ChE classroom

A ROBUST ALTERNATE TO LEAST SUM OF SQUARES FOR LINEAR REGRESSION

G. P. RANGAIAH
National University of Singapore
Singapore 0511, Singapore

G. P. Rangaiah received his Bachelors, Masters, and Doctorate degrees, all in chemical engineering, from Andhra University, IIT Kanpur, and Monash University, respectively. He worked in Engineers India Limited, New Delhi, from 1976 to 1978. He has been lecturing at the National University of Singapore since 1982. His research interests are in process simulation, optimization, and control.

Regression, which deals with fitting an appropriate equation to a set of experimental data, is often encountered in engineering and scientific analyses. Of all the regression problems, fitting a linear equation of the form $y = \alpha + \beta x$, containing two parameters (α and β), is the most common one. This problem is discussed in linear regression, while multiple linear regression and nonlinear regression deal with fitting linear equations containing three or more parameters and nonlinear equations respectively. Note that linear, in regression analysis, refers to parameters to be estimated based on experimental data.

LEAST SUM OF SQUARES

Given a set of data (x_i, y_i, for i = 1,2,...,n) and that the equation for representing the data is $y = \alpha + \beta x$, the main task in linear regression is to find the best estimates of α and β. The popular method used for this estimation is the familiar ordinary (or unweighted, or equally-weighted) least sum of squares (LSS). According to this method, estimates of parameters are obtained by minimizing the sum of squares of residuals,

$$\sum_{i=1}^{n}\left[y_i - \left(a + bx_i\right)\right]^2 \tag{1}$$

with respect to a and b. The quantities a and b are the estimates of α and β, respectively, since α and β cannot be found exactly because of inevitable experimental noise.

LSS analysis is almost two centuries old. By solving the minimization problem in Eq. (1), a and b can be obtained without making any assumptions. But for subsequent inferences (*e.g.*, on confidence intervals, goodness of fit), the following assumptions are generally implicit:[1]

- **Equation selected for fitting the data is correct.**
- **Independent variable, x, is free from noise.**
- **Noise in the dependent variable, y, follows normal distribution with mean zero and constant standard deviation (SD).**
- **Noise in y_i and y_j ($i \neq j$) is uncorrelated.**

When these assumptions are satisfied, the LSS estimator has nice features such as unbiasedness and minimum variance; one can also estimate confidence intervals and conduct goodness-of-fit tests.

However, LSS is sensitive to violation of the above assumptions. If one or more of these assumptions are not valid, then the estimates obtained by LSS may be biased and/or may have larger variances. In other words, accuracy and precision of LSS will be low. Although dictionary definitions of accuracy and precision are similar, there is a distinction in their scientific usage.[2] Accuracy and precision refer, respectively, to bias and variance of parameter estimates obtained by a method. (For the definition of bias and variance, see Equations 6 and 7, presented later.) Low accuracy implies larger bias (in absolute value), and low precision means larger variance.

Let us discuss to what extent the above assumptions are likely to be satisfied in practical applications. The first assumption is likely to be valid since knowledge of the physical phenomena relating to the experimental data (and hence the appropriate equation or model) is often available.

The second assumption relates to identifying the independent and dependent variables in a given set of data. The quantity which is free from noise or which has negligible noise should be selected as the independent variable, x. The other quantity, containing relatively large noise, will be the dependent variable or response, y. Frequently, this can be done with reasonable confidence. For example, quantities such as time, temperature, *etc.*, can be measured precisely; hence, one of these quantities should be x. Often, these are the quantities which are adjusted in engineering experiments. On the other hand, quantities such as flow rate, concentration, reaction rate, dimensionless groups, *etc.*, generally contain greater noise—so one of these quantities is a likely candidate for y. Therefore, the second assumption for LSS is often valid.

That leaves us with the last two assumptions. Variance (or square of SD) of noise in y may or may not be constant. One or more experimental points may be outliers or wild points (see references 3 and 4 for the definition of outliers). Or there may be systematic noise because of drift of instruments. Hence, one often does not know whether these two assumptions are valid or not. One might consider fitting data by LSS, and then analyzing residuals [*i.e.*, the differences between y_i and predicted value ($= a + bx_i$)] for the satisfaction of assumptions.[3] However, initial fitting by LSS is affected by outliers, if they are present. This in turn will affect residuals—thus complicating their analysis. This aspect, which is often not realized, will be demonstrated later. Also, analysis of residuals is sometimes qualitative and subjective.

A few points are worth mentioning here. First, if variance of noise in y is not constant, then the weighted least sum of squares is a potential alternate to (ordinary or unweighted) LSS. However, this requires knowledge of variation of SD of noise for determining the correct (relative) weights. Unfortunately, often this detailed information is either not available or is too expensive (in terms of experimental effort and cost) to obtain, thus precluding the use of weighted least sum of squares. Therefore, LSS hereafter refers to unweighted least-squares analysis unless otherwise stated.

Since some of the assumptions of LSS are unlikely to be valid, there is a need for alternative methods which are less affected by violation of the assumptions, or that require less stringent assumptions. These methods are generally known as robust methods.

Second, LSS can be used to obtain parameter estimates (by solving problems similar to Eq. 1) without worrying about the satisfaction of underlying assumptions. Apparently, this is what many users of LSS do. However, in such situations accuracy and precision of estimates obtained (as well as the use of LSS itself) is questionable. Finally, this article is concerned with point estimates rather than interval estimates (such as confidence intervals).

ROBUST METHODS

Since some of the assumptions of LSS are unlikely to be valid, there is a need for alternative methods which are less affected by violation of the assumptions, or that require less stringent assumptions. These methods are generally known as *robust methods*. A robust alternate to LSS should be comparable to LSS (in accuracy and precision) when all the assumptions stated above are satisfied, and it should be better than LSS when one or more assumptions are not valid. Further, the method should preferably be simple in principle and computations so that it has the potential to replace LSS (which can be readily done on many hand-held scientific calculators) in due course.

Many robust methods have been discussed in the literature.[4] One of the promising robust alternatives to LSS is the *median method* (MM), a nonparametric method.[5,6] A graphical version of MM, known as *direct linear plot*, is popular with biochemists.[7] The objective of this paper is to emphasize the underlying assumptions of LSS, to introduce MM, and to show how it compares with LSS for linear regression. All estimation methods can be expected to give nearly identical estimates if noise is negligible or if the number of points in the data set is very large. However, neither of these situations is true in many applications.

MEDIAN METHOD

In order to estimate the two parameters in $y = a + bx$, two points [say (x_i, y_i) and (x_j, y_j)] are sufficient, assuming $x_i \neq x_j$ (and absence of noise). The estimates a_{ij} and b_{ij} are given by

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j} \tag{2}$$

$$a_{ij} = y_i - b_{ij} x_i \tag{3}$$

The subscripts indicate that i^{th} and j^{th} points in the data set are used. These estimates can be termed as preliminary estimates. Generally, there will be more than two points. Hence, for n different x-values, one will have n(n-1)/2 sets of preliminary estimates. (If there are replicates, the number of sets will be less than this number.) The preliminary estimates (of either a or b) are very likely to be different from one another because of noise. In the absence of noise, all preliminary estimates will be identically equal to the true value.

In MM, the median of the above preliminary estimates is then taken as the (final) estimate. On the other hand, estimates of a(b) by LSS is equal to the weighted average of all a_{ij} (b_{ij}), the weights being proportional to $(x_i - x_j)^2$. Hence, in MM, median replaces the weighted average of LSS. Since median is robust towards extreme values compared to weighted average, MM can be expected to be less affected by the violation(s) of the assumptions of LSS.

EXAMPLES

A set of simulated data with $\alpha = 2$ and $\beta = 1$ is presented as Case A in the second column of Table 1; noise in this data set satisfies all the assumptions underlying LSS. The data in the third column (Case B) are similar to those in the second column except that two points (corresponding to $x_i = 8$ and 9) are simulated as outliers.

Application of LSS to the two data sets in Table 1 is well known and need not be described. Use of MM for Case A is briefly described here. In this set of data there are 10 points, and all x_i are different. Hence one can calculate $10 \times 9/2 = 45$ pair-wise intercepts and slopes using Equations 2 and 3 (with i=1, j=2,3,...,10; i=2, j=3,4,...,10; ...; i=9, j=10). If all x_i are not different, then the number of pair-wise estimates will be less than 45. The final estimate of α (or β) by MM is the median of the 45 preliminary estimates of intercept (or slope). The common definition of median is valid. For the data set under consideration, the final estimates of a and b can be found to be 1.66 and 1.03, respectively.

Estimates of α and β obtained by both LSS and MM in Cases A and B are shown in Table 2. For Case A, estimates by LSS are marginally closer to true values than those by MM. However, either method seems to be satisfactory in the ideal situation, wherein the underlying assumptions of LSS are valid.

The results for Case B are shown graphically in Figure 1. The fitted line by LSS is strongly influenced by the two outliers at $x_i = 8$ and 9. This is reflected in estimates of parameters (Table 2). The difference between α and a is about 45%, while that between β and b is about 20%. These large errors in estimates by LSS affect subsequent analysis of residuals.

Residuals are nothing but vertical differences between experimental points and the fitted line (by LSS) in Figure 1, and these are plotted in Figure 2.

TABLE 1
Typical Data With and Without Outliers

x_i	y_i	
	Case A	Case B
1.0	2.68	2.68
2.0	3.74	3.74
3.0	4.79	4.79
4.0	5.76	5.76
5.0	5.60	5.60
6.0	8.54	8.54
7.0	9.08	9.08
8.0	9.80	12.8
9.0	11.2	14.2
10.0	11.0	11.0

TABLE 2
Estimates of α and β by LSS and MM

	Parameter	Estimates by	
		LSS	MM
Case A			
	α	1.72	1.66
	β	1.00	1.03
Case B			
	α	1.12	1.57
	β	1.22	1.08

The largest (absolute) residual corresponds to the good point at $x_i = 10$, and is -2.32. Residuals corresponding to the two outliers at $x_i = 8$ and 9 are 1.92 and 2.1, respectively. Hence it is not easy to distinguish the outliers from the rest of the data. In this situation, one may opt to reject the point with the largest (absolute) residual, which happens to be a good point in the present example. If the point ($x_i = 10$, $y_i = 11$) is rejected and LSS is reapplied to the remaining 9 points, then the estimates of α and β are 0.41 and 1.41, respectively. Now, the error in the estimates is much more. This type of influence of outliers on LSS, in the first place and on subsequent analysis of residuals, is often not realized.

For the data of Case B, estimates obtained by MM are presented in Table 2, and the fitted line is shown in Figure 1 along with the corresponding results for LSS. Estimates obtained by MM are closer to the true values, compared to those by LSS. Further, estimates by MM in Cases A and B are nearly equal. Hence MM, unlike LSS, is almost unaffected by the outliers.

The robust performance of MM over LSS cannot be concluded based on a single set of simulated data. Therefore, LSS and MM are evaluated through Monte Carlo tests. A detailed description of these tests and results obtained on the Arrhenius equation are presented elsewhere.[8] Also, see reference 4 for some results obtained through Monte Carlo tests. Here, Monte Carlo tests are briefly described, and typical results on the linear equation $y = \alpha + \beta x$ are presented below.

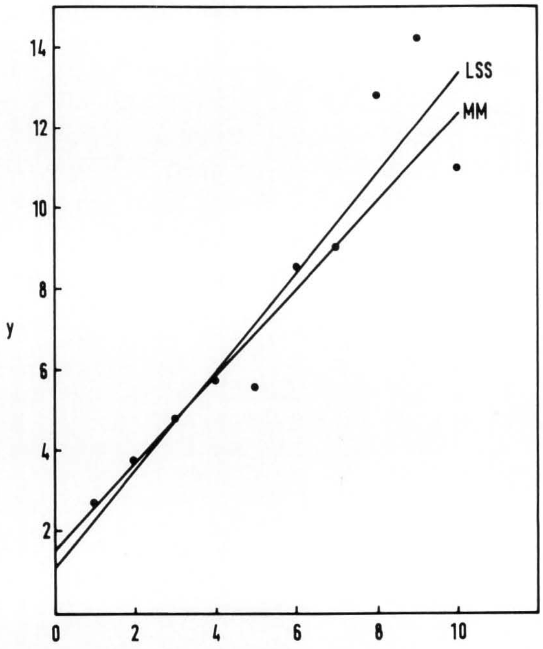

FIGURE 1. *Case B—Data and fitted lines by LSS and MM*

MONTE CARLO TESTS

First of all, let us write the linear equation including noise in the dependent variable, y:

$$y_i = a + b x_i + e_i \tag{4}$$

Noise, e_i, is the origin of all difficulties in finding precise parameter values and the need for a good estimation method. In the tests, noise with different assumptions regarding its distribution is simulated. The steps in the test are as follows:

1. **Assume α, β, n, x_i, and distribution for e_i. In the present tests, α, β, and n are 2, 1, and 10, respectively, while values of x_i are those shown in Table 1.**
2. **Simulate noise, e_i (satisfying the assumed distribution), and then calculate y_i (= $\alpha + \beta x_i + e_i$) for i = 1,2,...,n. Note that any two noises (e_i and e_j, $i \neq j$) will be different.**
3. **For the simulated data (x_i and y_i, i = 1,2,...,n), evaluate a and b by both LSS and MM. The estimates by the two methods are unlikely to be identical.**
4. **In order to obtain the average performance of the methods, repeat steps 2 and 3 many times — say, 4000 times.**

Hence, for each method and for each parameter there will be 4000 estimates; these estimates are likely to be different from one another as well as different from the true (or exact) value. The variation of 4000 estimates is analyzed in terms of mean square error (MSE). For example, mean square error of b is given by

$$\text{MSE of } b = \frac{\sum_{k=1}^{4000} (b_k - \beta)^2}{4000} \tag{5}$$

It can be easily shown that MSE is the sum of square of bias and variance[4], defined as follows

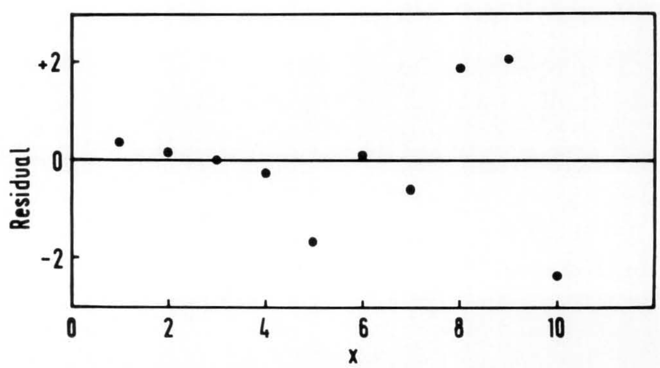

FIGURE 2. *Plot of residuals—Case B and LSS*

$$\text{Bias of } b = \bar{b} - \beta \qquad (6)$$

$$\text{Variance of } b = \frac{\sum_{k=1}^{4000} \left(b_k - \bar{b}\right)^2}{4000} \qquad (7)$$

where $\bar{b}$ is the arithmetic mean of all the 4000 estimates. The bias is usually small, and the main contributor to MSE is the variance. Because of this, a small MSE generally implies higher precision. Therefore, the best (or most precise) method is one having the smallest MSE.

RESULTS AND DISCUSSION

The results of a few Monte Carlo tests are summarized in Table 3. The variation in these tests is essentially in the distribution of noise, which is always assumed to be normally distributed but with a different mean and/or variance. This is because normal distribution is justified in many situations. In the first test in Table 3, SD of noise in y is constant and equals to 0.6, while mean of noise is zero; further, there is no noise in x. Hence, the underlying assumptions of LSS are all satisfied. In the remaining tests, generated noise is such that one assumption of LSS is relaxed or violated for evaluating the robustness of MM compared to LSS.

In tests 2 to 5, SD of noise in y depends on the true value of dependent variable, y^* ($= \alpha + \beta x$); the assumed SD of noise is stated in Table 3. In the sixth test, 2 points (selected at random) out of 10 are simulated as outliers. Mean and SD of noise for outliers are chosen as 3.0 and 0.6, respectively, while the corresponding quantities for regular points are 0.0 and 0.6, respectively. In the last test, unlike tests 1 to 6, noise with mean 0.0 and SD 0.6 is included in y as well as in x. That is, both x and y are subject to noise with the same distribution. However, noise in x_i is neither equal nor correlated to noise in the corresponding y_i.

The results in Table 3 show that in the first test, MSE of either a or b by LSS is lower than that by MM. This is expected since the assumptions of LSS are all satisfied. In tests 2 to 5, in which SD is not constant, MSE for MM is generally smaller than that for LSS. That is, MM is more precise than LSS. MM is also superior to LSS when there are outliers (test 6). When the same noise is present in both x and y (test 7), LSS is better than MM.

Hence, performance of MM is less affected by variation in SD of noise or by outliers. However, when SD is constant or when noise is present in both x and y, MM is marginally inferior to the popular LSS. Results in Table 3 indicate that error in a is generally much more than that in b. This means it is more difficult to estimate the intercept precisely.

TABLE 3
Results of Monte Carlo Tests on the Performance of LSS and MM

Test #	Noise Characteristics (see text for details)	MSE of a		(MSE of b) x 100	
		LSS	MM	LSS	MM
1	SD = 0.6	0.164	0.216	0.421	0.477
2	SD = $0.2y^*$	0.621	0.435	3.34	3.20
3	SD = $3.0 / y^*$	0.259	0.268	0.471	0.397
4	SD = $0.006 (y^*)^2$	0.047	0.012	0.358	0.247
5	SD = $9.0 / (y^*)^2$	0.203	0.095	0.363	0.134
6	2 Outliers	1.19	0.931	2.11	1.56
7	Noise in both x and y	0.342	0.436	0.889	0.993

CONCLUDING REMARKS

Ordinary (or unweighted) least sum of squares has been the workhorse for simple linear regression during the past two hundred years. However, its precision is affected when noise in the dependent variable has variable standard deviation, or when there are outliers. A promising robust method for these situations is the median method, where the principle and computations are simple. This article brings this method to the attention of chemical engineers and presents some results to show its robustness. Interesting research is progressing on regression diagnostics for outlier detection and on robust (or resistant to outliers) regression. The recent book by Rousseeuw[4] describes both these topics, as well as another promising robust method known as *least median of squares.*

REFERENCES

1. Gunst, R. F., and R. L. Mason, *Regression Analysis and Its Application,* Marcel Dekker, New York (1980)
2. Bevington, P. R., *Data Reduction and Error Analysis for Physical Sciences,* McGraw-Hill Book Company, New York (1969)
3. Draper, N. R., and H. Smith, *Applied Regresion Analysis,* John Wiley and Sons, New York (1981)
4. Rousseeuw, P. J., and A. M. Leroy, *Robust Regression and Outlier Detection,* John Wiley and Sons, New York (1987)
5. Sen, P. K., *J. Amer. Statist. Ass.,* **63**, 1379-89 (1968)
6. Hollander, M., and D. A. Wolfe, *Nonparametric Statistical Methods,* John Wiley and Sons, New York (1973)
7. Cornish-Bowden, A., *Fundamentals of Enzyme Kinetics,* Butterworths, London (1979)
8. Rangaiah, G. P., *Chem. Eng. J.,* **29**, 159-166 (1984) ❒