RECENT DEVELOPMENTS IN SEQUENTIAL DATA ANALYSIS STRATEGY

John A. Sonquist, University of Michigan

Abstract

Within the past few years computerized sequential algorithms have been developed and used to search out and display non-additive relationships implicit in survey data. The objective is the display of information relevant to the problem of how to specify a multiple classification analysis model purporting to explain the phenomenon in question. Experimentation with one such algorithm (AID) has led to a classification of simple additive and interactive multivariate models related to elementary Boolean operators. These models are described and illustrated; extensions to the AID algorithm to facilitate specification of complex models, for dealing with covariate models arising from "crucial" variables or over-time survey data and for imposing symmetry restrictions are discussed. Preliminary experimentation with the improved algorithm is reported.

Keywords: Survey, Multivariate analysis, model specification, statistical interaction, computer program, prediction, analysis of variance, multiple regression, multiple classification analysis, data analysis strategy, simulation.

The task of interpreting a mass of non-experimental data, such as that generated by modern cross-sectional and longitudinal survey methods has remained a difficult one. Two reasons for this appear to stand out as particularly worthy of examination. One is that the statistical methods used by most data analysts are oriented toward deductive testing of isolated hypotheses rather than toward the more frequent task of extraction of information from data and the development of models on an inductive basis. In addition, the full power of "second generation" computing equipment has not yet been utilized (let alone "third generation"); it has the ability to extend the scope of the analyst's logic as well as perform computations rapidly for him.

The present paper is a modest attempt to surmount these obstacles by extending the idea of coupling the calculating power of modern computing equipment to a formal sequential algorithm, the objective of which is the generation of information about the data. In particular, our concern is with obtaining summary information organized for use by the analyst in the inductive stages of his research; that is, in the stages of model formulation and development. The objective of the procedure is the display of information relevant to the problem of how to formulate or specify a model purporting to explain the phenomenon in question.

One frequently appearing and concrete form taken by this general specification problem is the question of choosing and of specifying the form of the terms to be included in a least squares multiple classification equation (MCA). Such an equation is commonly used as a statistical model representing the simultaneous and direct explanatory effects of a correlated set of presumed causal factors on a single dependent variable. It is of the form

$$Y_{ijk...} = \bar{Y} + a_i + b_j + \cdots e_{ijk...} \quad (1)$$

where

- Y_{ijk} = the score (on the dependent variable) of individual k, who falls in category i of predictor A and into category j of predictor B, etc.;
- $\bar{Y}$ = the grand mean of the dependent variable;
- a_i = the "effect" of membership in the i-th category of predictor A expressed as a deviation from the grand mean and adjusted for the intercorrelations of predictor A with the other terms in the equation;
- b_j = the corresponding "effect" associated with predictor B;
- $e_{ijk...}$ = the error term for the k-th individual.

Any given term in the equation, say b_j, may represent the effect of a two-way or higher-order interaction term between explanatory factors not otherwise represented in the equation.[1]

---

[1] For a more detailed exposition of this basic technique, see Andrews, Morgan and Sonquist (1967). A simple n-way analysis of variance model is not used because the number of observations in the cells of an n-way cross-classification of the predictors are not equal or proportional. This condition ordinarily occurs in survey or other non-experimental data in which there are non-zero correlations between explanatory characteristics.

It is this problem of choosing terms to be included in the equation to which we address ourselves. We shall review some of the previous work in this area, outline several improvements in an algorithm proposed several years ago, and present an illustration of the improved technique.

Previous Work in This Area

In a previous paper before this group, two of the present authors reported preliminary results from a new method of data analysis. We had undertaken to design, program, test, and document a large scale computer program. The algorithm incorporated into the program was aimed at extracting information bearing on the need to introduce interaction terms into a multivariate analysis in which a set of correlated predictor variables were to be related simultaneously to a criterion. Starting from a consideration of some of the problems inherent in applying multivariate statistical techniques (particularly multiple regression) to cross-sectional survey data, we had concluded that the use of Multiple Classification Analysis (MCA) permitted adequate handling of most data characterized by intercorrelated predictors, nominal scales and non-linearities, but could not deal with interaction effects. The computerized procedure that was developed at that time attacked the problem of locating interaction terms by asking a different kind of statistical question of the data than is implied by the immediate choice of MCA.[1]

The technique (termed Automatic Interaction Detection [AID]) employs a sequential decision procedure to divide the sample into a mutually exclusive set of sub-groups through a series of successive binary partitions, each formed by combining observations which are alike on certain characteristics. At each stage every observation is a member of exactly one such sub-group. These groups are formed so that at each stage in the process their means account for more of the total sum of squares (i.e., reduce the error sum of squares more) than the means of any other pair of such sub-groups obtainable by the algorithm. Thus, at each stage in the branching process, the set of groups developed at that point represents, according to the criteria of the model, the best currently available scheme for predicting the dependent variable from the information available. The branching process terminates when all existing sub-groups are either so homogeneous that the desired accuracy has been achieved, or no variable can be found which will enable an improvement in prediction sufficient to warrant its use.[2]

An examination of the statistics computed during the tree-like partitioning process provides evidence in support of, or against, the introduction of additivity assumptions, as well as providing some indication of what kind of interaction term should be generated, if required.

Experimentation with the method on data which were constructed to violate additivity assumptions revealed that asymmetric tree structures were associated with the existence of predictors which interacted.[3] On the other hand, additivity was associated with symmetric structures.

Interactive data were also found to be associated with the presence of differential changes in explanatory power displayed by corresponding predictors in different branches of the tree. Additivity was associated with similar changes in explanatory power.

Inspection of the profiles of mean values defining the effects of a given predictor in various branches of the tree also provided evidence of the appropriateness of making additivity assumptions. Similarity of effect profiles of a predictor in various parts of the tree structure implies additivity; but interactive data produce incongruent profiles.

AID was first proposed as a substitute for MCA. However, experience with the method in actual use and knowledge gained through experimentation with known data structures led to the development and publication of a strategy for using the AID and MCA techniques

---

1 Preliminary results are reported in Morgan and Sonquist (1963b).

2 Details of the process, flow charts, the computer program, sample input and output, formulas, nine illustrative analyses and recommendations for interpreting the output of the program are reported in Sonquist and Morgan (1964).

3 This experimentation is reported in Sonquist (1969a).

jointly, to supplement each other.[1] In addition it appeared that substantial gaines in analytic power might be made from implementing a somewhat more sophisticated sequential algorithm.[2]

Since its original implementation in a large scale program (Sonquist and Morgan, 1964) the algorithm has been translated by others to run on a number of other computers (Campbell, 1965; Aptakin, 1965; Land, 1965; Kay, 1966; Marks, 1966). Versions of the AID algorithm were also adapted to European equipment (Biervert, 1966, Arpi, 1967). The methods have been used by economists (e.g., Snowbarger, 1967; Gensemer, Lean and Neenan, 1965), by sociologists (Ross and Bang, 1965), by political scientists, (e.g., Sarlvik, 1968), by marketing researchers (e.g., Arpi, 1967), by engineers (e.g., Carlson, 1967), and by psychologists, (Caplan et al, 1966).

Revisions in the Algorithm

Experience in using the original algorithm over a period of five years, discussions with others who have used it, and experimentation with contrived data have all led to a rather large collection of proposed improvements in the techniques. Most of these are minor from an analytic standpoint, such as obtaining output in improved form.[3] They are obtained simply as a byproduct of the reprogramming necessary to incorporate the major changes proposed below.

The major revisions include extending the original algorithm to deal with covariance as well as a profile of means, providing for the imposition of an optional premium for symmetric partitioning, and extending the sophistication of the search algorithm to include a "look-ahead" to several successive partitions. In the latter case, the mechanism would not proceed blindly, always trying to maximize the explained variation in the current partition, but, like a chess player, would explore the possibility of sacrificing present payoff in favor of even greater gains from subsequent "moves." It would also attempt to provide the simplicity and parsimony of a symmetric model where the data appeared to warrant its use. Each of the revisions is discussed below in some detail.

Covariance Search Routine

There are many situations in economic, sociological and psychological research where a multivariate analysis is required, but where there exists one dominant explanatory or control variable. The explanatory factor in question may be of particular importance to some theoretical edifice, it may be subject to control more easily than other factors, or it may simply have been measured much more readily and reliably than the others. Where the data come from an experiment and not a survey, the obvious procedure is covariance analysis.

However, with non-orthogonal survey data, one may want to search out subgroups in which there are different relationships between the dependent variable and this dominant explanatory variable. For instance, in much analysis of cross-section survey data, the economist is often interested in the effect of personal or family income on some behavioral variable, and on whether that "income effect" (as represented by the slope of the regression of the behavioral variable on income) varies with other circumstances. The answer to this question will help to determine whether it is necessary to disaggregate the data in models used for forecasting, and the optimal way to do it.

Sociologists and psychologists often face similar problems in which the purpose of the investigation requires isolating the effect of a particular variable under a wide variety of combinations of circumstances. For instance, intelligence, alienation and authoritarianism have each been the subject of repeated investigations in which the object has been to relate that particular factor to specific consequences in such a way as to specify the form of the relationship under various conditions and for particular types of people.

Another illustration is in the analysis of changes taking place over time.

---

1 The strategy developed was suggested in Andrews, Morgan and Sonquist (1967), and then elaborated further by Sonquist (1969a, 1969b).

2 See Sonquist (1967).

3 Some of the programming revisions might include the use of Fortran IV for compatibility with various manufacturers' equipment; a trichotomous partition option; improved input flexibility with regard to both data and control language; additional output including predictor summary tables; and improved variable transformation capabilities.

The initial value of a phenomenon under study clearly affects its value measured at a subsequent time. This is why the residuals from the regression of its current (t_2) value on its initial (t_1) value are often used as a measure of change, instead of the raw increments.[1] However, this "initial value" effect might not be the same for all sub-groups in the population. If, then, a single equation were to be fitted, a downward bias would be exerted on the correlations between change and those factors thought to be responsible for it. Thus, when residualizing a variable for study, a search should be made to determine if this initial value effect is homogeneous throughout the population.

For all these reasons, a variant of the earlier sequential data analysis algorithm has been developed in which the criterion for sequential subdivision of a sample is changed from the sum of squares explained by two sub-group means (instead of one pooled mean) to the sum of squares explained by two simple sub-group regressions (instead of one simple regression of the pooled data). The search algorithm otherwise has basically the same framework: each group potentially to be divided is examined using all feasible partitions based on each explanatory characteristic.[2] The difference is that the quantity maximized when a group is divided into sub-groups is the variation explained by the regression of the dependent variable on the covariate within each sub-group. If the two best fitting regression lines differ as to intercept or slope, then the unexplained variation would be reduced, and the split with the largest difference between the two regressions would reduce it the most.

Given this conceptualization of the problem, a partition may be chosen so as to maximize any one of three quantities. One may evaluate reductions in unexplained variation due to differences between means, differences in regression lines, or both taken together. The original AID algorithm sought to maximize the sum of squares explained by means of the sub-groups resulting from the partition of the "parent" group, i.e., it maximized the expression

$$N_1\bar{Y}_1^2 + N_2\bar{Y}_2^2 - N\bar{Y}^2 \qquad (2)$$

The rationale for this is easily seen in Table 1.

It can be seen from Table 2 that a second quantity could also be maximized.[1] This is the expression

$$(N_1-1)r_1^2 s_{y1}^2 + (N_2-1)r_2^2 s_{y2}^2 - (N-1)r^2 s_y^2 \qquad (3)$$

This is the sum of squares of the two regression estimates around the two group means resultinf from the partition.[2] It reduces to:

$$\sum_{i=1}^{2} \left[\frac{[\Sigma(Y-\bar{Y}_i)(X-\bar{X}_i)]^2}{\Sigma(X-\bar{X}_i)^2} \right] - \frac{[\Sigma(Y-\bar{Y})(X-\bar{X})]^2}{\Sigma(X-\bar{X})^2} \qquad (4)$$

A third quantity which may be maximized is obtained from the sum of these two and would represent the effect of both the variable used in the partition and the X covariate. The variation explained by the sub-group means and that explained by the regression are both maximized. This expression is

$$N_1\bar{Y}_1^2 + (N_1-1)r_1^2 s_{y1}^2 + N_2\bar{Y}_2^2 + (N_2-1)r_2^2 s_{y2}^2 - N\bar{Y}^2 - (N-1)r^2 s_y^2 \qquad (5)$$

The existing algorithm has also been modified to permit an examination of the effects of one crucial categorical predictor in various parts of the sample without permitting it to be used in the partitioning process. Permitting the analyst to retain the ability to conceptualize the effects of the crucial variable in terms of slopes and intercepts as well as the ability to use profiles of sub-group means gives a desirable simplicity.

---

1 For a thorough discussion of this problem see Lord (1967).

2 If the order of the (k) classes of the explanatory characteristic is maintained, there are only (k-1) ways of forming two groups on the basis of that predictor. However, if one reorders them on the basis of the means of the dependent variable there are many more possible ways of forming the two sub-groups.

1 Computational formulas are, of course, somewhat different and are not given here. In maximizing any of these expressions the last term is constant over all possible partitions and can be ignored. See Walker and Law (1953) pp 210-216.

2 See Walker and Law (1953), pp 242-244.

TABLE 1.

Analysis of Variance for Differences in Means

Source of Variation	d. f.	Sum of Squares	Mean Square
Observations around grand mean	N-1	$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(Y_{ij}-\bar{Y})^2 = SST$	s_y^2
Between group means	k-1	$\sum_{j=1}^{k} N_j(\bar{Y}_j-\bar{Y})^2 = SSB$	MSB
Within Groups	N-k	$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(Y_{ij}-\bar{Y}_j)^2 = SSW$	MSW

The covariate problem is illustrated below. Owning a home (as opposed to renting, etc.) is related not only to income, but to age, family size, and place of residence (urbanization). More important, a very large fraction of older people outside the large urban areas own a home regardless of their income; i.e., income differences have no effect at all on home ownership in this group. However, among young families with children, small increases in income appear to lead to substantial in creases in the probability that the family will soon own its home. Furthermore, among young people with no children at all, home ownership is rare at any income level. Clearly, economic projections of home ownership need to be based not just on aggregate statistics of income increases, but on who received them.

The effective use of survey data to shed light on such problems requires concentrating attention on the differential relation of income to economic activity as well as studying its effect in the aggregate.

These details are further illustrated by an examination of Figure 1. In the total sample the regression of Y on X_1 is

$$Y = a_1 + b_1X_1 + u_1 \qquad (6)$$

However, when the sample is split on variable X_2 into 2 groups, one with $X_2 = 1,2$ and the other $X_2 = 3,4$, for the latter group this regression is

$$Y = a_3 + b_3X_1 + u_3 \qquad (7)$$

and if, in addition, this group is split on variable X_4, then for $X_4 = 4$ or 5, the regression of Y on X_1 is

$$Y = a_7 + b_7X_1 + u_7. \qquad (8)$$

This is illustrated in Figure 1.

If we also have

$$\Sigma u_1^2 > (\Sigma u_2^2 + \Sigma u_3^2) \qquad (9)$$

and

$$\Sigma u_2^2 > (\Sigma u_4^2 + \Sigma u_5^2) \qquad (10)$$

and

$$\Sigma u_3^2 > (\Sigma u_5^2 + \Sigma u_7^2) \qquad (11)$$

it is clear that the various effects of X_1 on Y as revealed by the differences in the slopes (b_i) and the intercepts (a_i) are associated with the joint occurrence of the conditions denoted by the indicated values of variables X_2, X_3 and X_4. Thus one must devise a means of searching various sample sub-groups in order to learn whether these differential effects exist and what forms they take, under adequate constraints to reduce the probability of detecting differences which are spurious. It is this problem which lends itself to solution via a formal sequential decision process programmed to run on a

TABLE 2.

Analysis of Variance for Regression

Source of Variation	D. F.	Sum of Squares	Mean Square
Regression estimates around $\bar{Y}$	1	$\sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2$	$(N-1)r^2 s_y^2$
Observations around regression estimates	N-2	$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$	$\frac{N-1}{N-2}(1-r^2)s_y^2$
Observations around $\bar{Y}$	N-1	$\sum_{i=1}^{N} (Y_i - \bar{Y})^2$	s_y^2

Where $r^2 = \dfrac{[\Sigma(X-\bar{X})\ (Y-\bar{Y})]^2}{\Sigma\ (X-\bar{X})^2 \Sigma (Y-\bar{Y})^2}$

$\hat{Y} = \bar{Y} + b_{yx}(X-\bar{X})$

$s_y^2 = \dfrac{\Sigma(Y-\bar{Y})^2}{N-1}$

$b_{yx} = \dfrac{\Sigma(X-\bar{X})\ (Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$

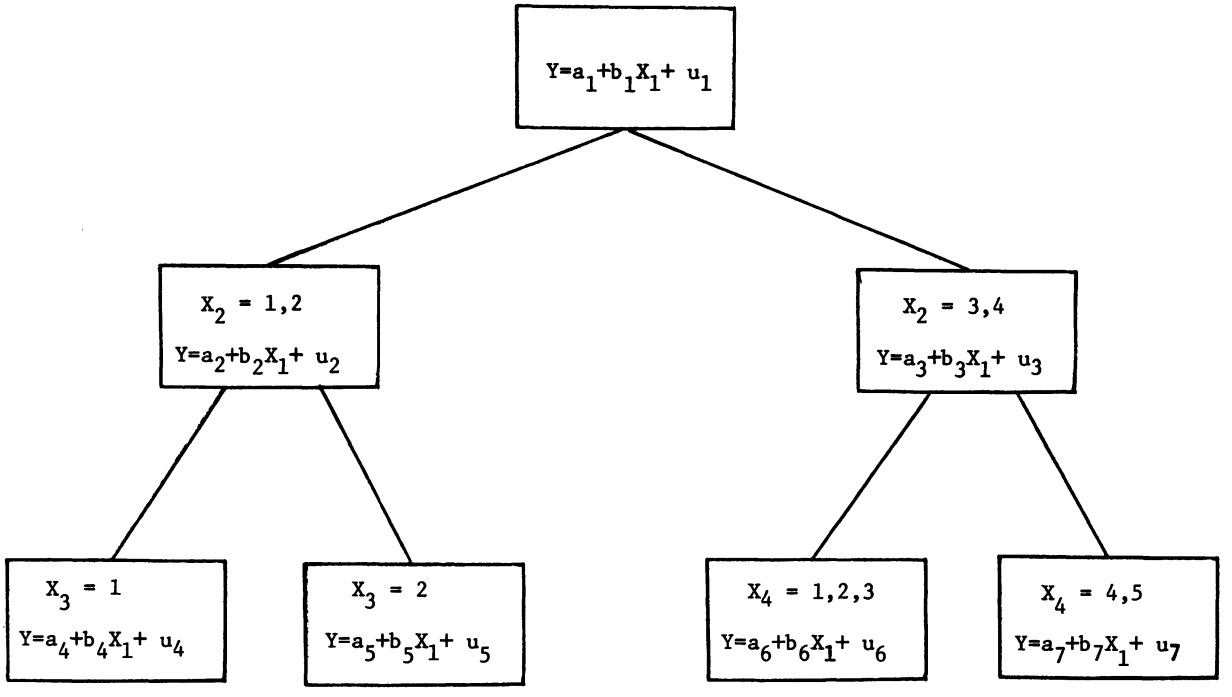

Figure 1. Differential Effects of X_1 on Y_1.

TABLE 3.

Logical Models for Two Dichotomous Predictors and a Trichotomous Dependent Variable

Configuration	$\bar{A}\bar{B}$	$A\bar{B}$	$\bar{A}B$	AB Y=0	AB Y=1	AB Y=2
Column(j) / Row(i)	1	2	3	4	5	6
1	0	0	0	A[a]	CS	CS
2	1	0	0	CS	BICOND	(M)BICOND
3	2	0	0	CS	(M)BICOND	BICOND
4	0	1	0	CS	A	(M)CS
5	1	1	0	A	SC	R
6	2	1	0	(M)CS	R	(M)BICOND
7	0	2	0	CS	(M)CS	A
8	1	2	0	(M)CS	A	(M)SC
9	2	2	0	A	(M)SC	SC
10	0	0	1	CS	A	(M)CS
11	1	0	1	A	SC	R
12	2	0	1	(M)CS	R	(M)BICOND
13	0	1	1	EXOR	SC	A
14	1	1	1	SC	A	CS
15	2	1	1	A	CS	BICOND
16	0	2	1	(M)EXOR	R	(M)SC
17	1	2	1	R	CS	A
18	2	2	1	(M)SC	A	SC
19	0	0	2	CS	(M)CS	A
20	1	0	2	(M)CS	A	(M)SC
21	2	0	2	A	(M)SC	SC
22	0	1	2	(M)EXOR	R	(M)SC
23	1	1	2	R	CS	A
24	2	1	2	(M)SC	A	SC
25	0	2	2	EXOR	(M)EXOR	SC
26	1	2	2	(M)EXOR	EXOR	SC
27	2	2	2	SC	SC	A

[a] A means additive; CS means cumulative upwardly and substitutive downwardly; SC is the inverse of CS; (M) means "modified;" BICOND means "biconditional;" R means reversal and EXOR refers to "exclusive or". Table adapted from Sonquist (1969a).

large scale computer.

Extending the Search Algorithm - A Multi-step Look-ahead

Experimentation with the original algorithm to determine its behavior under known conditions was carried out using contrived data. This process of working out tests under a variety of different conditions led to the development of a typology of multivariate models. From this viewpoint, multivariate interactive and additive AID and MCA models could be viewed as eighty-one variants of the same basic structure, which, in its simplest form (two dichotomous predictors and a trichotomous dependent variable) could be defined almost entirely in terms of the fundamental operators of Boolean algebra, (see Table 3)[1], Eliminating permutations and inverses reduces the number of models to seven non-trivial ones, (see Table 5)[2].

This typology also delineated exactly certain limitations of the original AID algorithm, some of which could not have been noticed earlier. For instance, while the published procedure was found to be capable of dealing adequately with many two-way interactions, others were identified as being difficult for it to deal with (Sonquist and Morgan, 1964). These were seen to consist of interactive relations characterized by consistency, i.e., by balance or symmetry. One such example is the biconditional model, illustrated in Table 5.

Table 4.

Occurrence of Effect "C" Associated with the Joint Occurrence or Joint Absence of Two Causal Factors, A and B.

Condition "A"	Condition "B"	
	B	Not-B
A	C	Not-C
Not-A	Not-C	C

[1] A more extensive discussion of these models is given in Sonquist (1969a)

[2] The eighth is a constant in all cells.

Although the frequency with which variants of this model actually occur in real data is not known, it is notable that at least one realization has received considerable attention in the recent sociological literature, the concept of status inconsistency.[1] The problem of developing analysis techniques to deal with this class of models is of importance for economic, educational and psychological research as well as for sociology.

It can be seen that the earlier sequential partitioning algorithm which examines only the "zero-order" effects of A and B separately could not discover the consistency effect in these data. There are really two A "effects" and they cancel each other out in the total group. Moreover, the additive assumptions implied by the choice of Multiple Classification Analysis would also tend to conceal the real state of the world.

However, the extended AID algorithm partitions the sample tentatively, first on one causal variable and then on the other (as well as making tentative partitions on other variables). The actual partition is made so as to maximize the effects of several successive partitions. This makes it possible first to reveal a consistency effect to the analyst by means of appropriate output, then to make an appropriate partition, and finally, to continue with the rest of the sequential search procedure.

In general, such a two-split scanning algorithm appears capable of providing information adequate for the analyst to identify the two-way interactions existent in the data. It also appears able to provide leads or clues to the existence of three-way interactions. This is seen to be a simple extension of the way in which the present algorithm provides clues to the existence of two-way interactions. Thus, an algorithm which examines the cross-classifica-

[1] For an example, see Blalock (1966).

TABLE 5

Seven Logical Models for Two Dichotomous Predictors and a Trichotomous Dependent Variable

Key = Cell Averages

H = high cell mean
L = low cell mean
M = middle cell mean

1. Univariate

	B_1	B_2
A_1	H	H
A_2	L	L

2. Cumulative

	H	L
	L	L

3. Modified Cumulative

	H	M
	L	L

4. Additive

H	M
M	L

5. Biconditional

H	L
L	H

6. Modified Biconditional

H	M
L	H

7. Reversal

H	M
L	M

tion of p predictors simultaneously appears able to reveal terms composed of p raw variables regardless of the symmetry of the term. However, such an algorithm also appears capable of revealing a term involving p + 1 raw variables if the term is asymmetric.

For instance, if we have the three variable asymmetric model,"if A and B and C, then Y = 0, otherwise Y = 4," the algorithm using a two-split strategy would produce the sequence of partitions illustrated in Figure 2.

Of course the amount of computing required to search out combinations of three or more variables increases as an exponential function of the number of variables considered simultaneously. Hence constraints have to be put on the process to permit the eliminations of unpromising leads and thus the examination of the subsequent partitions. However, this does not appear necessary in the three variable case.

Extending the Search Algorithm - Premium for Symmetry

The original algorithm represented a step in the direction of specifying a statistical model so it fits the data rather closely. The introduction of the look-ahead principle moves further in this direction. However, it can be anticipated that increasing the "wiggling" ability of a model being fitted will also increase the probability of the analyst basing his theoretical model on data largely reflecting idiosyncratic characteristics of the sample under investigation. Hence additional constraints are needed that would tend to guard the analyst against over-fitting his model. One such constraint is a premium for symmetry. Thus, when a look-ahead is employed a capability can be provided for increasing the probability that if a given predictor is used to make a partition in a given way on one branch of a partition sequence it will also tend to be used similarly in the parallel branch. This principle of constraint toward symmetry is illustrated in Figure 3.

Once groups four and five have been created using variable B the symmetry question arises. The proposed partition of group two, the "Not-A's" into new groups could be accomplished using, say, variables B, C, or D, but not variables E, F, or G, since the latter show insufficient explanatory power. In the previous algorithm, the choice of a variable on which to base a partition would have been to compare B, C, and D and then choose the one capable of producing the greatest reduction in the unexplained sum of squares. The present proposal would alter the algorithm to require that if C or D were chosen over B, it would have to achieve a certain ratio of explanatory power when compared with that resulting from a partition based on B identical to the one already performed in the parallel group.[1] The comparison ratio would be supplied by the analyst at the time of execution of the program. Setting the ratio to 1.0 would simply cause the regular algorithm to take effect. Setting it larger than 1.0 would bias the algorithm toward symmetry; setting it at less than 1.0 would tend to prevent symmetry. For instance, a value of 1.25 would require that a non-symmetric partition explain 25 percent more variation than a symmetric partition in order to be actually used in a split.[2]

Preliminary Examination of the New Algorithm

As a preliminary investigation into the power of the new algorithm with respect to the look-ahead and covariance options, tests were made using all combinations of the seven logical models of Table 5 applied to both means and slopes (see Tables 6a and 6b).[3]

---

[1] In the case of dichotomous predictors the identical partition is the only possible one. This is not the case where the variable has three or more classes. We focus on total symmetry, i.e., forming identical sub-groups based on the same predictor.

[2] Appropriate values of the symmetry premium for actual use are yet to be worked out by experimentation.

[3] For each of the means model in Table 5a, a dependent variable was generated, $Y_i = a + e_i$, $i = 1,\ldots,100$, where (a) represents the mean of a given cell (25 observations were generated for each call) and e_i is a random error $e_i \sim (N(0,.5)$. The e_i remains constant for all 7 models. A covariate $x_i \sim N(0,1)$, $i = 1, \ldots,100$, was also generated, and for all 49 combinations of means and slopes the dependent variable $y_i = a + bx_i + e_i$, $i = 1, \ldots,100$, was formed, where (a) and (b) are the mean and slope values for the appropriate cell. The means-alone model, $y = a + e_i$ was also evaluated, making the total number of experiments fifty-six.

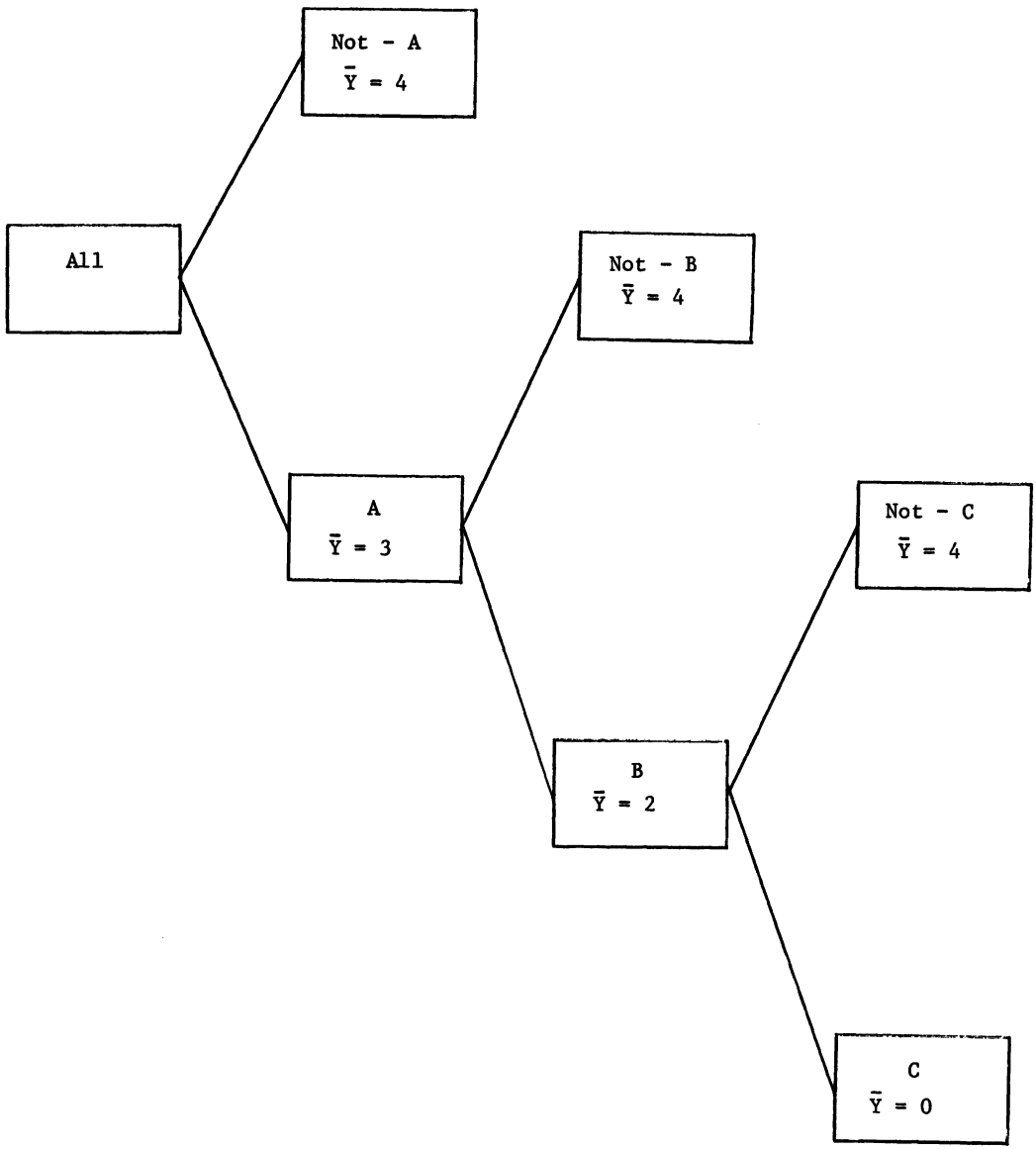

Figure 2. ABC Implies Y = 0, Else Y = 4

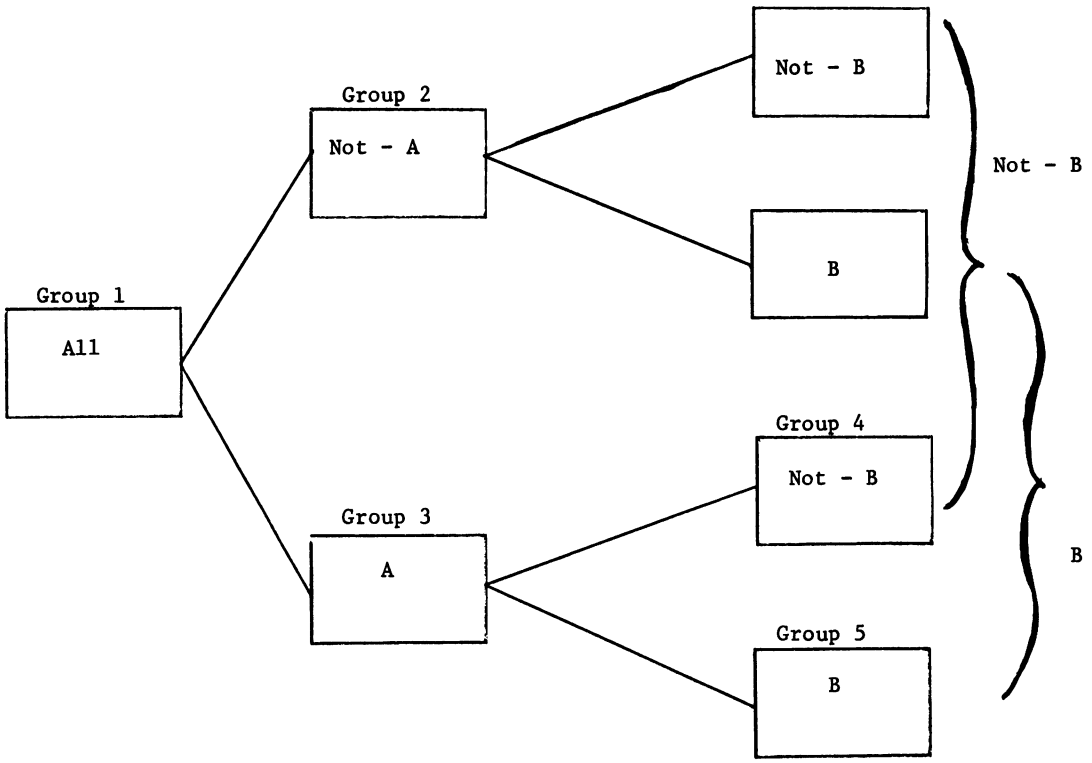

Figure 3. Symmetric Partition.

TABLE 6a

7 Logical Models Applied to Means

1. Univariate

	B_1	B_2
A_1	10	10
A_2	0	0

2. Cumulative

10	0
0	0

3. Modified Cumulative

10	5
0	0

4. Additive

10	5
5	0

5. Biconditional

10	0
0	10

6. Modified Biconditional

10	5
0	10

7. Reversal

10	5
0	5

TABLE 6b

7 Logical Models Applied to Slopes of Y on X

1. Univariate

	B_1	B_2
A_1	1	1
A_2	-1	-1

2. Cumulative

1	-1
-1	-1

3. Modified Cumulative

1	0
-1	-1

4. Additive

1	0
0	-1

5. Biconditional

1	-1
-1	1

6. Modified Biconditional

1	0
-1	1

7. Reversal

1	0
-1	0

Since this was a preliminary study, "masking" factors such as intercorrelations and noise in the predictors were not considered.[1] However, to provide a test of the algorithm's ability to pick out explanatory factors from a noisy background two uncorrelated dichotomous dummy variables were generated along with the two variables defining the model. Each experiment was made using the two real factors and the two noise factors as predictors, utilizing the "look-ahead" with two splits (creation of three groups).

For the means-alone cases, the sum of squares between the three terminal groups was maximized, i.e., from Table 1,

$$\sum_{i=1}^{3} N_i \bar{Y}_i^2 - N\bar{Y}^2 \qquad (12)$$

For the means and slopes cases, the quantity maximized between the three goups was the sum of squares resulting from both the reduction in means and regression, i.e., from Table 2,

$$\sum_{i=1}^{3}\left[N_i\bar{Y}_i^2 + (N_i-1)r_i^2 s_{yi}^2\right] - \left[N\bar{Y}^2 + (N-1)r^2 s_y^2\right] \qquad (13)$$

As was expected, with the exception of the biconditional model, the look-ahead with two splits yielded results similar to those from no look-ahead; the final groups in both cases were identical, but occasionally the look-ahead would lead to splits on the variables in reverse order from a parallel analysis using no look-ahead.

The biconditional-means model still presented some problems, however. For the no-slope version of the model as well as the univariate, cumulative, modified cumulative, additive and reverse slope models, the algorithm with no look-ahead could make incorrect splits; that is, it could use the dummy variables by mistake. it is significant, however, that the look-ahead identified the models correctly. However, it occasionally made subsequent partitions on the dummy variables. This proved to be capable of remedy by an adjustment of the reducibility criteria controlling the termination of the partitioning process.[1] Proper choice of this criterion still permits legitimate partitions to take place, but prevents subsequent spurious ones. For the no-look-ahead case, this fraction p_o apparently should be in the range $.005 \leq P_o \leq .008$. For the one-step look-ahead (two splits) this fraction P_1 apparently has a lower bound of $P_1 \leq .016$. The look-ahead apparently does the job it was designed for.

Our findings from this initial experimentation with covariance models suggest that the differences in means may be much more powerful in determining what split is to be made than are differences in slopes. In fact, in most of the cases where (a) the two groups had no differences in means and (b) one of these groups had slope zero, the two regressions resulting from tentative partitions were not sufficient to meet the reducibility criterion with, or without, a look-ahead. For instance, in the example given below in Table 7, the algorithm would split the sample on variable A, but would not split either of the resulting groups.

Table 7.

An Example of a Remaining Problem

	Means B_1	Means B_2		Slopes B_1	Slopes B_2
A_1	10	10	A_1	1	0
A_2	0	0	A_2	0	-1

[1] Further work is to include tests with predictors of varying levels of intercorrelation and skewness as well as the assessment of the algorithm activity to deal with correlated "noise."

[1] The criterion is the fraction of the total sum of squares from the total input sample that a split (or sequence of splits in the look-ahead case) must explain in order for the split actually to be made. The maximized function [equation (12) or (13)] must be greater than p_i times TSS. The behavior of the algorithm is in keeping with previous results from the original algorithm. For $p_o < .005$, there is a tendency to split on the dummy variables after the correct splits have been made. For $p_i < .016$ the look-ahead would occasionally split on a dummy variable even before making the correct splits.

While this result is not entirely unexpected, it may imply that maximizing differences in means alone or in regression slopes alone may be a more powerful tool[1] than the combination of the two. Further experimentation with these models using only the slopes is obviously indicated.

Significance of These Findings

This extended algorithm represents a continuation of our attempts to develop better methods for adequate handling of the problem inherent in applying multivariate statistical techniques in the analysis of cross-sectional survey data with large numbers of cases (1000 or more). The covariance capability is relevant for the analysis of panel as well as one-time cross-sectional data. The increasing volume of survey and other non-experimental data and the growing needs of researchers in many disciplines to utilize it in solving practical problems makes the development of adequate data-handling methods even more imperative. The procedures are frankly oriented toward the inductive phases of research in which model-building is the objective rather than toward the deductive model-testing phase. It is felt that a focus here will be of benefit both to theory builders and to those who must find immediate solutions to pressing problems.

---

[1] It is possible that this behavior results largely from the sizes of the mean and slope effects relative to the noise level. This remains a problem for further exploration.

References

Andrews, F.M., Morgan, J.N., and Sonquist, J.A. Multiple Classification Analysis. Ann Arbor: University of Michigan, Institute for Social Research. 1967.

Aptakin, Peter. Automatic Interaction Detector. New York: Service Bureau Corporation, Computing Sciences Division. 1965.

Arpi, Bo. En Modell for Optimal Marknadssegmentering. Lund Business Studies No. 4. Studentlitteratur Lund. 1967.

Biervert B., Dierkes, M., and Walzel, A. Automatischer Split zur Optimalen Kombination Erklarender Faktoren. Forschunsstelle fur Empirische Sozialökonomik und Rechenzentium, University of Cologne, March, 1968.

Blalock, H.M. The identification problem and theory building: The case of status inconsistency. Amer. Soc. Rev., 31, 1, 52-61.

Campbell, Robert H. Modifications to the Automatic Interaction Detector COMCOM/Simulation Memo No. 26. Program operational on the MIT-FMS system. Cambridge: Massachusetts Institute of Technology. 1965.

Caplan, N., Lippitt, R., Gold, M., Mattick, H., Suttles, G., and Deshaies, D. The Outcome of the Chicago Youth Development Project. New York: Boys Club of America Convention. 1966.

Carlson, W.L. Identification of the Problem Driver from Driver Records: A Preliminary Analysis. Ann Arbor: Highway Safety Research Institute, University of Michigan. 1968.

Gensemer, Bruce L., Lean, Jane A., and Neena, William B. Awareness of marginal income tax rates among high-income taxpayers. National Tax Journal, XVIII, September, 1965, 258-267.

Kay, Kevin, Automatic Interaction Detector: "AID" Translater into CDC 3600 Fortran and Compass, Technical Report 46. East Lansing: Michigan State University, Computer Institute for Social Science Research. 1966.

Land, K.C. The Sociology Program Library, A User's Manual. Austin: Population Research Center, Department of Sociology, University of Texas. 1968.

Lord, F.M. Elementary models for measuring change. In Problems in Measuring Change, C.W. Harris (ed.) Madison: University of Wisconsin Press, pp. 21-39. 1962.

Marks, E. NCHS AID Program. Unpublished manuscript. Wharton School, University of Pennsylvania. 1965.

Morgan, J.N., and Sonquist, J.A. Problems in the analysis of survey data: And a Proposal. Journ. Amer. Stat. Assoc., 1963(a), 58, 415-434.

Morgan, J.N., and Sonquist, J.A. Some results from a non-symmetrical branching process that looks for interaction effects. Proceedings of the Social Statistics Section, American Statistical Association, 1963(b).

Ross, J., and Bang, S. Predicting the adoption of family planning. Studies in Family Planning, No. 9. 1966.

Sarlvik, Bo. Socio-Economic Predictors of Voting Behavior: Research Notes from a Study of Political Behavior in Sweden. Unpublished Manuscript. 1968.

Snowbarger, Marvin. An Interaction Analysis of Consumer Durable Expenditures. Unpublished Ph. D. thesis, University of Michigan. 1966.

Sonquist, J.A. Automatic Interaction Detection and Multiple Classifiecation Analysis: The Validation of a Search Strategy. Unpublished Ph. D. dissertation, University of Chicago. 1969(a).

Sonquist, J.A. Finding variables that work. Public Opinion Quarterly, 1969(b), Vol. XXXIII, No. 1, pp. 83-96.

Sonquist, J.A. Simulating the research analyst. Social Science Information, 1967, Vol. 6, No. 4, p. 207-215.

Sonquist, J.A., and Morgan, J.N. The Detection of Interaction Effects. Ann Arbor: Survey Research Center Monograph No. 35, Institute for Social Research, University of Michigan. 1964.

Walker, H.M., and Lev, J. Statistical Inference. New York: Holt, Rinehart and Winston. 1953.