Research of a New Method for Solving Linear Regression

Yang Yu
Shandong University of Science and Technology, Qingdao 266590, China
1010891653@qq.com

Abstract. In this paper, we propose two new models for the absolute error of a linear regression. One of which is for the average value of all data points connected the slope, and then find their average value can be obtained slope, finally use the mean point can get a linear regression equation. The other is to use the regression line after the mean value, and each point deviates from the slope of the regression line to get the average value, that is, the slope of the regression line. The least squares method is sensitive to outliers, so in the case of outliers, the new method outperforms the least square method under certain circumstances, and in some cases, the absolute error and the prediction error are better than the least square method.

Keywords: New regression model; Least squares method; Total absolute error; Precision accuracy.

1. Introduction

Regression analysis is a classical data processing method in statistics. It can predict future value. Only the regression of two variables is called one linear regression. It is also common in practical application. There are many methods to solve it. The least square method is the most commonly used method to solve the model formula. [1]

In 1806, French scientist Legendre invented the "least square method [2]". independently, but it was unknown to the world. The least square method is a mathematical optimization technique, which minimizes the square of the error [3] and finds the best function matching of the data. The least square method is the best way to solve linear regression, but its absolute error [4] is not the least. In Xiru Chen's [5] thesis, the merits and demerits of the minimum absolute method are discussed. It is pointed out that the least square criterion is greatly influenced by outliers and the minimum absolute criterion is small, but the minimum multiplication has no fixed formula, so it is not often used. One of the important reasons that least squares regression is widely used is that it is simple to calculate and can be expressed by formula. A new linear regression model is proposed in this paper. In some cases, the absolute error of the model is less than the absolute error of the least square method, and the predicted value is more accurate.

2. New Linear Regression Model

2.1 Derivation of a New Linear Regression Model

The main idea of the new model is to assume that the model passes through a straight line and a mean point. All samples are connected to intercept to form their respective slopes. These points must be evenly distributed on the upper and lower ends of the line. The slope of all points and intercept connections is subtracted from the assumed straight-line slope. Add all of them and make them equal to zero. Finally, simultaneous equations are used to solve the intercept and slope of the assumed line.

Now suppose independent variable x_i and dependent variable $y_i (i = 1,2 \cdots n)$ satisfy $y_i = k_\theta x_i + b_\theta$,and $\bar{Y} = k_\theta \bar{X} + b_\theta$,The slope and intercept are unknown, and fig.1 is helpful to understand the model. $\bar{X}$ and $\bar{Y}$ are mean values. According to the new model, the following equations can be obtained.

$$\begin{cases} \dfrac{\sum_{i=1}^{n} \dfrac{y_i - b_\theta}{x_i}}{n} = k_\theta \\ \bar{Y} = k_\theta \bar{X} + b_\theta \end{cases} \tag{1}$$

It can be obtained from (1): $\sum \frac{y_i}{x_i} - \sum \frac{b_\theta}{x_i} = nk_\theta$,the common factor b_θ is obtained:

$$b_\theta = \frac{\sum \frac{y_i}{x_i} - nk_\theta}{\sum \frac{1}{x_i}}$$

Combining two equations: $\begin{cases} b_\theta = \dfrac{\sum \frac{y_i}{x_i} - nk_\theta}{\sum \frac{1}{x_i}} \\ b_\theta = \bar{Y} - k_\theta \bar{X} \end{cases}$ 。

Finally, the following formula is worked out:

$$k_\theta = \frac{\sum_{i=1}^{n} \frac{y_i}{x_i} - \bar{Y} \sum_{i=1}^{n} \frac{1}{x_i}}{n - \bar{X} \sum_{i=1}^{n} \frac{1}{x_i}}, \qquad b_\theta = \bar{Y} - k_\theta \bar{X} \tag{2}$$

The slope of the new model is defined as: Each point deviates from the slope of the regression line and then the average value is obtained. That is the slope of the regression line. It can also be understood as the slope produced by the intercept between the points above and below the regression line minus the slope of the regression line. Add it all together and make it equal to zero. Finally, simultaneous equations can be used to solve the slope and intercept of the model. The new model can be understood by Figure 1.

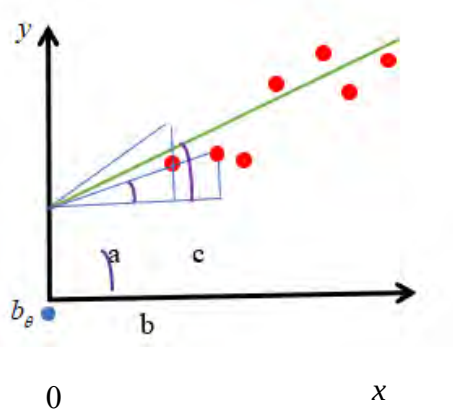

Fig.1 New linear regression model

The slope c in the graph is minus the slope b, plus the slope c minus the slope a. the other red dots were all subtracted from the slope c. Then add them all together. The sum equals zero, and then uses the mean to connect two equations. Then we can find out the slope and intercept of the regression line.

Now, the formula for solving the unitary linear regression by the least square method is as follows.

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, \quad b = \bar{Y} - a\bar{X} \tag{3}$$

Where a is slope, b is intercept. Next, compare the fitting effects of the two models. Compare absolute error, residual squared sum and prediction error.

3. Numerical Experiment

Known straight line $y_w = x_w + 1$ passes through $(1,2)$, $(2,3)$, $(4,5)$, $(5,6)$, $(6,7)$, $(7,8)$, $(9,10)$ and $(10,11)$ so on, Add outliers $(3,7)$ and $(8,6)$,and point $(11,12)$ as a prediction value.

Use the least squares method and the new model to fit these points respectively. Than compare the errors produced by the two models. Fig. 2 is a comparison chart of two models fitting effect, and Table 1 is a comparison of three errors of two models.

The regression equation obtained by least squares fitting is as follows:

$y_v = 0.818181818181818\, x_v + 1.999999999999999$

The sum of squares of the total residuals is 15.273, and the total absolute error is 8.727. When $x_v = 11$, the absolute error of 12 is 1.000.

The regression equation fitted by the new model is:

$y_g = 0.897697379104283\, x_g + 1.562664414926443$

The sum of squares of the total residuals is 15.794, and the total absolute error is 7.535. When $x_g = 11$, the absolute error of 12 is 0.563.

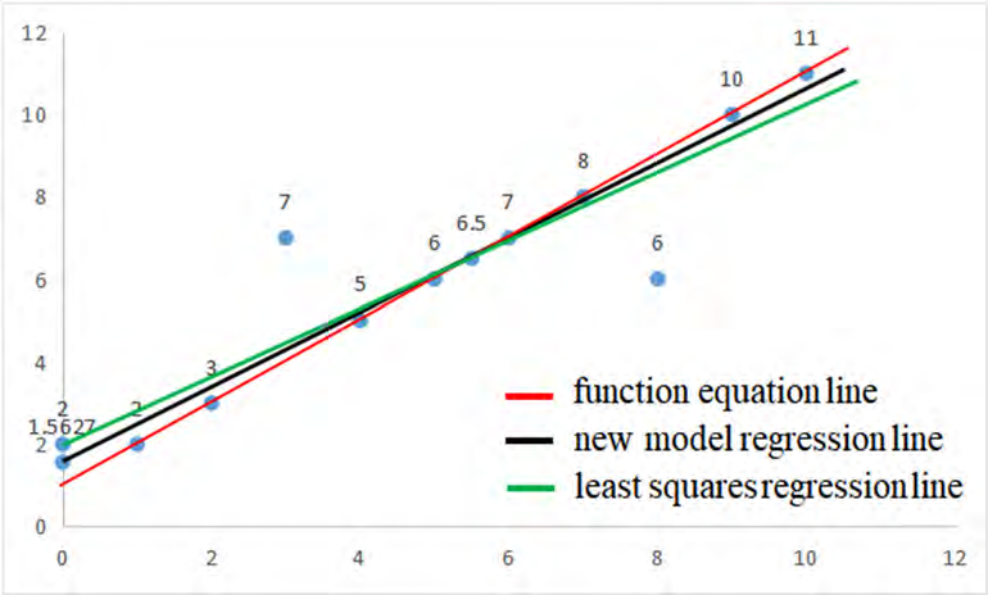

Fig.2 Comparison of straight-line fitting effects between two models

Table 1. Comparison of three errors between the two models

Model	Sum of squares of total residuals	Total absolute error	Absolute error of prediction point (11,12)
New model (2.3)	15.794	7.535	0.563
least square method(2.4)	15.273	8.727	1.000

As you can see from the above diagram, the new model is nearest to the given line. With the extension of the straight line, the absolute error and prediction error of the least square method will be bigger and bigger. The fitting effect of the new model is better than that of the least square method. The least square method is the best when there is no outlier, but it is not the best choice when there is outlier data. From the above diagram, we can see that the new model is closest to the given line, with the extension of the line, the absolute error and prediction error of the least square's method will become larger and larger, the fitting effect of the new model is better than the least squares method. The least square method is the best without outliers, but it is not the best choice when outliers exist. The new model can also be applied to some occasions. As can be seen from Table 1, the prediction error of the new model is much smaller than that of the least square method, so it is a good choice for the new model to predict the future value in some occasions.

4. Summary

The sum of squares of residuals in fitting a linear regression with least square method is the smallest in any case, but the absolute error is not absolutely optimal. Moreover, the error of prediction is sometimes very large. The least squares method is very sensitive to outliers. So, it is not good to fit straight lines in the case of abnormal values. When the outliers are in a certain degree, the fitting effect of the new model is better than that of the least square method. If the abnormal value has seriously deviated from the normal range, no matter what method is used to fit it, the fitting line will seriously deviate from the data points to be fitted, unless the great abnormal value is eliminated and then fitted. Linear regression is a method of data statistical collation. The analysis and prediction of statistical data is the core content of statistics. It is a process of exploring the inherent law of data through statistical description and statistical inference. Precision is the most important index to predict the future value with the existing data, and this method is more accurate in some cases, so it is suitable in some cases.

References

[1]. Yumao Li. Theory and Application of Univariate Linear Regression Method [J]. Journal of Chifeng University (Natural Science Edition), 2017,33 (15): 1-2. (in Chinese)

[2]. Goldberger, Arthur S. (1964). Classical Linear Regression. Econometric Theory. New York: John.

[3]. Editorial Committee. Mathematic Dictionary (Volume IV) [M]. Press: ShanXi Education Press,20-02:(403-404) (in Chinese)

[4]. Yidan Shen. Simple and clear mathematical dictionary[M]. Press: Beijing Institute of Technology-y Press,2007 (in Chinese).

[5]. Chen Xiru. Least absolute linear regression (I) [J]. mathematical statistics and management, 1989, (05): 48-55. (in Chinese)

[6]. Long Tian. Least-squares Method Piecewise Linear Fitting[J]. Computer Science,2012(S1):482-484 (in Chinese)

[7]. Trevor Hastie. The elements of Statistical Learning: Springer,2009: P11.