SOME RESULTS FROM A NON-SYMMETRICAL BRANCHING PROCESS THAT LOOKS FOR INTERACTION EFFECTS 1/

James N. Morgan and John A. Sonquist
Survey Research Center
Institute for Social Research
University of Michigan

This paper presents some results of a data reduction process designed, programmed and operative on the IBM 7090. 2/ It was designed with sample survey data in mind, that is, for data characterized by several thousand cases, a large number of explanatory variables or classifications, moderate intercorrelations among the predictors, and a continuous dependent variable, not badly skewed, but with a good deal of unexplained variation or noise.

Proponents of a new statistical procedure are always in danger of claiming too much for their method, both in terms of how original it is and in terms of how good it is. The program is original in scope but not in essence. It represents a simulation, with some added breadth and quantification, of what careful researchers have done for years by hand or using an IBM sorter and tabulator when investigating a new set of data.

Simulation of human behavior by a computer is not new. Ours is a particular kind of simulation not so much designed to gain insight into the behavior simulated, but to do a particular kind of job better than the human has the time or patience to do. We have systematically examined the behavior of a social science researcher tackling a particular kind of data analysis problem, making decisions, isolating interesting subgroups and computing statistics. We then stated the behavior explicitly and formally as a sequential set of decision rules and extended them to what the researcher might do if he had the time and patience. An examination of the preliminary results of the program indicates, among other things, that the decision rules of a researcher are more complicated than he realizes. We are now incorporating changes into our model and into the program to reflect these more sophisticated rules. But we have made a start in what we feel will be a fruitful line of development; the simulation of the researcher by machine. There are some unsolved problems of optimal strategy which we raise in the hope of stimulating further work along these lines.

So the basic idea is not new. What is new is the formalization of the analysis procedure and the capacity to apply it systematically and rigorously, so that unrealized compromises and arbitrary choices do not occur, and the results are completely reproducible.

There are some things this analysis technique will not do. It will not locate the best functional form in a set of data with a limited number of numerical predictors and a relatively low level of error or noise. That is what Professor Westervelt's process does well, and you will hear about that later on this program.

Our procedure also will not answer the question whether a particular variable has a significant effect on the dependent variable, the other variables somehow controlled or "held constant". It was designed for a set of data full of interaction effects, and wherever there are interaction effects, by and large, it is not meaningful to ask about the direct effects of one variable at a time. It is difficult to give up the habit of testing for the effect of one variable after another. Yet in much behavioral research we measure not the theoretical factors in which our interest lies, but rather the measurable, proxy factors which, we hope, may reflect more basic characteristics. These often must interact (in the statistical sense) to be able to represent the theoretical construct adequately. For instance, "family size" may be used to represent "the amount of housing space needed," but in combination with "income" also may be used as a proxy for "ability to pay for housing." The procedure can, however, minimize the noise for a tight test of the effect of a single factor.

The procedure is not designed for a very large number of highly correlated predictors such as batteries of attitude questions. Also, since at present, it does not look ahead more than one step at a time, it will not locate certain completely offsetting symmetrical negative interaction effects where neither factor has any effect by itself. For example:

% Who go to Hospital			
	Men	Women	
Young	2%	8	5
Old	8	2	5
	5	5	5%

But let's see what the program will do. We shall describe the process in terms of its purpose. (The formal algorithm is appended). The program divides the sample through a series of binary splits into a mutually exclusive set of subgroups. Every observation is a member of exactly one of these groups. These subgroups are chosen so that their means account for more of the total sum of squares (reduce the predictive error more) than the means of any other set of subgroups. The stopping point is subject to arbitrary decision and is set by parameters at the beginning of the computer run. (The present set of rules for stopping represents the best we have so far, but may not necessarily be optimal in terms of research strategy. Further work needs to be done in this area.)

At any 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 in that sample, from the information available. If the sample is representative, this is the best scheme for the population.

There are some minor qualifications to these claims, but for large samples and without certain symmetrical negative interactions, they seem to be valid.

In deciding which split to make, the rule is to scan all feasible splits and select the one which reduces the error sum of squares the most. This is a rule of importance, not significance. Subgroups which are significantly different, but which are so small that isolating them does not help in predicting a randomly selected individual (or group), are not split off. Why the rule of importance rather than significance?

Multivariate statistical techniques have been developed to the point where the argument can be made that significance tests are of doubtful value because of the large number of variables tried. With samples in the range of 1,000 to 3,000 observations many factors show up as statistically significant which are not important, in terms of their contribution to reducing predictive error. (They may, of course, have theoretical importance, for some reason.) The process we are using of scanning for all feasible splits at each stage vastly increases the number of things tried, so that the whole notion of degrees of freedom seems useless in this model. Formalizing the process makes this more obvious.

At each stage, the computer selects the group with the largest sum of squares within it, locates the best way of splitting it into two subgroups using each predicting classification, then takes the best of the best (on the basis of the largest between-groups sum of squares). We do not need to examine all possible combinations of classes of each predictor when separating a group into two parts, since it can be shown that after rearranging the classes into descending (or ascending) sequence according to the size of the class means of the dependent variable, it never pays to combine non-adjoining classes. [3] Thus, with k classes for predictor X: only k-1 feasible splits exist. But a little algebra shows that with a dozen predicting classifications of, say, eight classes each (7 feasible splits each), and with this scanning repeated at each split, the number of different trees theoretically possible is, to put it mildly, somewhat larger than the sample size. There are clearly no degrees of freedom left.

The rules for stopping overlap. A safety precaution puts a maximum on the number of final, unsplit groups at, say, 20. No group containing only a small internal sum of squares (less than 2% of the original total sum of squares) is examined further. And no split is allowed unless it reduces the overall predictive error by at least a visible amount like ½%. (That is, the between sum of squares for the split must be more than ½% of the original total unexplained sum of squares for the whole sample).

We can summarize the rules then:

> Take the group with the largest unexplained sum of squares within it, so long as it is more than 2% of the original sample's unexplained sum of squares.
>
> Considering all predictors, find the best binary division of that group, in terms of the "between splits sum of squares", so long as it is greater than ½% of the original sample's unexplained sum of squares. (If no split on this group is worth while, try the group with the next largest internal sum of squares.)
>
> If no group can be found worth examining, or if none of those which are found can be profitably split, then stop.

The details are in our article in the June Journal of the American Statistical Association. [4]

Note that all predictors are treated as classifications, even if this means making classes out of a continuous variable. It has been felt that the small loss of information from grouping is offset by the flexibility in discovering nonlinear relationships. Dummy-variable regression models are similarly attractive in this respect. [5]

We turn now to some actual results. Chart 1 and Table 1 are a relatively simple analysis of a dichotomous dependent variable: whether or not the spending unit owns its own home. Comparisons between the tree and multiple regression findings are not easy, but one way is to ask which predictors appear most important. For this purpose we use from the new analysis the total reduction in error from splitting on that predictor, whether used once or more than once. For dummy variable regression we use a partial beta coefficient, squared to put it in the same dimensions. [6] This can be thought of as the partial beta coefficients one would get if he took the dummy variable coefficients for each predictor to create a new scale or variable, and ran a multiple regression with these new scaled variables.

We give also the gross beta coefficient squared, which can be thought of as the square of a dummy variable multiple correlation coefficient using dummies to represent all the classes

(except one) of that particular predictor. It tells how much of the unexplained sum of squares can be accounted for by using only that one predicting classification in all its detail.

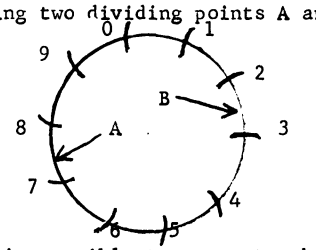
It is apparent that roughly the same variables appear important in both analyses. Does the tree really tell us anything that we didn't already know from the regression? Clearly it tells us that married people, and people with higher incomes become home owners earlier in life. In other words, the effects of age on home ownership depend on other things. Stated another way, the effect of income on home ownership depends on age and family status. There are interaction effects.

The total explanatory power of the two methods cannot be compared, since the regression uses all classes of each predictor, and since the newer analysis could always "explain" more if we allowed it to continue splitting. It is not even legitimate to argue that the new method explains more per unit of information utilized, since it has, in trying and discarding, used all the information the regression used, and with no restrictive additivity assemptions. But the tree provides an economical way to summarize a lot of data, and to make comparisons with other times or places.

Chart 2 and Table 2 move to a numerical dependent variable, annual medical expenses, based on a sample of individuals in Michigan.[7] Here again the same variables seem important, but while the regression indicated that good insurance coverage was associated with heavier utilization of medical services, the tree indicates that this is true to an important extent only for adult females. This makes sense to the analyst and certainly to someone concerned with keeping insurance costs down, this knowledge is of more use than the regression finding that insurance affects utilization.

Some of the problems with our original strategy can be seen in the tree illustrated in Chart 2. Why the particular combination of family sizes that puts 2.0 and 4.0 equivalent adults[8] together? It turns out that these are mostly people with 1 or 3 children. Perhaps they are the women who had a child born last year, but if so, the sub-group combinations appear partly fortuitous. We have allowed reordering of the subclasses of a predictor so that, for instance, the middle-aged could be separated from the old and young. The tendency for adult women with either low or high income to have higher medical bills may be real. But there may be places where we want to retain the order, if only for simplicity.

One way out of this impasse, which we are thinking of building into a new program, is to prevent the linear re-ordering of the classes of some predictors according to their means. Instead, an option would be provided to treat them as a circle, with identifying codes running from 1 to 9, thence to 0 and back to 1 again. Dividing a group into two sub-groups then requires determining two dividing points A and B. For example:



Thus, it is possible to separate the middle from both extremes without allowing erratic combinations. We shall surely add to the program a still more restrictive option that preserves the original linear scale ordering and allows only a split along that ordering. This option is being incorporated into the computer program in such a fashion that the researcher may specify which predictors are to have a split restriction, if any, placed on them.

Turn now to Table 3. The first use of the new program, with hourly earnings as the dependent variable, kept to the predicting classifications used in the regression analysis presented in the book Income and Welfare in the United States, by J. N. Morgan, M. David, W. Cohen and H. Brazer.[9] Here we have a relatively obvious example of the problem of variables at different stages in a causal process. Some of the explanatory variables used come earlier in time or are otherwise logically prior, and can help determine the levels of other explanatory variables, but cannot be affected by them. Age, race, sex are determined at birth, and will influence how much education a person gets. All these, including his education, will influence his occupation, and the cumulative set will help determine his present hourly earnings. In a case like this the analysis should be sequential if it is intended to explain the process, not merely predict. The clearly prior (exogenous) variables should be used first, and the residuals run against a more inclusive set of predictors. It is necessary to reintroduce such exogenous variables as race into the next stage because they may well mediate the effect of other things (education) on the dependent variable (earnings). The revisions being introduced into the program make such a sequential analysis easier.

At any rate, it is clear that if one puts both occupation and education in the same stage, occupation (being a more powerful predictor) "takes over" and, in the sequential design we use, education cannot fight its way back into the analysis as it does in multiple regression. This is a characteristic of sequential procedures when several predictors are intercorrelated, and the analysis problems result from the nature of the question being asked. If one wants only the best prediction, then even causal patterns among the predictors can be ignored. If one

wants to unravel the causal mechanism then a several-stage analysis is called for. Even among correlated predictors at the same stage, the one that is best at an early split may explain enough of the variance so that the others may not show up at all.

Chart 3 shows the tree omitting two predictors, occupation and supervisory responsibility. Here education, age, and sex are clearly important. There are some other interesting findings. Achievement motivation, a variable interesting for theoretical reasons, turns out to be important in explaining hourly earnings only for middle aged college graduates. This is an interesting and meaningful finding, since these tend to be the people who are best able to affect their own hourly earnings by their attitudes and efforts. Most other people have to work longer hours or take a second job if they want to earn more money. This type of interpretation is, of course ex post facto, and should be validated by examining its implications and performing additional analyses to test the interpretation.

With regressions, it is possible to compute the sampling error of each dummy variable regression coefficient, or to make F-tests for each set. These tests are of doubtful validity when applied to multi-stage clustered samples, often weighted to adjust for sampling and response rate variations. There seems to be even less theoretical basis for computing sampling errors for one of our trees. Another sample might produce an entirely different sequence of splits. But the sampling stability of the branching process is, nevertheless, of some interest.

One way to investigate the stability of the branching process is to repeat it on split half samples. The results from three such subsamples are given in Table 3. The predictors which are seen to be important seem not to vary from sample to sample. On the other hand, the trees are different, sometimes even at the first split. The similarity occurs in terms of the groups which finally result, because one can isolate groups which are nearly the same by splitting in different orders. For instance, one can select first college graduates, then middle aged, then men, or first men, then college graduates, then middle aged, etc., arriving the same place by a different route.

Another way of investigating stability, which we are working on, is to take a tree derived from one sample and ask how well those final groups predict in a different sample.

Chart 3 still has predictors at several stages, sometimes combined into a single classification such as where the head of the unit grew up and where he lives now. This was done originally to build some interaction effects into the regression analysis, that is, to investigate mobility. How do we know that more basic things like race are not being pushed "out of the tree" by other things which are largely the result of race in the first place? To answer problems of this type, we must use a multi-stage analysis.

Chart 4 and Table 4 represent the first stage of a more detailed step-wise analysis where we use only the predictors that were determined in early childhood, (N/Ach is so determined, in theory). The smaller number of predictors with fewer subclasses leads to a tree that is easier to look at and interpret. It is clear that there are interaction effects, for the tree would be symmetrical if there were not. Indeed, one gets the impression that the interactions are of a particular kind which can be interpreted by saying that disadvantages are substitutes for one another while advantages are complements. Having one or two disadvantages is enough, and further splits on the others will not explain additional variation and hence will not appear. Being old, or young, or uneducated, or a woman, or from a southern or rural background, or nonwhite, are alternative barriers to high earnings. In this analysis those predictors which affect many people severely, tended to be used early in the tree. Those affecting a minority, like race, tend not to appear because we can explain enough by knowing the other things. This does not mean there is no prejudicial discrimination. Rather it reflects a characteristic of the analysis that it does not test each explanatory factor holding all the others constant, but asks whether a factor is needed more than any other factor, given the group currently under consideration.

The extreme case of such substitutability among predictors would be a tree where once a group was split off as having one disadvantage, it would never be split again. Whenever groups are split reflecting extreme disadvantages (being very old, very young, or a woman) they tend not to be split again in Chart 4. Similarly those with very low education tend not to split further.

Persons accustomed to one method of analysis always like to translate problems back into solutions with the familiar method. With regression analysis, such patterns would clearly require not simple cross-product terms, but a new set of dummy variables like:

(1) The man has one of the following disadvantages:
(2) He has two or more of them

If interactions behave like those described above, then such variables will take over and little credit will be left for each of the separate components.

In our original trees we tried to include the card counts and standard deviations of the means, but it was too much to look at. These

data are all part of the output, however. The program also provides details of the subclass means just before each split is made, and the original subclass means for each predictor.

In looking for ways to present the data, we have tried one other method which is illustrated in Tables 5 and 6. Here we select one predictor and show the way the splits were made and the definitions of the subgroups just before they were split.

Table 5 shows the three ways in which education was used, once to divide the whole sample, and then on two different subgroups. Perhaps one reason why a second split on education was not made earlier, before looking at where the individual grew up, is that the critical educational level seems to differ depending on the person's background. For those without the disadvantage of a farm or southern background, it appears to be graduation from high school that matters, for the others whether they even learned to read.

Table 6 shows the four groups split on age, and how the splits were made. Clearly "middle-aged" people earn more in general, but for some groups (generally those with more education and fewer disadvantages) "middle age" starts at 35, while others reach it at 25. (The college graduate group contained one person 75 or older who was making $4.04 an hour and was grouped together with the "middle aged" by the computer.)

There is one problem which is not well illustrated in the data we present here: With a large number of predictors of many classes each, there is an increased possibility of fortuitous or untrustworthy splits. When the investigator looks at them, he immediately becomes aware that something is missing in a strategy which is willing to make <u>any</u> split that is important enough to reduce the error sum of squares by half a per cent. In other words, in arguing that it is importance that matters, not statistical significance, we tend to over-simplify the research problem. A better rule might be to proceed according to the importance of splits, but disregard any split, even if it appears important, if it is not significant, i.e., may be a fluke.

How can something be important but not significant? Whenever there are any extreme cases, or sufficient flexibility in combining codes on any of many variables and for smaller and smaller subgroups, this will allow the process to isolate some subgroup consisting of a few extreme cases. The formal relations between the number of cases, the between sum of squares, and the F-test need to be worked out, but it is clear from the fact that the square root of N appears in the denominator of the estimated sampling error, that subgroups of fewer than ten cases are unlikely to have means that differ <u>significantly</u>. Hence we are building into the program a side rule that no split is allowed if one of the resulting groups contains fewer than n (for example 10) cases. Further work clearly needs to be done in this area. Is it the size of the subgroup before splitting which should be above some minimal point? Should the smaller of the two subgroups split off? Answers are not yet forthcoming. One may also argue that small subgroups split off from large ones are deviant cases which <u>should</u> be removed from the main analysis and then examined in great detail.

Where does one go from here? Ideally the subgroups identified by the branching process should lead to the development of some new theoretical constructs, new variables that are combinations of the measured factors and have theoretical meaning as well as practical significance. Having defined these new variables, one could use them in an ordinary regression analysis for presentation purposes (and marvel at how well one explained things). The theoretical question raised is "why are these variables important". And, again, further analysis must provide the answers.

Some studies of the forecasting stability of this method compared with multiple regression would also be useful, as we have pointed out.

SUMMARY

It is clear, at least to us, that for purposes of discovering the structure of relations in a body of data, what really is related to a dependent variable, under what conditions, and through what intervening processes, this procedure offers some real advantages. Its strategy is to focus on what can be found out from the data with some assurance, rather than on testing the significance of effects of many factors and their cross-products, the results of many such tests being basically inconclusive rather than negative.

There are clearly some neat unsolved statistical problems of optimum strategy, or at least consistent strategy in setting the various arbitrary cut-off points. The 2% of total sum of squares before a group is examined, the ½% reduction in error before a split is allowed, and the minimal number of cases in a subgroup, should all depend on the sample size, the number of predictors, the constraints on rearranging scales, and the variance of the dependent variable, relative to its mean.

The trees look formidable at first, but are basically simpler than multiple regression results; the results to be presented being the definition of a subgroup and its mean on the dependend variable

We hope we have now started to come full circle to the point where the computer is doing what we want to do better, rather than doing incredible amounts of arithmetic, the results of which often do not meet the real needs of the analyst.

FOOTNOTES

1. A paper presented at the Meetings of the American Statistical Association, Case Institute of Technology and Western Reserve University, Cleveland, Ohio, September, 1963.

2. A great deal of credit is due to Kathleen Goode and Wen Chao Hsieh of the ISR programming section, who did the programming. The program is identified as the Automatic Interaction Detector (AID) Model 1. It is written in MAD for a 32K IBM 7090. It operates using the U. of M. Executive System Monitor. Our thanks are due to Dr. R. C. F. Bartels of the U. of M. Computing Center on whose equipment this experimentation took place.

3. A proof of this is due to Professor William Erickson of the U. of M., Computing Center and Mathematics Department.

4. Morgan, J. N., and Sonquist, J. A., Problems in the Analysis of Survey Data - and a Proposal, JASA, 58, (June, 1963), pp. 415-34.

5. See Suits, D. E., Use of Dummy Variables in Regression Equations, JASA, 52, (dec. 1957,) pp. 548-551.

6. See Andrews, F. M., The Revised Multiple Classification Analysis Program, Institute for Social Research, University of Michigan, August 1963, 13 pp. Multilith.

7. See Grover Wirick, Robin Barlow and James Morgan "Population Survey: Health Care and its Financing" in Walter J. McNerney et al, Hospital and Medical Economics (2 Vols.) Vol. 1, pp. 61-360, Chicago, Hospital Research and Educational Trust, 1962.

8. In the scale, the second adult, and children under 12 are counted as ½ each.

9. New York, McGraw Hill, 1962.

TABLE 1 --

WHETHER SPENDING UNIT OWNS ITS HOME

Predictors	Gross Beta Coefficients[2]	Multiple Classification Partial Beta Coefficients[2]	AID Analysis-Reduction in TSS(I)/TSS(T)
Age of heads	.111	.099	.107
Income	.088	.068	.040
Number of persons	.088	.062	.084
Unusual income last year	.039	.010	.000
Race	.014	.003	.000
Number of persons earning $600	.011	.003	.000
Education of heads	.001	.002	.000
R^2		.251	.231

CHART 1 --

AID ANALYSIS OF
WHETHER SPENDING UNIT OWNS ITS HOME

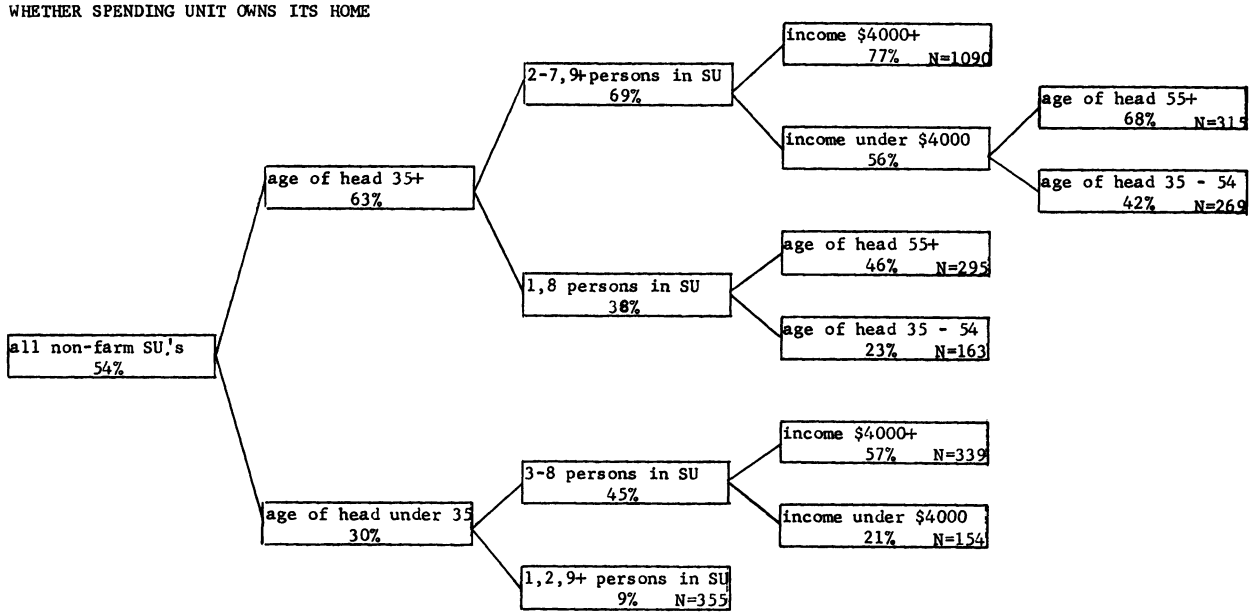


TABLE 2 --

INDIVIDUALS' MEDICAL EXPENSES

Predictors	Gross Beta Coefficients[2]	Multiple Classification Analysis- Partial Beta Coefficients[2]		AID Analysis- Reduction in TSS(I)/TSS(T)
		Males	Females	
Sex	.012	-----	-----	.016
Age	.041	.033	.069	.043
Health insurance coverage	.010	.011	.017	.007
Family income	.014	.006	.008	.005
Equivalent adults in family	.019	.004	.002	.007
Attitude toward early care	.002	.004	.004	.000
Education of head	.003	.003	.002	.000
Region where head grew up	.007	.001	.002	.000
Service level	.005	-----	-----	.000
R^2	----	.077	.089	.078

CHART 2 --
AID ANALYSIS
OF INDIVIDUALS' ANNUAL
MEDICAL EXPENSES

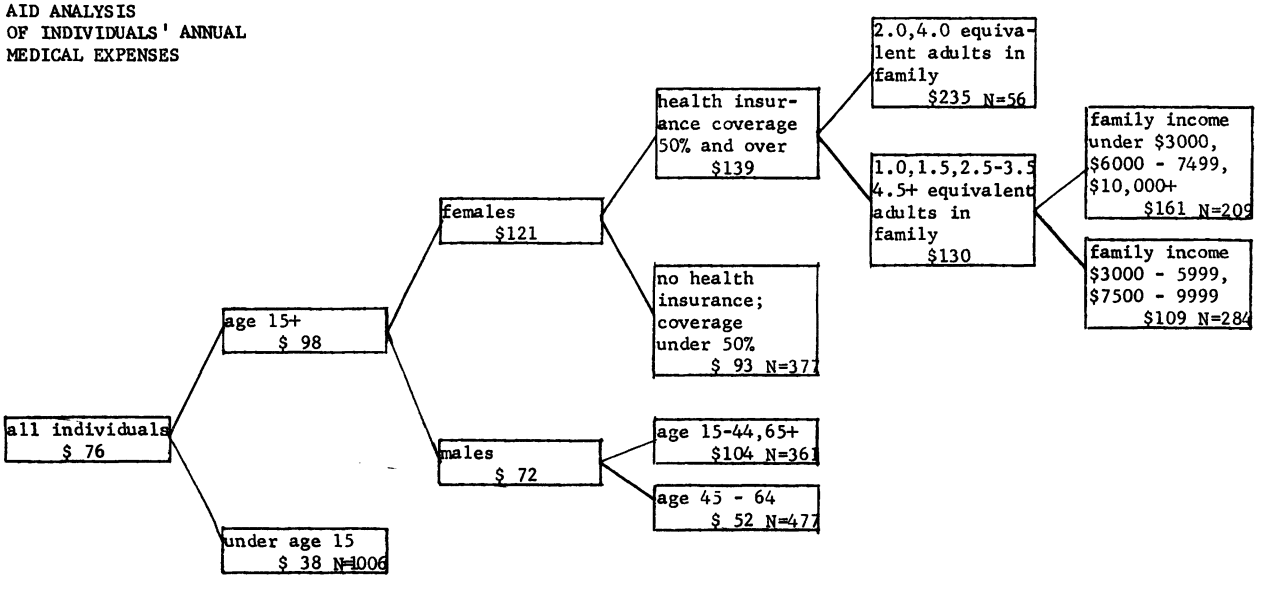


CHART 3 --
AID ANALYSIS
OF HOURLY EARNINGS
OF SPENDING UNIT HEADS

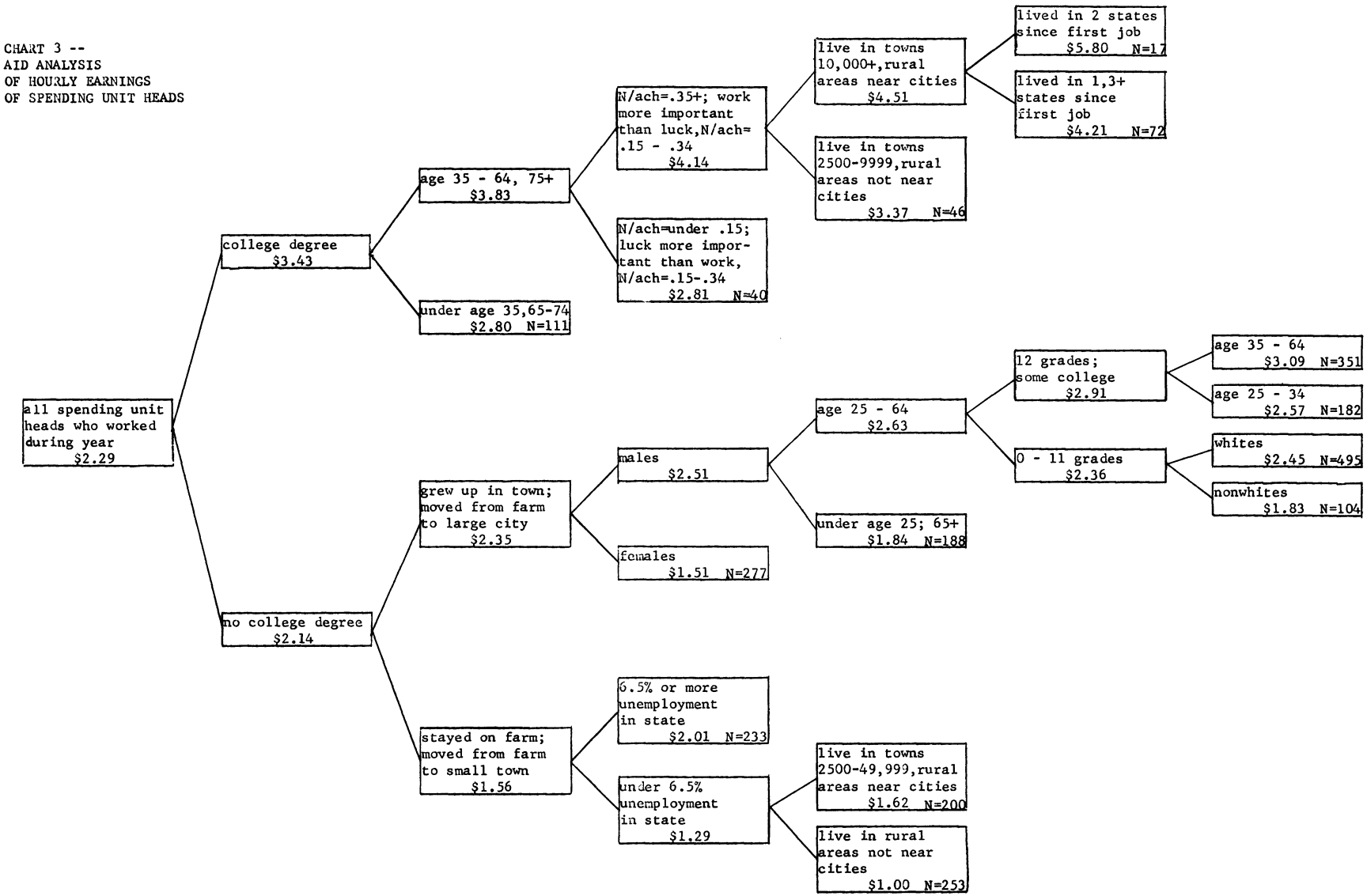


TABLE 3--

HOURLY EARNINGS OF SPENDING UNIT HEADS

Predictors	Gross Beta Coefficients[2]	Multiple Classification Analysis-- Partial Beta Coefficients[2] With Occupation	AID Analysis-- Reduction in TSS(I)/TSS(T) With Occupation	Without Occupation Total	Split Half #1	Split Half #2	Split Half #3
Education	.133	.055	.014	.097	.073	.101	.083
Age	.039		.035	.040	.052	.039	.068
Sex	.045	.048	.039	.042	.062	.046	.057
Occupation	.159	.042	.211	----	----	----	----
Population of Cities	.063	.032	.000	.014	.028	.029	.051
Urban-rural migration	.079	.015	.025	.051	.088	.067	.101
Movement out of Deep South	.038	.010	.000	.000	.027	.017	.000
Unemployment in states	.024	.009	.000	.013	.011	.038	.042
Supervisory responsibility	.065	.007	.005	----	----	----	----
Attitude toward hard work, need-achievement score	.030	.005	.000	.011	.023	.013	.016
Race	.025	.004	.000	.005	.000	.000	.005
Ability to communicate	.032	.004	.000	.000	.000	.000	.000
Geographic mobility	.007	.003	.000	.007	.023	.013	.008
Physical Condition	.016	.003	.000	.000	.022	.009	.000
Rank and Progress in School	.052	.001	.000	.000	.000	.000	.000
R^2		.359	.329	.280	.409	.372	.431

TABLE 4 --

HOURLY EARNINGS OF SPENDING UNIT HEADS -- EXOGENOUS FACTORS ONLY

Predictors	Gross Beta Coefficients[2]	AID Analysis- Reduction in TSS(I)/TSS(T)
Education	.133	.108
Age	.039	.039
Sex	.045	.046
Background	.069	.045
[illegible]	.020	.008
Race	.025	.000
Physical condition	.016	.000
Rank and progress in school	.052	.000
Religion	.040	.000
R^2	----	.246

CHART 4 --
AID ANALYSIS
OF HOURLY EARNINGS
OF SPENDING UNIT HEADS
(EXOGENOUS FACTORS ONLY)

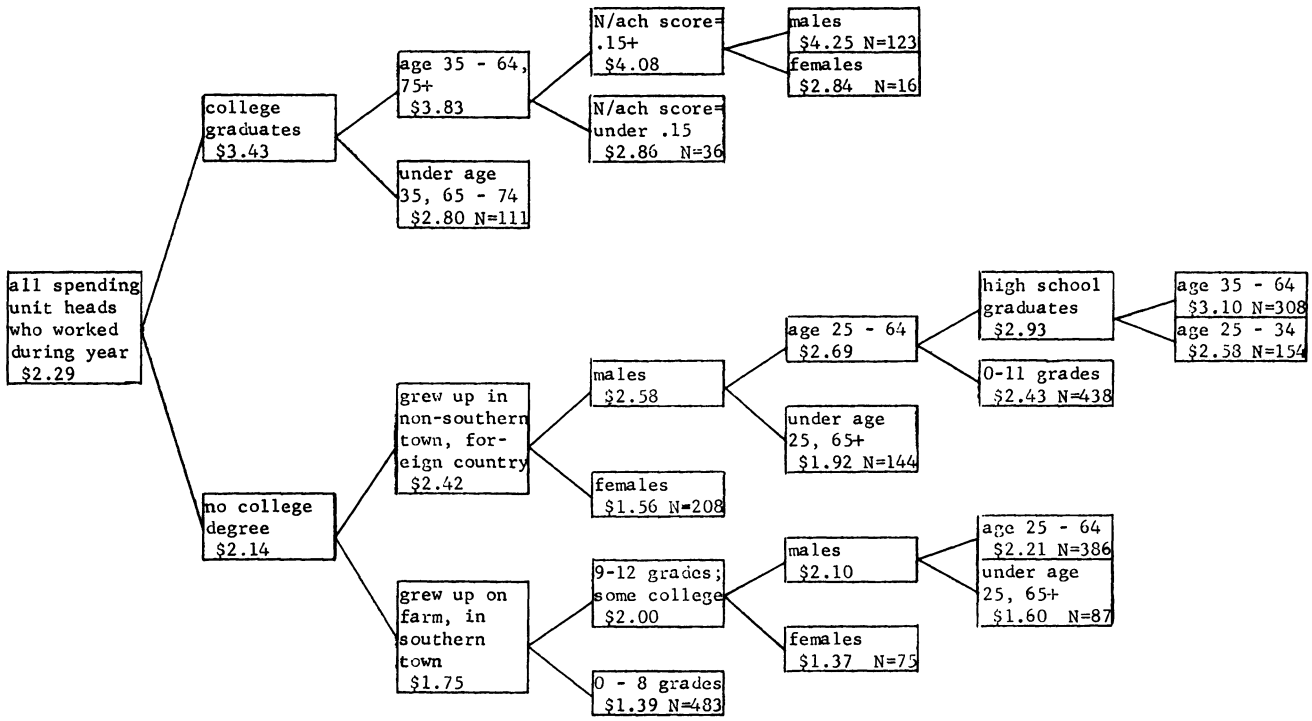


TABLE 5-- AID EDUCATION SPLITS ON HOURLY EARNINGS FOR SPENDING UNIT HEADS WHO WORKED IN 1959 (MEAN AMOUNTS FOR EACH SUBGROUP)

Education	All	Non-college graduates, Grew up outside Deep South in small town or city, male, aged 25-64	Non-college graduates, Grew up in Deep South or on a farm
None	$1.04	$2.57	$.85
1 - 8 grades	1.72	2.32	1.42
9 - 11 grades	2.14	2.52	1.92
12 grades	2.35	2.82	2.06
12 grades plus non-academic	2.48	3.03	1.99
College, no degree	2.51	3.01	2.17
College, bachelor's degree	3.25		
College, advanced degree	3.98		
Number of cases	2569	900	1031
Mean for group	$2.29	$2.69	$1.75

Non-College graduates $\bar{y}$ = $2.14

0 - 11 grades $\bar{y}$=$2.43

12 grades or more y=$2.93

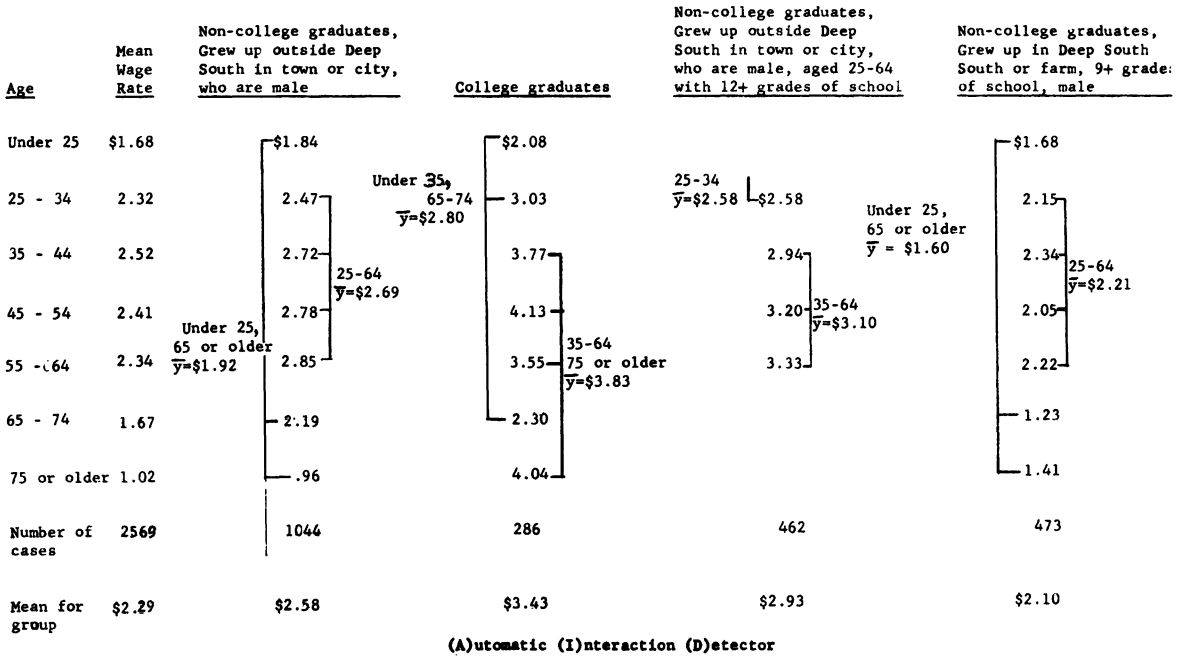
College graduates $\bar{y}$=$3.43

0 - 8 grades $\bar{y}$=$1.39

9 grades or more $\bar{y}$=$2.00

TABLE 6 --

AID AGE SPLITS ON HOURLY EARNINGS FOR SPENDING UNIT HEADS WHO WORKED IN 1959 (MEAN AMOUNTS FOR EACH SUBGROUP)



(A)utomatic (I)nteraction (D)etector

Model 1

Algorithm: Condensed Form.

PRELIMINARY READ IN. STEPS 0 AND 1.

0. Read in all parameters and all input observations, including all predictors and the dependent variable y.

1. To start, identify all input observations as belonging to group number one. Group number one is the current candidate group. Go to Step 5.

TEST FOR TERMINATION OF THE PROCEDURE. STEP 2.

2. Determine whether or not the current number of un-split groups is about to exceed the maximum permissible number; if so, go to Step 20, as the problem cannot proceed further.

DETERMINE WHICH GROUP SHOULD BE SELECTED FOR ATTEMPTED PARTITIONING. STEPS 3-5.

3. Considering all groups constructed so far, find one of them such that

   a. the total sum of squares (TSS_i) of that group is greater than or equal to 2 percent of the total sum of squares for the input observations (TSS_t);

   b. the group has not already been split up into two other groups;

   c. there has been no previous failure to split up the group;

   * d. the total sum of squares of that group is not smaller than the sum of squares for any other group that meets the above three criteria.

4. If there is no such group, go to Step 21; the problem is complete.

5. The group selected is the current candidate group, which will be the subject of an attempted split. Identify it with its group number (i) and, by option, print out N_i $\sum y_i$ $\sum y_i \bar{Y}_i$ and TSS_i. These statistics are always printed out if the group number one is the current candidate group.

PARTITION SCAN OVER ALL PREDICTORS. STEPS 6-17

6. Set j = 1 and go to Step 8.

7. Increment j by 1. If j is larger than the number of predictors being used in the analysis, the partition scan is complete; go to Step 18.

8. Compute N_{ijc}, Σy_{ijc}, Σy_{ijc}, $\bar{Y}_{ijc}$ for each class c of predictor j over group i.

9. Determine whether or not there exist two or more classes c, such that $N_{ijc} \neq 0$. If not, predictor j is a constant over group i; print an appropriate comment and go to step 7.

10. Sort the statistics produced in Step 8, together with the class identifiers for predictor j, into descending sequence using $\bar{Y}_{ijc}$ as a key.

PARTITION SCAN OVER THE c CLASSES OF PREDICTOR j.
STEPS 11 - 15

11. Set p = 1 and go to Step 13.

12. Increase p by 1. If p is larger than $(c_j - 1)$, where c_j is the number of classes in the jth predictor, then go to Step 16 as all possible feasible splits have been examined.

13. If $\Sigma N_k = N_1 = 0$ for k = 1...p, or if $(N_i - N_1) = N_2 = 0$, go to Step 12 as this split cannot be made because of empty classes in this group for predictor j. Otherwise, compute BSS_p, the between-groups sum of squares for the attempted binary split of group i on predictor j between the sorted classes (1,..,p) and the adjacent sorted classes (p+1,..,c).

*14. If this BSS_p is not larger than any BSS_p previosuly computed for this predictor over this group, go to Step 12.

15. This is the largest BSS_p encountered so far for this predictor. Remember BSS_p and the partition number p; then go to Step 12.

DETERMINATION OF BEST PREDICTOR. STEPS 16-17.

*16. Was the maximum BSS_p for predictor j larger than the largest BSS_p obtained from any of the other predictors previously tested over group i? If not, go to step 7.

17. This is the best BSS_p produced by any of the predictors tested so far over group i. Remember this partition and this predictor and then go to Step 7.

IS THE BEST PREDICTOR WORTH USING? STEPS 18-19

*18. Was the maximum BSS retained after the scan of all predictors over group i equal to at least 1/2 percent of the total sum of squares? If not, mark group i as having failed in a split attempt and then go to step 3.

19. Group i is to be split into two new groups and destroyed. Using the class identifiers and the partition rule remembered from Step 17, split the observations in group i into two parts. Identify the two new groups as having been created. Identify group i as having been split. Print the statistics from the successful partition attempt. Increase the total number of groups created so far by the quantity 2. Increase the current number of un-split groups by one. Then go to Step number 2.

TERMINATION OF THE ALGORITHM. STEPS 20-21

20. The maximum number of permissible un-split groups has been reached. Print an appropriate comment and go to Step 22.

21. There are no more groups eligible for further splitting. Print an appropriate comment and go to Step 22.

22. Print out a summary record of all groups created in the process of splitting, including the group number, its parent group, the values of the predictor class identifiers that were used in the partition which constructed the group, the predictor number used in this partition, an indication of whether or not this present group was ever split, and N_i, Σy_i Σy_i^2, and Tss_i. Stop.

FORMULAS

$$\bar{Y} = \Sigma y/N$$

$$TSS = \Sigma y^2 - \frac{(\Sigma y)^2}{N}$$

$$BSS = \frac{(\Sigma y_1)^2}{N_1} + \frac{(\Sigma Y2)^2}{N_2} - \frac{(\Sigma y)^2}{N}$$

$$WSS = TSS - BSS$$

* These decision rules constitute the crucial steps in the process which may be described in more global terms as follows.

1. Select that sample subgroup which has the largest total sum of squares, $TSS_i \geq .02\ (TSS_T)$

$$TSS_i = \Sigma X_i^2 - \frac{(\Sigma X_j)^2}{N_i}$$

The total sample is considered the first (and indeed, only) such group at the start.

2. Find the division of the classes of any single characteristic such that the partition p of this group into two subgroups on this basis provides the largest reduction in the unexplained sum of squares. Choose a division so as to maximize $(N_1\bar{X}_1^2 + N_2\bar{X}_2^2)$ with the restrictions that (1) the classes are ordered in descending sequence using their means as a key and (2) observations belonging to classes which are not contiguous after sorting are not placed together in one of the new groups to be formed.

3. For a partition p on variable k over group i to take place after the completion of (2), it is required that:

$$(N_1\bar{X}_1^2 + N_2\bar{X}_2^2) - N_i\bar{X}_i^2 \geq .005\ (\Sigma X_T^2 - N\bar{X}_T^2)$$

Otherwise group i is not capable of being split. No variable is "useful" over this group. The next most promising group (TSS_i = max) is selected.

4. If there are no more groups such that $TSS_i \geq .02\ (TSS_T)$ of if for the groups that meet this criterion there is no "useful" variable, or if the number of un-split groups exceeds a specified number, the process terminates.

Note 1: Eligibility criteria (2% for trying, ½% for an acceptable split) can be changed and should be for varying sample sizes and numbers of predictors. The program handles weighted data; the formulas being easily derivable.

Note 2: For an extended explanation of this algorithm see J. N. Morgan and J. A. Sonquist, Problems in the Analysis of Survey Data -- and a Proposal, JASA, Vol. 58 No. 302, June 1963, pp 415 - 434.