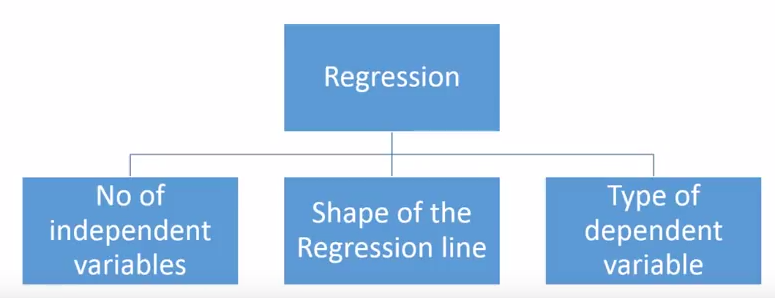
**[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Regression_Type.png)How many types of regression techniques do we have?**

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). We’ll discuss them in detail in the following sections.

For the creative ones, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven’t used before. But before you start that, let us understand the most commonly used regressions:

**1. Linear Regression**

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be [continuous or discrete](https://en.wikipedia.org/wiki/Continuous_and_discrete_variables), and nature of regression line is linear.

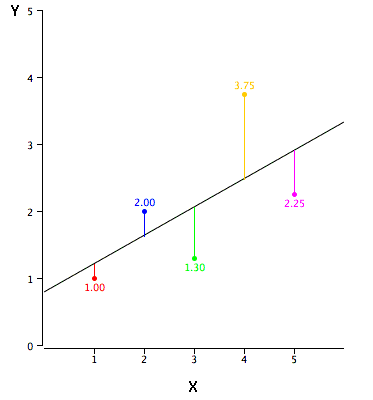
Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b\*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Linear_Regression1.png)

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.  Now, the question is “How do we obtain best fit line?”.

**How to obtain best fit line (Value of a and b)?**

[least square, regression line](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Least_Square.png)[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/reg_error.gif)This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

We can evaluate the model performance using the metric **R-square**.

**Important Points:**

* There must be **linear relationship** between independent and dependent variables
* Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
* Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
* Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
* In case of multiple independent variables, we can go with **forward selection**, **backward elimination** and **step wise approach** for selection of most significant independent variables.

**2. Logistic Regression**

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

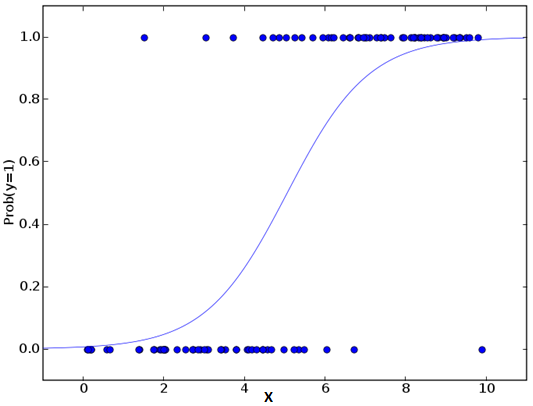
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation?”.

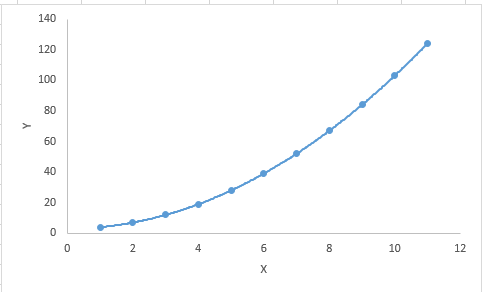
Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is [**logit**](https://en.wikipedia.org/wiki/Logistic_function) function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Logistic_Regression.png)**Important Points:**

* It is widely used for **classification problems**
* Logistic regression doesn’t require linear relationship between dependent and independent variables.  It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
* To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
* It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
* The independent variables should not be correlated with each other i.e. **no multi collinearity**.  However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
* If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**
* If dependent variable is multi class then it is known as **Multinomial Logistic regression**.

**3. Polynomial Regression**

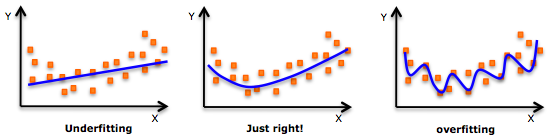
A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Polynomial.png)y=a+b\*x^2

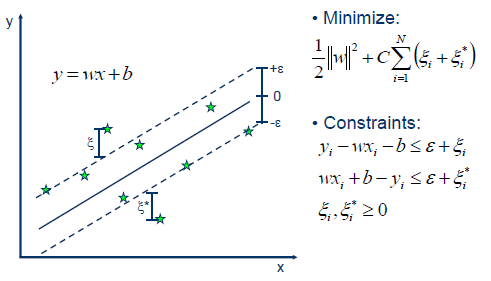
In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

**Important Points:**

* While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help:
* Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing wierd results on extrapolation.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/02/underfitting-overfitting.png)

### SVR_1.png4. Support Vector Machine - Regression (SVM)

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

# 5. Multiple Regression

**Multiple regression** involves a single dependent variable and two or more independent variables. It is a statistical technique that simultaneously develops a mathematical relationship between two or more independent variables and an interval scaled dependent variable.

The general form given for the multiple regression model is:

**Y= ß0 + ß1X1+ ß2X2+ …….. + ßkXk+ e.**

This multiple regression model is estimated using the following equation:

**= a + b1X1+ b2X2+ …….. + bkXk.**

There are certain statistics that are used while conducting the analysis.

* The R2 is the coefficient of the multiple determination. This coefficient measures the strength of association.
* The F test in multiple regression is used to test the null hypothesis that the coefficient of the multiple determination in the population is equal to zero.
* The partial F test is used to test the significance of a partial regression coefficient. This incremental F statistic in multiple regression is based on the increment in the explained sum of squares that results from the addition of the independent variable to the regression equation after all the independent variables have been included.
* The partial regression coefficient in multiple regression is denoted by b1. This denotes the change in the predicted value per unit change in X1,when the other independent variables are held constant.

In SPSS, multiple regression is conducted by the researcher by selecting “regression” from the “analyze menu.” From regression, the researcher selects the “linear” option. When the linear regression dialogue box appears, then theresearcher enters one numeric dependent variable and two or more independent variables and then finally he will carry out multiple regression in SPSS.

The following assumptions are made in multiple regression statistical analysis:

1. The first assumption involves the proper specification of the model. This assumption is important in multiple regression because if the relevant variables are omitted from the model, then the common variance which they share with variables that are included in the mode is then wrongly characterized with respect to those variables, and hence the error term is inflated.
2. The second assumption is that the residual errors are normally distributed. In other words, the residual errors in multiple regression should follow the normal population having zero as mean and a variance as one.
3. The third assumption is that of unbounded data. The regression line produced by OLS (ordinary least squares) in multiple regression can be extrapolated in both directions, but is meaningful only within the upper and lower natural bounds of the dependent.