A Comment on Pohlmann's Algorithm for Subset Selection in Multiple Regression Analysis

by

MaryAnn Ross
George P. McCabe

Purdue University

Department of Statistics
Division of Mathematical Sciences
Purdue University
Mimeograph Series #403

January 1975

A Comment on Pohlmann's Algorithm for Subset Selection in Multiple Regression Analysis

by

MaryAnn Ross
George P. McCabe
Purdue University

Abstract

Pohlmann's algorithm for incorporating cost criteria into the variable selection problem in multiple regression is examined. It is pointed out that this algorithm has the property that the choice of the optimal subset can be artificially changed by the addition of another variable. An example is included to illustrate this property.

Pohlmann (4) has presented an algorithm which incorporates cost information into the selection of a subset of variables in multiple regression analysis. Other approaches to this problem are given in (2) and (3). We suggest that Pohlmann's proposed method of reducing the costs and losses due to lack of predictability to a common scale of measurement may not be the most desirable.

The loss function to be considered is

$$L_I = k_1(c_1 \times \text{cost}_I) + k_2(c_2 \times (1-R_I^2)).$$

The weighting coefficients, c_1 and c_2, suggested are

$$c_1 = \left(\sum_{I=1}^{J} \text{cost}_I\right)^{-1}$$

and

$$c_2 = \left(\sum_{I=1}^{J} (1-R_I^2)\right)^{-1}$$

where

J = the total number of subsets under consideration.

Using the above definitions, the revised costs and losses due to lack of predictability and in turn the loss function for any particular subset are dependent on the total number of subsets under consideration. In other words, if an additional variable is added to or deleted from the original set of predictor variables, the loss function for the original subsets can be changed. This may, in fact, lead to the choice of a different subset from the original collection as the optimal subset. Ideally, the relative loss corresponding to any given subset, as compared to the loss for another given subset should not be dependent on the cost and lack of predictability of extraneous variables.

An Example

The following example uses data presented by Hald (1) with arbitrarily assigned cost values. In Table 1 the data are analyzed as suggested by Pohlmann. The chosen subset in this case is that containing predictor variates 3 and 4. Now let us assume that we can measure the dependent variable precisely for $100 per observation, but further, that $100 is the maximum amount per observation that we are willing to spend. Table 2 contains the analysis for this problem. It can be seen that the chosen subset in this case is that containing variables 1, 3 and 4. Even though the added variable is not a feasible choice because of its high cost, the preferred subset according to this method has changed to another subset of the original set.

An Alternative Method of Standardizing Cost and Validity Data

In order to overcome the above problem it is suggested that the weighting coefficients in the loss function might be defined as:

$$c_1 = (c_{max})^{-1}$$

where c_{max} is the maximum cost which the investigator is prepared to spend, and

$$c_2 = 1.0.$$

In this way, the revised values of both cost and loss due to lack of predictability are between 0.0 and 1.0 and the combined loss value for any subset will not depend on the number of subsets under consideration.

REFERENCES

1. Hald, A., Statistical Theory with Engineering Applications. New York: Wiley, 1952.

2. Lindley, D. V., The choice of variables in multiple regression. Journal of the Royal Statistical Society, Series B, 30, 1968, 31-35.

3. McCabe, G. P. and Ross, M. A., A stepwise algorithm for selecting regression variables using cost criteria. To appear in "Proceedings, Computer Science and Statistics: Eighth Annual Symposium on the Interface", 1975.

4. Pohlmann, J. T., Incorporating cost information into the selection of variables in multiple regression analysis. Viewpoints, 1973, 4, 18-26.

TABLE 1. ANALYSIS OF HALD DATA.

	Cost A	Revised Cost B	$1-R^2$ C	Revised Loss D	B+D E	B+3D F
1	5	.052	.47	.118	.170	.406
2	5	.052	.33	.083	.135	.301
3	1	.010	.72	.181	.191	.553
4	1	.010	.32	.080	.090	.250
1,2	10	.104	.02	.005	.109	.119
1,3	6	.062	.45	.113	.175	.401
1,4	6	.062	.03	.008	.070	.086
2,3	6	.062	.15	.038	.100	.176
2,4	6	.062	.32	.080	.142	.302
3,4	2	.021	.06	.015	.036	.066*
1,2,3	11	.115	.02	.005	.120	.130
1,2,4	11	.115	.02	.005	.120	.130
1,3,4	7	.073	.02	.005	.078	.088
2,3,4	7	.073	.03	.008	.081	.097
1,2,3,4	12	.125	.02	.005	.130	.140
NULL	0	.0	1.00	.251	.251	.753
TOTAL	96	.998	3.98	1.000		

TABLE 2. ANALYSIS OF HALD DATA WITH ADDITIONAL VARIABLE.

	Cost A	Revised Cost B	$1-R^2$ C	Revised Loss D	B+D E	B+3D F
5	100	.510	.00	.000	.510	.510
1	5	.026	.47	.118	.144	.380
2	5	.026	.33	.083	.109	.275
3	1	.005	.72	.181	.186	.548
4	1	.005	.32	.080	.085	.245
1,2	10	.051	.02	.005	.056	.066
1,3	6	.031	.45	.113	.144	.370
1,4	6	.031	.03	.008	.039	.055
2,3	6	.031	.15	.038	.069	.145
2,4	6	.031	.32	.080	.111	.271
3,4	2	.010	.06	.015	.025	.055
1,2,3	11	.056	.02	.005	.061	.071
1,2,4	11	.056	.02	.005	.061	.071
1,3,4	7	.036	.02	.005	.041	.051*
2,3,4	7	.036	.03	.008	.044	.060
1,2,3,4	12	.061	.02	.005	.066	.076
NULL	0	.000	1.00	.251	.251	.753
TOTAL	196	1.002	3.98	1.000		