Bike Sharing Assignment on Linear Regression

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

The below are the inferences from the analysis:

- The Count of bike rentals is increased during Summer and fall seasons and started decreasing during winter and spring.

- There is a gradual increase in bike rentals from 2018 to 2019.

- September has the highest number of rental counts and there is a fall in the median after that till December and a gradual increase from January

- The median of the rental counts is almost same for all the days.

- The bike rental count is high when the sky is clear and is gradually decresing when its cloudy or raining.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer:

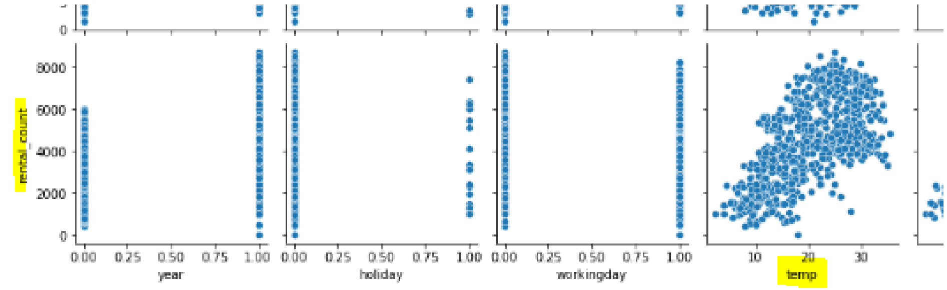
When we have categorical variables of ‘n’ levels, the idea of dummy variables is to build ‘n-1’ variables, in order to achieve this, we are using drop_first=True.

Example if we have 3 levels in a categorical variable we will have 2 dummy variables, if the first 2 categories are not present then by default it is the 3rd category.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

We can infer from pair-plot that temp (independent variable) has highest correlation with rental count (Target variable)

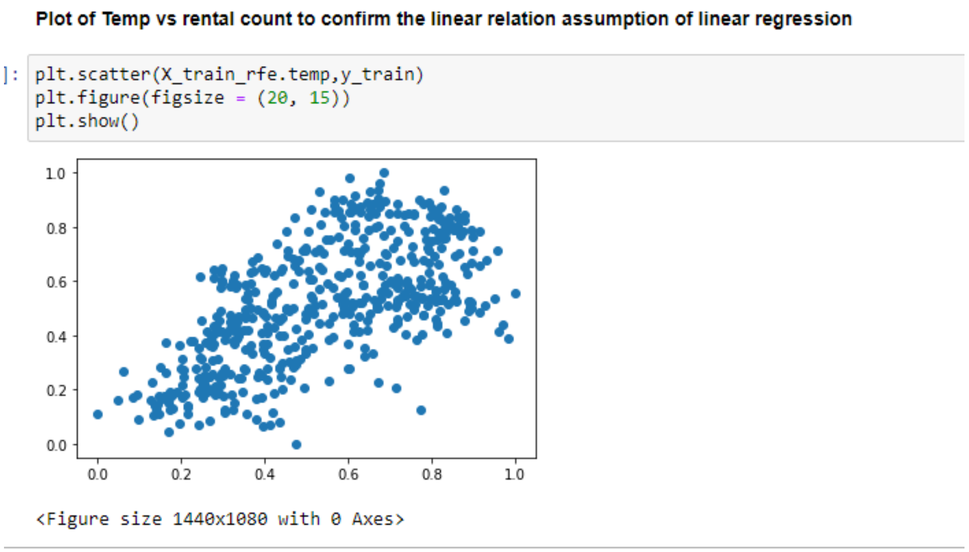


Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

I have validated the assumption of linear regression by the following details

- There is a linear relationship between dependent and independent variable

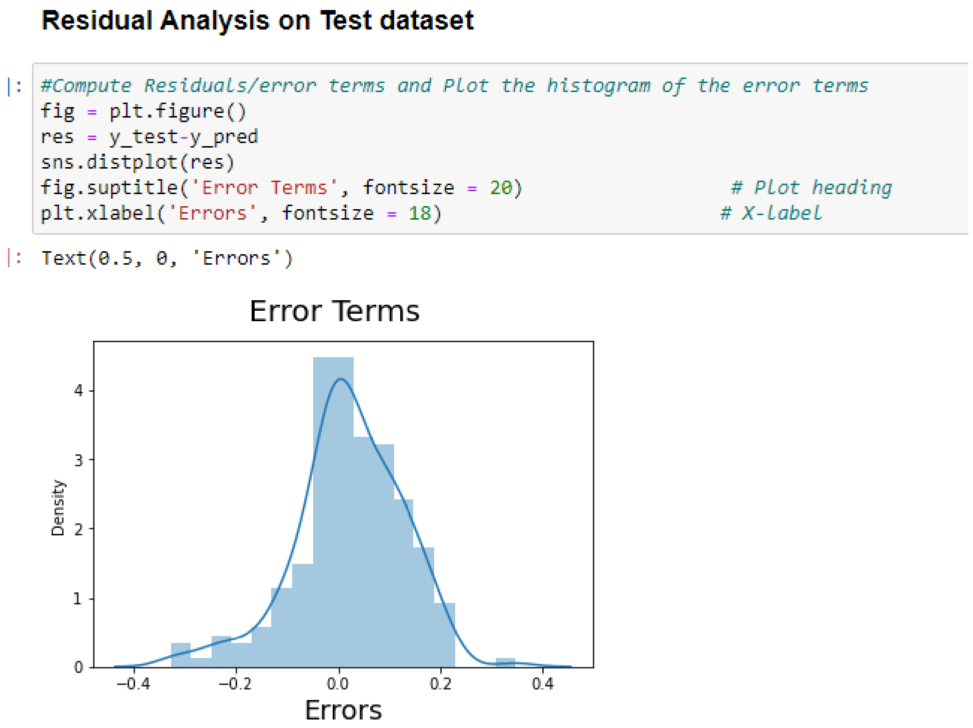


Plot of Temp vs rental count to confirm the linear relation assumption of linear regression

```
]: plt.scatter(X_train_rfe.temp,y_train)
   plt.figure(figsize = (20, 15))
   plt.show()
```

```
<Figure size 1440x1080 with 0 Axes>
```

- Error terms of test and train data has normal distribution

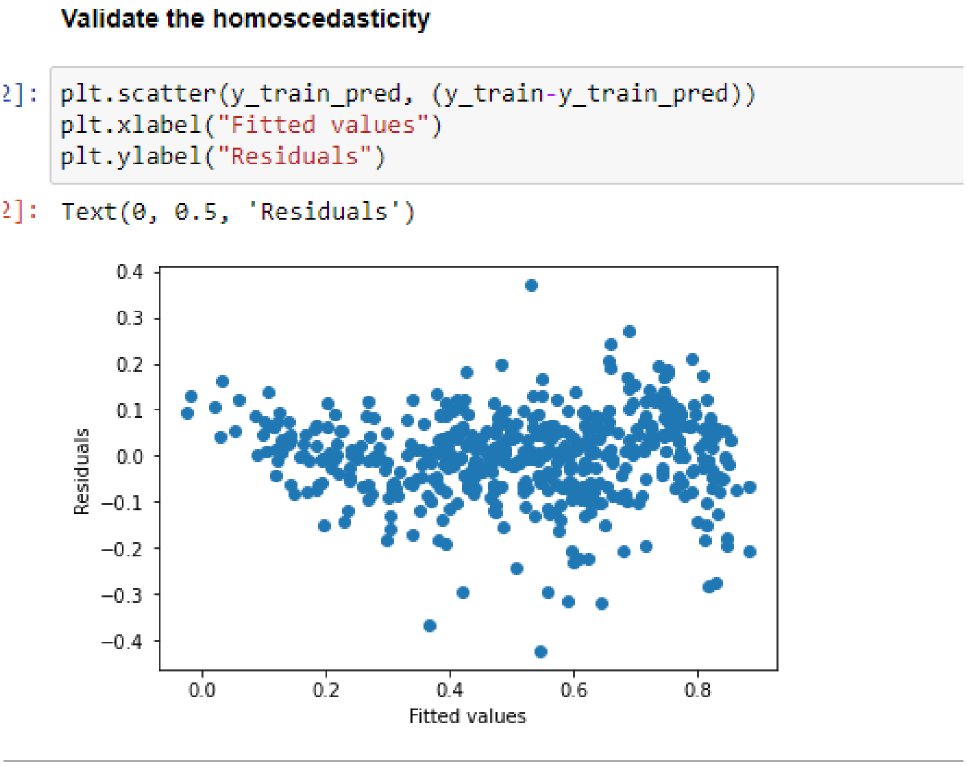


Residual Analysis on Test dataset

```
]: #Compute Residuals/error terms and Plot the histogram of the error terms
   fig = plt.figure()
   res = y_test-y_pred
   sns.distplot(res)
   fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
   plt.xlabel('Errors', fontsize = 18)                        # X-label
```

```
]: Text(0.5, 0, 'Errors')
```

- Error terms have constant variance that is homoscedastic



Validate the homoscedasticity

```
2]: plt.scatter(y_train_pred, (y_train-y_train_pred))
    plt.xlabel("Fitted values")
    plt.ylabel("Residuals")
```

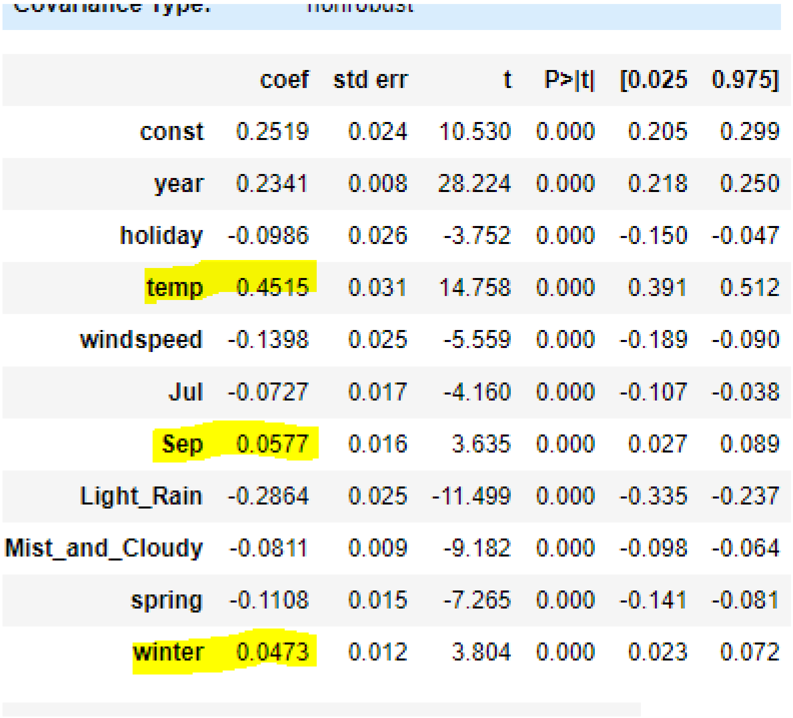
```
2]: Text(0, 0.5, 'Residuals')
```

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

From the final model built we can see that the Top 3 features are:

1. Temperature with a coefficient (B1 of 0.4515)
2. Season Winter with a coefficient (B1 of 0.0473)
3. Month September with a coefficient (B1 of 0.0577)



Covariance Type: nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2519	0.024	10.530	0.000	0.205	0.299
year	0.2341	0.008	28.224	0.000	0.218	0.250
holiday	-0.0986	0.026	-3.752	0.000	-0.150	-0.047
temp	0.4515	0.031	14.758	0.000	0.391	0.512
windspeed	-0.1398	0.025	-5.559	0.000	-0.189	-0.090
Jul	-0.0727	0.017	-4.160	0.000	-0.107	-0.038
Sep	0.0577	0.016	3.635	0.000	0.027	0.089
Light_Rain	-0.2864	0.025	-11.499	0.000	-0.335	-0.237
Mist_and_Cloudy	-0.0811	0.009	-9.182	0.000	-0.098	-0.064
spring	-0.1108	0.015	-7.265	0.000	-0.141	-0.081
winter	0.0473	0.012	3.804	0.000	0.023	0.072

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is a type of supervised machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

When the model has only 1 independent variable then it is **simple linear regression**

When the model has more than 1 independent variable then it is **Multiple linear regression**

Mathematically, we can write a linear regression equation as:

$y = a + bx$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

The Assumption of linear regression are:

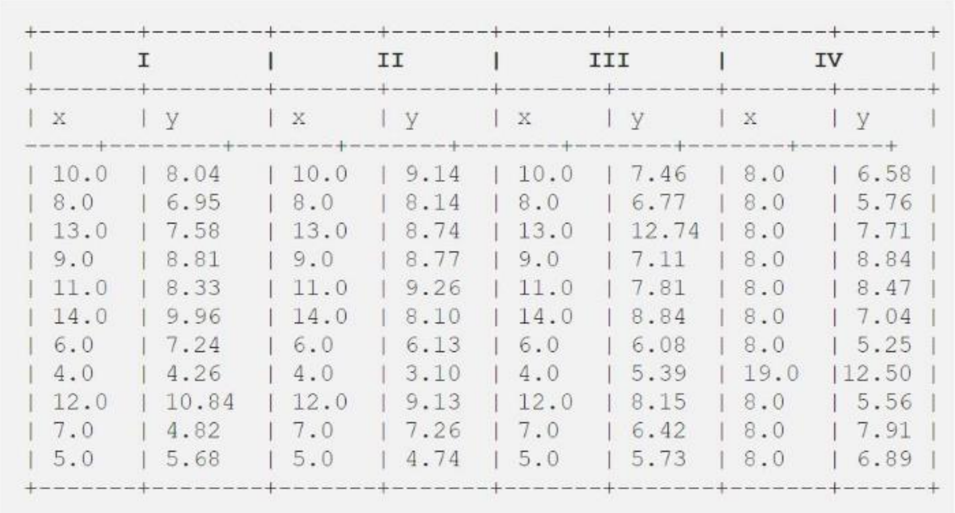
1. There should be a linear relation between dependent and independent variables
2. The error terms should be normally distributed
3. The error terms should be independent of each other
4. The error terms should have constant variance(homoscedastic)

Q2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer:

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

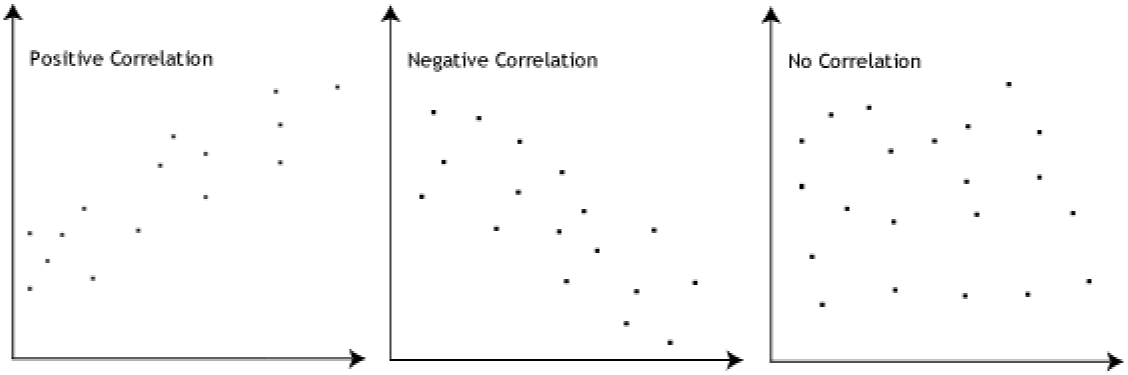
Q3. What is Pearson’s R? (3 marks)

Answer:

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by 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.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula:

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is the step of processing the data to a normalized/standardized form with in a particular range.

Scaling is performed on the dataset so that the regression algorithm finds it easy to calculate and speeds up the process

Normalization typically means rescales the values into a range of [0,1].

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.

2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

Answer:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. 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).

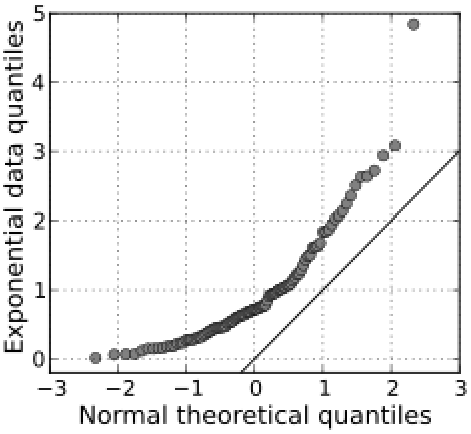
Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Answer:

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 Q Q 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.

A Q Q plot showing the 45-degree reference line:



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. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

--

Thank You

Submitted by,

Hanumanth A