International Journal of Engineering &Technology, 7 (3.12) (2018) 960 -962

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Least Square Regression for Prediction Problems in Machine Learning using R

Anila.M[1, 2*], G. Pradeepini[3]

[1]*Research scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh*
[2]*Assistant professor, Department of IT, MLR Institute of Technology, Hyderabad, Telangana*
[3]*Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh*
** Corresponding Author E-mail: [1]anilarao.m@gmail.com*

Abstract

The most commonly used prediction technique is Ordinary Least Squares Regression (OLS Regression). It has been applied in many fields like statistics, finance, medicine, psychology and economics. Many people, specially Data Scientists using this technique know that it has not gone with enough training to apply it and should be checked why & when it can or can't be applied.
It's not easy task to find or explain about why least square regression [1] is faced much criticism when trained and tried to apply it. In this paper, we mention firstly about fundamentals of linear regression and OLS regression along with that popularity of LS method, we present our analysis of difficulties & pitfalls that arise while OLS method is applied, finally some techniques for overcoming these problems.

Keywords: *Independent variable, Dependent variable, Least square regression.*

1. Introduction

Let us explain it with small scenario in which we need to find person's height using their weight (in pounds) information & when such situations have to dealt we use framework for regression i.e. relationship between two or more variables where one is called dependent variable y (also called as explained variable, output variable) that we want to predict.
According to our example, 'y' represent person height (in inches) and let x1, x2…, xn (generally referred as explanatory variable, input variable) be the independent variable which we use to predict for y. For example, let us take two variables say,x1 and x2 where x1 represent age in years and x2 represent weight in pounds. Also assume that we have some set of values for x1,x2,…xn represents set of independent variables.so here n, can also be referred as dimension for feature space. As regression main motto is to make an attempt to predict values of dependent variable 'y' from 'x1, x2…,xn'.
The main objective here is to predict people's heights using x1(age's) and x2(weights), which can be done using training data set that consists of weight, age and height. Regression algorithm [5] will learn from this dataset later when any test data containing only weight and age is given, it can be able to predict their heights. This could also be treated as a ranking task, as we can see same kind of methodology in classification technique when we attempt to categorize employees into three groups ('black', 'white', 'red') using some related training data. Linear Regression always tries to solve any problem with an assumption that dependent variable is a linear function of independent variable, i.e. we can be able to estimate 'y' using formula
y=c0+c1x1+c2x2+…+cnxn

2. Regression

The generalized equation for linear regression is y=c+m*x, where m and c are constants that can be calculated using regression algorithm.
As discussed earlier the main aim of linear regression is to find best choices of values for the constants m and c to make out formula more accurate. It must be important to mention why we are calling this as linear model, for the coefficients m and c, we plot function y(x) i.e. y(x)=c+m*x1 and it forms a line, for example consider plot y(x)=3+2*x1, then we get the below plot. The corresponding code in R language [7] is also written.

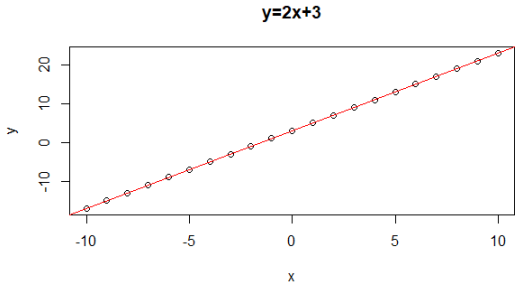


Fig. 1: Plot for y=2*x+3

R script for linear regression

```
x<-c(-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7,8,9,10)
y<-2*x+3
rel_xy<-lm(y~x)
View(rel_xy)
```

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

rel_xy
plot(x,y)
title(main='y=2x+3')
abline(rel_xy,col='red')
As we increase number of independent variables from to three or more, then linear functions will form planes to hyperplanes, which are further generalizations of lines to higher dimensional feature spaces.
Now we come to our actual point that the aim of regression is to find the best values for the constants c0, c1…cn and make our formula more precise to predict dependent variable 'y' using the 'x' values in the dataset. To let this thing happen consider we have few person's weight (in pounds) and age (in years) and height (in inches) and use the below formulation:
height = c0 + c1*weight + c2*age

2.1 OLS Regression

It is mandatory to choose the constants c0, c1, c2..., cn to show that our linear modelling method is as accurate to act as a predictor of height as possible. The word "accurate" in the former sentence means here is, determining the best and possible values for the constants by using the training dataset information so that the 'y' values can be predicted in an efficient way. And this can be achieved only by using OLS Regression.
The Least squares method says that we need to choose these constants making every instance in our training data to minimize the sum of the squared differences between the original dependent variable and it's predicted value. we need to select c0, c1, c2, …, cn to minimize the sum of the values (actual y – predicted y)^2 for each training point i.e.
(y – (c0 + c1*x1 + c2*x2 + c3*x3 + … + cn*xn))^2
for each training point of the form (y, x1, x2, x3, …).
To better understand this, let us consider an example training data and apply least square method to it. Our dataset contains of each person's (height, weight, age) data (for about 20 instances are used here).

Following is the R command to install required packages
install. packages("scatterplot3d", repos="http://R-Forge.R-project.org")
then use scatterplot3d function in Rstudio
Dataset contains nine persons data where 'y' representing 'height', 'x1' representing 'weight', 'x2' representing 'age'

Table 1: Dataset

Y	x1	x2
65	150	18
68	171	25
72	177	23
69	176	69
69	139	14
67	169	59
73	178	42
72	178	23
67	172	60

Our training data can be visualized in a 3D plot as below, where one axis represent y, another x1 and the other one x2.

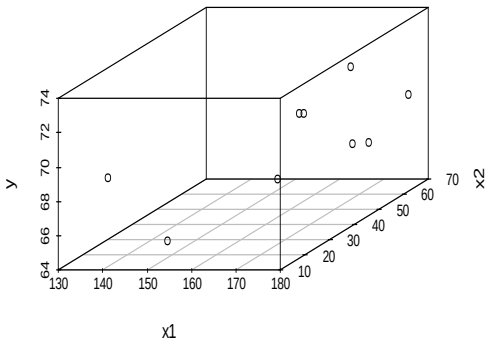


Fig. 2: 3-Dimension plot showing y,x1,x2 values.

The main objective of OLS regression is to minimize the sum of the squared error between the values of the dependent variables in our training data, and our model's predictions for these values. As we have out dataset mentioned in the above table, we look for determining values of c0,c1 & c2 to minimize squared error, these values can be calculated using the following command in R
lm(formula = y ~ x1 + x2)
where 'lm' means linear modelling
so, the values for c0, c1 and c2 are:
c0=46.19366 (inches)
c1=0.15157 (inches per pound)
c2=-0.06789 (inches per year)

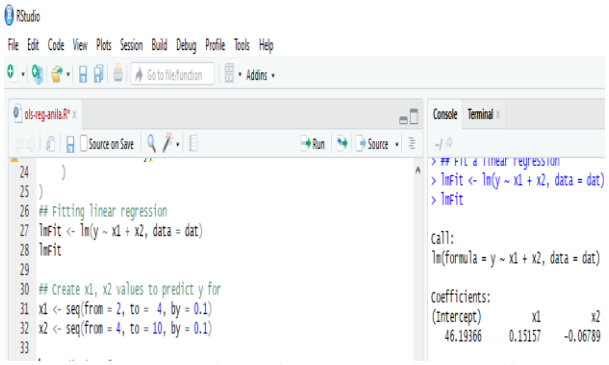


Fig. 3: Implementation of lm() function in RStudio (coefficient values derived are also shown at right hand side)

Sum of Squared Errors is also calculated and is given below
SSE= 15.893
Now, by using the above determined values of coefficients we can use the following formula to predict height of a person when weight and age is given.
height = 46.19366-0.06789*age+0.15157*weight

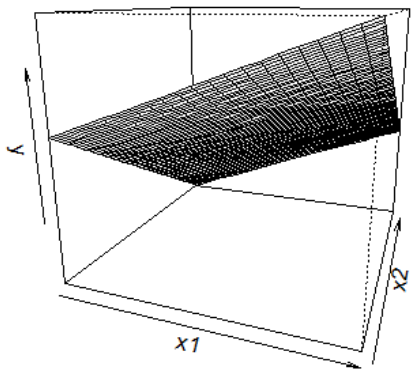


Fig. 4: Regression plane

From the above plane, height of person can be derived when weight and age are given.

3. Popularity of Least Squares Method

OLS Regression can be also called as a method that not only predicts dependent variable also acts as a technique for measuring "accuracy" (i.e. by using the sum of squared errors) and this fact makes it different from other forms of linear regression. Because of this reason, most of the people also call it as "least square regression" as it has attributed many applications and advantages listed below.

- Provides an easier method to analyze than any other regression technique.
- As it has got very basic level formulae, it is not much difficult to understand.
- Solutions can be instituted quickly and easily interpretable (in case of determining values for constants).

- One of the earliest general prediction methods.
- Firstly, invented by Carl Friedrich Gauss in 1795, later re-invented by Adrien-Marie Legendre in 1805 being very useful approach for making predictions.

While some of these explanations for using least squares are convincing under certain circumstances, our goal should be to find the model that does the best job at making predictions given our problem's design and constraints (such as prediction time, limited training points, processing time, and computer memory).

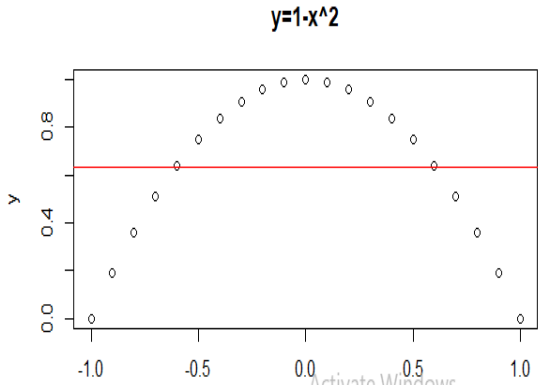
4. Pitfalls of OLS Regression

4.1 Participation of Outliers

We can know that OLS performs in an odd way when training data contains either extremely large or small values in it. The outlier we add vividly changes the least squares solution and hence will lead to much less precise predictions. Better solutions for this issue are discussed below:

- One best way to solve this issue is to measure accuracy in a way that does not square errors. Least absolute square method is one of the finest technique that help to overcome the problem of least square while outliers are involved in dataset.
- Least median of squares [9] is the other technique that determines median error made on training data.
- The finest method is to pre-process the data using any outlier detection algorithm [12] so as to avoid or remove outliers while applying linear regression.

4.2 Non-linearities



The line drawn in above plot depicts least square solution we can observe that it's not perfectly linear. This is the major drawback of OLS is same, that is linear regression is not always rectilinear, still it tries to fit in later attempts to do that.

In general, least squares method trains the datasets very well. But, in the problem statement that we have mentioned earlier in our discussion approximately tends y values to be zero as the dataset is too large. So, these kinds of problems can be solved by transforming the data so that it becomes perfectly linear, this would be possible only when user has better understanding of the system. If independent variables are transformed then it is okay, but when transformations are applied to dependent variables then it might project the results of least squares in falsehood, leading our regression method to fail in predicting values, there by generating objectionable outputs.

An optimal approach to avoid such non-linearities is to apply non-parametric regression [4] method such as kernel-regression method [6] or local linear regression. These approaches are non-parametric based, hence require setting up of model parameters, and it can be done by using cross validation.

OLS users must know when there is a strong correlation among independent variables least square method yields poor predictions, in fact poor performance on the test data.

4.3 Handling of Noise in the Independent Variables

Least square regression can better handle the noise of output variables, but it cannot be possible to do the same with independent variables. Whether it is independent or dependent variable, noise creates many difficulties while predictions happen. There might be errors like measurement error, rounding error or uncertainty of data. Etc., Total least squares method can be used when there is presence of extensive noise in the independent variables. Alternate method is least products regression [10].

4.4 Heteroscedasticity

It's a systematic pattern in the errors where the variances of the errors are not constant. This occurs when variance of the error terms differs across the instances, which is actually a big problem to deal with especially when the variances are not equal (because the corresponding reliability will also be unequal) [11].

There are some consequences while using OLS when heteroscedasticity is present.

- It is still linear and not efficient.
- Generation of incorrect standard errors.
- OLS estimation yields unbiased coefficient values.

So, in the presence of heteroscedasticity [13], variance of OLS estimators are not provided by the usual formulas. But if we still tend to do that, it misleads and results in inaccurate conclusions.

5. Conclusion

OLS regression is of course indisputably a convenient approach, even it has observed with some imperfections, but not an optimal method for implementing on real situations. So, there is still a requirement of discovering a top and perfect approach or method that accurately generates solutions for prediction problems.

References

[1] Breiman, L. (1984). Classification and Regression Trees. New York: Routledge.
[2] Book: "Regression Analysis and Linear Models: Concepts, Applications, and Implementation" by Richard B. Darlington, Andrew F. Hayes
[3] A. M. Bagirov, C. Clausen, M. Kohler, "Estimation of a regression function by maxima of minima of linear functions", *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 833-845, Feb'2009.
[4] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, New York, NY, USA: Springer-Verlag, 2002.
[5] D. W. Hosmer, S. Lemeshow, R. X. Sturdivant, Applied Logistic Regression, New York, NY, USA: Wiley, 2013.
[6] Widely Linear Complex-Valued Kernel Methods for Regression by Rafael Boloix-Tortosa ; Juan José Murillo-Fuentes ; Irene Santos ; Fernando Pérez-Cruz, published in IEEE Transactions on Signal Processing (Volume: 65, Issue: 19, Oct.1, 1 2017).
[7] Basics of R: https://www.udemy.com/r-basics/
[8] Heteroscedastic Max–Min Distance Analysis for Dimensionality Reduction- Xiaoqing Ding, Changsong Liu, Ying Wu
[9] D. Buchczik, Least Median of Squares in Multivariate Calibration, 2005.
[10] Least product relative error estimation - Chen, Kani - Lin, YuanyuanWang, Zhanfeng Ying, Zhiliang - Journal of Multivariate Analysis, VL - 144, 2016, DA - 2016/02/01/, 0047-259X
[11] The SAGE Handbook of Regression Analysis and Causal Inference, Henning Best, Christof Wolf,2014
[12] C. C. Aggarwal, Outlier Analysis, New York, NY, USA: Springer, 2013.
[13] P. Chen, L. Jiao, F. Liu, J. Zhao, Z. Zhao, S. Liu, "Semi-supervised double sparse graphs-based discriminant analysis for dimensionality reduction", Pattern Recognit., vol. 61, pp. 361-378, Jan. 2017.