Submitted by: Barkha Saxena (Roll no. DML2020080)

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression is statistical method used for predictive analysis and modelling. It is a type of parametric regression where the aim is to determine the strength and character of the relationship between one outcome variable (dependent) and a few other variables (independent, also known as predictors). Based on some assumptions, a polynomial equation of degree 1 is derived of the form:

$$y' = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots + \beta_n x_n$$

where

y' = the predicted value of dependent / outcome variable

x = the independent variable (predictors)

β = coefficient of the respective independent variable

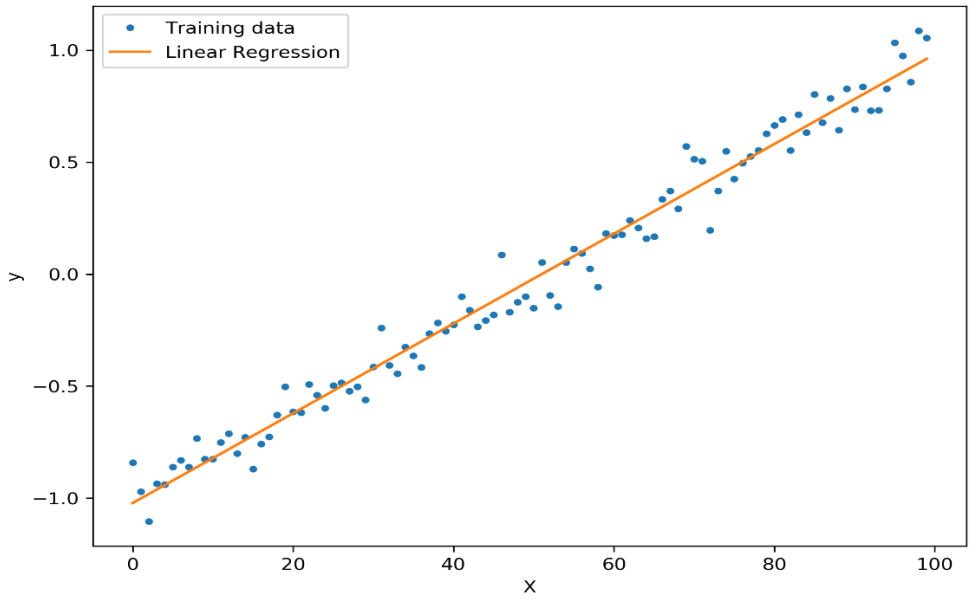


Figure showing a linear regression model (line) having one parameter. [image source: internet]

Algorithm Steps:

Step 1: Find the independent variables(predictors) related to y. Eliminate variables which are highly correlated to any other variables or can be explained by a combination of variables. Prepare categorical variables by encoding them if they are ordinal (here assuming that they are equally spaced) or by creating dummy variables. Feed only relevant numerical variables as predictors to model.

Step 2: Build and train a model on train data, based on the predictors (calculate coefficients and summary statistics). Check for significance and multicollinearity.

Step 3: Minimize cost by updating the coefficients by adding/removing predictors, checking the significance of model and each coefficient along with the models R^2 and adjusted R^2 values.

Error analysis:

Residual (error for point i) = $y'_i - y_{actual}$

Cost function = $\frac{1}{n}\sum_{i=1}^{n} y' - y_{actual}$

To minimize cost function, we update the coefficients of the equation using gradient descent.

To train the model, two measures can be used:

- RMSE: Root mean square of errors. Minimize RMSE for a better model
- R^2: Gives the measure of variance explained by the model. Maximize R^2 for a better model (also check adjusted R^2)

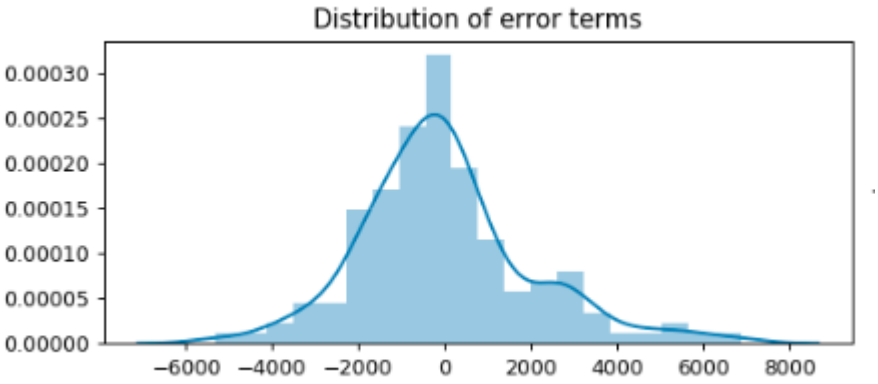
Step 4: Once an optimized model is ready, predict the value of dependent variable from test data

Step 5: Residual Analysis and model evaluation by plotting the predicted values and checking R2 score.

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

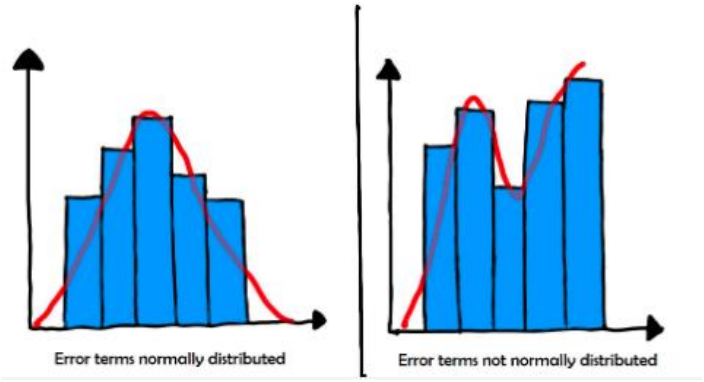
Ans. Assumptions related to linear regression to:

1. Fit the line on the data points:
   a. Outcome (dependent) variable has a linear relationship with the predictor variables
2. Infer about population from sample (*assumptions regarding residuals*):
   a. Error terms are normally distributed with a mean of 0 and constant standard deviation
      i. To check plot the distribution plot of error terms, a bell-shaped curve with mean at zero should be observed



ii.

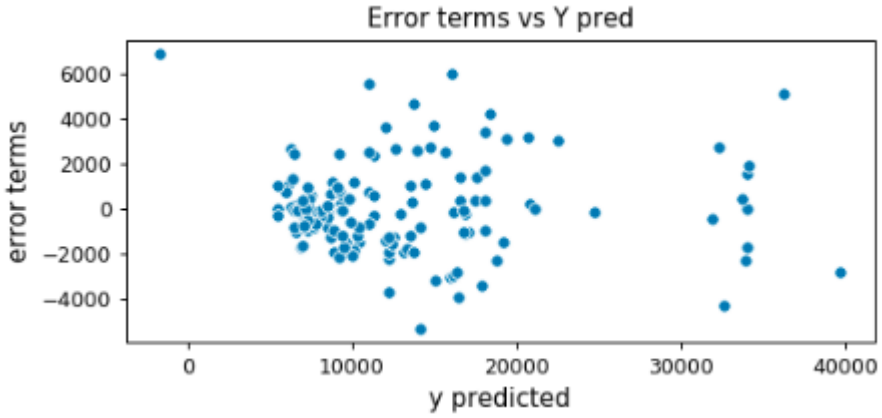
[source: own linear regression assignment]



[source: upgrad]

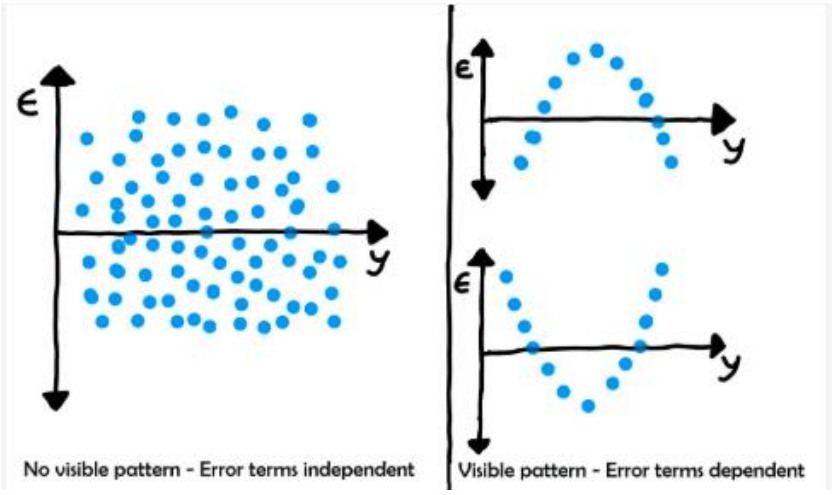
b. Error terms are independent of each other

i. To check plot a scatter plot of error terms, no pattern should be seen between the points



ii.

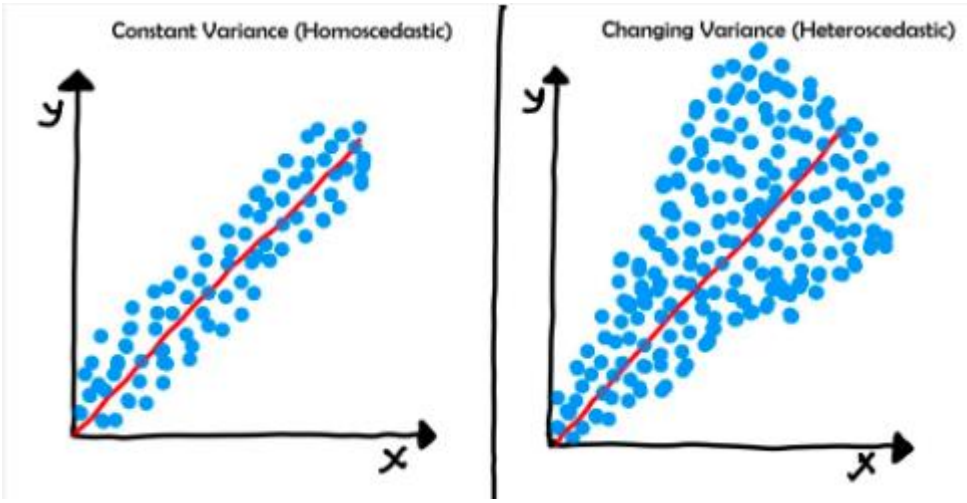
[source: own linear regression assignment]



[source: upgrad]

c. Error terms have constant variance

i. To check this intuitively, plot a scatter plot of error terms and see if the data points are not digressing towards the end or beginning.



[source: upgrad]

ii. To check this numerically take 2 points and calculate the standard deviation of error terms at each point. Both the values should be nearly equal.

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

Ans. Coefficient of correlation (R): Coefficient of correlation between two variables quantifies the strength and nature of relationship between two variables. Its value ranges from – 1 to 1. The sign of R tells about the proportionality of relationship between the variables. Positive sign means they are positively correlated and vice-versa. For example:

R= 0 represents that two variables are not related,

R**= -**1 represents a strong negative correlation between two variables meaning if one of the variables increases the other one is very likely to decrease,

R= 1 represents a strong positive correlation between two variables meaning if one of the variables increases the other one is very likely to increase,

|R |= 0.2 represents two variables are feebly related.

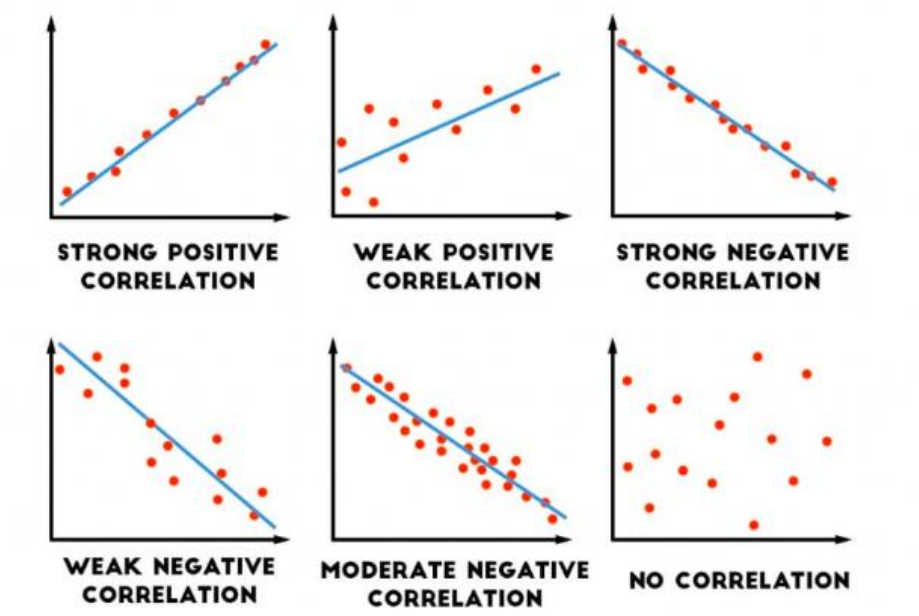


Figure showing points (y vs x) and the linear regression linear for one predictor model. [source: internet]

Coefficient of determination (R^2): It describes the fraction of variance of the outcome variable being explained by the predictor variables of the model. It ranges from zero to one, denoting the strength of linear relationship between y and predictors.

$$R^2 = \frac{explained\ variance}{total\ variance}$$

For example,

R = 0.91, then R^2=0.8281, which means 82.81% of the total variation in y could be explained by the model (linear regression line). It represents the percentage of data that is closest to the line of best fit. Hence it measures how well the line represents the data.

If the line passes through all the point in the scatter plot, it explains all the variance hence $R^2 = 1$.The further the line is away from the points, the less will be the value of R^2.

In the figure above the

coefficient of determination of subplot 1 (strong positive correlation) **>** coefficient of determination of subplot 2 (weak positive correlation).

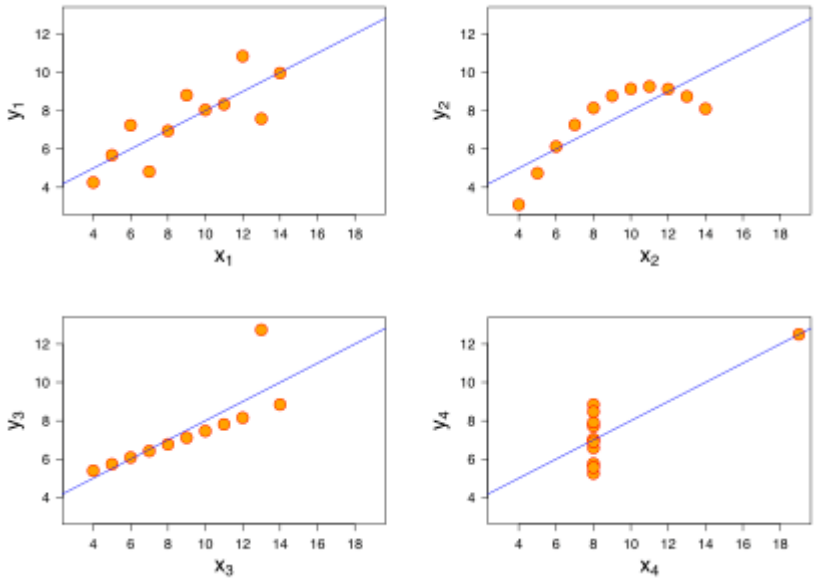
4. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is a classic example of how summary statistics can be misleading at times. A statistician Francis Anscombe demonstrated the value of visualizing the data before analysing it and the effect of outliers and some influencing data points on statistical summary.

With this statistical summary there can be more than one type of distribution,

Property	Value
Mean of x	9
Sample variance of x : σ^2	11
Mean of y	7.50
Sample variance of y : σ^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67

[Source: Wikipedia]



[Source: Wikipedia]

These were the four distributions demonstrated by Anscombe. We can observe that even though the first subplot follows a linear relationship and the second seems to follow a quadratic relationship, their mean and variance comes out to be same. The third and fourth subplots have completely different distributions as well even though their summary statistics are exactly same.

Hence summary statistics can identify a distribution individually. It is important to visualize the points (preferably on a scatter plot) to understand the distribution better.

5. What is Pearson's R?

Ans. Pearson's R is the Coefficient of correlation between two variables that quantifies the strength and nature of relationship between two variables. Its value ranges from – 1 to 1. The sign

of R tells about the proportionality of relationship between the variables. Positive sign means they are positively correlated and vice-versa. Formula for Pearson correlation for a sample is given as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{(Eq.3)}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

[Source: Wikipedia]

For example:

R= 0 represents that two variables are not related,

R= -1 represents a strong negative correlation between two variables meaning if one of the variables increases the other one is very likely to decrease,

R= 1 represents a strong positive correlation between two variables meaning if one of the variables increases the other one is very likely to increase,

|R |= 0.1 represents two variables are feebly related.

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

Ans. Scaling is a method to transform the independent features (having numerical values) such that they represent values on same scale.

Reasons to perform scaling on the predictors of a linear model:

1. Better comparison and interpretation of coefficients in the results

   For example, in model with predictors heights (in cm) and distance (in km), the larger coefficient for distance doesn't simply mean it is more important than height (we can't directly compare as they are not on same scale). For the model, a height of 125 (in cm) would seem greater than the distance of 2 (in km). Hence to unify the scales across all variables scaling is required. Then coefficients can be directly compared.

2. Faster Gradient descent
   The gradient descent algorithm tries to reach the local minima of cost function by adding /subtracting the gradients from the coefficients. If the coefficients are not at same scale, as can be seen in the above example, it will take more time and more steps to reach the minima.

Difference between normalized scaling and standardized scaling:

Normalized scaling: It rescales the data distribution of a feature between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardized scaling: It assumes that the data is normally distributed with respect to each feature. It scales the data of each feature to have a normal distribution centred around 0 and having a standard deviation of 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

Ans. VIF is calculated to check if a predictor variable can be explained by a combination of other predictor variables. VIF is calculated by the formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

If VIF is infinity, that implies denominator of this equation is zero or R_i^2 = 1.

This means that other predictor variables (or their combination) can explain 100% variance in this predictor variable.

8. What is the Gauss-Markov theorem?

Ans. According to the Gauss-Markov theorem, if a linear regression model fulfils the Gaus-Markov assumptions then the ordinary least squares (OLS) regression yields unbiased predictions that will have the least variance of all other linear estimators.

To rephrase if Gauss-Markov assumptions hold then OLS is BLUE.

BLUE means the Best linear unbiased estimator i.e OLS coefficient estimates follow the tightest possible sampling distribution of unbiased estimates compared to other linear estimation methods.

Gauss Markov assumptions (also called *conditions*):

1. Linear Parameters: the parameters used in the OLS methods must themselves be linear.
2. Random sample: the data set should be randomly sampled from the population
3. Non-Collinearity: the predictors used in the model should not be perfectly correlated with each other.
4. Exogeneity: the predictors are not correlated with the error terms.
5. Homoscedasticity: the error terms should have constant variance across the distributions of predictors.

9. Explain the gradient descent algorithm in detail.

Ans. Gradient Descent is an optimization algorithm used to minimize cost functions in machine learning. It is based on iteratively moving in the direction of steepest descent. Gradient descent is used to update the parameters of the model.

Every machine learning model has a cost function. Like in case of the following regression model:

$$y' = \beta_0 + \beta_1 x_1$$

The cost function with be the sum of least squares. Since the cost function is a function of Beta.

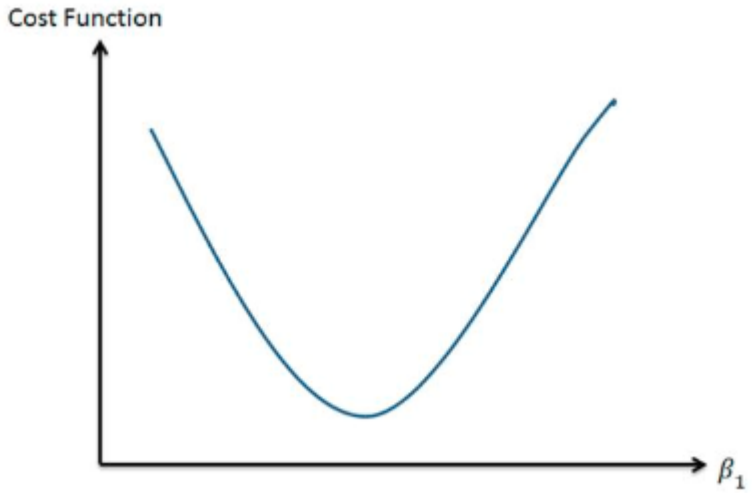
Error analysis:

Residual (error for point i) = $y'_i - y_{actual}$

Cost function = $\frac{1}{n} \sum_{i=1}^{n} y' - y_{actual}$

$$\text{Cost function} = \frac{1}{n} \sum_{i=1}^{n} \beta_0 + \beta_1 x_1 - y_{actual}$$

On integrating and solving the above equation we get quadratic equation in terms of β_1. On plotting the cost function with respect to β_1:



Depending on the number of parameters, this could be a contour of multi-dimensional figure (like a mountainous region in a 3 dimensions). Now, Imagine we leave an agent somewhere (random) on this multi-dimensional figure (which is called the initialization) and this agent can figure out the gradient of the point where it is. learning rate defines the length of step that the agent will take. So now when the agent gets the information about gradients and the length of step, if the agent iteratively moves towards the steepest descent then it will end up on the lowest point.

Mathematically,

Step 1: At point 1 calculate slope of the curve

Step 2: If the slop is negative, move forward by an amount of step length given. If the slope is positive, move backwards by an amount of step length given

Step 3: calculate if reach minima (cost is minimum, or first derivative is zero and second derivative is positive)

Step 4: Repeat Steps 1-3 till you reach either the minima or the maximum number of steps given.

Step 5: After reaching the minimum cost, the values of coefficients is your output.

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

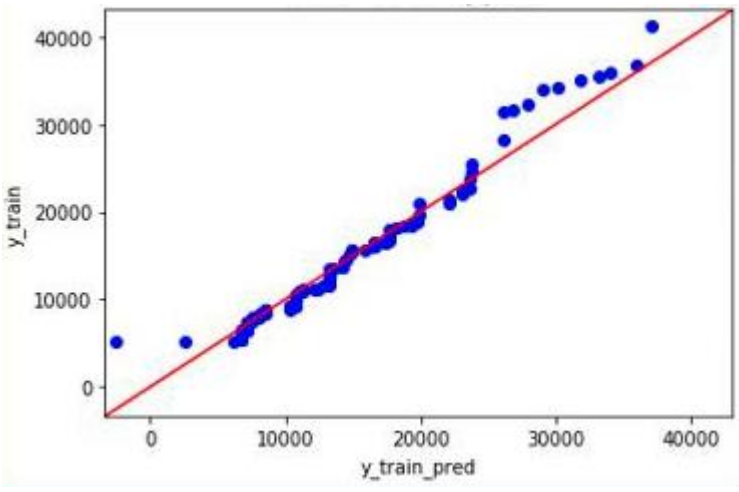
Ans. Q-Q plot or quantile-quantile plot is a probability plot made by plotting quantile values of two probability distributions against each other to compare them.

If the Q-Q plot looks like the line y = x, then distributions being compared are same.

If Q-Q plot is a line (but not y = x), then it indicates some linear relationship between the two distributions.

In regression, we are working on two datasets namely train and test under the assumption that they were sampled correctly, and they belong to the same population. We can use the QQ plot to know if they have same distribution or not. Which in turn can define if they belong to same population or not. If a QQ plot between test and train data follows y=x line, then it implies that they have same distribution and are very likely to be from the same population.

It can also be used to check if the predicted values follow the same distribution as actual values of the dependent variable in linear regression. For example, the plot given below,



[Source: internet]

Here the y_train (actual values of y – dependent variable) and y_train_pred (predicted values of y from model) almost follows a y = x line. Though towards the higher end, y_train is higher than the y_train_pred.

End of Assignment.