Efficiency Improvement for Ordinary Least Square and Orthogonal Regression - An Application in Chemical Engineering

Khurshid A. Bhat

1 State Education Department, Kashmir, Jammu and Kashmir, India.

Abstract

Regression analysis plays an indispensable role in QSAR/QSPR, chemical Engineering, science & technology, and research projects. Best fit regression models are constantly a challenge to the researchers, efforts are taken to minimize the error components so that the predictability and efficiency of models increase. Presence of high error component eventually upset the future research and forecasting of the facts. In this paper a technique is introduced that reduces the error component and improves the predictability and efficiency of the model.

Keywords: Regression analysis; internal variable relation; efficiency; orthogonal regression.

1. INTRODUCTION

Regression analysis plays an important role in engineering fields, science & technology, and other related fields. Many methods are used to fit the best models [1,2]. In case of linear regression models, Ordinary Least Square (OLS), Orthogonal Regression (OR) and Geometric Mean Regression (GMR) methods are extensively used and have seen a fair share of its applications in Aerosol sciences [3]; geology [4]; dietary assessment [5]; bioinformatics [6];
social science [7] and physics [3]. OLS method assumes that errors are confined to the dependent variable, while as OR is on the standard linear regression method to correct for the effects of measurement error in predictor. Different types of orthogonal regression models are available depending on different assumptions [8,9]. The method of OR has a long and distinguished history in statistics and economics. The method, which involves minimizing the perpendicular distance between the observations and the fitted line, has been viewed as superior to OLS in two different contexts. Firstly, the independent and dependent variables in a two-variable linear regression cannot be predetermined because of the minimizing of perpendicular distance do not depend on a specific axis [10-12]. Secondly, when used, there are errors in the independent variables called the errors-in-variables mode [13]. In the present study an internal linear combination method is introduced that increases the efficiency of the model by reducing the sum of square error (SSE) and improves R².

2. METHODS

Let us assume the two variable regression model

\[ Y + \varepsilon_y = a + b(X + \varepsilon_x) + \mu \]

Here \( \varepsilon_y \) and \( \varepsilon_x \) are measurement errors of Y and X both with mean zero, 'a' is the intercept, 'b' is the slope and '\( \mu \) is the equation error with zero mean.

Introduce a internal linear combination of the variables i.e

Let \( P = \{(X_i + X_{i+1})/2\} \) and \( Q = \{(Y_i + Y_{i+1})/2\} \)

This relationship reduced the error sum of square and improves the efficiency of the model.

2.1 Theorem

If \((X,Y)\) is bi-variate data set and \( V(X) \), \( V(Y) \), \( r_{xy} \) are the variances and correlation coefficient of X and Y then for the linear combination \( P = \{(X_i + X_{i+1})/2\} \) and \( Q = \{(Y_i + Y_{i+1})/2\} \), \( V(P) \leq V(X) \), \( V(Q) \leq V(Y) \) and \( r_{pq} \geq r_{xy} \).

2.2 Proof

Let \((X,Y)\) is a bi-variate data set having 'n' observations.

Let,

\[ P = \{(X_i + X_{i+1})/2\} \) and \( Q = \{(Y_i + Y_{i+1})/2\} \) be the two varaites with 'm' number of observations \((m < n)\).

\[ E(P) = E((X_i + X_{i+1})/2) \]
\[ E(P) = (n/m)E(X) – (1/2m)(X_i + X_{i+1}) \]
\[ V(P) = E(P^2) – E(P)^2 \]
\[ E(Q) = E(Y) \]
\[ V(Q) = (n/2m)V(Y) \]
\[ Cov(P,Q) = (n/2m)Cov(X,Y) \]
\[ \sum_{i=1}^{m} \{ \sum_{j=1}^{n} a_{ij}x_{ij} \} – \{E(P)E(Q)\} \]

Now,

\[ Cov(P,Q) = E(PQ) – \{E(P)\} \{E(Q)\} \]
\[ Cov(P,Q) = (n/2m) - \{E(P)\} \{E(Q)\} \]

2.3 Special Case

If,

\[ X_i = Y_i = k \] (any constant value) \]
\[ E(P) = E(Q) = k \]
\[ V(P) = (n/2m)V(X) \]
\[ V(Q) = (n/2m)V(Y) \]
\[ Cov(P,Q) = (n/2m) Cov(X,Y) \]

The factor \((n/2m)\) is always less than one.

Hence, \( V(P) \leq V(X) \), \( V(Q) \leq V(Y) \)

Correlation Co-efficient \((P,Q) \) \((r_{pq}) \) \( \geq \) correlation Co-efficient \((X,Y) \) \((r_{xy}) \).

Using this linear combination, the co-efficient of correlation is improved, consequently reduces the error sum of squares and increase \( R^2 \). An ideal quantitative structure – property or structure – activity relationship and test the three different models least square (LS) orthogonal regression (OR1 and OR2). Here 'X' is a descriptor (a connectivity \( \chi \) index) and 'Y' is a property or activity (P) [9].

Now apply the above internal linear combination (Table 2), \( P = \{(X_i + X_{i+1})/2\} \) and \( Q = \{(Y_i + Y_{i+1})/2\} \) the original data is changed and the new set of data is formed then use same models, the regression equations are depicted in Table 3.

The graphical representation of the predicted values before and after the method is in (Fig. 3a, b), it clearly shows a good difference, the predicted values after applying the methods are more consistent than the previous one.
Fig. 1. Original data and the three regression lines by using three models (LS, OR1 and OR2)

Fig. 2. Original data and the three regression lines by using three models (LS, OR1 and OR2) after applying the new method
Fig. 3. (a) Predicted values before (b) Predicted values after

Table 1. Experimental data

<table>
<thead>
<tr>
<th>X</th>
<th>0.86</th>
<th>1.57</th>
<th>2.53</th>
<th>4.32</th>
<th>6.13</th>
<th>7.42</th>
<th>9.19</th>
<th>10.47</th>
<th>12.65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y&lt;sub&gt;exp&lt;/sub&gt;</td>
<td>0.22</td>
<td>0.82</td>
<td>1.22</td>
<td>1.24</td>
<td>3.96</td>
<td>2.49</td>
<td>4.38</td>
<td>2.90</td>
<td>3.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>13.25</th>
<th>15.43</th>
<th>15.96</th>
<th>16.25</th>
<th>18.24</th>
<th>18.53</th>
<th>20.07</th>
<th>21.97</th>
<th>25.56</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y&lt;sub&gt;exp&lt;/sub&gt;</td>
<td>6.46</td>
<td>7.85</td>
<td>4.80</td>
<td>6.53</td>
<td>6.42</td>
<td>6.43</td>
<td>10.35</td>
<td>10.15</td>
<td>14.41</td>
</tr>
</tbody>
</table>

Table 2. Regression lines

<table>
<thead>
<tr>
<th>Method</th>
<th>Regression equation (I)</th>
<th>SSR</th>
<th>R&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>Y&lt;sub&gt;cal&lt;/sub&gt; = 0.4760X – 0.6050</td>
<td>33.66355</td>
<td>86.14</td>
</tr>
<tr>
<td>OR1</td>
<td>Y&lt;sub&gt;cal&lt;/sub&gt; = 0.4905X – 0.7825</td>
<td>33.85659</td>
<td>86.06</td>
</tr>
<tr>
<td>OR2</td>
<td>Y&lt;sub&gt;cal&lt;/sub&gt; = 0.5129X – 1.0563</td>
<td>34.91823</td>
<td>85.628</td>
</tr>
</tbody>
</table>

Table 3. Regression lines after applying method

<table>
<thead>
<tr>
<th>Method</th>
<th>Regression equation(II)</th>
<th>SSR</th>
<th>R&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>Y&lt;sub&gt;cal&lt;/sub&gt; = 0.4513043X – 0.399455</td>
<td>33.54431</td>
<td>86.194</td>
</tr>
<tr>
<td>OR1</td>
<td>Y&lt;sub&gt;cal&lt;/sub&gt; = 0.4589537X – 0.4926821</td>
<td>33.62053</td>
<td>86.16</td>
</tr>
<tr>
<td>OR2</td>
<td>Y&lt;sub&gt;cal&lt;/sub&gt; = 0.4731X – 0.66</td>
<td>33.82074</td>
<td>86.08</td>
</tr>
</tbody>
</table>

3. CONCLUSION

In this paper a technique is used to minimize the error components in the model so that the predictability and efficiency of models increase. The technique used here is named as internal linear combination technique in which every average pair is consider the new data for the model. The theorem provides the complete proof for the reduction of error in regression model by applying this average method. These techniques can be used in chemical engineering and every field of science and technology where regression models are used.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES


4. Jones TA. Fitting straight lines when both variables are subject to error. I. Maximum likelihood and least square estimation. Mathematical Geology. 1979;11:1-25.


Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle3.com/review-history/46138