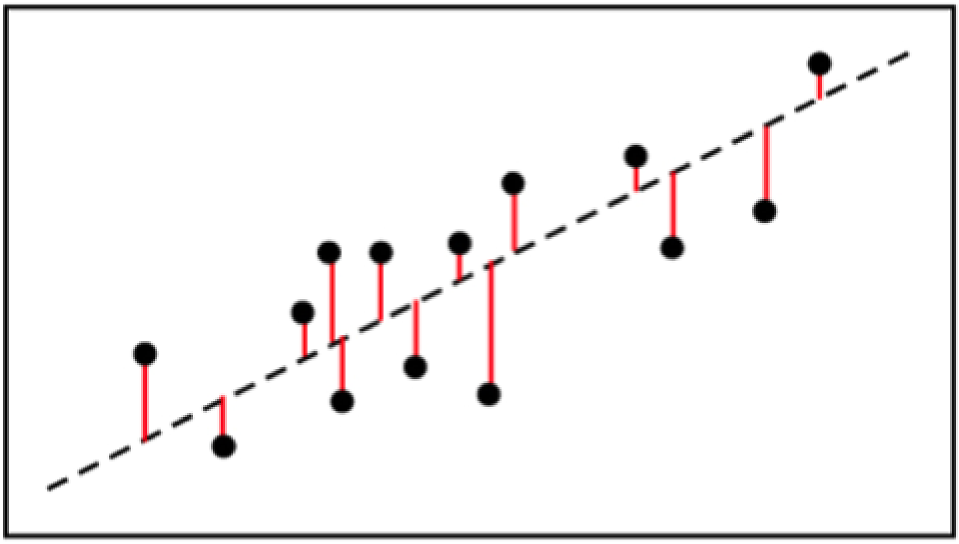
R-squared

R Squared

R-squared, which sometimes is also known as the **coefficient of determination**, defines the degree to which the variance in the dependent variable (target or response) can be explained by the independent variable (features or predictors).

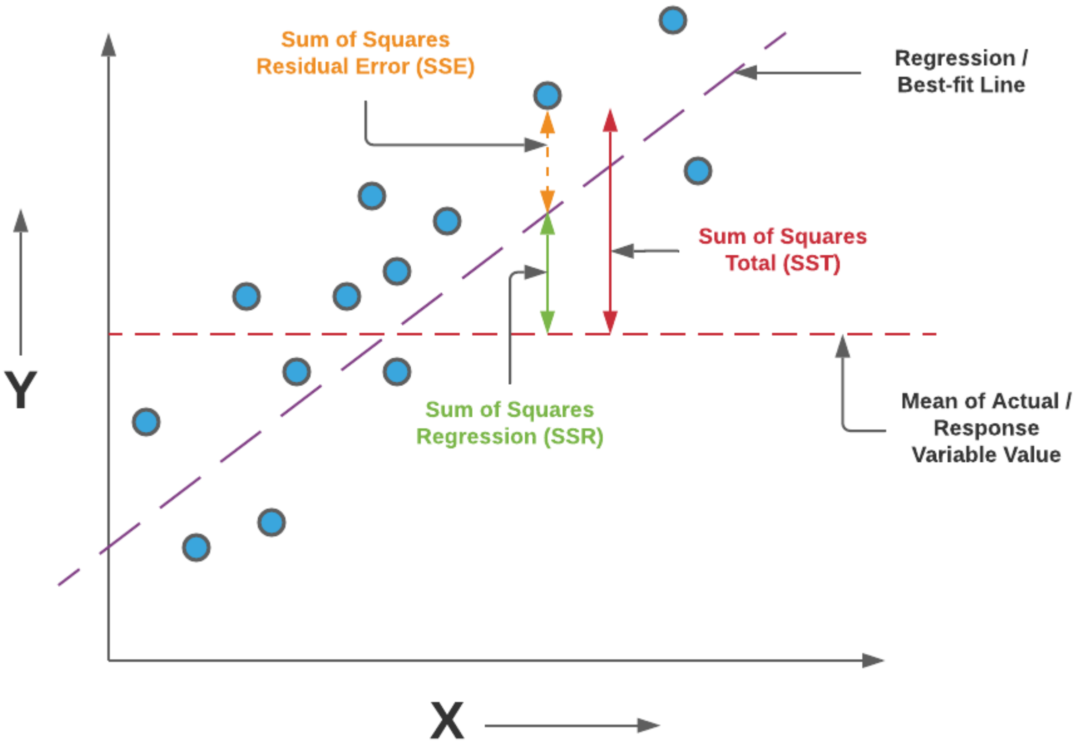
Normally we to find the ordinary least squares, we would take all the red line distances and square them and then sum them, once we have sum of the squares we need to find the minimum among the sum of all the other lines we have made. Now this sum is also called **Sum of Residual Squares.**

$$\sum_i (y_i - \hat{y}_i)^2$$



Y is the point and Y^ is the point on the line.

Total sum of squares is when we calculate the distances from the average line**(in red)** to the point in the graph.



Now to find out R-squared, we need to use this formula,

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}}$$

The R2 is calculated by dividing the sum of squares of residuals from the regression model (given by SSres) by the total sum of squares of errors from the average model (given by SStot) and then subtracting it from 1.

- **Thus we can say that the higher the R-squared value, meaning the more close the R-squared value to 1, the better is our model. If the R-squared values is far from 1, then out model is bad.**

The average line is also a trend line, it can be used to fit to dataset, but this wont give us the best model. But with the trend line generated from least value of sum of residuals will give us the best model. **R-squared will tell us how good is our SSres in comparison to SStot.**

R-squared can be negative, when our SSres fits our dataset in the worst way possible.

Adjusted R-squared

We spoke of R-squared for a simple linear regression, the same thing can be done for multiple linear regression.

R-squared is a goodness of fit parameter, but the problem is when we add more variables to our model.

For example, salary could be based on many variables, like experience, business brought in the year, etc.

If we add another variable, we find the R-squared again, the general observation is R-squared will never decrease, because R-squared is 1- **SSres/SStot**, here the **SSres** will be reduced because of addition of the new variable, hence reducing the **SSres/SStot** value, and the overall value will increase.

Regardless of the addition of variables, because of this problem we will never know if it is helping our model or not.

Here's where adjusted R-squared comes in:

$$Adjusted\ R^2 = \{1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]\}$$

- **n** represents the number of data points in our dataset
- **k** represents the number of independent variables, and
- **R** represents the R-squared values determined by the model.

▼ When k increases, the denominator decreases, which consequently increases overall ratio, the increased ratio value multiplied with (1-R-squared) will become big. 1- this increased value will result is smaller Adjusted R-squared value.

> 💡 Adjusted R-squared penalizes the model if the added variables are of no help to the model.

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine

whether adding new variables to the model actually increases the model fit.

This is a fair way to judge a good ness of the model.