What is Regression Analysis?

Let's 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.

It helps us to answer the following questions -

1. Which of the drivers have a significant impact on sales.
2. Which is the most important driver of sales
3. How do the drivers interact with each other
4. What would be the annual sales next year.

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.

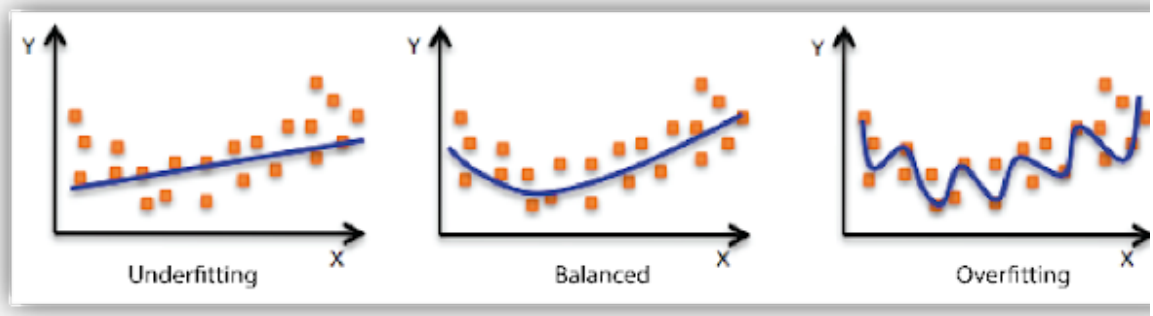
2. Multicollinearity
When the independent variables are highly correlated to each other then the variables are said to be multicollinear. Many types of regression techniques assumes multicollinearity should not be present in the dataset. It is because it causes problems in ranking variables based on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

3. 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

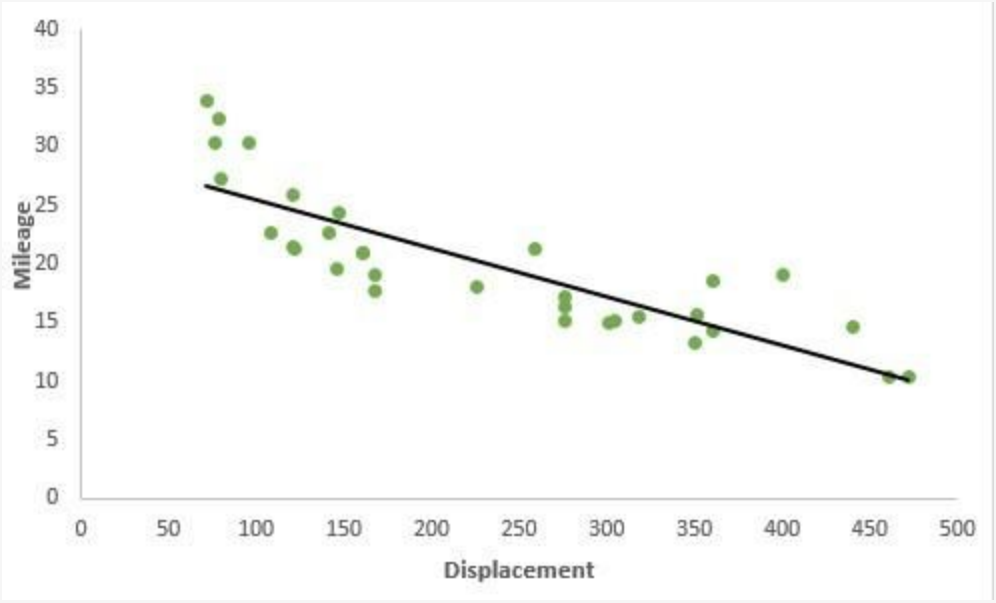
fig 2 is balanced i.e. such a fit can work on the training and test sets well, while in fig 3 the fit will lead to low errors in training set but it will not work well on the test set.



Regression : Underfitting and Overfitting

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 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



Regression Analysis

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 -

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

Here 'y' is the dependent variable to be estimated, and X are the independent variables and ε is the error term. βi's are the regression coefficients.

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.

Estimating the parameters
To estimate the regression coefficients βi's we use principle of least squares which is to minimize the sum of squares due to the error terms i.e.

$$\text{Min} \sum \varepsilon^2 = \text{Min} \sum (y - (\beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k))^2$$

On solving the above equation mathematically we obtain the regression coefficients as:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Interpretation of regression coefficients
Let us consider an example where the dependent variable is marks obtained by a student and explanatory variables are number of hours studied and no. of classes attended. Suppose on fitting linear regression we got the linear regression as:
Marks obtained = 5 + 2 (no. of hours studied) + 0.5(no. of classes attended)
Thus we can have the regression coefficients 2 and 0.5 which can interpreted as:

1. If no. of hours studied and no. of classes are 0 then the student will obtain 5 marks.
2. Keeping no. of classes attended constant, if student studies for one hour more then he will score 2 more marks in the examination.

3. Similarly keeping no. of hours studied constant, if student attends one more class then he will attain 0.5 marks more.

Simple vs. Multiple Linear Regression

Linear regression can be simple linear regression when you have only one independent variable . Whereas Multiple linear regression will have more than one independent variable.

Regression Equation

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots + b_kX_k$$

Measures of Model Performance

1. R-squared

It measures the proportion of the variation in your dependent variable explained by all of your independent variables in the model. It assumes that every independent variable in the model helps to explain variation in the dependent variable. In reality, some variables don't affect dependent variable and they don't help building a good model.

2. Adjusted R-squared

It measures the proportion of variation explained by only those independent variables that really affect the dependent variable. It penalizes you for adding independent variable that do not affect the dependent variable.