**LINEAR REGRESSION:**

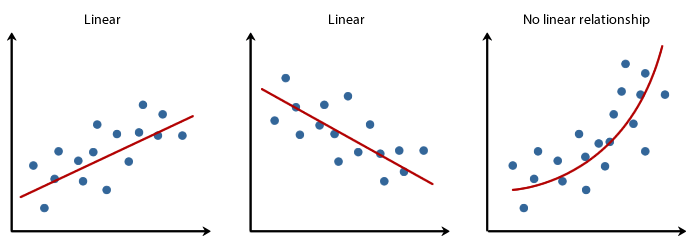
**DEFN:**

Linear regression is the supervised machine learning model in which model finds best fit line between independent and dependent variable.

OR

**DEFN:**

Linear regression attempts to model the relationship between two variables by fitting a linear equation (a straight line to the observed data).



There are two types linear regression,

1. Simple Linear Regression.
2. Multi-Linear Regression.

**SIMPLE LINEAR REGRESSION:**

A statistical method in which only one independent variable is present and the model has to find the linear relationship of it with the dependent variable.

* One variable, denoted x, is regarded as the explanatory or independent variable.
* One variable, denoted y, is regarded as the response, predictor, outcome or dependent variable.

General EQN,

Consider y as your income and as your education, the more education you have the higher income you will get.

works as multiplier and quantifies the income, while is constant value (consider it as minimum wage), so if you have 0 education you will get a minimum wage.

is the error (on average error is 0).

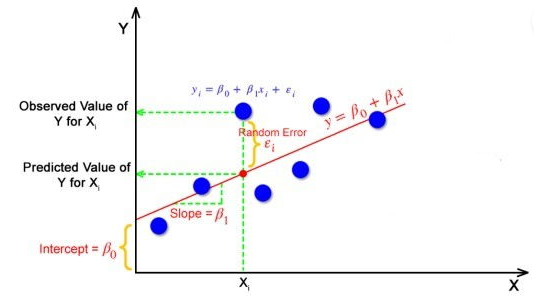
Linear Regression EQN,

**:** Estimated predicted value.

**:** Intercept / constant.

**:** Slope / quantifier.

**:** Sample data for independent variable.



**MULTI-LINEAR REGRESSION:**

A statistical method can be used to analyze the relationship between single dependent variable and multiple independent variable.

* Multiple variable, denoted is regarded as the explanatory or independent variables.
* One variable, denoted y, is regarded as the response, predictor, outcome or dependent variable.

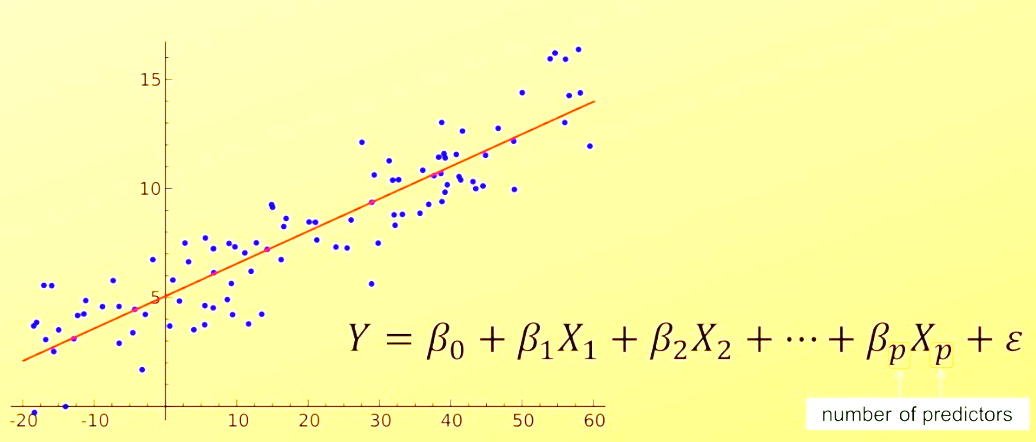
Multi-Linear Regression EQN,

**:** Estimated predicted value.

**:** Intercept / constant.

**:** Slope / quantifier.

**:** Sample data for independent variable.



**ASSUMPTIONS OF LINEAR REGRESSION:**

Regression is a parametric approach, which means that it makes assumptions about the data for the purpose of analysis. For successful regression analysis, it’s essential to validate the following assumptions.

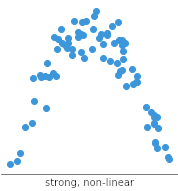
There are following five assumptions of linear regression,

1. Linearity.
2. No Endogeneity.
3. Normality and Homoscedasticity.
4. No Auto-Correlation.
5. No Multi-collinearity.

**LINEARITY:**

Linear regression needs the relationship between the independent and dependent variables to be linear.

The linearity assumption can best be tested with scatter plots, the following two examples depict two cases, where no linearity and linearity is present.

**Linear
**

Non-Linear Relationship

Linear Relationship

If no linearity is observed, transform the data (using log transformation, exponential transformation, reciprocal transformation, square root transformation, boxcox transformation etc.).

**NO ENDOGENEITY**

**Endogeneity**refers to situations in which a predictor (e.g. ) in a linear regression model is **correlated to the error term**.

https://towardsdatascience.com/causal-inference-with-linear-regression-endogeneity-9d9492663bac

**NO MULTI-COLLINEARITY**

Multicollinearity is observed when two or more independent variables are correlated to one another.

If we found any variable causing collinearity problem we must drop it.

Consider a kid who loves watching tv and eating snacks. Now, the kid is watching tv while eating snacks, can you tell what activity (either watching tv or eating snacks) does the kid love more? No, right? because he loves both therefore we can’t decide as watching tv and eating snacks are correlated, in order to decide we have to drop one variable so that we can decide.

The best method to test for the assumption is the Variance Inflation Factor method.

Here’s link where you can check how can we exclude variables sharing high correlation.

[How to Calculate VIF in Python - Statology](https://www.statology.org/how-to-calculate-vif-in-python/)

**NOTE:**

A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.  
Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

**Queries:**

How do we know which line is best fit line?

For starters you can try dropping multiple lines on a graph and calculating distance of each line with different data points called margin of error ().

In the end a line having the minimum margin of error will be chosen as a best fit line.