1**. Explain the linear regression algorithm in detail.**

1.**Linear Regression:** Linear Regression is a linear approach to modeling the relationship between a dependent variable or y and one or more explanatory variables or independent variables or x.

The linear regression model can be represented by the following equation:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

where, Y is the predicted value

θ_0 is the bias term.

θ_1,...,θn are the model parameters

x_1, x_2,...,xn are the feature values.

The above hypothesis can also be represented by

$$Y = \theta^T x$$

Where, θ is the model's parameter vector including the bias term θ_0; x is the feature vector with x_0 =1

Y (pred) = b0 + b1*x

The values b0 and b1 must be chosen so that the error is minimum. If sum of squared error is taken as a metric to evaluate the model, then the goal is to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

If we don't square the error, then the positive and negative points will cancel each other out.

For a model with one predictor,

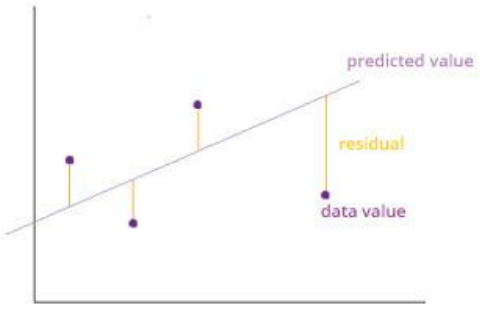
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

If b1 > 0, then x (predictor) and y(target) have a positive relationship. That is an increase in x will increase y.If b1 < 0, then x (predictor) and y(target) have a negative relationship. That is an increase in x will decrease y.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. The error is the distance between the point to the regression line.

The best fit line is considered to be the line for which the error between the predicted values and the observed values is minimum. It is also called the regression line and the errors are also known as residuals. The figure shown below shows the residuals. It can be visualized by the vertical lines from the observed data value to the regression line.



Process of Linear Regression:

1. **Simple Linear Regression**

With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.

This requires that you calculate statistical properties from the data such as means, standard deviations, correlations and covariance. All of the data must be available to traverse and calculate statistics.

Residual Analysis: Simple linear regression models the relationship between the magnitude of one variable and that of a second—for example, as x increases, y also increases. Or as x increases, y decreases. Correlation is another way to measure how two variables are related. The models done by simple linear regression estimate or try to predict the actual result but most often they deviate from the actual result. Residual analysis is used to calculate by how much the estimated value has deviated from the actual result.

Null Hypothesis and p-value: During feature selection, null hypothesis is used to find which attributes will not affect the result of the model. Hypothesis tests are used to test the validity of a claim that is made about a particular attribute of the model. This claim that's on trial, in

essence, is called the null hypothesis. A p-value helps to determine the significance of the results. p-value is a number between 0 and 1 and is interpreted in the following way:

A small p-value (less than 0.05) indicates a strong evidence against the null hypothesis, so the null hypothesis is to be rejected.

A large p-value (greater than 0.05) indicates weak evidence against the null hypothesis, so the null hypothesis is to be considered.

p-value very close to the cut-off (equal to 0.05) is considered to be marginal (could go either way). In this case, the p-value should be provided to the readers so that they can draw their own conclusions.

Ordinary Least Square

Ordinary Least Squares (OLS), also known as Ordinary least squares regression or least squared errors regression is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters for a linear function, the goal of which is to minimize the sum of the squares of the difference of the observed variables and the dependent variables i.e. it tries to attain a relationship between them.

Multiple Linear Regression

Data preparation for Linear Regression

Step 1: Linear Assumption

The first step for data preparation is checking for the variables which have some sort of linear correlation between the dependent and the independent variables.

Step 2: Remove Noise

It is the process of reducing the number of attributes in the dataset by eliminating the features which have very little to no requirement for the construction of the model.

Step 3: Remove Collinearity

Collinearity tells us the strength of the relationship between independent variables. If two or more variables are highly collinear, it would not make sense to keep both the variables while evaluating the model and hence we can keep one of them.

Step 4: Gaussian Distributions

The linear regression model will produce more reliable results if the input and output variables have a Gaussian distribution. The Gaussian theorem states that states that a sample mean from an infinite population is approximately normal, or Gaussian, with mean the same as the underlying population, and variance equal to the population variance divided by the sample size. The approximation improves as the sample size gets large.

Step 5: Rescale Inputs

Linear regression model will produce more reliable predictions if the input variables are rescaled using standardization or normalization.

Linear Regression with statsmodels

OLS method in stats models library will be used.

Example:

```
import statsmodels.formula.api as smf

# Initialise and fit linear regression model using `statsmodels`

model = smf.ols('Sales ~ TV', data=advert)

model = model.fit()
```

Once we have fit the simple regression model, we can predict the values of sales based on the equation we just derived using the .predict method and also visualise our regression model by plotting sales_pred against the TV advertising costs to find the line of best fit.

```
# Predict values

sales_pred = model.predict()

# Plot regression against actual data

plt.figure(figsize=(12, 6))

plt.plot(advert['TV'], advert['Sales'], 'o')       # scatter plot showing actual data

plt.plot(advert['TV'], sales_pred, 'r', linewidth=2)   # regression line

plt.xlabel('TV Advertising Costs')

plt.ylabel('Sales')
```

plt.title('TV vs Sales')

plt.show()

Linear Regression with scikit-learn

Build linear regression model using TV and Radio as predictors

 # Split data into predictors X and output Y

predictors = ['TV', 'Radio']

X = advert[predictors]

y = advert['Sales']

 # Initialise and fit model

lm = LinearRegression()

model = lm.fit(X, y)

print(f'alpha = {model.intercept_}')

print(f'betas = {model.coef_}')

alpha = 4.630879464097768

betas = [0.05444896 0.10717457]

model.predict(X)

2. **What are the assumptions of linear regression regarding residuals?**

Assumptions about the residuals:

Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.

Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.

Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

3. **What is the coefficient of correlation and the coefficient of determination?**

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

1 indicates a strong positive relationship.

-1 indicates a strong negative relationship.

A result of zero indicates no relationship at all

A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The coefficient of determination (denoted by R2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.

With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.

An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.

An R2 of 1 means the dependent variable can be predicted without error from the independent variable.

An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

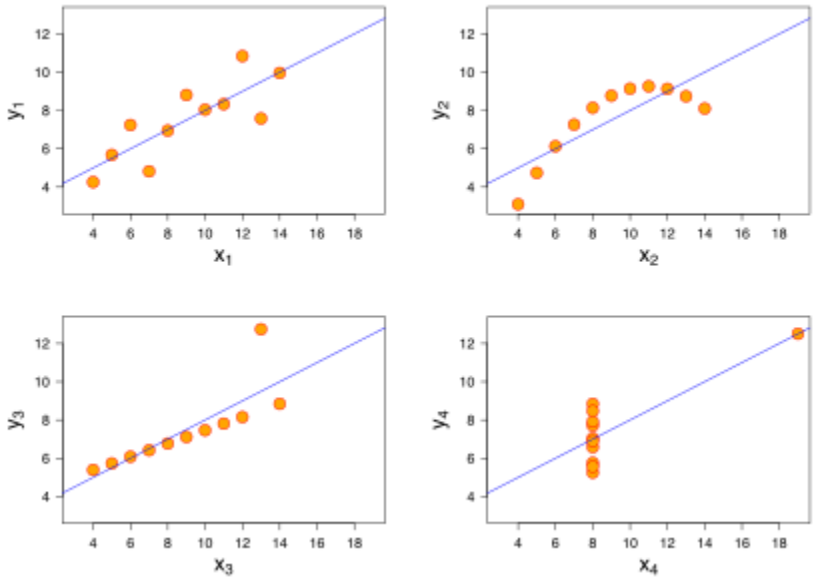
Coefficient of determination. The coefficient of determination (R2) for a linear regression model with one independent variable is:

R2 = { (1 / N) * Σ [(xi - x) * (yi - y)] / (σx * σy) }2

where N is the number of observations used to fit the model, Σ is the summation symbol, xi is the x value for observation i, x is the mean x value, yi is the y value for observation i, y is the mean y value, σx is the standard deviation of x, and σy is the standard deviation of y.

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

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 before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact

Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	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	0.67	to 2 decimal places

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) is not distributed normally; 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 distribution 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.

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.[2][3][4][5][6]

The datasets are as follows. The x values are the same for the first three datasets.[1]

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

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

There are several types of correlation coefficient formulas.

One of the most commonly used formulas in stats is Pearson's correlation coefficient formula.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

Potential problems with Pearson correlation.

The PPMC is not able to tell the difference between dependent variables and independent variables. For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also get the same result with the variables switched around. In other words, you could say that diabetes causes a high calorie diet. That obviously makes no sense. Therefore, as a researcher you have to be aware of the data you are plugging in. In addition, the PPMC will not give you any information about the slope of the line; it only tells you whether there is a relationship.

Real Life Example

Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson's correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

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

Some machine learning algorithms are sensitive to feature scaling while others are virtually invariant to it

Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. Take a look at the formula for gradient descent below:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

Distance algorithms like KNN, K-means, and SVM are most affected by the range of features. This is because behind the scenes they are using distances between data points to determine their similarity.

For example, let's say we have data containing high school CGPA scores of students (ranging from 0 to 5) and their future incomes (in thousands Rupees):

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

Since both the features have different scales, there is a chance that higher weightage is given to features with higher magnitude. This will impact the performance of the machine learning algorithm and obviously, we do not want our algorithm to be biassed towards one feature.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0

On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1

If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

Differences:

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization,

standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

VIF = Variance Inflation Factor

A simple approach to identify collinearity among explanatory variables is the use of variance inflation factors (VIF). VIF calculations are straightforward and easily comprehensible; the higher the value, the higher the collinearity. A VIF for a single explanatory variable is obtained using the r-squared value of the regression of that variable against all other explanatory variables:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where the VIF for variable j is the reciprocal of the inverse of R^2 from the regression. A VIF is calculated for each explanatory variable and those with high values are removed. The definition of 'high' is somewhat arbitrary but values in the range of 5-10 are commonly used.

When R2 is equal to 1 then VIF becomes infinity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables .

8. **What is the Gauss-Markov theorem?**

The Gauss–Markov theorem states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero.[1] The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance

The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate (BLUE) possible.

Gauss Markov Assumptions

There are five Gauss Markov assumptions (also called conditions):

Linearity: the parameters we are estimating using the OLS method must be themselves linear.

Random: our data must have been randomly sampled from the population.

Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.

Exogeneity: the regressors aren't correlated with the error term.

Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by

$y_i = x_i' \beta + \varepsilon_i$

and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

$E\{\varepsilon_i\} = 0, i = 1, \ldots, N$

$\{\varepsilon_1 \ldots\ldots \varepsilon_n\}$ and $\{x_1 \ldots\ldots, x_N\}$ are independent

$cov\{\varepsilon_i, \varepsilon_j\} = 0, i, j = 1, \ldots., N\ I \neq j.$

$V\{\varepsilon_1 = \sigma^2, i = 1, \ldots.N$

The first of these assumptions can be read as "The expected value of the error term is zero.".

The second assumption is collinearity,

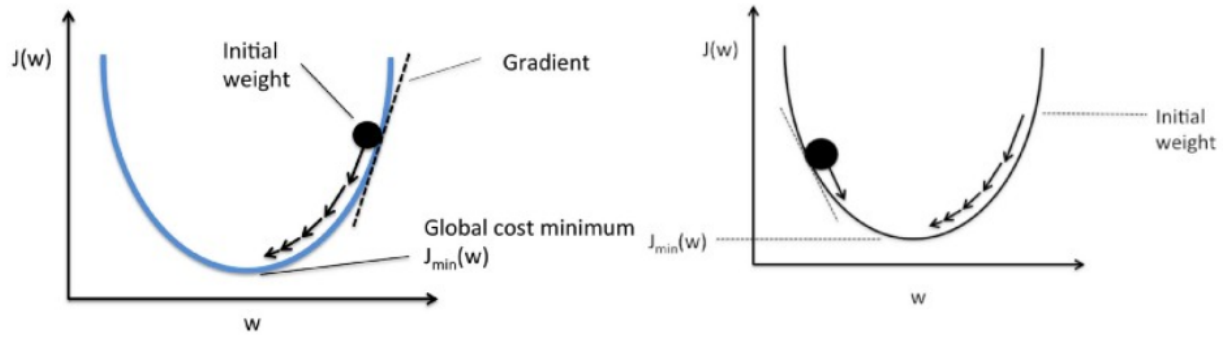
the third is exogeneity,

and the fourth is homoscedasticity.

9. **Explain the gradient descent algorithm in detail.**

Gradient descent is an optimization algorithm. In linear regression, it is used to optimize the cost function and find the values of the βs (estimators) corresponding to the optimized value of the cost function.

Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).



Gradient Descent

Mathematically, the aim of gradient descent for linear regression is to find the solution of

ArgMin J(Θ0,Θ1), where J(Θ0,Θ1) is the cost function of the linear regression. It is

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Here, h is the linear hypothesis model, h=Θ0 + Θ1x, y is the true output, and m is the number of data points in the training set.

Gradient Descent starts with a random solution, and then based on the direction of the gradient, the solution is updated to the new value where the cost function has a lower value.

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}).\, x_j^{(i)} \quad \text{for j = 1,2,...,n}$$

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

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.

Few advantages:

a) It can be used with sample sizes also

b) 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 —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

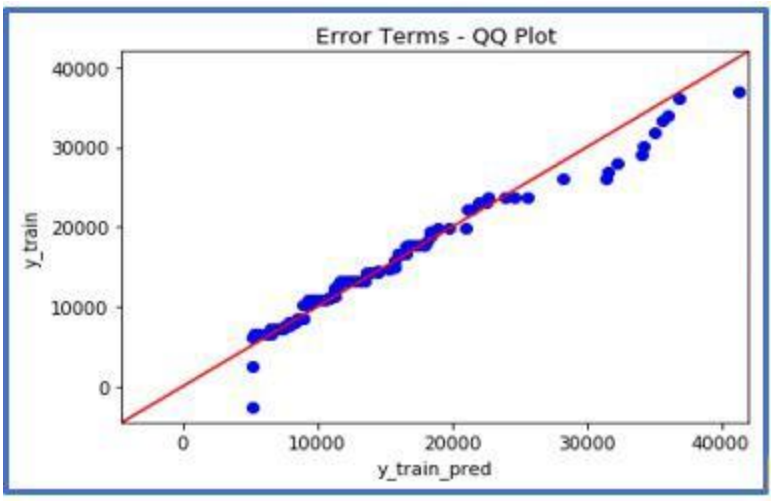
iv. have similar tail behavior

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

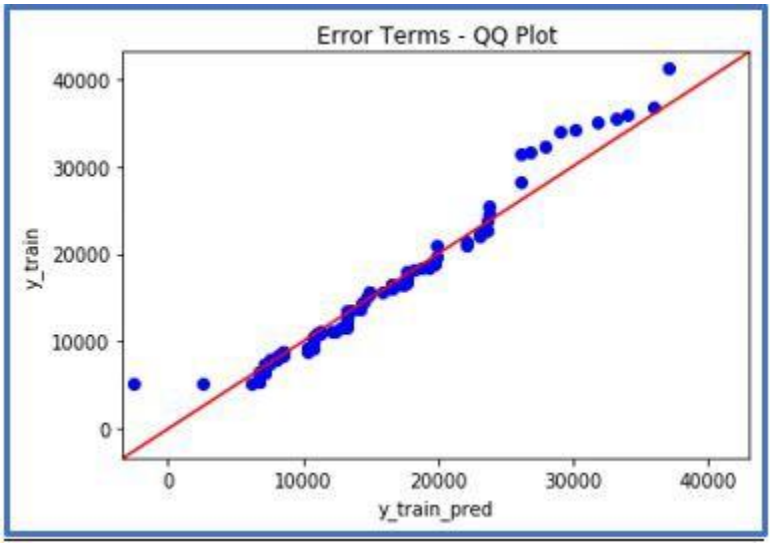
Below are the possible interpretations for two data sets.

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.