# Performance of Linear regression on Blood Transfusion service center Dataset

*Shrikant Patro*

*School of Computer Science and Engineering, VIT Chennai, Tamil Nadu – 600 127*

*Email: shrikantjagannath.2018@gmail.com*

## Abstract:

The machine learning model can be broadly classified into Supervised, Unsupervised, Semi-Supervised and Reinforcement learning. Out of all the above mentioned the Linear regression is one of the way to achieve the Supervised learning. In which Model is trained prior testing. The ultimate aim is to achieve the best fitting line across the data point plotted on given plane. Once the model is trained based on the value of explanatory variable we can predict the predictor. The Ordinary least square (OLS) regression method is used to predict the estimated value for the current Value of the explanatory variable. There are also other enhanced way to predict the same.

## Introduction:

Linear regression is one of the common mechanism of performing supervised learning. This model works in two different phases. This Phases are training and testing. In training phase the model is built using two type of variable. The two category of variable are explanatory and predictor. This is also called as independent and dependent variable. The value held by the independent and dependent variable is used to construct the best fit straight line. In visualization. Scatterplot is used to plot the data point on the given plane. Then the difference between the data point plotted and the predicted point that connect the straight line between two variable is used to measure the goodness of fit. The Least square error is also calculated for the same. The least square method is used in order to find the error for given predictor line. The straight line for which the least square error is minimum is considered as best fit. The best fitting line need not always will properly classify the data point. If the data point is such that it is not linearly separable. Then , It tries to find the straight line will generate minimal least square error and also ensure misclassification of data point must be minimum.

## Methodology

The goal is to obtain the linear equation of the straight line. The linear equation of the straight line is given in following form:

$$\mathbf{Y=w_0 + w_1X}$$

Y is dependent variable and X is independent variable. While $W_0$ and $W_1$ are weight that is evaluated. $W_1$ is slope that is evaluated using **dy/ dx** and $W_0$ is the intercept the point where the straight line intersecting the y-axis. Different values are tested for the $W_0$ and $W_1$ and instance when the least square error predicted between the straight line and data point found to be minimum is considered as best line. If data is linearly separable then data point belonging to one class and another class is completely separated. While, the case of the non-linearly separable dataset like iris dataset then it will try to minimize the misclassification of the data point. The line that is able to minimize the misclassification

of the data point and least square error is the best line. This best-fitting line is also called the regression line, and the vertical lines from the regression line to the sample points are the so-called offsets or residuals—the errors of our prediction.

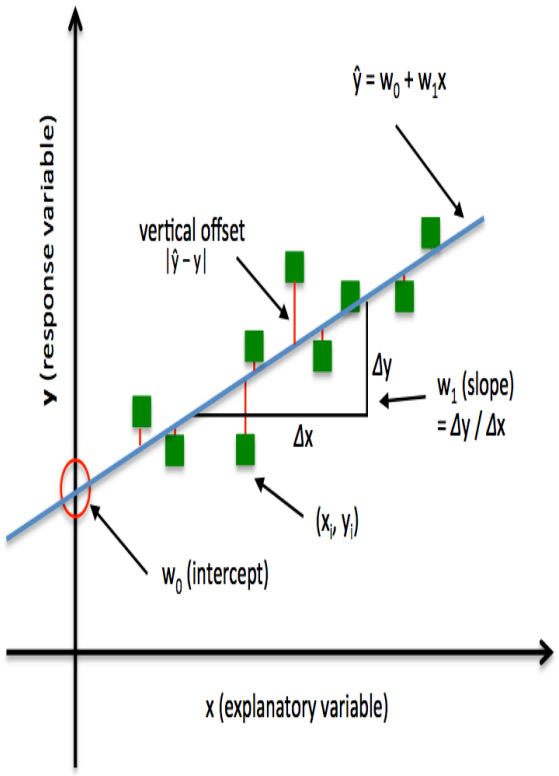



Fig 1: Regression line

The special case of one explanatory variable is also called simple linear regression, but of course we can also generalize the linear regression model to multiple explanatory variables. Hence, this process is called multiple linear regression:

$$y = w_0 x_0 + w_1 x_1 + \ldots + w_m x_m = \sum_{i=0}^{n} w_i x_i = w^T x$$

Here, $W_0$ is the y axis intercept with $x_{0=1}$

## Database- Woman's Fertility dataset

There are different CSV file for different phenomenon associated with woman fertility dataset. This dataset is collected by WHO with the help of 100 volunteers a semen sample analyzed according to WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits. Dataset characteristics: Multivariate, Attribute characteristics: Real, Associated task: Classification, Regression, Number of instances: 100, Number of attribute: 10, Missing values: NA, Area: Life, Date Donated: 2013-01-17 and number of web hits: 142847.

The further description is specifically regarding the "Blood Transfusion Service Center". This include field like Recency (Months), Frequency (Times), Monetary (c.c. blood), Time (months) and donated. Donated is binary attributed stating in terms of 1 and 0 to represent whether donated or not.

## Algorithm

In order to find the regression line we must follow the following steps:

**Step 1:** Import the pandas, numpy , sklearn , matplotlib Libraray. From sklearn import datasets , linear models. From matplotllib import pyplot.

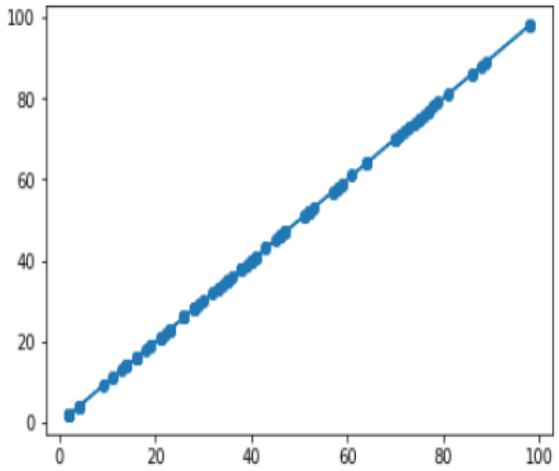
**Step 2:** Read dataset into pandas dataframe.

**Step 3:** Split training set and test set using train_test_split module imported from sklearn. Model _selection.

**Step 4:** Call linear regression function from linear_model and fit the model using training set.

**Step 5:** perform prediction using test set.

**Step 6:** Finally plot the data point and regression using pyplot module inn matplotlib.




**Conclusion:**

Linear Regression is one of the supervised learning method that is widely used for training the model for the prediction and It performed exceptionally well for linearly separable dataset while for non-linearly separable dataset its performance is marginal. There are different mechanism to fit the regression model for robust fitting of the regression model that is least affected by the outliers RANSAC is used.