



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

**Khurshid A. Bhat<sup>1\*</sup>**

<sup>1</sup>*State Education Department, Kashmir, Jammu and Kashmir, India.*

## **Author's contribution**

*The sole author designed, analysed, interpreted and prepared the manuscript.*

## **Article Information**

DOI: 10.9734/JERR/2019/v4i116893

### Editor(s):

(1) Dr. Felipe Silva Semaan, Professor, Department of Analytical Chemistry, Fluminense Federal University, Niterói, Rio de Janeiro, Brazil.

(2) Dr. Syamsul Bahari Bin Abdullah, Senior Lecturer, Faculty of Chemical and Natural Resources Engineering, Universiti Malaysia Pahang, Malaysia.

### Reviewers:

(1) R. Shenbagavalli, India.

(2) M. M. Jeeva, University of Madras, Chennai, India.

Complete Peer review History: <http://www.sdiarticle3.com/review-history/46138>

**Original Research Article**

**Received 30 October 2018**

**Accepted 08 February 2019**

**Published 07 March 2019**

## **ABSTRACT**

Regression analysis plays 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 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];

\*Corresponding author: Email: [jkstatuk@yahoo.co.in](mailto:jkstatuk@yahoo.co.in);

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 depends 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 pre-determined 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-invariables 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  $R^2$ .

## 2. METHODS

Let us assume the two variable regression model

$$Y + \epsilon_y = a + b(X + \epsilon_x) + \mu$$

Here  $\epsilon_y$  and  $\epsilon_x$  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

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

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,

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

be the two variables with 'm' number of observations ( $m < n$ ).

$$E(P) = E\{(X_i + X_{i+1})/2\}$$

$$E(P) = (n/m)E(X) - (1/2m)(X_1 + X_n)$$

$$V(P) = E(P^2) - \{E(P)\}^2$$

$$V(P) = (n/2m)V(X) + (n/2m)E(X^2) - (1/4m)(X_1^2 + X_n^2) + (1/2m)\{\sum_m X_i X_j\} - \{E(P)\}^2$$

Now,

$$\text{Cov}(P,Q) = E(PQ) - \{E(P)\}\{E(Q)\}$$

$$\text{Cov}(P,Q) = (n/2m) \text{Cov}(X,Y) + (n/2m)E(X)E(Y) - (1/4m)(X_1 Y_1 + X_n Y_n) + (1/4m)$$

$$\{\sum_m (X_i X_j + X_j Y_i)\} - \{E(P)E(Q)\}$$

### 2.3 Special Case

If,

$$X_i = Y_i = k \text{ (any constant value)}$$

$$E(P) = E(Q) = k \text{ and}$$

$$V(P) = (n/2m)V(X)$$

$$V(Q) = (n/2m)V(X)$$

$$\text{COV}(P,Q) = (n/2m) \text{Cov}(X,Y)$$

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

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

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.

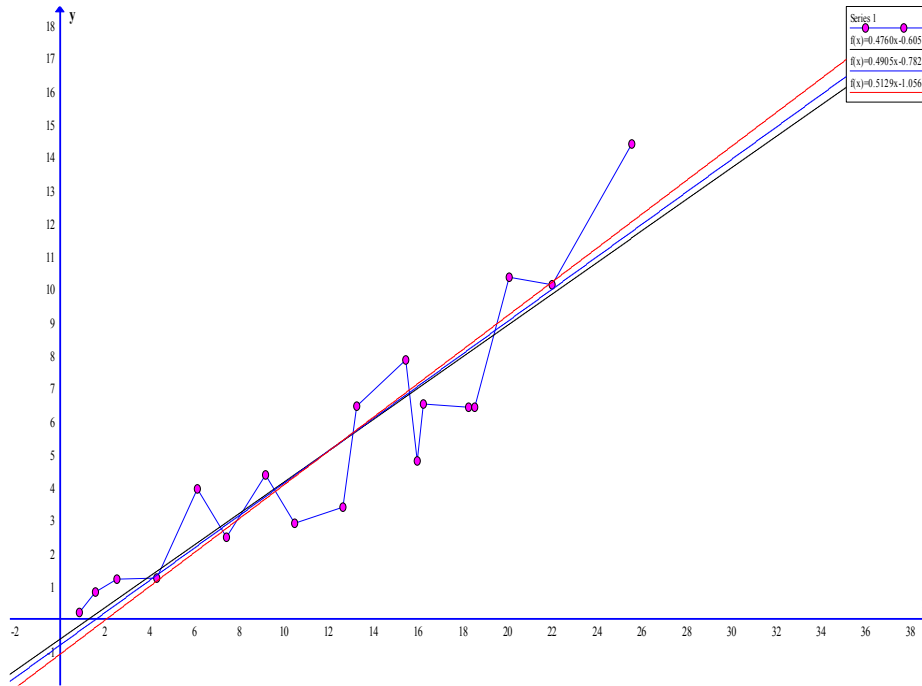


Fig. 1. Original data and the three regression lines by using three models (LS, OR1 and OR2)

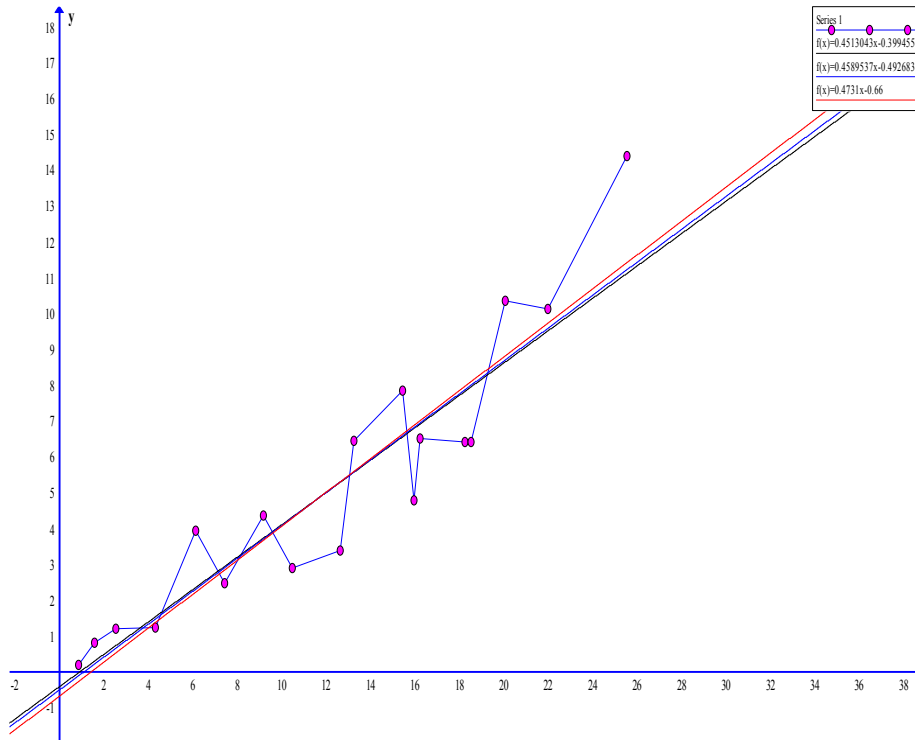


Fig. 2. Original data and the three regression lines by using three models (LS, OR1 and OR2) after applying the new method

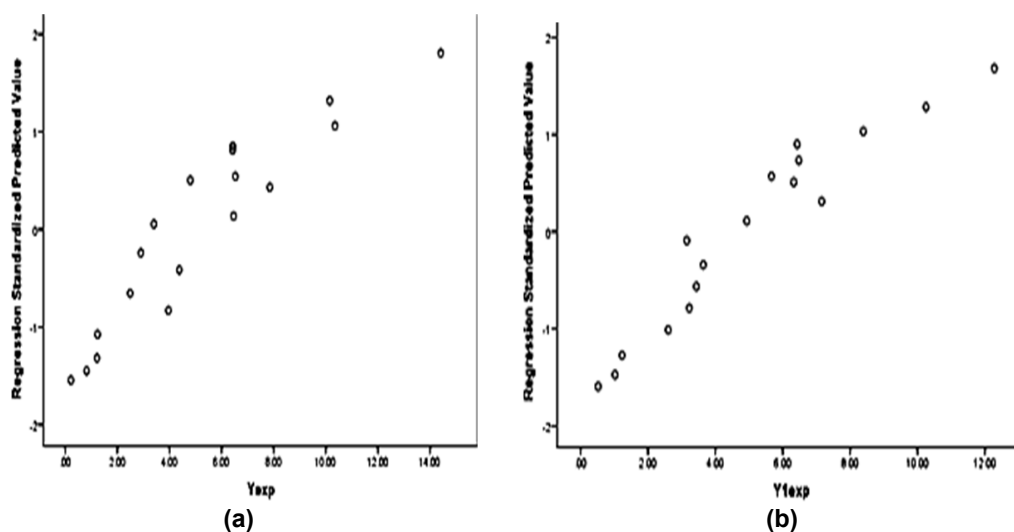


Fig. 3. (a) Predicted values before (b) Predicted values after

Table 1. Experimental data

<b>X</b>	0.86	1.57	2.53	4.32	6.13	7.42	9.19	10.47	12.65
<b>Y<sub>exp</sub></b>	0.22	0.82	1.22	1.24	3.96	2.49	4.38	2.90	3.40
<b>X</b>	13.25	15.43	15.96	16.25	18.24	18.53	20.07	21.97	25.56
<b>Y<sub>exp</sub></b>	6.46	7.85	4.80	6.53	6.42	6.43	10.35	10.15	14.41

Table 2. Regression lines

Method	Regression equation (I)	SSR	R <sup>2</sup>
LS	$Y_{cal} = 0.4760X - 0.6050$	33.66355	86.14
OR1	$Y_{cal} = 0.4905X - 0.7825$	33.85659	86.06
OR2	$Y_{cal} = 0.5129X - 1.0563$	34.91823	85.628

Table 3. Regression lines after applying method

Method	Regression equation(II)	SSR	R <sup>2</sup>
LS	$Y_{cal} = 0.4513043X - 0.399455$	33.54431	86.194
OR1	$Y_{cal} = 0.4589537X - 0.4926821$	33.62053	86.16
OR2	$Y_{cal} = 0.4731X - 0.66$	33.82074	86.08

### 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

- Garcia-Domenech R, de Julian-Ortiz JV, Pogliani L. Some new trends in chemical graph theory. Chem. Rev. 2008;108:1127-1169.
- Sprent P. Models in regression and related topics, Methuen's statistical monographs. Matheun & Co Ltd., London; 1969.

3. Ling Leng, Tianyi Zhang, Lawrence Kleinman, Wei Zhu. Ordinary least square regression, orthogonal regression. Geometric mean regression and their applications in aerosol science. J. Phys. Conference Series. 2007;78:1-5.
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.
5. Freedman LS, et al. Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. American Journal of Epidemiology.
6. Novikov E, Barillot E. An algorithm for automatic evaluation of the spot quality in wo-color DNA microarray experiments, BMC Bioinformatics. 2005;6:293-311.
7. Jackson JD, Dunlevy JA. Orthogonal least squares and the interchangeability of alternative proxy variables in the social sciences. The Statistician. 1988;37(1): 7-14.
8. Besalu E, Julian-Ortiz JV de, Pogliani L. Trend and plot methods in MLR studies. J. Chem. Inf. Model. 2007;47:751-760.
9. Besalu E, Julian-Ortiz JV de, Pogliani L. Ordinary and orthogonal regression in QSAR/QSPR and chemistry-related studies. MATCH Commun. Math. Comput. Chem. 2010;63:573-583.
10. Smyth DJ, Boyes WJ, Peseau DE, The measurement of firm size: Theory and evidence for the United States and the United Kingdom. The Review of Economics and Statistics. 1975;57:111-114.
11. Shalit SS, Sankar U. The measurement of firm size. The Review of Economics and Statistics. 1977;59:290-298.
12. Reza AM. Geographical differences in earnings and unemployment rates, The Review of Economics and Statistics. 1978; 60:201-208.
13. Fuller WA. Measurement error models, New York: John Wiley; 1987.

© 2019 Bhat; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

*Peer-review history:*

*The peer review history for this paper can be accessed here:*

*<http://www.sdiarticle3.com/review-history/46138>*