Linear Regression

Mumpy Bhajipale

Master of Computer Science, Pace University, New York City,USA

Abstract

A linear regression analysis is a procedure to estimate a linear connection between one or more independent variables and the dependent variable. In the specimen of statistics and machine learning, it has been acknowledged as one of the most well-known and understood algorithms. This article cites the basic concept of linear regression. This includes the concept of math used for the implementation of an algorithm, also the use of it in the market.

Keywords— *Regression, Analysis, Machine learning, share market, Trend forecast.*

1. INTRODUCTION

A linear regression notion was originated in 1894 by Sir Francis Galton. At the core of regression analysis, linear regression is a predictive analysis that shows the interdependence of scalar response and many explanatory variables. More specifically, it assumes a linear connection between the input variable(X) and the single output variable (Y). In simple Linear regression, we are interested in things like Y= MX+C. That is, every value of X has a corresponding value of Y only if it is in continuous form.

One of the most prominent uses of linear regression is in the field of stock market. It has been always a challenging role to analysis the share prices in the stock market. “Listed companies” is the term for the companies which are acceptable to the market for trading. Investors always look out for profitable investments and always try to hit for maximum profit either by purchasing or selling shares. To inspect a share index value also the fluctuation of the index value, many studies are been performed. Several techniques and algorithms are been used to witness the nature of the share market. In this case, a linear regression is used as a tool to determine the trend directions. In simple, it can be view as, when the share index price is below the linear regression line it’s not the favorable time to invest. Whereas, when the price is above the line, this could be viewed as a promising time to either sell or purchase the shares.

2. LINEAR REGESSION CONCEPT

Liner itself is defined as progression from one stage to another stage in a series of steps. This means it is used with continuous variables. But the drawback of linear regression is that it’s not good for the classification of different models.

If suppose we have an independent variable on X-axis and dependent variable on Y-axis. Let’s take the data point on the X-axis and Y-axis which is increasing. Here, we will get a positive linear regression line. Suppose we have dada variables that are increasing on X-axis and decreasing on Y-axis, then, in this case, we will catch a negative linear regression line. Once we add all the data points on the graph, the biggest task is to create the best fit line. After the line of regression is drawn using Y=MX+C, now the task is to predict the values. In this, the main goal is to reduce the error between the actual value and the estimated values, i.e. to reduce the distance between predicted or real value

and search for the best fit line. The best fit line is the one that has the least error or a least distance between the actual value and the estimated value. So, in simple words, we must minimize the error.

For instance, we have speed on the X-axis and the distance covered on the Y-axis, with a time as constant. If suppose we draw a graph between the speed of travel and the distance covered by the vehicle in a fixed unit time, we will get a positive relation. Therefore, the equation of line we use here is Y=MX+C, where Y is the distance traveled in a fixed duration of time, X is the speed of a vehicle, M is a positive slope of a line and C is a Y-intercept of the line. In contrast to this, we can put this in another form by keeping the distance constant. Here, if we plot a graph between the speed of a vehicle and the time taken to travel of fixed distance, it will get a line with a negative relationship. For this, we have a negative slope of the line, i.e. Y= -MX+C, where y is time taken to travel a fixed distance, X is the speed of the vehicle, M is a negative slope of a line and finally C as y-intercept of line.

3. MATHEMATICAL IMPLEMENTATION

Let's take the value of X and Y respectively, finally solve the problem for linear regression step by step.

STEP 1: - Plot the value for X and Y,
STEP 2:-Calculate the mean for plotted points of the X and Y-axis.
STEP 3: - Calculate the equation for slope (M)
STEP 4: - After calculating the slope, head for calculating the constant variable (C)
STEP 5: - Once all the values for the equation Y=MX+C is known, calculate the values of Y for every independent values of X.

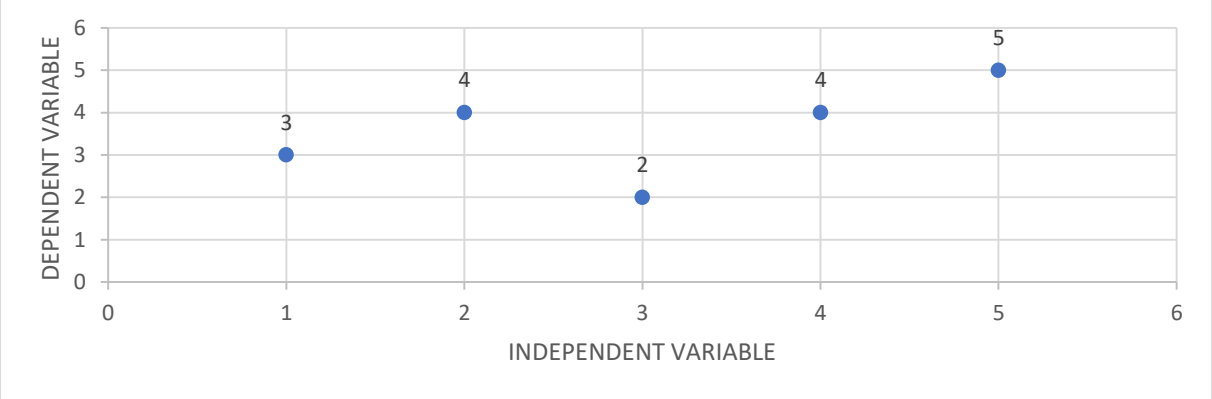


Fig 1: - Plotting of Variables.

X	Y	X-$\bar{X}$	Y-$\bar{Y}$	$(X-\bar{X})^2$	(X-$\bar{X}$)(Y-$\bar{Y}$)
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
M= 3	M= 3.6			Σ = 10	Σ= 4

Fig 2: - Mathematical Part of Variables.

Formula for calculating the slope is $\mathbf{M} = \frac{\Sigma(X-\bar{X})\ (Y-\bar{Y})}{\Sigma\ (X-\bar{X})^2}$

Where, (X-$\bar{X}$) = Distance of all the points through the line Y= 3,
(Y-$\bar{Y}$) = Distance of all the points from the line X= 3.6.

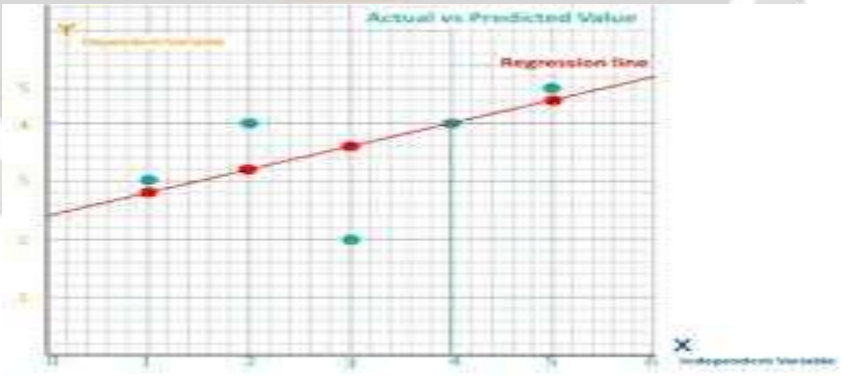
Once all the parts of the formula are done, we need to get the summation of last two-column,
Therefore, $\mathbf{M = \frac{\Sigma(X-\bar{X})\ (Y-\bar{Y})}{\Sigma\ (X-\bar{X})^2} = \frac{4}{10} = 0.4}$

Now, since Y = MX+C, we need to calculate the value for a constant C.
Y = MX+C
3.6 = 0.4(3) + C
3.6 = 1.2 + C
C = 3.6 – 1.2
C = 2.4

Y = 3.6, X = 3, M = 0.4, C = 2.4
So, **Y = 0.4X + 2.4**

Let's predict the value for Y for X = 1,2,3,4,5
Y = 0.4(1) + 2.4 = 2.8
Y = 0.4(2) + 2.4 = 3.2
Y = 0.4(3) + 2.4 = 3.6
Y = 0.4(4) + 2.4 = 4.0
Y = 0.4(5) + 2.4 = 4.4

Hence, the line passing through all these predicting points and cutting Y-axis at 2.4 is the line of regression.



Fir 3: - Linear Regression Line.

Now, comes the important task of reducing the distance. In other words, we need to reduce the error between the predicted value and the actual value. The line having the least error is known as the linear regression line or the best fit line. The simple logic for this, for different values of M it will perform N number of an iteration. As the value for M changes the line of regression also changes. In simple term, an iteration starts from first and then keep continuing to the end. So, for every value of iteration, it will calculate the predicted value according to the line and predict the distance of an actual value to the predicted value. The distance for which the value of M is minimum is considered as the best fit line.

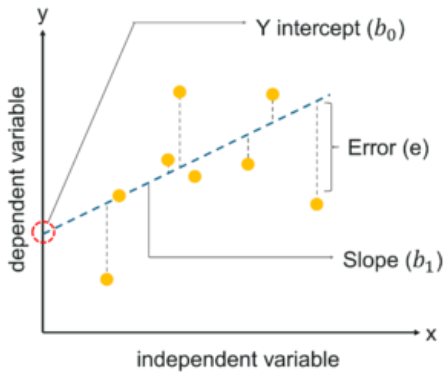


Fig 4: - Graph for Best Fit Line Calculation.

4. USE OF LINEAR REGRESSION

A. ANALYZING OF TRENDS AND SALES ESTIMATES

Linear regression can be considered as one of the best algorithms for the estimation of trends and sales. For instance, if a company has steadily increased sales every month for the past few years, then we can calculate a linear regression for the sales data with monthly sales on the Y-axis (dependent variable) and time on the X-axis (independent variable). This will help us to find the line which can predict an upward trend in the sale. With the estimated trend line, it's easy for the company to use the slope of the line to focus on sales for a future month and long run

B. ESTIMATING THE IMPACT OF PRICE CHANGES

We can always rely on a linear regression algorithm to analyze the effect of pricing on customer behavior. For example, if a company changes certain product prices, then it can record all set of the quantity by itself, also check the level of price and can finally perform the linear regression algorithm with quantities as a dependent value and price as the independent value. The resulted line can help us to predict the reduction of product consumption as the price keeps increasing. This is the best way of deciding future prices.

C. COSTING OF RISK FACTOR IN INSURANCE AND FINANCIAL SERVICES

As the estimation of sales and changing of price, we can also use linear regression for analyzing the risk factor. Suppose a health insurance company needs to conduct a linear regression algorithm. It can be done by plotting a graph of a number of claims per person with customers' age. This then helps them to realize that the old customers claim for more health insurance. Such results are beneficial for an important decision of a business. It aids to cut off the risk factors for the long term in business.

5. CONCLUSION

Linear regression is a lengthy step of procedure, but it has a painless and an undemanding calculation to conduct. We can find different methods to calculate linear regression in different fields. Because of its simplicity, the use of a linear regression is intense in many sectors. A linear regression gives us a clear picture and provides a ground for selection criteria like data quality, classification and regression capabilities, comprehensible and transparent and

computational complexity. This article depicts a brief information on the efficiency of a linear regression algorithm in the market.

6. REFRENCES

[1] http://www.j-pcs.org/article.asp?issn=2395-5414;year=2018;volume=4;issue=1;spage=33;epage=36;aulast=Kumari

[2] https://medium.com/x8-the-ai-community/practical-aspects-linear-regression-in-layman-terms-8981c60a848c

[3] https://machinelearningmastery.com/linear-regression-for-machine-learning/

[4] https://mobiledevmemo.com/when-why-and-how-you-should-use-linear-regression/

[5] https://people.cs.pitt.edu/~milos/courses/cs2750-Spring03/lectures/class6.pdf

[6] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm