EFFICIENCY IMPROVEMENT FOR ORDINARY LEAST SQUARE AND ORTHOGONAL REGRESSION - AN APPLICATION IN CHEMICAL ENGINEERING

Khurshid A. Bhat
State Education Department, Kashmir, J&K, India.
e-mail jkstatuk@yahoo.co.in

UNDER PEER REVIEW

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.

Key Word: Regression analysis, internal variable relation, efficiency, orthogonal regression.

Introduction:

Regression analysis plays important role in engineering fields, science & technology and other related fields. Many methods are used to fit the best models. 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 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.

Method

Let us assume the two variable regression model

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

Here ε_y and ε_x are measurement errors of Y and X both with mean zero, 'a' is the intercept, 'b' is the slope and 'μ' 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.

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 $r_{pq} \geq r_{xy}$.

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).

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

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

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

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

Now

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

$$Cov(P,Q) = (n/2m)Cov(X,Y) + (n/2m)E(X)E(Y) - (1/4m)(X_1Y_1 + X_nY_n) + (1/4m)$$

$$\{\textstyle\sum_m (XiXj + XjYi)\} - \{E(P)E(Q)\}$$

Special case

If $X_i = Y_i = k$ (any constant value)

$E(P) = E(Q) = k$ and $V(P) = (n/2m)V(X)$

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

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) 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 χ index) and 'Y' is a property or activity (P) [9].

UNDER PEER REVIEW

Table I (Experimental Data)

X	0.86	1.57	2.53	4.32	6.13	7.42	9.19	10.47	12.65
Y_{exp}	0.22	0.82	1.22	1.24	3.96	2.49	4.38	2.90	3.40

X	13.25	15.43	15.96	16.25	18.24	18.53	20.07	21.97	25.56
Y_{exp}	6.46	7.85	4.80	6.53	6.42	6.43	10.35	10.15	14.41

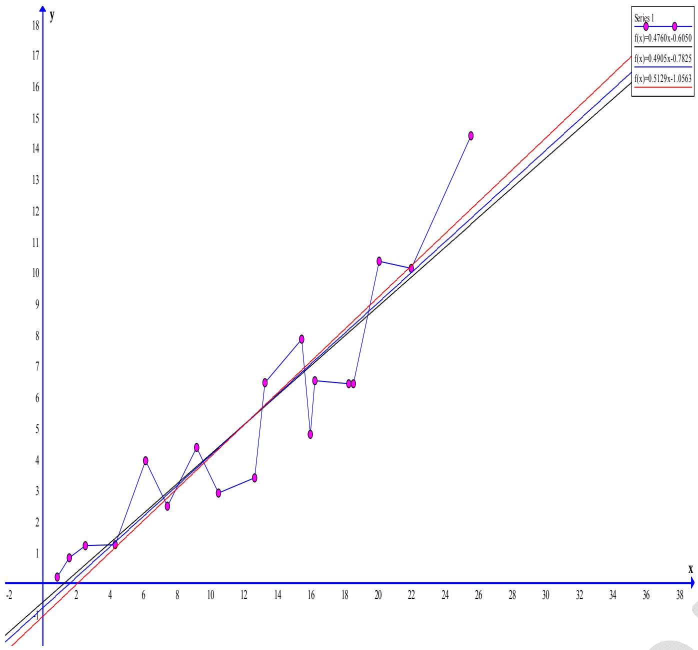
Table II (Regression lines)

Method	**Regression Equation (I)**	**SSR**	R^2
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

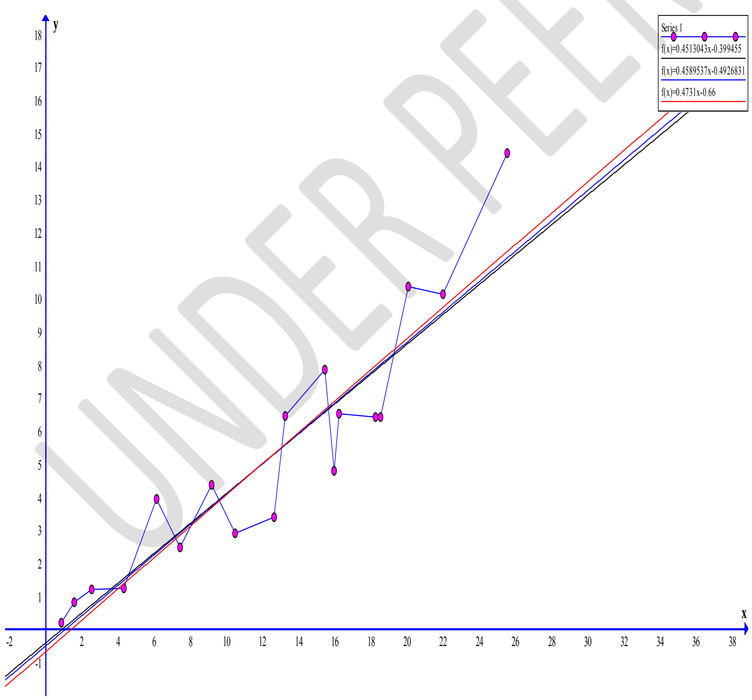
Now apply the above internal linear combination, P = $\{(X_i + X_{i+1})/2\}$ and Q = $\{(Y_i + Y_{i+1})/2\}$the original data is changed and the new set of data is formed then use same models, the regression equations are as under:

Table III (Regression lines after applying method)

Method	**Regression Equation(II)**	**SSR**	R^2
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**



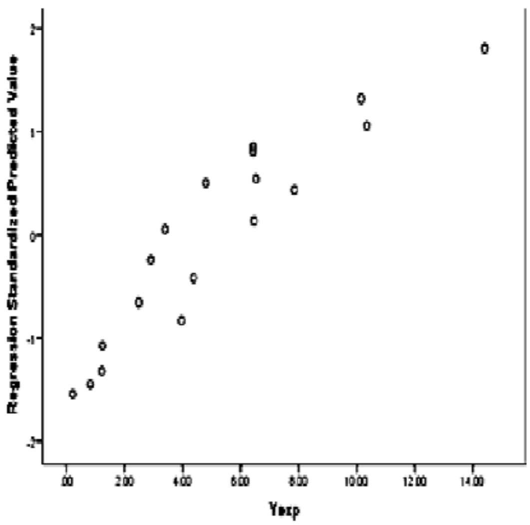
The original data and the three regression lines by using three models (LS, OR1 and OR2)



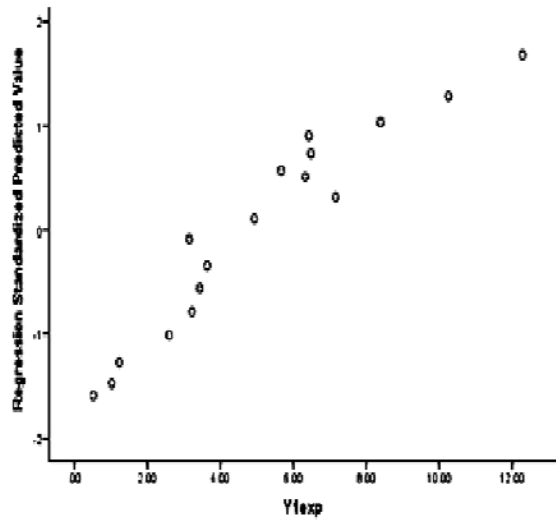
The original data and the three regression lines by using three models (LS, OR1 and OR2) after applying the new method.

UNDER PEER REVIEW

The graphical representation of the predicted values before and after the method is in Fig. (a, b), it clearly shows a good difference, the predicted values after applying the methods are more consistent than the previous one.



(a)



(b)

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.

UNDER PEER

References:

[1] R. Garcia-Domenech, J. V. de Julian-Ortiz, L. Pogliani, "Some new trends in chemical graph theory", Chem. Rev. 108 (2008) 1127-1169.

[2] P. Sprent. “Models in Regression and Related Topics”, Methuen's Statistical Monographs. Matheun & Co Ltd, London, 1969.

[3] Ling Leng, Tianyi Zhang, Lawrence Kleinman and Wei Zhu. “Ordinary least square regression, orthogonal regression, geometric Mean Regression and their applications in Aerosol Science”, J. Phys. Conference Series 78 (2007) 1-5.

[4] T. A. Jones, “Fitting straight lines when both variables are subject to error. I. Maximum likelihood and least square estimation”, Mathematical Geology 11(1979), 1-25.

[5] L. S. Freedman, et al. “Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake”, American Journal of Epidemiology.

[6] E. Novikov, E. Barillot, “An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments”, BMC Bioinformatics 6 (2005) 293-311.

[7] J. D. Jackson and J. A. Dunlevy. “Orthogonal Least Squares and the Interchangeability of Alternative Proxy Variables in the Social Sciences”. The Statistician, 37(1): (1988) 7-14,.

[8] E. Besalu, J. V. de Julian-Ortiz, L. Pogliani, “Trend and plot methods in MLR studies”, J. Chem. Inf. Model. 47 (2007) 751-760.

[9] E. Besalu, J. V. de Julian-Ortiz, L. Pogliani, “ Ordinary and orthogonal regression in QSAR/QSPR and chemistry–related studies”, MATCH Commun. Math. Comput. Chem. 63(2010) 573-583.

[10] D. J. Smyth., W. J. Boyes, and D. E. Peseau, “The measurement of firm size: theory and evidence for the united states and the united kingdom”, The Review of Economics and Statistics 57, (1975) 111-114.

[11] S. S. Shalit, and U. Sankar, “The Measurement of Firm Size”, The Review of Economics and Statistics 59, (1977) 290-298.

[12] A. M. Reza, “Geographical Differences in Earnings and Unemployment Rates”, The Review of Economics and Statistics 60, (1978) 201-208.

[13] W. A. Fuller, “Measurement Error Models”, New York: John Wiley, (1987).

UNDER PEER REVIEW