Pergamon

Computers Ops Res. Vol. 25, No. 12, pp. 1137–1143, 1998

Printed in Great Britain
PII: S0305-0548(98)00016-1
0305-0548/98 $19.00 + 0.00

SOLVE LEAST ABSOLUTE VALUE REGRESSION PROBLEMS USING MODIFIED GOAL PROGRAMMING TECHNIQUES

Han-Lin Li†
Institute of Information Management, National Chiao Tung University, Hsinchu 30050, Taiwan, ROC

(Received July 1997; accepted January 1998)

Scope and Purpose—Least absolute value (LAV) regression methods have been widely applied in estimating regression equations. However, most of the current LAV methods are based on the original goal program developed over four decades. On the basis of a modified goal program, this study reformulates the LAV problem using a markedly lower number of deviational variables than used in the current LAV methods. Numerical results indicate that for the regression problems with hundreds of observations, this novel method can save more than 1/3 of the CPU time compared to current LAV methods.

Key words: Goal programming, least absolute value regression

1. INTRODUCTION

Since Charnes *et al.* [1] formulated the least absolute value (LAV) regression problems as linear programs, numerous algorithms have been developed to solve LAV regression problems. Dodge [2] and Dielman [3] thoroughly review these algorithms.

A LAV regression problem can be stated as follows: consider a linear regression model with n observations

$$y_i = \beta_0 + \sum_{j=1}^{m} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

where $(x_{i1}, x_{i2}, \ldots, x_{im}, y_i)$ are the ith observation.

To find estimates of $\beta_0, \beta_1, \ldots, \beta_m$ so that the sum of the absolute values of the differences between the true and fitted y_i values is a minimum. In addition, β_j should satisfy some constraints. That is, choose $b_0, b_1, \ldots, b_m$ to solve the following goal programming problem:

Problem 1:

$$\text{Minimize} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{1}$$

where $\hat{y}_i = b_0 + \Sigma_{j=1}^{m} x_{ij}b_j$ and b_j, $j = 0, 1, 2, \ldots, m$, are variables, which are unrestricted in sign. x_{ij} and y_i are constants for $i = 1, 2, \ldots, n$ and $j = 1, \ldots, m$.

Most LAV regression algorithms [1, 4–6] restate Problem 1 as the following linear programming based on the conventional goal programming techniques [1, 7]:

Problem 2:

$$\text{Minimize} \sum_{i=1}^{n} (d_i^+ + d_i^-)$$

$$\text{subject to } y_i - \left(b_0 + \sum_{j=1}^{m} x_{ij}b_j + d_i^+ - d_i^-\right) = 0, \quad i = 1, 2, \ldots, n, \tag{2}$$

†Tel.: +886-35-72-8709; Fax: +886-35-72-3792; E-mail: hlli@ccsun2.cc.nctu.edu.tw

where d_i^+, $d_i^- \geq 0$ and b_j ($j = 0, 1, \ldots, m$) $\in F$ (F is a feasible set). The d_i^+ and d_i^- is, respectively, the positive and negative deviation variable associated with the ith observation.

The dual problem of Problem 2 is stated below

$$\begin{aligned} &\text{Maximize } \sum_{i=1}^{n}(\rho_i y_i - y_i) \\ &\text{subject to } \sum_{i=1}^{n}\rho_i x_{ij} = \sum_{i=1}^{n} x_{ij}, \quad j = 1, 2, \ldots, m, \\ &\qquad 0 \leq \rho_i \leq 2,\ \rho_i \in F, \quad i = 1, 2, \ldots, n \end{aligned} \tag{3}$$

Armstrong and Kung [8] provided a dual algorithm to solve the LAV regression problem. Their algorithm was slower than algorithms using the primal problem (i.e. Problem 2) unless a "good" dual feasible solution was available.

Many of the timing comparisons of the LAV algorithms are summarized in Refs. [3, 9]. These investigators confer that the primal LP approach proposed by Armstrong *et al.* [4], which is formulated as Problem 2, appears to be best among the LAV regression algorithms.

However, two computational difficulties are encountered in solving Problem 2:

(i) Problem 2, which contains equality constraints, can only be solved by the big-M method or the two-phase method [10]. Both the big-M method and the two-phase method take a markedly longer time than the simplex method in terms of solving linear programming problems [11].

(ii) Problem 2 involves too many deviation variables (i.e. d_i^+ and d_i^-). Assume that there are n observations, then the number of deviation variables becomes $2n$. Such an event causes heavy computational burden for large data sets.

An alternative LAV model is proposed by Gonin and Money [12], which formulates Problem 1 as the following linear programming problem:

Problem 3:

$$\begin{aligned} &\text{Minimize } \sum_{i=1}^{n} d_i \\ &\text{subject to } y_i - \left(b_0 + \sum_{j=1}^{m} x_{ij} b_j - d_i\right) \geq 0, \quad i = 1, 2, \ldots, n, \\ &\qquad y_i - \left(b_0 + \sum_{j=1}^{m} x_{ij} b_j + d_i\right) \leq 0, \quad i = 1, 2, \ldots, n, \\ &\qquad a_i \geq 0 \text{ and } b_j \in F. \end{aligned}$$

Compared with Problem 2, Problem 3 uses half of the deviation variables to reformulate the same Problem. However, the number of constraints in Problem 3 is twice that in Problem 2. Problem 3 is therefore not superior to Problem 2 from the perspective of computational efficiency.

In light of the above developments, this paper reformulates a LAV regression problem as a linear program based on modified goal programming techniques. Advantages of the proposed formulation over the current formulation as expressed in Problem 2 are as follows: first, the proposed formulation can be directly solved by the simplex method instead of by the big-M method or the two-phase method. Second, in terms of solving the regression problem involving n observations, the proposed formulation only requires adding n deviation variables instead of adding $2n$ deviation variables. The proposed method is therefore more convenient to find a good start basis leading to significant efficiencies. Numerical results presented herein demonstrate that the proposed formulation is more computationally efficient than the current formulation.

2. REFORMULATION OF LAV REGRESSION PROBLEMS

Before a linear program is solved by the simplex algorithm, the linear program must be converted into a standard form. Such a standard form is a proper form for Gaussian elimination. Some conditions of this standard form are listed below [13]:

(i) Each constraint is converted into an equation by adding a slack variable and/or a surplus variable. Such a slack variable or surplus variable is served as the initial basic variable.

(ii) Each equation has only one basic variable, which does not appear in the objective function or in any other equation.

(iii) An "$=$" constraint is converted into an equation by adding a surplus variable. For instance, if row i is $ax_1 + bx_2 = c$ $(c \geq 0)$, then it is converted into the following equation

$$ax_1 + bx_2 + u_i = c \quad (u_i \geq 0)$$

(iv) A "$\leq$" constraint is converted into an equation by adding a slack variable. For instance, if row i is $ax_1 + bx_2 \leq c$ $(c \geq 0)$, then it is converted into the following equation

$$ax_1 + bx_2 + s_i = c \quad (s_i \geq 0)$$

(v) A "$\geq$" constraint is converted into an equation by adding a slack variable and a surplus variable. For instance, if row i is $ax_1 + bx_2 \geq c$ $(c \geq 0)$, then it is converted into

$$ax_1 + bx_2 - s_i + u_i = c, \quad s_i \geq 0,\ u_i \geq 0,$$

where s_i is a slack variable and u_i is a surplus variable.

Surplus variables u_i in (iii) and (v) although ultimately having the value of zero, should appear in the standard form to obtain a feasible basic solution. The big-M method or the two-phase method is applied to ensure that these surplus variables will ultimately become zeroes.

Problem 2 is examined as follows. Since d_i^+ and d_i^- appear in the equality constraint as well as in the objective function, either d_i^+ or d_i^- can not be regarded as basic variable [following condition (ii)]. A basic variable cannot appear in the objective function primarily owing to that it may erroneously choose the wrong non-basic variable to enter into the simplex table [11]. In addition, converting Problem 2 into a standard form which is readily solved by the simplex method requires adding a surplus variable to each of its equality constraints.

The big-M method is more commonly used than the two-phase method [10]. The standard form of Problem 2 solvable by the big-M method is as follows:

Problem 4:

$$\text{Minimize } \sum_{i=1}^{n}(d_i^+ + d_i^-) + M\sum_{i=1}^{n}u_i$$

$$\text{Subject to } y_i - b_0 - \sum_{j=1}^{m}x_{ij}b_j - d_i^+ + d_i^- + u_i = 0, \quad i = 1, 2, \ldots, n,$$

$$d_i^+,\ d_i^-,\ u_i \geq 0 \text{ and } b_j\ (j = 0, 1, \ldots, m) \in F\ (F \text{ is a feasible set})$$

where u_i is a surplus variable and M is a large positive number used to ensure u_i close to zero.

The computational difficulties of solving Expression (4) include the following: first, if the chosen M is too small then u_i will not close to zero, possibly leading to an infeasible solution. Second, if the chosen M is too large, the resulting "optimal" solution may not be optimal for the original objective owing to round off error. Winston [11] reported that the big-M method often takes a much longer time to solve a linear program than the common simplex algorithms.

A novel means of reformulating a LAV regression problem is described as follows. By referring to a modified goal program method [14], we have the following proposition:

Proposition 1. *Problem 1 can be converted into the following linear program*

$$\text{Minimize } \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} (y_i - \hat{y}_i + 2\varepsilon_i)$$

$$\text{subject to } y_i \geq \hat{y}_i - \varepsilon_i, \quad i = 1, 2, \ldots, n,$$
$$\varepsilon_i \geq 0, \quad i = 1, 2, \ldots, n. \tag{5}$$

Proof. Denote I^+ as the set of observations where $y_i - \hat{y}_i \geq 0$ and I^- the set of observations where $y_i - \hat{y}_i < 0$.

Then Problem 1 is equivalent to the following program

$$\text{Minimize } \sum_{i \in I^+} (y_i - \hat{y}_i) + \sum_{i \in I^-} (-y_i + \hat{y}_i) \tag{6}$$

Subject to the same constraints in Problem 1.

Next, we prove that Expression (5) is equivalent to Expression (6).

Consider Expression (5), since ε_i only appears in the ith constraint and in the objective function, ε_i is independent of ε_j for $i \neq j$. Two cases are analyzed:

Case 1: For $i \in I^+$ (i.e. $y_i \geq \hat{y}_i$).

Since ε_i needs to be minimized, at the optimal solution ε_i will be zero and z_i becomes $z_i = y_i - \hat{y}_i$.

Case 2: For $i \in I^-$ (i.e. $y_i < \hat{y}_i$).

In order to fit the ith constraint in Expression (5) and to minimize ε_i, at the optimal solution ε_i will be $\varepsilon_i = \hat{y}_i - y_i$ and z_i then becomes $z_i = -y_i + \hat{y}_i$.

Both cases ensure that $\Sigma_{i=1}^{n} z_i$ in Expression (5) is equivalent to the following form:

$$\sum_{i=1}^{n} z_i = \sum_{i \in I^+} (y_i - \hat{y}_i) + \sum_{i \in I^-} (-y_i + \hat{y}_i),$$

which is exact the form of Expression (6). The proposition is then proven.□

To illustrate the usefulness of Proposition 1, consider the following simple data set: $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (2, 2.5)$, $(x_3, y_3) = (2.5, 1.7)$, $(x_4, y_4) = (3, 2.1)$ and $(x_5, y_5) = (4, 3)$.

The relationship between x_i and y_i is assumed as

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, 5.$$

By using the current LAV regression method (Problem 2 and Problem 3), b_0 and b_1 values are obtained by solving Problem 2′ and Problem 3′.

Problem 2′:

$$\text{Minimize } \sum_{i=1}^{5} (d_i^+ + d_i^-)$$

$$\begin{aligned}
\text{subject to } & 1 - (b_0 + b_1 + d_1^+ - d_1^-) = 0 \\
& 1.5 - (b_0 + 2b_1 + d_2^+ - d_2^-) = 0 \\
& 1.7 - (b_0 + 2.5b_1 + d_3^+ - d_3^-) = 0 \\
& 2.1 - (b_0 + 3b_1 + d_4^+ - d_4^-) = 0 \\
& 3 - (b_0 + 4b_1 + d_5^+ - d_5^-) = 0 \\
& d_i^+, d_i^- \geq 0, \quad i = 1, 2, \ldots, 5,
\end{aligned}$$

b_0 and b_1 are un-restricted in sign.

Problem 3′:

$$\text{Minimize } \sum_{i=1}^{5} d_i$$

$$\begin{aligned}\text{subject to } & 1-(b_0+b_1-d_1)\geq 0, \quad 1-(b_0+b_1+d_1)\leq 0\\ & 71.5-(b_0+2b_1-d_2)\geq 0, \quad 1.5-(b_0+2b_1+d_2)\leq 0\\ & 1.7-(b_0+2.5b_1-d_3)\geq 0, \quad 1.7-(b_0+2.5b_1+d_3)\leq 0\\ & 2.1-(b_0+3b_1-d_4)\geq 0, \quad 2.1-(b_0+3b_1+d_4)\leq 0\\ & 3-(b_0+4b_1-d_5)\geq 0, \quad 3-(b_0+4b_1+d_5)\leq 0\\ & d_i\geq 0, \quad i=1,\,2,\,\ldots,\,5,\end{aligned}$$

b_0 and b_1 are un-restricted in sign.

Problem 2′ involves 10 deviation variables and 5 constraints, whereas Problem 3′ contains 5 deviation variables but 10 constraints. By using Proposition 1, this regression model could be simplified as:

Problem 4′:

$$\begin{aligned}\text{Minimize } & -5b_0-12.5b_1+2d_1+2d_2+2d_3+2d_4+2d_5\\ \text{subject to } & 1-(b_0+b_1-d_1)\geq 0, \quad 1.5-(b_0+2b_1-d_2)\geq 0\\ & 1.7-(b_0+2.5b_1-d_3)\geq 0, \quad 2.1-(b_0+3b_1-d_4)\geq 0,\\ & 3-(b_0+4b_1-d_5)\geq 0,\\ & d_i\geq 0,\ i=1,\,2,\,\ldots,\,5,\end{aligned}$$

b_0 and b_1 are unrestricted in sign.

Problem 4′ although equaling Problem 2′ and Problem 3′, only contains 5 deviation variables and 5 constraints. These three problems have the same optimal solution (i.e. $b_0=0.3$ and $b_1=0.6$).

Recall that a constraint with standard form in linear programs should have non-negative right-hand-side value. Consider y_i in Expression (5). By assuming $y_i\geq 0$ for all of i, the "$\leq$" type constraints in Expression (5) can be directly formed as equality constraints by adding slack variables. However, if $y_i<0$ for all of i, then Expression (5) must be changed as follows:

$$\text{Minimize } \sum_{i=1}^{n}\left(-y_i+b_0+\sum_{i=1}^{m}x_{ij}b_j+2d_i\right)$$

$$\text{subject to } -b_0-\sum_{j=1}^{m}x_{ij}b_j-d_i\leq -y_i, \quad i=1,\,2,\,\ldots,\,n,$$

$$d_i\geq 0 \text{ and } b_j\in F. \tag{7}$$

Assume that Problem 1 contains k ($k<n$) observations with $y_i\geq 0$ ($i=1,\,2,\,\ldots,\,k$) and $n-k$ observations with $y_i<0$ ($i=k+1,k+2,\ldots,n$), then Problem 1 can be converted into the following:

$$\text{Minimize } \sum_{i=1}^{k}\left(y_i-b_0-\sum_{j=1}^{m}x_{ij}b_j+2d_i\right)+\sum_{i=k+1}^{n}\left(-y_i+b_0+\sum_{j=1}^{m}x_{ij}b_j+2d_i\right)$$

$$\text{subject to } b_0+\sum_{j=1}^{m}x_{ij}b_j-d_i\leq y_i \quad i=1,\,2,\,\ldots,\,k,$$

$$-b_0-\sum_{j=1}^{m}x_{ij}b_j-d_i\leq -y_i \quad =k+1,\,k+2,\,\ldots,\,n,$$

$$d_i\geq 0\ (i=1,\,2,\,\ldots,\,n) \text{ and } b_j\in F. \tag{8}$$

where y_i are constants, $y_i\geq 0$ for ($i=1,2,\ldots,k$) and $y_i<0$ for $i=k+1,k+2,\ldots,n$.

The standard form of Expression (8) becomes
Problem 4:

$$\text{Minimize } \sum_{i=1}^{k}\left(y_i - b_0 - \sum_{j=1}^{m} x_{ij}b_j + 2d_i\right) + \sum_{i=k+1}^{n}\left(-y_i + b_0 + \sum_{j=1}^{m} x_{ij}b_j + 2d_i\right)$$

$$\text{subject to } b_0 + \sum_{j=1}^{m} x_{ij}b_j - d_i + s_i = y_i, \quad i = 1, 2, \ldots, k,$$

$$-b_0 - \sum_{j=1}^{m} x_{ij}b_j - d_i + s_i = -y_i, \quad i = k+1, k+2, \ldots, n,$$

$$d_i \geq 0,\ s_i \geq 0,\ y_i \geq 0\ (i = 1, 2, \ldots, k),\ y_i < 0\ (i = k+1, k+2, \ldots, n) \text{ and } b_j \in F. \quad (9)$$

where s_i are slack variables used to convert the "$\leq$" type inequality constraints into the standard form solvable by a simplex method.

Comparing Problem 3 in Expression (4) with Problem 4 in Expression (9) yields:

(i) Problem 3 contains $2n$ deviation variable (i.e. d_i^+ and d_i^-), while Problem 4 contains n deviation variables (i.e. d_i).

(ii) Problem 3 must be solved by the big-M method which is more time consuming than the simplex algorithm applied to solve Problem 4.

3. NUMERICAL EXPERIMENTS

In applied regression analysis there has been a trend toward larger and longer data sets. Two types of data sets are generally used in testing: randomly generated and "real world". Most studies involving the comparison of various algorithms for regression have used randomly generated data sets. However, randomly generated data sets are often criticized because these simulated data may exhibit different properties from real world data sets. A problem with the use of real-world data sets, however, is that the number of data sets is frequently quite small. To obtain the advantages of real-world data and randomly generated data, Gentle *et al*. [9] proposed a method of "data based" random number generation due to Taylor and Thompson [15]. In this study, we apply their methods to generate testing data sets to compare the computational efficiency of Problem 3 with Problem 4.

Two real-world data sets used herein can be found elsewhere [16, 17]. Based on these two data sets, different sizes of data sets are generated by the following processes as proposed by Gentle *et al*. [9].

Assume that the (vector-valued) observation x_i are in the rows of the real data set. An observation, x_j, is chosen randomly; its nearest k neighbors, $x_{j1}, x_{j2}, \ldots, x_{jk}$, are determined and the mean, $\bar{x}_j$, of those neighbors is calculated. Next a random sample $u_1, u_2, \ldots, u_k$ is generated from a uniform distribution with lower bound $1/k - \sqrt{3(k-1)/k^2}$ and upper bound $1/k + \sqrt{3(k-1)/k^2}$. The random variable is delivered as

$$\sum_{l=1}^{k} u_l(x_{jl} - \bar{x}_j) + \bar{x}_j.$$

By using the value of random variable delivered herein, many pseudo-observations can be generated as desired. The same process can be used to generate new pseudo-variables as well as observations.

Table 1. Results using published data

			Problem 3		Problem 4	
Data set	n	m	CPU time* (s)	No. of iterations†	CPU time* (s)	iterations†
Weisberg [16]	48	4	4.83	77	3.41	44
Gunst and Mason [17]	60	15	5.28	98	4.22	55

*CPU time is reported by Mathematica [18].
†Number of iterations is reported by LINDO [13].

Table 2. Results using published data to generate random test sets

Data set	n	m	Problem 3		Problem 4	
			CPU time* (s)	No. of iterations†	CPU time* (s)	No. of iterations†
Weisberg [16]	100	4	18.51	157	13.24	69
	200	4	90.85	273	66.08	171
	400	4	460.95	620	380.50	334
Gunst and Mason [17]	100	15	24.51	202	18.25	150
	200	15	120.45	371	90.75	210
	400	15	560.90	750	401.88	375

*CPU time is reported by Mathematica [18].
†Number of iterations is reported by LINDO [13].

Problem 3 and Problem 4 are solved by Mathematica [18] and LINDO [13], i.e. two widely used linear program packages, on a 586 personal computer. Table 1 summarizes the results of solving Problem 3 and Problem 4 using published data. Table 2 displays the results of using generated test data sets to solve the two problems. Mathematica [18] can only report the CPU time of reaching the optimal solution; LINDO [13] can only show the number of iterations of solving the problem. Experiments demonstrate that although both problems have the same solutions, Problem 4 takes less CPU time and less number of iterations to reach the optimum solutions than Problem 3.

4. CONCLUSIONS

This study reformulates an LAV regression problem as a linear program using lower number of variables. The linear program proposed herein can attain a feasible basic solution. Numerical experiments confirm that the reformulated form is more computationally efficient than the conventional form of the LAV regression problems.

REFERENCES

1. Charnes, A., Cooper, W. W. and Ferguson, R., Optimal estimation of executive compensation by linear programming. *Manage. Sci.*, 1995, **1,** 138–151.
2. Dodge, Y., *L_1: Statistical Analysis and Related Methods*. Elsevier Science, Amsterdam, 1992.
3. Dielman, T.E., Computational algorithms for least absolute value regression. In *L_1: Statistic Analysis and Related Methods*, eds. Y. Dodge. Elsevier Science, Amsterdam, 1992, pp. 35–58.
4. Armstrong, R. D., FRAME, E. L. and Kung, D. S., A revised simplex algorithm for the absolute deviation curve fitting problem. *Commun. Statist. Simul. Comput. B*, 1979, **8,** 175–190.
5. Abdelmalek, N. N., A FORTRAN subroutine for the L_1 solution of overdetermined systems of linear equations. *ACM Trans. Math. Software*, 1980, **6,** 228–230.
6. Bloomfield, P. and Steiger, W., Least absolute deviations curve: Fitting. *SIAM J. Sci. Statist. Comput.*, 1980, **1,** 290–300.
7. Ignizio, J., *Introduction to Linear Goal Programming*. Sage, Newbury Park, California, 1985.
8. Armstrong, R. D. and Kung, M. T., A dual algorithm to solve linear least absolute value problems. *J. Oper. Res. Soc.*, 1982, **33,** 931–936.
9. Gentle, J. E., Narula, S. C., and Sposito, V. A., Testing software for robust regression. In *L_1: Statistic Analysis and Related Methods*, eds. Y. Dodge. Elsevier Science, Amsterdam, 1992, pp. 134–158.
10. Bazaraa, M. S., Jarvis, J. J. and Sherali, H. D., *Linear Programming and Network Flows*. Wiley, New York, 1990.
11. Winston, W., *Operations Research: Applications and Algorithms*. Duxbury Press, Boston, 1987.
12. Gonin, R. and Money, A. H., *Nonlinear LP-Norm Estimation*. New York, Marcel Dekker, 1989.
13. Schrage, L., *LINDO: User's Manual*. Scientific Press, San Francisco, California, 1991.
14. Li, H. L., Technical note: An efficient method for solving linear goal programming problems. *J. Opt. Theory Appl.*, 1996, **90,** 465–469.
15. Taylor, M. S. and Thompson, J. R., Database random number generation for a multivariate distribution via stochastic simulation. *Comput. Statist. Data Anal.*, 1986, **4,** 93–101.
16. Weisberg, S., *Applied Linear Regression*. John Wiley and Sons, New York, 1980.
17. Gunst, R. F. and Mason, R. L., *Regression Analysis And Its Application: A Data-Oriented Approach*. Marcel Dekker, New York, 1980.
18. *Mathematica User's Guide for Microsoft Windows*. Wolfram Research, New York, 1994.

AUTHOR BIOGRAPHY

Han-Lin Li is Professor in the Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan. He has a Ph.D. degree from the University of Pennsylvania, U.S.A. His research include operational research, decision support system and geographic information system. His research work is published in *Decision Support Systems, Operations Research, Journal of Operational Society, European Journal of Operation Research, Journal of Information Science and Engineering, Fuzzy Sets and Systems* and *Computers and Operations Research*.