# Performance Of Linear Regression Algorithm On Student Performance Data

**Shubham Barudwale**
School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: barudwaleshubham.dinesh2015@vit.ac.in

## Abstract-

Linear regression is useful model where the target function is real valued. Insted of just classification into two parts like perceptron it gives all real values. Which is more useful in deciding the accurate values.

In this example, we will try to fit the values abscences versus average marks and G1 vs G2. Linear regression mean squared error has to be minimized to fit the line.

We also calculated the minimum number of iterations to fit the data, which came less than 5 in both examples. We can see the stright line and the slope and intercept of respective fitting line.

## 1. Introduction

Unsuperwised learning is the learning from the data where the target value is not given. Linear regression is also a algorithm which uses the unsuperwised learning method to fit and learn some knowledge out of the given data.

In linear regression we try to fit the data with a stright line which reduces the mean sqquare errors between the points and line. To minimize the calculations we use the pseudo-inverse method. In which the X is matrix of attribute values and y is our hypothesis line.

We also used the RANSAC algorithm which reduces the effect of outliers on fitting.

## 2. Methadology

The linear regression algorithm is based on minimizing the squared error between h(x) and y.2

$$Eout\ (h) = E\ [(h(x) - y)^2]$$

where the expected value is taken with respect to the joint probability distribution P(x, y). The goal is to find a hypothesis that achieves a small Eout (h).Since the distribution P(x, y) is unknown, Eout (h) cannot be computed. Similar to what we did in classification, we resort to the in-sample version instead,

$$E_{in}(h) = (1/N)\ \Sigma\ (h(x_n)-Y_n)^2$$

In linear regression, h takes the form of a linear combination of the components of x. That is,

$$h(x) = \Sigma W_i\ X_i = w^T\ x,$$

substituting we get

$$E_{in}(x)=(1/N)\ (w^T\ X^T\ Xw - 2w^T\ X^T\ y + y^T\ y),$$

Where $W^T X^T w = X^T y$

If $X^T X$ is invertible, $w = X^\dagger y$ where $X^\dagger = (X^T x)^{-1} X^T$ is the pseudo-inverse of X. The resulting w is the unique optimal solution

## 4. Algorithm

1: Construct the matrix X and the vector y from the data set (x1 , Y1 ) , · · · , (xN , YN), where each x includes the xo = 1 bias coordinate, as follows

2: Compute the pseudo-inverse xt of the matrix x. If XTX is invertible

$$X^\dagger = (X^T X)^{-1} X^T$$

3: Return $W_{lin} = X^\dagger y$.

Which calculates the squared error by using previous mean square formula to reduce it.

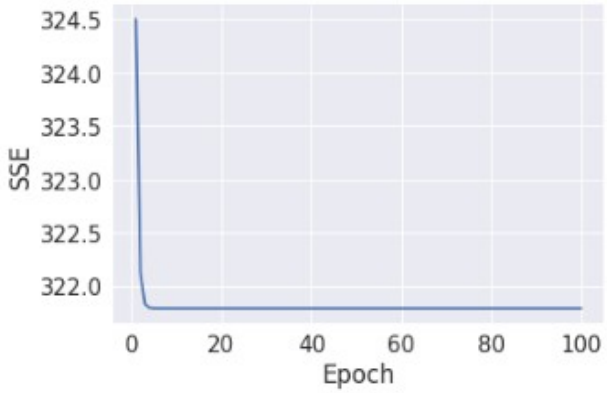
## 5. Experiments

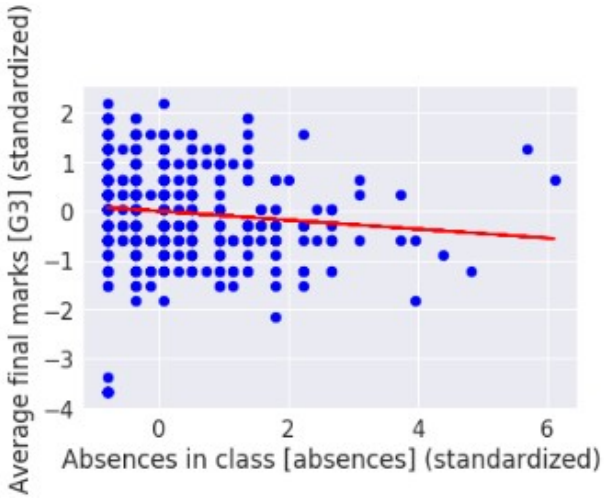
We have to fit a linear regression with the given points with the least square error method. Here we first plotted the datapoints which are Absences in class versus the average final marks in the exam. We fitted the straight line such that the mean square is least. We have also calulated the SSE vs Epoch and plotted it. It gives the least iterations required to get a linear regrassion which fits the data points. Slope and intercept of the line also has been calculated.

The same procedure is done with the data which represents the relation between the marks in two subjects.

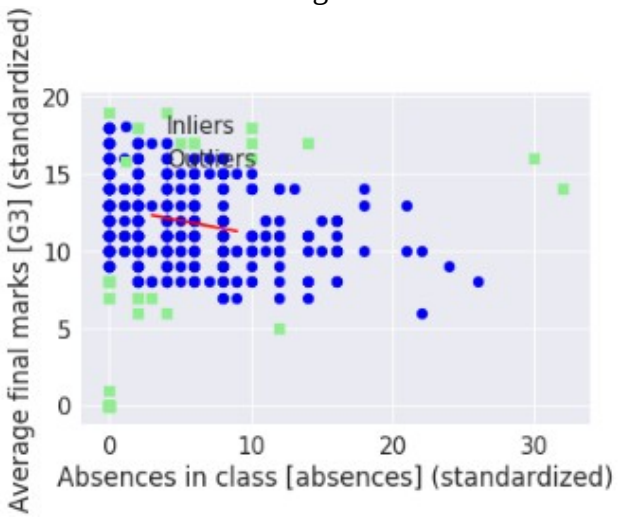
I have also used RANdom SAmple Consensus (RANSAC) algorithm which takes only the points which are close to the line and neglects the outliers since the proper result might get affected by outliers.

1) Graphs for Absences vs Marks



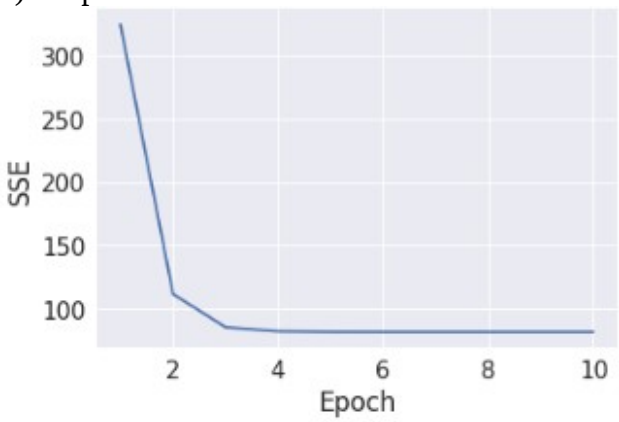



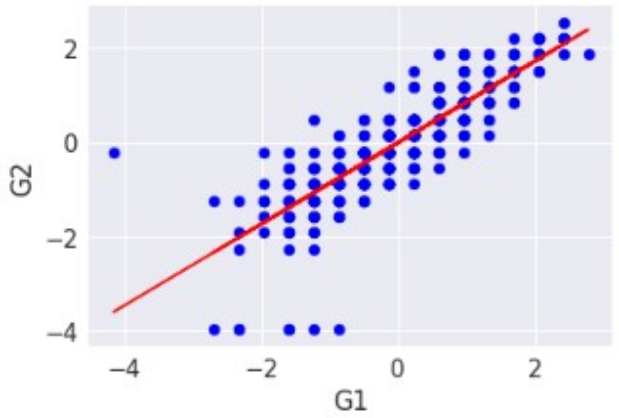

Linear Regression



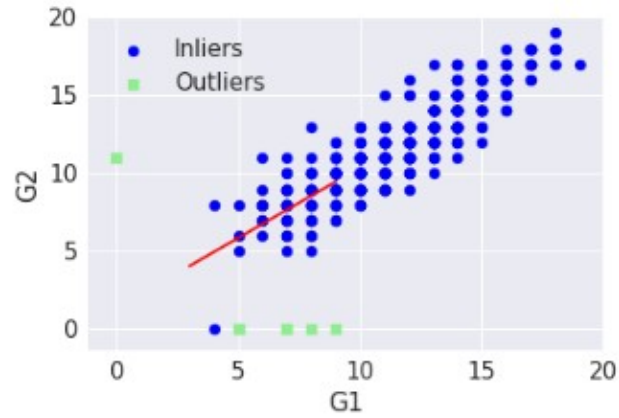

RANSAC

2) Graphs for G1 vs G2




Linear Regression




RANSAC

## 6. Conclusion

We used linear regression to fit the given data points. For first experiment where we fitted the data abcenses vs marks it took around 3 iterations to fit the data and to achieve the least squared errored line which passes through the point s given. It has slope -0.091 and the intercept is 0.0. The intercept is zero since the data we took for plotting the graph was standardized. Where the data was not standardized the intercept is 12.139 and slope is -0.064 which is almost the same. In RANSAC

algorithm we can see the outliers which the algorithm has neglected to find the line.

In second data which is G1 vs G2, the datapoints were already linear except some of the points. Where the slope of line is 0.865 and intercept is 0 since the standardized data. In non standardized data slope was 0.918 and intercept is 1.105. Here since the data was aleady linear the outliers found using RANSAC algorithm are very few.

## 7. Referances


1) Python Machine Learning by Sebastian Raschka Chapter 10

2) http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

3) http://stackoverflow.com