Regression

Regression techniques are one of the most popular statistical techniques used for predictive modeling and data mining tasks. On average, analytics professionals know only 2-3 types of regression which are commonly used in real world. They are linear and logistic regression. But the fact is there are more than 10 types of regression algorithms designed for various types of analysis. Each type has its own significance. Every analyst must know which form of regression to use depending on type of data and distribution.

What is Regression Analysis?

Lets take a simple example : Suppose your manager asked you to predict annual sales. There can be a hundred of factors (drivers) that affects sales. In this case, sales is your dependent variable. Factors affecting sales are independent variables. Regression analysis would help you to solve this problem

In simple words, regression analysis is used to model the relationship between a dependent variable and one or more independent variables.

Terminologies related to regression analysis

1. Outliers
Suppose there is an observation in the dataset which is having a very high or very low value as compared to the other observations in the data, i.e. it does not belong to the population, such an observation is called an outlier. In simple words, it is extreme value. An outlier is a problem because many times it hampers the results we get.

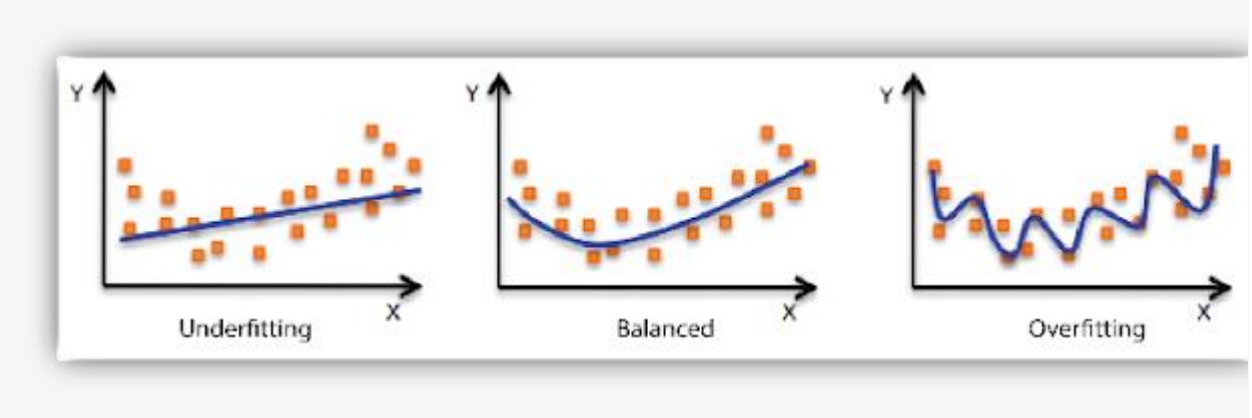
2. Multicollinearity
When the independent variables are highly correlated to each other then the variables are said to be multicollinear. Many types of regression techniques assumes multicollinearity should not be present in the dataset. It is because it causes problems in ranking variables based on its importance. Or it makes job difficult in selecting the most important independent variable (factor).

3. Heteroscedasticity
When dependent variable's variability is not equal across values of an independent variable, it is called heteroscedasticity. **Example -**As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times eat expensive meals. Those with higher incomes display a greater variability of food consumption.

4. Underfitting and Overfitting When we use unnecessary explanatory variables it might lead to overfitting. Overfitting means that our algorithm works well on the training set but is unable to perform better on the test sets. It is also known as problem of **high variance.**

When our algorithm works so poorly that it is unable to fit even training set well then it is said to **underfit the data.**It is also known as **problem of high bias.**

In the following diagram we can see that fitting a linear regression (straight line in fig 1) would underfit the data i.e. it will lead to large errors even in the training set. Using a polynomial fit in fig 2 is balanced i.e. such a fit can work on the training and test sets well, while in fig 3 the fit will lead to low errors in training set but it will not work well on the test set.

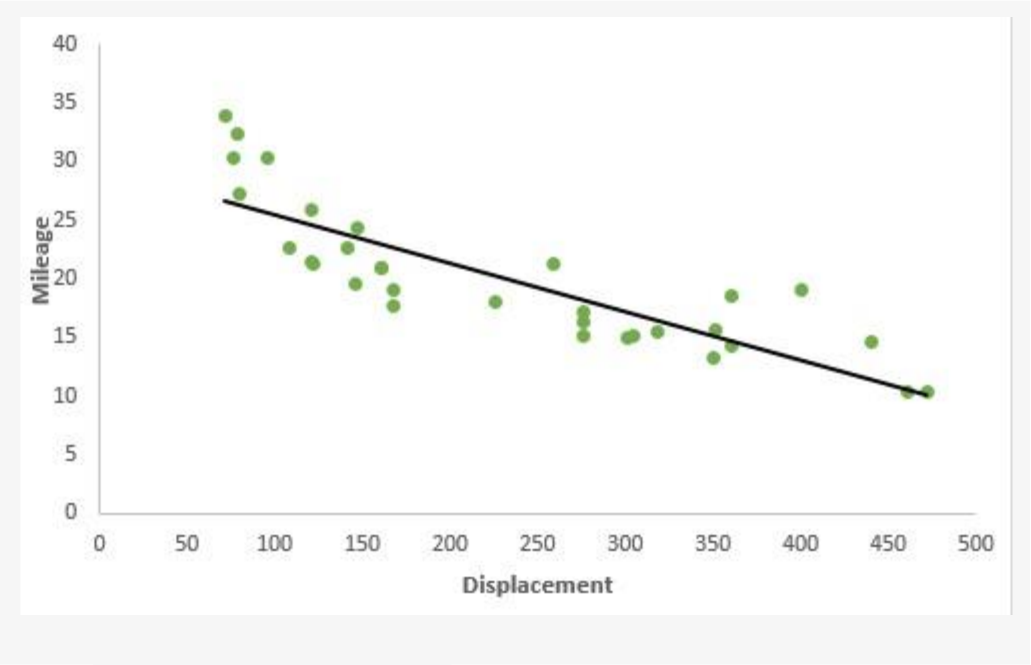


Types of Regression

Every regression technique has some assumptions attached to it which we need to meet before running analysis. These techniques differ in terms of type of dependent and independent variables and distribution.

1. Linear Regression

It is the simplest form of regression. It is a technique in which the dependent variable is continuous in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature. We can observe that the given plot represents a somehow linear relationship between the mileage and displacement of cars. The green points are the actual observations while the black line fitted is the line of regression



When you have only 1 independent variable and 1 dependent variable, it is called simple linear regression.
When you have more than 1 independent variable and 1 dependent variable, it is called Multiple linear regression.

The equation of multiple linear regression is listed below -

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$$

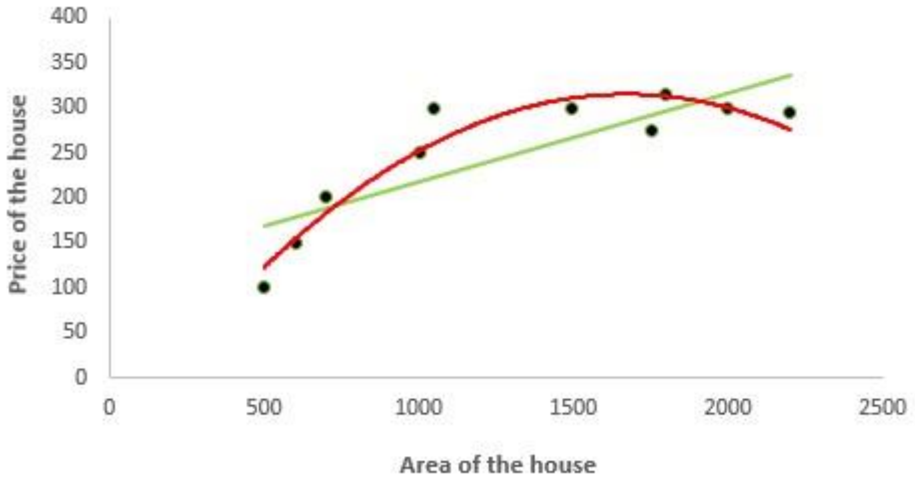
Assumptions of linear regression:

1. There must be a linear relation between independent and dependent variables.
2. There should not be any outliers present.
3. No heteroscedasticity
4. Sample observations should be independent.
5. Error terms should be normally distributed with mean 0 and constant variance.
6. Absence of multicollinearity and auto-correlation.

2. Polynomial Regression

It is a technique to fit a nonlinear equation by taking polynomial functions of independent variable.

In the figure given below, you can see the red curve fits the data better than the green curve. Hence in the situations where the relation between the dependent and independent variable seems to be non-linear we can deploy **Polynomial Regression Models.**



Thus a polynomial of degree k in one variable is written as:

$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon$$

Here we can create new features like

$$X_1 = x, X_2 = x^2, \ldots, X_k = x^k$$

and can fit linear regression in the similar manner.

In case of multiple variables say X1 and X2, we can create a third new feature (say X3) which is the product of X1 and X2 i.e.

$$X_3 = X_1 * X_2$$

Disclaimer: It is to be kept in mind that creating unnecessary extra features or fitting polynomials of higher degree may lead to overfitting.

3. Logistic Regression

In logistic regression, the dependent variable is binary in nature (having two categories). Independent variables can be continuous or binary. In multinomial logistic regression, you can have more than two categories in your dependent variable.
Here my model is:

$$p = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+---+b_kX_k)}}$$

logistic regression equation

Why don't we use linear regression in this case?

- Homoscedasticity assumption is violated.
- Errors are not normally distributed
- y follows binomial distribution and hence is not normal.

Examples

- **HR Analytics :** IT firms recruit large number of people, but one of the problems they encounter is after accepting the job offer many candidates do not join. So, this results in cost over-runs because they have to repeat the entire process again. Now when you get an application, can you actually predict whether that applicant is likely to join the organization (Binary Outcome - Join / Not Join).

- **Elections :** Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (response) variable is binary (0/1); win or lose. The predictor variables of interest are the amount of money spent on the campaign and the amount of time spent campaigning negatively.
  - Predicting the category of dependent variable for a given vector X of independent variables.Through logistic regression we have -
  - $P(Y=1) = exp(a+ B\square X)/ (1+ exp(a+ B\square X))$
  -

    Thus we choose a cut-off of probability say 'p' and if P(Yi = 1) > p then we can say that Yi belongs to class 1 otherwise 0.

Interpreting the logistic regression coefficients (Concept of Odds Ratio)

- If we take exponential of coefficients, then we’ll get odds ratio for ith explanatory variable. Suppose odds ratio is equal to two, then the odds of event is 2 times greater than the odds of non-event. Suppose dependent variable is customer attrition (whether customer will close relationship with the company) and independent variable is citizenship status (National / Expat). The odds of expat attrite is 3 times greater than the odds of a national attrite.

Stepwise Selection Stepwise regression is a combination of the forward and backward selection techniques. It was very popular at one time, but the Multivariate Variable Selection procedure described in a later chapter will always do at least as well and usually better. Stepwise regression is a modification of the forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a nonsignificant variable is found, it is removed from the model. Stepwise regression requires two significance levels: one for adding variables and one for removing variables. The cutoff probability for adding variables should be less than the cutoff probability for removing variables so that the procedure does not get into an infinite loop.