Linear Regression Assignment Questions and Responses

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?**

Answer:

Please find below inferences based on my analysis

- ✓ **Year**: 62.33 % of total Bike hired in year 2019. So, we can say, bike hiring increases by every year
- ✓ **Month**: More number of bike hiring happened between May and Oct (between ~32K & ~35K). In which Aug month has high.
- ✓ **Season**: Highest number of bike hiring is in fall and then summer seasons compare to winter and spring
- ✓ **Weathersit**: Around 69% of total bike hiring is in clear / partial cloudy weather and then 30.24% in the mist weather. If it's raining, the bike hiring is almost not happening
- ✓ **Holidays**: Around 97.62% of bike hiring is on non holidays
- ✓ **Working Days**: Around 69.6% of bike hiring is on working days. So, people use more bike hiring on working day
- ✓ With good weather conditions in fall and summer season and in the month between May and Oct on working days, bike hire will be more

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

Answer:

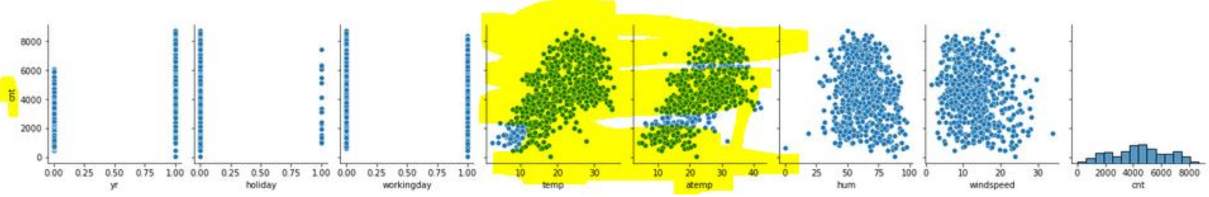
drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

In our assignment, we can take example of season variable. Where, it created dummy variable with only 3 values ('season_spring', 'season_summer','season_winter') but not fall.

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

Answer:

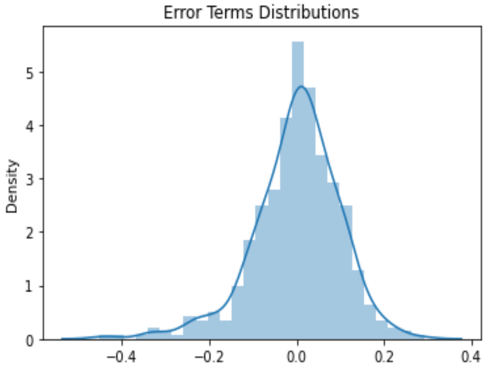
Temperature (temp) and Actual Temperature (atemp) has highest correlation with the target variable (cnt).

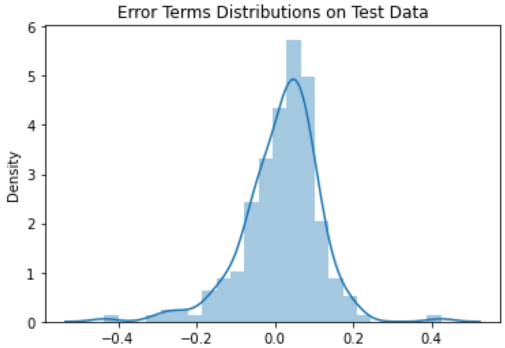


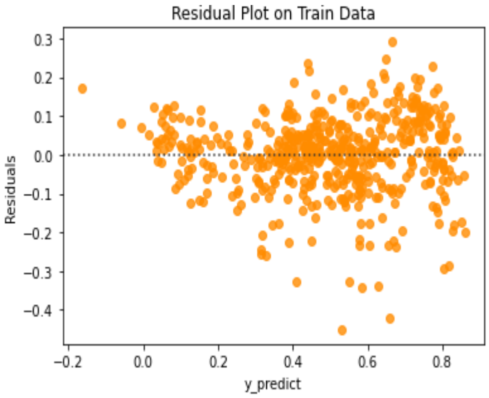
4) How did you validate the assumptions of Linear Regression after building the model on the training set?

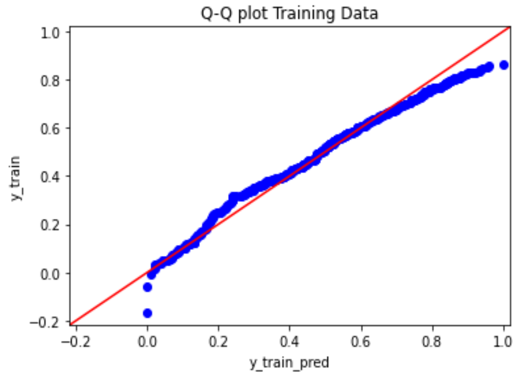
Answer:

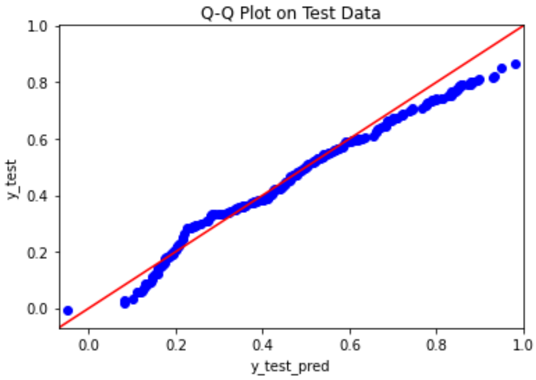
- ✓ The distribution of error terms approximately centered around 0 ie Residuals are normally distributed. Hence our assumption for Linear Regression is valid.
- ✓ From the residual plot, we could infer that the residuals didn't form any pattern. So, the residuals are independent of each other.
- ✓ The residuals have constant variance. Variance doesn't seem to increase/decrease constantly with the y_predict value.
- ✓ Residuals are normally distributed as the Q-Q plot of residuals will be a straight line. Points are very close to straight line of 45 degree
- ✓ Perform the Linearity check plotted actual and predicted values to test the linearity

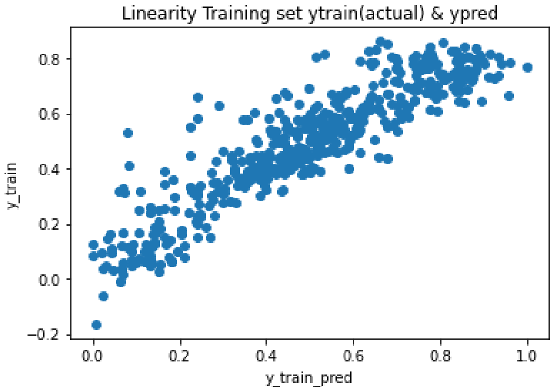


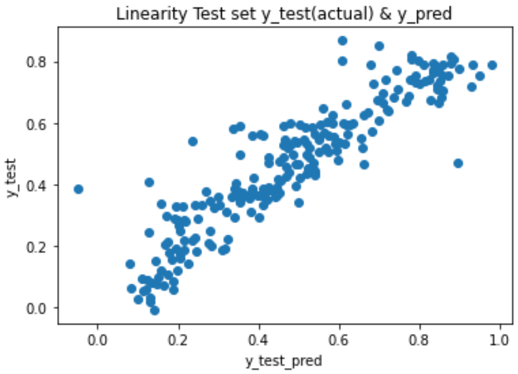








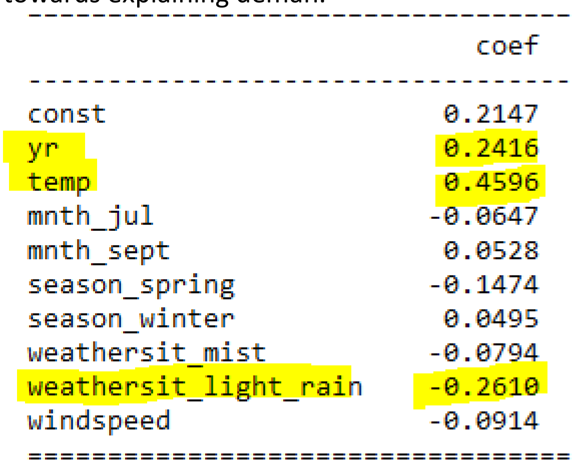




5) **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Answer:

Temperature, Year and weathersit (negatively) are the three variables contributing significantly towards explaining deman.



```
----------------------------------
                              coef
----------------------------------
const                       0.2147
yr                          0.2416
temp                        0.4596
mnth_jul                   -0.0647
mnth_sept                   0.0528
season_spring              -0.1474
season_winter               0.0495
weathersit_mist            -0.0794
weathersit_light_rain      -0.2610
windspeed                  -0.0914
==================================
```

General Subjective Questions

1) Explain the linear regression algorithm in detail?

Answer:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour 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.

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

Assumptions:

Before implementing linear regression, we should check whether the data is following these assumptions:

- ✓ Data should be linear
- ✓ No Multicollinearity
- ✓ No auto-correlation
- ✓ Homoskedasticity should be there

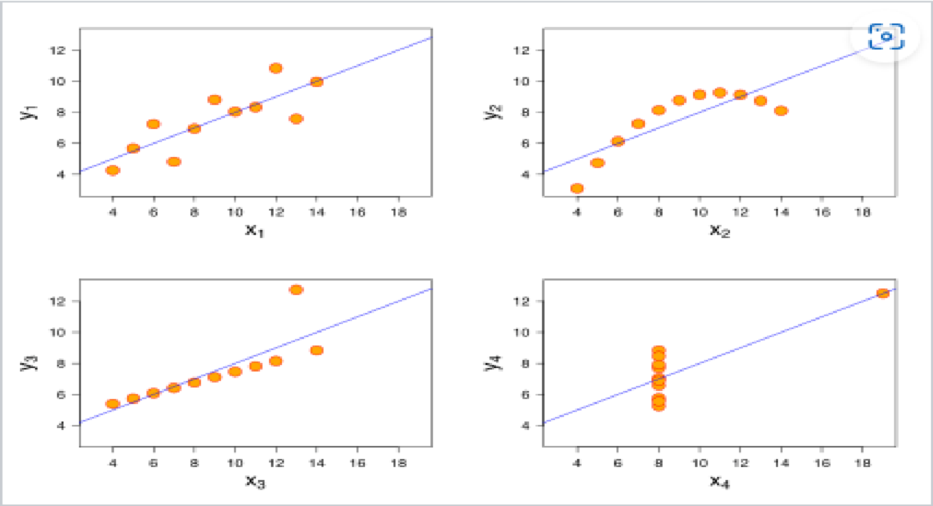
2) Explain the Anscombe's quartet in detail

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 when analyzing it, and the effect of outliers and other influential observations on statistical properties.

Four datasets:

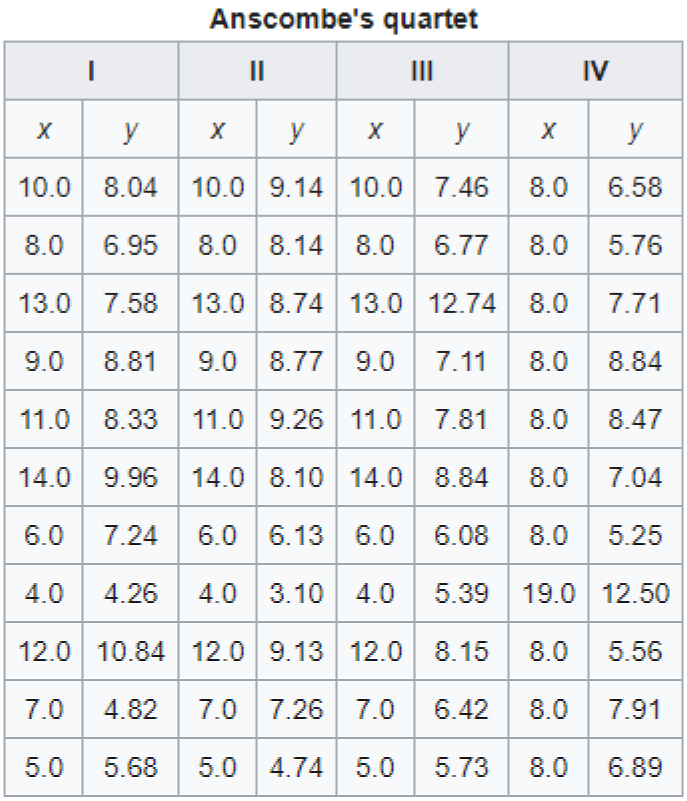
- ✓ The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- ✓ The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- ✓ In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- ✓ Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

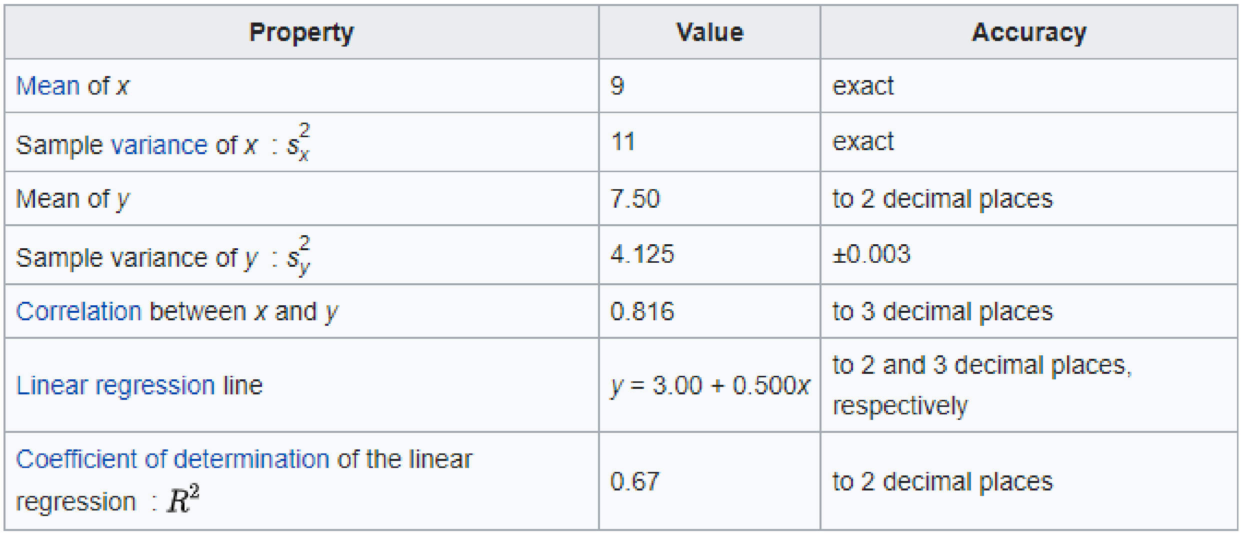
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The x values are the same for the first three datasets.



Anscombe's quartet

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



Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Usage wrt Data Visualization:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3) What is Pearson's R?

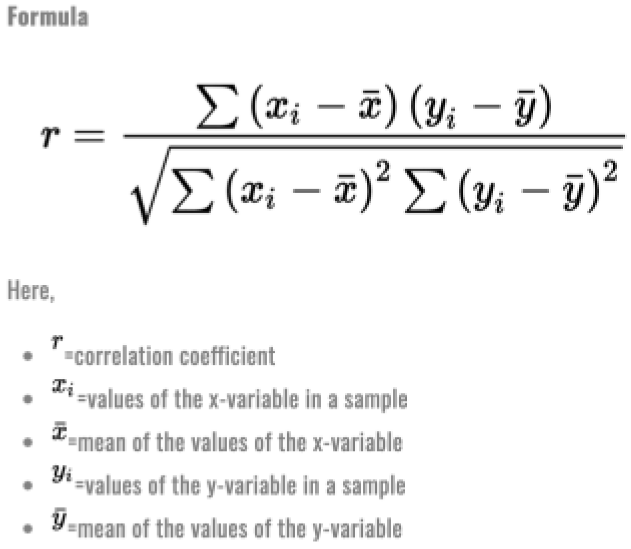
Answer:

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 correlation coefficient has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. This linear relationship can be positive or negative.

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

a. **Perfect**: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
b. **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
c. **Moderate degree**: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
d. **Low degree:** When the value lies below + .29, then it is said to be a small correlation.
e. **No correlation**: When the value is zero.

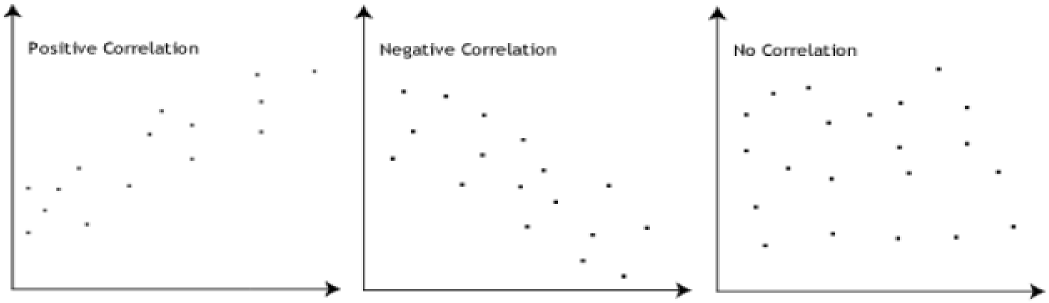


Formula

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

Here,

- r=correlation coefficient
- x_i=values of the x-variable in a sample
- $\bar{x}$=mean of the values of the x-variable
- y_i=values of the y-variable in a sample
- $\bar{y}$=mean of the values of the y-variable



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

Answer:

Scaling is a technique of bringing down the values of all the independent features of our dataset on the same scale. Feature selection helps to do calculations in algorithms very quickly. It is the important stage of data preprocessing.
If we didn't do feature scaling then the machine learning model gives higher weightage to higher values and lower weightage to lower values. Also, takes a lot of time for training the machine learning model.

Why Scaling?
When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

- ✓ Ease of interpretation
- ✓ Faster convergence for gradient descent methods

If a feature in the dataset is big in scale (Salary) compared to other (Age) then in model it gives higher weightage to higher values and lower weightage to lower values. this big scaled feature becomes dominating and needs to be normalized/standardized.

Difference between normalized scaling and standardized scaling?

Standardizing:
The variables are scaled in such a way that their mean is zero and standard deviation is one.

x = [x – mean(x)] / sd(x)

Use-case of Standardized Scaling:

- ➢ In most of the Machine Learning models and it outperform MinMaxScaler(Normalization).

- Anywhere, where there is no need to scale features in the range 0 to 1.
- Since, it transforms the normal data distribution to standard normal distribution, which is the ideal & expected to have, most of the time it is the best to use in machine learning models

Normalization:
The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data .

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Use-case of Normalization:

- Every situation where the range of features should be between 0 to 1. For example, in Images data, there we have color pixels range from 0 to 255(256 colors in total), here Normalizer is the best one to use.
- There can be multiple scenarios where this range is expected, there it is optimal to use MinMaxScaler.

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

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)

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

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions

Advantages:

- It can be used with sample sizes also.

- ➢ Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets

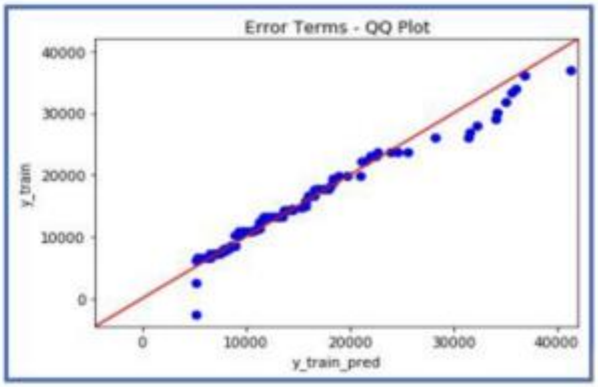
- ✓ come from populations with a common distribution
- ✓ have common location and scale
- ✓ have similar distributional shapes
- ✓ have similar tail behavior

<u>Interpretation</u>

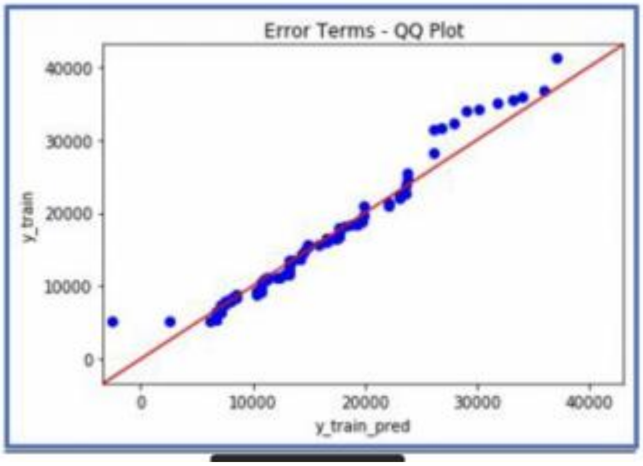
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis