Using Simulated Annealing for Selection in Multiple Regression Analysis

Zvi Drezner
George A. Marcoulides
California State University - Fullerton

This note presents the simulated annealing heuristic search procedure as an alternative variable selection method for use in multiple regression analysis. The procedure performs better than traditionally used model selection techniques.

Drezner, Marcoulides and Salhi (1999) recently illustrated the heuristic Tabu search procedure as an alternative variable selection method for use in multiple regression analysis. The Tabu search procedure was compared to traditionally used regression analysis procedures (e.g., maximum R^2 and stepwise selection). The results of the study indicated the superiority of the Tabu search procedure over other model selection procedures in multiple regression analysis and comparability to the all-possible regression that may require prohibitive computer time. Using simulated data sets, Tabu search found the optimal solutions for all test problems examined without any computational difficulty.

The purpose of this note is to present the simulated annealing search procedure, which is a different heuristic search technique, for model selection in multiple regression analysis. To examine the capabilities of the simulated annealing search procedure, the same simulated data sets used by Drezner et al. (1999) were analyzed.

Simulated Annealing for Model Selection

Consider a multiple linear regression model with n observations and k independent variables. The most commonly used criterion to help in choosing between alternative equations in multiple regression is the R^2 (adjusted or unadjusted), the F-ratio based on R^2, along with the statistical significance of the F-ratio (Schumacker, 1994). Obviously, this criterion could easily be replaced by any other selection from those available in the literature. Based upon the selected criterion, the objective is to find the subset of independent variables that yields the lowest significance level among all possible subsets. For example, with 26 independent variables 2^{26} = 67,108,864 possible subsets must be calculated along with their significance levels. As such, it should be obvious that a very large number of equations need to be examined even when the number of independent variables is relatively small.

Simulated annealing (SA) is ideally suited for solving all types of large-scale optimization problems (Kirkpatrick, Gelat, & Vecchi, 1983). The process simulates the annealing of metals by starting with a high temperature and cooling the metal off. The process of simulated annealing has been successfully used for the solution of numerous optimization problems in the field of operations research (see Salhi, 1998 for a review and detailed description of the method).

The general simulated annealing (SA) approach is described below. Following the general description, we present the particular parameters used to solve the multiple regression model selection issue examined in this note. A FORTRAN coded computer program for model selection in multiple regression is available upon request from the authors.

The General SA Approach

1. A starting solution is selected.
2. A starting temperature T_0 is selected.

   (T_i is the temperature in iteration i.)
3. The following iterations are repeated N times.
4. At iteration i:

   a. A perturbation of the current selected set is randomly generated.

   b. The difference between the values of the objective function of the current set and the perturbed set, Δf , is calculated.

   c. If the perturbation results in a better objective function, it is accepted and the set of selected variables updated.

   d. If the perturbation results in a worse objective function, the quantity $\delta = \Delta f / T_i$ is calculated.

   e. The perturbed set is accepted with a probability of $e^{-\delta}$. Otherwise, the selected set remains unchanged and the perturbation ignored.

   f. The temperature T_i is changed to T_{i+1}.

Specific Parameters needed for Multiple Regression

1. The empty set was selected as a starting solution (i.e., no independent variables).
2. The starting temperature was set to $T_0 = 1$ This means that if the perturbation doubles the significance level, it is accepted 37% of the time.
3. A perturbation of the current selected set is created by randomly selecting an independent variable. If the variable is in the current set, it is moved out, and if it is not in the current set it is put in.
4. The number of iterations was set to N=10,000.
5. Since our objective function is a significance level, which varies a lot among problems, we replaced the change in the objective function Δf with the relative change in the objective function $\Delta f / f$ where f is the value of the objective function of the current set.
6. The last selected set was selected as the solution. One may keep the best solution encountered throughout the iterations as the solution.
7. The success of the simulated annealing procedure depends on the selection of the starting temperature T_0, the way the temperature is lowered, and the number of iterations. We kept the temperature constant for blocks of 100 iterations each. When a block of 100 iteration is completed, the temperature is multiplied by the value 0.95. One hundred blocks of 100 iterations each were executed for a total of 10,000 iterations. This lead to a final temperature of 0.006. At the end of the procedure, a deterioration in the significance level by a factor of 1.05 is accepted with probability of only 0.0002.

Computational Results

The simulated annealing procedure was tested on the simulated data sets examined by Drezner et al (1999). The data sets used had 50 observations and k variables ranging in $17 \leq k \leq 26$. The data for the smallest problem with $k = 17$ independent variables and $n = 50$ are presented in Table 1 (the remaining data sets are available upon request from the authors). Using this data set, the optimal subset of independent variables includes #2, #6, #12, and #17. It is important to note that the proposed simulated annealing procedure found this optimal solution. In contrast, stepwise regression produced the set #1, #4, #5, #7, #12, #13, #17 (when the entry selection level was set to 0.15), the set #1, #5, #12, #17 (when the entry selection level was set to 0.05), whereas maximum R^2 found the set #2, #5, #7, #1, #13, #17. Interestingly, variable #6, which is in the optimal group, was never identified by any of the other procedures, and variable #5, which is not in the optimal group, was included by the other procedures. Table 2 presents the results of the comparison between the SA procedure and the maximum R^2 and stepwise procedures. As can be seen in Table 2, the SA procedure found the best subset for all the data sets examined. In contrast, the other procedures were not very systematic in selecting the optimal solution. It is important to note that the results obtained using the simulated annealing procedure were identical to those obtained by Drezner et al. (1999) using their proposed Tabu search procedure which is a different local search procedure.

References

Drezner Z., Marcoulides, G.A., & Salhi, S. (1999). Tabu search model selection in multiple regression analysis. *Communications in Statistics - Computation and Simulation*, *28*(2), 349-367.

Kirkpatrick S., Gelat, C.D., & Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, *220*, 671-680.

Salhi S. (1998). Heuristic search methods, in G.A. Marcoulides (Ed.) *Modern Methods for Business Research*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Publishers.

Schumacker, R.E. (1994). A comparison of the Mallows *Cp* and principal component regression criteria for best model selection in multiple regression. *Multiple Regression Viewpoints*, *21*(1), 12-22.

Table 1. Data for the 17 Variable Problem

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	y
10	12	11	22	25	29	33	34	32	26	28	33	23	21	19	25	31	182
22	22	22	16	20	16	15	17	12	17	12	14	12	15	15	15	20	129
35	35	33	36	38	28	28	25	22	26	24	17	17	14	15	18	27	188
41	42	30	31	23	19	19	21	28	33	36	27	27	21	17	20	21	170
12	13	14	16	18	27	32	25	19	22	18	27	19	19	21	28	33	167
31	36	34	24	27	20	26	19	17	22	27	20	16	20	22	23	27	149
0	1	11	7	15	21	19	15	20	18	13	23	25	30	24	23	28	125
27	24	20	22	17	24	29	33	35	36	27	21	18	18	27	23	16	130
10	19	15	25	24	30	21	24	28	23	28	25	24	25	18	13	19	135
29	21	26	31	32	23	17	14	18	27	19	21	18	23	18	26	31	159
4	14	13	11	16	24	21	19	27	22	27	31	25	18	20	18	14	151
21	15	15	18	20	28	29	29	21	21	29	28	23	18	20	27	20	141
46	32	25	20	21	23	26	19	13	23	28	31	23	30	27	33	28	181
10	16	21	14	20	17	20	18	19	15	20	25	20	14	10	20	22	132
37	36	36	32	29	23	25	27	19	25	19	23	27	19	24	20	25	180
41	38	40	35	25	19	19	18	25	17	26	28	29	27	19	20	25	164
13	18	16	25	18	20	16	17	23	19	14	21	24	25	22	15	21	155
27	28	26	27	33	33	33	25	19	24	21	15	17	26	31	28	24	156
39	29	25	19	28	23	27	25	28	26	29	27	29	29	34	37	35	195
37	36	26	22	30	30	30	31	30	35	34	24	18	17	19	14	17	176
4	13	18	19	27	20	17	12	21	19	18	25	24	29	21	14	22	149
10	8	11	11	16	22	21	23	30	30	21	26	25	30	23	21	18	158
14	10	19	15	23	26	20	27	30	33	36	31	27	28	23	18	22	143
31	27	22	23	23	21	16	15	19	22	29	30	30	24	25	31	25	184
47	41	40	34	32	28	32	25	19	15	19	21	20	24	29	23	30	206
34	38	35	33	30	28	22	18	21	16	24	19	20	15	24	31	31	186
34	34	34	38	35	34	25	20	17	17	15	11	21	16	25	31	23	193
21	20	18	16	24	28	31	25	29	21	20	24	18	22	27	20	25	146
22	16	16	15	24	27	27	32	28	28	25	23	24	25	28	20	27	157
14	24	27	21	21	19	21	21	21	28	29	23	16	16	19	20	26	134
12	21	17	14	20	17	12	13	12	9	6	8	15	21	19	15	24	98
3	5	14	12	19	19	13	13	22	27	31	33	27	32	25	22	29	143
24	27	21	22	29	20	18	16	16	13	20	19	27	23	18	19	22	156
47	40	39	35	27	20	26	28	29	31	24	17	13	16	22	17	26	186
39	30	23	19	18	23	26	30	21	18	27	31	26	24	21	23	20	151
5	16	24	24	21	23	26	32	26	21	21	16	14	21	23	24	22	146
20	22	18	14	20	21	26	28	32	28	33	34	27	30	25	28	29	194
13	10	13	16	24	19	27	22	30	34	27	18	21	17	20	25	24	146
28	32	35	35	35	29	30	29	34	33	27	21	17	24	24	28	28	192
32	37	34	31	22	16	18	23	23	16	22	25	30	35	34	32	25	172
49	49	35	37	37	27	27	27	33	33	32	33	23	19	25	22	26	206
34	36	29	30	26	27	32	25	19	19	24	29	32	32	27	33	24	182
35	28	24	18	18	20	22	16	11	14	20	17	12	14	17	15	24	155
10	10	11	12	9	10	7	12	16	16	12	9	11	16	24	26	21	90
19	18	24	24	31	32	22	28	31	29	30	28	23	27	20	19	15	152
8	17	21	26	18	13	20	28	30	31	24	24	18	26	21	17	25	140
1	8	14	24	30	33	33	26	32	36	31	25	22	21	27	32	28	151
15	21	22	16	22	27	29	22	20	27	22	26	30	31	25	21	17	165
48	45	35	29	26	21	26	24	27	24	25	18	23	21	26	22	23	173
3	2	5	13	15	20	24	20	19	21	20	14	11	17	20	16	13	98

Table 2. Comparison of Simulated Annealing to Other Regression Procedures

Number of Variables	Variables in Optimal Solution	SA Procedure	Stepwise Procedure		Max R^2 Procedure	
			Include	Exclude	Include	Exclude
17	2, 6, 12, 17	*identical*	1,4,5,7, 13	2,6	5,7,13	6
18	1, 6, 12, 13, 16, 17	*identical*	7	—	7	—
19	2, 6, 12, 13, 17	*identical*	1	—	8,10,15	—
20	1, 5, 7, 12, 13, 16, 17	*identical*	4	—	4,8,15	
21	1, 6, 12, 13, 17, 18	*identical*	*identical*		*identical*	
22	1, 4, 6, 7, 9, 13, 17, 18, 22	*identical*	*identical*		*identical*	
23	1, 2, 3, 6, 12, 13, 17, 22	*identical*	4	—	4	—
24	1, 4, 5, 6, 9, 11, 12, 13, 16, 18, 22, 24	*identical*	*identical*		*identical*	
25	1, 3, 4, 6, 8, 10, 12, 13, 15, 17, 19, 22, 25	*identical*	5,18,23	3,4,8,10, 15,19	2,7	1
26	1, 3, 6, 12, 13, 16, 17, 18, 20, 22, 24, 26	*identical*	5	—	8	—

Note: The term ***'identical'*** indicates that the final set of variables selected by that procedure is the same as the optimal set. The columns headed by **'Include'** indicate that the given procedure includes variables not in the optimal solution, and those headed by **'Exclude'** indicate that the given procedure excludes variables which are members of the optimal set.