ALGORITHM 615
The Best Subset of Parameters in Least Absolute Value Regression

R. D. ARMSTRONG
University of Georgia
P. O. BECK
Southern Methodist University
and
M. T. KUNG
California State University

Categories and Subject Descriptors: G.1.6 [**Numerical Analysis**]: Optimization—*linear programming*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*statistical computing, statistical software*

General Terms: Algorithms

Additional Key Words and Phrases: Regression, least absolute value criterion, branch and bound

DESCRIPTION

The purpose of this algorithm is to determine the best subset of parameters to fit a linear regression under a least absolute value criterion. A complete description of the algorithm is given in [2]. The program consists of seven subroutines written in standard FORTRAN.

During the initial phases of data analysis it is frequently desirable to consider different mathematical model formulations. One common technique in linear regression analysis is to obtain the "best" model when including exactly k independent variables. A generalization of this approach is to obtain the best subset for $k = p, p + 1, \ldots, m$ parameters in the model, where m is the total number of independent variables observed. The solution algorithms for this best subset problem are fairly well known when the least-squares criterion is used

Received 6 April 1982; accepted 15 September 1983
Authors' addresses: R.D. Armstrong: College of Business, University of Georgia, Athens, GA 30602; P.O. Beck, Cox School of Business Administration, Southern Methodist University, Dallas, TX 75275; M.T. Kung, Department of Management Science, California State University, Fullerton, CA 92634.
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
© 1984 ACM 0098-3500/84/0600-0202 $00.75

Table I. A Comparison of KBEST with SUBSET

	$n = 275$, $m = 6$		$n = 250$, $m = 8$		$n = 200$, $m = 10$	
	SUBSET	KBEST	SUBSET	KBEST	SUBSET	KBEST
Best 1 to m	6.38	2.555	23.379	4.599	47.519	12.770
	862	665	2610	1208	4992	2510
Best 3 to m	4.026	1.404	17.795	3.633	44.746	11.596
	488	327	1852	825	4368	1998
Best 5 to m	2.258	0.456	9.102	0.988	20.007	6.678
	234	58	991	56	1895	542
Best 7 to m	—	—	3.153	0.290	4.437	2.372
			291	22	476	108
Best 9 to m	—	—	—	—	2.341	0.380
					233	25

The upper number in each row is the CPU time in seconds; the lower number is the number of iterations required, n is the number of observations, and m is the number of parameters.

(see [4, 6, 7]). This paper presents a best subset algorithm when the evaluation criterion is least absolute values. Special purpose codes for other least absolute value problems are found in [1, 3].

Let $(x_{i1}, x_{i2}, \ldots, x_{im}, y_i)$, $i = 1, 2, \ldots, n$, be given. The least absolute value regression problem is to find $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ to

$$\text{minimize} \left| \sum_{i=1}^{n} y_i - \sum_{j \in J} x_{ij}\beta_j \right|, \tag{1}$$

where J is the index set of independent variables included in the model.

Charnes and Cooper [5] have shown that (1) is equivalent to the following linear-programming problem:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} (P_i + N_i) \\ &\text{subject to} && \sum_{j \in J} x_{ij}\beta_j + P_i - N_i = y_i, \end{aligned} \tag{2}$$

where $P_i \geq 0$, $N_i \geq 0$, and $i = 1, 2, \ldots, n$. The best subset for a given number of independent variables k where $k \leq m$, is one which yields the minimum objective value of all possible subsets of k variables from among the set of m variables under consideration.

The algorithm presented uses a branch-and-bound technique to find the best subset regression. Each node of the solution tree corresponds to a linear programming problem of the form given by (2). This problem is solved using a special-purpose revised simplex algorithm. A detailed description of the strategies used in developing the solution tree can be found in Armstrong and Kung [2].

An available option allows for specification of a percentage deviation from optimality. When the input parameter POPT is not equal to zero, the obtained subsets are only guaranteed to have a sum of absolute values within (100-POPT) percent of the optimal sum of absolute values. As was shown in [2], this option can provide significant savings in computer time at the expense of suboptimality in some cases. Another option allows the user to force any of the parameters to be included in every model.

COMPUTATIONAL RESULTS

The algorithm was tested together with the Narula and Wellington [8] computer code SUBSET, which is designed to solve the same problem. Several runs were made with randomly generated problems of various dimensions and the results are summarized in Table I. Sample problems taken from the literature on regression analysis were also solved with similar results. In terms of numerical accuracy, for the problems we solved, all objective values corresponded to seven digits. All runs were performed on the CDC Dual Cyber 170/750 computer with a sixty-bit word at the University of Texas at Austin Computation Center using an MNF compiler.

ACKNOWLEDGMENT

The authors are indebted to Dr. A. Buckley for making many helpful suggestions that improved the form and structure of the algorithm.

REFERENCES

1. ABDELMALEK, N.N. L_1 solution of overdetermined systems of equations. *ACM Trans. Math. Softw. 6*, 2 (1980), 220–227.
2. ARMSTRONG, R.D., AND KUNG, M.T. An algorithm to select the best subset for a least absolute value regression problem. In *Optimization in Statistics Volume, TIMS Studies of the Management Sciences 33* (1982), 931–936.
3. BARRODALE, I., AND ROBERTS, F.D.K. Solution of the constrained L_1 linear approximation problem. *ACM Trans Math. Softw 6*, 2 (1980), 231–235.
4. BEALE, E.M L., KENDALL, M.G., AND MANN, D.W. The discarding of variables in multivariate analysis *Biometrica 54* (1967), 357–366.
5. CHARNES, A., AND COOPER, W.W. *Management Models and Industrial Applications of Linear Programming*, vols. I and II, Wiley, New York, 1961.
6. FURNIVAL, G.M., AND WILSON, R.W. Regression by leaps and bounds *Technometrics 16* (1974), 499–512.
7. KENNEDY, W.J., AND GENTLE, J.E. *Statistical Computing*, Marcel Dekker, New York, 1980.
8. NARULA, S C., AND WELLINGTON, J.F. Selection of variables in linear regression using the minimum sum of weighted absolute errors criterion. *Technometrics 21* (1979), 299–306.

ALGORITHM

[A part of the listing is printed here. The complete listing is available from the ACM Algorithms Distribution Service (see page 215 for order form).]

```
      SUBROUTINE KBEST (X,Y,M,N,ITER,IFAULT,POPT,MININ,NMAX,MMAX,BVAL
     1,IDEX,ISTAT,ZL)
C
C***********************************************************************
C
C
C     THE PURPOSE OF THIS PROGRAM IS TO DETERMINE THE BEST SUBSET OF
C     PARAMETERS TO FIT A LINEAR REGRESSION UNDER AN LEAST ABSOLUTE
C     VALUE CRITERION. THIS PROGRAM UTILIZES THE SIMPLEX METHOD OF
C     LINEAR PROGRAMMING WITHIN A BRANCH-AND-BOUND ALGORITHM TO
C     SOLVE THE BEST SUBSET PROBLEM.
C
C
```

```
C     THE ALGORITHM IS BASED ON THE PUBLICATION:
C        ARMSTRONG,R.D. AND M.T. KUNG "AN ALGORITHM TO SELECT THE BEST
C        SUBSET FOR A LEAST ABSOLUTE VALUE REGRESSION PROBLEM,"
C        OPTIMIZATION IN STATISTICS, TIMS STUDIES OF THE MANAGEMENT
C        SCIENCES.
C
C   FORMAL PARAMETERS
C
C   X       REAL ARRAY          INPUT:  VALUES OF INDEPENDENT VARIABLES
C           (NMAX,MMAX)                  SUCH THAT EACH ROW CORREPSONDS TO
C                                        AN OBSERVATION
C
C   Y       REAL ARRAY          INPUT:  VALUES OF THE DEPENDENT VARIABLES
C             (NMAX)
C
C   M       INTEGER             INPUT:  NUMBER OF DEPENDENT VARIABLES
C
C   N       INTEGER             INPUT:  NUMBER OF OBSERVATIONS
C
C   ITER    INTEGER             OUTPUT: NUMBER OF ITERATIONS
C
C   IFAULT  INTEGER             OUTPUT: FAILURE INDICATOR
C                                 =0     NORMAL TERMINATION
C                                 =1     OBSERVATION MATRIX DOES NOT HAVE
C                                        FULL ROW RANK (RANK M)
C                                 =2     PROBLEM SIZE OUT OF RANGE
C                                 =3     NO PIVOT ELEMENT FOUND IMPLYING
C                                        NEAR SINGULAR BASIS
C
C   POPT     REAL               INPUT:  PERCENTAGE DEVIATION FROM
C                                       OPTIMALITY ALLOWED
C
C   MININ    INTEGER            INPUT:  MINIMUM NUMBER OF PARAMETERS IN THE
C                                      MODEL.  BEST SUBSET OF SIZE MININ TO
C                                      M IS OBTAINED.
C
C   NMAX     INTEGER            INPUT:  DIMENSION OF ROWS IN X (ALSO Y)
C
C   MMAX     INTEGER            INPUT:  DIMENSION OF COLUMNS IN X
C
C   BVAL     REAL ARRAY         OUTPUT: ARRAY OF OPTIMAL BETA VALUES FOR
C                                       EACH SUBSET. THE BETA VALUES FOR
C                                       THE SUBSET OF SIZE M ARE STORED
C                                       IN POSITIONS BVAL(1),BVAL(2),...,
C                                       BVAL(M), FOR THE SUBSET OF SIZE
C                                       M-1 THE VALUES ARE STORED IN
C                                       POSITIONS BVAL(M+1),BVAL(M+2),...,
C                                       BVAL(2M-1). IN GENERAL, THE BETA
C                                       VALUES FOR THE OPTIMAL SUBSET OF
C                                       SIZE K ARE STORED IN POSITIONS
C                                       BVAL(L),...,BVAL(L-K+1) WHERE
C                                          L=(M*(M+1)-K*(K+1))/2 + 1
C
C  IDEX        INTEGER ARRAY    OUTPUT: BETA INDEX SET FOR THE OPTIMAL
C        (((MMAX+1)*MMAX)/2)            SUBSET. THIS ARRAY IS A  PARALLEL
C                                       ARRAY FOR BVAL; I.E., IF BVAL(J)=2.
C                                       AND IDEX(J)=7  THEN BETA(7)=2.7 IN
C                                       THE ASSOCIATED OPTIMAL SUBSET.
C
C  ISTAT       INTEGER ARRAY    INPUT:  PARAMETER STATUS ARRAY.
C                (MMAX)
C
C                                           1 IF BETA(J) IS REQUIRED
C                                             IN EVERY MODEL
```

```
C                    ISTAT(J)=
C
C                                              Ø OTHERWISE
C
C  ZL       REAL ARRAY     OUTPUT: BEST OBJECTIVE VALUE FOR EACH SUBSET
C            (MMAX)
C                                 ZL(J) GIVES THE BEST OBJECTIVE VALUES
C                                 FOR THE SUBSET WITH M-J+1 PARAMETERS
```