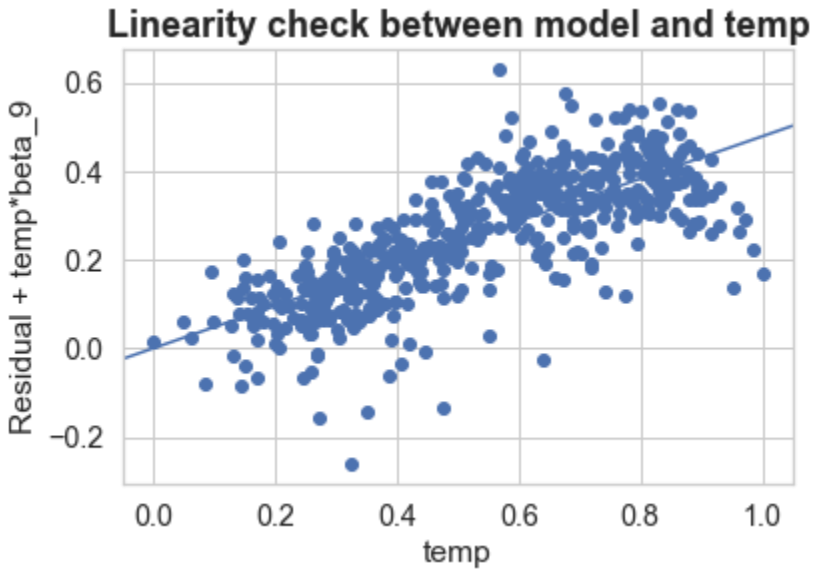


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

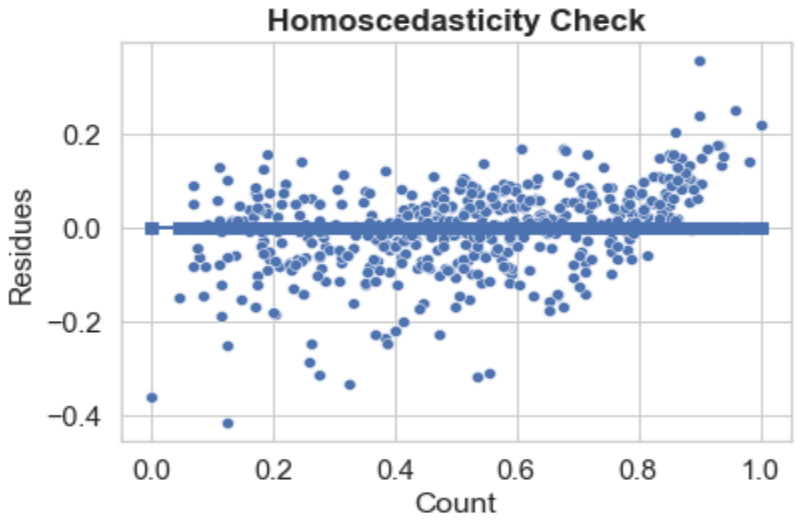
Answer:
Assumptions of linear regression are,

1. **Linear Relationship**
   There should be a linear relationship between dependent and independent variable.

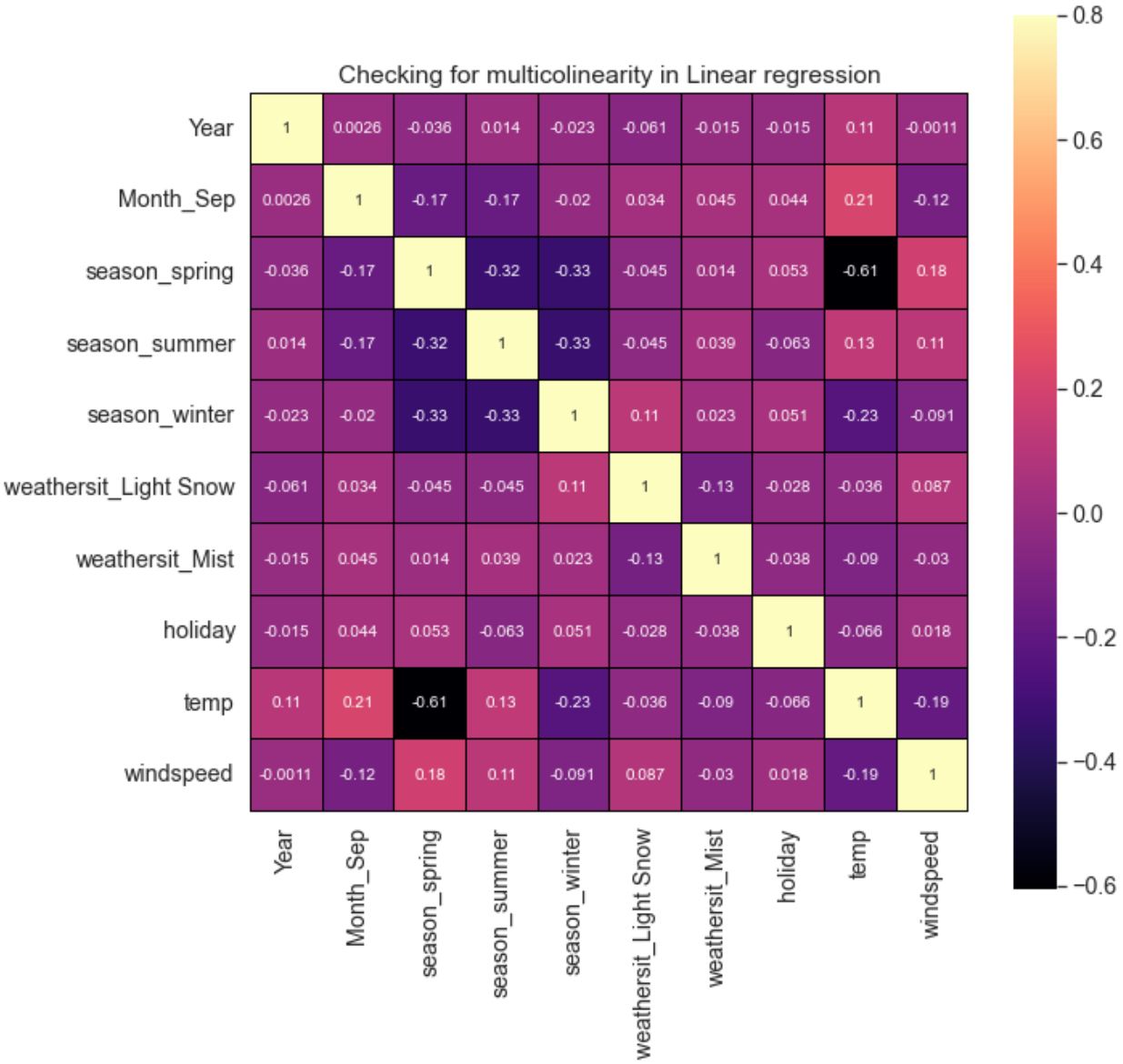
   Below graph shows linear relationship in our model with respect to Temperature (temp).
   Linear relationship holds true.



2. **Homoscedasticity**
   Variance of the error term should be constant.



3. **Absence of Multicollinearity**
   There should not be correlation between independent variables in our model.

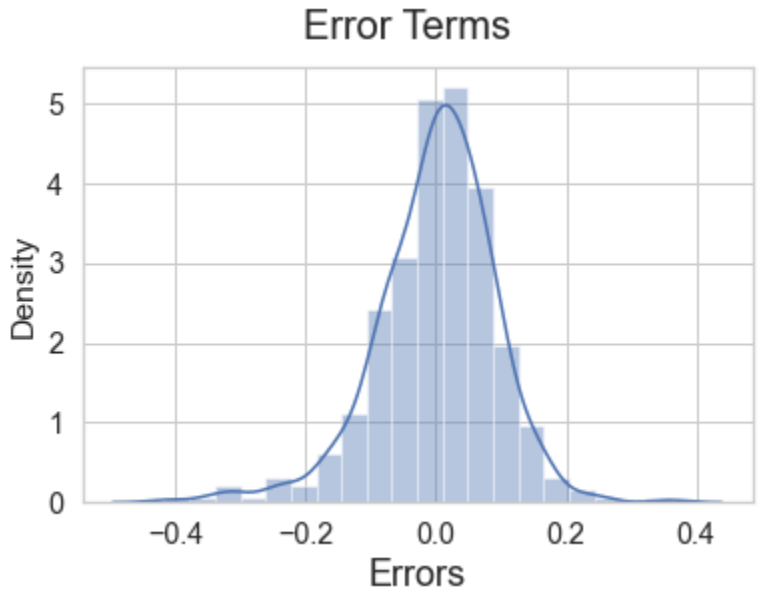


4. **No autocorrelation of errors**
   - There should not be auto correlation of errors. Error terms in linear regression should be independent. It is achieved by Durbin Watson test.
   - Acceptable values from Durbin Watson test is in the range 1.8 to 2.2
   - We can find this by

```
from statsmodels.stats.stattools import durbin_watson
durbin_watson(bike_lm7.resid)
```

   - Our model has **Durbin-Watson: 2.07570**

5. **Normality of Errors**
    - In our model error terms are normally distributed.



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

Answer:
Top three features that are contributing in explaining the demand of the shared bikes are

1. temp
2. yr : year (0: 2018, 1:2019)
3. Month_Sep (mnth : 9 for September)

General Subjective Questions

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

Answer:

Linear regression is a basic and commonly used type of predictive analysis. Linear regression examine the following.

- ❖ A set of predictor variable (independent variable) do a good job in predicting an outcome (dependent variable).
- ❖ Identifies which variable are significant in predicting the dependent variable, and how it impacts the target variable.

In linear regression, there will be one target variable and one or more independent variable.

Linear Regression model fits a linear relationship between the independent variable and dependent variable.

We can classify linear regression into two types

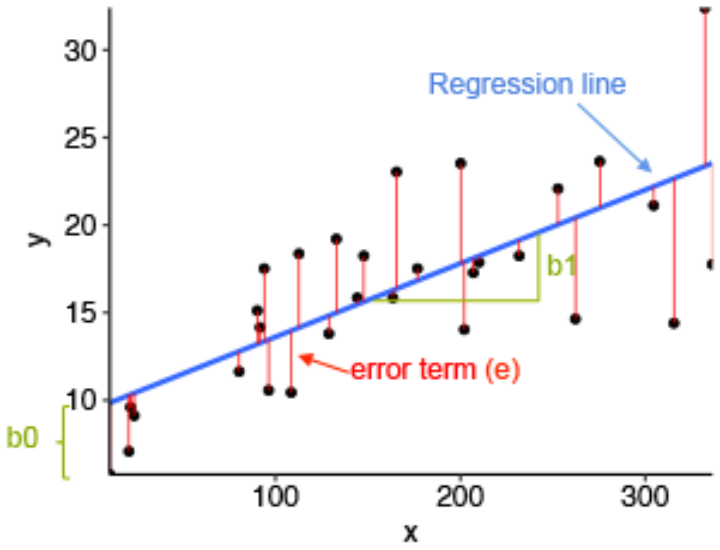
1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression

- Simple Linear Regression predicts the dependent variable (y) with one independent variable(x).

Equation of simple linear regression

$$y = B0 + B1 * x + e$$



Where,

y – predicted value

B0 and B1 are regression beta coefficients or parameters.

B0 - intercept

B1 – slope

e – error term also known as residual errors

Multiple Linear Regression

- Multiple Linear regression predicts the dependent variable or target variable with two or more independent variable.

Equation of multiple linear regression

$$y = B0 + B1 * X1 + B2 * X2 + \ldots\ldots\ldots + Bn * Xn + e$$

Where,

y – predicted value

B0 – intercept

B1 X1 – the regression coefficient (B1) of the first independent variable (X1)

Bn Xn – Last regression coefficient of independent variable

e – error term

Uses of Linear Regression

- Determining the strength of predictors
- Forecasting an effect
- Trend Forecasting

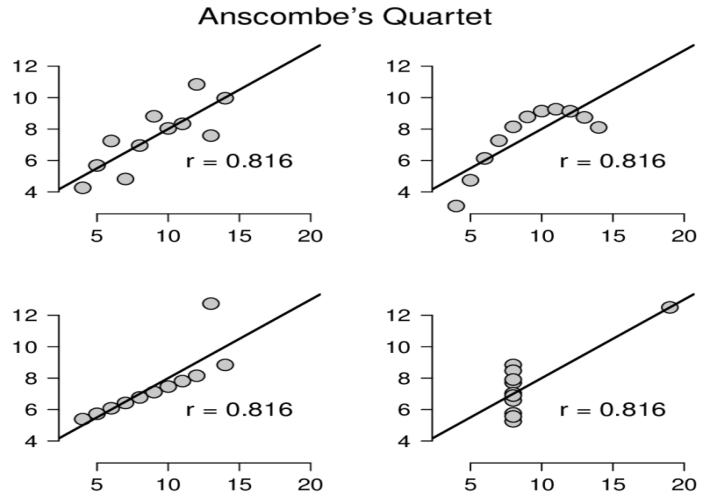
Assumptions of Linear Regression

- Linear relationship between independent and dependent variable.
- Homoscedasticity
- Absence of multicollinearity
- No autocorrelation of errors
- Normality of errors

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

Answer:

Anscombe's quartet is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs.



- Average x value is 9 for each dataset. Average y value is 7.50 for each dataset.

- Variance for x is 11 and variance for y is 4.12.
- Correlation between x and y is 0.816.
- Linear regression for each dataset follows the equation y = 0.5x + 3

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

Answer:

- Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r.
- Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables.

Pearson Correlation coefficient Formula

$$r = \Sigma (xi- \bar{x})(yi- \bar{y}) / \sqrt{\Sigma(xi- \bar{x})^2 \Sigma (yi- \bar{y})^2}$$

Where,

r = correlation coefficient
xi = x-values in the sample
$\bar{x}$ = mean of x variable
yi = x-values in the sample
$\bar{y}$ = mean of y variable

- Pearson correlation coefficient 'r' can take values from -1 to +1.
  - A value of 0 indicates that there is no association between the two variables.
  - Value greater than 0 indicates a positive association. Which means when value of one variable increases, other variable also increases.
  - Value less than 0 indicates negative association. Which means when value of one variable increases, the value of other variable decreases.

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

Scaling

Feature scaling is a technique of bringing down the values of all the independent features of our dataset on the same scale. Feature scaling helps to do calculations in algorithms very quickly.

Why Scaling need to be performed?

If scaling is not performed our model gives high weightage to higher values and low weightage to lower values. It takes more time for training a model.

Normalized Scaling	Standardized Scaling
1. Values are rescaled in the range 0 to 1	1. There is no range in standard scaling
2. Minimum and maximum values of features are used for scaling.	2. Mean and standard deviation are used for scaling.

3. MinMaxScaler from Scikit-Learn is used for Normalization.	3. StandardScaler from Scikit-Learn is used for Standardization.
4. Formula X' = (X-Xmin)/(Xmax-Xmin)	4. Formula X' = (X- μ)/ σ

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

(3 marks)

Answer:

When VIF = Infinity, it denotes perfect correlation between the explanatory variables. Which means the corresponding variable may be expressed exactly by a linear combination of other variables.

Formula to calculate VIF

$$VIF = 1 / (1- R2)$$

If we calculate R2 we get R2 = 1, when VIF is infinity.

We need to overcome this problem by eliminating one variable that causes multi collinearity and calculate VIF again.

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

(3 marks)

Answer:

Quantile-Quantile (Q-Q) plot, plots the quantiles of a sample distribution against quantiles of a theoretical distribution. This helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

Advantages of Q-Q plot

- To find whether two populations are of same distribution.
- Skewness of distribution
- It can be used with sample sizes
- Presence of outliers, distributional aspects like shifts in location, shifts in scale, changes in symmetry.

Interpretation of two data sets

- Similar Distribution : If all point of quantiles lies on or close to straight line at an angle of 45 degree from x-axis
- Y-value < X-value: If y quantiles are lower than x quantiles.
- X-value < y-value: If x quantiles are lower than y quantiles.
- Different Distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x-axis.