Moment of inertia based simple linear regression

Xu Li-Li
School of Economics and Management, Henan Polytechnic University, Jiaozuo
Henan 454000, China

ABSTRACT

Variables on economic activities are of random and correlation generally, while ordinary least square regression (OLSR) can not reflect such data. The reason is OLSR defines only the dependent variable as the one-dimensional random variable, without considering the random variable, and the regression results are affected by the chosen coordinate. To improve it and reflect real economic variables, a new independent-coordinated linear regression method -- the rotating inertia method is presented, which is based the dynamics nature of rigid plate fixed-axis rotation. The simulation example verified and compared with the least squares method to conventional method regression system deviation is small, the regression accuracy. Moment of inertia method is a numerical solution, the computation of the advantages make it has a broad application prospect.

KEYWORDS

Simple linear regression (SLR); Ordinary least squares regression (OLSR); Moment of inertia, independence of coordinates.
INTRODUCTION

Linear regression analysis\[^{1,2}\] is one of the most basic research method in mathematical statistics, and usually used to study the variables correlation. In the field of social economy, the relationship between many variables is not linear even on the macro level, while can be approximated as linear processing on the micro level. Moreover, sometimes if the variables are pretreated logarithmically, the nonlinear relationship between variables can be transformed to the linear relationship. At present, the main analysis of the statistical numerical calculation software based on matrix computation. Therefore, it’s useful to make linear regression variables with high precision.

Based on number of variables dependent variables linear regression can be divided into simple linear regression (one-element linear regression) and multiple regression analysis, in which, simple linear regression is one of the most simple and the most basic questions, summarized as follows. Assume \(x, y\) as variables existing a linear relationship

\[
y = \beta_0 + \beta_1 x + \varepsilon
\]  

In which, \(\beta_i (i = 0,1)\) is constant, \(\varepsilon\) is a random error. \(N\) observations on variables are made, observations are following.

\[
X = (x_1, x_2, \ldots, x_n)'
\]
\[
Y = (y_1, y_2, \ldots, y_n)'
\]

The above data and scattered point sets \(S = \{(x_i, y_i) | i \in [1, n]\}\) are equivalent. The regression line between \(x\) and \(y\) based on above observation data is\[^3\]

\[
\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0
\]  

Matrix form of equation (1) is

\[
Y = (1, X)B + E
\]  

In which

\[
B = (\beta_0, \beta_1)', \quad E = (\varepsilon_1, \ldots, \varepsilon_n)'
\]

Most commonly used solution of a linear regression is the linear regressions based on ordinary least squares regression (OLSR), \(y\) is regarded as the dependent variable, the only random variable, while \(x\) is as the independent variable and is not as random variables. Maximum likelihood estimation for the parameter matrix \(B\) is\[^4\]

\[
\hat{B} = ((1, X)'(1, X))^{-1}(1, X)'Y
\]  

In which \((\hat{\beta}_0, \hat{\beta}_1)\)'

Results of least squares regression (OLSR) are not of coordinate independence. The coordinate independence refers to the operation location coordinates orthogonal transformation (translation and / or rotation) does not affect the result of the operation. As shown in Figure 1, the lines \(L\) and \(L'\) are respectively for the same set of data the regression results in the least squares coordinates \(xOy\) and \(x'O'y'\). \(L\) and \(L'\) are apparently not coincidence.
The author thinks, social economic variables value is seldom "pure" argument without random. When the angle of observation, observation equipment, data definition and summarization method is different, the same economic phenomenon may have very different observational data form. But by a linear transformation and even simple coordinate transformation, data often exhibit obvious equivalence. Therefore, to get the results of regression data set having the same equivalence relation, it is necessary to develop the linear with coordinate invariance regression method.

**A LINEAR REGRESSION METHOD BASED ON MOMENT OF INERTIA**

The following two-element linear regression method is inspired by a classical dynamics nature that as rotating shaft is axis of symmetry, the moment of inertia of rigid plate with uniform thickness symmetric shape is to get the minimum value.

Definition1: The paired observation value of variable $x, y$ is expressed as a set of points $P = \{(x_i, y_i)|i \in [1, n]\}$. Regard a given point $x_i, y_i$ as the particle whose mass is 1, then with respect to the line $y = kx + b$ in any plane the moment of inertia of the point is

$$J = \sum_{i=1}^{n} \frac{(y_i - kx_i - b)^2}{1 + k^2}$$  \hspace{1cm} (5)

Get the minima of $J (k,b)$

$$\begin{cases} 
    k = \frac{-G + \sqrt{G^2 + 4F^2}}{2F} \\
    b = \bar{Y} - k\bar{X}
\end{cases}$$  \hspace{1cm} (6)

In which

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$  \hspace{1cm} (7)

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$  \hspace{1cm} (8)

$$F = \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})$$  \hspace{1cm} (9)

$$G = \sum_{i=1}^{n} [(x_i - \bar{X})^2 - (y_i - \bar{Y})^2]$$  \hspace{1cm} (10)
In the formula (2), assume
\[
\begin{aligned}
\hat{\beta}_1 &= k \\
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}
\end{aligned}
\tag{11}
\]

The above one-element linear regression method is called “one-element linear regression method based on moment of inertia”, referred to as the Moment of Inertia Based Regression, MIBR. According to the definition of MIBR, results with coordinate independence can be obtained.

**THE SIMULATION EXPERIMENT**

Comparing MIBR to least square method can prove its advantages. To simplify the calculation, the mathematical model\(^7\) is constructed, as shown in Figure 2.

a. Assume \(D_1, \ldots, D_5\) independent each other, and \(D_i \sim N(0, 0.4), d_i\) is the observed values of random variables \(D_i, i = 1, \ldots, 5\).

b. Assume \(x, y\) as the variables, the observation vector respectively \(X = (x_1, x_2, x_3, x_4, x_5)'\), 
\(Y = (y_1, y_2, y_3, y_4, y_5)'\)
In which
\[
\begin{aligned}
x_i &= 5 + i - \frac{\sqrt{5}}{2} d_i \\
y_i &= 6 + i + \frac{\sqrt{2}}{2} d_i
\end{aligned}
\tag{12}
\]

Obviously, the theoretical relationship of variables \(x, y\) is considered
\[y = 1 + x + \varepsilon\]

(13)

The above data were calculated respectively by MIBR and the least square regression method, and the corresponding regression errors were got by comparing calculated results with the formula (13). When the regression line angle is large, small changes in the position of the line will cause dramatic changes of the slope, so defining \(\beta_1\) as a measurement variable of error is not appropriate, this paper adopts \(\Delta \alpha\) (error of regression line angle) as regression error metrics. Additionally, since \(\beta_0\) (the intercept of the regression line) and \(\beta_1\) (or \(\alpha\)) is not independent, it’s not of practical significance to analysis their errors separately. Therefore, only one index is defined to measure the performance of each regression method. Making \((X, Y)\) 30 independent observations and linear regression with the two methods, the regression line is obtained, whose angle is as shown in TABLE 1, \(\overline{\Delta \alpha}\) (the average error of three kinds of regression methods) and \(|\Delta \alpha|\) (the mean absolute error) are as shown in TABLE 2, among them
\[
\overline{\Delta \alpha} = \frac{1}{30} \sum_{i=1}^{30} (\alpha_i - 45\,')
\tag{14}
\]
\[
|\overline{\Delta \alpha}| = \frac{1}{30} \sum_{i=1}^{30} |\alpha_i - 45\,'| \tag{15}
\]
TABLE 1: Obliquity data of unary linear regressions (°)

<table>
<thead>
<tr>
<th>OLSR</th>
<th>MIBR</th>
<th>OLSR</th>
<th>MIBR</th>
<th>OLSR</th>
<th>MIBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>36.4614</td>
<td>36.7430</td>
<td>51.1583</td>
<td>52.5040</td>
<td>34.3672</td>
<td>36.5952</td>
</tr>
<tr>
<td>44.9712</td>
<td>46.0789</td>
<td>42.2432</td>
<td>43.4865</td>
<td>44.1101</td>
<td>45.4543</td>
</tr>
<tr>
<td>39.6035</td>
<td>47.3888</td>
<td>57.4840</td>
<td>57.6469</td>
<td>52.4824</td>
<td>57.6470</td>
</tr>
<tr>
<td>29.3945</td>
<td>30.1362</td>
<td>36.8264</td>
<td>38.4114</td>
<td>48.5504</td>
<td>50.4036</td>
</tr>
<tr>
<td>53.9081</td>
<td>54.6807</td>
<td>49.1087</td>
<td>55.8194</td>
<td>33.3938</td>
<td>35.8162</td>
</tr>
<tr>
<td>41.9061</td>
<td>44.3928</td>
<td>46.9750</td>
<td>49.2606</td>
<td>42.2317</td>
<td>44.1155</td>
</tr>
<tr>
<td>47.1428</td>
<td>50.2315</td>
<td>43.0093</td>
<td>44.6703</td>
<td>26.8489</td>
<td>27.5804</td>
</tr>
<tr>
<td>31.7569</td>
<td>33.8584</td>
<td>35.8210</td>
<td>36.1897</td>
<td>44.7187</td>
<td>47.1474</td>
</tr>
<tr>
<td>42.4422</td>
<td>43.0127</td>
<td>46.0340</td>
<td>52.6152</td>
<td>40.3706</td>
<td>41.4210</td>
</tr>
<tr>
<td>51.9599</td>
<td>52.1456</td>
<td>35.0894</td>
<td>36.0983</td>
<td>48.0692</td>
<td>51.4395</td>
</tr>
</tbody>
</table>

TABLE 2: Angle error data of unary linear regressions (°)

<table>
<thead>
<tr>
<th>OLSR</th>
<th>MIBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δα</td>
<td>-2.3854</td>
</tr>
<tr>
<td></td>
<td>6.2436</td>
</tr>
</tbody>
</table>

The absolute value of $\overline{\Delta \alpha}$ (mean angle error) by MIBR is smaller than by OLSR, as shown in TABLE 2, which means MIBR has unbiased advantages.

APPLICATION EXAMPLES

Data on the annual reports of the listed coal enterprises in China are collected, in which those non-fixed assets value being between RMB25 000 000 000 yuan and RMB150 000 000 000 yuan are shown in TABLE 3.

TABLE 3: Data in some China listed coal corporates’ annual reports

<table>
<thead>
<tr>
<th>Corporate name</th>
<th>Year</th>
<th>Main business income p(RMB)</th>
<th>Non-fixed capital k(RMB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shenhua shares</td>
<td>2012</td>
<td>27985000000</td>
<td>26417400000</td>
</tr>
<tr>
<td>Lu’an Environmental</td>
<td>2011</td>
<td>22426300000</td>
<td>27582110000</td>
</tr>
<tr>
<td>Energy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lu’an Environmental</td>
<td>2012</td>
<td>13980400000</td>
<td>32268490000</td>
</tr>
<tr>
<td>Energy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jizhong energy</td>
<td>2012</td>
<td>30072400000</td>
<td>26613100000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yanzhou coal</td>
<td>2009</td>
<td>21500352215</td>
<td>45172821500</td>
</tr>
<tr>
<td>Yanzhou coal</td>
<td>2010</td>
<td>34844400000</td>
<td>54495300000</td>
</tr>
<tr>
<td>Yanzhou coal</td>
<td>2011</td>
<td>48768300000</td>
<td>76592900000</td>
</tr>
<tr>
<td>Yanzhou coal</td>
<td>2012</td>
<td>59673500000</td>
<td>96623500000</td>
</tr>
<tr>
<td>China Coal Energy</td>
<td>2006</td>
<td>28346700000</td>
<td>36712800000</td>
</tr>
<tr>
<td>China Coal Energy</td>
<td>2007</td>
<td>36823300000</td>
<td>41069800000</td>
</tr>
<tr>
<td>China Coal Energy</td>
<td>2008</td>
<td>52282566000</td>
<td>72945635000</td>
</tr>
<tr>
<td>China Coal Energy</td>
<td>2009</td>
<td>53729503000</td>
<td>83103856000</td>
</tr>
<tr>
<td>China Coal Energy</td>
<td>2010</td>
<td>71268400000</td>
<td>92494400000</td>
</tr>
<tr>
<td>China Coal Energy</td>
<td>2011</td>
<td>88872400000</td>
<td>129312600000</td>
</tr>
<tr>
<td>China Coal Energy</td>
<td>2012</td>
<td>87291700000</td>
<td>143319700000</td>
</tr>
<tr>
<td>China Shenhua Energy</td>
<td>2007</td>
<td>82107000000</td>
<td>121975000000</td>
</tr>
<tr>
<td>China Shenhua Energy</td>
<td>2008</td>
<td>107133000000</td>
<td>146466000000</td>
</tr>
<tr>
<td>China Shenhua Energy</td>
<td>2009</td>
<td>121312000000</td>
<td>164152000000</td>
</tr>
</tbody>
</table>
Scatter diagram of data shown in TABLE 3 is expressed in Figure 2. All data were divided into two groups, of which, scatters are marked “*” and made regression analysis if the non-fixed asset value is less than the RMB100 000 000 000 yuan, the others are marked “O” and used for verification of the results of the regression.

Data of group “*” are made least squares regression, the result is

\[ \hat{p}_1 = 58.3537 + 0.5980k \]  

(16)

The unit of variable in above formula is RMB100 million yuan. Regression line of OLSR is dotted line as shown in Figure 2.

While data of group “*” are made regression by MIBR, the result is

\[ \hat{p} = 41.4890 + 0.6288k \]  

(17)

The unit of variable in above formula is RMB100 million yuan. Regression line of MIBR is the solid line as shown in Figure 2.

To compare regression quality of two kinds of regression method, data of non-fixed assets value being more than RMB100 000 000 000 yuan in TABLE 3 are used to validate regression effect. Based on the maximum likelihood hypothesis, considering main business income of group “O” as theoretical value, the error of the predictive value compared with the theoretical value is calculated on the basis of the two kinds of regression. The regression error of OLSR is recorded as \( \epsilon_1 \), while the regression error of MIBR is taken as \( \epsilon_2 \), the mean regression error is denoted as \( \bar{\epsilon} \), and the mean absolute value of regression errors is written as \( \overline{\epsilon} \). The calculation of regression error results as shown in TABLE 4.

The mean regression error of OLSR

\[ \bar{\epsilon}_1 = -71.609 \]

The absolute mean regression error of OLSR
The mean regression error of MIBR

\[ \bar{e}_1 = 88.612 \]

The absolute mean regression error of MIBR

\[ e_2 = 45.048 \]

The mean regression error \( \bar{e} \) can be considered system deviation, according to \( \bar{e}_2 < \bar{e}_1 \), the system deviation by MIBR is narrower than by OLSR.

**TABLE 4** : Errors of regressions

<table>
<thead>
<tr>
<th>p(亿元)</th>
<th>888.724</th>
<th>872.917</th>
<th>821.070</th>
<th>1071.330</th>
<th>1213.12</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{p} )</td>
<td>831.662</td>
<td>915.426</td>
<td>787.782</td>
<td>934.241</td>
<td>1040.006</td>
</tr>
<tr>
<td>( e_1 )</td>
<td>-57.062</td>
<td>42.509</td>
<td>-33.288</td>
<td>-137.089</td>
<td>-173.114</td>
</tr>
<tr>
<td>( \hat{p} )</td>
<td>854.610</td>
<td>942.687</td>
<td>808.471</td>
<td>962.471</td>
<td>1073.681</td>
</tr>
<tr>
<td>( e_2 )</td>
<td>-34.114</td>
<td>69.770</td>
<td>-125.990</td>
<td>-108.859</td>
<td>-139.439</td>
</tr>
</tbody>
</table>

The absolute mean regression error can be regarded as standard deviation of the regression error, according to \( \bar{e}_2 < \bar{e}_1 \), standard deviation of the regression error by MIBR is smaller than by OLSR. In another word, MIBR is more stable than OLSR.

**CONCLUSION**

Both the simulation experiment and regression examples have proved MIBR is more precise and stable than OLSR in one-element linear regression. Solution of MIBR is analytical and demands less computational complexity, so MIBR could be used broadly.

This paper only discusses the application of MIBR in one-element linear regression, as for its application in a multiple regression will be discussed in another paper.

**ACKNOWLEDGEMENTS**

This paper is funded by national natural sciences fund youth fund of the People’s public of China (51274087), Projects of national natural science foundation of the People’s public of China (51104055), Doctor fund of Henan Polytechnic University (B2012-086), the views expressed are authors’ alone.

**REFERENCES**


