What is R-squared?

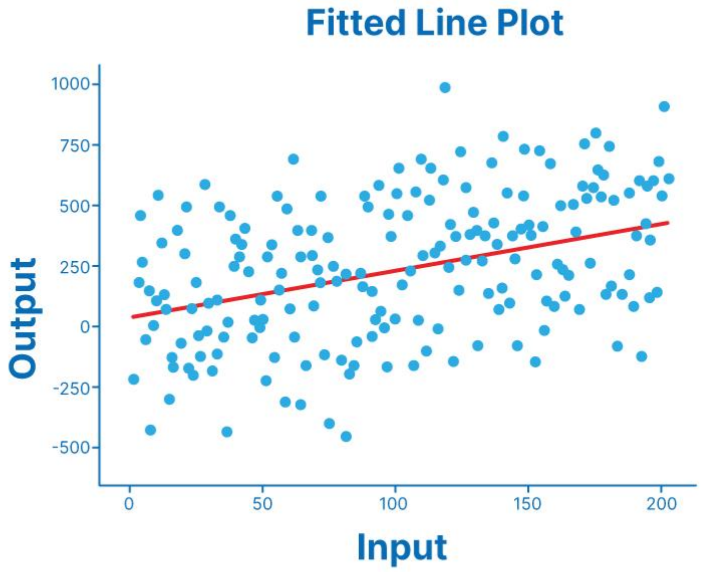
R squared or Coefficient of determination, or R² is a measure that provides information about the goodness of fit of the regression model. In simple terms, it is a statistical measure that tells how well the plotted regression line fits the actual data. R squared measures how much the variation is there in predicted and actual values in the regression model.

What is the significance of R squared

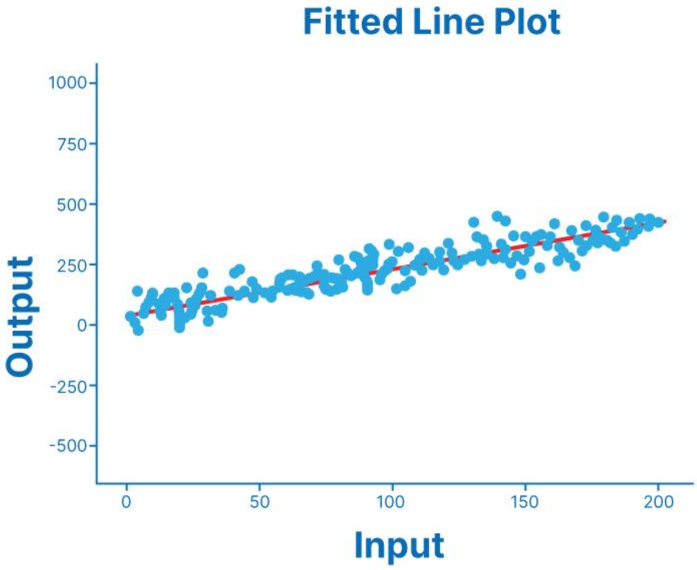
- R-squared values range from 0 to 1, usually expressed as a percentage from 0% to 100%.
- And this value of R square tells you how well the data fits the line you've drawn.
- The higher the model's R-Squared value, the better the regression line fits the data.

Note: R-squared values very close to 1 are likely overfitting of the model and should be avoided.

- So if the model value is close to 0, then the model is not a good fit
- A good model should have an R-squared greater than 0.8.
- So if the R squared value is close to 0, then you will get plot like this.



If the R squared value is close to 1, then you will get plot like this



In this the first image has the data points scattered and away from the regression line, and in this case, the prediction will not be accurate. So this clearly shows R squared value will be less than 0.5 or near 0. In the second image, the values are very close to the regression line, which means the prediction will be good. So the value of R squared will be close to 1.

Note- As the R squared value increases, the difference between actual and predicted values decreases.

When to Use R Squared

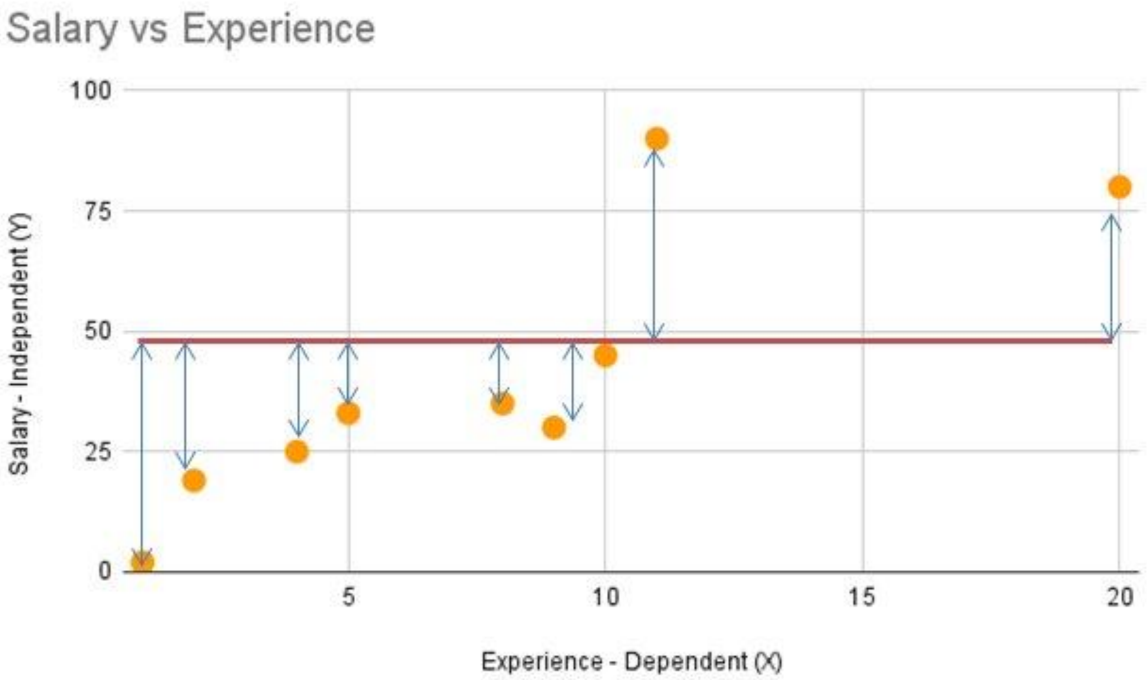
Linear regression is a powerful tool for predicting future events, and the r-squared statistic measures how accurate your predictions are. But ***when should you use r-squared?***

- Both independent and dependent variables must be continuous.
- When the independent and dependent variables have linear relationship (+ve or -ve) between them.

How to calculate R squared in linear regression?

Problem Statement:

Let’s take a look at an example. Say you’re trying to predict salary based on the number of years of experience. First, You will draw a mean line as shown in fig below.

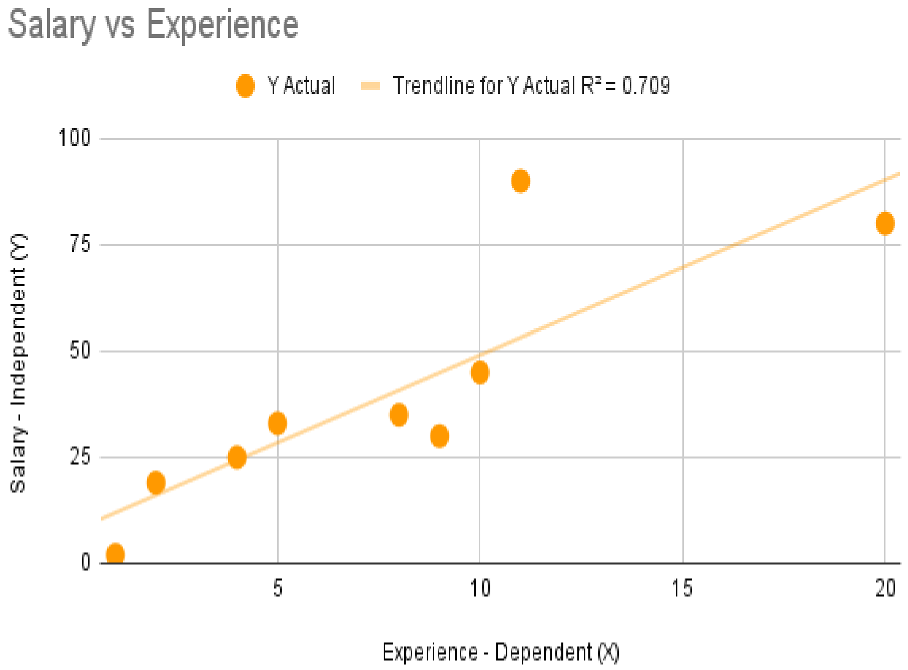


- Here we have actual values(orange points) and a horizontal/mean line. This line represents the mean of all values of the response variable in the regression model. It is represented on the graph as the average of actual variable values. Now calculate the distance of actual values(all) from the mean line.
- The deviation of an actual value from the mean is therefore calculated as the sum of the squared distances of the individual points from the mean. This is also called the ***total sum of squares*** (SST). SST is also called ***total error***. Mathematically, SST is expressed as:

$$\sum_{i=1}^{n}(yi\;\;-yi^{\;-}\;)^{2}$$

where yi represents the average(mean) and yi represents the actual value.

- Draw the best line that fits in as shown in fig below.



This straight line is called the **regression line or best fit line** and has the equation ***ŷ= mx +c***, where ***ŷ*** represents the predicted line, **c** is the intercept, and **m** is a slope. By using ***the least square method***, we can calculate the slope and intercept.

A **least square regression line** is a straight line that minimizes the vertical distance from the data points to the regression line.

- The total variation of the actual values from the regression line is expressed as the sum of the squared distances between the predicted values from the regression line, also known as the Residual sum of squared error ***(SSR)***. SSR is sometimes called explanatory error or explanatory variance. Mathematically, SSR is expressed as yi represents the predicted value and yi represents the actual value.

$$\sum_{i=1}^{n} (yi - \hat{yi})^2$$

The formula to calculate R squared is:

$$\text{R-Squared} = 1-\left(\frac{SSR}{SST}\right)=1- \frac{\sum_{i=1}^{n} (yi - \hat{yi})^2}{\sum_{i=1}^{n} (yi - \bar{yi})^2}$$

where:

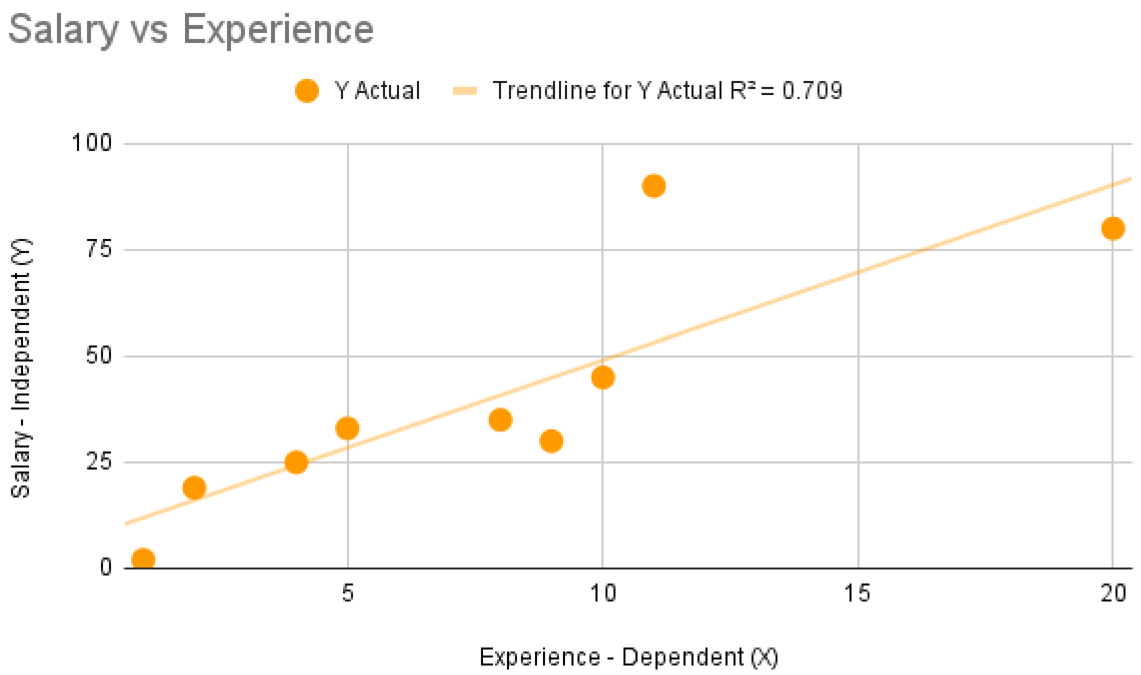
SSR is the sum of squared residuals (i.e., the sum of squared errors)

SST is the total sum of squares (i.e., the sum of squared deviations from the mean)

Example of R squared

Suppose we have a data set having values of X representing experience and Y representing salary.

The below data can be represented by this graph in which we have a regression line.



x_i	y_i	xi- x_i^{-}	**y_i-y_i^{-}**	**(xi- x_i^{-})^2**	**(xi- x_i^{-}) (y_i-y_i^{-})**
11	90	3.22	50.11	10.38	161.47
10	45	2.22	5.11	4.94	11.36
2	19	-5.78	-20.89	33.38	120.69
8	35	0.22	-4.89	0.05	-1.09
4	25	-3.78	-14.89	14.27	56.25
20	80	12.22	40.11	149.38	490.25
1	2	-6.78	-37.89	45.94	256.80

9	30	1.22	-9.89	1.49	-12.09
5	33	-2.78	-6.89	7.72	19.14
x_i^{-} =**7.78**	**y_i^{-} =39.89**			**267.56**	**1102.78**

Suppose we have a data set having values of X and Y.

1. We have to find Xi(mean) and Yi(mean).

2. Calculate Xi-Xi and Yi-Yi and then do (Xi-Xi)2

3. Now calculate **(Xi-Xi)(Yi-Yi)**

Now we have to calculate R squared so let’s try to calculate it

$y_i^\wedge$	**y_i – $y_i^\wedge$**	**SSR**	**y_i-y_i^{-}**	**SST**
53.17	36.83	1356.46	50.11	2511.12
49.05	-4.05	16.39	5.11	26.12
16.07	2.93	8.56	-20.89	436.35
37.10	-2.10	4.39	-4.89	23.90
20.61	4.39	19.29	-14.89	221.68
86.56	-6.56	42.97	40.11	1608.90
8.24	-6.24	38.98	-37.89	1435.57
41.22	-11.22	125.82	-9.89	97.79
24.73	8.27	68.39	-6.89	47.46
		1681.24		**6408.89**

4. Now first we have to calculate $y_i^\wedge$ i.e. Predicted value. This column includes predicted values. This predicted value is calculated by using formula ***y^=mx+c***, where m is **slope** and c is **intercept**.So now we have to calculate m,which can be calculated by using least squares formula.

5. Find y_i-$y_i^\wedge$. This value tells us how much the predicted values are away from actual values.

5. Now calculate **SSR** by using the formula

$$\sum_{i=1}^{n} (yi \ -\hat{yi} \ )^2$$

Where yi represents the predicted value or Y predicted and yi represents the actual value.

Adn SST by using the formula

$$\sum_{i=1}^{n} (yi \ -\bar{yi} \ )^2$$

where yi^{-} represents the average(mean) and yi represents the actual value.
6. Now simply put t it the formula

R-Squared = 1-(SSR/SST). You will get R squared value.We get **R square= 0.74**,Which shows that the prediction values are somehow close to the actual values
After doing calculation we will get these values

Slope	3.068922306
Intercept	13.46783626