Introduction to Linear Regression

Linear regression is a machine learning algorithm used to predict the value of continuous response variables.

The predictive analytics problems that are solved using linear regression models are called supervised learning problems as it requires that the value of response/target variables must be present and used for training the models.
"continuous" represents the fact that the response variable is numerical in nature and can take infinite different values. Linear regression models belong to a class of **parametric models.**
Linear regression models work great for data that are linear in nature. In other words, the predictor / independent variables in the data set have a linear relationship with the target/response / dependent variable. The following represents the linear relationship between response and the predictor variable in a simple linear regression model.

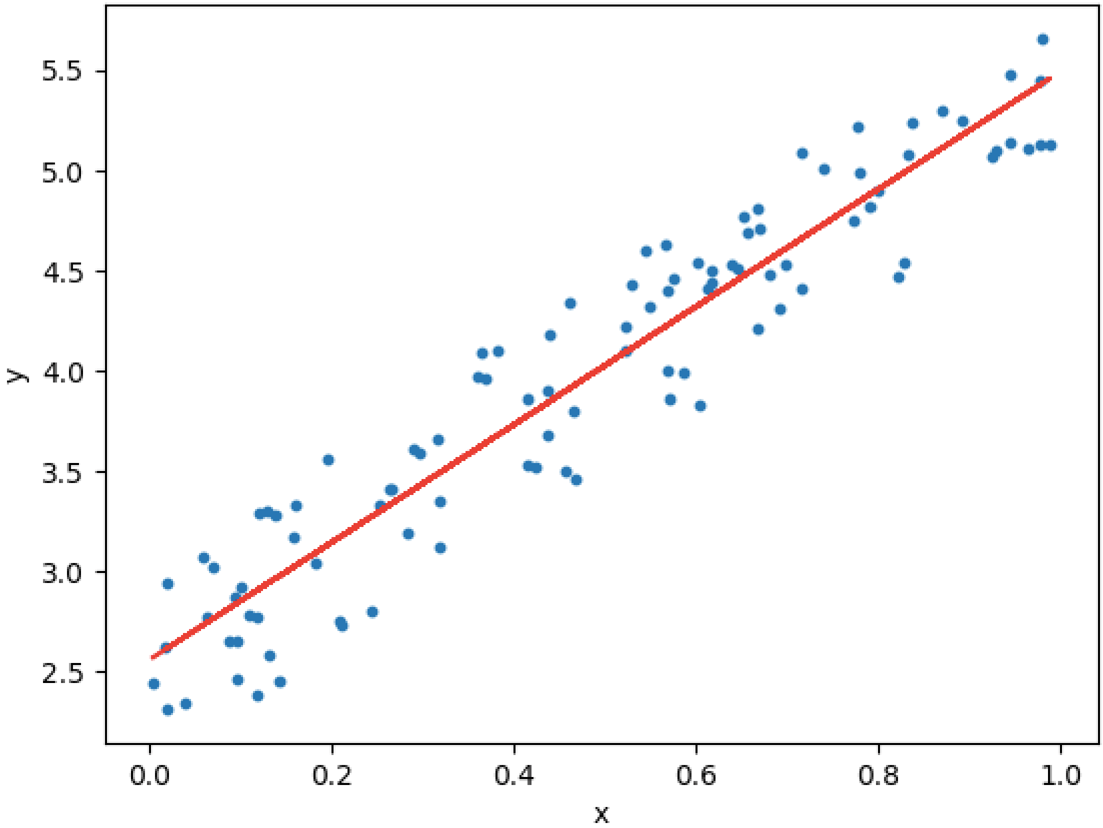


Fig 1. Simple linear regression model
The red line in the above diagram is termed as **best-fit line** and can be found by training the model such as **Y = mX + c**
Linear regression models is of two different kinds. They are simple linear regression and multiple linear regression.

- **Simple linear regression**: When there is just one independent or predictor variable such as that in this case, Y = mX + c, the linear regression is termed as simple linear regression.
- **Multiple linear regression**: When there are more than one independent or predictor variables such as

$$Y = w_1x_1 + w_2x_2 + \ldots + w_nx_n,$$

the linear regression is called as multiple linear regression.

Linear Regression Concepts / Terminologies

- **Residual Error**: Residual error is the difference between the actual value and the predicted value. When visualizing in terms of best fit line, if the actual value is above the best-fit line, it is called the positive residual error and if the actual value is below the best fit line, it is called the **negative residual error**. The figure below represents the same.

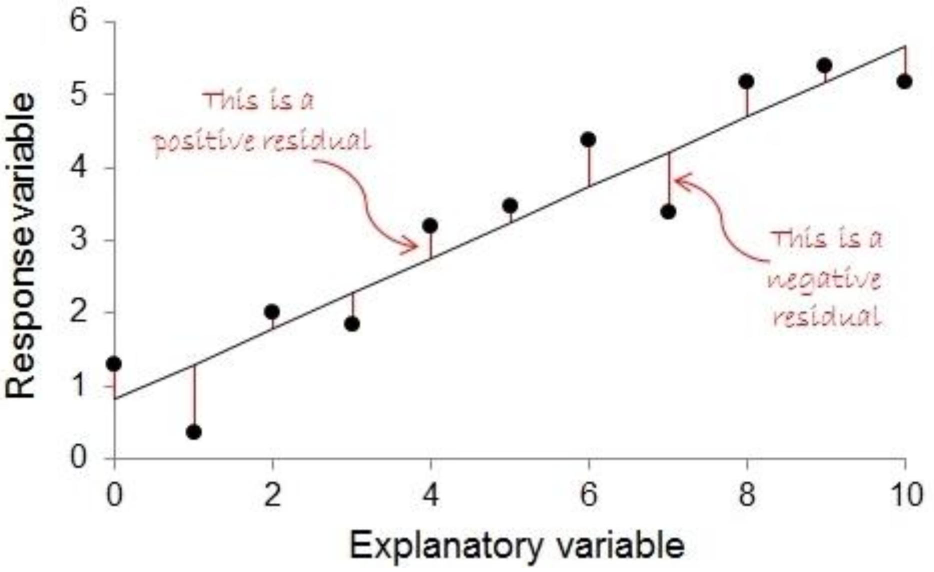


Fig 2.

SST, SSE, SSR: The following are key concepts when dealing with the linear regression model. The following diagram is the representation of SST, SSE, and SSR

- **Sum of Square Total (SST)**: Sum of Squares Total is equal to the sum of the squared difference between actual values related to the response variable and the mean of actual values. It is also called the variance of the response. Recall how you calculate variance – the sum of the squared difference between observations and the mean of all observations. It is also termed as Total Sum of Squares (TSS).
- **Sum of Square Error (SSE):** Sum of Square Error or Sum of Square Residual Error is the sum of the squared difference between the actual value and the predicted value. It is also termed as Residual Sum of Squares.
- **Sum of Square Regression (SSR)**: Sum of Square Regression is the sum of the squared difference between the predicted value and the mean of actual values. It is also termed as **Explained Sum of Squares (ESS)**

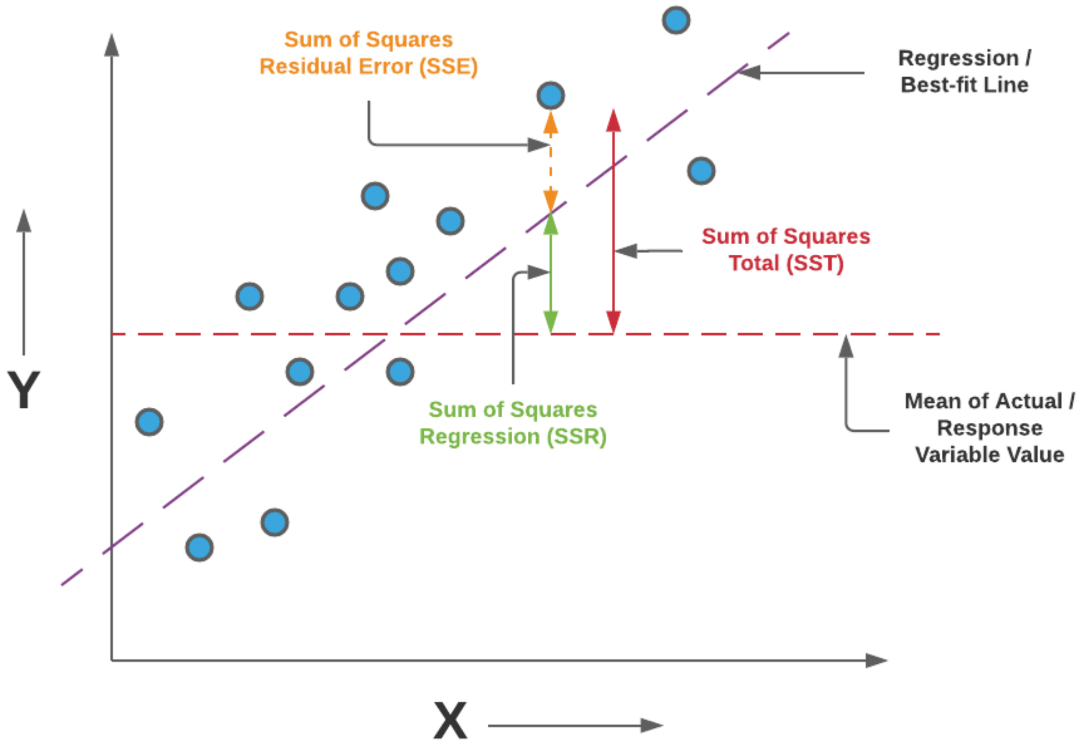


Fig 3. SSR, SSE and SST Representation in relation to Linear Regression

- **How are SST, SSR, and SSE related?**

Here is how SST, SSR, and SSE are related. The same could be comprehended using the diagram in fig 3.

SST = SSR + SSE

R-Squared: R-squared is a measure of how good is the regression or best fit line.
It is also termed as the **coefficient of determination.** Mathematically, it is represented as the ratio of Sum of Squares Regression (SSR) and Sum of Squares Total (SST).
R-Squared = SSR / SST = (SST – SSE) / SST = 1 – (SSE / SST)

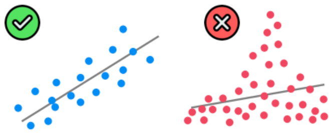
The greater the value of R-Squared, the better the regression line as higher is the variation explained by the regression line.

The value of R-squared is a statistical measure of goodness of fit for a linear regression model. Alternatively, R-squared represents how close the prediction is to the actual value.

Assumptions of Linear Regression

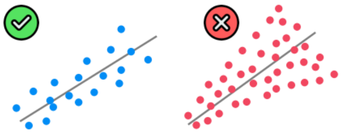
1. Linearity

(Linear relationship between Y and each X)



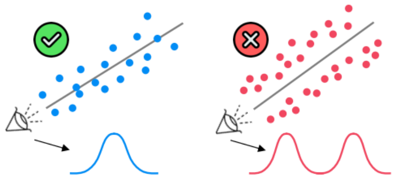
2. Homoscedasticity

(Equal variance)



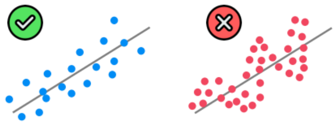
3. Multivariate Normality

(Normality of error distribution)



4. Independence

(of observations. Includes “no autocorrelation”)



5. Lack of Multicollinearity

(Predictors are not correlated with each other)

✓ $X_1 \nsim X_2$ ✗ $X_1 \sim X_2$

6. The Outlier Check

(This is not an assumption, but an “extra”)