The Application Research of Least Square Linear Fitting Based on Logarithmic Transform

Mengqiu Kong

Guizhou Minzu University 550025 HuaXi, Guiyang, China

410828640@qq.com

Keywords: Linearity after logarithmic, transformation, least square method

Abstract: This paper theoretically proves that the least square method for linear fitting after logarithmic transformation of data is essentially based on the principle of reducing relative error. Through empirical analysis and comparison with the traditional least square method, it is found that the least square method after logarithmic transformation of data has a higher precision in model fitting effect, and at the same time makes it possible to use the least square method after logarithmic transformation of data to fit the model accurately. Fitting the straight line gives better consideration to the information of all observation points.

1. Introduction

When the traditional least squares method is used to fit regression, it often considers eliminating the larger data which has great influence on the regression estimation (the original data is correct), but if the sample data is small, while it is removed, the sample will lose part of the information. If it is not removed, the regression line will shift to larger data points. As a result, the fitting accuracy of regression model is not high or the desired results cannot be obtained. Yimin Wang (1997) [1] thinks that the accuracy of curve fitting can be improved by considering relative error, which has been proved in engineering test. Bing Li(2007) [2] thinks that when the observed data are abnormal, Jacobi iteration pretreatment of the iteration matrix and the least square method can improve the accuracy of parameter estimation and fitting accuracy. At the same time, more scholars have applied least squares curve fitting to finance, physics, engineering and other fields, after processing data by various methods such as logarithmic transformation, exponential transformation, trigonometric function transformation, normalization, standardization, interpolation and so on, the least square method is used for curve fitting or prediction to improve the accuracy of the model [3-7].

2. The Principle of Traditional Least Square Method

There is a linear relationship between a random variable y and $p-1$ independent variable $x_1, x_2, \cdots, x_p$, which satisfies the relationship

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon \\ \mathrm{E}(\varepsilon) = 0, D(\varepsilon) = \sigma^2 \end{cases} \tag{1}$$

Give n groups of sample observation values $\left(x_{i1}, x_{i2}, \cdots, x_{i,p-1}; y_i\right) i = 1, \cdots, n$, the equation (1) can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1,2,\cdots,n \tag{2}$$

The equivalent form is

$$\begin{cases} Y = X\beta + \varepsilon \\ \mathrm{E}(\varepsilon) = 0, D(\varepsilon) = \sigma^2 I_n \end{cases} \tag{3}$$

Where Y is the variable observation vector of $n \times 1$, X is a known design matrix of $n \times p$, β is an unknown parameter vector of $p \times 1$ and ε is a random error vector.

Take the length of the error vector $\varepsilon = Y - X\beta$ squared $\|Y - X\beta\|^2$ and minimize it, the error here refers to the total error. Remember as

$Q(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)'(Y - X\beta)$,

The Principle of Finding Extremum by Calculus, take the partial derivative of β and set it to zero, get the equations

$X'X\beta = X'Y$

Get the estimate of β

$\hat{\beta} = (X'X)^{-1} X'Y$

Thus, the linear regression model can well fit the known data and make prediction.

3. Error vector calculation after logarithmic transformation

From Formula (2), logarithmic to Y

$$\ln y_i = \beta_0' + \beta_1' x_{i1} + \cdots + \beta_{p-1}' x_{i,p-1} + \varepsilon_i', \quad i = 1, 2, \cdots, n \tag{4}$$

Order $y_i' = \ln y_i, \quad i = 1, 2, \cdots, n$,

$y_i^0 = \beta_0' + \beta_1' x_{i1} + \cdots + \beta_{p-1}' x_{i,p-1}, \quad i = 1, 2, \cdots, n$

$$y_i' = \beta_0' + \beta_1' x_{i1} + \cdots + \beta_{p-1}' x_{i,p-1} + \varepsilon_i' = y_i^0 + \varepsilon_i', i = 1, 2, \cdots, n \tag{5}$$

In matrix form, equation (5) is transformed into

$$\begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0' \\ \beta_1' \\ \vdots \\ \beta_n' \end{pmatrix} + \begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_n' \end{pmatrix} \tag{6}$$

To differentiate from the preceding, order $T = (y_1', y_2', \cdots, y_n')'$,by the same method as above

$T = X\beta'$

Take the length of the error vector $\varepsilon' = T - X\beta'$ squared $\|T - X\beta'\|^2$ and minimize it, the error here refers to the total error. Similarly, the estimate of β' is obtained

$\hat{\beta}' = (X'X)^{-1} X'T$.

4. Relative error analysis

For convenience, definition of relative error $D\varepsilon$,the traditional least square method defines the relative error as follows

$$D\varepsilon_i = \frac{\varepsilon_i}{y_i} \tag{7}$$

Obviously, here ε_i and y_i are independent, the traditional least square method considers the global error. Every each error ε_i has the same weight, it doesn't vary by size. Therefore, it is difficult to meet the actual needs in real life.

The relative errors after logarithmic transformation are as follows:

$$D\varepsilon_i^{'} = \frac{\varepsilon_i^{'}}{y_i^{'}} \tag{8}$$

After reducing it to an exponent, $e^{y_i^{'}}$ is the observed value (actual value), $e^{y_i^{0}}$ is the predicted value, then the upper formula can be transformed into

$$D\varepsilon_i^{'} = \frac{e^{y_i^{'}} - e^{y_i^{0}}}{e^{y_i^{'}}} \tag{9}$$

From formula (6) $y_i^{'} = y_i^{0} + \varepsilon_i^{'},\ \ i = 1, 2, \cdots, n$

The substitution of the upper form can be transformed into

$$D\varepsilon_i^{'} = \frac{e^{y_i^{'}} - e^{y_i^{0}}}{e^{y_i^{'}}} = \frac{e^{y_i^{0}+\varepsilon_i^{'}} - e^{y_i^{0}}}{e^{y_i^{0}+\varepsilon_i^{'}}} = 1 - \frac{1}{e^{\varepsilon_i^{'}}} \tag{10}$$

Obviously, the relative error $D\varepsilon_i^{'}$ here is only related to $\varepsilon_i^{'}$, that is, the size of each $e^{y_i^{'}}$ is different, its relative error $D\varepsilon_i^{'} = 1 - \frac{1}{e^{\varepsilon_i^{'}}}$,decision by $\varepsilon_i^{'}$,in formula (6), each error a has the same weight or is equally important. Therefore, its relative error is also of equal weight or importance, so logarithmic transformation is equivalent to relative error and minimum.

5. The empirical analysis

5.1 data sources

Data are derived from textbooks on linear statistical models [7].An experimental vessel relies on steam to provide heat, thus keeping the temperature constant. Figure 1 shows the relation between steam and temperature.

As shown in Figure 1 of the scatter plot, this case belongs to the linear least square fitting method.

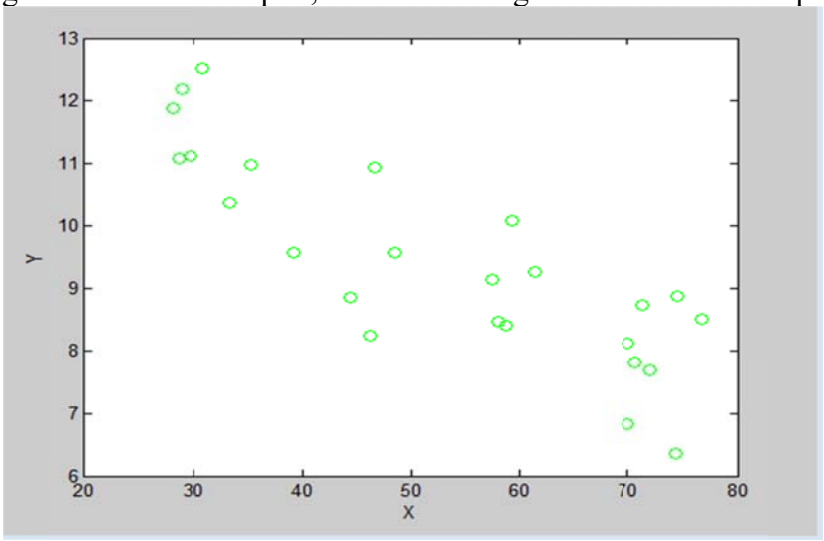

Fig. 1 The relationship between steam and temperature

5.2 The fitting results of two kinds of least squares

Data from Table 1, MATLAB R2014a Software Programming, The linear regression equation obtained by the traditional least square method is

$$y = 13.623 - 0.0798x \tag{11}$$

The linear regression equation fitted by the least square method based on the logarithmic transformation of dependent variables is as follows

$$y = 2.6730 - 0.0084x \tag{12}$$

Formula (7) shows that the amount of vapor required per unit time decreases by 0.0798 (L) for every increase of ambient temperature around the container at 1 (℃).

From formula (8), it shows that the amount of steam required per unit time decreases by 0.0084 (L) for every 1 ℃rise in ambient temperature around the container.

5.3 Analysis of Empirical Results

As can be seen from Table 2, The total and relative errors of logarithmic transformation of dependent variables are much smaller than those of traditional methods.in fact, the improved algorithm can allow the existence of large errors for large data.as for the smaller data, the error will be smaller, so that the final line can take into account all observation points at the same time. It is not that the straight line obtained by traditional method will migrate to larger observation points or need to exclude larger data than others, and the fitting accuracy is higher. similarly, the green line in Fig.2 represents the linear fitting of the traditional least squares method. While the blue line is the linear fitting of the data after logarithmic transformation. Obviously, the latter has higher fitting accuracy and can better take into account all observation points.

Table. 2 Comparison of Empirical Analysis Results

Method	Overall error	relative error
traditional method	18.2234	0.2433
After logarithmic transformation	17.5805	0.2335

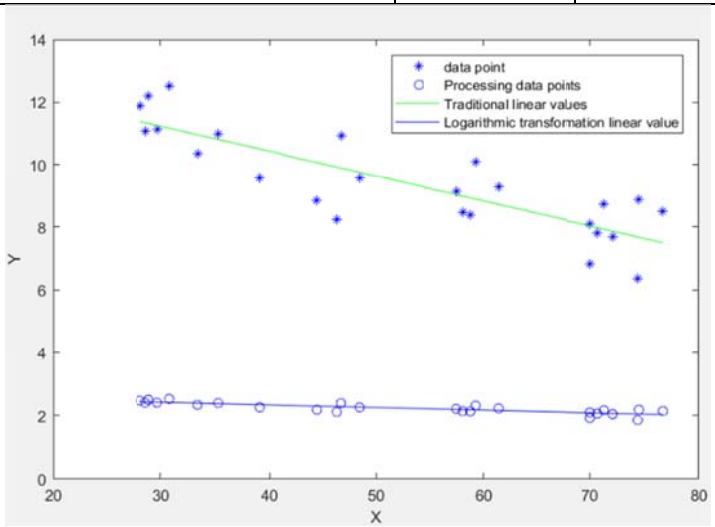

Figure. 2 Linear fitting of data

6. Conclusion

Based on the traditional least squares method, this paper calculates the logarithmic transformation of dependent variables. After deduction and analysis, the example validates that, compared with the

traditional theory of the method, for the data model with large difference, there is no need to eliminate the large dispersion or abnormal data. The method after processing data has the advantages of high fitting accuracy, and the relative error and the overall error are reduced. It can better take into account the information of all observation points. It can be used in finance, physics, engineering and other fields.

References

[1] Yimin Wang, Ying Ma.Defects and Improvements of Traditional Least Square Curve Fitting [J]. Journal of Electric Power, 1997 (01): 51-54.

[2] Bing Li, Ning Zhu, Wenfang Tang, Fujian Duan. Improved least squares estimation to determine high-precision parameter model [J]. Statistics and decision-making, 2007 (23): 9-11.

[3] Jian Zhou, Weimin Fan, Longdao Chen. Least squares algorithm and its application in improving instrument test accuracy [J]. Journal of Zhejiang University (Natural Science Edition), 1998 (No. 6).

[4] Liangzhang Deng. Fitting and Application of Least Square Method [J]. Journal of Lanzhou Institute of Education, 2012, 28 (08): 109-110+131.

[5] Dexiu Hu, Pan Guo, Shiyi Chen, Lin Cheng, Zhiming Zhao, Li Ran.Monitoring Data Analysis Method Based on Minimum Cut Square Sum Estimation [J]. Mathematical Statistics and Management, 2017 (Phase 4).

[6] Pinghong Zhou, Zhifu Li. Application of Moving Least Square Method in Elevation Anomaly Fitting [J]. Global Positioning System, 2018, 43 (01): 85-90.

[7] Songgui Wang ET al. Linear regression and variance analysis of linear statistical model [M]. Beijing: Higher Education Press. 1999.