2192

THE DETECTION OF INTERACTION EFFECTS

A REPORT ON A COMPUTER PROGRAM FOR THE SELECTION OF OPTIMAL COMBINATIONS OF EXPLANATORY VARIABLES

35

BY

JOHN A. SONQUIST
JAMES N. MORGAN

ISR

SURVEY RESEARCH CENTER

Institute for Social Research, The University of Michigan, Ann Arbor, Michigan

THE DETECTION OF INTERACTION EFFECTS

A report on a computer program for the selection of optimal combinations of explanatory variables

by

John A. Sonquist
James N. Morgan

Monograph No. 35
Survey Research Center
Institute for Social Research
The University of Michigan

Library of Congress Catalog Card No. 64-63935

Printed by Cushing-Malloy Inc., Ann Arbor, Michigan
Manufactured in the United States of America

PREFACE

The motivation for the development of the computer program described in this report comes from two sources. First, is the belief that the multivariate statistical techniques in common usage are often inadequate for the analysis of the rich body of data from a cross section sample survey, and second is the conviction that a large-scale digital computer can be used for more than just a high-speed adding machine.

Modern data-collection techniques produce a wide variety of data. These range from classifications through rankings to continuous variables which sometimes approach near-normality in their distributions. Generally, they contain a variable amount of error, with little evidence as to its size or extent of randomness. When data come from a complex probability sample, serious questions arise as to the proper application of statistical tests of significance which usually assume simple random sampling models. Intercorrelations between explanatory variables make their effects difficult to assess and, when complex interaction effects and departures from linearity are present, the analyst has a difficult task indeed. Finally, some explanatory variables are logically prior to others, in that they can affect them, but cannot, in turn be affected.

Given the large amount of data, the essence of research strategy is to put some restrictions on the process in order to make it manageable. The more theoretical or statistical assumptions one is willing to impose on the data, the more the complexity of the analysis can be reduced. But the restrictions imposed in advance through the use of most conventional multivariate techniques cannot be tested. It appears to us to be desirable not to impose advance assumptions of linearity, absence of interaction and normality, yet to be able to consider the simultaneous effects of thirty or forty variables.

We have tried to break away from the habit of asking the question, "What is the effect of x on y when everything else is held constant?" This has been replaced with, "What do I need to know most in order to reduce predictive error a maximum amount?"

This is the type of question that might be asked by a research scientist working in a substantive area in which theory is not yet very precise. Once he receives an answer, he may well ask, "Now that I know this, what additional information would help to reduce predictive error still further?" and so on. He would certainly ask other questions as his results came back, but he would be unable to explore very many variables in this fashion without the aid of powerful machine techniques.

We have felt that one approach to the development of more satisfactory multivariate analysis techniques might be to start with the analysis strategy a scientist might use in exploring the system of relationships among a few variables, formalize it, and extend it to more variables by simulating the formal model on the computer.

The strategy implemented in what follows is admittedly very limited, and deliberately so, but it seems to work. What is clear is that sequential data-analysis strategies far more sophisticated than the present one can be programmed, and that the modern computer can provide an extension of the analytic capabilities of the research scientist in addition to being an extension of his pencil.

We would like to express our appreciation to the various people and organizations who have made important contributions to this work:

i. to Kathleen Goode, Keith Mather and David Schupp of the Institute for Social Research Data-Processing staff, and especially to Wen Chao Hsieh who did the programming.

ii. to Professors L. J. Savage and William Ericson for their advice and help. Professor Ericson's Note on Partitioning for Maximum Between Sum of Squares, a proof of the sufficiency of the partitioning algorithm, is incorporated herein.

iii. to the staff of the University of Michigan Computing Center, under Dr. R. C. F. Bartels for a powerful programming language and a readily accessible, though busy, computer.

iv. to Nancy Baerwaldt, Susan Sprunk, Elizabeth Goodrich and Alice Sano for their valuable assistance in the preparation of input data and in the preparation and editing of this manuscript.

v. to Frank M. Andrews, David Goldberg, and John Lansing for their helpful comments and criticisms.

vi. to the National Science Foundation for its generous financial support.

John A. Sonquist

James N. Morgan

Ann Arbor, Michigan

TABLE OF CONTENTS

LIST OF TABLES

LIST OF CHARTS

LIST OF APPENDICES

CHAPTER I

THE PROBLEM AND THE PROGRAM

Section 1.1 Abstract and Indexing Description

This report describes a computer program written in MAD and UMAP, for the IBM 7090, operating under the University of Michigan Executive System. The program is useful in studying the interrelationships among a set of up to 37 variables. Regarding one of the variables as a dependent variable, the analysis employs a nonsymmetrical branching process, based on variance analysis techniques, to subdivide the sample into a series of subgroups which maximize one's ability to predict values of the dependent variable. Linearity and additivity assumptions inherent in conventional multiple regression techniques are not required. Some examples of its use are presented, as are formulas, accompanying research strategy and some unsolved problems. Indexing Descriptors: Computer Program, IBM 7090, Multivariate statistical analysis, Statistical interaction, analysis of survey data, prediction, analysis of variance, data analysis strategy, sequential decision procedures, simulation.

Section 1.2 Introduction

This computer program (Identified as the (A)utomatic (I)nteraction (D)etector, Version 2) operates under the University of Michigan Executive System (1). It is focused on a particular kind of data-analysis problem, characteristic of many social science research situations, in which the purpose of the analysis involves more than the reporting of descriptive statistics, but may not necessarily involve the exact testing of specific hypotheses. In this type of situation the problem is often one of determining which of the variables, for which data have been collected, are related to the phenomenon in question, under what conditions, and through what intervening processes, with appropriate controls for spuriousness.

The data-model to which the present procedure is applicable may be termed a "sample survey model," in which values of a set of predictors X_1, X_2, ... X_p, and a dependent variable Y, have been obtained over a set of observations, or units of analysis, U_1, U_2, ... U_α ... U_n. A weight, W_α, may also be established for U_α if sampling models are not representative and self-weighting are used, or if one observation is considered to be more reliable than another. Data may be considered "missing" or undefined on any of the X_i or on Y. In particular, this analysis situation is defined to be one in which the X_i are a mixture of nominal and/or ordinal scales (or coded intervals of an equal-interval scale) and Y is a continuous, or equal-interval scale. The X_i variables may consist of a mixture of "independent variables" and also "specifiers" (conditions) and "elaborators" (intervening variables). Thus, the problem is similar to the accounting or explanatory analysis described by Hyman (2).

The objective is to explain the variance of the dependent variable Y. Where the number of predictors is small, the problems of isolating the relationships between the X_i and Y are manageable, but when the number of predictors is large, which is typical of many survey data analysis problems, then an analysis of the joint effects of the X_i on Y presents serious problems. Many of these have been extensively discussed on the methodological literature. One summary is presented in

Morgan and Sonquist (3). Tukey (4) presents a searching critique of present data analysis techniques.

Data-analysis problems are translated into a variety of statistical questions. For instance, multiple regression techniques and other statistical procedures based on them attempt to answer the questions, "What is the effect of predictor variable X_i on the dependent variable, holding 'constant' or removing the linear effects of the other predictors?" and "Are these effects 'significant' after taking into account the intercorrelations of the predictors?" The objective in an explanatory analysis is to ascribe the correct amount of the explained variation in Y to each predictor, within the limitations of the linear and additive assumptions of the model, using least squares criteria. Thus, one way of handling the problem of determining the joint effects of a large number of predictors is to introduce linearity and absence-of-interaction assumptions and then ask the above questions. The problem is that in view of the present state of much theory, one typically doesn't know in advance which transformations (e.g., X_i^2) or interaction terms (e.g., X_iX_k) to introduce into the regression model, in order to produce a multi-dimensional surface over which the residuals are not only normally distributed, but in which extreme values of the residuals are scattered randomly over the surface [Ezekiel and Fox (5)]

A great deal of work has been done in several fields which are related to the problem focussed upon here. Belson (6) has suggested a sequential, nonsymmetrical division of the sample for the purpose of matching two groups on various characteristics used as controls in order to compare them. Tanimoto and Loomis (7) have developed a computer program which forms clusters of observations which are similar along a number of dimensions. Reiter (8) presents a stochastic algorithm for optimizing payoff functions. Alexander and Manheim (9) have developed a computer program for the analysis of correlational data. The intercorrelations between variables are represented as lines on a linear graph, which is broken into components using a "hill-climbing" algorithm based on the information-transfer between variables.

There are also studies going on in the selection of test items to get the best prediction with a limited set of predictors (10), usually using multiple regression. Westervelt (11) has developed an interesting approach to the problem of maximizing predictability with a minimum number of terms by using a step-regression model combined with artificial intelligence.

Group-screening methods have been suggested by Watson (12) and by Box (13) in which a set of factors is lumped and tested and the individual components checked only if the group seems to have an effect. These procedures have some similarity to the sequential process suggested here.

Our approach bears some resemblance to a formal decision procedure proposed by Duncan, Ohlin, Reiss and Stanton (14), using cost-utility curves and also to a sequential procedure suggested and tried by Danière and Gilboy (15). Earlier related work has been done by Wright (16) and by Kitagawa (17). Kretschmer and Vinton (18) have programmed an "Information-Theoretic Seive" procedure which partitions a sample universe into two or more segments which are mutually exclusive and which minimize conditional uncertainty.

Each of these analysis schemes represents a specific statistical question. One such question is, "Given the units of analysis under consideration, what single predictor variable will give us a maximum improvement in our ability to predict values of the dependent variable?" This question, embedded in an iterative scheme is the basis for the algorithm used in this program. See (3, 19) for an extensive discussion of the rationale behind its development and implementation. The program divides the sample, through a series of binary splits, into a mutually exclusive series of subgroups. Every observation is a member of exactly one of these subgroups. They are chosen so that at each step in the procedure, their means account for more of the total sum of squares (reduce the predictive error) than the means of any other equal member of subgroups. The procedure may be described as follows.

Section 1.3 Description of the Algorithm

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

2. Select that unsplit sample group, group i, which has the largest total sum of squares

$$TSS_i = \sum_{\alpha=1}^{N_i} Y^2 - \frac{\left(\sum_{\alpha=1}^{N_i} Y_\alpha\right)^2}{N_i} \qquad (1.3.1)$$

such that for the i'th group

$$TSS_i \geq R\ (TSS_T) \quad \text{and} \quad N_i \geq M \qquad (1.3.2)$$

where R is an arbitrary parameter (normally $.01 \leq R \leq .10$) and M is an arbitrary integer (normally $20 \leq S \leq 40$).

The requirement (1.3.2) is made to prevent groups with little variation in them, or small numbers of observations, or both, from being split. That group with the largest total sum of squares (around its own mean) is selected, provided that this quantity is larger than a specified fraction of the original total sum of squares (around the grand mean), and that this group contains more than some minimum number of cases (so that any further splits will be credible and have some sampling stability as well as reducing the error variance in the sample).

3. Find the division of the C_k classes of any single predictor X_k such that combining classes to form the partition p of this group i into two nonoverlapping subgroups on this basis provides the largest reduction in the unexplained sum of squares. Thus, choose a partition so as to maximize the expression

$$(n_1\bar{y}_1^2 + n_2\bar{y}_2^2) - N_i\bar{Y}_i^2 = BSS_{ikp} \qquad (1.2.3)$$

where $N_i = n_1 + n_2$

and $\bar{Y}_i = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{N_i}$

for group i over all possible binary splits on all predictors, with restrictions that (a) the classes of each predictor are ordered into descending sequence, using their means as a key and (b) observations belonging to classes which are not contiguous (after sorting) are not placed together in one of the new groups to be formed. Restriction (a) may be removed, by option, for any predictor X_k.

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

$$BSS_{ikp} \geq Q(TSS_T) \qquad (1.3.4)$$

where Q is an arbitrary parameter in the range $.001 \leq Q < R$, and TSS_T is the total sum of squares for the input sample. Otherwise group i is not capable of being split; that is, no variable is "useful" in reducing the predictive error in this group. The next most promising group (TSS_j = maximum) is selected via step 2 and step 3 is then applied to it, etc.

5. If there are no more unsplit groups such that requirement (1.3.2) is met, or if, for those groups meeting it, requirement (1.3.4) is not met (i.e., there is no "useful" predictor), or if the number of currently unsplit groups exceeds a specified input parameter, the process terminates.

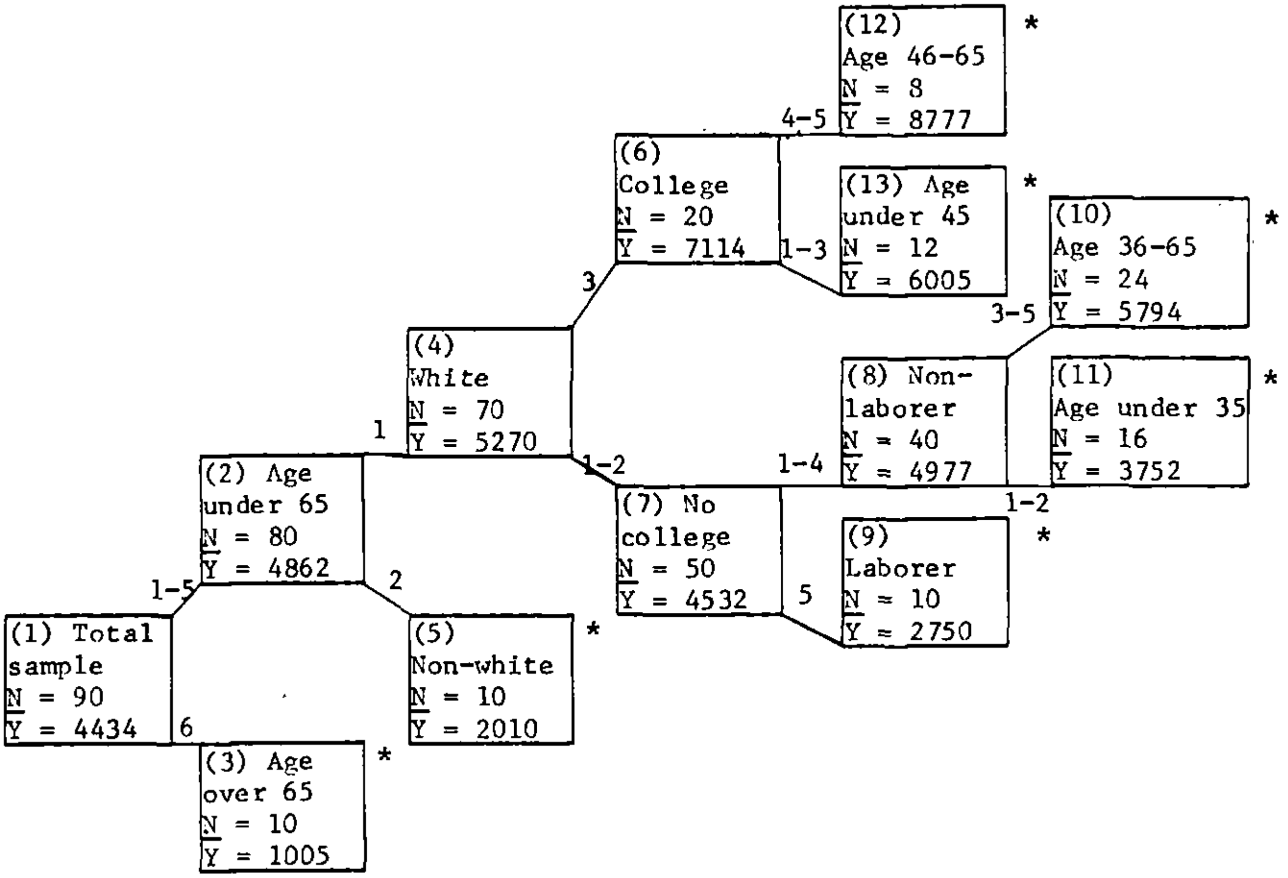
Section 1.4 Output Illustration

The following results, contrived, but realistic, will illustrate the basic output of the procedure. Suppose that Age, Race, Education, Occupation, and Length of Time in Present Job, are used in an analysis to predict Income. Age is an ordered series of categories represented by the numbers [1,2, ..., 6]. Race is coded [1 or 2], Occupation is coded [1,2, ..., 5], Education is coded [1,2,3], and Time on Job is coded [1,2, ..., 5]. We find the following mutually exclusive groups whose means may be used to predict the income of observations falling into that group:

Group	Type	N	Mean Income	σ
12	Age 46-65, white, college	8	$8777	$773
13	Age under 45, white, college	12	6005	812
10	Age 36-65, white, no college, nonlaborer	24	5794	487
11	Age under 35, white, no college, nonlaborer	16	3752	559
9	Age under 65, white, no college, laborer	10	2750	250
5	Age under 65, nonwhite	10	2010	10
3	Age over 65	10	1005	5
Total		90	4434	2263

A one-way analysis of variance over these seven groups would account for 95 per cent of the variation in income.

These results are arrived at by the following procedure, as represented by the tree of binary splits:



When the total sample (group 1) is examined, the maximum reduction in the unexplained sum of squares is obtained by splitting the sample into two new groups, "age under 65" (classes 1-5 on age) and "age 65 and over" (those coded 6 on age). Note that each group may contain some nonwhites and varying education and occupation groups. Group 2, the "under-65" people are then split into "white" and "nonwhite." Note that group 5, the "nonwhites" are all under age 65. Similarly the "white, under age 65" group is further divided, into college and non-college individuals, etc. A group which can no longer be split is marked with an asterisk and constitutes one of the above final groups. The variable "Length of Time in Present Job" has not been used. At each step there existed another variable which proved more useful in explaining the variance remaining in that particular group.

The predicted value Y_α for any individual for any individual α is the mean, $\overline{Y}_i$, of his final group. Thus $Y = \overline{Y}_i + \varepsilon$, where ε is an error term. Prediction of income on the basis of age, education, occupation and race would provide a considerable reduction in error. Variables which "work" are, of course, the most logical candidates for inclusion in a theoretical framework.

We now turn to a description of the computer program, its organization, and use.

CHAPTER II

USING THE PROGRAM

Section 2.1 Program Organization

The program is written in MAD (Michigan Algorithm Decoder), a compiler language developed by Galler, Arden and Graham (20) for the IBM 704, 709 and 7090 systems. It uses several subroutines written in UMAP (University of Michigan Assembly Program), which is a modification of the standard assembly programs available through the IBM user's organization SHARE. MAD and UMAP are contained in the University of Michigan Executive System (1). Loading the program, program segmentation, input and output, and the need for numerous subroutines contained in the System require that AID (2) be operated in the context of the U. of M. System. The System, MAD, and UMAP are available through the IBM user's organization, SHARE. The program requires a 32k system with 8 tape units.

AID (2) is organized into three program segments, the Editor or control segment, the Iterator or processing segment, and the Final Output Segment. Control originates in the Editor, is passed to the Iterator, then to the Output Segment and is then returned to the Editor, or to any program segment which may precede it on the program segment tape.

The functions of the Editor are to:

1) Read in control cards which describe
    a) the location of the input data (tape or cards) and where it is to be stored.
    b) which variables are to be used in the analysis and what they are to be used for.
    c) what subset of the input data is to be used in the analysis.
    d) other aspects of the current problem.

2) Read in the data and store it on tape if necessary.

3) Store the data to be used in the analysis into the appropriate positions of core storage.

4) Compute various statistics needed by the Iterator.

If errors occur, such as control cards out of sequence, problem too big for the program, illegal data, etc., the Editor provides appropriate diagnostic comments and then exits to the U. M. Executive System monitor.

The Iterator performs the analysis indicated by the parameters on the data provided for it by the Editor and provides intermediate output as requested. Threaded lists (21) are employed in the algorithm implementing the partitioning process.

The final output segment then calculates various statistics and prints out a summary of the results. It also calculates predicted values of the dependent variable and residuals for each unit of analysis. It then returns control of the computer to the Editor (or to any other program segment which the user desires to place in front of the Editor).

The tapes used are listed below.

Tape Number	Function
1	U. M. Executive System Tape
2	AID Program Segment Tape
3	Scratch Tape--used by AID
4	Scratch Tape--used by AID
5	Not used
6	Output Tape
7	Input Tape
8	U. M. Executive System Tape

Since a large number of variables may be read in and stored on a scratch tape, several analyses may be performed in succession. An attempt has been made to provide considerable flexibility with respect to data formats, multi-stage analyses using residuals as the dependent variable, and selection of subsets of the input data for analysis.

Section 2.2 Data Input Requirements

It is assumed that the data have been punched on IBM cards; one or more cards per observation. Input data may be punched anywhere on the card except in column 1. Column 1 of the data cards may contain any legal character except an alphanumeric E. A legal character is defined as any punching pattern obtainable from a single depression of a key on a keypunch.

Since several analyses may be performed during one machine run, it is desirable to list the types of variables that may be entered into the computer. Each analysis may use its own subset of the variables. Variables entered into the computer are of five types:

1) Identifiers
2) Sample subset selectors (filters)
3) Predictors
4) Dependent variables
5) Weights

With the exception of identifiers, any variable may be used for purposes two through five above, provided it meets the restrictions made by the program on the values that variable may legally assume.

There are no restrictions on where any of the variables may be placed on the data cards, except that no variable to be used in an analysis may be punched in column 1.

Since the card reading equipment associated with the IBM 7090 operates in BCD mode, no data cards may be used by the program which have punching patterns anywhere on the card that do not constitute legal IBM characters.

The input data may be any file of (match-merged) data cards conforming to the above rules which can be described by nine cards of MAD format information. The MAD format information describes one unit U_α of data.

All input variables except those which are to be used as identifiers are supplied to the program in Integer mode. Variables which are to be used as identifiers must be supplied to the program in Character (BCD) mode. Hence they may be used only for that purpose.

Thus, for any purpose except that of observation (unit) identifier, variables must be punched on the data cards in such a way as to permit their representation inside the computer as integers. Consequently, classes of predictors may not be represented as alphanumeric characters. However, there are certain special cases in which the characters + and - may be represented in the computer as integers. These are described in Appendix G. In general, the user is advised to represent his data on the IBM cards using only the characters 0 through 9, with the exception of variables to be used as dependent variables which may be signed numbers.

When several analyses are to be performed on the same set of data, machine costs will be somewhat reduced if all the variables to be used in all of the analyses are read in at the time of the first analysis, saved on tape, and subsequent analyses performed using the data from tape.

If card output of residuals is not desired and if, in addition, the analyses are to be performed on a subset of the sample, it will, in general, be cheaper to sort out the unwanted observations before setting up the run. If, however, punched residuals are desired, it is recommended that the entire sample be entered into the computer and the unwanted observations screened out using the sample subset selector (which will be described below). When residuals are requested, it is generally advisable to punch them, even if subsequent analyses are made of them from tape, since it is not always possible to anticipate which additional variables should be used in subsequent analyses of the residuals.

Section 2.3 Program Capacity

Though data may be stored on tape, in the interest of computing efficiency, all of the information for any particular analysis, including predictors, dependent variable, weights, etc., are kept in core storage. Thus, the following limits apply:

Maximum number of input variables = 100

Maximum number of dependent variables for any one analysis = 1

Maximum number of predictor variables for any one analysis = 36

Maximum allowable number of groups into which the input observations may be split = 63

Range of any predictor $0 \leq V_p \leq 63$

Range that may be legally taken on by the dependent variable before scaling $-99999 \leq V_y \leq 999999$

Range that may be legally taken on by the weight associated with any given observation $0 < V_w \leq 9999$

Range that may be legally taken on by any variable used as a sample subset selector $-99999 \leq V_f \leq 999999$

Maximum number of merged input data decks = no limit, except that they must be able to be described by the MAD format statement

Maximum number of cards in the MAD format statement = 9

Minimum number of observations that must be contained in the i'th group if that group is to become a candidate for splitting $2 \leq N_i \leq 999$

Maximum number of input observations: limits are determined by single-precision accuracy of 7090 floating point computations. A six-digit dependent variable, weighted by a three-digit weight is probably not subject to serious rounding error in calculating the total sum of squares until the sample size exceeds 5000. No exact rounding and truncation error analysis has been performed.

Storage requirements are such that for any problem to be entered into AID, the amount of storage per observation (3-8 words) times the number of observations must be less than 20000. Maximum sample sizes for all possible numbers of predictors are listed below. Determine which category the problem falls into, based on the number of predictors. The second line gives the maximum sample size.

	Category					
	(a)	(b)	(c)	(d)	(e)	(f)
Number of Predictors (NP) in any one analysis	1 - 6	7 - 12	13 - 18	19 - 24	25 - 30	30 - 36
Maximum permissible number of observations in analysis	6666	5000	4000	3333	2857	2500

If a problem is too large either the number of input observations must be cut down or the number of predictors must be reduced enough to put it into a different category. For instance, a problem with twelve predictors and a given sample size takes up as much space as one with seven predictors and the same sample size since both fall into category (b).

Section 2.4 Recommended Steps in Setting up Runs for AID (Model 2)

1. Complete page one of the Run Specification form (see Section 2.7). Make sure the data cards to be used as input are free of illegal punching patterns and that match-merging has been properly accomplished. Normally, the input data deck sequence is match-merged on interview or data-unit number. Cards may be sorted into subgroups for entry into the computer, but in this case, a complete AID run must be made on each group separately. An accurate count of the number of input observations (N) must be made, as AID will throw the job off the computer if the parameter N does not agree with the actual number of observations read in.

2. Complete page two of the Run Specification form, the data-description. List all variables to be used in all of the AID analyses, including predictors, dependent variables, identifiers, filters, and weights, starting from the left side of the input data cards and working toward the right. Then number the variables, sequentially, starting with the integer 1. There are no restrictions in AID as to where the predictors or dependent variable or weight, etc., must be located, except for the fact that no variable may be in column 1. Column 1 on the first card of each data set may not contain the character E. There are no other restrictions on data except the usual ones. That is, no multiple punches may occur in the columns which are to be used as predictors, filters, weights or dependent variables; and no illegal punching patterns may appear anywhere on the card. If + or - punches occur in the predictors or the dependent variable as other than a sign for the dependent variable, confer with an experienced programmer before proceeding further (see Appendix G).

   List for each variable (a) a name (up to 12-characters), e.g., AGE, INSURANCE, etc., and (b) the column numbers in which the variable is located. For all variables which are to be used as predictors, filters, or dependent variables, list all of the possible values that variable can legally have. NOTE: NO VARIABLE

TO BE USED AS A PREDICTOR MAY LEGALLY HAVE A VALUE LARGER THAN 63. Nor may any predictor have a negative value. Note values of the dependent variable that should be omitted as missing data. Use the intended usage column to indicate the function (identifier, predictor, filter, etc.) that variable is to perform. Illegal values of input data will result in an automatic exit from AID and a memory dump.

AID has the capacity to omit observations that have certain specified values of the dependent variable. All observations having the dependent variable $V_y = -0$ are automatically omitted. All values larger than a certain specified value may be omitted. In addition, all observations equal to either of two other specified values may be omitted. These values should be indicated on the data-description form. Unless there is a great deal of missing data, it is desirable to leave such observations in the deck and have the computer throw them out of the analysis, rather than sorting them out beforehand.

The purpose of the analyst's recording this information on these forms is to inventory all necessary information about the run in one place to prevent the inadvertent forgetting of a necessary piece of it. If any of the variables have the characters + or - used for anything other than a sign for a dependent variable, one of the special input formats provided by subroutine IRFORM must be used. Confer with an experienced programmer.

3. Use the collected information on the Data Description forms to fill out the remainder of the Run Specification forms which establish control card punching and the input file sequence.

Section 2.5 Control Card Punching

Label Card--Type 1 card

Column one of this label card must be punched with a one (1). Punch 78 characters of Alphanumeric run identification (anything you can punch with one depression of a key on a keypunch) in columns 2-79.

If card residuals are requested, place the research project identifier starting in column 2 of this card, followed by a deck identification number. The contents of columns 2-13 of this card will be punched into the card residual output from this analysis. All BCD characters are legal project and deck identifiers. The entire contents of cols. 2-79 will be printed on the output statistics. If this run is a subset of the sample, i.e., a partial data deck has been entered into the computer, this information should be punched on the label card somewhere after col. 13.

Main Parameter Card--Type 2 card

This card contains a series of parameters identifying the location of the input data (cards or tape) that is to be used on the analysis, the number of observations to be expected from this source and the number of input variables. The remainder of the card is a series of parameters that form a sentence. This sentence defines what subset of the input observations are to be used in the analysis. It will be referred to as an input subset selector or "filter." The subset selector has no effect on what is transmitted to tape, but only defines what observations are to be packed into core storage for the present analysis.

A description of the parameters, their permissible values and their purposes follows:

Column	Name of Parameter		Remarks
1	Card type	=	Must be punched with a 2.
7	LOCDAT	=	C if input data is on cards and is not to be saved on tape.

Column	Name of Parameter	Remarks
		= W if input data is on cards and is to be written on binary tape and saved for subsequent analysis.
		= T if the input data is already on binary tape as the result of a previous run.
10-13	N	= The number of observations in the input file. It must agree exactly with the actual number read into the computer. This parameter may not necessarily be equal to the number of physical cards read in if merged data-decks are used.
17-19	NV	= The total number of all input variables including identifiers, predictors, sample subgroup selection variables, dependent variables and weight(s) AND when LOCDAT = T, this includes residuals left on tape from any previous analyses performed on this run. This number is first determined using the Computer Input Data-Description form. Take all the variables to be used in all analyses and assign integers to them, starting with the left-most variable on the first (merged) data deck as variable number 1. The last field number on the last merged deck is variable NV. The first residual placed on tape is numbered one larger than this, the second residual is numbered two larger than this. Subsequent values of NV for succeeding runs must be increased by the appropriate amount. Range ($2 \leq NV \leq 100$).

The following describes the input sample subgroup selector. Some of the terms are self-explanatory. A closed interval is defined conventionally as one in which the boundary values are considered to be part of the interval, not outside it. Thus, the interval (-556 to 1089) includes the integers (-556, -555, ... -1, +0, +1, 2, ... 1088, 1089). Minus zero (-0) is specifically defined to be in this interval. The numbers -557 and 1090 lie outside the interval. (-556 to 1089) are considered to be inside the interval. In this example, the lower bound is defined as -556 and the upper bound is 1089.

A subscript or index of an input variable is that integer assigned to it when input variables are numbered from left to right across the (merged) data decks as described above. An input variable which is stored in the computer in BCD mode (identifiers are normally in this mode) may not be used in the sample subgroup selector.

The words AND and OR appear in the selector sentence. The two terms correspond to common English usage. Specifically, OR is inclusive, rather than exclusive. For completeness, they are described as follows:

		B	
		True	False
A OR B → A	True	True	True
	False	True	False

		B	
		True	False
A AND B → A	True	True	False
	False	False	False

The sentence contains a command (INCLUD, EXCLUD); a first condition, called condition A; a connector (AND, OR); and a second condition, called condition B. For example:

1) "INCLUDe in this analysis all input observations which are
2) OUTside the closed interval which runs from 1 (lower bound) up to 4 (upper bound) on the variable whose input number is 5
3) OR which have values such that they are
4) INside the closed interval which runs from 2 (lower bound) up to 2 (upper bound) on the variable whose input number is 6."

(1) above is a command; (2) is condition "A" which is either true or false for any given input observation U_α; (3) is a connector; and (4) is condition "B" which is either true or false for the observation U_α.

The above example specifies that if condition A is true or if condition B is true, or if both of them are true, then the observation will be included in the analysis. If both are false for that observation, then it will not be used.

Another way of stating this is to say that the conjunction of conditions A and B for any observation U_α is either true or false. If it is true, then the action specified by (1) is taken. If it is false, then the action complementary to that specified in (1) is taken. The actions which may be specified are INCLUDe or EXCLUDe. They are complementary.

It may be desired to establish only one condition for entry of an observation into the analysis. In this case the connector is left blank and the program ignores the parameters referring to condition B. Then, if condition A is true for observation U_α, the action specified in the command is taken. If condition A is false, then the complement of the action specified in the command is taken.

It may be desired to use all of the input observations in the analysis. In this case, the command itself is left blank and all observations will be used, bypassing the subgroup selection process completely.

The exact description of the filter (input subset) parameters is given in a Section 2.7 entitled "AID (2) Run Specifications, Input File Assembly."

It should be noted that several other conditions will cause an observation to be excluded from an analysis. These conditions involve values of the dependent variable which are declared to be "missing data," and will be described later. These conditions operate independently of the sample subset selector.

Input variables which are to be used as identifiers for punched residual output may not be used as "filter" variables in the sample subgroup selector, since they are BCD in mode. However, if it is desired to perform an analysis on a subset of the input data definable in terms of the observation identifier, and if, in addition, match-merged data decks are used as input, one of the unit identifier fields may be read in as an integer and used for a filter variable provided it contains only the characters zero through nine in the field. Another identifier may then be stored in BCD (Character mode) and used for identifying the residuals.

Secondary Parameter Card--Type 3 card

This card contains the remaining parameters describing the analysis to be performed, with the exception of the list of variables that are to be used for predictors.

Column	Name of Parameter		Remarks
1	Card type	=	Must be punched 3
6-7	NP	=	Punch a count of the number of predictor variables that are to be used in this analysis, e.g., 07 if seven predictors are to be used. Restriction: $1 \leq NP \leq 36$
11-13	WT	=	This is an index number. If the field to be used as a weight is the ninth variable listed on your Computer Input Data Description form, punch 009. If 000 is punched here, the run will be unweighted. If the run is weighted, punch the index number of the input variable to be used as a weight. Restriction: $0 \leq WT \leq NV$.
14-19	P1	=	This should be set at .00001, essentially deactivating it. This will allow P2 to control the termination of the splitting process.
20-25	P2	=	The best split on the i'th candidate group must reduce the unexplained sum of squares by P2 proportion of the total sum of squares or that

Column	Name of Parameter		Remarks
			group will not be split, and it will not become a candidate group again even though it may meet the P1 requirement above. The range is: $P1 \leq P2 \leq$ .99999. The decimal point is punched in this field as above. The proportion .0060 has been found to work well with samples of 1500-3000. Other values may be punched at the user's option. Increase P2 to at least .01 with sample sizes of 200-300 or less.
26-28	MAXGP	=	The maximum allowable number of final groups into which the input data may be split, regardless of P1 or P2. Thus, the splitting process will always stop when the sample has been divided into MAXGP number of unsplit subgroups. MAXGP may not be larger than 63, i.e., 62 splits, 125 groups in all. Fifty has been found to be a satisfactory maximum number of splits. The range is: $1 \leq$ MAXGP ≤ 63.
29-31	MSIZE	=	This is the minimum number of observations that must be contained in a group if that group is to become a candidate for splitting. Its purpose is to prevent small groups with somewhat unstable means from being further split, since the splits are likely to be heavily influenced by sampling errors. Normally MSIZE should not be smaller than 25. Range: $001 \leq$ MSIZE ≤ 999.
35-37	Y	=	This is the index number of the variable to be used as the dependent variable. For example, if the dependent variable is the 14th variable on your Input Description form, then punch 014 here. If the dependent variable is number four, punch 004. NO VARIABLE TO BE USED AS A DEPENDENT VARIABLE MAY HAVE A VALUE LARGER THAN 999,999 or less than -99999 before scaling.

Column	Name of Parameter		Remarks
38-49	YNAME	=	Punch alphanumeric information here. The name of the dependent variable, e.g., INCOME, WIFE'S WAGE, etc.
50-55	YMAX	=	This is a "missing-data" code. For some observations, there may be no information on the dependent variable. Or there may be large values which are to be screened out. These may be left in the computer input file. YMAX is for preventing them from being used in the analysis. Thus, any observation whose dependent variable has a value algebraically *larger than* YMAX will be read, but not used by the computer in this analysis. YMAX is scaled by the input scale factor before being used. If you do not wish to use YMAX, leave it blank.
56-61	MD1	=	This is an additional method of throwing missing data out of the analysis. Any observation such that the dependent variable is *exactly equal* to MD1 will not be used in the analysis. MD1 is scaled by the input scale factor before being used. If you do not wish to use MD1, leave it blank.
62-67	MD2	=	The same as MD1. Do not use MD2 without using MD1 also. Leave it blank if you do not use it.

Note on missing data: regardless of what is punched in YMAX, MD1 and MD2, *AID will omit all observations such that the dependent variable has the value minus zero*. If all of your missing data are coded in this fashion, or if you have no missing data, then leave YMAX, MD1, and MD2 blank. All undefined residuals have the value -0.

Column	Name of Parameter		Remarks
68-70	CDRES	=	If it is desired to compute residuals for this analysis and punch them on cards, this parameter is punched CRD, otherwise it must be left blank. If residuals are to be punched on cards, columns 2-13 of the label card (type 1) must contain research project and deck number information. An identifier variable must be included as part of the set of input variables and must be made available to the program in BCD (character) mode. This variable must be indicated by a nonzero value for the parameter INTNO described below.
71-73	TPRES	=	If it is desired to compute residuals and write them on tape for a subsequent analysis, this parameter is punched TAP, otherwise it must be left blank. This option may be exercised regardless of whether the input data for this analysis is on cards or on tape. If it is exercised, then the residual is written on tape as variable NV + 1, where NV is defined as above. IF A SUBSEQUENT ANALYSIS IS TO BE PERFORMED ON TAPE, THE PARAMETER NV ON THE FOLLOWING ANALYSES MUST BE ADJUSTED ACCORDINGLY, as there is now one more input variable.
74-76	INTNO	=	This is the index, or subscript of the input variable (identifier) to be punched in the interview number field of the output cards containing residuals. If card residuals are being obtained from this analysis, this parameter must lie in the range $1 \leq \text{INTNO} \leq \text{NV}$. If card residuals are not being obtained from this analysis, then INTNO may be left blank or set to zero. It will not be interrogated.

Column	Name of Parameter	Remarks
77-78	SCFIN	= This is an input scale factor to be applied to Y, to YMAX and to MD1 and MD2. It is that power of ten by which Y is to be multiplied before being used in computation. Thus, the characters 12345 read as a five-column Integer (I) field on a data card, or from tape, will have the internal value of 12.345, if this parameter has the value -3. The purpose of this parameter is to determine where the decimal appears in the printed output. For analysis of residuals, where a previous SCFOUT has moved the decimal point to carry more significant digits, SCFIN is used to put the decimal point back in the right place for this analysis stage. In this case SCFIN equals the <u>previous</u> SCFOUT with opposite sign. Range: $-9 \leq$ SCFIN $\leq +9$.
79-80	SCFOUT	= This is an output scale factor which is applied to Y, the predicted value of Y and the output residual, after computation and before punching or the writing of the residual on tape takes place. It is that power of ten by which these terms are to be multiplied before being output as integers. It will generally be desirable to provide more significant digits in the residuals than there were in the <u>original</u> dependent variable. Therefore, SCFOUT is normally equal to [(-SCFIN) + 2], reducing the dependent variable to its original form and adding two more significant digits. Range: $-9 \leq$ SCFOUT $\leq + 9$. The purpose of SCFOUT is to move the decimal point in the (previously scaled) dependent variable into a place suitable for punching or writing on tape.

Predictor List Cards

The user must supply information to AID telling it which of the input variables are to be used as predictors. (The information on the main parameter cards has indicated which input variables are to be used as the dependent variable and the weight, if desired.) Each predictor list card contains information on up to four predictors. The last predictor card is the only card that may contain information on less than four predictors. Any input variable may be used as a predictor provided it is stored in the computer in integer mode, never exceeds the value 63 and is never negative in value.

The predictors may be listed in any order desired by the user, since the order listed is irrelevant for the program. Three types of information are punched for each predictor: its index, a type code and its name. The index is obtained from the Data Description sheet. It is the field number established by numbering the NV variables from left to right across the merged input decks. The name of the variable should be punched as up to 12 characters representing a suitable mnemonic reference to the substantive meaning of the variable, e.g., AGE, SEX, INCOME, REGION, RISK SCALE, etc. A blank is counted as a character.

The predictor type is punched as M (monotonic), or F (free). Predictors identified as type "M" will have the order of their coded values (0, 1, ..., k, ..., 62, 63) maintained during the partition scan. In this case the classes of the predictor will not be re-arranged by sorting them into descending sequence using the mean value of Y for each class as a key. In designating a predictor, say V_p a type M predictor, the user assumes that though the function $Y = \bar{Y}_{kp}$ may not be linear it is at least monotonic. The usual use for a type M restriction is to apply it to an ordinal scale, or to class-interval codes established for a continuous variable with an expected monotonic effect on the dependent variable.

Predictors identified as type "F" will have their classes re-arranged during the partition scan. They will be sorted into descending sequence using the mean value of Y for each class as a key.

The usual use for a type F predictor classification is for variables that are nominal scales, or for other cases in which it is suspected that the function $Y = \overline{Y}_{kp}$, where k is the predictor class code, is not monotonically increasing or decreasing. A useful strategy may be to classify all predictors as type F, determine whether partitions appear that look fortuitous, and then to restrict the offending predictor(s) in a subsequent analysis.

Punch only as many predictor cards as are needed. Up to 36 predictors may be used (nine predictor cards). Each card should be completely filled in, except the last one, which will have some blank spaces at the end if NP is not an exact multiple of four. The format of the predictor list cards is described hereafter in the AID (2) Run Specification, Input File Assembly.

MAD Format Statement

The MAD format statement is punched in columns 2-72 inclusive, on up to nine (9) cards. IT MUST BE COMPLETELY ENCLOSED IN PARENTHESES, as it is read in by subroutine IRFORM. There must be exactly NV field descriptions, in addition to the

(C1,

that starts the format statement. Included are all predictor(s), the dependent variable(s), identifier(s), filter(s) and the weight(s) for all analyses to be performed, together with the appropriate S (skip) and / (go to the next card) characters. All columns of the input data starting with column 1 of the first merged deck and continuing to the last (rightmost) variable of the last merged deck must be accounted for. The first column on the first merged deck is accounted for by the

(C1,

on the first card of the MAD format statement. The format statement ends with the characters

*)

All fields used for input must be specified in integer (I) mode, except for identifiers which are character (C) in mode. Insert only as many format cards as needed. See the MAD manual (20) for additional details.

See Appendix G of this write-up for a description of Subroutine IRFORM which reads in the MAD Format information, especially if the data cards to be used contain other than the characters 0-9 in the variables to be used as predictors or filter variables, or if the dependent variables contain punching patterns other than signed numbers or minus zeroes. An example follows for NV=7 and one input deck:

(C1, S3, 4I1, I4, I2, S60, C6*)

For two merged input decks and NV=7 one might write:

(C1, S3, 4I1 / S8, I4, I2, S60, C6*)

In the first example variables are located as follows:

Index No.	Cols.	Function
1	5	Predictor
2	6	Predictor
3	7	Predictor
4	8	Predictor
5	9-12	Dep. Var. Y
6	13-14	Weight
7	75-80	Identifier

In the second example variables are located as follows:

Index No.	Cols.
1	5
2	6
3	7
4	8
Next Card	
5	9-12
6	13-14
7	75-80

Section 2.6 Input File Assembly Sequence

An AID Run Requires the Following Input File

(1) Two computing center job cards (See U. of M. Executive System Write-Up [reference (1)] for a description)

(2) A systems card $EXECUTE, DUMP, I/O DUMP, BINARY

(3) The AID program decks, in binary form

(4) A systems card $DATA

(5) An AID label card (type 1 card)

(6) An AID main parameter card (type 2 card)

(7) An AID secondary parameter card (type 3 card)

(8) Up to nine (9) AID predictor list cards (type 4 cards). Insert as many as needed, no more

(9)* Up to nine (9) cards containing a MAD format statement enclosed in parentheses. Insert only as many cards as needed, no more.

(10)* A DATAFOLLOWS card

(11)* The match-merged data-decks

(12)* A Type E trailer packet

(13)* As many repetitions of (5) - (12) above as desired

*These cards are omitted if the data are already on tape from a previous analysis.

Section 2.7 AID (2) Run Specifications, Input File Assembly

These forms were developed as an aid to taking an inventory of all the information necessary to initiate a run on AID (2). Taken together, and properly completed, they provide the user with the source material necessary for keypunching his control cards and assembling his input file.

7090 COMPUTER

AID(2) RUN SPECIFICATIONS, INPUT FILE ASSEMBLY

WRITTEN BY: ______________________ PHONE __________

CHECKED BY: ______________________ PHONE __________

DATE: ______________________

CHARGE TO STUDY # __________

MTR # ________ FOR

STUDY # __________

THIS SECTION FOR DATA PROCESSING USE

DATE STARTED __________

DATE COMPLETED __________

OPER. INIT. __________

Data Processing JOB NUMBER

C. C. # ______

PLEASE INCLUDE COMPLETE IDENTIFICATION AND CARD COUNT OF ALL DECKS USED. INDICATE WHICH COLUMNS CONTAIN THE STUDY NUMBER, DECK NUMBER, AND INTERVIEW NUMBER.

COMPUTER PROGRAMS(s) TO BE USED: (A)utomatic (I)nteraction (D)etector (Model 2)

PREREQUISITES: ______________________

Purpose: (description of dependent variables, predictors, whether multi-stage run, etc.)

Number of file assembly packets (pages 3-8) included in this run = ☐

Input Data Decks:	Study Number:	Study Identif. in columns:	Deck Number:	Deck Identif. in columns:	Deck N

Sight check identification and verify all N's on sorter before proceeding further.

Special Instructions: (match-merging of decks, cards to be omitted from computer input file, request for checking of invalid punching, etc.)

Number of observations in computer input file = ☐ (Control card 2, col. 8-13)

NOTE: Prior to any 7090 run, all decks should be checked for blank columns and double punches. If this has not been done previously, request deck checks as a preliminary step in this request.

EDPM Form #18a
Rev. 3/64

Page One of ______ Pages

Computer Input Data Description
AID(2)

FIELD NO. (1)	VARIABLE NAME (2)	IN COLS. (3)	MAD FORMAT* (4)	POSSIBLE CODE VALUES (5)	INTENDED USAGE (6)

*Filled in by programmer (Insert extra pages as necessary)

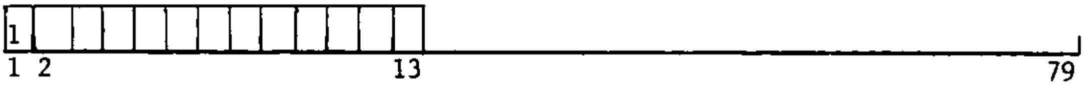
April 1, 1964

Page_____of_______

AID(2) Run Specifications, Input File Assembly

(A)utomatic (I)nteraction (D)etector
MODEL 2
File Assembly Packet
(1 APR 64)

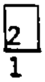
I. <u>Label Card: Type 1 input parameter card.</u> Column 1 <u>Must</u> be punched with a 1.

Col.


(78 Characters of Alphanumeric Run Identification. Eligible characters include 0-9, A-Z, $ + - / * = , . ( ) and blanks. If punched residuals are requested, then columns 2-13 of this card will be punched into cols. 1-12 of all residual card output for this run. Thus, alphanumeric study and deck information can be transferred to output.) Cols. 2-79 of the label card are <u>always</u> printed on the output.

PUNCH ALL PARAMETERS RIGHT-ADJUSTED IN THE FIELD ON ALL FOLLOWING PARAMETER CARDS

II. <u>Type 2 parameter card:</u>

Col.


= Card type = Must be punched with a 2

Col.


(Punch C, W, or T in col. 7)

= LOCDAT = C if input data is on <u>cards</u> and <u>IS NOT</u> to be saved on tape

= W if input data is on <u>cards</u> and <u>IS</u> to be written on binary tape and saved for another run.

= T if input data is on binary tape as a result of saving it from a previous run.

= N = The number of observations in the input file.

= NV = Total number of all input variables including the predictors, interview number, dependent variable, weight (if any), AND (when LOCDAT=T) INCLUDING RESIDUALS LEFT ON TAPE FROM ANY PREVIOUS RUNS.

INPUT SAMPLE SUB-GROUP SELECTOR (Filter)

These parameters define (if desired) a subset of the input observations which are either to be included or excluded from this run. Cross out <u>filled in options</u> which do <u>not apply</u>.

INCLUD
EXCLUD in this run all input observations which are
Col. 20 25

▨▨▨OUT - side
▨▨▨▨IN - side the closed interval which runs from
26 31

☐☐☐☐☐☐ (lower bound), up to
32 37

☐☐☐☐☐☐ (upper bound), on the variable whose input number is
38 43

▨▨▨☐☐☐ (subscript of filter variable),
44 49

▨▨▨▨OR
▨▨▨AND which have values such that they are
50 55

▨▨▨OUT - side
▨▨▨▨IN - side the closed interval which runs from
56 61

☐☐☐☐☐☐ (lower bound), up to
62 67

☐☐☐☐☐☐ (upper bound), on the variable whose input number is
68 73

▨▨▨☐☐☐ (subscript of filter variable).
74 79

☐ Leave blank
80

Note: If cols. 20-25 are blank, then <u>all</u> input observations will be used in the run and the parameter cols. 26-79 are ignored. If cols. 50-55 are blank, then the parameters in cols. 56-79 are ignored. The above sample sub-group selection does <u>not</u> affect which observations are written on tape (LOCDAT = W). It only determines which observations will be allowed to enter this analysis. DO <u>NOT</u> USE COLS. 56-79 UNLESS YOU ALSO USE COLS. 20-55. Do <u>not</u> use interview number as a filter variable.

III. Parameter Card Type 3

3 = Card Type Must be punched 3
Col. 1

Col. 2 7 = NP = Number of predictors to be used in the analysis. $1 \leq NP \leq 36$.

8 13 = WT = Index of the variable to be used as a weight. If WT = 0, the run is unweighted. Otherwise, $0 < WT \leq NV$.

14 19 = P1 = This should be set at .00001, essentially deactivating it. This will allow P2 to control the termination of the splitting process.

20 25 = P2 = The best split on the ith candidate group must reduce the unexplained sum of squares by P2 proportion of the total sum of squares, or that group will not be split and will not become a candidate group again. The decimal point is punched in the field, e.g., .006 (split reducibility criterion).

26 28 = MAXGP = The maximum allowable number of final groups into which the input data may be split, regardless of P1 or P2. ($002 \leq MAXGP \leq 063$). Normal = 050.

29 31 = MSIZE = Minimum number of observations that must be contained in the ith group if that group is to become a candidate for splitting. Normally MSIZE = 25.

32 37 = Y = Index of the variable to be used as the dependent variable. ($1 \leq Y \leq NV$).

38 49 = YNAME = Alphanumeric name of the dependent variable.

50 55 = YMAX = Any observation such that the dependent variable $V_{y\alpha}$ >YMAX will be omitted from the analysis and printed, thus YMAX is an integer. ($-99999 \leq YMAX \leq 999999$). <u>If nothing is to be omitted, leave blank.</u>

☐☐☐☐☐☐ = MD1 = Any observation such that $V_{y\alpha}$ = MD1 will not be used in the analysis. Note: The program also automatically bypasses all observations such that $V_{y\alpha}$ = -0. (-99999 ≦ MD1 ≦ 999999) Note 2: Leave MD1 blank if not to be used.
Col. 56 61

☐☐☐☐☐☐ = MD2 = Any observation α such that $V_{y\alpha}$ = MD2 will not be used in the analysis. (-99999 ≤ MD2 ≤ 999999). Leave MD2 blank if not to be used.
62 67

☐☐☐ = CDRES = If it is desired to punch residuals from this analysis on cards, this parameter is punched CRD, otherwise it must be left blank.
68 70

☐☐☐ = TPRES = If it is desired to compute residuals and write them on tape for a subsequent analysis, this parameter is punched TAP, otherwise, it must be left blank.
71 73

(Note: If this option is exercised, then the residual is written on tape as variable NV+1. AND THE PARAMETER NV ON THE FOLLOWING ANALYSIS MUST BE ADJUSTED ACCORDINGLY, as there is now one more input variable.)

☐☐☐ = INTNO = Index of the input variable to be punched in the Interview Number field of output cards containing residuals. This field MUST lie in the range 1 ≤ INTNO ≤ NV whenever residuals are to be punched, otherwise it may be left blank or set to zero.
74 76

☐☐ = SCFIN = Input scale factor. This is that power of ten by which Y is to be multiplied before being used in computation. Use 0 for one-zero "dummy" variable dependent variables. Use -4 for residuals of such dummies. Range (-9 ≤ SCFIN ≤ +9). For residuals, normally SCFIN = the previous SCFOUT with the opposide sign.
77-78

☐☐ = SCFOUT = Output scale factor. This is that power of ten by which Y, predicted Y and the residual are to be multiplied before the punching or writing of the residual on tape takes place. Range (-9 ≤ SCFOUT ≤ +9). Use +4 for one-zero "dummy" variable dependent variables. Use +4 for residuals of such dummies. Normally SCFOUT = [(-SCFIN) + 2]
79-80

IV. Type 4 Parameter cards: Predictor list cards.

Column 1 MUST be punched with a 4. on all predictor cards. Insert only as many predictor list cards as are needed to account for NP predictors. Values of all variables used as predictors must lie in the range $0 < X_i \leqslant 63$. The maximum number of predictors is 36. Do not leave any blank fields on the predictor cards except on the physically last card after the last predictor. The Variable Number is the input field number. The predictor type is punched M if the predictor code ordering for that predictor is to be maintained during the partition scan. If the categories of that predictor are to be sorted into descending sequence on their means, then the predictor is a type F (Free) predictor, and the parameter T_p is punched F. Names may include all eligible characters listed on page one.

	Var.No.	Tp.	Name	Var.No.	Tp.	Name
4 Col. 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77
4 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77
4 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77
4 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77
4 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77
4 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77
4 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77
4 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77
4 1	3 5	7	9 20	22 24	26	28 39
	41 43	45	47 58	60 62	64	66 77

AID(2) Run Specifications, Input File Assembly

THIS PAGE TO BE FILLED IN BY PROGRAMMER

IV. MAD Format Statement:* Up to 9 cards, cols. 2-72.

Cols. ENCLOSE IN PARENTHESES

1	2	72
▨	(C1,	
▨		
▨		
▨		
▨		
▨		
▨		
▨		
▨		*)

There must be exactly NV field descriptions, including all predictors, the dependent variable and the weight (if the latter is to be used), together with appropriate S (skip) and / (go to next card) characters. If residuals are to be punched, either on this run, or on any subsequent runs using tape input, then the NV field descriptions MUST include an observation identification field (normally the interview or subject number). This is a C (character) field in mode. Column 2 of the first data card through the units position of the rightmost variable on the last (merged) deck must all be accounted for. All fields are integer (I) in mode except for the interview number field, which is a C (character) field. Insert only as many format cards as are needed. Note the first MUST start with (C1, and the last MUST end with *).

V. DATAFOLLOWS Card:*

Cols.

1 11	12 80
DATAFOLLOWS	

VI. Insert (match-merged) data deck(s) here.* The number of observations must agree with the parameter N. The number of cards is D(N), where D is the number of merged decks. D $\neq$ 0.

VII. Type E Data Trailer Packet:* Insert a packet of D cards, where D is the number of (merged) input decks. An E is punched in column 1 of the first card. The remaining cards are blank.

VIII. As many repetitions of I through VII as desired.

*OMIT THESE CARDS IF INPUT DATA ARE ON TAPE FROM AN EARLIER RUN (LOCDAT = T)

Section 2.8 Program Timing Estimates

Timing examples:

N = 2770	9 predictors	2.5 minutes
N = 2800	14 predictors	3.1 minutes
N = 2997	12 predictors	2.8 minutes
N = 1059 Residuals obtained & results on cards	23 predictors	5.6 minutes
N = 1059	23 predictors	2.5 minutes
N = 2997 Residual obtained--first stage	8 predictors	5.4 minutes
N = 2997 Second stage	18 predictors	3.5 minutes

Section 2.9 Output Page Estimates

$$P = 4 + Q + (M \times Q) + \frac{3M}{2} + \frac{M}{2}$$

$$\text{where } Q = \frac{\sum_{x=1}^{NP} C_x + 5NP}{25}$$

M = maximum allowable groups

ΣC_x = the sum of C_x over all predictors where C_x is the largest (numeric value) code in the predictor x

NP = the number of predictors

Section 2.10 Printed Output Available from the Program

For each analysis:

1. An identifying label.
2. Number of input observations, number of input variables, number of predictors, index of the variable used as a weight, group split-eligibility criterion, group split reducibility criterion, minimum group size, maximum number of allowable groups, index and name of the dependent variable, a definition of missing-data values of the dependent variable which were used in deleting such observations, decimal point locators (scale factors), location of input data, and a definition of which subset, if any, of the input observations were specified for use in the analysis.
3. A dictionary of where the variables came from on the data cards and a record of the mode in which they were stored in the computer, and program timing information.
4. A listing of all of the predictors used in the analysis, their maximum values and the type of predictor (free or monotonic).
5. Statistics for the total number of observations in the analysis including total read, total deleted, total used, and, for the latter, the total sum of weights, sum and sum of squares of the dependent variable, its mean and standard deviation, the total sum of squares (TSS) for the analysis, and the two values PA and PB, that is, the sum of squares that must be contained in a group if it is to be split, and the sum of squares that must be transferred from within to between-group sums of squares for a split to take place.
6. a. A record of the statistics for all attempted partitions of the entire sample (group 1), over all classes of all predictors.

   b. These statistics include, for each class, the number of observations, the sum of the weights, ΣY, ΣY^2, $\overline{Y}$, σ, and the BSS_{ikp} for each possible partition between adjacent classes, and the total sum of squares in the group under attempted partitioning.

Final output for each analysis consists of:

1. A complete definition of each group created during the partitioning process, including the group identification number, the identification of the 'parent' group from which it was split, identification of the variable used to split off this group, the classes of the partitioning variable forming the group, and an indication whether the group was retained as a final group; for the group, the statistics N, Σw, $\overline{Y}$, ΣY, σ, ΣY^2, deviation of the group mean from the grand mean, weighted proportion of the total observations used which are in the group, weighted mean square for the group, the proportion of the total sum of squares in the group, and the sum of squares for the group.

2. A one-way analysis of variance table over the final groups. This should be interpreted with extreme caution, especially when weighted data are used.

3. By option, residuals (discrepancies between observed and predicted values of the dependent variable) may be punched or written on tape, or both, for subsequent analysis. Punched output includes identifying information supplied by the user, the observation number, the identification number of the AID final group into which the observation fell, the predicted value for the observation, its actual value on the dependent variable, and the residual score. Scale factors punched on the control cards provide for the desired number of significant digits in the residuals.

Section 2.11 Residual Output on Punched Cards

If residuals are requested in punched card form, the following output will result for each analysis. One card will be punched for each observation initially read into the computer whether it was used in the analysis or not. These cards have the following format. (Note that the dependent variable is read into the computer as an integer.)

Columns	Content	Remarks
1-12	Identifying information	Obtained from cols. 2-13 of control card 1
13-18	Observation number	Obtained from the input variable identified as the observation number on control card 3. This is a BCD character (C) field and is punched left-justified in the field established for it in the output card.
19-21	Group number	The identification number of the final group of which this observation is a member. If the observation was not used in the analysis, this is zero.
22-29	Predicted Value of Y	This is the mean of the final group of which the observation is a member. If the observation was not used, then this has the value -0. When present this quantity is obtained by computing the group mean to 8-place floating point accuracy, multiplying the result by the output scale factor (decimal point locator) and then rounding it to the nearest integer for punching. (Input values of the dependent variable must be integer in mode, but may be scaled appropriately via a control card parameter.
30-35	Actual Value of the Dependent Variable Y	Obtained from the input variable designated as Y by the parameter on control card 3.

Column	Content	Remarks
36-43	Residual	$R = Y - \overline{Y}_i$, where i is the group number of the final group of which the observation is a member. It is computed to 8 place accuracy, multiplied by the output scale factor (decimal point locator), and then rounded to an integer for punching. If the observation was not used in the analysis, the residual is set to -0.
44-50	Weight	This is the constant 1 if the run is unweighted. Otherwise, it is obtained from the input variable designated as a weight on control card 3.

Note: Normally the contents of cols. 1-12 on the output residual cards should contain the research project number starting in column 1, followed immediately by the deck identification number of the new deck produced by the computer. The residuals are punched in exactly the same order as the observations in the input file. For group means, values of the dependent variable and also for residuals, signs are punched to the left of the most significant nonzero digit and the remainder of the field to the left is blank.

Section 2.12 Residual Output on Tape

If residuals are requested for storage on tape, they will be computed as indicated above and then stored on a data-tape along with all variables entered as input. Thus, if the input consists of NV variables, predictor variables, dependent variables, weights, filters, etc., then the residuals will be written on the tape as variable NV + 1. Residuals which are undefined, either because the dependent variable is undefined (missing data) for that observation or because the observation was prevented from being used in the analysis by the use of the "filter" will have a value of -0. They will be omitted automatically from any subsequent analysis which uses the data on the tape and which specifies these residuals as the dependent variable.

Each analysis specifying residuals to be left on tape will result in additional variable; the residual from that analysis will be left on the tape. Thus, suppose four analyses are performed. The first specifies card output of residuals. The data input consists of 56 variables. It is requested that the data be saved on tape. At the end of this analysis, there will be 56 variables on tape. The control cards for the following analysis will specify 56 input variables. If they specify residuals to be left on tape, then, at the end of that analysis, the tape will have 57 variables on it, the 57th being the residuals requested. A further analysis using the tape must specify 57 variables as input. If tape residuals are again requested, then after the termination of this analysis, there will be 58 variables on the tape. The fourth analysis, if it is to use the tape, must specify 58 variables as input. A fifth analysis, if the data come from cards, may either write a new tape or ignore it, but may not add additional residuals to it. Thus, any time data come in from cards and either tape residuals or the saving of the data on tape are requested, a new tape is written and the old one destroyed. There are no provisions in the program for saving tapes which have been written. It is assumed that the primary data-storage mode is on cards.

CHAPTER III

ILLUSTRATIONS AND EXAMPLES

Section 3.1 Introduction

We present a series of thumbnail analyses drawn from computer runs that were made on the program. Our objective is to illustrate the output available from the program, analysis strategy with respect to its interpretation, and to point out the sensitivity that the method has when problems occur, such as a skewed dependent variable or un-interpretable splits associated with predictors of considerable conceptual complexity.

A number of the trees presented use sets of predictors that had previously been employed in a multiple classification analysis. This technique (22) is equivalent to a dummy-variable multiple regression (23). One objective has been to determine whether the findings based on the trees were consistent with previous analyses, and whether additional information about the structure of relationships between the variables could be extracted from the trees. With a few exceptions, which will be noted later, these expectations appear to be fulfilled.

Nine examples are presented. The first is a two-stage analysis where the objective is a stringent test of the effectiveness of a factor (occupation) known to have a very powerful effect on average hourly earnings. Complete documentation of the entire run is presented, including a listing of the input, codes for the variables and the computer output.

The second example (home ownership) illustrates the use of a dependent variable which is dichotomous, rather than equal-interval. Parsimonious explanation is achieved, together with clear evidence that neither family size nor age are uniform in their effects throughout the population.

The following example (plans to move) introduces an assumption of an underlying continuum. The concept of alternative inhibiting factors is illustrated. The fourth example (nonfamily contributions) illustrates the type of analysis problems that arise when the dependent variable is badly skewed. An analysis strategy for handling this problem is presented. The effects of using predictors, which are themselves complex indices representing several dimensions, are illustrated. Several questions which the analyst should raise when interpreting the tree output are suggested.

The next example (expected family size) constitutes a re-analysis of data which have been extensively studied, to determine whether the behavior of the variables in the trees were consistent with previous findings. Generally, this was found to be the case. However, the importance of keeping the number of classes in the predictors to a minimum and of constraining the ordering of those which have a natural ordering to them is clear. The illustration emphasizes the need for predictors which are as uni-dimensional as is possible. The sensitivity of the procedure to this type of conceptualization problem indicates its possible use in locating concepts in need of refinement. Coding the offending variable somewhat differently may then be possible, leading to better discriminatory power for it when used.

The following example (average completed education) illustrates the use of several methods of displaying the results for further analysis, together with a hypothesis suggested by one of the splits.

The seventh example (disposable income) illustrates a nonsymmetric effect by a series of handicaps and cumulative advantages. The stability of the procedure is investigated by applying a tree to a subsequent sample.

The next example illustrates use of the procedure to locate interaction terms for inclusion in a multiple regression analysis. Interpretation problems from the inclusion of indices representing complex interactions as predictors are noted.

The final example (number of hours worked) provides another illustration of a two-stage analysis. Variables which were felt to be early in a possible causal chain (in the sense that they could influence other predictors, but could not themselves be influenced by the other predictors) were put into the first stage. The results provide an interesting picture of constraints operating to reduce the number of hours worked, rather than of motivational factors.

When going through these examples, the reader should keep in mind that, unlike a multiple regression technique, this procedure allows predictors to substitute for each other in explaining variation in the dependent variable. Thus, when examining each split, the question, "What are the reasons why the split was made on this variable, rather than on one of the other predictors?" should be kept in mind.

Section 3.2 Average Hourly Earnings

A complete analysis is presented, illustrating various aspects of the revised computer program (AID Model 2), and several strategies which may be employed in the interpretation of the results. The objective was to replicate a previous analysis (24) of average hourly income. Some of the variables (e.g., place where head grew up) are multidimensional, since they were used previously in an additive model and interactions had been suspected between their components. A two-stage analysis was employed for the dual purpose of separating out exogenous factors from more current situational factors and providing a stringent test of the explanatory power of occupation. The latter was accomplished by putting it in with the second-stage predictors. A listing of the computer input, the complete output and supporting documents are included. (See Appendices K, L and M.)

An equivalent hourly earnings measure was computed for the heads of spending units for a national sample (the quotient of head's total wage income divided by hours worked x 100). Where the head had no wage income, the value was assigned to this variable. These observations (N = 451) were omitted from the analysis.

A two-stage analysis strategy was adopted. All variables to be used in both stages were used as input to the program. These variables are identified and described in detail below. The following variables were used in stage one:

Variable Number	Name	Number of Classes
1	Physical Condition of Head	4
3	Education of Head	8
8	Rank in School	8
11	Race	2
12	Age	7
22	Sex	2
23	Religion	4
24	N/Ach (need-achievement score)	4
25	Background (place where head grew up)	6

Since a multi-stage probability sample with varying sampling fractions was used, the analysis employed weights attached to each observation to adjust for differences in sampling and response rates. At the end of stage one, residuals were computed. These residuals were used as the dependent variable (with the same sample weights) in stage two. The following variables were used in stage two:

Variable Number	Name	Number of Classes
2	Geographic Mobility (number of states lived in)	6
3*	Education	8
4	Immigration (of head or father)	3
5	Occupation	10
6	Supervisory Responsibility on Job	3
7	Frequency of Unemployment	8
9	Religion x Church Attendance	7
10	Attitude toward work x N/Ach (achievement motivation index)	7
11*	Race	2
13	Education difference between Head and Wife	7
14	Urban-Rural Migration	5
15	North-South Migration	6
16	Family Composition (sex, marital status, number of children)	8
17	Plans to help parents and children	4
18	Interviewers' rating on ability to communicate	4
19	Size of place (city size)	6
20	Educational difference between Head and Father	4

The variables used in stage one were suspected to be logically prior to those used in stage two. The starred items, Education and Race were used in both stages. They were included in the second stage on the hypothesis that they were likely to have both direct and indirect effects, and they were likely to interact with occupation in explaining variation in the residuals. The index of achievement motivation, and religion, were each reintroduced in combination with an allied variable.

The stage-one tree is presented in Chart 1. The total reduction in prediction error from these variables is .242, which corresponds roughly to a multiple R^2 of that size. Physical condition, rank in school, race and religion were not actually used by the program.

Stage one shows the powerful effects of education, age and sex. Achievement motivation appears important only for college graduates over 35 years of age. Rural-urban-north-south background appears important only for noncollege graduates.

Structure of the Tree

After the initial division of the sample into three parts (groups 3, 4 and 5), the branching process follows a "trunk-twig" pattern. That is, successive branches isolate a subgroup, which is not split further.

The reasons why these groups are not split further is of some theoretical importance. Either the number of observations is too small to warrant splitting the group, or the proportion of variation in it, compared to the variation in the total sample is too small, or no predictor in the analysis is capable of reducing the unexplained variation in that group the requisite amount.

If we consider groups 14, 21, 23 and 25, we find that the latter three are either too small to split, or do not have sufficient internal variation to warrant an attempted split. Group 14 cannot be split further, even though it has sufficient internal variation to warrant an attempt. No predictor "works." Age comes closest, but does not reduce the unexplained variation enough for the split actually to take place.

Two other groups, group 8 and group 19, did not have sufficient internal variation to warrant an attempted split, though they contained 95 and 73 persons respectively. For the other final groups, 7, 10, 16, 17, 12 and 18, no predictor "worked." If a group has a small variance, it has been explained. If it has a large one and no predictor works, then additional variables are needed in the analysis.

The tree illustrates a complex interaction between age, education, sex, N/Ach, and background. The trunk-twig structure indicates what one might call the "alternative barrier" situation with respect to achieving high hourly income. If one is a college graduate, being under 35 years old is a "barrier," which cannot be surmounted by being characterized by any set of classes of the predictors used in the analysis, at least under the split criteria set up. If one passes this hurdle, then the absence of middle or high achievement motivation constitutes a barrier, etc. (see Table 1).

The same description applies to the noncollege-graduates who did not grow up on a farm or in the south. Being a woman (group 7), or being young (group 8), or failing to complete high-school (group 10), constitute alternative barriers (see Table 2).

Similarly, for noncollege-graduates who grew up on a farm or in the south, completing less than nine grades of school is a barrier, as is being a woman (see Table 3). Considering groups 7, 8, 10 and 16, it is clear that there are different sets of barriers for men than there are for women, since group 7 (women) was not split further, though eligible (education was almost good enough to be used to split group 7).

This stage one tree illustrates the extent to which variables may substitute for one another in the analysis, depending on how they are correlated with the dependent variable. For instance, an examination of Table 1 indicates that the Urban-Rural-Farm-Nonfarm background of the Head was almost as good as Education in the split of group one into two and three. It was not used at that stage, but did not have its relationship to the dependent variable reduced enough by the split to prevent its being used in the split of group two into four and five. However, in group three, its relationship to the dependent variable has dropped considerably, and it was not used in further splits on this trunk. It appears important, from an analysis standpoint, to make a careful examination of those variables which were not used in the tree, but which, as it were, "almost made it."

Rank in School is another case in point. Examination of Table 2 indicates it was second-best in a number of branches, and would have been used if group 10 had been permitted to split by lowering the reducibility criterion.

Education as a Substitute for Race

Another example of substitutability is the variable Race. There is plenty of evidence that being white or nonwhite affects one's wage rate. In this sample, the mean wage rate for whites is $2.38, for nonwhites it is $1.60. Moreover, in each of the final groups except one (see Table 4) there are white/nonwhite discrepancies between group means ranging from $.11 to $1.49. Some of the N's in these groups are too small to put much trust in, but the replicated discrepancies point overwhelmingly to an important race effect. Furthermore, the mean residuals for nonwhites are -$.35. If race exercised no effect, this would be closer to zero. Clearly there is a race effect. Why doesn't it show up in the tree?

We may reason as follows. Race may be considered to affect wage rates directly, and also indirectly, through its effects on other variables, which in turn affect wage rates. This combination of effects is undoubtedly quite complex and a detailed analysis is beyond the scope of this discussion. However, a discussion of race, education and wage rates will serve to illustrate an analysis strategy based on the algorithm.

We may hypothesize that race affects wage rates partly through its effect on education. Education is clearly a powerful predictor; but other things than race affect education. If this indirect effect is occurring, we should expect to find that nonwhites tend to have less education than whites. A stringent test of the hypothesis that this indirect effect is occurring would be to examine the relationship between race and education in each of our final groups. If nonwhites tend to have less education, the hypothesis of the existence of this indirect effect would be confirmed. An examination of the bivariate frequency distributions between race and education for each of the final groups tends to confirm this interpretation. In groups 14, 23, 25, 7, 8, 10, 16, 17, and 12, whites tended to have a higher proportion of individuals in the upper educational categories. For instance, in group 17, we find (percentages based on weighted data):

	N	Per cent having only a high-school education or less	Per cent having additional vocational or college training	Total
White	297	45%	55%	100%
Nonwhite	14	62	38	100
Total	311	45	55	100

Group 21 had no nonwhites. In group 24 (college graduates) nonwhites had a slightly higher proportion of persons with advanced degrees. In this group, as in most of the others, however, the N's are relatively small. Groups 18 and 19 show a somewhat different pattern indicating that for rural and/or southern noncollege-graduate males, the pattern of relationships between race, education and sex is somewhat more complex. There is a larger proportion of high school drop-outs among nonwhites than among whites. Nonwhites who got education past high school tended to go to college rather than get other types of training. Perhaps there are a number of factors influencing the types of post high school education obtained by these males. The statistics for group 18 are as follows (percentages are based on weighted data).

	N	Some High School	High School Graduate	High School plus Noncollege Training	College, No Degree	Total
White	410	45%	26%	13%	16%	100%
Nonwhite	58	60	10	6	24	100
Total	468	46	25	12	17	100

The fact that a very similar pattern is repeated in group 19 (females with similar backgrounds) lends credence to the complexity notion.

Interrelationships between the Predictors

Additional hints as to the structure of interrelationships among the variables may be found in a manner similar to that used in constructing Table 4, by running frequency distributions on the predictors not used by the program. For instance 50 per cent of the American-born sons of immigrants are in groups 10 and 17, approximately 25 per cent in each. The proportion of persons in group 10, high school drop-outs, scoring low, intermediate and high on the N/Ach predictor is, contrary to what might be expected, almost exactly the same as that for the total sample. Referring to Table 2, we see that Rank in School and Race were almost powerful enough to split group 10.

Groups 10, 12 and 18, and to a lesser extent, group 17, are similar in that they constitute relatively large numbers of respondents and are not splittable in terms of the algorithm and the split criteria. No single variable "works." And the analyst must consider the possible reasons why these groups could not be split. This suggests a possible revision of the program algorithm to consider the effects of each pair of predictors simultaneously for this type of group since there may exist negative, offsetting interactions. This might be done in either of two ways, which are similar, but not identical. One method would involve the treatment of a two-way analysis of variance table so that the methods outlined in the present algorithm are used on both the rows and columns simultaneously. An alternative would be to postpone the actual splitting process until the split rules which produce minimum within group variation in all possible "grandchildren" of the parent group under consideration, have been determined. This would constitute a "look-ahead" one step down each branch of the tree.

Stage Two

Only two of the variables allowed as predictors in stage two were used, occupation and husband-wife educational differential. That occupation should pass this severe test of its effectiveness as a predictor is to be expected. The selection and use of the husband-wife educational differential is somewhat surprising. Group 5 was 36 per cent female heads of spending units and 64 per cent males. All of the females from group 2 are located in group 5. Thus, the split reflects partly a male-female differentiation. In group 5, 26 per cent of the respondents are single males. Thus, 62 per cent of this group are single. The remainder are married male heads of spending units. As we might expect (see Table 5), family composition is almost as good a predictor as husband-wife educational differential, in the attempted split of group 2.

One way of interpreting this is to examine the nature of the two variables. Husband-wife educational differential may be considered to be tapping at least three sources of variation; sex, marital status among males, and husband-wife educational differentials among married males. Family composition taps only two of these sources, sex and marital status. But we note that in the program output detail for the split of group 2 into groups 4 and 5, that we do, apparently have an educational-differential effect, as is indicated below.

	Mean	N	Group
Education of Wife N.A.	+.61	9	4
Wife has two or more levels of education more than head	+.56	149	
Wife has one more level than head	+.39	244	
Wife has same level	+.27	496	
Wife has one less level	+.27	264	
Wife has two or less levels	+.04	251	5
No wife present (male and female)	-.06	408	

This variable apparently had no further effects in groups 6 and 7 (see Table 5), but after the farmers were separated out of group 3, it still showed some effects in groups 8 and 9.

Summary

This example has been presented to illustrate the use of a two-stage analysis to provide a stringent test of the effects of a variable which is known to be of considerable theoretical importance (occupation) and which has high correlations with other important variables, such as education.

The difference between two types of final groups, homogeneous or small, and unsplittable has been described.

The "trunk-twig" or alternative barrier tree structure as opposed to a more symmetric or "trunk-branch" structure, has been discussed.

Several examples of the substitutability of variables as a characteristic of the analysis algorithm have been presented and their implications for interpretation have been discussed. A strategy for investigating the extent to which a variable which has been used in a split is substituting for other variables is presented, together with its converse, a strategy for investigating why a variable which has considerable outside evidence as to its effects--does not get used.

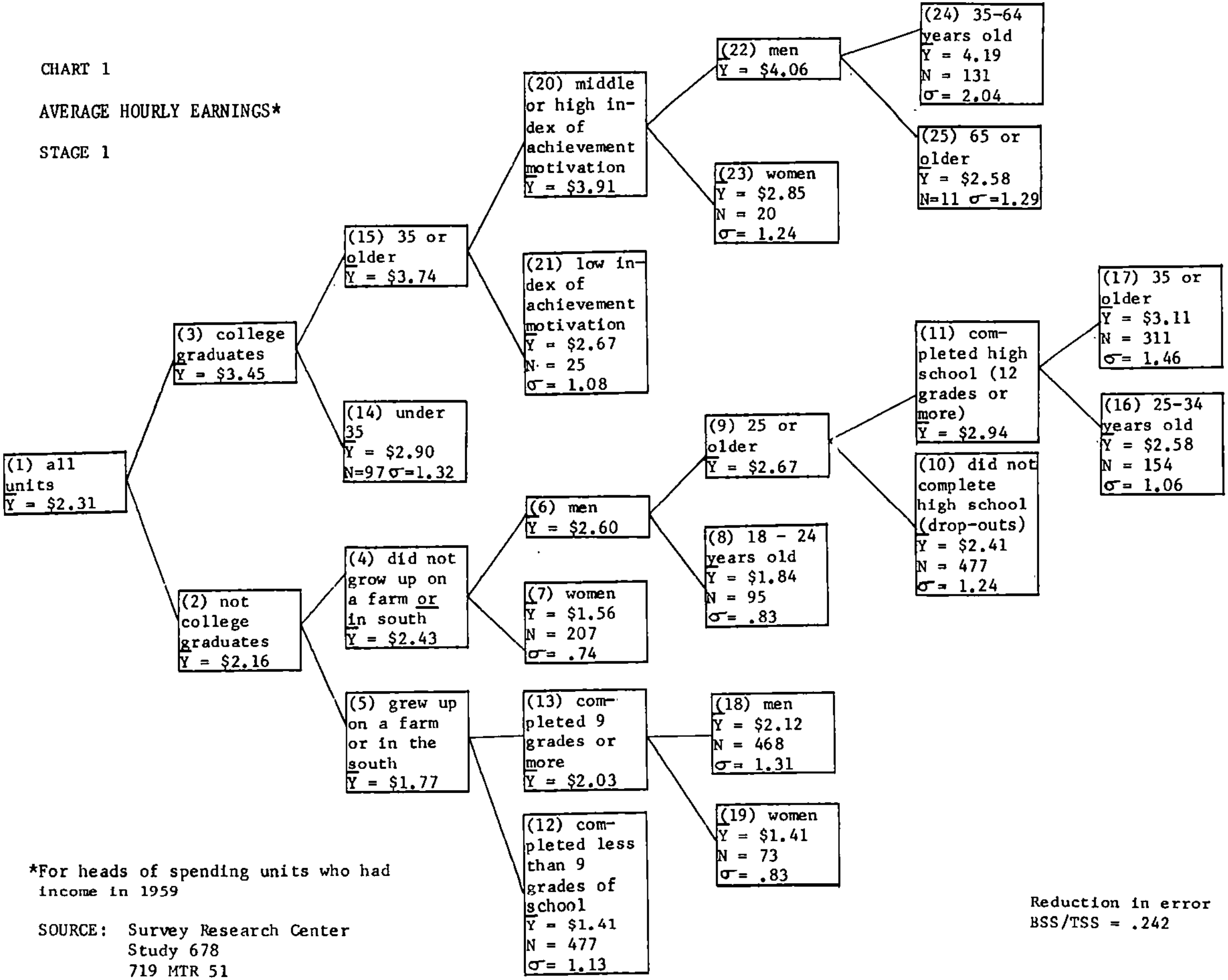
It is recommended that all output options be exercised, including the punching of residuals as an aid to simplifying further analyses.

Some suggestions for further possible revisions in the analysis algorithm are made.

CHART 1

AVERAGE HOURLY EARNINGS*

STAGE 1



Reduction in error
BSS/TSS = .242

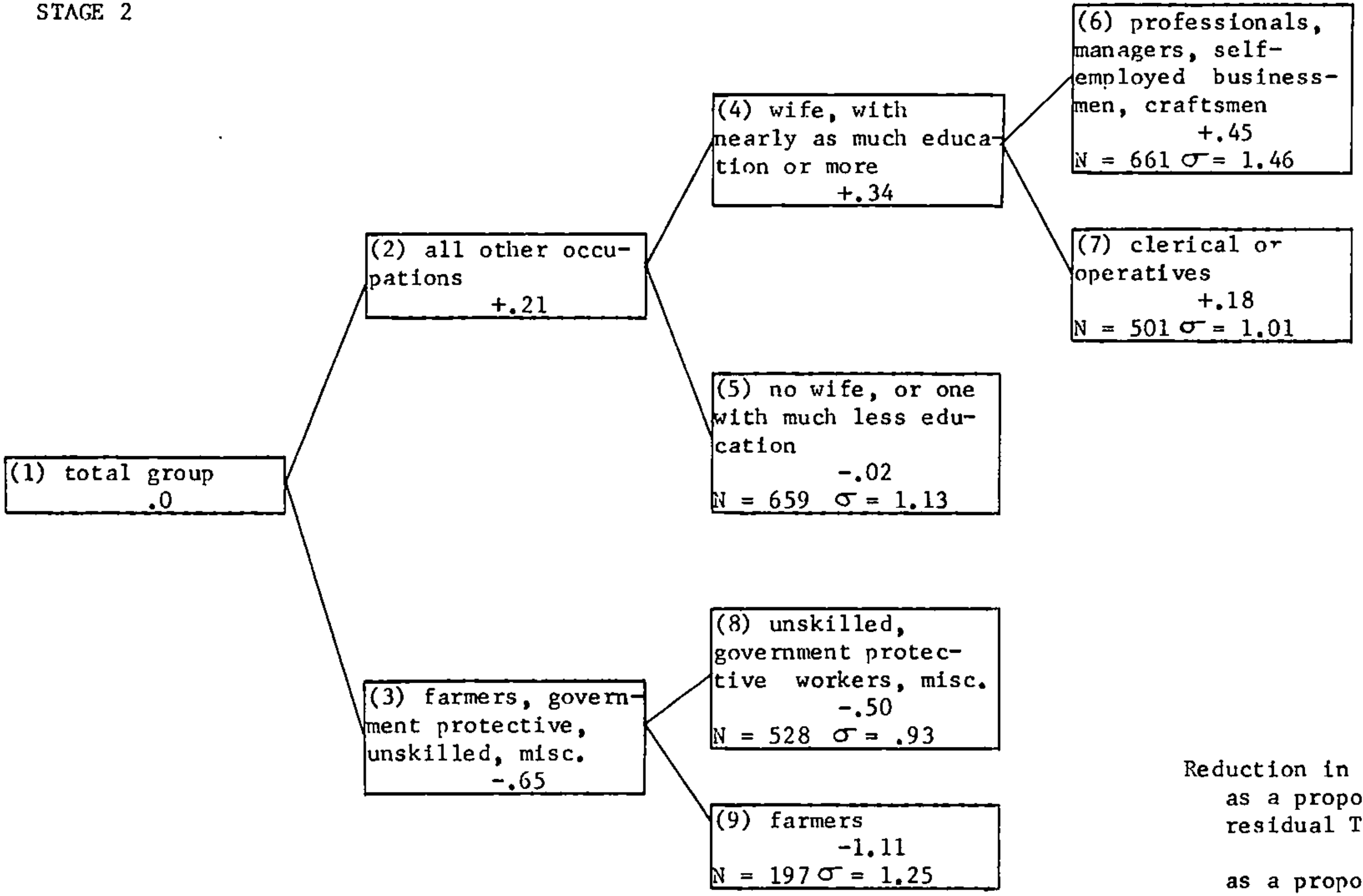
*For heads of spending units who had income in 1959

SOURCE: Survey Research Center
Study 678
719 MTR 51

CHART 2

RESIDUALS - AVERAGE HOURLY EARNINGS*

STAGE 2



Reduction in error
as a proportion of
residual TSS = .114

as a proportion of
original TSS = .087

Total reduction in error
from two-stage analysis
as a proportion of original
TSS = .329

*For heads of spending units who had income in 1959

SOURCE: Survey Research Center
Study 678
719 MTR 51

Table 1

WAGE RATE ANALYSIS STAGE 1
COLLEGE GRADUATES ONLY

	Group Number									
	1	3	14*	15	21*	20	23*	22	25*	24*
Physical condition	.015	.001	.002	.002	NA	.001	NA	.000	NA	.001
Education	.083	.029	.016	.023		.016		.015		.012
School rank	.041	.012	.001	.026		.014		.009		.005
Race	.027	.004	.011	.016		.023		.029		.026
Age	.020	.049	.079	.038		.038		.046		.012
Sex	.047	.021	.000	.043		.042		Constant		Constant
Religion	.030	.021	.030	.021		.025		.023		.014
N/Ach	.017	.017	.015	.048		.007		.007		.012
Background	.067	.019	.047	.025		.022		.023		.011
N	2546	284	97	187	25	162	20	142	11	131
TSS_i/TSS_T	1.0	.180	.033	.138	.006	.125	.006	.114	.004	.105
MEAN	2.31	3.45	2.90	3.74	2.67	3.91	2.85	4.06	2.58	4.19

Proportion of variation in that group explainable for each predictor $(BSS/TSS)_i$

→ = Split made on this variable.

xxx = Next best BSS/TSS.

* = Final group.

↷ = Split attempted but not made.

NA = Split not attempted.

Source: ISR Study 678, Deck 35
719, MTR 51

Table 2

WAGE RATE ANALYSIS STAGE 1
NONCOLLEGE GRADUATES WHO DID NOT GROW UP ON A FARM OR IN THE SOUTH

	Group Number										
	1	2	4	7*	6	8*	9	10*	11	16*	17*
Physical condition	.015	.014	.003	.028	.002		.003	.005	.008	.007	.008
Education	.083	.049	.023	.092	.029		.040	.009	.003	.027	.002
School rank	.041	.022	.016	.076	.023		.025	.019	.013	.032	.012
Race	.027	.027	.011	.017	.009	NA	.009	.017	.004	.009	.000
Age	.020	.016	.038	.017	.032		.009	.013	.033	Const.	.013
Sex	.047	.053	.086	Const. F	Const. M		Const. M	Const. M	Const. M	Const. M	Const. M
Religion	.030	.027	.012	.070	.014		.012	.014	.013	.023	.005
N/Ach	.017	.013	.007	.036	.010		.010	.008	.008	.047	.006
Background	.067	.060	.003	.006	.004		.005	.004	.000	.016	.003
N	2546	2262	1244	207	1037	95	942	477	465	154	311
TSS_i/TSS_T	1.0	.738	.415	.022	.358	.013	.334	.146	.174	.034	.134
MEAN	2.31	2.16	2.43	1.56	2.60	1.84	2.67	2.41	2.94	2.58	3.11

Proportion of variation in that group explainable for each predictor $(BSS/TSS)_i$

↷ = Split made on this variable.
(xxx) = Next best BSS/TSS.
* = Final group.
↻ = Split attempted but not made.
NA = Split not attempted.

Source: ISR Study 678, Deck 35
719, MTR 51

Table 3

WAGE RATE ANALYSIS STAGE 1
NONCOLLEGE-GRADUATES WHO GREW UP ON A FARM OR IN THE SOUTH

	Group Number						
	1	2	5	12*	13	18*	19*
Physical condition	.015	.014	.033	.026	.017	.015	NA
Education	.083	.049	.059	.009	.004	.006	
School rank	.041	.022	.026	.028	.002	.003	
Race	.027	.027	.027	.014	.020	.011	
Age	.020	.016	.016	.017	.019	.020	
Sex	.047	.053	.034	.038	.033	Constant	
Religion	.030	.027	.011	.005	.007	.011	
N/Ach	.017	.013	.017	.011	.015	.017	
Background	.067	.060	.018	.004	.011	.020	
N	2546	2262	1018	477	541	468	73
TSS_i/TSS_T	1.0	.738	.278	.094	.167	.153	.009
MEAN	2.31	2.16	1.77	1.41	2.03	2.12	1.41

Proportion of variation in group explainable for each predictor $(BSS/TSS)_i$.

→ = Split made on this variable.

(.xxx) = Next best BSS/TSS.

* = Final group.

↷ = Split attempted but not made.

NA = Split not attempted.

Source: ISR Study 678, Deck 35
719, MTR 51

Table 4

WAGE RATE ANALYSIS STAGE 1
MEAN INCOME BY RACE WITHIN GROUP

Group	N	White		Nonwhite		Discrepancy
		Mean	N	Mean	N	
14	97	2.87	93	3.64	4	-.77
21	25	2.67	25	--	0	--
23	20	2.86	19	2.65	1	+.11
24	133	4.29	125	2.80	8	+1.49
25	11	2.69	10	1.43	1	+1.26
7	207	1.60	180	1.28	27	+.32
8	95	1.86	87	1.58	8	+.28
10	477	2.45	439	1.77	38	+.68
16	154	2.62	135	2.30	19	+.32
17	311	3.12	297	3.00	14	+.12
12	477	1.49	328	1.17	149	+.32
18	468	2.16	410	1.71	58	+.45
19	73	1.54	51	.92	22	+.62
Total	2548	2.38	2199	1.60	349	+.78

Source: ISR Study 678, Deck 35
719, MTR 51

Table 5

WAGE RATE ANALYSIS STAGE 2
RESIDUALS

	Group Number								
	1	2	5*	4	6*	7*	3	8*	9*
Geogr. Mobil.	.007	.007	.005	.008	.011	.004	.005	.005	.006
Education	.002	.001	.008	.006	.006	.004	.007	.017	.009
Immigr.	.000	.000	.001	.000	.003	.008	.000	.000	.006
Occup.	.084 →	.010	.016	.010 →	.001	.000	.064 →	.001	--
Supv. Resp.	.017	.008	.007	.007	.007	.002	.041	.003	.012
Freq. Unempl.	.023 ◁	.007	.021 ↶	.002	.003	.007	.023	.006	.004
Rel. x Att.	.008	.006	.013	.009 ◁	.012 ◁	.016	.007	.012	.016
Work x N/Ach	.006	.002	.009	.002	.001	.003	.008	.012	.010
Race	.009	.003	.005	.002	.000	.005	.000	.003	.000
H-W Educ.	.012	.019 →	.002	.005	.008	.005	.005	.022	.029 ◁
Urb-Rur Mig.	.013	.008	.011	.007	.005	.026 ↶	.045 ◁	.037 ↶	.010
N-S Mig.	.006	.005	.004	.007	.010	.017 ◁	.025	.025 ◁	.051 ↶
Family Comp.	.016	.017 ◁	.020 ◁	.004	.007	.003	.008	.014	.012
Help Par. & Child	.011	.006	.009	.001	.001	.004	.006	.002	.029 ◁
Comm. Abil.	.008	.002	.002	.005	.006	.003	.001	.003	.010
Size of Place	.029 ◁	.007	.019	.007	.013 ↶	.003	.042	.013	.006
H-Fa. Educ. D.	.002	.001	.011	.000	.004	.010	.002	.006	.009
N	2546	1821	659	1162	661	501	725	528	197
TSS_i/TSS_t	1.0	.748	.225	.509	.370	.134	.168	.098	.060
MEAN	.783	.205	-.021	.336	.452	.185	-.653	-.495	-1.110

Proportion of variation in each group explainable for each predictor $(BSS/TSS)_i$

→ = Split made on this variable.
◁ = Next-best BSS.
* = Final group.
↶ = Split attempted but not made.
-- = Variable is constant in this group.

Source: ISR Study 678, Deck 35
719, MTR 51

Section 3.3 A Dichotomous Dependent Variable--Home Ownership

Home ownership in early 1959 was analyzed using data from the 1959 Survey of Consumer Finances (25) in which 2980 nonfarm spending units were interviewed. They were weighted to account for varying sampling and response rates. The explanatory classifications allowed (all free to be rearranged) were:

Characteristic	Number of subclasses, including missing information
Age of head of unit	7
Number of people in the unit	10
Income	9
Education of head of unit	7
Race	4
Number of "major" earners ($600 or more)	6
Whether income last year was unusual (a combination of reported income change, unemployment in 1958, and whether head was in the labor force)	8

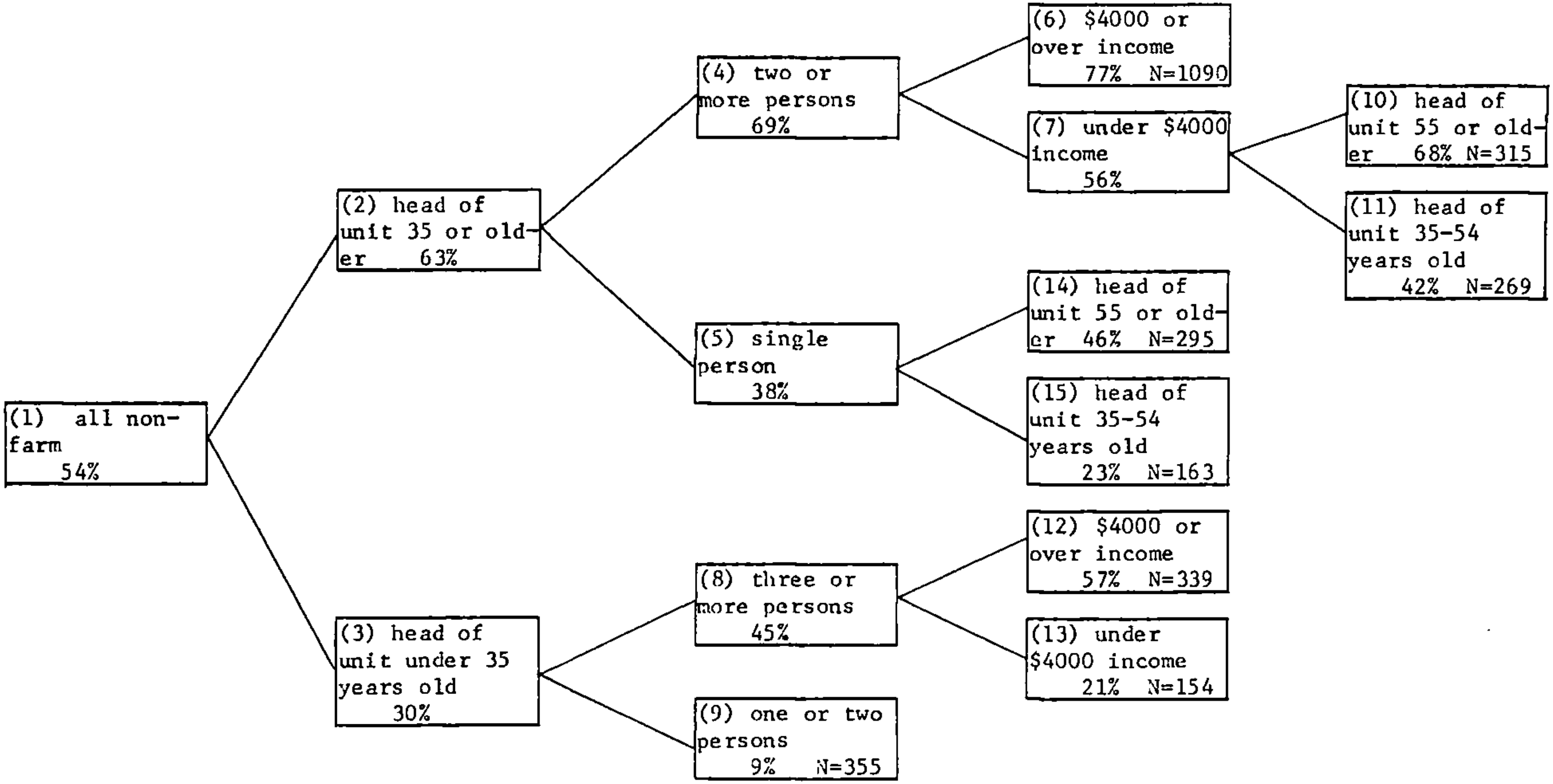
The eight final groups in the tree (see Chart 3) accounted for some 23 per cent of the total sum of squares, making use of only three of the seven factors: age, income, and number of people in the unit. A previously-run multiple regression using the same predictors found age, income, number of persons, race, and "whether last year's income was unusual" all significant, and explained the same fraction of the total sum of squares. According to either analysis, the proportion of home owners increases with age, with additional influences from higher income and larger families. What the tree adds is the impression that it takes a wife and children to push the young to home ownership, and then only if their income is adequate, whereas among the older people marriage is enough, with the single people becoming home owners mostly after they are 55 years old.

It is certainly more economical to explain home ownership with eight combinations of three characteristics, rather than the 45 subclasses of seven predictors used in the regression. More important, we are kept from assuming that there is a single uniform effect of family size on home ownership, or of age on home ownership. Interestingly enough, however, the best income division to discriminate older couples as to their home ownership was the same as the best division for younger families (most of which have children).

It should be noted that even though age was used in an early split, it was still eligible to be used again, and was used in a later split. The program does not discard a variable after using it once.

CHART 3

HOME OWNERSHIP IN EARLY 1959
BY CHARACTERISTICS OF SPENDING UNITS



SOURCE: ISR Project 719
MTR 20

Section 3.4 Plans to Move

An example (26) of a relatively simple scaled dependent variable resulted from asking 2384 people who were in the labor force in August 1962 or November 1962 the questions:

> Do you think there is any chance you will move away from (town or place where now living) in the next year?
>
> If some chance: Would you say you definitely will move, you probably will, or are uncertain?

Those who said they definitely or probably would move were coded "2," those who were uncertain were coded "1," and those who indicated little chance of moving were coded "0." The assumption is made explicit that these points are deemed to represent approximately equal intervals on an underlying continuum "probability of moving."

The prior multiple regression analysis was done separately for four subgroups on the assumption that there might be interaction effects, i.e., that other factors might operate differently on each of them. The four were:

Mean Score	
.22	People under 35 years of age living in a redevelopment area
.28	People under 35 years of age living elsewhere
.10	People 35 or older living in a development area
.11	People 35 or older living elsewhere

There appeared a tendency for one variable (having relatives living nearby) to affect mostly the young. Another (whether moved in last five years) affected mostly those not in redevelopment areas. Two variables (whether unemployed last year, whether owns home) tended to affect only those 35 and over and not in a redevelopment area. One (whether a college graduate) affected only those under 35 not in a redevelopment area and one (being very young, 18-24 years old) affected only those under 35 and in a redevelopment area.

Chart 4 using the same explanatory factors, gives quite a different impression. Neither of the two factors assumed to be crucial in the four regression subgroups appear in the tree. The first split makes use of past mobility (which was significant in only two of the regressions). The variables used were:

Variables	Number of classes
Age of head of the spending unit	7
Education of head	8
Whether a redevelopment area (a county or pair of counties designated by the Area Redevelopment Administration as having sufficiently low income or sufficiently high unemployment to qualify for assistance)	4
Financial reserves (assets)	7
Whether owns a car	2
Whether has children in school	2
Whether wife works	2
Whether moved in the past 5 years (Chart 4 only)	2
Whether pay is perceived as higher elsewhere	2
Whether unemployed during the last year	2
Whether has relatives living nearby	2
Whether would lose some pension rights by changing jobs	2
Whether owns home	2

Having children in school, which appeared significant in none of the regressions, makes an important split among those who have moved in the past five years. Other splits use car ownership, significant nowhere in the regressions, and education which was significant only in one.

Two problems are apparent with this tree. First, the combinations of education are difficult to interpret. Second, and more important, there is some circularity in using past mobility to explain expected mobility. For predictive purposes this may be all right, but it does not "explain" mobility.

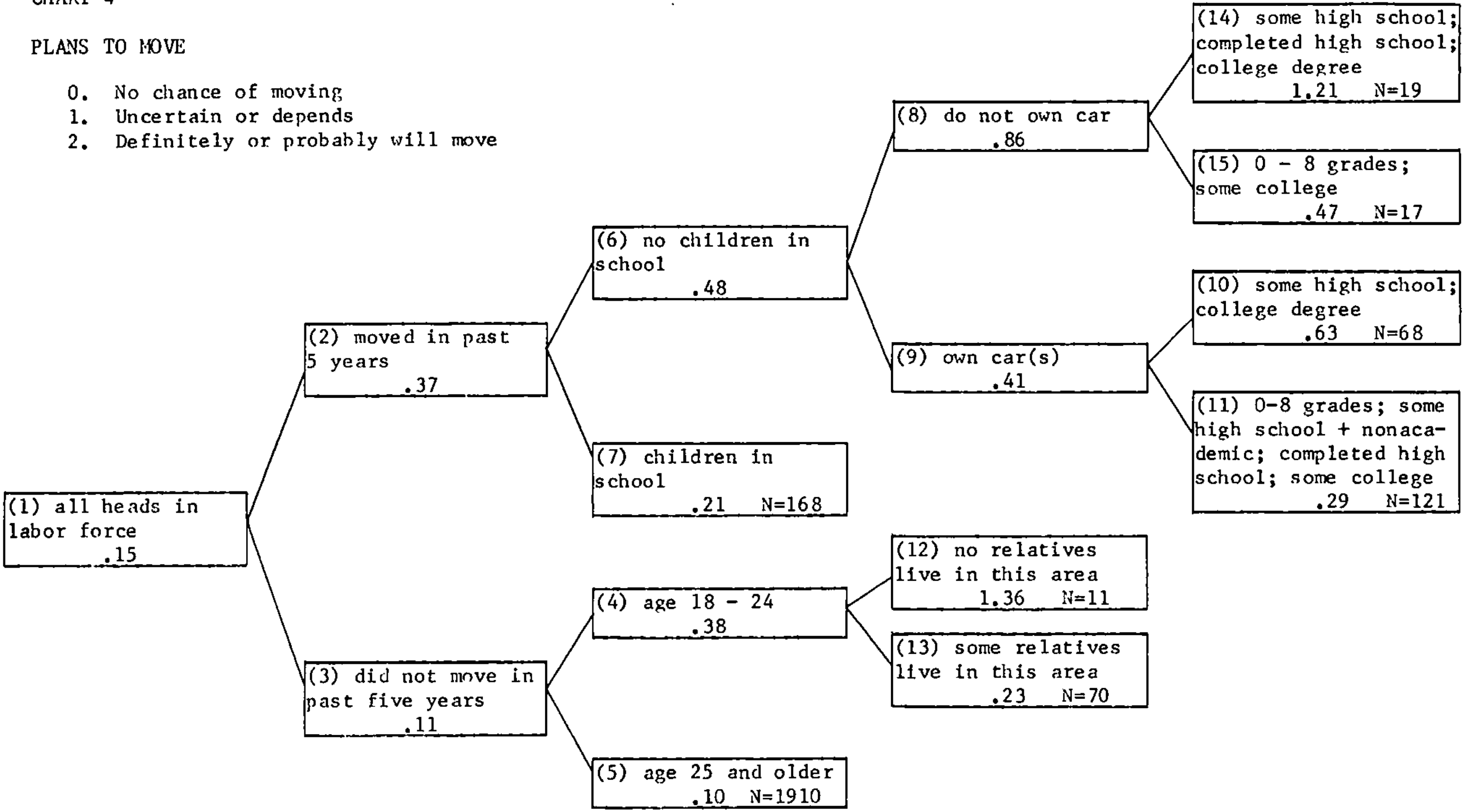
A second analysis was made, omitting only "whether moved in the past five years," and is presented in Chart 5. Instead of the full tree with the .005 reducibility criterion, this tree has been truncated (some very small final groups were combined into their parent). The results have an intuitive appeal to them, and provide a vivid impression of alternative inhibitors of moving: age, relatives nearby, children in school, or owning a home.

In such a situation, the particular sequence of splits may well be unstable, since once one factor is used, the other can only influence the nonhampered group. Subsequent analysis might well be done developing a new variable: "Any one of the following inhibiting factors is present," and would involve an analysis of the correlations between the predictors.

CHART 4

PLANS TO MOVE

0. No chance of moving
1. Uncertain or depends
2. Definitely or probably will move

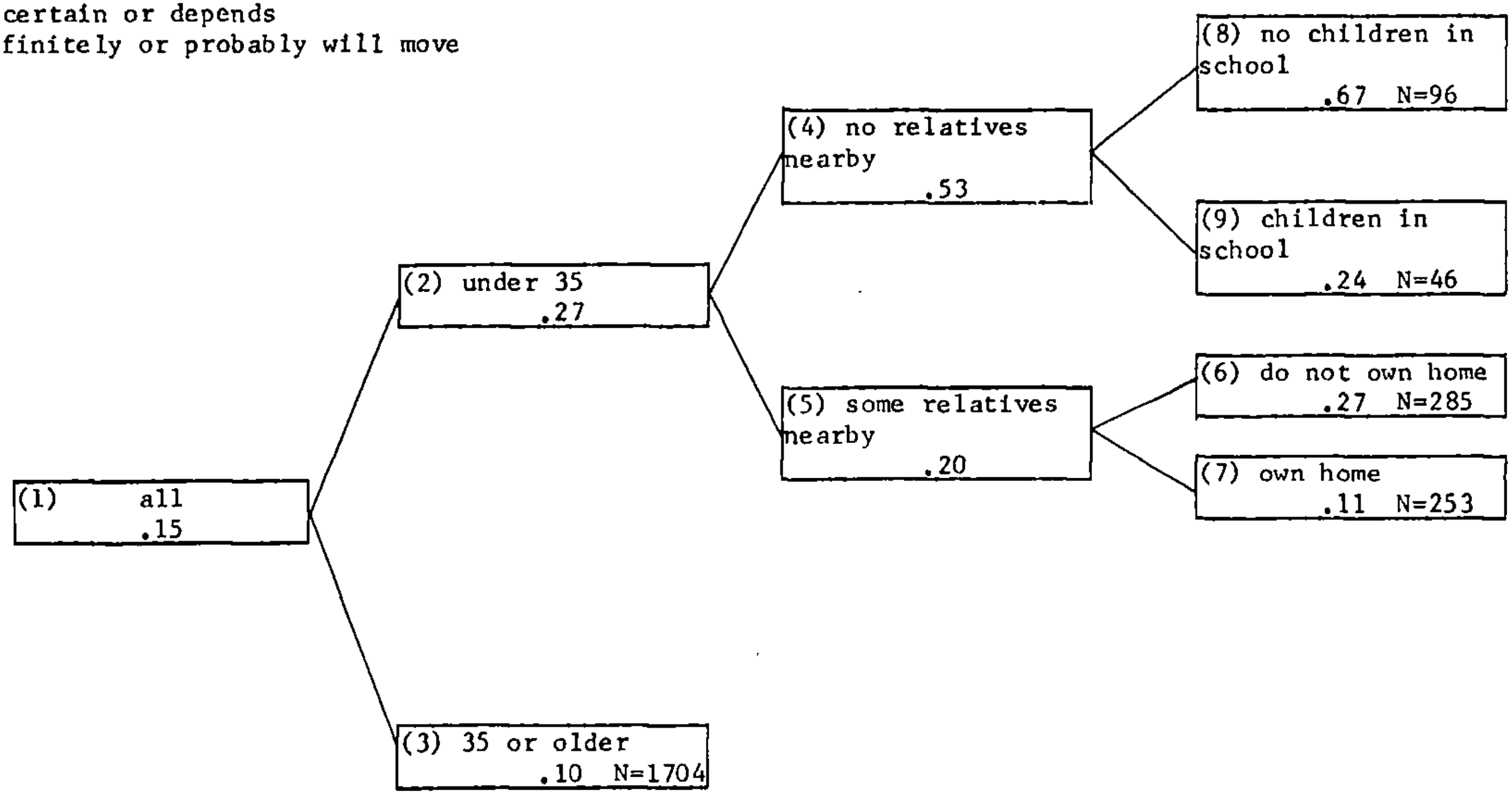


SOURCE: ISR Project 719
MTR 23

CHART 5

PLANS TO MOVE

0. No chance of moving
1. Uncertain or depends
2. Definitely or probably will move



SOURCE: ISR Project 719
MTR 33

Section 3.5 A Skewed Distribution-Nonfamily Contributions

Dollar contributions during 1959 reported by families as made to charity, church, and relatives not living in the household, had been analyzed by using an additive dummy-variable regression technique (27).

The prior analysis used some variables representing interactions between the original classifications--combinations of religious preference and church attendance, and combinations of race, age, education and farmer status entitled "earning potential."

The badly skewed nature of the dependent variable had been ignored in the original analysis, but showed up immediately in the AID results. Sixteen of the twenty-two final groups contained ten or fewer observations. Eliminating 33 cases of the original 2800 where contributions of $3,000 or more were reported, reduced the standard deviation of the dependent variable from $725 (mean was $315) to $419 (with mean of $254).

Table 6 gives the classifications used, which were purposely kept the same even in a second AID run which excluded the 33 extreme cases. Neither of the trees is given here because they are difficult to read. In addition to the problem of small groups split off, which remained even after eliminating the most extreme cases, the introduction of complex classifications such as "earning potential" into the AID analysis lead to combinations of combinations which were extremely difficult to describe and interpret. A revised program allowed us to constrain such factors as "number of children" against reordering of the scale.

Consequently, a third AID run was made, using the components of the complex classes separately: religion, church attendance, race, age, education, labor force status. The results are given in Table 7 and in Chart 6. There is a clear preponderance of income as an explanatory factor, but also a clear tendency for those over 45 years old to contribute more to others.

The problem of skewness still remains, as can be seen from the two remaining cases where a group of two or three is split off, reducing the error sum of squares by more than 1 per cent in each case.

An examination of the extremely large contributors revealed that they tended to have quite high incomes and either dependent parents or children living away from home (in college, just married) to whom they were making gifts. For very high income people such gifts are an important method of avoiding estate and inheritance taxes. The persistence of small groups would indicate that some transformation of the dependent variable into logs or percentages of income might be necessary. There are disadvantages to any of these transformations, however, when it comes to interpreting the results.

It is not the purpose here to provide a thorough analysis and interpretation of the results of each of these exploratory analyses. Two general questions are always to be asked:

1. At any stage, are there competing factors correlated with the one actually used in the split, and subsequently made unimportant?

In this case, an analysis of the between-sums-of-squares for the best split on each predictor at each stage indicates that whenever a second factor was almost as good as the one used, it tended to come into its own and be used later on in at least one of the branches. This, however, must necessarily be a property of the particular set of variables used in the analysis and depends on the orthogonality of the predictors.

2. Do the results suggest hypotheses which, for final testing, would require new information?

The importance of age in the analysis, with those over 45 persistently contributing more, raises questions whether this is the result of more assets, or more children, relatives and organizations making claims on older people, or whether it reflects a historical process of the passing away of private philanthropy--the younger generation being more willing to leave it to the government. No data are available on the existence of relatives who need aid. The attitudes toward government responsibility for the aged, and toward level of unemployment compensation benefits are not strikingly different between the young and old (24).

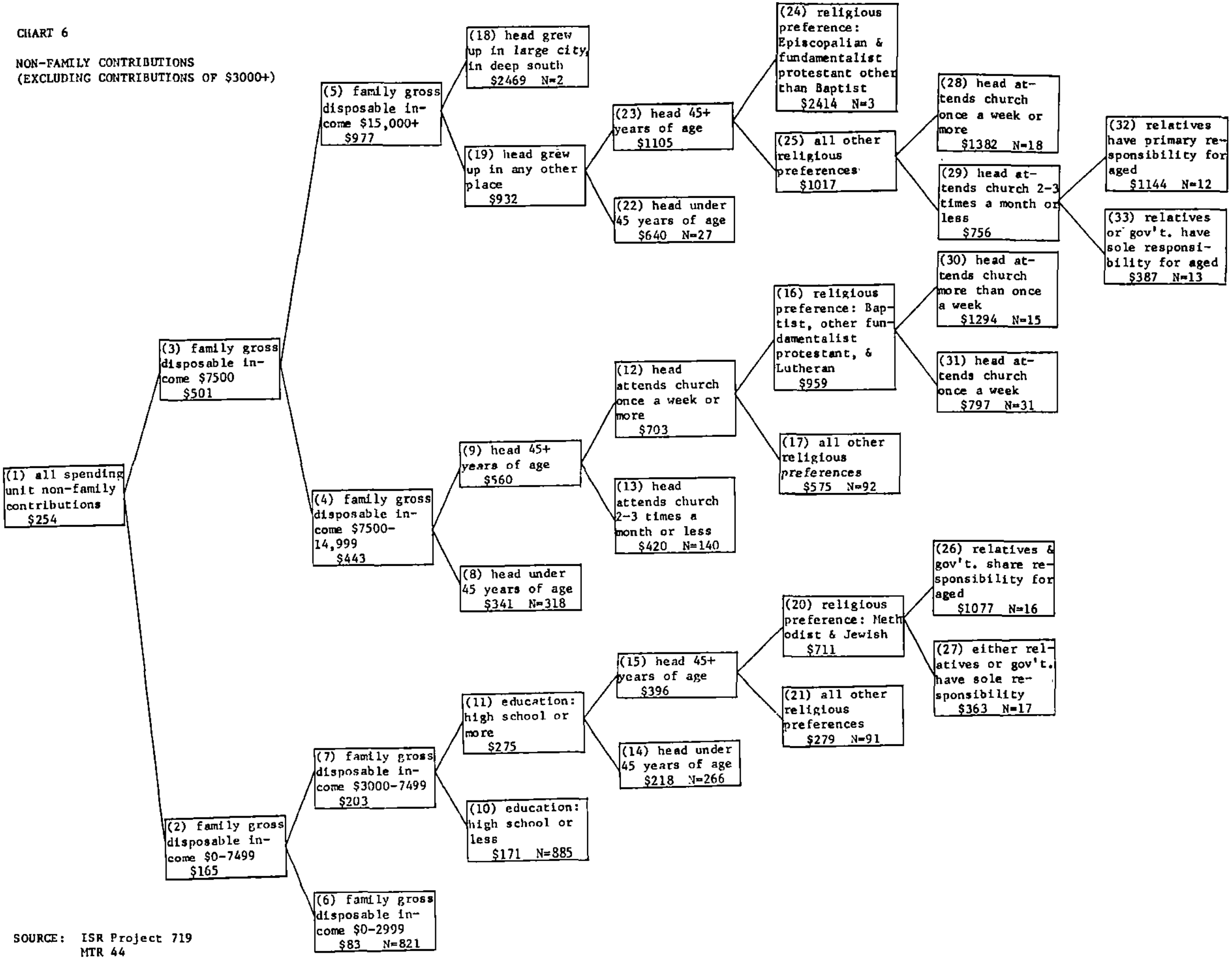
One of these attitudes was used directly in the analysis and comes in only at the end and with older people.

The output of the AID program gives the subgroup means in detail at each split so that one can observe whether anything was lost by maintaining the order of the age groups. Nothing was, since only one small age group would have switched to the other side. It is also possible to look at the competing factors at each stage to see which factors nearly succeeded. Near the ends of the trees there were some cases where geographic background and current marital status nearly "made it," but in both cases the N's were quite small.

The importance of church attendance is not surprising, though there are problems whether it is cause or effect, or a joint result of some more basic factor.

CHART 6

NON-FAMILY CONTRIBUTIONS
(EXCLUDING CONTRIBUTIONS OF $3000+)



SOURCE: ISR Project 719
MTR 44

Table 6

PROPORTION OF VARIATION IN NONFAMILY CONTRIBUTIONS EXPLAINED BY FAMILY CHARACTERISTICS

Characteristics of families	Type	All families	Families with contributions of $0-2999 only	Squared Beta coefficients from additive regression analysis*
Gross disposable income	F	.190	.187	.173
Earning potential of heads	F	.105	.028	.016
Religious preference and church attendance of heads	F	.053	.022	.012
Number of living children of heads	F	.031	.016	.004
Political preference	F	.063	.009	.004
Age of heads at birth of eldest living child	F	.016	.000	.002
Number of siblings of heads	F	.039	.006	.002
North-South-Urban-Rural background of heads	F	.006	.016	.001
Attitude of heads toward who should have primary responsibility for aged; government or relatives	F	.000	.006	.001
Sex of heads	F	.000	.000	.001
Family provides housing for nonnuclear family members in household	F	.031	.000	.000
Total proportion of variation explained		.534	.290	.22**
Mean contributions		$315	$254	
Standard deviation of contributions		$725	$419	
Number of observations		2800	2767	

*See (27) for a description of these coefficients.

**Beta coefficients do not add. This is an adjusted R^2.

Source: ISR Study 678, Deck 33

Table 7

PROPORTION OF VARIATION IN NONFARMILY CONTRIBUTIONS EXPLAINED BY FAMILY CHARACTERISTICS
(Contribution of $0-2999 only)

Characteristics of families	Type	Number of classes	Proportion of variation explained
Family gross disposable income	M	10	.179
Marital status of heads	F	6	.000
Labor force status of heads	F	7	.000
Age of heads	M	7	.030
Sex of heads	F	2	.000
Race of heads	F	4	.000
Education of heads	M	8	.007
Number of siblings of heads	M	5	.000
Number of living children of heads	M	5	.000
Attitude of heads toward government or relatives having responsibility for aged	F	7	.018
Religion of heads	F	10	.032
Church attendance of heads	M	7	.027
North-South-Urban-Rural background of heads	F	6	.011
Family provides housing for nonnuclear-family members in household	F	2	.000
Total proportion of variation explained		.304	
Mean contributions		$254	
Standard deviation of contributions		$419	
Number of observations		2767	

Section 3.6 Expected Family Size

Data from Friedman, Whelpton and Campbell (28) were used in an analysis of expected family size. The input consisted of responses from all wives married ten or more years, whose fecundity was not classified as indeterminate. Three analyses were run, as illustrated in Charts 7, 8 and 9.

The first analysis (see Chart 7) included twelve predictors (see Table 8). All were left free in mode, that is, the class orderings were not constrained. The analysis explained thirty-seven per cent of the variation in the dependent variable, with number of years worked by wife, husband's education, fecundity status, husband's occupational status, wife's education, and an interaction of religious preference and attendance accounting for over thirty per cent of the variation. The results are generally in conformity with those reported by Friedman, Whelpton and Campbell. However, this tree serves to illustrate several properties of the AID algorithm. This analysis contained variables with two classes (fecundity, wives' age at marriage) to ten (wives' work experience, husband's occupational status, and education of both husband and wife). The tree indicates that wives who have worked zero through three years have a mean of 3.5 on the dependent variable, and those who have worked four or more years have a mean of 2.2. But the interpretation of the extremely powerful effect of this variable is difficult. It taps variation associated with the work-enabling situation of sterility and/or children in school. It may well be the result of a decision to work rather than care for more children. This decision is a complex function of attitudes toward family size limitations, economic aspirations, attitudes toward the appropriate role of an adult woman, job opportunities, etc. Thus, it may be interpreted as an <u>effect</u> of family size, rather than a <u>link</u> in a causal chain explaining family size. Family size may be an enabling condition for working.

These issues arise because of the question which should be asked at each split. "Why should <u>this</u> variable be more highly correlated with the dependent variable than any <u>other</u> one in the analysis for this particular group?" The answer may be that this variable is very highly correlated with one or more other variables which have not been measured directly, and which are very close to the dependent variable in a causal chain, either as a cause or as an effect. Another answer is that the more classifications (in this case, ten) encompassed by a variable, the more likely it is for the algorithm to find a permutation of the class means that will produce a high between-groups sum of squares. However, constraining the order of the classifications would not, in this case, have caused another variable to be used at this stage.

The same type of problem may be seen later on in this tree in the behavior of the variables wives income, husbands occupational status, wives education and husbands education. Husbands occupational status is a derived measure based on occupation, salary, and education, for which a score between 0 and 99 is computed. The measure coded for use as a predictor consists of the ten deciles of that score distribution. On this basis, the splits in the tree do not make sense. When a relatively small group is partitioned on the basis of an unconstrained predictor with a large number of categories, the sampling variation of the class means will be large because of the small number of observations in each class. The probability of a fortuitous split is relatively high.

We are led to a conservative rule of thumb. Predictors which have a rank ordering to their classifications should be constrained to that ordering during the partition process, and unordered predictors should not have more than five or six classes. The exception to the rule of constraining rank ordered predictors is the case where the possibility of a U-shaped or inverted U relationship between that predictor and the dependent variable is suspected, in which case adjoining classes should be combined to form a maximum of five and the variable left unconstrained.

Charts 8 and 9 are identical runs except that all predictors are unconstrained in the first run (8), and both education variables, husband and wife have a constrained status in the second run. Also, in Chart 9 no group where $N_i < 25$ was permitted to split. Both runs used six predictors, a subset of the predictors listed in Table 8. They were husband's education, wife's education, size of city lived in, attitudes toward family limitation, fecundity status and the interaction variable religion and attendance.

In Chart 8, the tree produced an R^2 of .259 as compared with .216 in Chart 9. Here, we have a clear effect of the constraints on the ordering on the ranked variables having a large number of categories. In Chart 8, one suspects that the later splits on education tend to be susceptible to influence by sampling variation. The constraints are not present. There are more final groups in this tree. Variation is being attributed to education which probably does not belong there. The fact that several splits appeared in which a very small group was separated from a large one leads one to suspect a skewed or very spread out rectangular distribution. These extreme observations should undoubtedly be subjected to a careful deviant case analysis to see if they have something in common that is not used as a predictor in the tree.

Other somewhat unexpected findings appear, and are associated with the interaction variable religion x attendance. The expected relation between Catholicism, church attendance and expected family size is not found. Regular attenders who are Catholic show up as having <u>fewer</u> expected number of children than those who only report attending often. These may represent measurement errors, sampling errors, or a genuine finding.

There appears to be evidence in the tree presented in Chart 8 that the variable place of residence is somewhat differently related to expected family size in the three subgroups in which it was used as a criterion for splitting. Table 9 illustrates the differential behavior of the variable. In the total sample of wives married ten or more years, the clearest difference is between the rural farm wives vs. the remainder. This is also characteristic of group A, the sterile wives, and group B, the fecund wives with 9 or more grades of education who do not disapprove of family limitation. Group B is most like the total sample. The effect of sterility is clearly shown by an examination of the lowered means in group A, compared to the total group. Its effect is more pronounced with increasing urbanization. But in both group A and group B, the maximum binary split was the rural farm vs. all others.

A somewhat different pattern appears in group C, fecund wives with only one through eight grades of education. Here it is the twelve largest central city and suburban people who are quite different from the remainder. The somewhat surprising change in the rank-ordering of the means in this group, between the small towns and cities over 50,000 is consistant with results found by Friedman, et al., and may be explained by the fact that the place-code for metropolitan areas other than the twelve largest include the entire county in which the central city of over 50,000 is located, and probably contain uneducated persons who should more properly be classified as rural farm and rural nonfarm.

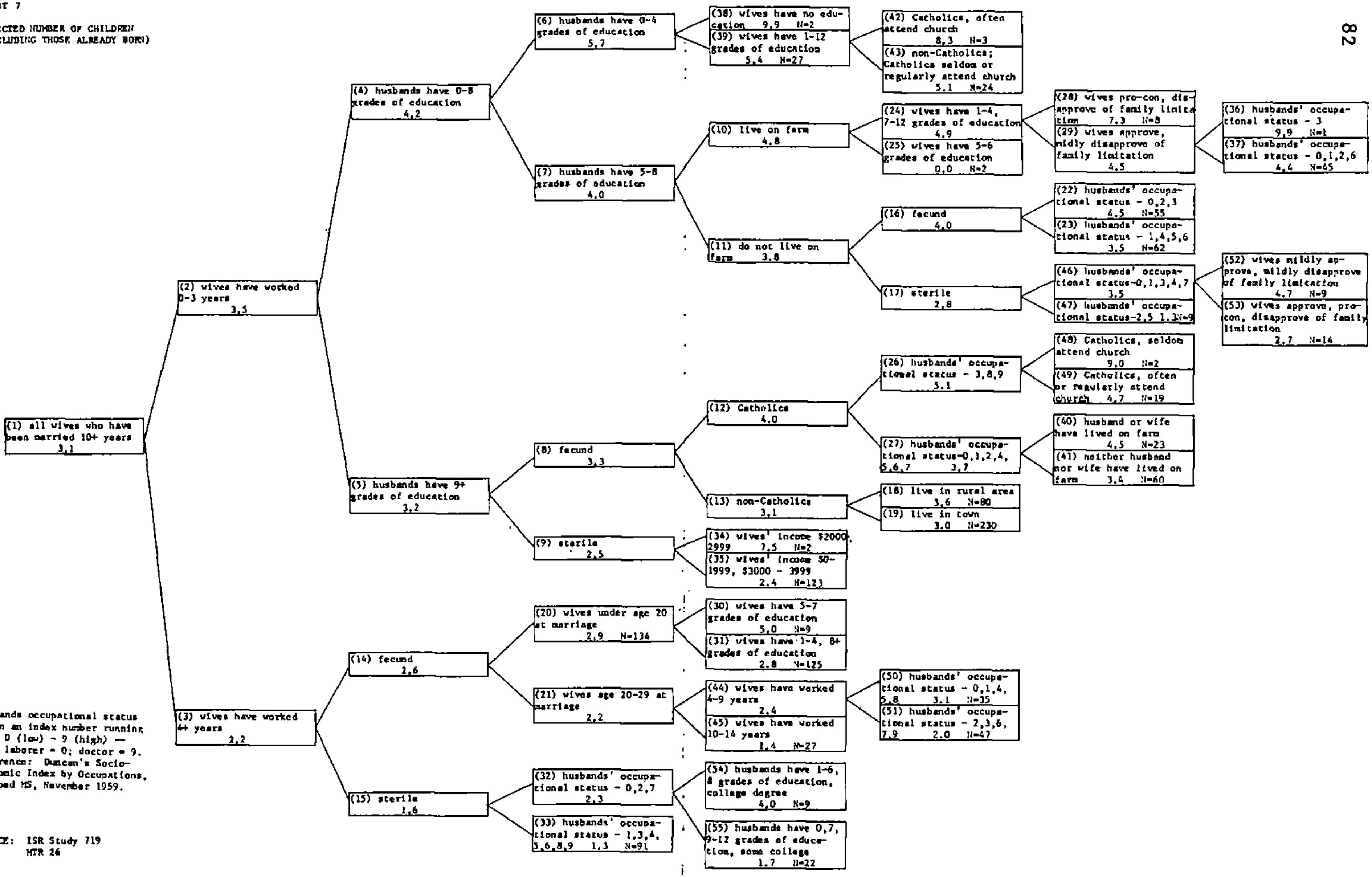
The implication of this finding for the further use of the algorithm is that in the initial stages of analysis, it may be desirable to leave all predictors unconstrained, and to use the program as a device for locating conceptual problems. It is quite likely that classes such as the 50,000+ code for place of residence which, when used as an index of urban-rural residence do not conceptualize all of the population properly. In this case, it is probably true that those living outside the city of 50,000+, but inside the county in which it is contained, are really living in a rural-farm or rural-nonfarm community situation. It is also quite likely that there is a fairly heavy concentration of low-education people in these areas outside these small central cities. Thus, it is implied that the urban-rural variable, as coded, tends to place better educated persons more accurately along the rural-urban dimension than persons with low education.

One possible use for the procedure is to scan the data for variables which do not "behave" as expected. When unexpected findings appear, one possible interpretation involves the relation between the indicator, or variable used as a predictor and the underlying concept which it operationalizes. There may be some classes of the sample for which the variable does not correspond to the concept. One must, of course, decide whether the split represents covariation, conceptualization, coding errors, sampling variability or a genuine finding.

The purpose of this discussion has been to focus on the need for a careful examination of the relationship between the underlying concept and the indicator (predictor) as it behaves in the analysis.

CHART 7

EXPECTED NUMBER OF CHILDREN
(INCLUDING THOSE ALREADY BORN)

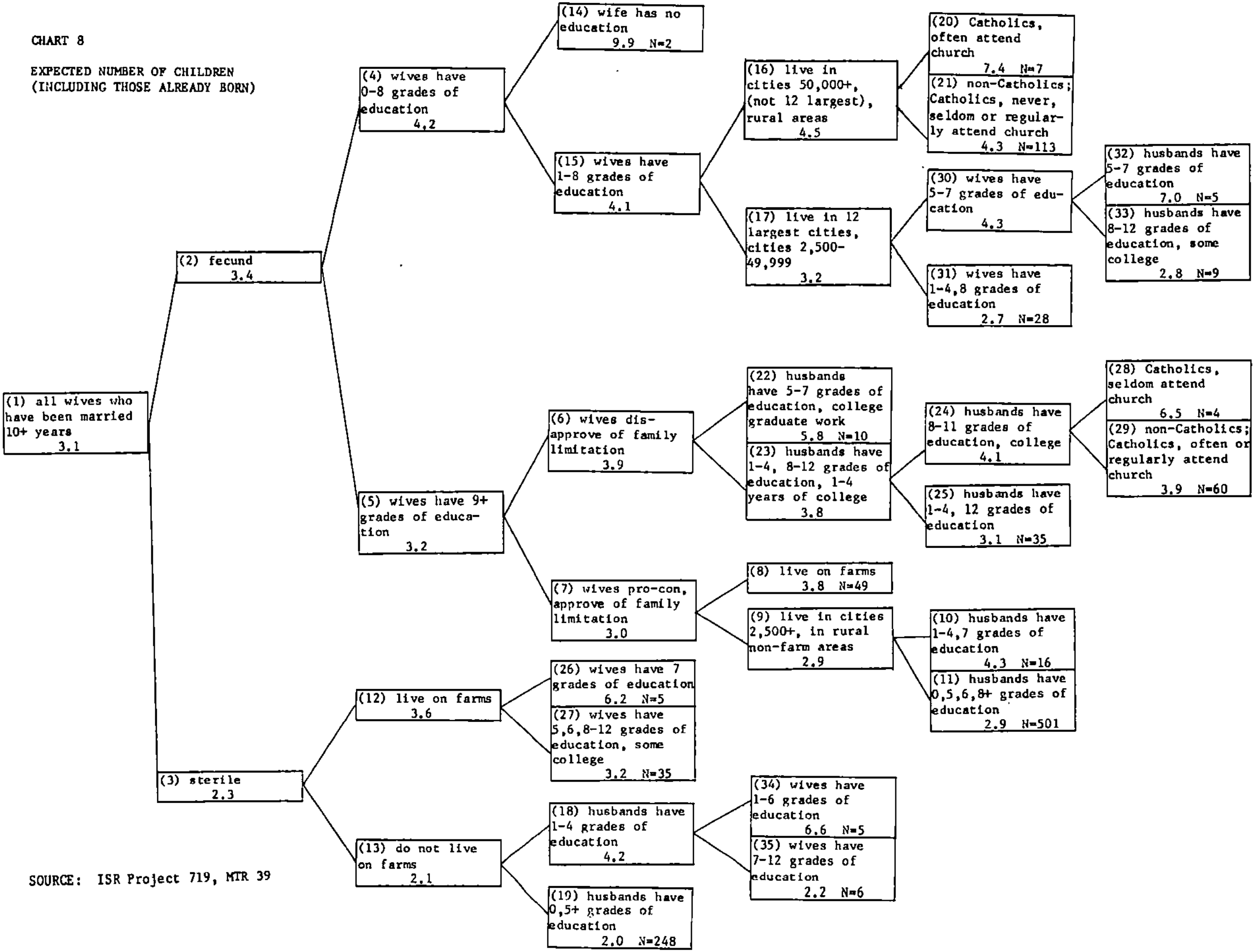


* husbands occupational status given an index number running from 0 (low) - 9 (high) -- e.g. laborer = 0; doctor = 9. Reference: Duncan's Socio-Economic Index by Occupations, dittoed MS, November 1959.

SOURCE: ISR Study 719
MTR 26

CHART 8

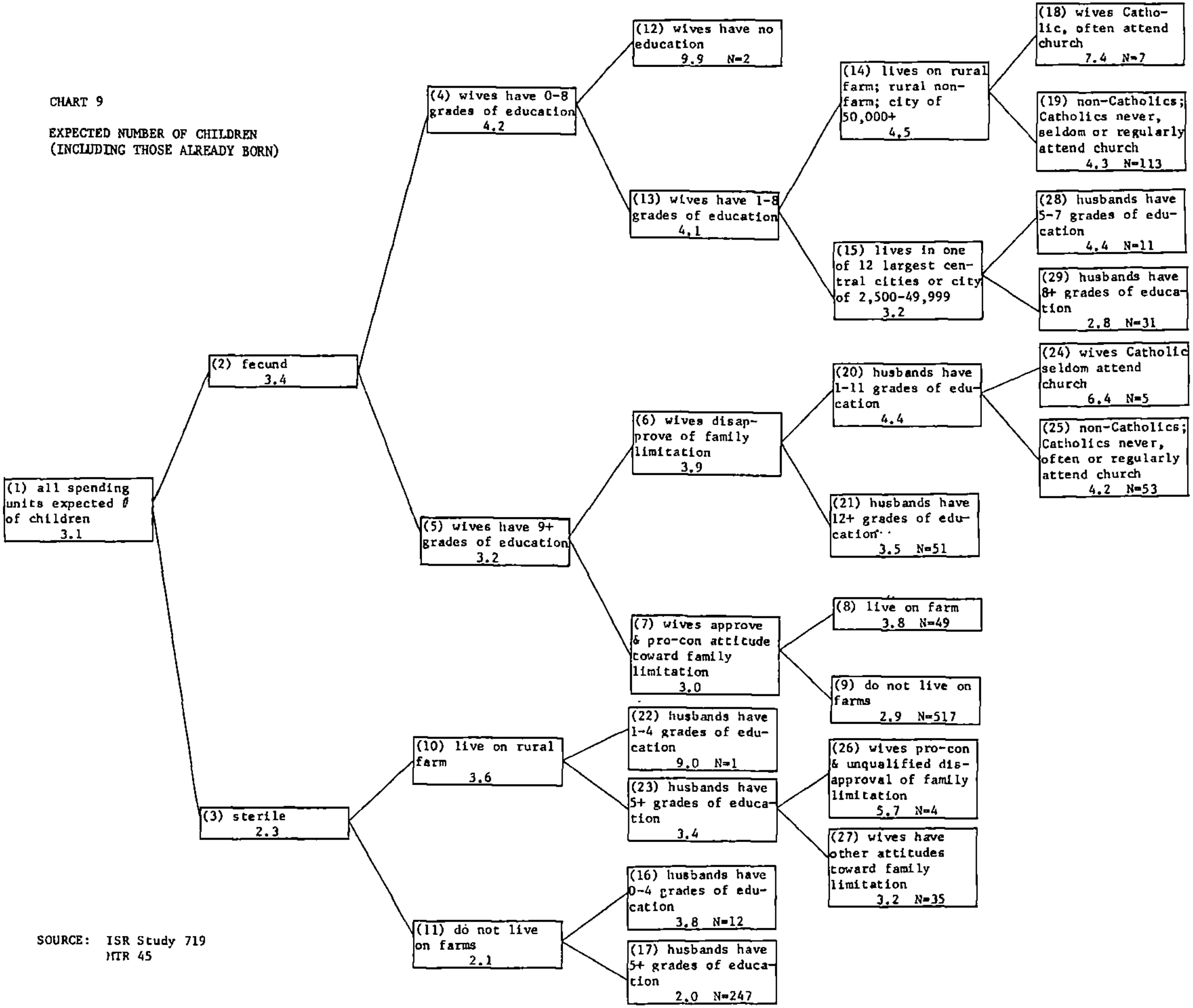
EXPECTED NUMBER OF CHILDREN
(INCLUDING THOSE ALREADY BORN)



SOURCE: ISR Project 719, MTR 39

CHART 9

EXPECTED NUMBER OF CHILDREN
(INCLUDING THOSE ALREADY BORN)



SOURCE: ISR Study 719
MTR 45

Table 8

RELATIVE POWER OF VARIABLES PREDICTING EXPECTED NUMBER OF CHILDREN

Predictor	AID reduction in TSS(I)/TSS(T)	Number of classes
Number of years wives have worked	.097	10
Husbands education	.072	10
Wives fecundity status	.044	2
Husbands occupational status	.041	10
Education of wives	.030	10
Religion x attendance of wives	.027	5
Attitude of wives toward family limitation	.018	5
Farm background of husbands and wives	.015	4
Wives income	.012	8
Age of wives at marriage	.008	2
Present place of residence (urban-rural)	.007	5
Discrepancy in income between husbands and wives	.000	3
R^2	.371	

Mean	3.09
σ	1.92
N	1138

Source: ISR Project 719, MTR 26

Table 9

MEAN EXPECTED FAMILY SIZE FOR THREE GROUPS, BY SIZE OF PLACE OF RESIDENCE

	Total			A			B			C		
	N	$\bar{\bar{Y}}$	σ	N	$\bar{\bar{Y}}$	σ	N	$\bar{\bar{Y}}$	σ	N	$\bar{\bar{Y}}$	σ
Rural farm	135	4.0	2.2	40	3.6	2.3	49	3.8	1.9	36	4.9	2.4
Rural nonfarm	189	3.3	2.0	54	2.7	2.0	83	3.2	1.6	37	4.4	2.2
Places of 2500-49,999	180	2.9	1.7	55	2.2	1.6	92	3.0	1.4	20	3.7	2.1
Cities 50,000+ and suburban rings which are not 12 largest cities	396	2.8	1.8	88	1.9	1.8	214	2.7	1.4	47	4.2	2.2
12 largest central cities and suburban rings	238	2.9	1.7	62	1.9	1.3	128	3.0	1.5	22	2.8	2.0
Total--all wives married 10 years or more	1138	3.1	1.9	299	2.3	1.9	566	3.0	1.5	162	4.1	2.3

Group A Sterile wives married 10 or more years

Group B Fecund wives married 10 or more years with 9 or more grades of education who do not disapprove of family limitation

Group C Fecund wives married 10 or more years with 1 through 8 years of education

[ ] Groups placed together in the partition process

Source: ISR Project 719, MTR 39

Section 3.7 Average Number of Grades of School Completed by Children in the Spending Unit

A major vehicle for transmission of economic status from one generation to the next is formal education. A previous multivariate analysis (24) using the dummy-variable regression model, employed the explanatory factors listed in Table 10.

Table 10 gives the beta coefficients from a multiple classification analysis (22), squared for comparability of dimension, and the proportionate reduction in error sum of squares attributable to the same predictors when used in the AID analysis. It is interesting that every one of the predictors is used to make at least one split in Chart 10. This suggests that there really are many forces at work which are not so highly correlated with one another that the division of the sample on one makes the other unnecessary.

Again, however, there are problems when variables which themselves represent interactions are used, since the resulting splits involve combinations of combinations, frequently difficult to interpret. There are also some relatively small groups split off. However, most of the splits go in the expected directions.

In the right center of the chart is an interesting sequence in which first, those with a high index of achievement motivation are split off, and among the rest, those who go to church frequently (or are non-Christian). Are transmitted achievement motivation and a religiously oriented sense of responsibility alternative forces inducing people to provide more education for their children?

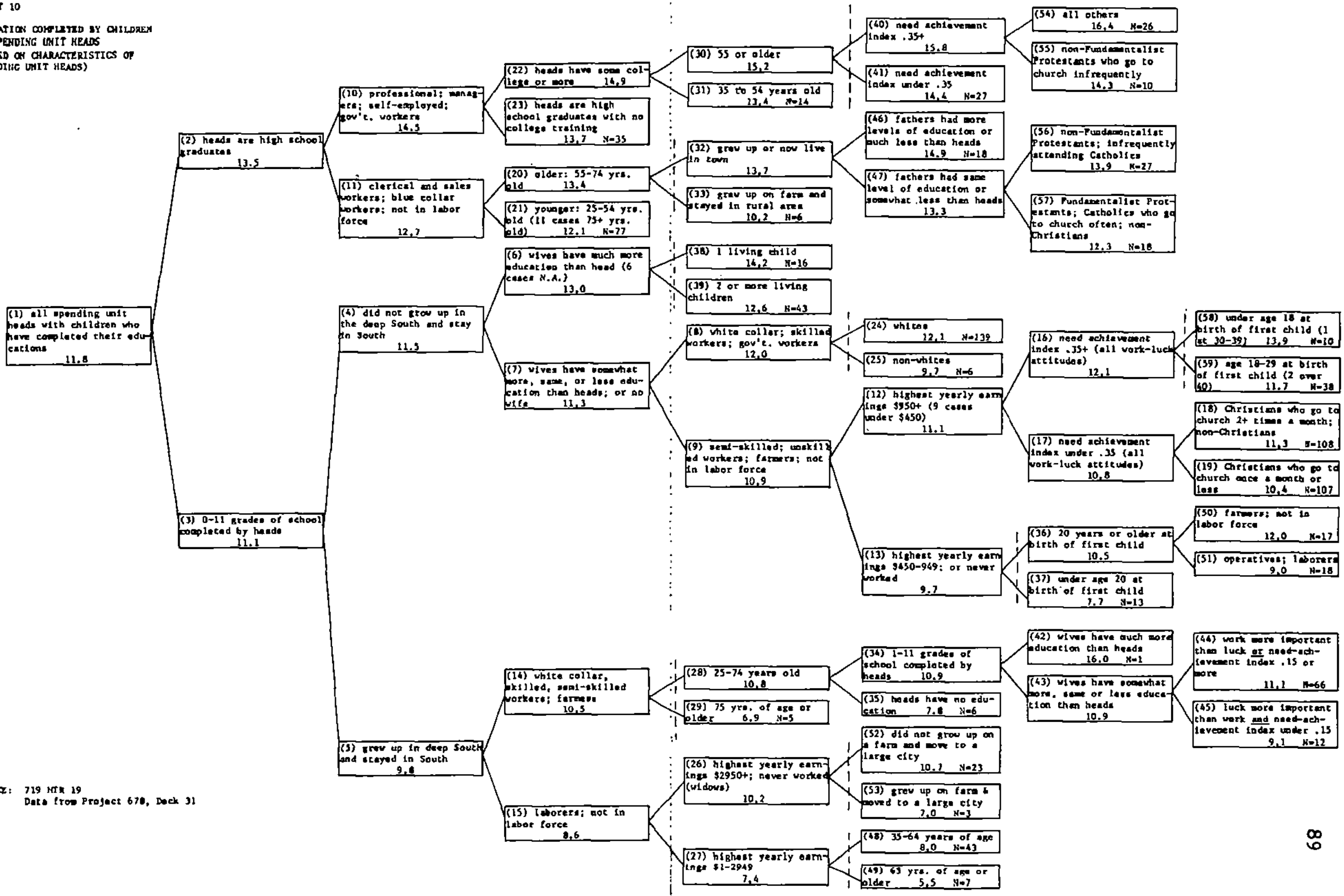
In a number of places one may wonder whether the variable used is really a proxy for one of the others, i.e., "grew up in the deep South and stayed there," meaning "mostly nonwhites." The program as now set up, provides a distribution on each predictor at each split so that one can tell to what extent a competing variable came close to being used.

There is a group where the father had some college training and was a professional, manager, self-employed, or government employee where children of fathers 55 and older had clearly more education than those under this age, and where an examination of the group before splitting indicated a continuous trend across five age groups. One implication is that while in earlier generations the children of college educated fathers were almost certain to go to college, the strength of this effect has been getting smaller. (If colleges rely more on merit and grades and admit fewer of the "gentlemen" school, this finding might be real.)

Table 11 describes the final groups resulting from the AID analysis listed in decreasing order of the mean education of children in that group. The distribution of educational levels for spending units with living children who have completed their education is presented in Table 12.

CHART 10

EDUCATION COMPLETED BY CHILDREN
OF SPENDING UNIT HEADS
(BASED ON CHARACTERISTICS OF
SPENDING UNIT HEADS)



SOURCE: 719 HTR 19
Data from Project 678, Deck 31

Table 10

AID AND MULTIPLE CLASSIFICATION ANALYSIS OF AVERAGE COMPLETED EDUCATION OF CHILDREN

Characteristics of spending unit heads	AID--Reduction in TSS(I)/TSS(T)	MCA Analysis--Beta coefficients (squared)	Number of classes
Education of head of unit	.192	.140	8
Difference in education between head and wife	.034	.035	7
Occupation of head	.090	.023	10
Number of living children	.005	.015	4
Whether grew up in the deep South, and whether now lives in the South	.052	.013	6
Whether hard work is seen as more important than luck and help from friends, and level of need-achievement score (a complex measure of motivation)	.021	.011	7
Highest income ever earned by the head of the unit	.028	.096	10
Religious preference and church attendance	.019	.093	7
Age of head at birth of eldest living child	.017	.083	7
Difference in education between head and his father	.006	.076	4
Race	.005	.048	2
Whether grew up on farm, and whether now lives in rural area	.019	.020	7
Age of head of spending unit	.034	.014	4

$R^2 = .522$ $R_A^2 = .41$

$\overline{Y} = 11.8$

$\sigma = 2.6$

$N = 939$

Source: ISR Study 719, MTR 19

Table 11

COMPLETED EDUCATION OF CHILDREN FINAL (TRUNCATED) GROUPS IN RANK ORDER BY THEIR AVERAGES

Group number	Number of cases	Average years of education	Characteristics of parents
(30)	63	15.2	Father had some college and is a professional, or manager, or government employee or is self-employed and is aged 55 or older.
(23)	35	13.7	Father is professional, manager, or government employee who finished high school, but had no college.
(31)	14	13.4	Father had some college, is a professional, or manager, or government employee or is self-employed and is 35-54 years of age.
(20)	69	13.4	Father finished high school or has additional education and is blue collar worker or clerical and is 55-74 years of age.
(6)	59	13.0	Father did not finish high school, did not grow up and remain in the South, mother had two or more levels of education than the father.
(21)	77	12.1	Father finished high school or has additional education, is blue collar worker or clerical and is aged 25-54 or over 74 years of age.
(16)	48	12.1	Father did not finish high school, did not grow up and remain in the South, was unskilled worker or farmer, did not have two or less levels of education than mother, had a highest income that was not in the lowest category, and scored high on achievement motivation.

Table 11--(CONTINUED)

Group number	Number of cases	Average years of education	Characteristics of parents
(8)	145	12.0	Father did not finish high school, did not grow up and remain in the South, did not have two or less levels of education than mother, was a white collar, skilled worker, or government employee.
(18)	108	11.3	Father did not finish high school, did not grow up and remain in the South, did not have two or less levels of education than mother, had not always had low income, was low on achievement motivation, was a Christian and attended church regularly or was non-Christian.
(14)	90	10.5	Father did not finish high school, grew up in the South and stayed there, was not a laborer.
(19)	107	10.4	Father did not finish high school, did not grow up and remain in the South, did not have two or less levels of education than mother, was unskilled worker or farmer, had not always had low income, was low on achievement motivation, was a Christian who attended church infrequently.
(26)	26	10.2	Father did not finish high school, grew up in the South and stayed there, was a laborer, had highest income over $3000.
(13)	48	9.7	Father did not finish high school, did not grow up and remain in the South, did not have two or less levels of education than mother, was unskilled worker or farmer, had a very low highest previous income.
(27)	50	7.4	Father did not finish high school, grew up in the South and stayed there, was a laborer, had never earned more than $3000.

Table 12

COMPLETED EDUCATION OF CHILDREN
OF SPENDING UNITS EXISTING IN EARLY 1960**
(FOR THOSE WHO HAVE LIVING CHILDREN)

Average completed education of children*	Per cent of units
Six grades or less	2.9
Seven or eight grades	9.9
Nine through eleven grades	25.8
Twelve grades (high school)	33.6
Thirteen through fifteen grades (some college)	17.3
Sixteen or more grades (college graduates)	10.5
Total	100.0

*In most families the children had similar education, and the averages tend to cluster around integers, hence all averages have been rounded downward (twelve grades includes 12.0 through 12.9).

**Source ISR Project 678 (D: 31, MTR 54). This is a national probability sample of 2999 spending units, 34 per cent of which had at least one child who had completed his education. These, of course, tend to be spending units in which the head of the unit is older than the average.

Section 3.8 A Somewhat Skewed Variable-- Spending Unit Disposable Income

An alternative to separate analyses of components of income, such as labor force employment of each member, hours of work, and hourly earnings, is to analyze the resulting combination of incomes, even though the causes may work through one or more of the components. It is important sometimes to see just what are the most important forces affecting an overall result. Data from 2033 spending units interviewed in early 1964 in the 1963 Survey of Consumer Finances (29) were used. The following explanatory variables were employed:

	Number of subgroups
Stage in the family life cycle	10
Education of the head of the unit	6
Age of head	6
Size of place of residence	6
Race	4
Income change over previous year	4
Region of the country	4

The twenty final groups accounted for half the total variance. The standard cutoff criteria which allowed any split which reduced the error by 1/2 per cent allowed one final (omitted) split which formed groups of one and four cases respectively. It is quite clear from Chart 11 that the income of spending units depends mostly upon whether they are married, educated, middle aged, live outside the South, and live in metropolitan areas. The first split points to those "married and not retired" which means at least one earner and in many cases two. The other group are handicapped by being extremely young or old, having children but no spouse, or (and here the causation may go the other way) by having no family responsibilities.

We may summarize the next split on education by saying that the group with advantages is depressed only by very low education, but the disadvantaged group as to family situation is redeemed only if the head is a college graduate.

Following up the top set of branches, we note that a combination of advantages cumulate into substantial incomes. The higher income among one group of college drop-outs than among college graduates may be explained by their age (45-54), which means that they dropped out during the great depression. This may be a chance fluctuation, however, since a reverse effect is apparent among the same cohort living in small towns or rural areas, as well as among other age groups. This problem could be pursued further by a deviant case analysis with the object of determining what factor(s) are common to members of each of these two apparently contradictory groups.

As one way of assessing the stability of the resulting subgroupings, an analysis was made of spending unit disposable income for three separate Surveys of Consumer Finances covering incomes for the years 1952, 1957 and 1962. In addition, the 1957 subgroups were formed with the 1962 data to see how well they could explain data from which they had not been derived.

In different years, there was a good deal of agreement as to which predictors accounted for most of the reduction in the unexplained sum of squares, except that age, education, and stage in the life cycle increased greatly in explanatory power over time (but the last resulted from a more detailed coding of life cycle). It turned out (30) that there was a real change toward a greater earning payoff from education that took place during the period (see Table 13).

The order in which the branching took place varied from year to year. The reason is probably that there are several alternative ways to achieve roughly the same subgroups--one can separate the college graduates, then the middle aged among the college graduates, or start by selecting the middle aged, then separate the college graduates. Sampling variability may well be influencing which of two almost equally good predictors will be used.

This means that the proper focus in investigating sampling stability should be on the composition of the final groups, the interpretation

Table 13

AID ANALYSIS OF SPENDING UNIT DISPOSABLE INCOME --1952, 1957, 1962

Predictors	Reduction in TSS(I)/TSS(T)			Gross Beta Coefficients[2]		
	1952	1957	1962	1952	1957	1962
Place of residence	.034	.029	.042	.042	.033	.032
Age of head	.029	.034	.059	.064	.081	.124
Education of head	.124	.114	.171	.127	.126	.179
Race	.005	.000	.008	.030	.033	.034
Region	.021	.000	.016	.016	.003	.003
Life cycle	.095	.128	.201	.107	.135	.197
Income change	.021	.010	.007	.018	.028	.036
R^2	.329	.315	.504			

Source: ISR Study 719, MTR's 28-30

of the combinations of factors (pedigree) they represent, and on the explanatory power of the predictors at different stages in the tree rather than on the paths. It also means that even the total explanatory power assigned to various factors is stable only in a rough sense.

One may also compare the total explanatory power of the 1957-derived subgroups for 1962 data. The proportions of total sum of squares accounted for are presented in Table 14.

Table 14

AID ANALYSIS OF SPENDING UNIT DISPOSABLE INCOME, 1957, 1962

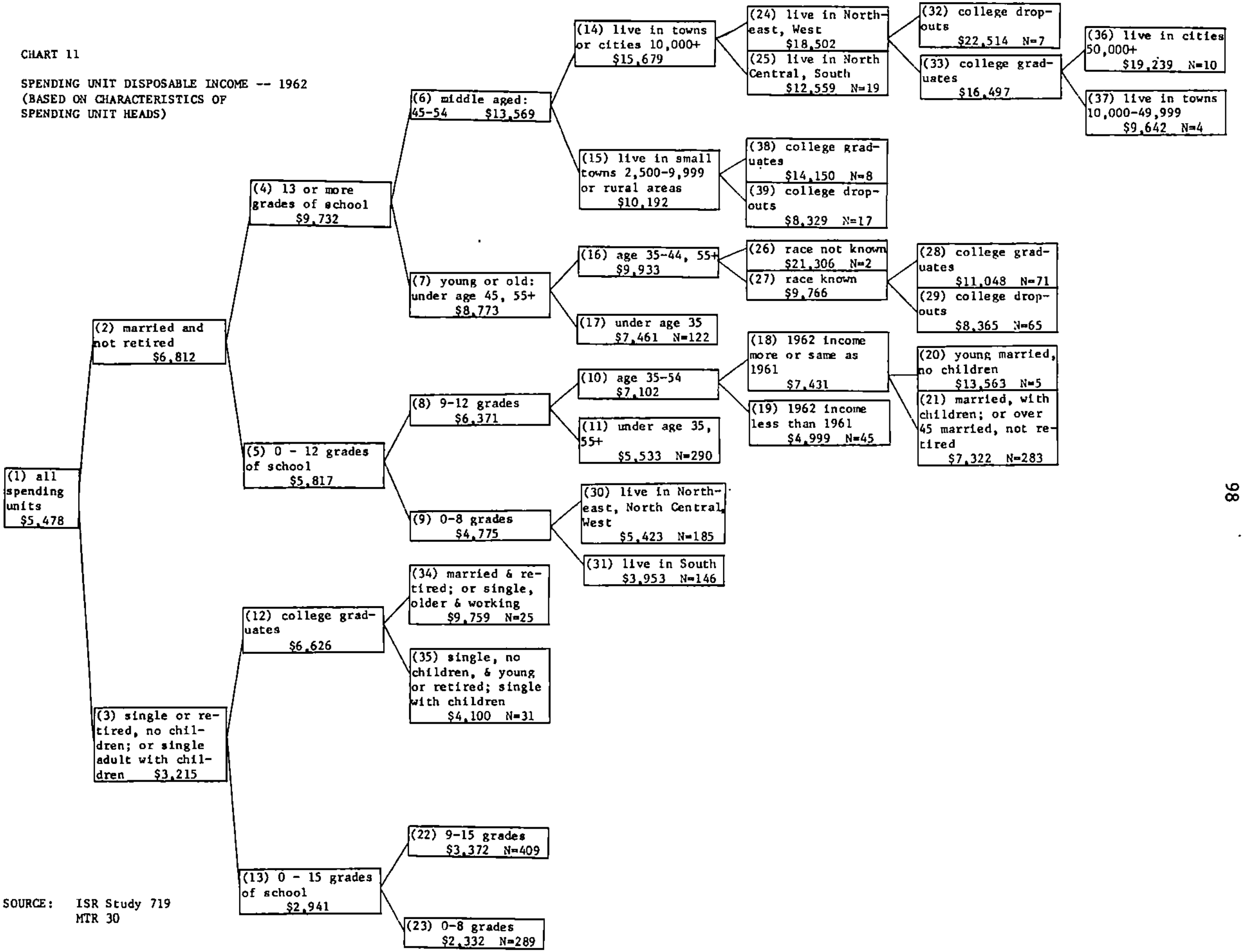
	R^2
1957 tree, 1957 data	.315
1957 tree, 1962 data	.366
1962 tree, 1962 data	.504

The increase in explanatory power of some of the factors over time makes it necessary to qualify any conclusions, but it is clear that the 1957 tree is not so good in 1962 as one based on the 1962 data, yet neither is it so inferior that one would regard it as an unstable, fortuitous breakdown of no use for prediction.

Another experiment involved split-half samples carefully designed to take account of the original stratification. Three different split halves were run on hours worked and three on hourly earnings. Again, while the way in which they were developed differed; the final groups were reasonably similar, and the ranking of factors by importance reasonably comparable. The proportions of unexplained sum of squares were much higher for the split halves, because the cut-off rules are less stringent with smaller samples. In other words, explaining 1/2 per cent of the total sum of squares of a smaller sample, using the same possible subclasses, leads to more subdividing and hence explains more of the variance.

CHART 11

SPENDING UNIT DISPOSABLE INCOME -- 1962
(BASED ON CHARACTERISTICS OF
SPENDING UNIT HEADS)



SOURCE: ISR Study 719
MTR 30

Section 3.9 Two Year Saving as Per Cent of Income

So far the initial analyses have been by multiple regression. The AID analysis sought to discover new things about the data not revealed by the regression. As an example of a more appropriate process, we turn now to a case where the AID analysis was used to determine which new (interaction) variables should be created and used in a regression analysis. The dependent variable was two-year saving as a percentage of two-year income from a panel study (34).

Earlier analyses had been run on the first version of the program with saving rate, discretionary saving rate, and an index based on an ordered series of saving rate classes as dependent variables, but the large number of classifications with 8, 9, or 10 subclasses combined with the relatively small sample provided many fortuitous combinations. The tree presented here (see Chart 12) made use of the option to maintain the order of subclasses for nine of the twenty-one predictors. It still tends to use predictors with too many subclasses, and combines clearly exogenous factors with some which might be results as well as causes. The variables used are listed in Table 15.

Sixteen of the twenty-one factors were used to form twenty-seven groups that accounted for 32 per cent of the total sum of squares. The AID analysis split first on home ownership (treating homes worth less than $2500 as not owning), then split both branches on whether the head was a self-employed businessman or farmer. Some other groups were split off from each of the nonentrepreneurial branches, notably low saving groups who had spent a lot on consumer investment items (cars, durables, additions and repairs), but this could be regarded as partly circular, i.e., as a decision to buy durables rather than save. The most important subsequent split was one which used initial assets, but split home owners and nonowners at a different level and revealed that owners with high initial assets saved more than other owners, while nonowners with some initial assets saved less than other nonowners.

This threw light on an ancient discussion about the effects of assets on spending and saving. Some economists had argued that assets facilitated spending, burning a hole in the man's pocket. Others said that those with assets were motivated to save and would persist in this behavior. Our analysis seemed to say that the best way to separate those with a persistent tendency to save from those with a high marginal propensity to consume, was to separate home owners from others.

The rest of the tree is complex and shows problems that arise when complex variables created ad hoc are introduced instead of built by the analysis from their components. Some of the later splits involve very small numbers of cases and have been recombined.

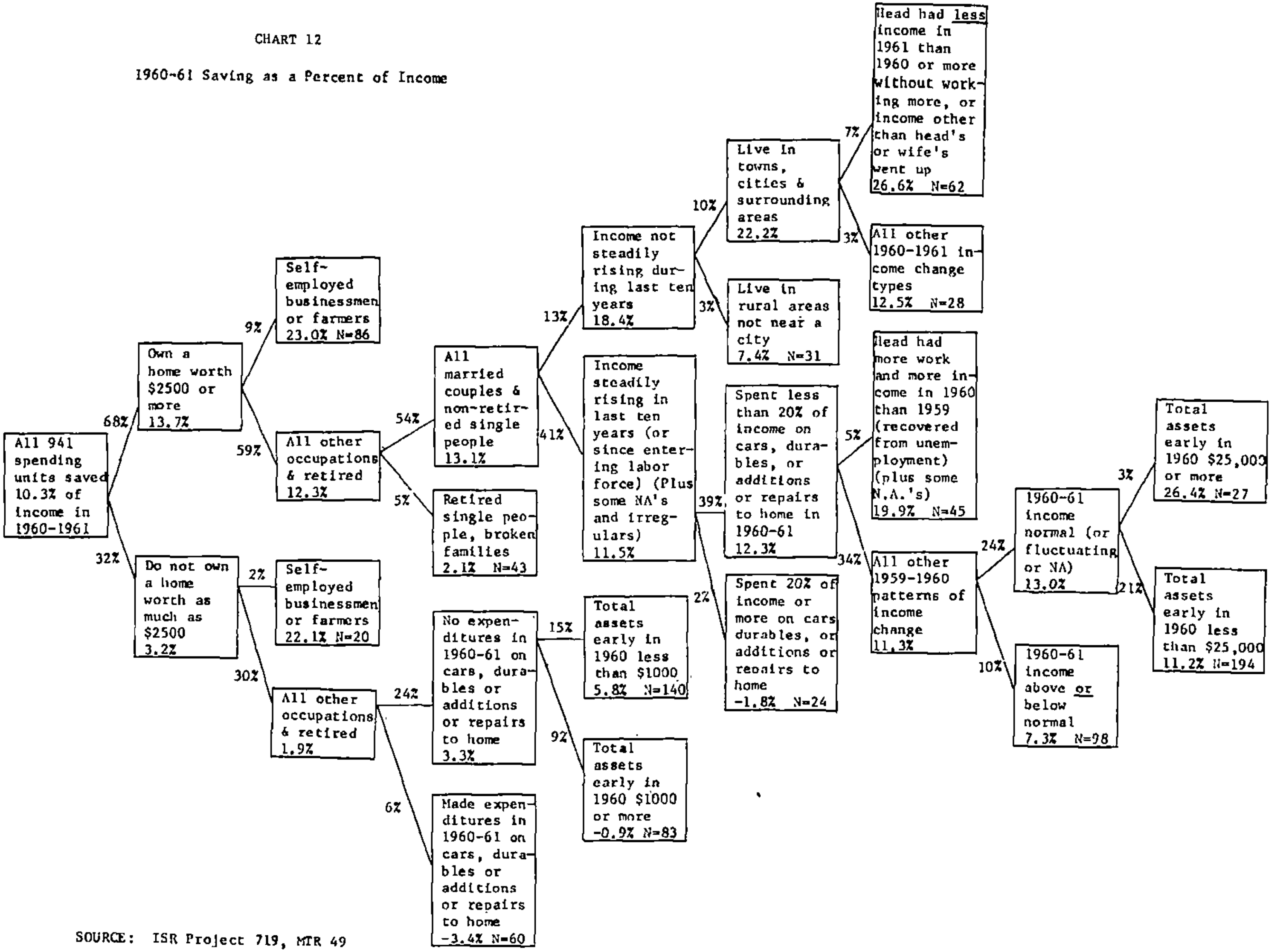
A neater format for presenting data, and a tightening up of this notion, required developing a set of dummy variables and putting them into a multiple regression, to assure that these relations could hold their own with other variables against a charge of spurious correlation. Several others of the sets of subgroups in the regression were developed deductively, others were unidimensional, and some had a long mixed history of development (stage in the family life cycle). The factors included and their partial beta coefficients are presented in Table 16.

To have put in "whether a self-employed businessman or farmer" as a separate dummy variable would have been to assume that home-ownership and assets affected the saving of these people too. A glance at the AID tree will reveal that this is not the case.

Needless to say, no significance tests should be applied to variables derived from a second analysis of the same set of data, and there is even a question about those derived by analysis of similar sets. On the other hand the five subgroups have reasonable and meaningful differences. They also serve the purpose of controlling on some factors (removing unwanted "noise") in a test of other factors in the regression. The unadjusted saving ratios, and the ratios adjusted by regression are given in Table 17 below. (Regression adjustment means adding the constant term to the dummy variable regression coefficient, the result being what the saving ratio of that group would be if it were like the whole population in its distribution on all the other variables.)

CHART 12

1960-61 Saving as a Percent of Income



SOURCE: ISR Project 719, MTR 49

Percent of all Spending Units is Given on Lines

Table 15

VARIABLES USED TO PREDICT TWO-YEAR SAVING AS A PER CENT OF INCOME

	Subclass order free or monotonic	Number of subclasses
Stage in family life cycle	F	10
Number of people in the spending unit	M	9
Occupation	F	5
Age of head	M	7
How long lived in this residence	M	8
Bracket value of house	M	10
Home ownership status	F	6
Education of head	M	6
Anticipated course of income over next ten years	F	6
Course of income over past ten years	F	9
Level of optimism in early 1961	F	3
Level of optimism in early 1960	F	3
Two-year expenditures on cars, durables and additions and repairs as per cent of two-year income (bracket)	M	10
Two-year income (bracket)	M	11
Size of place (city size)	M	6
Expected income change in 1962	F	4
Income change from 1958 to 1959 (memory)	F	4
Sources of income change 1958-1959	F	10
Sources of income change 1959-1960	F	10
Pattern of past and expected income change	F	5
Total assets in early 1960 (bracket)	M	7

Table 16

RELATIVE IMPORTANCE OF 14 SETS OF DUMMY VARIABLES IN A MULTIPLE REGRESSION

(N = 1001)

Characteristic	Relative importance partial β^2	Number of subgroups
Occupation-house-value-assets	.060	5
Stage in family life cycle	.022	9
Two year income	.021	10
Pattern of past and expected income change	.015	5
Age of head of unit	.011	6
Sources of income change 1960-1961	.010	10
Sources of income change 1959-60	.010	10
Size of place of residence (city size)	.009	6
Changes in optimism	.009	6
Anticipated income change 1961-1962	.006	4
Years lived at present address	.005	6
Educational attainment of head	.004	6
Course of income over past ten years	.003	9
Anticipated course of income over next ten years	.002	6

Table 17

UNADJUSTED AND ADJUSTED SAVING RATIOS

Characteristic	N	Two year saving as per cent of two year income	
		Actual	Adjusted
All self-employed businessmen artisans and farmers	137	27%	29%
Nonentrepreneurial home owners with home worth $2500 or more			
And total assets at start of $25,000 or more	91	20	16
And total assets at start of less than $25,000	499	11	10
Nonentrepreneurial people who do not own a home worth as much as $2500			
And total assets at start of $1,000 or more	102	-5	-3
And total assets at start of less than $1,000	172	5	8

Source: ISR Project 715, Deck T

Section 3.10 A Two-Stage Analysis: Hours Worked-Head

Charts 13 and 14 provide another illustration of a two-stage analysis. The data are taken from a national sample of spending units (24). The dependent variable is the number of hours worked by the head of the spending unit during the year. The analysis is performed on only those units where the head worked during the year.

The mean of the original distribution analyzed is 2092 hours. Its standard deviation is 797.

Predictors were divided into two parts, those felt to be early, or basic in a causal chain which might explain variation in hours worked, and those which were regarded as probably having later, more direct effects (see Table 18). The residuals from the first analysis were computed and were used as input to the second stage.

The interpretation of the stage one tree is straightforward. However, several things should be noted. In the split of group 3 into groups 12 and 13, those aged 75 and over are put together with the age 25-54 group. There are only six such observations. The split of group 7 into 8 and 9 is somewhat unexpected. Why should "having grown up on a farm outside the deep South" lead to long hours of work?

One plausible interpretation is the upward push of habits of work associated with farm background, uninhibited by the depressing effects of southern rural background (or associated race).

Notice that all the other splits in this tree involve separating off a group inhibited from working by some handicap, none of these groups being split again. The inference is that such handicaps are alternatives, any one being sufficient to keep a person from a full year's work.

This analysis accounted for sixteen per cent of the variation.

The second phase of the analysis included a large number of predictors, including some of those already used in the first phase. Four of them were constrained (monotonic). The dependent variable was the residual from the first phase. For each input observation, a large positive residual indicates that the dependent variable was larger than its predicted value. The mean of the dependent variable for this run was -5 (the departure from zero is due to truncation and rounding error). Its standard deviation was 732 hours. This second analysis explained 22 per cent of the variation in the residuals.

Chart 14 is quite plausible and meaningful. The most important factors reflect not motivations of the usual economic sort, but constraints, such as working for others (who set the hours of work) or being unemployed. After these effects are at least partially accounted for, it is clear that lower hourly earnings are associated with longer hours of work. Finally among a low-wage, self-employed group, those who migrated out of the South appear to work the longest hours of all. If these are people with ambition but not a great deal of education, many of them still farmers, the result makes sense.

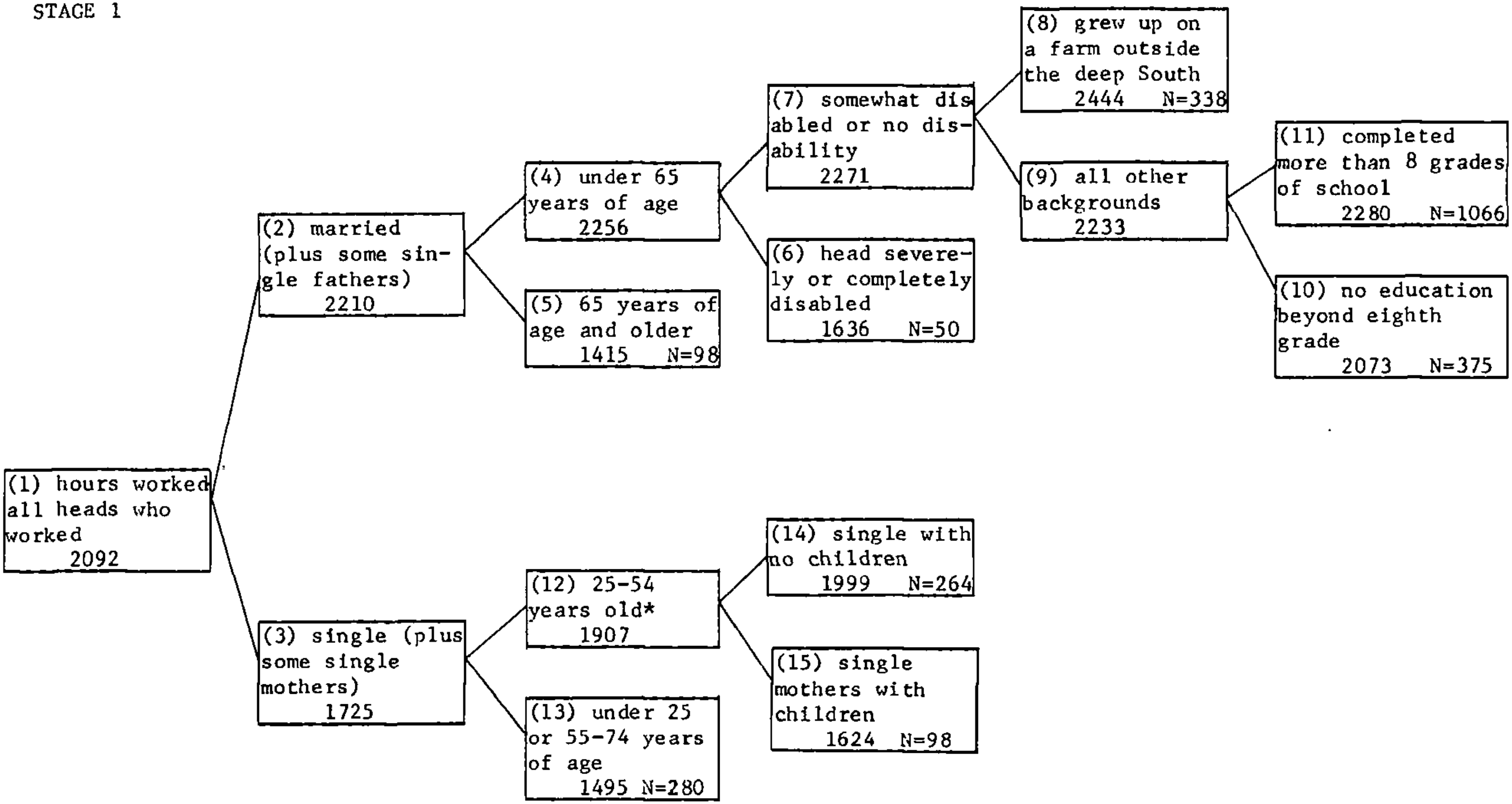
Unemployment experience can be thought of as not so much a cause for shorter hours, but as a joint result; both unemployment and short hours resulting from lack of basic skills or living in a labor surplus area. This serves as an explanation as to why some people work more than others, even after the main effects of age and education, etc., have been removed.

The tree was truncated by omitting two further splits using stage in the family life cycle, and selecting combinations of that combination which were difficult to interpret. This provides one more example of the need to restrict the explanatory factors to one dimension each.

CHART 13

HOURS WORKED -- HEAD
(EXCLUDES SU HEADS WHO DIDN'T WORK)

STAGE 1



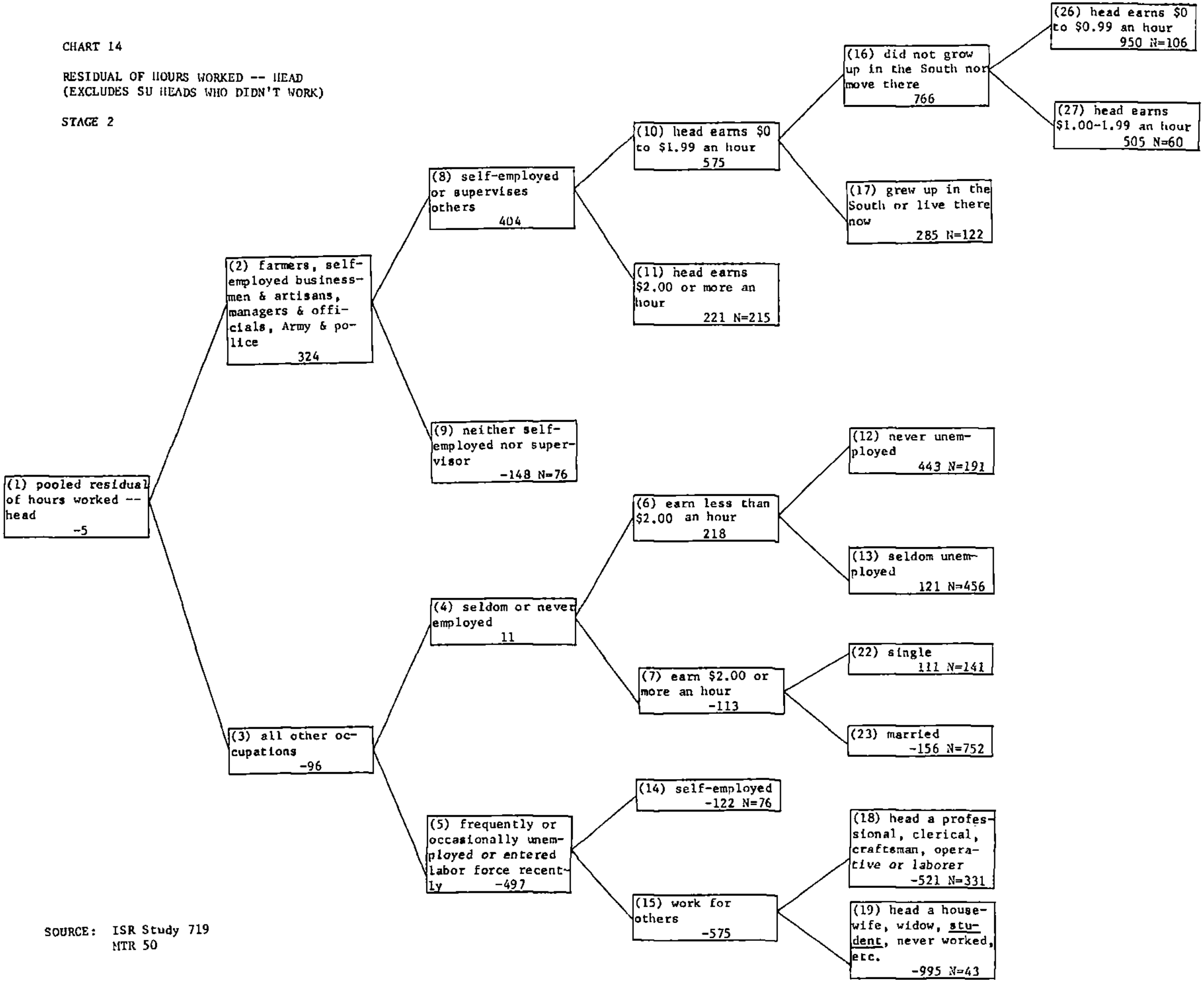
SOURCE: ISR Study 678
Deck 35, MTR 50

* Plus six cases 75 or older

CHART 14

RESIDUAL OF HOURS WORKED -- HEAD
(EXCLUDES SU HEADS WHO DIDN'T WORK)

STAGE 2



SOURCE: ISR Study 719
MTR 50

Table 18

VARIABLES USED IN THE ANALYSIS OF HOURS WORKED IN 1959 BY SPENDING UNIT HEAD

Number of subgroups	Monotonic or free	Description of Classification
First Stage		
4	M	Physical condition, whether or not a physical disability is reported
8	M	Education of heads
2	F	Race
7	F	Age
8	F	Sex, marital status, and number of children
4	F	Major religious affiliation (Protestants separated into Fundamentalist and non-Fundamentalist)
4	F	Index of need for achievement (three groups plus not ascertained)
6	F	Where heads grew up--deep South, rest of U.S., abroad, and whether on farm or not
Second Stage		
9	M	Wage rate of heads
6	M	Number of states lived in, and whether heads ever lived more than 100 miles from present residence
3	M	Whether heads or fathers grew up in a foreign country
10	F	Occupation of heads
3	F	Whether heads are self-employed, or supervise others, or neither
8	F	Frequency of unemployment of heads
7	F	Religion and frequency of attendance
7	F	Index of need for achievement and belief that hard work is more important than luck and help from friends
2	F	Race
9	F	Stage in the family life cycle (married, wife under or over 45, pre-school children, school children)
7	F	Difference in education between heads and their wives
5	F	Where heads grew up and where they live--urban rural migration
6	F	Where heads grew up (deep South?) and now live --north-south migration
6	F	Unemployment in the area--U.S.B.E.S. ratings
4	F	Plans to help parents or to send children to college
6	M	Size of place (city)
4	F	Difference in education between heads and their fathers
2	F	Sex of heads

CHAPTER IV

INTERPRETATION AND ANALYSIS STRATEGY

Section 4.1 Structure of the Trees

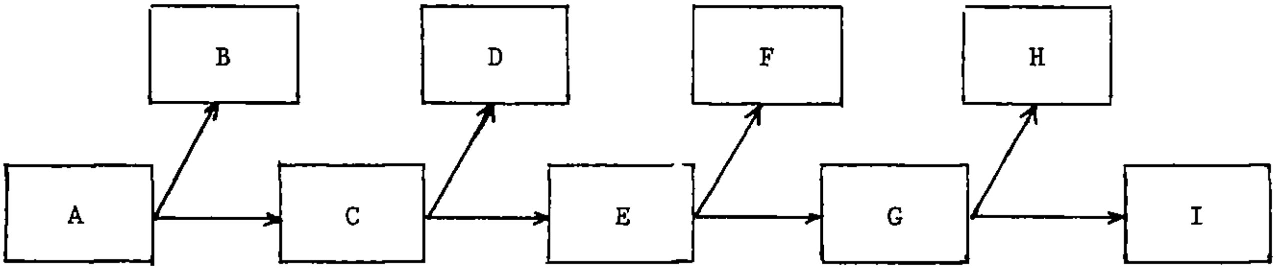
The analyses that have been presented show a series of characteristic tree patterns. Each tree has sections that can be described as a combination of two configurations, based on the useful convention of showing the group with the highest mean as the uppermost branch. One may be termed a trunk-twig structure, the other a trunk-branch structure.

The trunk-twig structure is a main branch from which small groups are split off from the main branch and are not themselves split again. This may take three forms, top-termination, bottom termination, and alternating termination. The top-termination structure may be termed an "alternative advantage" model. Group B consists of those observations possessing the "advantage" represented by that characteristic which split group A into groups B and C. Once group B has been established, it cannot be split further by the program.

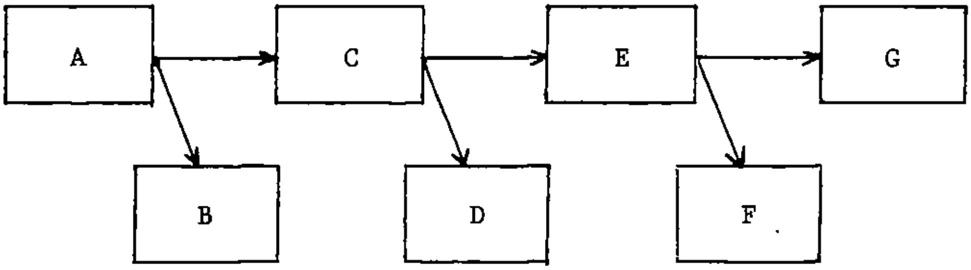
The bottom-determination structure may be termed an "alternative disadvantage" model, and is analogous. The possession of any one of a number of characteristics is enough to prevent an observation from achieving a high value on the dependent variable.

The interpretation of the alternating termination configuration is similar. In all three types, the interpretation to be made depends on the characteristics of the final groups themselves, especially on the number of observations in the group, its variance, and whether or not there existed predictor variables which "almost worked" in the attempt made by the program to split it.

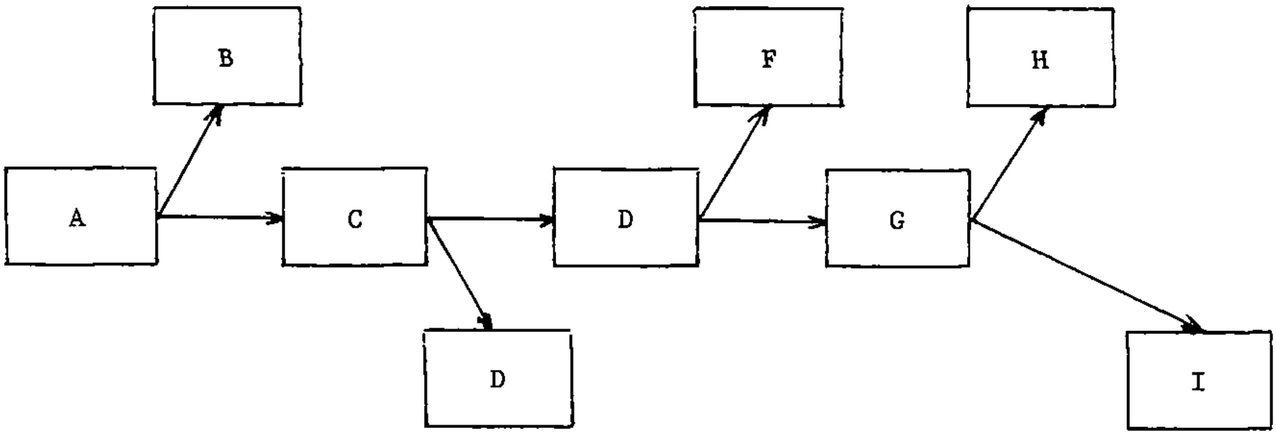
Another property of the tree is its symmetry or nonsymmetry in terms of the extent to which the same variables are used in the splits



TOP TERMINATION



BOTTOM TERMINATION

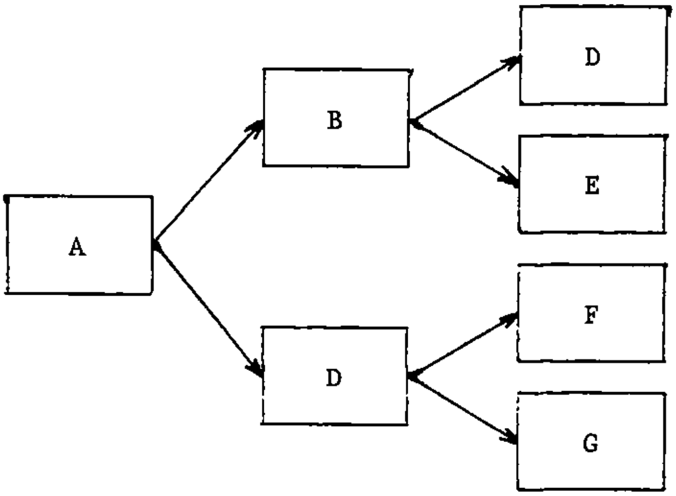


ALTERNATING TERMINATION

on the various trunks. Nonsymmetry implies interaction, i.e., effects of combinations of factors. If a variable is used on one of the trunks, and if it shows no actual or potential utility in reducing predictive error in another trunk, then there is clear evidence of an interaction effect between that variable and those used in the *preceding* splits. The utility of a predictor in reducing predictive error is evaluated by statistic $(BSS_{mpr}/TSS)_i$ for each predictor at each branch in the tree. This output is produced by the program and represents the proportion of the variation *in the group to which the predictor is being applied* that would be explained if it were used in a binary split of that group.

Trees may, of course, be symmetrical with respect to the way in which top-termination, bottom-termination and alternating-termination configurations appear in the main trunks.

The trunk-branch structure is usually typical of the first few splits of any tree. In this case, each group produced by a split is further subdivided.



TRUNK-BRANCH STRUCTURE

Some of the early groups may remain unsplit. If this is so, then the most important aspect of the interpretation of this structure has to do with the fact that there remains within-group variation which *can* be explained. At each step, the analytic question that should be asked is,"What are the reasons why there is as much variation in each of the groups as there is?" This question will be discussed below in more detail.

A further property of each tree is the number of final groups that result from the analysis. This is, of course, a function of the input sample size, the statistical properties of the algorithm, and the relationships between the characteristics of the predictor variables and the dependent variable.

Based on the present characteristics of the algorithm, we can distinguish three types of final groups: small groups, explained groups, and unexplainable groups. A small group is one containing too few observations to warrant an attempt to split it. An explained group is over this minimum size, but has too little variation in it (less than, say, 2 per cent of the original variation) to warrant an attempted split. An unexplainable group is sufficiently large and spread out, but no variable in the analysis is useful in reducing the unexplained variation contained within it. Each tree will generally have some of each of the three types. But the total number of final groups is heavily dependent on the rules used to stop the splitting process.

Section 4.2 The Rules for Stopping

What are the statistical considerations behind the choice of rules as to where to stop the splitting process? Just as there is no point in making any but the most important split at each stage--allowing other variables a later chance--so there is no point in making splits which are likely to be heavily influenced by sampling error.

It seems unreasonable to apply ordinary statistical tests at each split; that is, to insist that the split be a statistically significant difference between the two means. It is the best of a large number of possible splits at each stage. Even ignoring the re-ordering of subclasses, there are C_x-1 possible splits for each predictor at each stage (less some that have been eliminated because of previous splits), and the deductive logic associated with these tests does not apply.

The primary test is one of importance, i.e., the reduction in the error sum of squares. This is kept from being too arbitrary by expressing it as a per cent of the original total sum of squares. This is equivalent to saying that if there is a great deal of variation, the two new group means must be more disparate than if there is less variation. The use of error reduction also handles the problem of different numbers of observations in the two new groups, since the greater the disparity between group sizes (N's), the larger the difference between the means has to be to produce the same between-group sum of squares.

A separate test of significance, in addition to the test of importance, might be desirable in spite of the difficulties about degrees of freedom if there are likely to be splits which are not significant even on the boldest assumption, but which produce substantial error reductions. This tends to be true with skewed distributions and very small numbers of observations in a number of subgroups. But when this happens, there are serious problems no matter what multivariate technique is used.

The standard error of the difference between two means is:

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_1^2}{N_2}} \qquad (4.2.1)$$

If, in each split, we make the strong assumptions that the resulting two groups are roughly the same size and that their standard deviations are the same, then the standard deviation of the difference is approximately:

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \frac{\sqrt{2}\ \sigma}{\sqrt{N_1 + N_2}} = \frac{\sqrt{2}\ \sigma}{\sqrt{N_0}} \qquad (4.2.2)$$

where: $\sigma_1 = \sigma_2 \leqq \sigma$

and: $2N_1 = 2N_2 = N_0$, the size of the group being split.

Under the assumptions, this quantity differs from split to split, depending on the values of σ and N_0. But in practice, σ typically varies from split to split much less than N_0. It was for this reason that it was decided to add a cut-off criterion based on the size of the group to be split.

If a difference between two means is to be *significant* (*say, more* than twice its standard error), then it is required that:

$$(\bar{Y}_1 - \bar{Y}_2) > \frac{2\sqrt{2}\ \sigma}{N_0} \qquad (4.2.3)$$

and

$$\sqrt{N_0} > \frac{2\sqrt{2}\ \sigma}{(\bar{Y}_1 - \bar{Y}_2)} \qquad (4.2.4)$$

hence,

$$N_0 > \frac{(2\sqrt{2}\ \sigma)^2}{(\bar{Y}_1 - \bar{Y}_2)^2}$$

and

$$N_0 > \frac{8\ \sigma^2}{(\bar{Y}_1 - \bar{Y}_2)^2} \qquad (4.2.5)$$

Thus, one might not wish to make a split which was not "significant" even under these extremely lenient assumptions.

This implies that the minimum group size to be eligible for splitting really should depend on the standard deviation of the dependent variable, and on the size of the difference between the means of the two prospective new groups. In other words, in developing a rule about minimum group size, we should also pay some attention to the variance of the dependent variable.

But there is also the problem of the "number of things tried," which is relevant to the problem of fortuitous splits. The probability of this happening must be proportional to the number of possible splits at each step, since if we had enough classes available in the predictors, and a sufficient number of such predictors, we should be able to reduce the unexplained variation by half with each split. Thus, a term such as:

$$K = \sum_{i=1}^{NP} (C_i - 1) \qquad (4.2.6)$$

which is the total number of classes for each predictor (minus one), summed over all predictors, should also be taken into consideration. We note, however, that even this ignores the re-ordering of the classes during the partition scan.

In other words, for a group of any given size N_0, the larger the number of predictors and the more classes per predictor, the larger is the chance of finding a (fortuitous) split that is "important," but not "significant," particularly if one raised the significance levels to fit the situation (and the assumptions described by formulas 4.2.1-4.2.5 provide a situation which is one of the most powerful alternatives).

Clearly, there are a number of interesting problems in mathematical statistics raised here, the solution of which might lead to clearer rules about how to set the four cut-off criteria, the total number of final groups, the minimum interior sum of squares for a group to be eligible for splitting, the minimum number of cases for a group to be eligible for splitting, and the minimum between group sum of squares if a split is to be made at all.

Figure 1 provides an example of the relationship between the size of the split reducibility criterion (the program input parameter P2) and the number of final groups. The original analysis was run with this criterion set at .002. There were 13 predictors with a total of 70 classes and a sample size of 2569. The minimum group size rule was not used. The eligibility criterion P1 = .02 was used. The tree was then "pruned"; that is, it was determined how many final groups would have resulted if the reducibility criterion had been set at progressively higher levels. The resulting curve is a hyperbola which becomes asymptotic along the reducibility (R) axis at one, since there must be at least one final group, and which becomes asymptotic along the G (number of final groups) axis at about .002. The maximum possible number of final groups is, of course, N, the input sample size. The curve:

$$G \approx \frac{1}{\left(\frac{K-P}{\sqrt{N}}\right) R} \qquad (4.2.7)$$

where K is the total number of classes over all predictors,
where P is the number of predictors,
where N is the total number of input observations, and
where R is the split reducibility criterion,

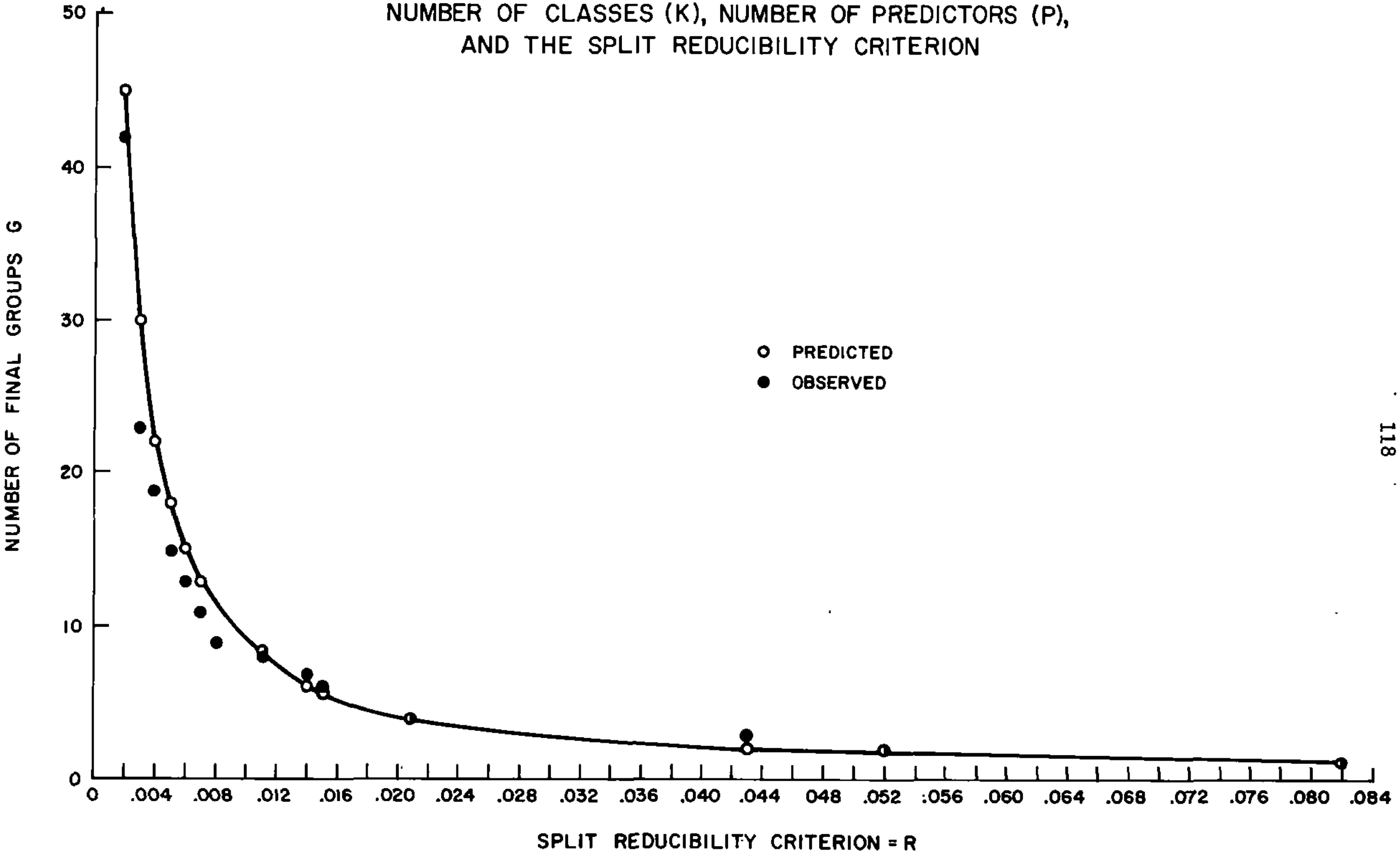
provides a reasonably good fit to the points observed. Further, if we plot G against X, where $X = (K-P)/\sqrt{N}$, with R held constant at .005, we find that (approximately) $G = 12X + 1$, so that for most cases, the relationship between G and X as defined is linear. Four cases do not fit this general model. All are truncated or badly skewed distributions.

Other analyses of this type have not been performed. This family of curves is suggested as an example of the lines along which some further investigation is needed.

All that can be stated here are some general rules which apply to the range of sample sizes, kinds of data and numbers of predictors which we have used.

FIGURE I

NUMBER OF FINAL GROUPS AS A FUNCTION OF $\sqrt{N}$, NUMBER OF CLASSES (K), NUMBER OF PREDICTORS (P), AND THE SPLIT REDUCIBILITY CRITERION



SOURCE ISR STUDY 678 DECK 35
719; MTR II

1. For typical survey data a minimum group size of 25 seems reasonable, since one hardly ever puts much credence in two subgroups whose combined N's add to 25 or less, however different their means. Any group i with $N_i<25$ should not be split. With other kinds of data with less error, however, a smaller number might be appropriate.

2. An eligibility rule that a group must contain at least two per cent of the total original sum of squares if it is to be considered eligible for splitting has the disadvantage that the program, as currently written, presents no data on the distributions of the predictors over that group. Indeed, this rule should be regarded as the least important of the four and kept low enough so that it is seldom, if ever, used. The minimum group size is more meaningful, and failure to find an acceptable split even more so. With minimum group size set at 25, and a reducibility criterion of .005 of the original total sum of squares, an eligibility limit of .015 seems to be low enough to assure that the other rules predominate.

3. The maximum number of groups should be regarded as a safety rule to cut off the program if something goes wrong, i.e., if the other rules were improperly set. There may be instances, however, where one wants only the best ten or twenty groups for some reason, such as developing procedures for assigning missing data, or developing a single new variable out of several raw variables.

4. The split reducibility criterion appears to be the crucial one to set; that is, the relative size of the between group sum of squares from a split which is necessary to allow that split to be made (after the best available one has been selected). The standard is like the one per cent or five per cent rule for significance tests. It is somewhat arbitrary.

   Our experience has been that with K less than 100 (formula 4.2.6) and samples of 2,000 to 3,000, and with a dependent variable that is not too badly skewed, the resulting trees seem manageable and interpretable with a requirement for error reduction of .005. With more predictors, or smaller samples, the criterion should be raised.

The skewness of the dependent variable also influences the number and type of the final groups produced. For example, in one case with a very large number of predictors and a sample of 1000, we produced a reasonable number of final groups using .005 as a reducibility criterion. But when we omitted 38 extreme cases (which accounted for 53 per cent of the total sum of squares), the same rules produced twice as many groups from the remainder.

These problems are not so serious as they might seem, since it is always possible to truncate (prune the tree) either for higher minimum group sizes or for a higher minimum split reducibility criterion. It is not possible to truncate on the basis of the size of the trunk-twig subgroups, since once one is split off, the remaining trunk is affected. Hence, if the dependent variable is skewed and a number of groups consisting of one or two observations are split off, these twigs cannot be pruned. In this case, the extreme cases should be removed, explained separately, and the analysis re-run without them, or else the dependent variable should be transformed into a somewhat more normally distributed form, perhaps using logarithms.

An added reason for using the split reducibility criterion rather than the others to do the real work of cutting off is that in this case all predictors are tried and the results printed out for each final group as well as the intermediate ones.

The analyst must decide whether each split shall be regarded as real or as containing fortuitous elements which should cause it to be disregarded. We have presented a rationale for setting the input parameters in such a fashion so as to minimize the probability of the occurrence of splits which are important (in the sense that they reduce the unexplained variation by a large amount), but not significant (in the sense that they could quite reasonably have occurred by chance).

We have investigated the sampling stability of the procedure in a limited fashion by using split-half techniques and by using the tree produced from one sample to predict values of the same dependent variable in another sample. Though the results seem quite encouraging, much more work needs to be done in this area.

In addition, we have also noted that the number of classes in the predictors are factors which contribute to the probability of a fortuitous split occurring, and that whether or not any given predictor has had its ordering constrained (is assumed to be monotonic) will affect the probability of its being involved in a fortuitous split. We have indicated a two-part rule for minimizing the effects of a large number of classes and the increased probability of a fortuitous split when a predictor is unconstrained.

a. Predictors which have a natural rank ordering to their classifications should be constrained to that ordering during the partitioning process, except where the possibility of a U-shaped or inverted U-shaped relationship between that variable and the dependent variable is suspected, in which case adjoining classes should be combined to form a maximum of five classes and the variable left unconstrained.

b. Un-ordered predictors should not have more than five or six classes, and should be left unconstrained.

We now turn to a description of techniques for displaying the results and to the problems of interpreting the behavior of the variables in the trees.

Section 4.3 Data-Display Techniques

A number of techniques can be used for summarizing and displaying the data produced by the AID (2) program. A number of these have been presented in previous sections. They may be described as follows:

1. The tree itself. A useful convention is to represent those groups with the higher means on the upper branches of the tree. A group may conveniently be represented as a box containing a short description of the predictor classes used in that particular split and which are included in the group, together with its mean, and standard deviation. We have included the N, or group size only on the final groups. A useful convention is to display the per cent of sample on each line leading to a box (see Chart 12). For convenience, an asterisk or other indicator may be used to mark a final group on which an attempted split was made, but which failed; that is, an unexplainable group.
2. The statistic $(BSS/TSS)_i$ can be examined for each predictor over each group created during the partitioning process (see Table 1) with suitable indicators to mark those variables used in the splits, other variables which were almost effective enough to be used, split fail attempts, and terminal groups.
3. For a rather gross, overall description of the behavior of the variables in the tree, a tabulation of the reduction in unexplained sum of squares attributable to the splits using β^2 for each predictor appears to be useful (see Table 14). This could also contain the statistic B^2, or gross effect of that variable if it were used in a one-way analysis of variance with all its classification detail, but without considering the effects of the other variables. This table also facilitates comparison with multiple regression statistics.
4. A detailed analysis of the behavior of one predictor is facilitated by the construction of a table (see Table 9) which shows all of the various classes of that predictor, and mean values of the dependent variable over these classes for the total input group,

and for the splits in which this predictor was involved. This table could also be developed for each subgroup occurring in the tree.

5. A description of the final groups listed in rank order on their means. This provides a summarization useful for presentation (see Table 11).

6. Frequency distributions of each of the predictors for each of the final groups provide additional information about the behavior of the variables. If residuals are punched from the program, obtaining such frequency distributions for variables not in the analysis is straightforward, and provides a method for investigating the extent to which variables have substituted for each other in the analysis. Distributions for the predictors used can also be produced by running a second-stage AID analysis using the residuals as the dependent variable.

These summarizations provide a number of devices for collecting the large amount of information produced by the program and organizing it in a fashion which facilitates the decision making process that constitutes the analysis.

Section 4.4 The Behavior of the Variables in the Trees

The analysis of the behavior of the predictors and their relationship to the dependent variable during the partitioning process can be approached through a series of questions, asked with reference to each partition.

Chance Factors

The first question is, "Given the minimum group size rule, split reducibility rule and split eligibility rule used, what is the likelihood that this split occurred by chance?" This problem may still occur even if the above-suggested rules have been used for minimizing the probability of its happening. If a variable actually used in the split is the only one which shows up as important, according to the criteria used, then the probability of its predictive power being based largely on sampling variability is relatively slight, unless it is an unconstrained variable with a large number of classes. When several variables are almost equally good as predictors, in any given split, then the likelihood is greater that sampling variability has had a hand in selecting one, rather than another, as that variable to be actually used in the split. The $(BSS/TSS)_i$ tabulation (display method 2, above) provides a guard against basing an interpretation only on those variables actually based in the partition process, since the explanatory power of the unused predictors is presented in all its detail.

The overall structure of the tree provides a clue as to the probability that sampling variability is operating together with a skewed distribution.

In the case where the dependent variable is badly skewed and has a tail extending toward the right (positive skewness), a top-terminating trunk-twig structure is likely to appear in several main branches of the tree. These terminal groups will have large, positive means, and will contain few (1-5) observations. Typically, they will result from splits on several different variables. Sooner or later the program will find some predictor which enables it to split out these extreme cases from the group in which they happen to be.

As we have mentioned previously, a careful re-reading of interviews may turn up a variable, certain values of which most of these extreme cases will have in common. This variable may then be inserted into a subsequent analysis. One may be reasonably confident that these observations will then be placed together in one group via a split on this variable. Good strategy would, therefore, dictate a preliminary investigation of the skewness of the dependent variable before the main analysis starts.

One might construct a dummy variable which has the value one if an observation is out in the skew tail and zero if it is not. A preliminary AID analysis, using this as the dependent variable, together with the predictors to be used in the main analysis will provide information as to which classes of the sample are out in the tail, rather than being in the main body of the distribution. It may be that one set of variables will be found optimal to explain being out on the tail of a distribution. Another set might prove best for explaining overall variation or variation in the main body of the destruction. This possibility would, of course, be of considerable theoretical importance.

Of course this technique need not be confined to observations out in a skew tail of the dependent variable distribution. For some analytic purposes it may be desirable to use this technique to determine what combination of variables are associated with an observation's being, say, in the second quartile of the distribution, or less than some specified value.

It should be noted that a variable which is not skewed in the total sample, may become skewed during the partitioning process. This cannot be caught in advance. Hence even when a preliminary investigation of skewness has been made, the analyst should be on his guard for the appearance of this particular trunk-twig structure (see Section 3.5). A bottom-terminating trunk-twig structure with small terminal groups would provide a signal for negative skewness.

Conceptualization Problems

A second question that should be asked is, "Does this split reflect conceptualization problems in applying the framework of predictor variables to the sample, or sections of it?" A number of interpretation problems in the trees may stem from measurement or coding errors, or from the use of variables that were designed for other statistical purposes. This technique is at its best when the predictors have a clear, uni-dimensional reference. We have presented one example of a conceptual problem that looked, initially, like a somewhat contradictory finding, until coding decisions were uncovered which appeared to misclassify uneducated people living on the fringes of cities of 50,000 and over, with respect to the rural or urban nature of their surroundings. Indices having several components also tend to behave in a somewhat peculiar fashion. Presumably, this is because the items in these indexes, though related both theoretically and statistically, may affect the dependent variable in different ways, particularly if some of them interact with other variables in the tree and others do not. Splits involving such variables may or may not "make sense." See Coombs (31) for a thorough discussion of scaling problems.

Perhaps the most important point to be made here is that problems like these are often revealed only by large standard errors that may accompany a multiple regression analysis. They tend to stand out quite clearly in the tree display of the AID results.

Substitution of Variables

A further question which should be asked with reference to any given split is, "Are there competing predictors correlated with the one actually used in the split? If so, does their explanatory power increase, decrease, or stay the same in subsequent splits?" The logic to be employed here is developed extensively by Hyman (2) in his discussion of spuriousness, and in his presentation of M- and P-type elaboration. He presents a formalization of the logic of examining the relationship between two variables when a third factor is introduced. The two factors under examination are referred to as x and y, and the third is called t. In our notation, x is the variable used to split group i into groups j and k; y is the dependent variable, and t is multiple and consists of each of the other predictors in the analysis. We are interested in the relationship between variable t and variable y, as represented by the statistics $(BSS/TSS)_i$, $(BSS/TSS)_j$ and $(BSS/TSS)_k$ for each predictor t. If, in addition, we consider whether or not there is a logical, theoretical justification for a correlation between x and t, and if so, whether x can be conceptualized as antecedent to t in a causal chain, we have a systematic application of the analysis strategies of:

1. Interpretation (t is an intervening variable)
2. Explanation (t is antecedent to x and is logically related to it)
3. Control for spuriousness (t is antecedent to x and cannot be related logically to it)
4. Specification (t is neither antecedent to x nor subsequent to it, but is logically related. Here x is a circumstance that affects the extent to which t is related to y.)

The reader is referred to Hyman (2) and to Blalock (32) for the details of the logic.

We note that we have reverted to a form of the analysis question, "Other things being equal, how does x affect y?" but in a somewhat different form. We now have the question, "When we extract variation associated with predictor x, how do the relationships between t_1, t_2, ..., t_p and y change?"

In providing an answer to this question that is meaningful, the question of the substitutability of variables in the analysis must be taken into consideration. This is the problem of intercorrelations between the predictors. Numerous examples may be seen in the trees. The variable "number of wage earners in the family" may really be serving to split off some old, retired people. The variable "pattern of income change" may really be splitting off people who are not in the labor force, i.e., old and retired. It is impossible here to consider all the problems associated with the relationship between a variable and the concept(s) it purports to represent, but a few points should be emphasized.

Some intercorrelations are built into the data by the coding process. Other high correlations may result because two predictors may themselves be the results of a third factor which may or may not be represented in the analysis by a variable. Still others are there because things go together in the real world. But it is on exactly this structure of relations that we are trying to get a grip. What is required is a strategy for minimizing the interpretation problems.

One way to deal with this is to put in the most clearly exogenous, most orthogonal and uni-dimensioned variables into a first-stage analysis, together with a relatively high reducibility criterion and fairly large minimum group size, and then use the richer matrix of predictors for an analysis of the residuals. Where a tight test is desired as to whether a variable which is of considerable theoretic importance has effects, this variable may be held out of the first-stage analysis and entered in the second stage to see whether it enables the explanation of residual variance. If a low eligibility criterion is used, the present algorithm will make a final sweep over all the final groups before dropping them from consideration, thus providing information on how all of the predictors are distributed within each group. (The present version of the program will not provide this, however, if the final group size (N_i) is under the specified minimum.) These distributions can be used to provide information as to whether the group occupies its present place because of its actual pedigree or because of some other factor(s) correlated with the ones used to form it.

Moreover, it would certainly be desirable to obtain information on the zero-order correlations among the predictors in the sample. Since they are classifications, this is not easy. A complete set of bivariate frequency distributions provides a general impression. Further improvements in the algorithm itself should provide for a satisfactory method of computing the intercorrelation matrix of predictors at each branch of the tree.

If there are some variables which, because of high intercorrelations, or low logical priorities, must be put into a second-stage analysis, one will not know (and has decided not to ask) what their influence would have been in the formation of the first-stage groups. The second stage, however, will show whether or not their influence on the dependent variable has already been accounted for. Re-introducing the first-stage variables into the second stage will also provide an answer to the question of whether there is a small, but universal, effect across all groups which will appear when they are pooled for the residual analysis.

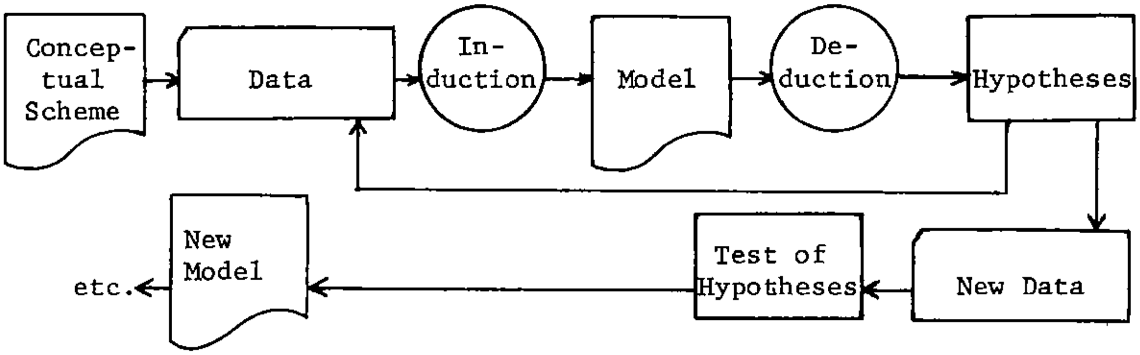
In some cases, the first-stage analysis will identify groups which are clearly constrained in some special way, and explained so clearly that they really should be eliminated from the subsequent analysis.

Concentrating on explaining the level of the dependent variable may tend to obscure other information contained in the tree which may be extremely important. The homogeneity of the final groups, especially if some of them appear after only a few splits, and are large in size, may be more interesting and important than their average on the dependent variable. Since the program produces the standard deviation as well as the mean of each group, one can examine the variance, or relative variance of each final group. If any group has a larger variance than the others, it raises the question of whether there is some other factor which affects this group, or varies more over it, but which was not included in the analysis.

The use of the tree strategy calls one's attention to the possibility that one or two variables may be sufficient for explaining the variation associated with some of the observations, whereas, additional theoretical sophistication may be required for an adequate explanation of the remainder of the sample.

Section 4.5 Overall Logic in Using AID (2)

The ongoing process of research in the social sciences involves both inductive and deductive reasoning (33). Theoretical orientations and conceptual schemes provide initial suggestions as to what type of data to collect. These ways of looking at the world often do not constitute a precise model, specifying exact or even probalistic relationships between clearly conceptualized and operationalized variables, nor are they often sufficiently precise to enable the deduction of specific hypotheses. But an ex-post-facto analysis of the data collection suggested by a conceptual scheme can serve as a basis for inductive reasoning, the results of which is a more precise model. Specific hypotheses can then be deduced and tried out (at least in a preliminary fashion) on the data which suggested the model from which they were deduced, and then tested on new data.



No multivariate analysis scheme can ever be a substitute for good, sound, theoretical work, but it seems clear that any one, including the AID algorithm, can be employed in both the inductive and deductive phases of research. In the inductive phase, it may be used as an aid to the formulation of a series of more precise statements about the behavior of the variables in the analysis. In the deductive phase, the tree must be consistent with the model or theoretical structure. This amounts to testing the whole model itself, rather than specific hypotheses deduced from it. The present procedure is focussed on the maximization of predictive ability. Its objective is to identify variables which discriminate between classes of observations for which predictability is good, and classes for which predictability is poor, while providing supplementary information suggesting model refinements to take care of the latter more adequately. It is based on the conventional idea that though correlation may not be sufficient to show causation, it is necessary.

CHAPTER V

POSSIBLE MODIFICATIONS TO THE PROGRAM

Section 5.1 Problems and Modifications

The work that we have done to date indicates that examining the strategy a scientist uses when working out the relationships between a few variables, formalizing it, and then extending it by means of a computer to many variables, can prove useful. The present programmed strategy is extremely limited. Certainly, additional experimentation in this type of simulation would be of value.

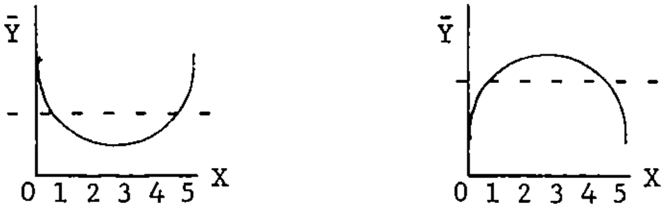
A number of unsolved problems with the present algorithm remain, and its usefulness could be extended by making it more sophisticated. We shall list some of the unsolved problems, propose some possible lines along which approaches to their solutions may lie, and sketch out some of the ways in which the present procedure might be extended. Then, finally, we shall take up the question of what additional modifications might be made to simulate a research analyst of somewhat greater sophistication.

The ability of the procedure to discriminate between classes of observations is based on some of the variables having important enough main effects to warrant their being used in a split. If any variable has only a very small main effect, but interacts with another variable which also has only a very small main effect, this procedure cannot discover it under certain conditions. As it stands, the class of discoverable interaction effects contains only those which involve variables, at least one of which has a detectable main effect, or which have detectable interactions with variables previously used in a split. One possible way out of this limitation would be to revise the algorithm to maximize the between-groups sums of squares one step ahead of the current step. This would involve an enumeration of all possible triads of splits on any given parent and its two children and the sacrifice of immediate predictability in favor of better predictions further on down in the tree. We note that the tree produced by the algorithm is not necessarily that one which is better than all other possible trees for the data under consideration. It is only optimum under the sequential algorithm used. But the closer one gets to explaining all of the variation, the more likely it is that sampling variation is being explained. One buys completeness with the coin of instability.

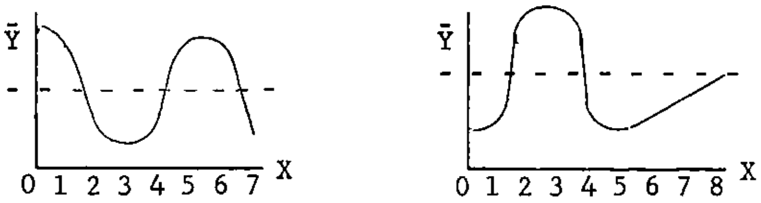
A second problem has to do with the flexibility of the constraints that may be placed on the predictors. They are presently specified to be in one of two modes, free or monotonic. One or more modes which intermediate between the two in constraints would be desirable for variables which have a natural ordering to them, that is, either bracketed equal interval scales or ranked classes. We consider the following cases:

Case 1. The slope of the regression of the ordered class means on their identifiers does not change in sign.

Case 2. This slope changes sign once.



Case 3. The slope changes sign twice.



The dotted line represents the desired split.

The cases where the slope changes sign more than twice probably represent sampling errors or a genuine absence of correlation between that predictor and the dependent variable, and will not be discussed. It would be desirable to be able to maintain the ordering of the predictor codes, yet permit splits of the type indicated. The present monotonic mode takes care of Case 1 adequately, but is inadequate for Cases 2 and 3. Leaving the predictor in free mode may allow sampling error to exercise an undesirable amount of influence on the rank ordering of the means in such a way as to produce an erroneous split. The problem is further complicated by the fact that missing information on some predictors is usually represented as a separate class, often as a nine or a zero. If a predictor is constrained to monotonic status and Case 2 or Case 3 represents the real state of its relationship to the dependent variable, then its utility will be severely and unduly limited.

We leave aside the question of missing information and consider Case 2 and Case 3. At least two strategies are possible. For Case 2, the algorithm could be modified to split the parent group into three parts, and then either combine the two groups which are most alike (the two high's or the two low's depending on whether the U is upright or inverted), or leave them as three separate groups. There are arguments for both strategies. Combining them tends to keep the group size large enough to permit a scan over all the predictors again. Keeping them separate might prove superior for theoretical reasons. If they are kept separate, subsequent frequency distributions may enable the conclusion that they have high values for different reasons. The same strategy may be extended to Case 3.

The presence of missing information complicates things somewhat. Into which group should these observations be placed? In the free mode, they are placed together with those observations whom they are most like on the dependent variable. An alternative strategy would be to distribute them among the newly created groups on a random basis. If the algorithm were modified to accept information about which predictors contained missing information and what characters had been used to represent this, either procedure could be used to handle missing information on variables of any mode.

The present procedure requires a dependent variable which is at least assumed to be an equal interval scale, or one which is dichotomous. It would be desirable to be able to handle a dependent variable which is a series of ordered, or ranked classifications. The statistic H, presented by Kruskal and Wallis (35) might be investigated for use here.

There is a great deal that is, as yet, unknown about the present procedure, especially with reference to the rules for setting the four cut-off criteria. Some preliminary work on the distribution of the number of final groups has been done, but the mathematical relationships between group sizes, variances, skewness, the number of predictors, the number of classes and the constraint status of the predictors have yet to be worked out.

A related question has to do with the sampling stability of the trees. The tree structure itself is probably subject to more variation than the models implied by it, since there is more than one way of arriving at nearly the same set of final groups. In general, it is likely that the more complex the tree, the greater the sampling variability that can be expected. This would be in line with findings reported by Ward (36), who found that when multiple regression equations developed one sample are applied to another sample, the correlations between predicted and actual values of the dependent variable tend to decrease more when complex functions are used than when simple linear regressions are used. When data come from a sample and the model leaves out a number of the sources of variation that occur in the real world, then increasing the complexity of the interaction terms to increase explained variation can only result in greater sampling instability, since one is fitting a very precise curve to a set of points whose values are partly random. One purchases completeness with the coin of instability. The answers are to get more data and to develop a model which takes these additional sources of variation into account.

The output of the present program could be made more useful by changing the logic to cause a final sweep over all final groups regardless of their size, amount of variation, or ability to be split, and printing the statistics for each final attempted split, together with an indication of which type of final group each is.

Two additional changes could be made in the present program, making the output more usable. One would be a summary print-out in table form, of the statistic $(BSS/TSS)_i$. Another would be better identification of final groups for which one or more of the predictors is a constant or is heavily clustered.

Another possibility would be the incorporation of a procedure for automatic scanning to detect the trunk-twig structure that indicates skewness. Or measures of skewness and kurtosis could be computed in advance of each attempted split and a decision made as to whether to attempt to locate a discriminator that would split off the observations in the tail, rather than explaining maximum variation. Examining the shape of the distribution of the dependent variable before each split might also provide the basis for a decision as to whether to split a group into two or three parts.

Still another addition to the algorithm which might be useful is the automatic pooling, or combining of final groups with similar mean values. This should probably be applied only to small, and to unexplainable final groups, and would involve the sacrifice of some explained variation, because the means of these groups would not be identical. But combining some of these groups might well make possible additional splits that would more than offset the losses. The subsequent groups might be very difficult to describe or explain, however.

Since the logic of elaboration and specification is heavily dependent on intercorrelations between predictors, it would be desirable to incorporate into the program the instructions necessary to compute an intercorrelation matrix of predictors associated with each split. This would enable the analyst to follow the patterns of change in the intercorrelations from split to split.

The reader should note that in some of the above suggested lines of modification, we are proposing to incorporate into the program some of the decisions that the analyst himself is, at present, making. This is particularly true in the detection of the trunk twig structure that indicates skewness. We ask the question, "What information does the analyst use to make a decision, what are his alternative lines of action, and on what basis does he choose one line of action over another?" If the information is already in the computer and if the basis for decision has a clear criterion and can be formalized, then it can be programmed.

One last line of promising development is suggested. Westervelt (11) has shown that artificial intelligence may be applied successfully to a sequential algorithm aimed at maximizing predictive power. He incorporated a simple learning procedure into the now well-known step-regression technique. Information about how to solve the problem is built up through experience with attempts to solve it. Thus, a further extension of the AID algorithm might well involve a series of trial trees which were not restricted to the best split at each stage, but chose on a random basis among those predictors which were almost equally good, and which produced information about what works and under what circumstances. By repeated iterations, modifying the probabilities with which each variable is used in each split on the combined basis of its effects in that split and the efficiency of subsequent splits, it may be possible to produce a tree which explains a great deal more of the variation in the dependent variable than that presently obtained.

Section 5.2 Strategy and Computers

The foregoing presentation has been based on the presently predominant method of using a large-scale digital computer, batch-processing. By this, we mean a machine-use mode in which problems are submitted to the computer in a stream--one after the other. In this mode, a problem is completely processed before another is started, and it is desirable for the analyst to get as much information out of a "job" as he can use.

This is not the only mode of machine organization. As computers increase in size and speed, the possibilities of the simultaneous processing of many problems grow highly probable. Indeed, at least one computer installation (37) is now experimenting with a remote console mode of operation. The analyst can then be brought into direct and immediate communication with the computer, re-acquiring not only the ability to intervene directly in the computing process (an ability severely lacking in the batch-processing mode of operation), but being able to do so with a great deal more power than he had when looking at banks of counters on a sorter. Moreover, programming techniques for translating problem-oriented languages similar to English into machine instructions are now developed to the point where direct on-line communication with a large scale computer operating in multi-processing mode is quite feasible.

This implies that far from being cloud-nine thinking, the distinct possibility of the analyst sitting at his desk with a console typewriter and requesting information from the computer is a realistic possibility. The following examples of possible requests might be typical of such a situation:

1. DISPLAY THE INTERCORRELATIONS BETWEEN X(1) AND X(2) IN GROUP 7.
2. DISPLAY AN UNSORTED, TENTATIVE, SPLIT OF GROUP 6 ON X(5).
3. CONTINUE AN AUTOMATIC ANALYSIS ON GROUPS 6, 9, AND 13.
4. DISPLAY THE $(BSS/TSS)_i$ TABLE FOR GROUPS 6, 7, AND 9.

This would allow the analyst to insert his hunches into the computing process.

CHAPTER VI

SUMMARY AND CONCLUSIONS

Section 6.1 Summary and Conclusions

Our starting point has been a consideration of some of the problems inherent in the application of multivariate statistical techniques to survey data (3). Most of the problems of analyzing this type of data have been reasonably well handled, except those revolving around the existence of interaction effects. A number of multivariate techniques are now in use, but this increased efficiency has been achieved primarily by imposing linearity and additivity assumptions. Since many useful concepts are classifications, their introduction into conventional multivariate procedures are difficult. Moreover, these procedures tend to obscure rather than illuminate errors in the measurement process. The fact that almost all survey samples are stratified and clustered leads to severe problems in the proper applications of statistical tests of significance. The intercorrelations between explanatory factors and interactions between them, make difficult the construction of precise theoretical models reflecting chains of causation, especially where the number of explanatory factors is large.

The procedure presented here represents an attempt to attack some of these problems by asking different kinds of statistical questions of the data than are implied by the choice of multiple regression techniques.

It is capable of handling a large number of predictors, will handle variables which are only nominal scales (i.e., mere classifications), and appears to be somewhat sensitive to measurement error. Linearity of relationships is not assumed. The problem of whether or not something could reasonably have occurred by chance will be with us as long as sampling techniques are used; but we suggest that the proper focus of the analysis should be on explanatory power, or importance, not significance. It is this focus which underlies what has been presented.

In the inductive phases of science the problem is to develop a model that fits the observed patterns of relationships between variables maximally. It is unlikely that a model which does not predict well for the sample upon which it is based will prove useful for very long without extensive modifications. Multivariate statistical methods are one of the tools used to develop such models. It is our hope that we have added a useful one to the tool-kit.

APPENDICES

APPENDIX A

REFERENCES

1. "University of Michigan Executive System for the IBM 7090 Computer," University of Michigan Computing Center, September 1963 (available from SHARE, IBM user's organization). (Mimeographed.)

2. Herbert Hyman, Survey Design and Analysis, Chap. VI and VII (Glencoe: Free Press).

3. James N. Morgan and John A. Sonquist, "Problems in the Analysis of Survey Data and a Proposal," Journal of American Statistical Association, 58 (June 1963), 415-34.

4. J. W. Tukey, The Future of Data Analysis, A Paper Presented at the 24th Annual Meeting of the Institute of Mathematical Statistics (Seattle, Washington, June 1961).

5. M. Ezekiel and K. A. Fox, Methods of Correlation and Regression Analysis (New York: Wiley, 1959), pp. 373-77.

6. W. A. Belson, "Matching and Prediction on the Principle of Biological Classification," Applied Statistics, VIII (1959), 65-75.

7. T. T. Tanimoto and R. G. Loomis, A Taxonomy Program for the IBM 704, program write-up (New York: IBM Applications Library, 590 Madison Ave.).

8. Stanley Reiter, "Choosing An Investment Program Among Interdependent Projects," Review of Economic Studies, 30, No. 1.

9. C. Alexander and M. L. Manheim, "HIDECS 2: An IBM 709/7090 Program for the Hierarchical Decomposition of a Set with An Associated Linear Graph," Behavioral Science, 8, No. 2 (April 1963), 168-70.

10. Paul Horst and Charlotte MacEwan, "Optimal Test-Length for Multiple Prediction, The General Case," Psychometrika, 22 (December 1957), 311-24 and references cited therein.

11. F. H. Westervelt, Automatic System Simulation Programming (Ann Arbor: University of Michigan College of Engineering, November 1960).

12. G. S. Watson, "A Study of the Group-Screening Method," Technometrics, 3 (August 1961), 371-88.

13. G. E. P. Box, *Integration of Techniques for Process Control*, Transactions of the Eleventh Annual Convention of the American Society for Quality Control, 1958.

14. O. D. Duncan, L. E. Ohlin, A. J. Reiss, Jr. and H. R. Stanton, "Formal Devices for Making Selection Decisions," *American Journal of Sociology* (May 1953).

15. Andre Daniere and Elizabeth Gilboy, "The Specification of Empirical Consumption Structures," in *Consumption and Saving*, eds. I. Friend and R. Jones, I (Philadelphia: University of Pennsylvania, 1960), 93-136.

16. Sewall Wright, "The Method of Path Coefficients," *Annals of Mathematical Statistics*, 5 (1934), 161-215.

17. Evelyn M. Kitagawa, "Components of a Difference Between Two Rates," *Journal of the American Statistical Association*, 50, No. 272 (December 1955).

18. K. S. Kretschmer and L. L. Vinton, "Some Empirical Studies in Discrimination Analysis" (unpublished manuscript, General Electric Co., New York, 1963).

19. J. N. Morgan and J. A. Sonquist, "Some Results from a Non-Symmetrical Branching Process That Looks for Interaction Effects," *Proceedings of the Social Statistics Section*, American Statistical Association, Cleveland, Ohio, September 1963.

20. B. Arden, B. Galler and R. Graham, *Michigan Algorithmic Decoder*, Programming Manual (Ann Arbor: University of Michigan Computing Center, January 1963).

21. J. Weizenbaum, *Symmetric List Processor*, Communications of the Association for Computing Machinery, 6 (September 1963), 524-44.

22. Frank M. Andrews, "The Revised Multiple Classification Analysis Program," Institute for Social Research, University of Michigan, Ann Arbor, August 1963. (Mimeographed, 13 pp.)

23. Daniel Suits, "The Use of Dummy Variables in Regression Equations," *Journal of the American Statistical Association*, 52 (December 1957), 548-51.

24. J. N. Morgan, M. David, W. Cohen, and H. Brazer, *Income and Welfare in the United States* (New York: McGraw-Hill, 1962).

25. "The 1959 Survey of Consumer Finances," *Federal Reserve Bulletin* (March, July, September 1959).

26. *Determinants of the Geographical Mobility of Labor*, Project 709, data collected by John Lansing and Eva Mueller, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, 1963.

27. Morgan, _et al_., _op. cit_., Table 18-8, pp. 267-68.

28. R. Friedman, P. Whelpton and A. Campbell, _Family Planning, Sterility and Population Growth_ (New York: McGraw-Hill, 1959).

29. G. Katona, E. Mueller, J. Lansing, C. Lininger, _The 1963 Survey of Consumer Finances_, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, 1964. In press.

30. J. N. Morgan and C. Lininger, "A Note on Education and Income," _Quarterly Journal of Economics_ (April 1964).

31. C. H. Coombs, _A Theory of Data_ (New York: Wiley, 1964).

32. H. Blalock, "Evaluating the Relative Importance of Variables," American Sociological Review, XXVI, No. 6 (December 1961).

33. M. Cohen and E. Nagel, _An Introduction to Logic and Scientific Scientific Method_ (New York: Harcourt Brace, 1934), Chaps. 14 and 20.

34. The reinterview panel was imbedded in the 1959, 1960 and 1961 Survey of Consumer Finances. Upper income units were sampled somewhat more heavily and weighted to eliminate biases. Of the 1059 units interviewed three times, extreme cases with apparent saving less than -49 per cent of income or greater than 99 per cent of income were excluded, leaving 941 cases. The main analysis of the panel data will be reported in a monograph to be published by the Survey Research Center in 1964.

35. W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," _Journal of the American Statistical Association_, XLIV (1952), 583-621.

36. Joe H. Ward, Jr., An Application of Linear and Curvilinear Joint Functional Regression in Psychological Prediction, Research Bulletin, Air Force Personnel and Training Center, Lackland Air Force Base, San Antonio, December 1954.

37. F. J. Corbato, _The Compatible Time-Sharing System: A Programmer's Guide_ (Cambridge, Mass.: M.I.T. Press, 1963), vii + 96 pp.

APPENDIX B

AID (Model 2) FORMULAS

FOR TOTAL:

Y_α = Value of dependent variable for the αth observation in the data

N = Total number of observations in the data

w_α = Weight value attached to the αth observation in the data

w_α = 1 if the data is unweighted

$$W = \sum_{\alpha=1}^{N} w_\alpha = \text{sum of weights*}$$

$$\Sigma Y = \sum_{\alpha=1}^{N} w_\alpha Y_\alpha = \text{sum of } Y$$

$$\Sigma Y^2 = \sum_{\alpha=1}^{N} w_\alpha Y_\alpha^2 = \text{sum of } Y\text{-squared}$$

$$\bar{Y} = \frac{\Sigma Y}{W} = \text{mean}$$

$$\sigma = \sqrt{\frac{1}{W}\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{W}\right)} = \text{standard deviation*}$$

$$TSS_T = \Sigma Y^2 - \frac{(\Sigma Y)^2}{W} = \text{total sum of squares}$$

$$BSS_T = \sum_{i=1}^{K} \frac{(\Sigma Y_i)^2}{W_i} - \frac{(\Sigma Y)^2}{W} =$$ between-group sum of squares where the summation ($i = 1, 2, \ldots, k-1, k$) is over the final unsplit groups

$$WSS_T = TSS_T - BSS_T = \text{within-group sum of squares}$$

---

*If W is small, say $W < 50$, and the run is unweighted, then it may be advisable to correct for small sample sizes and $\sigma_{adj} = \sqrt{N/(N-1)}$ where N is the number of observations over which summation has taken place.

Gross Beta Coefficient B_x^2 ; the proportion of variance which could be explained by predictor x alone in a one-way analysis of variance over its c_x classes

$$B_x^2 = \frac{BSS_x}{TSS_T}$$

where TSS_T is defined as above and

$$BSS_x = \sum_{k=1}^{C_x} \frac{(\Sigma Y_k)^2}{W_k} - \frac{(\Sigma Y)^2}{W}$$

and C_x is the number of classes defined by predictor x.

Partial Beta Coefficient β_x^2; the proportion of variance explained by predictor x in the tree.

$$\beta_x^2 = \frac{\sum_i TSS_{ix} - \sum_j TSS_{jx}}{TSS_T}$$

where i is over all parent groups split by predictor x, and j is over all new groups formed by splitting a parent group on predictor x. An equivalent computational formula using program output is:

$$\beta_x^2 = \sum_i \frac{TSS_{ix}}{TSS_T} - \sum_j \frac{TSS_{jx}}{TSS_T}$$

The total proportion of variance explained by the tree is:

$$R^2 = \sum_{x=1}^{NP} \beta_x^2 = \frac{BSS_T}{TSS_T}$$

where NP is the number of predictors used in the analysis.

The reduction in unexplained variation from any one split is

$$D = \frac{TSS_i}{TSS_T} - \left(\left[\frac{TSS_j}{TSS_T} \right] + \left[\frac{TSS_k}{TSS_T} \right] \right)$$

where i is the identifier of the group being split and j and k are the identifiers of the resultant groups.

Formulas for the i'th Group

N_i = number of observations in the i'th group

$W_i = \sum_{\alpha=1}^{N_i} w_\alpha$ = sum of weights in the i'th group

$\Sigma Y_i = \sum_{\alpha=1}^{N_i} w_\alpha Y_\alpha$ = sum of Y in the i'th group

$\Sigma Y_i^2 = \sum_{\alpha=1}^{N_i} w_\alpha Y_\alpha^2$ = sum of Y-squared in the i'th group

$\bar{Y}_i$ = mean of the i'th group

$TSS_i = \sum_{\alpha=1}^{N_i} w_\alpha Y_\alpha^2 - \frac{\left(\sum_{\alpha=1}^{N_i} w_\alpha Y_\alpha\right)^2}{W_i}$ = total sum of squares in the i'th group

$\sigma_i = \sqrt{\frac{TSS_i}{W_i}}$

$D_i = \bar{Y}_i - \bar{Y}$ = deviation of the mean of the i'th group from the grand mean

$\frac{TSS_i}{TSS_T}$ = proportion of the original total sum of squares still left in the i'th group

$\frac{W_i}{W} \times 100$ = per cent of total = (weighted) proportion of the observations in the i'th group

$\frac{\left(\sum_{\alpha=1}^{N_i} w_\alpha Y_\alpha\right)^2}{W_i}$ = weighted mean square for the i'th group

PA = PCT1 (an input parameter) times TSS_T. If the i'th group is to become a candidate for splitting, then $PA \leqq TSS_i$.

PB = PCT2 (an input parameter) times TSS_T. In order for an attempted split on group i to be allowed the requirement $PB \leqq BSS_{mpi}$ must be met.

$$BSS_{mpi} = \frac{\left(\sum_{j=0}^{m} Y_j\right)^2}{\Sigma W_j} + \frac{\left(\sum_{j=m+1}^{c} Y_j\right)^2}{\Sigma W_j} - \frac{\left(\sum_{j=0}^{c} Y_j\right)^2}{\Sigma W_j}$$

where m is the split between the m'th and m+lst classes of predictor p over group i. C is the maximum value attained by predictor p.

BSS_{mpi} is maximized over all classes of all predictors over group i.

The split of group i occurs after selection of the maximum BSS_{mpi} and occurs only if the criterion $PB \leqq BSS_{mpi}$ is met. There are C-1 elements in the BSS column produced by the partition scan output. These are the BSS_{mpi}. The C'th element is TSS_i for the group being split. Ratio of variation in group i explained by unsuccessful predictor r in attempted partitioning of group i, $BSS/TSS = BSS_{mpr}/TSS_i$.

APPENDIX C

A Note on Partitioning for Maximum Between Sum of Squares

11/10/62

by W. A. Ericson

1. The Problem

This note presents some results, both positive and negative, concerned with analysis of the following problem:

One is given $k > 2$ sets of observations, where

$$\bar{x}_i \,, \qquad i = 1, 2, \ldots, k$$

is the mean of the observations within the i'th set and

$$N_i \,, \qquad i = 1, 2, \ldots, k$$

is the number of observations in that set. The problem is to partition these k sets of observations into two nonempty classes such that the "between class sum of squares" is maximized. In other words, to find I, a set of any m $(1 \leq m < k)$ of the k indices $i = 1, 2, \ldots, k$, such that

$$N_I(\bar{x}_I - \bar{x})^2 + N_{\bar{I}}(x_{\bar{I}} - \bar{x})^2 \tag{1}$$

is maximized, where

$$N_I = \sum_{i \in I} N_i \,, \qquad N_{\bar{I}} = \sum_{i \notin I} N_i \,,$$

$$\bar{x}_I = \frac{1}{N_I} \sum_{i \in I} N_i \, \bar{x}_i \,, \qquad \bar{x}_{\bar{I}} = \frac{1}{N_{\bar{I}}} \sum_{i \notin I} N_i \bar{x}_i \,,$$

and $\bar{x}$ is the overall mean, i.e.,

$$\bar{x} = \frac{N_I \bar{x}_I + N_{\bar{I}} \bar{x}_{\bar{I}}}{N_I + N_{\bar{I}}} \,.$$

2. Previous Results

No literature search having been made, it is not known whether this problem has been researched by other investigators. This remains a point for further study.

3. Restatement and Assumptions

It is well-known that the problem outlined above is basically unchanged by the addition of the same arbitrary constant to each $\bar{x}_i$. It may thus be assumed without loss of generality that

$$\bar{x}_1 \geq \bar{x}_2 \geq \cdots \geq \bar{x}_k > 0 \tag{2}$$

Furthermore, it is easily seen that maximizing (1) by choice of I is equivalent to maximizing

$$f(I) \equiv \frac{(N_I \bar{x}_I)^2}{N_I} + \frac{(N_{\bar{I}} \bar{x}_{\bar{I}})^2}{N_{\bar{I}}} \tag{3}$$

4. A Negative Result

The following algorithm was suggested for finding I and its complement, $\bar{I}$, which maximizes (3):

a) Compute $f(I)$ for I taken, in turn to be $\{1\}$, $\{2\}$, ..., $\{k\}$.

b) Pick the maximum $f(I)$ over these I's. Suppose, e.g., $I = \{a\}$ maximizes $f(I)$ over the I's considered in (a).

c) Compute $f(I)$ for I taken in turn to be $\{a,1\}$, ..., $\{a, a - 1\}$, $\{a, a + 1\}$, ... $\{a, k\}$.

d) Choose that I, among those considered in (c) which maximizes $f(I)$, say $I = \{a,b\}$. If $f(\{a\}) > f(\{a,b\})$, stop and assert $I = \{a\}$

yields maximum value of (3), otherwise continue the process, looking next at f(I) for I's of the form $\{a,b,i\}$, $i \neq a$, $i \neq b$, repeating steps (c) and (d) above.

This procedure does not lead invariably to the optimum or maximizing partition, I. That this is so is demonstrated by the following counterexample:

Suppose $k = 5$ and the data are as shown below:

i:	1	2	3	4	5
$\bar{x}_i$:	3.1	3.0	2.0	2.0	1.0
N_i:	1	2	3	1	3

It is easily verified that

I	$\bar{I}$	f(I)
{1}	{2,3,4,5}	41.72111
{2}	{1,3,4,5}	42.85125
{3}	{1,2,4,5}	40.40142
{4}	{1,2,3,5}	39.31764
{5}	{1,2,3,4}	44.77285

Following the suggested algorithm we next look at $I = (5,i)$, $i = 1,2,3,4$, and obtain the following:

I	$\bar{I}$	f(I)
{5,1}	{2,3,4}	41.96916
{5,2}	{1,3,4}	40.84200
{5,3}	{1,2,4}	44.30250
{5,4}	{1,2,3}	44.25166

Each of these values of $f(I)$ being less that $f(\{5\})$, we conclude, as per the suggested algorithm, that $I = \{5\}$ maximizes (3). This is not true since it is easily shown that

$$f(\{1,2\}) = 44.88904 > f(\{5\}) = 44.77285$$

5. The Basic Result

It will be proved in this section that (3) is maximized over all possible I's by I^* where I^* is that set $I_m \equiv \{1,2, \ldots, m\}$, $1 \leq m < k$ for which $f(I^*) \geq f(I_m)$ for all m. Thus to find the maximizing partition one need only compute $f(I)$ for the $k - 1$ sets I_m and choose the maximum. Furthermore, I^*, obtained in this fashion, maximizes (3) over <u>any</u> partition of the $N = \sum_1^k N_i$ <u>individual observations</u> into two sets (assuming each individual observation within any set equals the set mean $\bar{x}_i$ say).

The present proof of these assertions, while straightforward, involves considerable tedious algebra. Further study may yield more succinct and more tidy demonstrations. The present proof is given in two parts. We first state and prove the theoretical results, in some degree of generality and then make the necessary identifications to the problem stated in §1 by which the assertions stated above become established.

We adopt the following notation: let

$$a_1 \geq a_2 \geq a_3 \geq \cdots \geq a_N \tag{4}$$

be any nonincreasing sequence of real positive numbers. Let P_m and P_n be <u>any</u> partition of the N a_i's, i.e., P_m is any set of m of the a_i's and P_n is the set of the remaining $n = N - m$ a_i's. Further, let H_m, L_m and M be respectively the set of the largest m a_i's, the set of the smallest m a_i's, and the $n - m$ middle a_i's. (It is assumed that $n \geq m$, hence M is null if $n = m$, otherwise not.) Thus

$$H_m = \left\{a_1, \ldots, a_m\right\}$$

$$L_m = \left\{a_{N-m+1}, \ldots, a_N\right\}$$

$$M = \left\{a_{M+1}, \ldots, a_{N-m}\right\}$$

The first result may then be stated as

Theorem A: At least one of the following is true:

a) $$\frac{(\Sigma(H_m))^2}{m} + \frac{(\Sigma(M) + \Sigma(L_m))^2}{n} \geq \frac{(\Sigma(P_m))^2}{m} + \frac{(\Sigma(P_n))^2}{n}$$

b) $$\frac{(\Sigma(L_m))^2}{m} + \frac{(\Sigma(M) + \Sigma(H_m))^2}{n} \geq \frac{(\Sigma(P_m))^2}{m} + \frac{(\Sigma(P_n))^2}{n} ,$$

where $\Sigma(H_m) \equiv \sum_{a_i \epsilon H_m} a_i$, etc.

▼ Proof: The theorem is obviously true if either $\Sigma(L_m) = \Sigma(P_m)$ or $\Sigma(H_m) = \Sigma(P_m)$. We then consider the other cases, i.e., $\Sigma(H_m) > \Sigma(P_m) > \Sigma(L_m)$, and show that if (a) fails then (b) holds. Straightforward algebra* shows that if (a) is false, then

$$[m\Sigma(P_n) + m(\Sigma(L_m) + \Sigma(M)) - n(\Sigma(H_m) + \Sigma(P_m))] > 0. \tag{5}$$

Similarly, (b) is true if

$$[m\Sigma(P_n) + m(\Sigma(M) + \Sigma(H_m)) - n(\Sigma(L_m) + \Sigma(P_m))] \geq 0. \tag{6}$$

That (5) implies (6) is obvious, since the left side of (6) is greater than or equal to the left side of (5). ▲

---

*The major hint needed in going from (a) and (b) to (5) and (6) is to replace

$$[\Sigma(m) + \Sigma(L_m)]^2 \text{ by } [\Sigma(m) + \Sigma(L_m)]\ [\Sigma(P_n) + \Sigma(P_n) - \Sigma(H_m)]$$

and to replace

$$[\Sigma(P_n)]^2 \text{ by } [\Sigma(P_n)]\ [\Sigma(m) + \Sigma(L_m) + \Sigma(H_m) - \Sigma(P_m)]$$

etc.

The second main result is given by the following:

Theorem B: Suppose

$$a_1 \geq \ldots \geq a_m > a_{m+1} = \ldots = a_{m+n}$$

$$= a_{m+n+1} = \ldots = a_{m+n+\ell} > a_{m+n+\ell+1} \geq \ldots \geq a_{m+n+\ell+r}$$

where $m + n + \ell + r = N$, $m \geq 0$, $n > 0$, $\ell > 0$, $r \geq 0$, and $m + r \geq 1$. Then at least one of the following statements is defined and true:

$$\text{d)} \quad \frac{1}{m}(\Sigma_m)^2 + \frac{1}{n+\ell+r}((n+\ell)a + \Sigma_r)^2 \geq \frac{1}{m+n}(\Sigma_m + na)^2 + \frac{1}{\ell+r}(\ell a + \Sigma_r)^2$$

or

$$\text{d)} \quad \frac{1}{m+n+\ell}(\Sigma_m + (n+\ell)a)^2 + \frac{1}{r}(\Sigma_r)^2 \geq \frac{1}{m+n}(\Sigma_m + na)^2 + \frac{1}{\ell+r}(\ell a + \Sigma_r)^2 ,$$

where $a \equiv a_i$, $i = m+1, \ldots, m+n+\ell$, $\Sigma_m \equiv \sum_{i=1}^{m} a_i$,

$\Sigma_r \equiv \sum_{i=1}^{r} a_{m+n+\ell+i}$.

Proof: If $m = 0$, it is immediately verifiable that (d) is true. Likewise, if $r = 0$, then (c) is true. Suppose then that m, n, r, and ℓ are all positive. Straightforward algebra shows that (c) is equivalent to

$$\text{c}') \quad A \equiv (\Sigma_m)^2 - 2ma\,\Sigma_m \geq \frac{m(m+n)}{(n+\ell+r)(\ell+r)}(\Sigma_r)^2 - \frac{2mr(m+n)}{(n+\ell+r)(\ell+r)}a\Sigma_r$$

$$+ \frac{m[(m+n)\ell - (\ell+r)n]^2 - m[(m+n)\ell^2 + (\ell+r)n^2]}{n(n+\ell+r)(\ell+r)}a^2 \equiv B$$

and (d) is equivalent to:

$$\text{d}') \quad A \equiv (\Sigma_m)^2 - 2ma \leq \frac{(m+n+\ell)(m+n)}{r(\ell+r)}(\Sigma_r)^2 - \frac{2(m+n+\ell)(m+n)}{(\ell+r)}a\Sigma_r$$

$$- \frac{\{[(m+n)\ell - (\ell+r)n]^2 - r[(m+n)\ell^2 + (\ell+r)n^2]\}}{\ell(\ell+r)}a^2 \equiv C .$$

To show that either (c) or (d) is true (or both) it suffices then to show that if (c') is false then (d') must be true. This is clearly established if the right side of the inequality in (c') is less than or equal to the right side of the inequality in (d'), i.e., if $C - B \geq 0$. But some simple but tedious algebra shows that

$$C - B = \frac{(m+n)[(n+\ell+r)(m+n+\ell) - mr]}{r(n+\ell+r)(\ell+r)} [\Sigma_r - ra]^2 \quad ,$$

which is obviously nonnegative.▲

To use these results for the problem stated in §1 above and to establish the assertions at the beginning of the present section one need only identify the following nonincreasing sequence with those sequences of a_i's referred to above:

$$\underbrace{\bar{x}_1, \ldots, \bar{x}_1}_{N_1}, \underbrace{\bar{x}_2, \ldots, \bar{x}_2}_{N_2}, \underbrace{\bar{x}_3, \ldots, \bar{x}_3}_{N_3}, \ldots, \underbrace{\bar{x}_k, \ldots, \bar{x}_k}_{N_k} .$$

Then it is clear that Theorem A establishes the fact that for <u>any</u> partition of these $N = \sum_1^k N_i$ $\bar{x}_i$'s into two sets of m and $n = N - M$ elements respectively will yield a value of "between sum of squares," (3), no larger than that for either the partition consisting of the m largest $\bar{x}_i$'s and the N-m remaining or the m smallest $\bar{x}_i$'s and the $N - m$ remaining. This result clearly includes the case where for every $i = 1, \ldots, k$ all N_i $\bar{x}_i$'s are put in the same one of the two sets forming the partition, i.e., the case where the partition is of the k sets of means rather than of the N individual means.

Theorem B then closes the remaining loophole, viz., it may be that some partition, I, $\bar{I}$, of the k sets of means into $N_I = \sum_{i \in I} N_i$ and $N_{\bar{I}} = N - N_I$ observations, respectively, has a sum of squares, (3), which is no larger than that for the partition consisting, say, of the largest N_I individual $\bar{x}_i$'s and the $N_{\bar{I}}$ remaining $\bar{x}_i$'s. However, this latter partition may very easily split one set of N_i identical $\bar{x}_i$'s. Theorem B then says that for any partition of the N individual $\bar{x}_i$'s into the m largest and $N - m$ remaining and where the partitioning point occurs within one of the k sets of observations then there is another partition into largest and smallest $\bar{x}_i$'s where the partitioning point occurs between two of the k sets of $\bar{x}_i$'s and which has a between sum of squares no smaller than the original partition.

Theorems A and B then together demonstrate that to find the partition which maximizes (3) one need only look at the $k - 1$ partitions, I_m, where $I_m = \{1,2, \ldots, m\}$, $1 \leq m < k$, and choose that one yielding the largest value of (3).

6. A Final Negative Result

It was further conjectured that perhaps (3), $f(\{I_m\})$, $m = 1,2, \ldots, k - 1$, treated as a function of m was well-behaved in the sense of say concavity and that, e.g., if $f(\{I_1\}) > f(\{I_2\})$ then one might be able to stop and assert $I^* = I_1$, and thus not look at all $k - 1$ I_m's. This is not the case, however, as witnessed by the following counter example:

i	1	2	3	4	5
$\bar{x}_i$	3.000	2.01000	2.0010	2.0001	1.0000
N_i	1	1	1	1	2

,

here one finds the following values for $f(\{I_m\})$, $m = 1, 2, 3, 4$:

I_m	$f(\{I_m\})$
$\{1\}$	21.84
$\{1,2\}$	21.55
$\{1,2,3\}$	21.72
$\{1,2,3,4\}$	22.30

7. Conclusions

The above results indicate that to find the partition which maximizes the between sum of squares, (3), one need only compute (3) for the $k - 1$ partitions consisting of the first set of size N_1 and all the rest, the first two sets of size $N_1 + N_2$ and all the remaining, etc., and choosing that one which maximizes (3). Further the partition found in this manner maximizes (3) over <u>all</u> partitions of the $N = \sum_1^k N_i$ individual observations (assuming each observations within any one of the k sets equals the mean of that set). Finally it does not seem possible to improve on this technique, in the sense of reducing the computational burden.

APPENDIX D

AID (2) ALGORITHM

Preliminary Read in. Steps 1 and 2.

1. Read in all parameters and all input observations, including all predictors and the dependent variable Y. Screen out observations where Y is missing data or it is not desired to use this observation. Save all observations on tape if necessary.

2. To start, identify all observations used in the analysis as belonging to group number one. Group number one is the current candidate group. Go to Step 6.

Test for Termination of the Procedure. Step 3.

3. Determine whether or not the current number of unsplit groups is about to exceed the maximum permissible number; if so, go to Step 22, as the problem cannot proceed further.

Determine Which Group Should Be Selected for Attempted Partitioning. Steps 4-6.

4. 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 R per cent of the total sum of squares for the input observations (TSS_t);

   b. the number of observations in the group is not smaller than MSIZE;

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

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

   e. the total sum of squares of that group is not smaller than the sum of squares for any other group that meets the above four criteria.

5. If there is no such group, go to Step 23; the problem is complete.

6. 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 print out N_i, ΣY_i, ΣY_i^2, $\bar{Y}_i$, and TSS_i.

Partition Scan Over All Predictors. Steps 7-19.

7. Set $j = 1$ and go to Step 9.

8. 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 20.

9. Compute N_{ijc}, ΣY_{ijc}, ΣY^2_{ijc}, $\bar{Y}_{ijc}$ for each class c of predictor j over group i.

10. 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 8.

11. If predictor j has been defined as monotonic, skip Step 12, do <u>not</u> sort the Step 9 statistics, go to Step 13 instead.

12. Sort the statistics produced in Step 9, 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 13-17.

13. Set $p = 1$ and go to Step 15.

14. Increase p by 1. If p is larger than $(c_j - 1)$, where c_j is the number of classes in the j'th predictor, then print the statistics for class c_j and go to Step 18 as all possible feasible splits have been examined.

15. If $\Sigma N_k = N_1 = 0$ for $k = 1, \dots p$, or if $(N_i - N_1) = N_2 = 0$, go to Step 14 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 $(1, \dots, p)$ and the adjacent sorted classes $(p + 1, \dots, c)$. Print the statistics for class p.

16. If this BSS_p is not larger than any BSS_p previously computed for this predictor over this group, go to Step 14.

17. This is the largest BSS_p encountered so far for this predictor. Remember BSS_p and the partition number p; print them and go to Step 14.

Determination of Best Predictor. Steps 18-19.

*18. 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 8.

19. 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 8.

Is the Best Predictor Worth Using? Steps 20-21.

*20. Was the maximum BSS retained after the scan of all predictors over group i equal to at least Q per cent of the total sum of squares? If not, mark group i as having failed in a split attempt and then go to Step 4.

21. Group i is to be split into two new groups and destroyed. Using the class identifiers and the partition rule remembered from Step 19, 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 unsplit groups by one. Then go to Step 3.

Termination of the Algorithm. Steps 22-26.

22. The maximum number of permissible unsplit groups has been reached. Print an appropriate comment and go to Step 24.

23. There are no more groups eligible for further splitting. Print an appropriate comment and to go Step 24.

24. 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.

25. Determine whether punched or tape residuals are desired. If so, go to Step 26, otherwise go to Step 1.

26. Compute predicted values of Y and residuals and, by option, punch them and/or write them on tape with the data. Then go to Step 1.

---

*These decision rules constitute the crucial steps in the algorithm.

Formulas

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

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

$$BSS = \frac{(\Sigma Y_1)^2}{N_1} + \frac{(\Sigma Y_2)^2}{N_2} - \frac{(\Sigma Y)^2}{N}$$

$$WSS = TSS - BSS$$

$$Y_\alpha = \bar{Y}_i$$

$$R_\alpha = \tilde{Y}_\alpha - Y_\alpha$$

AID (2) Algorithm: Summary

1. Considering the currently unsplit sample subgroups having at least 25 observations in them, select that sample subgroup which has the largest total sum of squares, such that $TSS_i \geq R(TSS_T)$

$$TSS_i = \Sigma Y_i^2 - \frac{(\Sigma Y)^2}{N_i}$$

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

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

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 are not placed together in one of the new groups to be formed. (3) The sorting of classes may be suppressed by option.

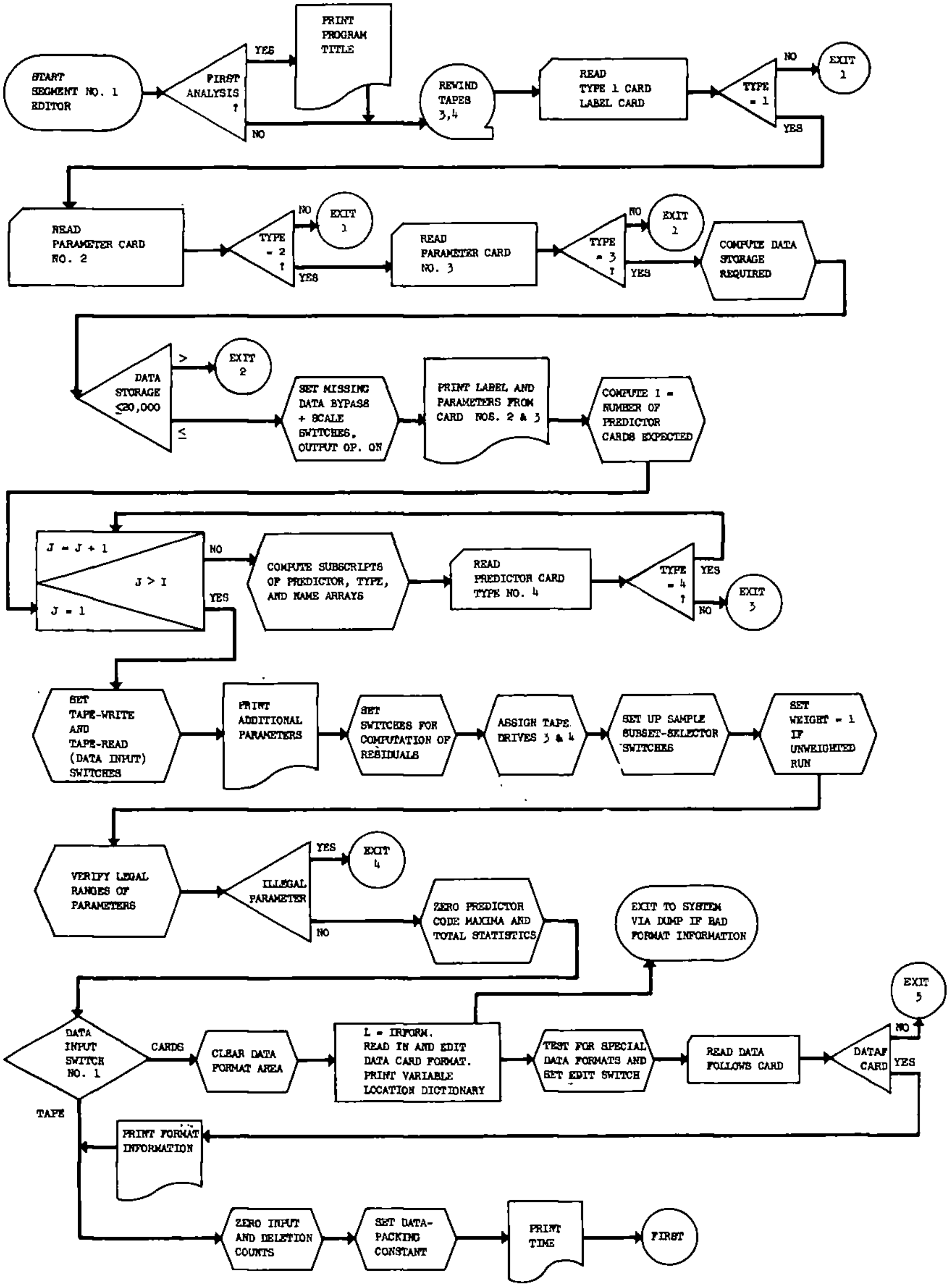
3. For a partition p on variable k over group i to take place after the completion of (2), it is required that:

$$(N_1\bar{Y}_1^2 + N_2\bar{Y}_2^2) - N_i\bar{Y}_i^2 \geq Q\ (\Sigma Y_T^2 - N\bar{Y}_T^2)$$

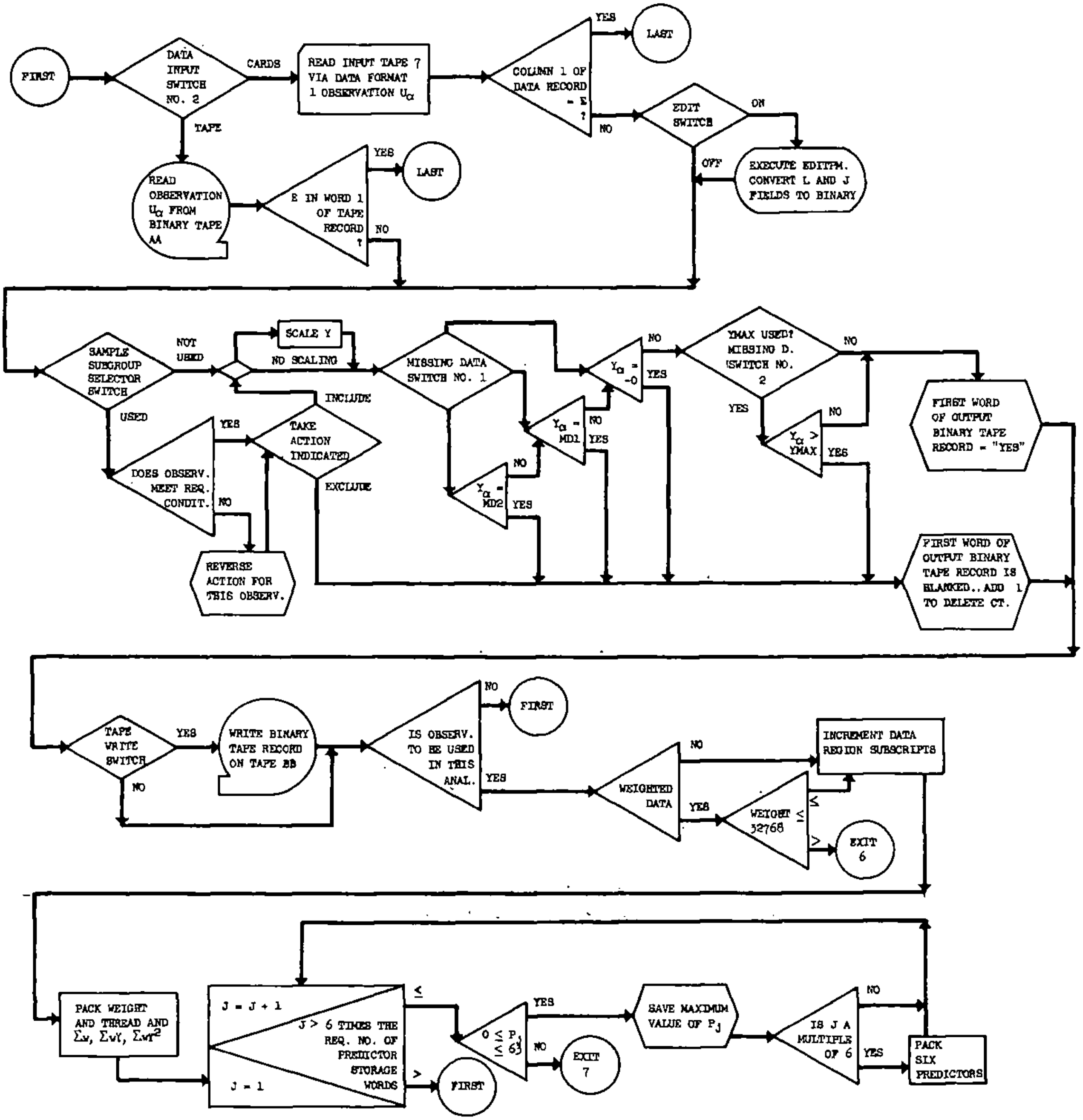
Otherwise group i is not capable of being split. No variable is "useful" over this group. The next most promising group (TSS_i = max) is selected.

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

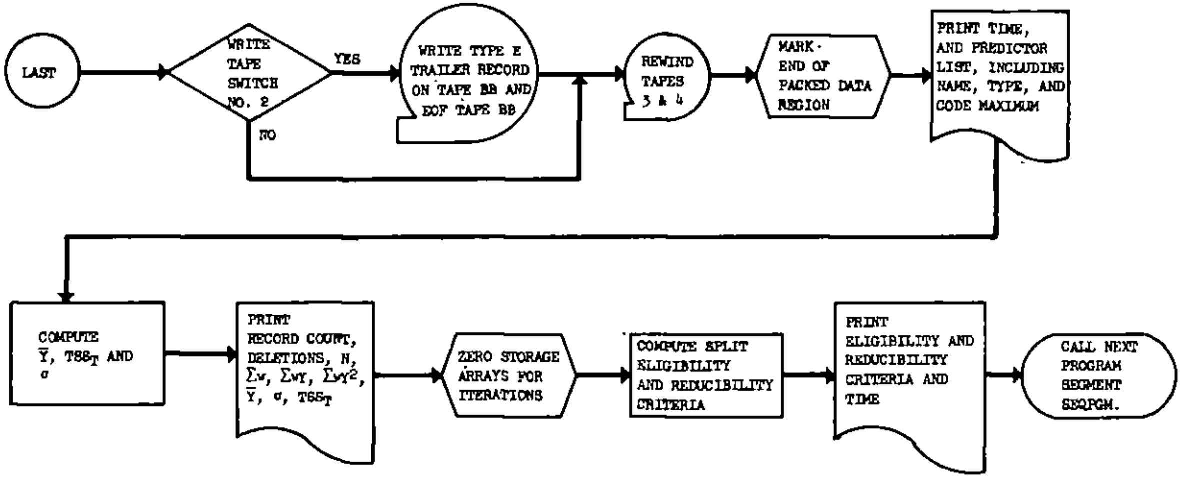
FLOW CHARTS
AID (2)
PROGRAM SEGMENT 1 (EDITOR)

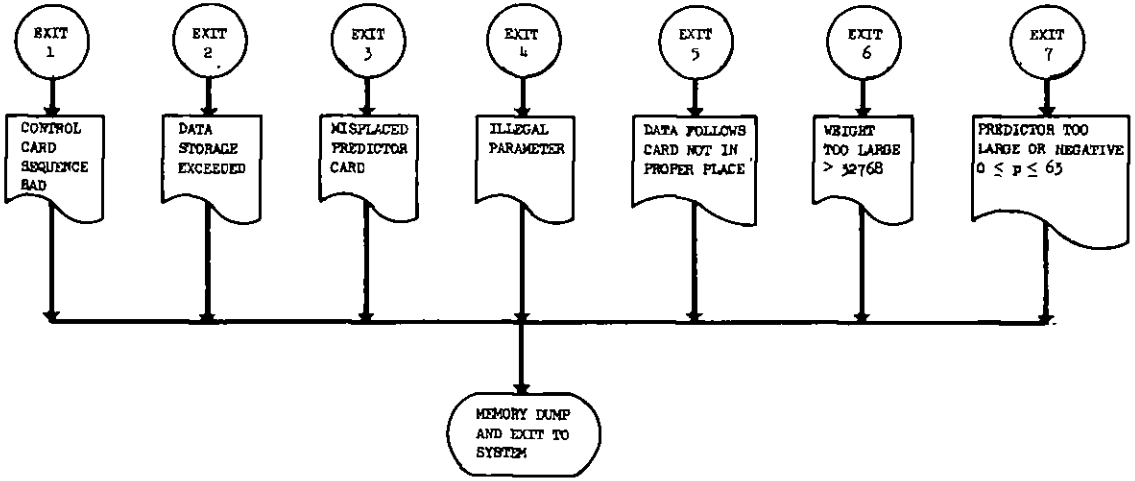


FLOW CHARTS
AID (2)
PROGRAM SEGMENT 1 (EDITOR)

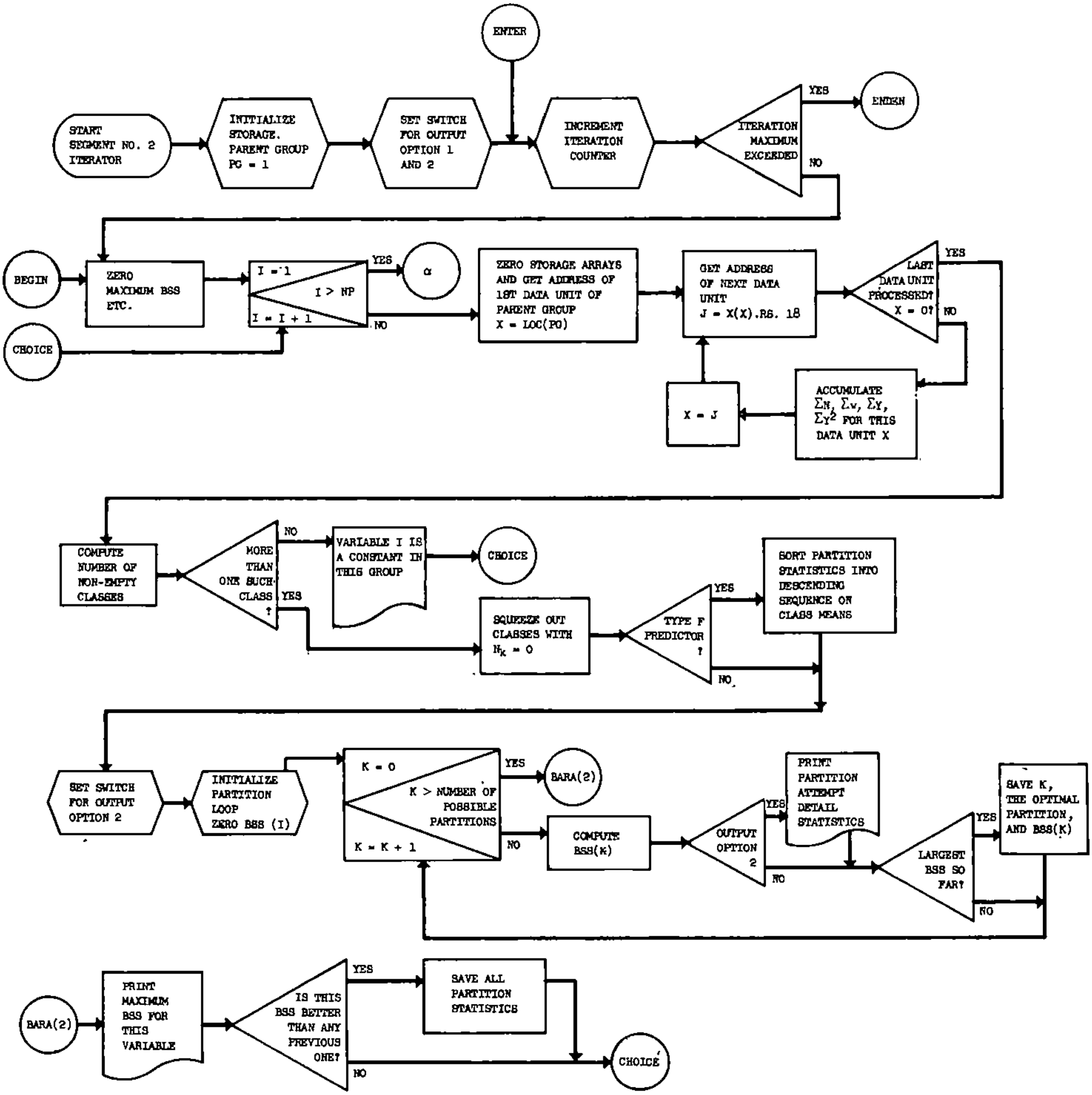


FLOW CHARTS
AID (2)
PROGRAM SEGMENT 1 (EDITOR)

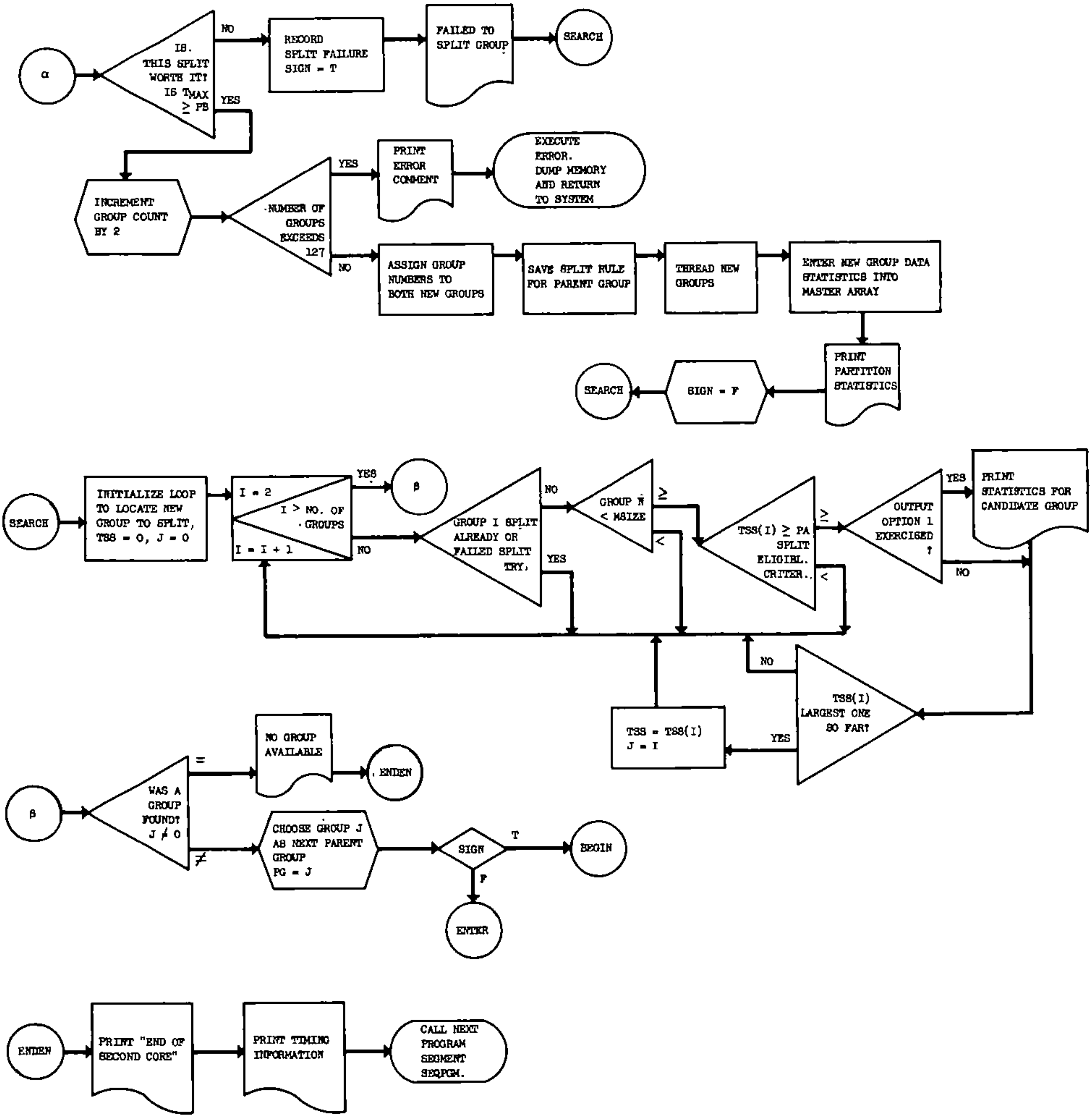




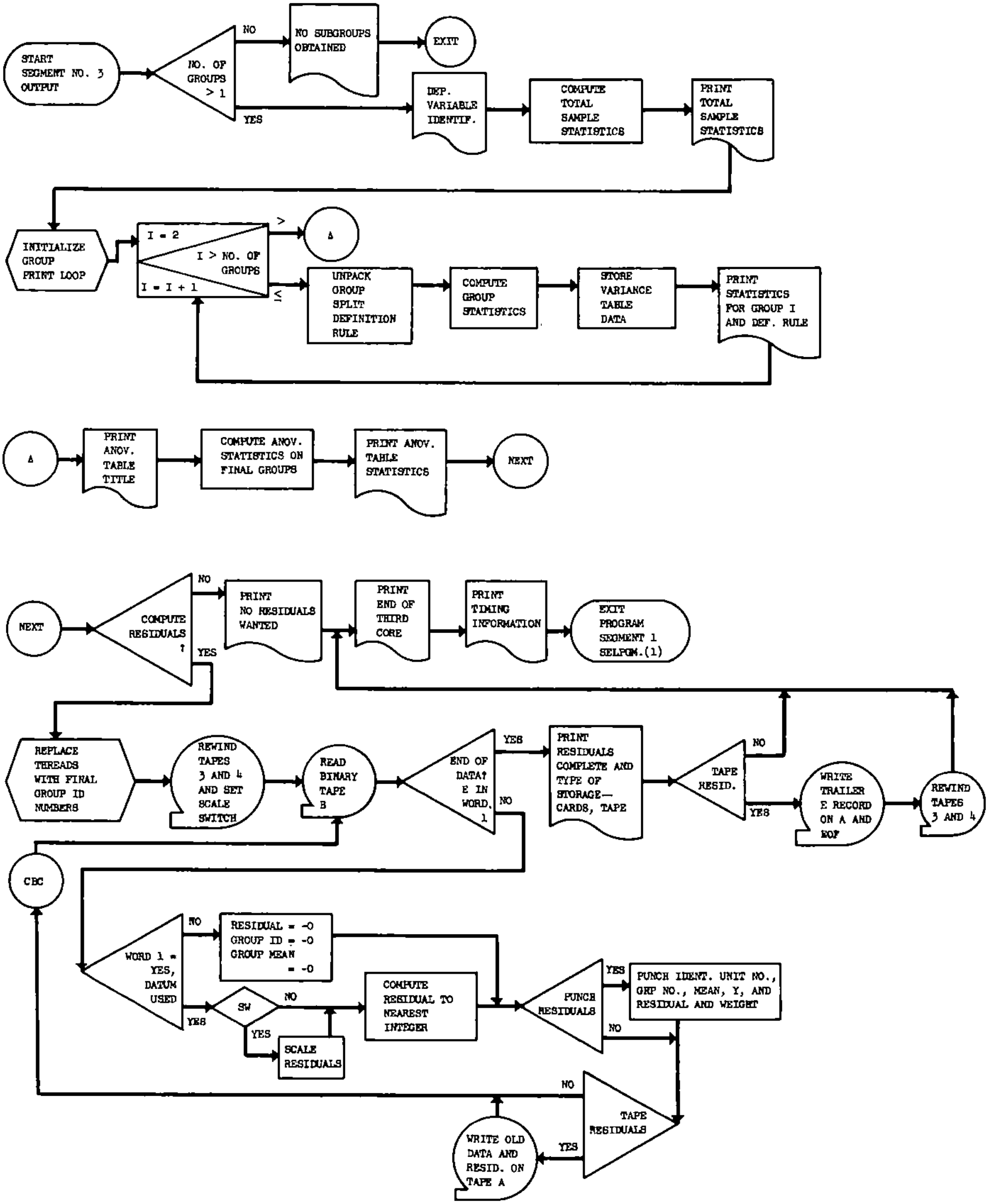
FLOW CHARTS
AID (2)
PROGRAM SEGMENT 2 (ITERATION)



FLOW CHARTS
AID (2)
PROGRAM SEGMENT 2 (ITERATION)



FLOW CHARTS
AID (2)
PROGRAM SEGMENT 3 (OUTPUT)



APPENDIX F

COMPUTER PROGRAM ARRAY STORAGE: AID (Model 2)

Array Name	Dimension	Function
ID	128	ID(I) contains the subscript in ID of the parent group from which group I was split.
INDEX	128	INDEX (I)ccontains the subscript of the input variable used on the parent group when creating group I.
HI	128	HI(I) contains the subscript of one of the groups created by splitting group I. This is the member of the pair with the (algebraically) largest mean.
LO	128	LO(I) contains the subscript of the other of the groups created by splitting group I. This is the member of the pair with the smaller mean.
TN = LAST	128	TN(I) contains the number of observations contained in group I.
TWT	128	TWT(I) contains the sum of weights for the observations in group I.
TY1	128	TY1(I) contains ΣY for the observations in group I.
TY2	128	TY2(I) contains ΣY^2 for the observations in group I.
FAIL = SIGN	128	FAIL (I) contains 0 if there has never been a failure to split group I given an attempted partition, it contains a 1 if an attempted partition has failed or if the group has already been split.
LOC	128	LOC(I) contains the subscript in D of the first observation in the threaded list comprising group I.
MEAN	128	MEAN(I) contains the mean Y for group i.

APPENDIX F--(CONTINUED)

Array Name	Dimension	Function
LIST = C	64	Temporary storage used during partitioning. LIST (I) contains the new group number to be assigned to all observations in the parent group for which the predictor used in the partition has the value appearing in the array KODE (I).
KODE	64	Temporary storage used during partitioning. KODE(I) contains the values of the predictor used in the partitioning of the current parent group, in the same order as the statistics (after sorting) in the partition scan arrays.
N	64	N(K) contains the number of observations in the k'th category of the predictor currently being used in an attempted partition.
W	64	W(K) contains the sum of weights for the observations in the k'th category of the predictor currently being used in an attempted partition.
Y1	64	Y1(K) contains the weighted Y of the observations in the k'th category of the predictor currently being used in an attempted partition.
Y2	64	Y2(K) contains the weighted ΣY^2 of the observations in the k'th category of the predictor currently being used in an attempted partition.
YBAR	64	YBAR(K) contains the weighted mean of the observations in the k'th category of the predictor currently being used in an attempted partition.
BSS	64	BSS(K) contains the between-group sum of squares for an analysis of variance performed by combining the ordered classes 0, 1, ..., k into one group and the classes (k+1) ... PMAX, into another group for the predictor currently being used in an attempted partition.

The following arrays correspond exactly to those described above, except they contain the statistics for the best available predictor for partitioning the group under consideration.

(BSSP,BSS), (CODE,KODE), (N,N1), (W,W1), (Y1,Y3), (Y2,Y4), (YBAR,Y5)

Array Name	Dimension	Function
P	36	Subscripts of predictor variables.
NAME1	36	First word of the alphanumeric name of that predictor
NAME2	36	Second word of the alphanumeric name of that predictor
TYPE	36	Type of predictor, monotonic or free.
LAB	12	Alphanumeric run identification.
CLASS	256	Contains partition rule split identification codes. The region is divided up into 128 blocks of two words. The 36 bits in each word are arbitrarily identified as follows.

For example:

0	0	0	0 0	1	1	0	1	0	Word 1
35	34	33	. . . .	4	3	2	1	0	[CLASS(9)]

0 0	0	0	0	0 0	0	0	0	Word 2
	63	62	61	. . . .	38	37	36	[CLASS(10)]

Each pair of words contains information identifying the values of the partition variable used in assigning observations into the group with which that pair of words is associated. If group 5 is created via a split of group 3 such that all observations in group 5 have the values 4 or 1 or 3 on predictor X_p used in the split, the words 9 and 10 in the Array CLASS would look as illustrated.

<table>
<tr><th>Array Name</th><th>Dimension</th><th>Function</th></tr>
<tr><td>D = X</td><td>20,000
Optional--depends on number of predictors</td><td>Contains data, including up to 36 predictors, the dependent variable and a weight to be used in the analysis. The structure is as follows:
<table>
<tr><th>Word</th><th colspan="6">Structure</th><th>Mode</th></tr>
<tr><td></td><td>Prefix</td><td colspan="2">Decrement</td><td>Tag</td><td colspan="2">Address</td><td></td></tr>
<tr><td>1</td><td>0</td><td colspan="2">Thread = subscript of 1st word of next observation in D which belongs to this group.
Thread = 0 if last observation in group</td><td>0</td><td colspan="2">Weight</td><td>Packed integer</td></tr>
<tr><td>2</td><td colspan="6">Y</td><td>Floating point</td></tr>
<tr><td>3</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>6</td><td>Packed logical</td></tr>
<tr><td>4</td><td>11</td><td>10</td><td>9</td><td>8</td><td>7</td><td>12</td><td>"</td></tr>
<tr><td>5</td><td>17</td><td>16</td><td>15</td><td>14</td><td>13</td><td>18</td><td>"</td></tr>
<tr><td>6</td><td>23</td><td>22</td><td>21</td><td>20</td><td>19</td><td>24</td><td>"</td></tr>
<tr><td>7</td><td>29</td><td>28</td><td>27</td><td>26</td><td>25</td><td>30</td><td>"</td></tr>
<tr><td>8</td><td>35</td><td>34</td><td>33</td><td>32</td><td>31</td><td>36</td><td>"</td></tr>
<tr><td></td><td>36--31</td><td>30--25</td><td>24--19</td><td>18--13</td><td>12--7</td><td>6--1</td><td></td></tr>
</table>
Up to 36 predictors (6 bits each) are packed in up to 6 words of storage. Thus, each observation takes up to 8 words of storage. Observations are stored sequentially in (X, D), starting at D(1).</td></tr>
<tr><td>V</td><td>100</td><td>Input Vector of data. Contains values of variables 1, 2, ..., (NV-1), NV, for observation U_α.</td></tr>
</table>

APPENDIX G

Program Write-Up
Institute for Social Research
The University of Michigan
IBM 7090

Program: Function IRFORM

Programmer: T. C. O'Brien

Source Language: UMAP

Date: November 1963

Function:

Much of the data collected by the Institute for Social Research prior to 1959 contain codes not easily capable of being handled by the format statements available in the MAD and FORTRAN II programming languages. Moreover, it is desirable for general purpose library programs to have data-input formats read in at execution time, rather than compiled with the program. It is also desirable for a dictionary of the locations of the input variables on the cards printed out for ease in interpretation of the output, and for ready checking of the correctness of the format statement read in.

This subroutine may be incorporated into any MAD, FORTRAN II or UMAP program compiled or assembled by the translators in the U. of M. Executive System for the IBM 7090. It accomplishes the following:

1. Reads and edits format information punched in columns 1-72 on a series of cards, either MAD or FORTRAN specifiers.
2. Prints out a dictionary of the locations of the variables on the input cards.
3. Supplies the edited format information to the calling program.
4. Edits subsequent data which is read in by the main program, employing several BCD to BINARY conversion schemes not easily available in FORTRAN or MAD.

Input:

Format information presented to IRFORM _must_ be enclosed in parentheses regardless of whether it consists of MAD field specifiers or FORTRAN specifiers. It must be punched in columns 1-72 of any number of consecutive cards. The word "FORMAT" is _not_ punched on the cards, nor are any continuation marks used. Thus, MAD format information is punched (......*).

The following list of FORTRAN field specifiers are permissible:

I, F, A, H, X, E, O

The following FORTRAN operators and symbols are permissible:

P , (/) . + - blank

The following MAD field specifiers are permissible:

I, F, C, H, S, E, K

The following MAD operators and symbols are permissible:

P , (/) . + - * blank

Any legal IBM character may occur inside an H string.

These characters and operators, when used in their proper form will be supplied to the program, with the following restrictions:

1. Parentheses may not be nested inside the format statement.
2. NO FIELD except an H string or a series of skip X (FORTRAN) or skip S (MAD) fields may be more than six (6) columns in width.

Several new field specifiers have been established. They have meaning similar to the ones above, except that BCD to BINARY conversion takes place in subroutine IRFORM instead of in the standard system input-output subroutines. When a special field indicator is read in, it is replaced as follows and a conversion switch is set.

Format:	Stored as:	
nLw	nAw	BCD to 12-base integer
nTw	nAw	BCD to 12-base floating point integer
nJw	nAw	BCD to 10-base integer
nGw.d	nAwbb	BCD to 10-base floating point number

(Where b = blank, d = decimal places, n is the field repetition operator and w is the field width in columns.)

Scale factors (P indicators) may not be used with G field specifiers.

The purpose of establishing these format field descriptors is two-fold.

1. The L and T specifiers permit single-punches in rows 12 (+) and 11 (-) of the IBM card to be read into the machine and used as integers. They may be stored in the machine in integer mode, or in floating point mode.
2. The J and G specifiers permit program control of punching patterns in fields which, though legal alphanumeric patterns, result in an I/O dump when read in through I and F field specifications.

In the L and T BCD to BINARY conversion rule, single punched columns in a field have the following internal machine values:

Card	Value
0 - 9	0 - 9
+	10
-	11
blank	-0

Thus, the value of a field read in and converted through an L or T specifier may be represented internally as follows:

$$a_1\ (12)^{n-1} + a_2(12)^{n-2} \ldots + a_{n-1}(12)^1 + a_n(12)^0$$

> where a_1, a_2, ... a_n are the internal machine values of the symbols as defined above and n is the number of columns in the field. Note: $(12)^0 = 1$. Thus, a two-column field goes into the machine as a positive integer in the range from 0 to 143. Note that all variables entering this way _must_ be positive integers, since the symbols normally used to differentiate between positive and negative numbers _are now part of the number system itself_. A table of conversions for two-column codes is appended.

In the J and G BCD to BINARY conversion rules, single-punched columns in a field have the following internal machine values:

Card	Value
0 - 9	0 - 9
+	-0
-	-0
blank	-0

The only difference between the special field designator J and the standard MAD or FORTRAN Integer (I) designator is in the treatment of nonnumeric characters when they appear on the data cards. The same is true of the G designator and the MAD or FORTRAN (F) field designators. When a J or G specifier is used, then all character patterns which are illegal in the corresponding I or F field are reduced to the value -0, rather than causing a halt of the computer, followed by an I/O dump. A flag is then set, which can be interrogated by the main program.

The following rules apply:

1. All data fields read through standard FORTRAN or MAD field specifications are not edited by IRFORM. The FORTRAN and MAD manuals describe what punching patterns may legally appear in these types of fields.

2. For the fields, L, T, J, and G, all illegal character patterns read from data cards result in the internal machine value of the field -0.

3. For the fields, L, T, J, and G, all characters other than +, -, 0-9 or blank are illegal, except for decimal points read in through a G field.

4. In addition, for L and T fields, any field containing a blank is illegal. All combinations of the set of characters [0 1 2 3 4 5 6 7 8 9 + -] are legal.

5. For J and G fields, the only legal character patterns are

   [blank(s)] followed by [at most one sign (+ or -)], followed by at least one digit 0123456789 and continuing with digits to the right end of the field. Items in brackets [ ] are optional. In a G field, a decimal point may appear immediately to the left or to the right of any digit, and take precedence over the number of decimals specified in the format description. An all blank field is illegal.

6. If at least one illegal character pattern is detected when the data are read in through the format statement, a signal is returned to the calling program.

To summarize:

The purpose of the L and T fields is to provide a means of converting nonblank fields containing patterns of single-punches into 12-base integers, under program control, rather than under I/O subroutine control. The purpose of J and G fields is to convert signed or unsigned numbers into decimal numbers in integer or floating point form, providing for the conversion of character patterns in these fields, which are <u>not</u> signed numbers (usually strings of characters, e.g., (--- or +++, etc.) into a representation usable by the computer, under program control, rather than under I/O subroutine control, since these are illegal and would cause an I/O dump.

Calling Sequences:

FORTRAN II

```
J = IRFORM (FMT, LEN, ISTART, IEND, EDIT, IDLEN)
```

MAD

```
J = IRFORM.(FMT, LEN, ISTART, IEND, EDIT, IDLEN)
```

where

FMT is the name of the first element of the vector in which the format statement is to be stored. (Mode is Floating in FORTRAN, Integer in MAD)

LEN is the dimensioned length of the array FMT in the calling program. (Integer mode)

ISTART is the number (subscript) to be printed out in the dictionary for the first field read by the format statement. (Integer mode)

IEND is the number (subscript) to be printed out in the dictionary for the last field read by the format statement. (Integer mode)

EDIT is the name of the first element of a one-dimensioned array or vector in the calling program in which format conversion codes will be stored by IRFORM. (Integer in MAD, Floating in FORTRAN)

IDLEN is the dimensioned length of the array EDIT. It should be one element longer than the array X used by the calling program to read in data. (Integer mode)

UMAP examples:

```
                CALL IRFORM               CALL EDITPM
                TXH  FMT                  TXH  X
                TXH  = 200                TXH  = 500
                TXH  = 0                  TXH  EDIT
                TXH  = 50
                TXH  = EDIT
                TXH  = 501
                 o
                 o
            FMT BES  200
           EDIT BES  501
                 o
                 o
```

Subroutine Functioning.

A call to IRFORM causes card images to be read from input tape 7 and the format statement is scanned, edited for special fields and stored away. Special fields are detected, converted to A (character) fields and an entry is made in the vector EDIT for each special field encountered. An exit flag (J) is returned to the calling program. If at least one special field is encountered, J = 1. J J = 0 if no special fields were encountered. Card reading and scanning continues until either a zero-level of parentheses (as many left as right) has occurred, or until the dimensioned length of FMT is about to be exceeded, or until the dimensioned length of EDIT is about to be exceeded. If either of the latter conditions occurs, the program cannot continue and a memory dump is initiated. J is returned as a FORTRAN or MAD integer depending on the language of the calling program. No variable which is more than six columns wide may be read in through an L, T, J, or G field specifier. Otherwise a dump will result.

The structure of the edit list is as follows: The first word in the list contains the number of elements which follow in the list.

Prefix	Decrement	Tag	Address
0	i	w	c

where i is the index of the variable to be edited in the input region defined by the calling program, w is the field width, and c is a conversion code. The conversion code specifies the input mode, the type of conversion desired and, for G fields, the number of (implied) decimal places in the field. The edit list is used by means of another entry into IRFORM which is executed after each vector of data is read into the computer using FMT. The function value J as defined above may be used to determine whether or not any special formats were read in, thus requiring an edit on each vector of data. Thus, conversion of data values is accomplished by the following calling sequences:

FORTRAN II

```
    L = EDITPM (X, LEN, EDIT)
```

MAD

```
    L = EDITPM. (X, LEN, EDIT)
```

In each case X is the appropriate MAD or FORTRAN base address of an input vector with length dimensioned at LEN and EDIT is defined as above. X may be integer or floating mode, LEN is integer, and EDIT is floating mode in FORTRAN and integer mode in MAD. Execution of this statement causes the necessary transformations to be made on those variables listed as requiring them in the edit list. The results are placed back in the corresponding positions in the X array. If an illegal field has been detected, the value of L is nonzero, otherwise it is zero. L is reset each time EDITPM is executed. L is an integer of the appropriate form returned to the calling program.

A typical FORTRAN code sequence might be as follows:

```
    DIMENSION X(100), EDIT (101), FMT (108)
    NX = 50
    J = IRFORM (FMT, 108, 0, NX, EDIT, 101
1   READ INPUT TAPE 7, FMT, CTYPE, (X(I), I = 1, NX)
    IF (J) 2,5,2
2   L = EDITPM (X, 100, EDIT)
3   IF (L) 4, 5, 4
4         locate undefined value of field.
          Some X_i = -0 and take appropriate action
5   CONTINUE
          process data card
6   GO TO 1
```

L & T CONVERSION TABLE FOR TWO COLUMN FIELDS

LOW ORDER COLUMN

		0	1	2	3	4	5	6	7	8	9	+	-
HIGH ORDER COLUMN	0	00	01	02	03	04	05	06	07	08	09	10	11
	1	12	13	14	15	16	17	18	19	20	21	22	23
	2	24	25	26	27	28	29	30	31	32	33	34	35
	3	36	37	38	39	40	41	42	43	44	45	46	47
	4	48	49	50	51	52	53	54	55	56	57	58	59
	5	60	61	62	63	64	65	66	67	68	69	70	71
	6	72	73	74	75	76	77	78	79	80	81	82	83
	7	84	85	86	87	88	89	90	91	92	93	94	95
	8	96	97	98	99	100	101	102	103	104	105	106	107
	9	108	109	110	111	112	113	114	115	116	117	118	119
	+	120	121	122	123	124	125	126	127	128	129	130	131
	-	132	133	134	135	136	137	138	139	140	141	142	143

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

‡                 750031               DECK
‡HSIEH                         J113N               002   050   111       TEST  ‡
$EXECUTE, DUMP, I/ODUMP                                                 AID20001
$ COMPILE MAD,PRINT OBJECT,PUNCH OBJECT                         AIDM2001AID20002
          R                                                             AID20003
          R          PROGRAM  NAME  --  A I D.     FIRST CORE.          AID20004
          R          WRITTEN  BY  ROBERT  WENCHAO  HSIEH  ON  1/31/63.  AID20005
          R            AID  -  MODEL  2 -   REWRITTEN  ON  AUGUST 1963. AID20006
          R                                                             AID20007
           DIMENSION ID(128),INDEX(128),HI(128),LO(128),TN(128),TWT(128)AID20008
          1 ,TY1(128),TY2(128),CLASS(256,DIM),LOC(128),FAIL(128),MAX(36)AID20009
          2,P(36),FMT(107),NAME1(36),NAME2(36),LAB(12),V(100),EDITV(101)AID20010
          3  ,X(20000),D(20000) ,TYPE(36)                               AID20011
          R                                                             AID20012
           BOOLEAN  SIGN, BIN, BOB, A, B, EZ, TK, RUN,CA,QA,T1,T2       AID20013
          1 ,MA,MB,YA,SFA,SFB                                           AID20014
          R                                                             AID20015
           PROGRAM COMMON NAME1,NAME2, NP,NV, LAB,                      AID20016
          0               ID,INDEX,HI,LO,TN,TWT,TY1,TY2,CLASS,LOC,FAIL, AID20017
          1 MAX,P,NOGP,ITR,ITRMAX,PA,PB,OP1,OP2,X,MSIZE,SCFIN,SCFOUT,   AID20018
          2  KONST,AA,BB,RUN,ZWANT,ZTYPE,ZTAPE,BOB,TYPE                 AID20019
          R                                                             AID20020
           FLOATING POINT TWT,TY1,TY2,PA,MEAN,WEIGHT,SQRT.,TSS,D,SIGMAY AID20021
          1,PB,MD1,MD2,YMAX,P1,P2,FY                                    AID20022
          R                                                             AID20023
           EQUIVALENCE(X,D), (V,INDEX(1)),(EDITV,HI(1)),(ZTAPE,EZ),     AID20024
          1 (TR,A),(KT,TK),(ZWANT,B),(AC,CA),(AQ,QA),(FILT1,T1),        AID20025
          2 (FILT2,T2),(SCFIN,SFA),(SCFOUT,SFB),(MD1,MA),(MD2,MB),      AID20026
          3 (YMAX,YA)                                                   AID20027
          R                                                             AID20028
           NORMAL MODE IS INTEGER                                       AID20029
          R                                                             AID20030
           WHENEVER BOB                                                 AID20031
           PRINT COMMENT$1                        *    (A)UTOMATIC  (I)NAID20032
          1TERACTION  (D)ETECTOR    --    MODEL  2.   *$                AID20033
           PRINT COMMENT$0                                  WRITTEN   INAID20034
          1 MAD  BY  ROBERT   W.  HSIEH    -    AUGUST  1963.$          AID20035
           END OF CONDITIONAL                                           AID20036
           BOB=0B                                                       AID20037
          R                                                             AID20038
           REWIND TAPE 4                                                AID20039
           REWIND TAPE 3                                                AID20040
          R          READ  A LABEL  --  TYPE 1.                         AID20041
           READ FORMAT CARD1,TYPE,LAB(0)...LAB(12)                      AID20042
           WHENEVER TYPE .NE. $1$,TRANSFER TO EXIT1                     AID20043
          R          READ  A  PARAMETER  CARD   --  TYPE 2.             AID20044
           READ FORMAT CARD2, TYPE,LOCDAT,N,NV,IXCLUD,INOUT1,LOW1,HIGH1,AID20045
          1FILT1,CONJ,INOUT2,LOW2,HIGH2,FILT2                           AID20046
           WHENEVER TYPE .NE. $2$,TRANSFER TO EXIT1                     AID20047
          R          READ  A  PARAMETER  CARD   --  TYPE 3.             AID20048
           READ FORMAT CARD3,TYPE,NP,WT,P1,P2,ITRMAX,MSIZE,Y, NAME1,    AID20049
          1 NAME2,YMAX,MD1,MD2,CORES,TPRES,INTNO,SCFIN,SCFOUT           AID20050
           WHENEVER TYPE .NE. $3$,TRANSFER TO EXIT1                     AID20051
          R                                                             AID20052
          R          CHECK  CORE  STORAGE.                              AID20053
          R                                                             AID20054
           IX=NP/6                                                      AID20055
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           WHENEVER (NP-IX*6) .NE. 0, IX=IX+1                           AID20056
           P=IX*6                                                       AID20057
           KONST=IX+2                                                   AID20058
           WHENEVER(N*KONST) .G. 20000,TRANSFER TO EXIT2                AID20059
          R                                                             AID20060
          R          SET  SWITCHES  FOR  MISSING  DATA.                 AID20061
          R                                                             AID20062
           WHENEVER CAS.(MD2,-0.) .NE. 0                                AID20063
           NOMD=3                                                       AID20064
           OTHERWISE                                                    AID20065
             WHENEVER CAS.(MD1,-0.) .NE. 0                              AID20066
             NOMD=2                                                     AID20067
             OTHERWISE                                                  AID20068
             NOMD=1                                                     AID20069
             END OF CONDITIONAL                                         AID20070
           END OF CONDITIONAL                                           AID20071
           QZ=0                                                         AID20072
           WHENEVER CAS.(YMAX,-0.) .E. 0, QZ=1                          AID20073
          R                                                             AID20074
          R             SET  SCALE  FACTOR  SWITCH.                     AID20075
          R                                                             AID20076
           FY=1.                                                        AID20077
           SQ=1                                                         AID20078
           WHENEVER SFA                                                 AID20079
           SQ=0                                                         AID20080
           FY=10.0.P.SCFIN                                              AID20081
           END OF CONDITIONAL                                           AID20082
          R             SCALE   MD1, MD2, AND YMAX                      AID20083
           TRANSFER TO PASS(SQ)                                         AID20084
PASS(0)    WHENEVER       MB, MD2=MD2*FY                                AID20085
           WHENEVER       MA, MD1=MD1*FY                                AID20086
           WHENEVER       YA,  YMAX=YMAX*FY                             AID20087
          R                                                             AID20088
PASS(1)    OP1=1                                                        AID20089
           OP2=1                                                        AID20090
           ID=Y                                                         AID20091
           INDEX=WT                                                     AID20092
          R                                                             AID20093
           PRINT FORMAT HEAD, LAB(0)...LAB(12),              N,NV,NP,WT,AID20094
          1P1,P2,ITRMAX,Y,NAME1,NAME2,YMAX                              AID20095
           I=NP/4                                                       AID20096
           WHENEVER (NP-I*4) .NE. 0, I=I+1                              AID20097
          R                                                             AID20098
          R          READ  PREDICTORS, TYPES, AND NAMES  --  TYPE 4.    AID20099
          R                                                             AID20100
           THROUGH FEED,FOR J=1,1,J .G. I                               AID20101
           K=(J-1)*4+1                                                  AID20102
           L=K+3                                                        AID20103
           READ FORMAT CARD4,TYPE,(M=K,1,M.G.L,P(M),TYPE(M),NAME1(M),   AID20104
          1NAME2(M))                                                    AID20105
           WHENEVER TYPE .NE. $4$,TRANSFER TO EXIT3                     AID20106
FEED       CONTINUE                                                     AID20107
          R                                                             AID20108
          R          SET  SWITCHES  FOR  INPUT  DATA.                   AID20109
          R                                                             AID20110
           TR=0                                                         AID20111
           WHENEVER LOCDAT .E. $W$, TR=1                                AID20112
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           KT=0                                                         AID20113
           WHENEVER LOCDAT .E. $T$,KT=1                                 AID20114
          R                                                             AID20115
          R          SET   SWITCHES   FOR   RESIDUAL   OUTPUT.          AID20116
          R                                                             AID20117
           ZWANT=0                                                      AID20118
           ZTYPE=0                                                      AID20119
           ZTAPE=0                                                      AID20120
           WHENEVER CDRES .NE. $      $                                 AID20121
           ZWANT=1                                                      AID20122
             WHENEVER TPRES .NE. $      $                               AID20123
             ZTYPE=2                                                    AID20124
             ZTAPE=1                                                    AID20125
             END OF CONDITIONAL                                         AID20126
           OTHERWISE                                                    AID20127
             WHENEVER TPRES .NE. $      $                               AID20128
             ZWANT=1                                                    AID20129
             ZTYPE=1                                                    AID20130
             ZTAPE=1                                                    AID20131
             END OF CONDITIONAL                                         AID20132
           END OF CONDITIONAL                                           AID20133
          R          SET   TAPE   WRITE   SWITCH                        AID20134
           TW=0                                                         AID20135
           WHENEVER A .OR. B, TW=1                                      AID20136
          R                                                             AID20137
           PRINT FORMAT HEAD1, MD1,MD2,OP1,OP2,MSIZE,DATA(KT)           AID20138
           PRINT FORMAT HEAD3, BAKA(ZWANT),OUT(ZWANT+ZTYPE)             AID20139
          R                                                             AID20140
          R        TAPE   NUMBER   TO   BE   ASSIGNED.   ****           AID20141
          R                                                             AID20142
           WHENEVER .NOT. TK                                            AID20143
           RUN=0B                                                       AID20144
           BB=4                                                         AID20145
           AA=3                                                         AID20146
             WHENEVER EZ , RUN=1B                                       AID20147
           OTHERWISE                                                    AID20148
           AA=4                                                         AID20149
           BB=3                                                         AID20150
             WHENEVER RUN                                               AID20151
           AA=3                                                         AID20152
           BB=4                                                         AID20153
             END OF CONDITIONAL                                         AID20154
           END OF CONDITIONAL                                           AID20155
          R                                                             AID20156
          R        ASSIGN   SWITCHES   FOR   FILTER.   ****             AID20157
          R                                                             AID20158
           AQ=1                                                         AID20159
           WHENEVER IXCLUD .E. $      $                                 AID20160
           AQ=0                                                         AID20161
           OR WHENEVER IXCLUD .E. $INCLUD$                              AID20162
           MQ=1                                                         AID20163
           OTHERWISE                                                    AID20164
           MQ=0                                                         AID20165
           END OF CONDITIONAL                                           AID20166
           TRANSFER TO INTO(AQ)                                         AID20167
INTO(1)    PRINT FORMAT HEAD2, IXCLUD,INOUT1,LOW1,HIGH1,FILT1,          AID20168
          1CONJ,INOUT2,LOW2,HIGH2,FILT2                                 AID20169
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           TRANSFER TO SPACE                                            AID20170
          R                                                             AID20171
INTO(0)    PRINT COMMENT$0    NO  FILTERS.$                             AID20172
SPACE      TRANSFER TO XR(AQ)                                           AID20173
XR(1)      WHENEVER CONJ .E. $      $                                   AID20174
           AC=0                                                         AID20175
           OR WHENEVER CONJ .E. $   AND$                                AID20176
           AC=2                                                         AID20177
           OTHERWISE                                                    AID20178
           AC=1                                                         AID20179
           END OF CONDITIONAL                                           AID20180
           AIN=0                                                        AID20181
           WHENEVER INOUT1 .NE. $    IN$,AIN=1                          AID20182
           BIN=1B                                                       AID20183
           WHENEVER INOUT2 .NE. $    IN$, BIN=0B                        AID20184
           WHENEVER QA                                                  AID20185
             WHENEVER .NOT. T1, TRANSFER TO EXIT                        AID20186
            WHENEVER CA                                                 AID20187
             WHENEVER .NOT. T2, TRANSFER TO EXIT                        AID20188
             END OF CONDITIONAL                                         AID20189
           END OF CONDITIONAL                                           AID20190
          R                                                             AID20191
          R    VERIFY DIMENSIONS ON ALL VARIABLES.                      AID20192
          R                                                             AID20193
XR(0)      WHENEVER NV .L. 2 .OR. NV .G. 100,TRANSFER TO EXIT4          AID20194
           WHENEVER NP .L. 1 .OR. NP .G. 36, TRANSFER TO EXIT4          AID20195
           WHENEVER ITRMAX .G.  63,TRANSFER TO EXIT4                    AID20196
           THROUGH BAR, FOR J=1,1, J .G. NP                             AID20197
BAR        WHENEVER P(J) .L. 1 .OR. P(J) .G.  NV, TRANSFER TO EXIT4     AID20198
           PRINT RESULTS SCFIN,  SCFOUT                                 AID20199
          R          SET  WEIGHT  SWITCH.                               AID20200
           WHENEVER WT .E. 0                                            AID20201
           R=2                                                          AID20202
           WHT=1                                                        AID20203
           OTHERWISE                                                    AID20204
           WHENEVER WT .G.  NV, TRANSFER TO EXIT4                       AID20205
           R=1                                                          AID20206
           END OF CONDITIONAL                                           AID20207
          R                                                             AID20208
           EXECUTE ZERO.(MAX(1)...MAX( P))                              AID20209
           TN=0                                                         AID20210
           TWT=0.0                                                      AID20211
           TY1=0.0                                                      AID20212
           TY2=0.0                                                      AID20213
           THROUGH ONE, FOR J=NP+1, 1, J .G. P                          AID20214
ONE        P(J)=0                                                       AID20215
          R                                                             AID20216
          R          READ  DATA  FORMAT  -  IF  KT=0.                   AID20217
          R                                                             AID20218
           TRANSFER TO ENTER(KT)                                        AID20219
ENTER(0)   THROUGH CLEAN, FOR J=1,1,J .G. 107                           AID20220
CLEAN      FMT(J)=$      .$                                             AID20221
           NZ=NV+1                                                      AID20222
           L=IRFORM.(FMT,107,0,NV,EDITV,NZ)                             AID20223
          R          SWITCH  FOR  EDITPM.                               AID20224
           WHENEVER L .NE. 0                                            AID20225
           Q=1                                                          AID20226
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           OTHERWISE                                                    AID20227
           Q=0                                                          AID20228
           END OF CONDITIONAL                                           AID20229
          R                                                             AID20230
           READ FORMAT CHECK,CARD                                       AID20231
           WHENEVER CARD .NE. $DATAFO$,TRANSFER TO EXIT5                AID20232
           PRINT COMMENT $4    INPUT-DATA FORMAT AS FOLLOWS.$           AID20233
           PRINT FORMAT FMTIN,FMT(0)...FMT(107)                         AID20234
          R                                                             AID20235
          R          SET  INPUT  COUNTER  AND  DELETION  COUNTER =0     AID20236
          R                                                             AID20237
ENTER(1)   CC=0                                                         AID20238
           CD=0                                                         AID20239
           N1=-KONST+1                                                  AID20240
           PRINT COMMENT $0    READ  DATA  BEGINS.$                     AID20241
           EXECUTE WRATIM.(0)                                           AID20242
          R                                                             AID20243
          R          DATA  INPUT  SWITCH                                AID20244
          R                                                             AID20245
FIRST      TRANSFER TO ENTRY(KT)                                        AID20246
          R          READ  INPUT  --  CARDS.     KT=0                   AID20247
ENTRY(0)   READ FORMAT FMT,KARD,V(1)...V(NV)                            AID20248
           WHENEVER KARD .E. $E$, TRANSFER TO LAST                      AID20249
           TRANSFER TO EDIT(Q)                                          AID20250
EDIT(1)    EXECUTE EDITPM.(V(1),NV,EDITV)                               AID20251
           TRANSFER TO EDIT(0)                                          AID20252
          R          READ  INPUT  --  TAPE.     KT=1                    AID20253
ENTRY(1)   READ BINARY TAPE  AA ,KARD,V(1)...V(NV)                      AID20254
           WHENEVER KARD .E. $E$,TRANSFER TO LAST                       AID20255
          R          NUMBER  OF  DATA  BEING  READ.                     AID20256
EDIT(0)    CC=CC+1                                                      AID20257
           TRANSFER TO TRA(AQ)                                          AID20258
          R                                                             AID20259
          R             FILTER  PROCESS FOLLOWS.                        AID20260
          R                                                             AID20261
TRA(1)     SIGN=0B                                                      AID20262
           V1=V(FILT1)                                                  AID20263
           V2=V(FILT2)                                                  AID20264
           TRANSFER TO JON(AIN)                                         AID20265
JON(0)     WHENEVER    V1 .GE. LOW1 .AND.    V1 .LE. HIGH1, SIGN=1B     AID20266
           TRANSFER TO DAVID                                            AID20267
JON(1)     WHENEVER    V1 .L. LOW1 .OR.    V1 .G. HIGH1,SIGN=1B         AID20268
DAVID      WHENEVER SIGN                                                AID20269
             WHENEVER AC .L. 2,TRANSFER TO OK(MQ)                       AID20270
DROPIN       WHENEVER BIN                                               AID20271
               WHENEVER V2.GE.LOW2 .AND. V2.LE. HIGH2,TRANSFER TO OK(MQ)AID20272
             OTHERWISE                                                  AID20273
               WHENEVER   V2.L.LOW2.OR.    V2.G.HIGH2,TRANSFER TO OK(MQ)AID20274
             END OF CONDITIONAL                                         AID20275
             TRANSFER TO NOP(MQ)                                        AID20276
           OTHERWISE                                                    AID20277
             WHENEVER AC .NE. 1, TRANSFER TO NOP(MQ)                    AID20278
             TRANSFER TO DROPIN                                         AID20279
           END OF CONDITIONAL                                           AID20280
OK(0)      TRANSFER TO BAD                                              AID20281
OK(1)      TRANSFER TO GOOD                                             AID20282
NOP(0)     TRANSFER TO GOOD                                             AID20283
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

NOP(1)     TRANSFER TO BAD                                              AID20284
TRA(0)     CONTINUE                                                     AID20285
GOOD       SIGN=0B                                                      AID20286
           D=V(Y)                                                       AID20287
           TRANSFER TO HANG(SQ)                                         AID20288
HANG(0)    D=D*FY                                                       AID20289
          R          CHECK  VALUE  OF  DEPENDENT  VARIABLE (Y).         AID20290
HANG(1)    TRANSFER TO SKIP(NOMD)                                       AID20291
SKIP(3)    WHENEVER CAS. (D,MD2).E.0 , TRANSFER TO BAD                  AID20292
SKIP(2)    WHENEVER CAS. (D,MD1).E.0 , TRANSFER TO BAD                  AID20293
SKIP(1)    WHENEVER CAS. (D,-0.).E.0 , TRANSFER TO BAD                  AID20294
           TRANSFER TO JUMP(QZ)                                         AID20295
JUMP(0)    WHENEVER D.G. YMAX, TRANSFER TO BAD                          AID20296
JUMP(1)    KARD=$YES$                                                   AID20297
           TRANSFER TO SKIP(0)                                          AID20298
          R          NUMBER  OF  DATA  BEING  DELETED.                  AID20299
BAD        CD=CD+1                                                      AID20300
           KARD=$ $                                                     AID20301
           SIGN=1B                                                      AID20302
          R          TAPE  WRITE  SWITCH.                               AID20303
SKIP(0)    TRANSFER TO WTAPE(TW)                                        AID20304
WTAPE(1)   WRITE BINARY TAPE  BB     , KARD, V(1)...V(NV)               AID20305
WTAPE(0)   WHENEVER SIGN, TRANSFER TO FIRST                             AID20306
          R          WEIGHT  SWITCH                                     AID20307
           TRANSFER TO ADD(R)                                           AID20308
ADD(1)     WHENEVER V(WT) .G. 32768,TRANSFER TO EXIT6                   AID20309
           WHT=V(WT)                                                    AID20310
          R          PUT  WEIGHT  AND  NUMBER  EACH  DATUM.             AID20311
ADD(2)     TN=TN + 1                                                    AID20312
           N1=N1+KONST                                                  AID20313
           N2=N1+KONST                                                  AID20314
           WHENEVER (N2-1) .G. 20000, TRANSFER TO EXIT2                 AID20315
           X(N1)=WHT .V.(N2 .LS. 18)                                    AID20316
           K=N1+1                                                       AID20317
           D(K)=D                                                       AID20318
           WEIGHT=WHT                                                   AID20319
           TWT=TWT+  WEIGHT                                             AID20320
           TY1=TY1+D*WEIGHT                                             AID20321
           TY2=TY2+D*D*WEIGHT                                           AID20322
          R                                                             AID20323
          R          PACK  DATA  INTO  CORE.                            AID20324
          R                                                             AID20325
           THROUGH PACK,FOR J=1,1,J .G. P                               AID20326
           V=V(P(J))                                                    AID20327
           WHENEVER V .L. 0 .OR. V .G. 63,TRANSFER TO EXIT7             AID20328
           WHENEVER V .G. MAX(J),MAX(J)=V                               AID20329
           WHENEVER(J-J/6*6) .NE. 0,TRANSFER TO PACK                    AID20330
           K=K+1                                                        AID20331
           X(K)=V(P(J)).V.(V(P(J-5)).LS.6).V.(V(P(J-4)).LS.12)          AID20332
          1 .V.(V(P(J-3)).LS.18).V.(V(P(J-2)).LS.24).V.(V(P(J-1)).LS.30)AID20333
PACK       CONTINUE                                                     AID20334
           TRANSFER TO FIRST                                            AID20335
          R                                                             AID20336
LAST       PRINT COMMENT $0      DATA  ARE  ALL  IN.$                   AID20337
          R          READ  INPUT  BEING  COMPLETED.                     AID20338
           TRANSFER TO REWIND(TW)                                       AID20339
          R                                                             AID20340
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          R          WRITE  TAPE  TRAILER,  E.O.F.  AND  REWIND  3,  4. AID20341
          R                                                             AID20342
REWIND(1)  WRITE BINARY TAPE  BB  ,KARD,V(1)...V(NV)                    AID20343
           END OF FILE TAPE   BB                                        AID20344
          R                                                             AID20345
REWIND(0)  REWIND TAPE  4                                               AID20346
           REWIND TAPE  3                                               AID20347
          R                                                             AID20348
           X(N1)=X(N1) .A. 77777K                                       AID20349
           EXECUTE WRATIM.(0)                                           AID20350
           PRINT COMMENT $1     * *    PREDICTOR    LISTING.     * * $  AID20351
          R                                                             AID20352
          R          PRINT  PREDICTOR  LIST  --  TYPE, CODE  MAX., ETC. AID20353
          R                                                             AID20354
           PRINT COMMENT $0     VARIABLE  NO.      DESCRIPTION      MAXIAID20355
          1MUM VALUE      T Y P E  $                                    AID20356
           THROUGH WXYZ,FOR J=1,1, J .G. NP                             AID20357
           PRINT FORMAT F1, P(J),NAME1(J),NAME2(J),MAX(J),TYPE(J)       AID20358
           WHENEVER TYPE(J) .E. $F$                                     AID20359
           TYPE(J)=0                                                    AID20360
           OTHERWISE                                                    AID20361
           TYPE(J)=1                                                    AID20362
WXYZ       END OF CONDITIONAL                                           AID20363
          R                                                             AID20364
          R          PRINT  AND  COMPUTE  BASIC  STATISTICS.            AID20365
          R                                                             AID20366
           PRINT COMMENT $2   * STATISTICS    FOR    TOTAL.$            AID20367
           MEAN=TY1/TWT                                                 AID20368
           TSS=TY2-TY1*MEAN                                             AID20369
           SIGMAY=SQRT.(TSS/TWT)                                        AID20370
           PRINT FORMAT F2,CC,CD,TN,TWT,TY1,TY2,MEAN,SIGMAY,TSS         AID20371
          R                                                             AID20372
          R        ZERO MASTER GROUP ARRAY.                             AID20373
          R                                                             AID20374
           THROUGH CLEAR, FOR I=1,1, I .G. 128                          AID20375
           ID(I)=0                                                      AID20376
           INDEX(I)=0                                                   AID20377
           LOC(I)=0                                                     AID20378
           HI(I)=0                                                      AID20379
           LO(I)=0                                                      AID20380
           TN(I)=0                                                      AID20381
           TWT(I)=0.0                                                   AID20382
           TY1(I)=0.0                                                   AID20383
           TY2(I)=0.0                                                   AID20384
CLEAR      FAIL(I)=0                                                    AID20385
           EXECUTE ZERO.(CLASS(1,1)...CLASS(128,2))                     AID20386
          R                                                             AID20387
           PA=P1*TSS                                                    AID20388
           PB=P2*TSS                                                    AID20389
           PRINT RESULTS PA,PB                                          AID20390
           HI=INTNO                                                     AID20391
           EXECUTE WRATIM.(0)                                           AID20392
          R                                                             AID20393
           EXECUTE  SEQPGM.(0)                                          AID20394
          R                                                             AID20395
          R     * *  ERROR  FLAGS.  **                                  AID20396
          R                                                             AID20397
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

EXIT1      PRINT COMMENT $0  *****  CONTROL CARD OUT OF ORDER.  ADIEU. $AID20398
           TRANSFER TO EXIT                                             AID20399
EXIT2      PRINT COMMENT $0  *****  DATA  STORAGE  EXCEEDED.    CIAO. $AID20400
           PRINT RESULTS N,KONST,NP,IX                                  AID20401
           TRANSFER TO EXIT                                             AID20402
EXIT3      PRINT COMMENT $0  *****  PREDICTOR CONTROL CARD MISPLACED.  $AID20403
           TRANSFER TO EXIT                                             AID20404
EXIT4      PRINT COMMENT $0  *****  SOME PARAMETER VALUES ARE SICK.    $AID20405
           TRANSFER TO EXIT                                             AID20406
EXIT5      PRINT COMMENT $0  *****  DATA FOLLOWS CARD IS MISSING. BYE. $AID20407
           TRANSFER TO EXIT                                             AID20408
EXIT6      PRINT COMMENT $0  *****  WEIGHT VARIABLE VALUE EXCEEDS 32768$AID20409
           TRANSFER TO EXIT                                             AID20410
EXIT7      PRINT COMMENT $0  *****  PREDICTOR VALUE EXCEEDS 63.        $AID20411
           PRINT RESULTS J, P(J), V(P(J)), V, TN                        AID20412
EXIT       EXECUTE ERROR.                                               AID20413
          R                                                             AID20414
          R       FORMAT  SPECIFICATIONS.       FIRST  CORE             AID20415
          R                                                             AID20416
           VECTOR VALUES DIM=2,1,2                                      AID20417
          R                                                             AID20418
           VECTOR VALUES CARD1=$C1,13C6*$                               AID20419
          R                                                             AID20420
           VECTOR VALUES CARD2=$C1,S5,C1,2I6,2C6,3I6,2C6,3I6*$          AID20421
          R                                                             AID20422
           VECTORVALUESCARD3=$C1,2I6,2F6.5,2I3,I6,2C6,3F6.0,2C3,I3,2I2*$AID20423
          R                                                             AID20424
           VECTOR VALUES CARD4=$C1,    4(S1,I3,S1,C1,S1,2C6)*$          AID20425
          R                                                             AID20426
           VECTOR VALUES FMTIN=$1H0,12C6*$                              AID20427
          R                                                             AID20428
           VECTOR VALUES CHECK=$C6*$                                    AID20429
          R                                                             AID20430
           VECTOR VALUES HEAD=$1H1/1H0,S20,13C6 /1H4,S4,20HNO.  OF  INPUAID20431
          2T  DATA,S12,I6/1H ,S4,18HNO.  OF  VARIABLES,S15,I5/1H ,S4,19HAID20432
          3NO.  OF  PREDICTORS,S14,I5/1H ,S4,21HWEIGHT  VARIABLE  NO.,S1AID20433
          42,I5/1H ,S4,27HSPLIT ELIGIBILITY CRITERION,S5,F6.4/1H ,S4,28HAID20434
          5SPLIT REDUCIBILITY CRITERION,S4,F6.4/1H ,S4,26HMAXIMUM  ALLOWAID20435
          6ABLE  GROUPS,S6,I6/1H0,S4,22HDEPENDENT VARIABLE IS ,I3,3H  (,AID20436
          72C6,1H)/1H ,S4,40HVALUES OF DEPENDENT VARIABLE LARGER THAN,  AID20437
          8E15.8,15H   ARE OMITTED. *$                                  AID20438
          R                                                             AID20439
           VECTOR VALUES HEAD1=$1H ,S4,40H  ..          ..         ..   AID20440
          1EQUAL   TO,E15.8,12H      ..      /1H ,S4,40H  ..         .. AID20441
          2    ..         ..     ,E15.8,7H     ../1H ,S4,20HOUTPUT  OPTIAID20442
          3ON 1  IS,I4,2H ./1H ,S4,20HOUTPUT  OPTION 2  IS,I4,2H . /1H0,AID20443
          4S4,21HMINIMUM SIZE REQUIRED,S12,I5/1H0,S4,20HINPUT  DATA  AREAID20444
          5  ON,S11,C6*$                                                AID20445
          R                                                             AID20446
           VECTOR VALUES F1=$1H ,S10,I3,S10,2C6,S10,I4,S14,C1*$         AID20447
          R                                                             AID20448
           VECTOR VALUES F2=$1H0,S4,26HTOTAL  NO.  OF  DATA  READ,S10,I5AID20449
          1/1H ,S4,22HNO.  OF  DATA  DELETED,S14,I5/1H ,S4,26HTOTAL  NO.AID20450
          2  OF  DATA  USED,S10,I5/1H S4,16HSUM  OF  WEIGHTS,S10,E15.8  AID20451
          3/1H ,S4,10HSUM  OF  Y,S16,E15.8/1H ,S4,17HSUM  OF  Y-SQUARE, AID20452
          4S9,E15.8/1H S4,10HMEAN     YS16,E15.8/1H S4,17HSTANDARD  DEV.AID20453
          5  Y,S9,E15.8/1H S4,26HTOTAL SUM OF SQUARES (TSS)E15.8*$      AID20454
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          R                                                             AID20455
           VECTOR VALUES YCARD=$1H ,S4,I5,30H TH  DATA  DELETED.  VARIABAID20456
          1LE(,I4, 5H ) = ,E15.8*$                                      AID20457
          R                                                             AID20458
           VECTOR VALUES DATA=$CARD$,$TAPE$                             AID20459
          R                                                             AID20460
           VECTOR VALUES BAKA=$  NOT $,$      $                         AID20461
          R                                                             AID20462
           VECTOR VALUES  OUT=$ NONE $,$ CARD $,$ TAPE $,$ BOTH $       AID20463
          R                                                             AID20464
           VECTOR VALUES HEAD2=$1H0,S4,C6,16HE DATA WHICH LIE,C6,21HSIDEAID20465
          1 OF INTERVAL FROM,I6,3H TO,I6,12H ON VARIABLE,I6/1H ,S4,C6,  AID20466
          210H WHICH LIE,C6,21HSIDE OF INTERVAL FROM,I6,3H TO,I6,12H ON AID20467
          3VARIABLE,I6*$                                                AID20468
          R                                                             AID20469
           VECTOR VALUES HEAD3=$1H0,S4,14HRESIDUALS  ARE,C6,33H REQUESTEAID20470
          1D  AND  OUTPUT  WILL  BE ,C6,1H.*$                           AID20471
          R                                                             AID20472
           VECTOR VALUES BOB=1B                                         AID20473
          R                                                             AID20474
           VECTOR VALUES RUN=0B                                         AID20475
          R                                                             AID20476
           END OF PROGRAM                                               AID20477
$ASSEMBLE, PUNCH OBJECT                                         IRFORM01AID20478
CODE   MACRO   A                                                        AID20479
       ZET     HALF                                                     AID20480
       TRA     INSERT                                                   AID20481
       STO     TEMPC                                                    AID20482
       CLA     A                                                        AID20483
       TRA     CODESV                                                   AID20484
CODE   END                                                              AID20485
       ENTRY   IRFORM                                                   AID20486
IRFORM SXA     IDX1,1                                                   AID20487
       SXA     IDX2,2                                                   AID20488
       SXA     IDX4,4                                                   AID20489
       STZ     RESULT                                                   AID20490
       CLA     1,4                                                      AID20491
       STA     STOAWY                                                   AID20492
       AXT     0,1                                                      AID20493
       STZ     MAD                                                      AID20494
       CLA*    2,4                                                      AID20495
       PAX     ,2                                                       AID20496
       TXH     *+3,2,0                                                  AID20497
       PDX     ,2                                                       AID20498
       STL     MAD                                                      AID20499
       SXD     LENTST,2                                                 AID20500
       CLA*    3,4                                                      AID20501
       PDX     ,2                                                       AID20502
       TXH     *+2,2,0                                                  AID20503
       PAX     ,2                                                       AID20504
       PXA     ,2                                                       AID20505
       STO     VARINO                                                   AID20506
       STO     VARIN1                                                   AID20507
       CLA*    4,4                                                      AID20508
       PDX     ,2                                                       AID20509
       TXH     *+2,2,0                                                  AID20510
       PAX     ,2                                                       AID20511
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

       SXD     LSTST1,2                                                 AID20512
       TXI     *+1,2,1                                                  AID20513
       SXD     LISTST,2                                                 AID20514
       CLA     5,4                                                      AID20515
       AXT     1,2                                                      AID20516
       SXD     TABPOT,2                                                 AID20517
       STA     TABPOT                                                   AID20518
       CLA*    6,4                                                      AID20519
       PDX     ,2                                                       AID20520
       TXH     *+2,2,0                                                  AID20521
       PAX     ,2                                                       AID20522
       SXD     LENEDT,2                                                 AID20523
       STZ     HALF                                                     AID20524
       TSX     READ,2                                                   AID20525
       STZ     COUNT                                                    AID20526
       AXT     0,2                                                      AID20527
       AXT     6,4                                                      AID20528
       LDQ     INPUT,2                                                  AID20529
SCHLPR ZAC                                                              AID20530
       LGL     6                                                        AID20531
       CAS     LPARN                                                    AID20532
       TRA     *+2                                                      AID20533
       TRA     FOUNDL                                                   AID20534
       TIX     SCHLPR,4,1                                               AID20535
       TXI     *+1,2,-1                                                 AID20536
       TXH     SCHLPR-2,2,-12                                           AID20537
ERRORF TSX     /TV/SPRINT,4                                             AID20538
       BLK     COMFOM,,7                                                AID20539
       TSX     /TV/SYSTEM,4                                             AID20540
COMFOM BCI     7, FORMAT NO STARTED BY END OF FIRST CARD.               AID20541
FOUNDL LGR     6                                                        AID20542
       CLA     BLANKS                                                   AID20543
       LGL     6                                                        AID20544
       TIX     *-1,4,1                                                  AID20545
       SLW     INPUT,2                                                  AID20546
       LDQ     INPUT,2                                                  AID20547
       AXT     6,4                                                      AID20548
TRY    ZAC                                                              AID20549
       LGL     6                                                        AID20550
       CAS     LPARN                                                    AID20551
       TRA     *+2                                                      AID20552
       TRA     LEFT                                                     AID20553
       CAS     RPARN1                                                   AID20554
       TRA     *+2                                                      AID20555
       TRA     RIGHT                                                    AID20556
INREC  TIX     TRY,4,1                                                  AID20557
       CLA     INPUT,2                                                  AID20558
STOAWY STO     **,1                                                     AID20559
       TXI     *+1,1,1                                                  AID20560
       TXI     *+1,2,-1                                                 AID20561
       TXH     TRY-2,2,-12                                              AID20562
       TSX     READ,2                                                   AID20563
       AXT     0,2                                                      AID20564
       TRA     TRY-2                                                    AID20565
LEFT   CLA     COUNT                                                    AID20566
       ADD     =1                                                       AID20567
       STO     COUNT                                                    AID20568
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

       TRA     INREC                                                    AID20569
RIGHT  CLA     COUNT                                                    AID20570
       SUB     =1                                                       AID20571
       TZE     OUT                                                      AID20572
       STO     COUNT                                                    AID20573
       TRA     INREC                                                    AID20574
OUT    CLA     INPUT,2                                                  AID20575
       STO*    STOAWY                                                   AID20576
LENTST TXL     OUTPUT,1,**                                              AID20577
       CALL    SPRINT                                                   AID20578
       BLK     LENERR,,4                                                AID20579
       CALL    SYSTEM                                                   AID20580
LENERR BCI     4, FORMAT IS TO LONG                                     AID20581
OUTPUT SXD     ENDTST,1                                                 AID20582
       CALL    SPRINT                                                   AID20583
       BLK     HEAD,,8                                                  AID20584
       CLA     =1                                                       AID20585
       STO     COLUMN                                                   AID20586
       STO     CARDNO                                                   AID20587
       TSX     CARDHD,1                                                 AID20588
BACIN  AXT     0,1                                                      AID20589
       AXT     6,4                                                      AID20590
       LDQ*    STOAWY                                                   AID20591
       STZ     INT                                                      AID20592
OP1    ZAC                                                              AID20593
       LGL     6                                                        AID20594
       CAS     TEN                                                      AID20595
       TRA     FILDTP                                                   AID20596
       NOP                                                              AID20597
       STO     TEMP                                                     AID20598
       CLA     INT                                                      AID20599
       ALS     2                                                        AID20600
       ADD     INT                                                      AID20601
       ALS     1                                                        AID20602
       ACL     TEMP                                                     AID20603
       STO     INT                                                      AID20604
NXTCHR TSX     INCRE,2                                                  AID20605
       TRA     OP1                                                      AID20606
HEAD   BCI     8,1     VARIABLE NUMBER     COLUMNS     TYPE             AID20607
CARDHD SXA     IDX1H,1                                                  AID20608
       SXA     IDX2H,2                                                  AID20609
       SXA     IDX4H,4                                                  AID20610
       CLA     CARDNO                                                   AID20611
       TSX     CONVTI,2                                                 AID20612
       STO     CDHEAD+1                                                 AID20613
       CALL    SPRINT                                                   AID20614
       BLK     CDHEAD,,2                                                AID20615
IDX1H  AXT     **,1                                                     AID20616
IDX2H  AXT     **,2                                                     AID20617
IDX4H  AXT     **,4                                                     AID20618
       TRA     1,1                                                      AID20619
CDHEAD BCI     2,0CARD                                                  AID20620
CONVTI SXA     IDX1I,1                                                  AID20621
       SXA     IDX2I,2                                                  AID20622
       SXA     IDX4I,4                                                  AID20623
       AXT     0,4                                                      AID20624
       LRS     35                                                       AID20625
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

       CLA     BLANKS                                                   AID20626
       STO     BUILD                                                    AID20627
CONT1  TXI     *+1,4,1                                                  AID20628
       CLM                                                              AID20629
       DVP     TEN                                                      AID20630
       SLW     TEMPI,4                                                  AID20631
       STO     TEMPI1                                                   AID20632
       CLA     TEMPI1                                                   AID20633
       TNZ     CONT1                                                    AID20634
       CAL     BUILD                                                    AID20635
       ALS     6                                                        AID20636
       ORA     TEMPI,4                                                  AID20637
       TIX     *-2,4,1                                                  AID20638
       SLW     BUILD                                                    AID20639
       CLA     BUILD                                                    AID20640
IDX1I  AXT     **,1                                                     AID20641
IDX2I  AXT     **,2                                                     AID20642
IDX4I  AXT     **,4                                                     AID20643
       TRA     1,2                                                      AID20644
TEMPI1 PZE                                                              AID20645
TEMPI  BES     5                                                        AID20646
INCRE  TIX     EDTST2,4,1                                               AID20647
       TXI     *+1,1,1                                                  AID20648
ENDTST TXL     EDTST1,1,**                                              AID20649
       LXA     VARINO,2                                                 AID20650
LSTST1 TXH     FINSH,2,**                                               AID20651
       ZET     HALF                                                     AID20652
       TRA     IDX1                                                     AID20653
       CLA     CARDNO                                                   AID20654
       ADD     =1                                                       AID20655
       STO     CARDNO                                                   AID20656
       TSX     CARDHD,1                                                 AID20657
       CLA     =1                                                       AID20658
       STO     COLUMN                                                   AID20659
       LXA     GROUP1,2                                                 AID20660
       TRA     RESTOR                                                   AID20661
EDTST1 LDQ*    STOAWY                                                   AID20662
       AXT     6,4                                                      AID20663
EDTST2 TRA     1,2                                                      AID20664
FILDTP AXT     TABLEN,2                                                 AID20665
       CAS     TAB+1,2                                                  AID20666
       TRA     *+2                                                      AID20667
       TRA*    SWITCA+1,2                                               AID20668
       TIX     FILDTP+1,2,1                                             AID20669
       TRA     ERR                                                      AID20670
       BCI     1,000001                                                 AID20671
       BCI     1,00000K                                                 AID20672
       BCI     1,00000F                                                 AID20673
       BCI     1,00000E                                                 AID20674
A      BCI     1,00000A                                                 AID20675
       BCI     1,00000H                                                 AID20676
       BCI     1,00000C                                                 AID20677
       BCI     1,00000S                                                 AID20678
       BCI     1,000000                                                 AID20679
       BCI     1,00000T                                                 AID20680
       BCI     1,00000L                                                 AID20681
       BCI     1,00000,                                                 AID20682
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

       BCI     1,00000/                                         AID20683
       BCI     1,00000(                                         AID20684
ABRPN  BCI     1,00000)                                         AID20685
POINTD BCI     1,00000.                                         AID20686
       BCI     1,00000X                                         AID20687
ASTRIX BCI     1,00000*                                         AID20688
       BCI     1,00000P                                         AID20689
BLANK  BCI     1,00000                                          AID20690
       BCI     1,00000$                                         AID20691
       BCI     1,00000G                                         AID20692
       BCI     1,00000J                                         AID20693
PLUS   BCI     1,00000+                                         AID20694
TAB    BCI     1,00000-                                         AID20695
TABLEN EQU     25                                               AID20696
       PZE     FILDCH                                           AID20697
       PZE     FILDCH                                           AID20698
       PZE     FILDCH                                           AID20699
       PZE     FILDCH                                           AID20700
       PZE     FILDCH                                           AID20701
       PZE     HOLTH                                            AID20702
       PZE     FILDCH                                           AID20703
       PZE     SPACE                                            AID20704
       PZE     FILDCH                                           AID20705
       PZE     FILDCT             (T)                           AID20706
       PZE     FILDCL             (L)                           AID20707
       PZE     WIDTH                                            AID20708
       PZE     SLASH                                            AID20709
       PZE     LEFTPR                                           AID20710
       PZE     RPARN                                            AID20711
       PZE     POINT                                            AID20712
       PZE     SPACE1                                           AID20713
       PZE     NXTCHR                                           AID20714
       PZE     ZERINT                                           AID20715
       PZE     NXTCHR                                           AID20716
       PZE     ERR                                              AID20717
       PZE     FILDCG             (G)                           AID20718
       PZE     FILDCJ             (J)                           AID20719
       PZE     NXTCHR             (+)                           AID20720
SWITCA PZE     NXTCHR                                           AID20721
FILDCH STO     TEMPC                                            AID20722
       STL     FILDWT                                           AID20723
       CLA     BLANK7                                           AID20724
       ORA     TEMPC                                            AID20725
       STO     LINE+6                                           AID20726
       STL     FIELD                                            AID20727
FILDH1 CLA     INT                                              AID20728
       TNZ     *+2                                              AID20729
       CLA     =1                                               AID20730
       STO     REPFLD                                           AID20731
ZERINT STZ     INT                                              AID20732
       TRA     NXTCHR                                           AID20733
POINT  STZ     DWIDTH                                           AID20734
       TSX     INCRE,2                                          AID20735
       ZAC                                                      AID20736
       LGL     6                                                AID20737
       CAS     BLANK                                            AID20738
       TRA     FILDTP                                           AID20739
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

       TRA     POINT+1                                                  AID20740
       CAS     TEN                                                      AID20741
       TRA     FILDTP                                                   AID20742
       TRA     *+1                                                      AID20743
       STO     TEMP1                                                    AID20744
       CLA     DWIDTH                                                   AID20745
       ALS     2                                                        AID20746
       ADD     DWIDTH                                                   AID20747
       ALS     1                                                        AID20748
       ADD     TEMP1                                                    AID20749
       STO     DWIDTH                                                   AID20750
       TRA     POINT+1                                                  AID20751
DWIDTH PZE                                                              AID20752
TEMP1  PZE                                                              AID20753
TABPOT PZE     ,4,1                                                     AID20754
HALF   PZE                                                              AID20755
HOLTH  CLA     INT                                                      AID20756
       STO     TEMP                                                     AID20757
HOLTH1 TSX     INCRE,2                                                  AID20758
       LGL     6                                                        AID20759
       CLA     TEMP                                                     AID20760
       SUB     =1                                                       AID20761
       STO     TEMP                                                     AID20762
       TNZ     HOLTH1                                                   AID20763
SKIP1  CLA     =1                                                       AID20764
       STO     REPFLD                                                   AID20765
SKIP   STQ     TEMP                                                     AID20766
       LDQ     REPFLD                                                   AID20767
       MPY     INT                                                      AID20768
       XCA                                                              AID20769
       ADD     COLUMN                                                   AID20770
       STO     COLUMN                                                   AID20771
       LDQ     TEMP                                                     AID20772
       STZ     FIELD                                                    AID20773
       TRA     ZERINT                                                   AID20774
SPACE  STZ     FIELD                                                    AID20775
       TRA     FILDH1                                                   AID20776
SPACE1 CLA     =1                                                       AID20777
       STO     REPFLD                                                   AID20778
       STZ     FIELD                                                    AID20779
       TRA     NXTCHR                                                   AID20780
LEFTPR CLA     INT                                                      AID20781
       STO     GROUP                                                    AID20782
       STO     GROUP1                                                   AID20783
       STQ     MQTEM                                                    AID20784
       SXA     DX1,1                                                    AID20785
       SXA     DX4,4                                                    AID20786
       TRA     ZERINT                                                   AID20787
RPARN  NZT     FIELD                                                    AID20788
       TRA     RPARN2                                                   AID20789
       ZET     HALF                                                     AID20790
       TRA     RPARN2                                                   AID20791
       ZET     FILDWT                                                   AID20792
       TSX     LINFLD,2                                                 AID20793
RPARN2 STZ     SPECFD                                                   AID20794
       STZ     DWIDTH                                                   AID20795
       LXA     GROUP,2                                                  AID20796
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

        TIX     REST1,2,1                                               AID20797
        ZET     FIELD                                                   AID20798
        TRA     SKIP                                                    AID20799
        TRA     ZERINT                                                  AID20800
REST1   SXA     GROUP,2                                                 AID20801
RESTOR  LDQ     MQTEM                                                   AID20802
DX1     AXT     **,1                                                    AID20803
DX4     AXT     **,4                                                    AID20804
        ZET     FIELD                                                   AID20805
        TRA     SKIP                                                    AID20806
        TRA     ZERINT                                                  AID20807
SLASH   NZT     FIELD                                                   AID20808
        TRA     SLASH1                                                  AID20809
        ZET     HALF                                                    AID20810
        TRA     SLASH1                                                  AID20811
        ZET     FILDWT                                                  AID20812
        TSX     LINFLD,2                                                AID20813
SLASH1  STZ     SPECFD                                                  AID20814
        STZ     DWIDTH                                                  AID20815
        ZET     HALF                                                    AID20816
        TRA     ZERINT                                                  AID20817
        TSX     SAVE,2                                                  AID20818
        CLA     CARDNO                                                  AID20819
        ADD     =1                                                      AID20820
        STO     CARDNO                                                  AID20821
        TSX     CARDHD,1                                                AID20822
        TSX     RTN,2                                                   AID20823
        CLA     =1                                                      AID20824
        STO     COLUMN                                                  AID20825
        TRA     ZERINT                                                  AID20826
SAVE    STQ     SAVMQ                                                   AID20827
        SXA     SAVDX1,1                                                AID20828
        SXA     SAVDX4,4                                                AID20829
        TRA     1,2                                                     AID20830
RTN     LDQ     SAVMQ                                                   AID20831
SAVDX1  AXT     **,1                                                    AID20832
SAVDX4  AXT     **,4                                                    AID20833
        TRA     1,2                                                     AID20834
WIDTH   NZT     FIELD                                                   AID20835
        TRA     SKIP                                                    AID20836
        ZET     HALF                                                    AID20837
        TRA     ZERINT                                                  AID20838
        TSX     LINFLD,2                                                AID20839
        STZ     FILDWT                                                  AID20840
        STZ     SPECFD                                                  AID20841
        STZ     DWIDTH                                                  AID20842
        TRA     ZERINT                                                  AID20843
LINFLD  SXA     EXITF,2                                                 AID20844
        TSX     SAVE,2                                                  AID20845
        LXA     REPFLD,1                                                AID20846
        NZT     SPECFD                                                  AID20847
        TRA     LINFL3                                                  AID20848
        CLA     CODEWD                                                  AID20849
        ADD     DWIDTH                                                  AID20850
        STO     CODEWD                                                  AID20851
        LXA     INT,4                                                   AID20852
        TXH     WDERR,4,6                                               AID20853
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
       PXD     ,4                                                       AID20854
       ARS     3                                                        AID20855
       STT     CODEWD                                                   AID20856
LINFL3 CLA     VARINO                                                   AID20857
       TSX     CONVTI,2                                                 AID20858
       LDQ     BLANKS                                                   AID20859
       LGL     12                                                       AID20860
       SLW     LINE+2                                                   AID20861
       CLA     VARINO                                                   AID20862
       ADD     =1                                                       AID20863
       STO     VARINO                                                   AID20864
       PAX     ,2                                                       AID20865
LISTST TXH     FINSH,2,**                                               AID20866
       NZT     SPECFD                                                   AID20867
       TRA     LINFL4                                                   AID20868
       LXA     VARINO,4                                                 AID20869
       TXI     *+1,4,-1                                                 AID20870
       SXD     CODEWD,4                                                 AID20871
       LXD     TABPOT,4                                                 AID20872
       CLA     CODEWD                                                   AID20873
       STO*    TABPOT                                                   AID20874
       TXI     *+1,4,1                                                  AID20875
       SXD     TABPOT,4                                                 AID20876
LENEDT TXH     SAVER,4,**                                               AID20877
LINFL4 CLA     COLUMN                                                   AID20878
LINFL1 EQU     LINFL3                                                   AID20879
       ADD     INT                                                      AID20880
       SUB     =1                                                       AID20881
       STO     TEMP                                                     AID20882
       SUB     COLUMN                                                   AID20883
       TZE     ONECOL                                                   AID20884
       CLA     COLUMN                                                   AID20885
       TSX     CONVTI,2                                                 AID20886
       STO     LINE+4                                                   AID20887
       CLA     TEMP                                                     AID20888
       TSX     CONVTI,2                                                 AID20889
       LGR     12                                                       AID20890
       CAL     MINUS                                                    AID20891
LINFL2 LGL     12                                                       AID20892
       LDQ     BLANKS                                                   AID20893
       LGL     18                                                       AID20894
       SLW     LINE+5                                                   AID20895
       CALL    SPRINT                                                   AID20896
       BLK     LINE,,7                                                  AID20897
       CLA     TEMP                                                     AID20898
       ADD     =1                                                       AID20899
       STO     COLUMN                                                   AID20900
       TIX     LINFL1,1,1                                               AID20901
       TSX     RTN,2                                                    AID20902
EXITF  AXT     **,2                                                     AID20903
       TRA     1,2                                                      AID20904
ONECOL CLA     BLANKS                                                   AID20905
       STO     LINE+4                                                   AID20906
       CLA     TEMP                                                     AID20907
       TSX     CONVTI,2                                                 AID20908
       LGR     12                                                       AID20909
       CAL     BLANKS                                                   AID20910
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

       TRA     LINFL2                                                   AID20911
ERR    CALL    SPRINT                                                   AID20912
       BLK     ERCOM,,3                                                 AID20913
       CALL    SYSTEM                                                   AID20914
IDX1   AXT     **,1                                                     AID20915
       CLA     TABPOT                                                   AID20916
       AXT     0,4                                                      AID20917
       ZET     MAD                                                      AID20918
       SSM                                                              AID20919
       STO*    TABPOT                                                   AID20920
       CLA     RESULT                                                   AID20921
IDX2   AXT     **,2                                                     AID20922
IDX4   AXT     **,4                                                     AID20923
       TRA     2,4                                                      AID20924
READ   TSX     /TV/SCARDS,4                                             AID20925
       BLK     INPUT,,EOF                                               AID20926
       TRA     1,2                                                      AID20927
EOF    TSX     /TV/SPRINT,4                                             AID20928
       BLK     EOFCM,,3                                                 AID20929
       TSX     /TV/SYSTEM,4                                             AID20930
EOFCM  BCI     3, END OF FILE.                                          AID20931
ERCOM  BCI     3, ILLEGAL CHARACTER                                     AID20932
VARINO PZE                                                              AID20933
COLUMN PZE                                                              AID20934
CARDNO PZE                                                              AID20935
TEN    PZE     10                                                       AID20936
INT    PZE                                                              AID20937
TEMP   PZE                                                              AID20938
BUILD  PZE                                                              AID20939
FIELD  PZE                                                              AID20940
REPFLD PZE                                                              AID20941
GROUP1 PZE                                                              AID20942
GROUP  PZE                                                              AID20943
MQTEM  PZE                                                              AID20944
TEMPC  PZE                                                              AID20945
SAVMQ  PZE                                                              AID20946
LINE   BCI     7,                                                       AID20947
MINUS  BCI     1,     -                                                 AID20948
BLANK7 BCI     1,     0                                                 AID20949
COUNT  PZE                                                              AID20950
LPARN  BCI     1,00000(                                                 AID20951
BLANKS BCI     1,                                                       AID20952
RPARN1 BCI     1,00000)                                                 AID20953
UNDWD  MZE                                                              AID20954
VARIN1 PZE                                                              AID20955
SPECFD PZE                                                              AID20956
CODEWD PZE                                                              AID20957
BLCHK  PZE                                                              AID20958
RESULT PZE                                                              AID20959
MAD    PZE                                                              AID20960
INTFLT PZE                                                              AID20961
DWIDT1 PZE                                                              AID20962
VALUE  PZE                                                              AID20963
DGSW   PZE                                                              AID20964
SIGNSW PZE                                                              AID20965
MZE    MZE                                                              AID20966
MASKT  OCT     700000                                                   AID20967
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

INPUT  BSS     30                                                        AID20968
FILDWT PZE                                                               AID20969
FILDCT CODE    =2                                                        AID20970
FILDCL CODE    =1                                                        AID20971
FILDCJ CODE    =3                                                        AID20972
FILDCG CODE    =10                                                       AID20973
CODESV STO     CODEWD                                                    AID20974
       STL     SPECFD                                                    AID20975
       STL     RESULT                                                    AID20976
       TRA     FILDCH+1                                                  AID20977
FINSH  ZET     HALF                                                      AID20978
       TRA     IDX1                                                      AID20979
       STL     HALF                                                      AID20980
       CLA     VARIN1                                                    AID20981
       STO     VARINO                                                    AID20982
       TRA     BACIN                                                     AID20983
INSERT LDQ*    STOAWY                                                    AID20984
       STZ     BLCHK                                                     AID20985
       ZAC                                                               AID20986
       SXD     *+2,4                                                     AID20987
       AXT     6,2                                                       AID20988
INSET1 TXL     INSET2,2,**                                               AID20989
       LGL     6                                                         AID20990
       TXI     INSET1,2,-1                                               AID20991
INSET2 ALS     6                                                         AID20992
       ACL     A                                                         AID20993
       SLW     TEMP                                                      AID20994
       LGL     6                                                         AID20995
       TNX     INSET3,2,1                                                AID20996
INSET5 CAL     TEMP                                                      AID20997
       ALS     6                                                         AID20998
       SLW     TEMP                                                      AID20999
       ZAC                                                               AID21000
       LGL     6                                                         AID21001
       CAS     BLANK                                                     AID21002
       TRA     INSET8                                                    AID21003
       TRA     INSET6                                                    AID21004
       CAS     TEN                                                       AID21005
       TRA     INSET4                                                    AID21006
       NOP                                                               AID21007
INSET7 ZET     BLCHK                                                     AID21008
       CLA     BLANK                                                     AID21009
INSET6 ACL     TEMP                                                      AID21010
       SLW     TEMP                                                      AID21011
       TIX     INSET5,2,1                                                AID21012
INSET3 CAL     TEMP                                                      AID21013
       SLW*    STOAWY                                                    AID21014
       CLA     BLANKS                                                    AID21015
       STO     TEMP                                                      AID21016
       TXI     *+1,1,1                                                   AID21017
       AXT     6,2                                                       AID21018
       LDQ*    STOAWY                                                    AID21019
       TRA     INSET5                                                    AID21020
BLOUT  STL     BLCHK                                                     AID21021
       TRA     INSET7                                                    AID21022
INSET4 CAS     ABRPN                                                     AID21023
       TRA     INSETA                                                    AID21024
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

       TRA     INSET8                                                   AID21025
       TRA     BLOUT                                                    AID21026
INSETA CAS     ASTRIX                                                   AID21027
       TRA     ERR                                                      AID21028
       TRA     INSET8                                                   AID21029
       TRA     ERR                                                      AID21030
INSET8 ACL     TEMP                                                     AID21031
       STQ     TEMP                                                     AID21032
       SXA     SAV4,2                                                   AID21033
       TNX     INSET9,2,1                                               AID21034
       LGL     6                                                        AID21035
       TRA     *-2                                                      AID21036
INSET9 SLW*    STOAWY                                                   AID21037
       LDQ     TEMP                                                     AID21038
SAV4   AXT     **,4                                                     AID21039
       TRA     ZERINT                                                   AID21040
       ENTRY   EDITPM                                                   AID21041
EDITPM SXA     IDX1,1                                                   AID21042
       SXA     IDX2,2                                                   AID21043
       SXA     IDX4,4                                                   AID21044
       STZ     RESULT                                                   AID21045
       CAL*    3,4                                                      AID21046
       STO     TABPOT                                                   AID21047
       STA     AROUND                                                   AID21048
       STZ     MAD                                                      AID21049
       PBT                                                              AID21050
       TRA     *+2                                                      AID21051
       STL     MAD                                                      AID21052
       LXD     TABPOT,1                                                 AID21053
       TXL     RETURN,1,1                                               AID21054
       TXI     *+1,1,-1                                                 AID21055
       SXD     TSTEND,1                                                 AID21056
       CLA*    2,4                                                      AID21057
       PAX     ,2                                                       AID21058
       ZET     MAD                                                      AID21059
       PDX     ,2                                                       AID21060
       SXD     TSTLEN,2                                                 AID21061
       AXT     1,2                                                      AID21062
       CLA     1,4                                                      AID21063
       ADD     =1                                                       AID21064
       STA     PICKUP                                                   AID21065
AROUND CLA     **,2                                                     AID21066
       PAX     ,1                                                       AID21067
       PDX     ,4                                                       AID21068
TSTLEN TXH     RETURN,4,**                                              AID21069
       STZ     DWIDTH                                                   AID21070
       TXL     PICKUP,1,9                                               AID21071
       TXI     *+1,1,-10                                                AID21072
       SXA     DWIDTH,1                                                 AID21073
       AXT     4,1                                                      AID21074
PICKUP LDQ     **,4                                                     AID21075
       TXH     TABERR,1,4                                               AID21076
       TRA     SWITCH,1                                                 AID21077
       TRA     GCONV              RETURN-NUMBER                         AID21078
       TRA     JCONV              IN ACC-DECIMAL                        AID21079
       TRA     TCONV              DIGITS IN IDX1                        AID21080
       TRA     LCONV                                                    AID21081
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

SWITCH NZT     MAD               INTEGER                                AID21082
       TRA     STOWY                                                    AID21083
       ALS     18                                                       AID21084
STOWY  STO*    PICKUP                                                   AID21085
       TXI     *+1,2,1                                                  AID21086
TSTEND TXL     AROUND,2,**                                              AID21087
RETURN CLA     RESULT                                                   AID21088
       LXA     IDX1,1                                                   AID21089
       TRA     IDX2                                                     AID21090
FLOAT  TXH     *+2,1,0                                                  AID21091
       LXA     DWIDTH,1                                                 AID21092
       ORA     FTCONT                                                   AID21093
       FAD     FTCONT                                                   AID21094
       FDP     TABP,1                                                   AID21095
       XCA                                                              AID21096
       TRA     STOWY                                                    AID21097
       DEC     1.0E6                                                    AID21098
       DEC     1.0E5                                                    AID21099
       DEC     1.0E4                                                    AID21100
       DEC     1.0E3                                                    AID21101
       DEC     1.0E2                                                    AID21102
       DEC     1.0E1                                                    AID21103
TABP   DEC     1.                                                       AID21104
FTCONT DEC     155B8                                                    AID21105
GCONV  STZ     INTFLT                                                   AID21106
       TRA     *+2                                                      AID21107
JCONV  STL     INTFLT                                                   AID21108
       STZ     DWIDT1                                                   AID21109
       ANA     MASKT                                                    AID21110
       ALS     3                                                        AID21111
       PDX     ,1                                                       AID21112
       STZ     VALUE                                                    AID21113
       STZ     TEMP                                                     AID21114
       STZ     DGSW                                                     AID21115
       STZ     SIGNSW                                                   AID21116
NXTDIG ZAC                                                              AID21117
       LGL     6                                                        AID21118
       CAS     BLANK                                                    AID21119
       TRA     UNDEFV                                                   AID21120
       TRA     NXTCH1                                                   AID21121
       CAS     MINUSP                                                   AID21122
       TRA     UNDEFV                                                   AID21123
       TRA     MSIGN                                                    AID21124
       CAS     PLUS                                                     AID21125
       TRA     DPOINT                                                   AID21126
       TRA     PSIGN                                                    AID21127
NXTCH3 CAS     =10                                                      AID21128
       TRA     UNDEFV                                                   AID21129
       TRA     UNDEFV                                                   AID21130
       STA     TEMP                                                     AID21131
       STL     DGSW                                                     AID21132
       CLA     VALUE                                                    AID21133
       ALS     2                                                        AID21134
       ADD     VALUE                                                    AID21135
       ALS     1                                                        AID21136
       ADD     TEMP                                                     AID21137
       STO     VALUE                                                    AID21138
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

NXTCH2 TIX     NXTDIG,1,1                                               AID21139
       NZT     DGSW                                                     AID21140
       TRA     UNDEFV                                                   AID21141
       LXA     DWIDT1,1                                                 AID21142
NXTCH4 ZET     INTFLT                                                   AID21143
       TRA     SWITCH                                                   AID21144
       TRA     FLOAT                                                    AID21145
MSIGN  ZET     SIGNSW                                                   AID21146
       TRA     UNDEFV                                                   AID21147
       STL     SIGNSW                                                   AID21148
       CLA     MZE                                                      AID21149
       STO     VALUE                                                    AID21150
       STO     TEMP                                                     AID21151
       ZET     DGSW                                                     AID21152
       TRA     UNDEFV                                                   AID21153
       TRA     NXTCH2                                                   AID21154
PSIGN  ZET     SIGNSW                                                   AID21155
       TRA     UNDEFV                                                   AID21156
       ZET     DGSW                                                     AID21157
       TRA     UNDEFV                                                   AID21158
       STL     SIGNSW                                                   AID21159
       TRA     NXTCH2                                                   AID21160
NXTCH1 ZET     DGSW                                                     AID21161
       TRA     UNDEFV                                                   AID21162
       ZET     SIGNSW                                                   AID21163
       TRA     UNDEFV                                                   AID21164
       TRA     NXTCH2                                                   AID21165
DPOINT CAS     POINTD                                                   AID21166
       TRA     UNDEFV                                                   AID21167
       TRA     *+2                                                      AID21168
       TRA     NXTCH3                                                   AID21169
       TXI     *+1,1,-1                                                 AID21170
       SXA     DWIDT1,1                                                 AID21171
       TXI     NXTCH2,1,1                                               AID21172
TCONV  STZ     INTFLT                                                   AID21173
       TRA     *+2                                                      AID21174
LCONV  STL     INTFLT                                                   AID21175
       ANA     MASKT                                                    AID21176
       ALS     3                                                        AID21177
       PDX     ,1                                                       AID21178
       STZ     VALUE                                                    AID21179
       STZ     TEMP                                                     AID21180
       STZ     SIGNSW                                                   AID21181
NXTCLT ZAC                                                              AID21182
       LGL     6                                                        AID21183
       CAS     BLANK                                                    AID21184
       TRA     UNDEFV                                                   AID21185
       TRA     LTCON1                                                   AID21186
       CAS     MINUSP                                                   AID21187
       TRA     UNDEFV                                                   AID21188
       TRA     EM                                                       AID21189
       CAS     PLUS                                                     AID21190
       TRA     UNDEFV                                                   AID21191
       TRA     EP                                                       AID21192
       CAS     =10                                                      AID21193
       TRA     UNDEFV                                                   AID21194
       TRA     UNDEFV                                                   AID21195
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

LTCON2 STA     TEMP                                                     AID21196
       CLA     VALUE                                                    AID21197
       ALS     1                                                        AID21198
       ADD     VALUE                                                    AID21199
       ALS     2                                                        AID21200
       ADD     TEMP                                                     AID21201
       STO     VALUE                                                    AID21202
       STL     SIGNSW                                                   AID21203
LTCON3 TIX     NXTCLT,1,1                                               AID21204
       NZT     SIGNSW                                                   AID21205
       TRA     UNDEFV                                                   AID21206
       TXI     NXTCH4,1,-1                                              AID21207
EP     CLA     =10                                                      AID21208
       TRA     LTCON2                                                   AID21209
EM     CLA     =11                                                      AID21210
       TRA     LTCON2                                                   AID21211
LTCON1 ZET     SIGNSW                                                   AID21212
       TRA     UNDEFV                                                   AID21213
       TRA     LTCON3                                                   AID21214
UNDEFV CLA     UNDWD                                                    AID21215
       STL     RESULT                                                   AID21216
       TRA     STOWY                                                    AID21217
TABERR CALL    SPRINT                                                   AID21218
       BLK     TABERC,,3                                                AID21219
       CALL    ERROR                                                    AID21220
WDERR  CALL    SPRINT                                                   AID21221
       BLK     WDERRC,,6                                                AID21222
       CALL    ERROR                                                    AID21223
SAVER  CALL    SPRINT                                                   AID21224
       BLK     SAVERC,,4                                                AID21225
       CALL    ERROR                                                    AID21226
WDERRC BCI     6, FIELD WIDTH MORE THAN 6                               AID21227
SAVERC BCI     4, EDIT TABLE EXCEEDED                                   AID21228
TABERC BCI     3, BAD EDIT TABLE                                        AID21229
MINUSP SYN     TAB                                                      AID21230
       END                                                              AID21231
$ASSEMBLE, PUNCH OBJECT                                         SAPTIM01AID21232
       ENTRY   WRATIM                                                   AID21233
SAVE   PZE                                                              AID21234
       PZE                                                              AID21235
       PZE                                                              AID21236
SIXTY  DEC     60                                                       AID21237
HRS    PZE                                                              AID21238
MIN    PZE                                                              AID21239
SEC    PZE                                                              AID21240
FRACT  PZE                                                              AID21241
WRATIM SXD     SAVE,4                                                   AID21242
       SXD     SAVE+1,2                                                 AID21243
       SXD     SAVE+2,1                                                 AID21244
       CALL    DAYTIM                                                   AID21245
       LRS     35                                                       AID21246
       DVP     SIXTY                                                    AID21247
       STO     FRACT                                                    AID21248
       ZAC                                                              AID21249
       DVP     SIXTY                                                    AID21250
       STO     SEC                                                      AID21251
       ZAC                                                              AID21252
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

       DVP     SIXTY                                                    AID21253
       STO     MIN                                                      AID21254
       ZAC                                                              AID21255
       DVP     SIXTY                                                    AID21256
       STO     HRS                                                      AID21257
       PRINT   FMA,HRS,MIN,SEC,FRACT,0                                  AID21258
OUT    LXD     SAVE,4                                                   AID21259
       LXD     SAVE+1,2                                                 AID21260
       LXD     SAVE+2,1                                                 AID21261
       TRA     2,4                                                      AID21262
FMA    BCI     *,12HOTIME IS NOW,4(I3,1H.)*                             AID21263
       END                                                              AID21264
$ASSEMBLE, PUNCH OBJECT                                           CAS001AID21265
*              FUNCTION CAS, UMAP, NOV 1961, SONQUIST                   AID21266
*              CHECKS TWO ARGS WITH A CAS                               AID21267
*              TO SEE IF THEY ARE EQUAL                                 AID21268
*              NORMAL USE IS WITH AN IF IN FORTRAN                      AID21269
*              IF(CAS(A,B))A.L.B,A=B,A.G.B                              AID21270
       ENTRY   CAS                                                      AID21271
CAS    CLA     1,4                                                      AID21272
       STA     GETA                                                     AID21273
       CLA     2,4                                                      AID21274
       STA     GETB                                                     AID21275
GETA   CLA     **                                                       AID21276
GETB   CAS     **                                                       AID21277
       TRA     AGR                                                      AID21278
       TRA     EQ                                                       AID21279
       TRA     ALES                                                     AID21280
AGR    CLA     PLONE                                                    AID21281
       TRA     3,4                                                      AID21282
EQ     CLA     ZER                                                      AID21283
       TRA     3,4                                                      AID21284
ALES   CLA     MONE                                                     AID21285
       TRA     3,4                                                      AID21286
PLONE  DEC     1.0                                                      AID21287
ZER    DEC     0.                                                       AID21288
MONE   DEC     -1.0                                                     AID21289
       END                                                              AID21290
$BREAK                                                                  AID21291
$ COMPILE MAD,PRINT OBJECT,PUNCH OBJECT                         AIDM2201AID21292
          R                                                             AID21293
          R          PROGRAM  NAME  --  A I D.    SECOND CORE.          AID21294
          R              WRITTEN  BY  ROBERT  W   HSIEH.                AID21295
          R            AID  -  MODEL  2  -  REWRITTEN  ON  AUGUST 1963. AID21296
          R                                                             AID21297
           NORMAL MODE IS INTEGER                                       AID21298
          R                                                             AID21299
           DIMENSION ID(128),INDEX(128),HI(128),LO(128),TN(128),TWT(128)AID21300
          1 ,TY1(128),TY2(128),CLASS(256,DIM),LOC(128),FAIL(128),MAX(36)AID21301
          2,P(36),SIGN(128),LAST(128),C(64),LIST(64),X(20000),D(20000), AID21302
          3 BSS(64),CODE(64),N(64),W(64),Y1(64),Y2(64),YBAR(64),        AID21303
          4 BSSP(64),KODE(64),N1(64),W1(64),Y3(64),Y4(64),Y5(64),       AID21304
          5 NAME1(36), NAME2(36) ,TYPE(36) ,LAB(12)                     AID21305
          R                                                             AID21306
           PROGRAM COMMON NAME1,NAME2, NP,NV, LAB,                      AID21307
          0               ID,INDEX,HI,LO,TN,TWT,TY1,TY2,CLASS,LOC,FAIL, AID21308
          1 MAX,P,NOGP,ITR,ITRMAX,PA,PB,OP1,OP2,X,MSIZE,SCFIN,SCFOUT,   AID21309
```

```
000000000111111111122222222223333333333444444444455555555556666666666777777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          2  KONST,AA,BB,RUN,ZWANT,ZTYPE,ZTAPE,BOB,TYPE                AID21310
          R                                                            AID21311
           FLOATING POINT TWT,TY1,TY2,PA,PB,Y,D,TSS,BSS,BSSP,TMAX,BMAX, AID21312
          1 Y1,Y2,Y3,Y4,Y5,W,W1,YA,YB,SQRT.,YBAR,WHT                   AID21313
          R                                                            AID21314
           BOOLEAN SIGN,OUTP1,OUTP2                                    AID21315
          R                                                            AID21316
           EQUIVALENCE(X,D),(TN,LAST),(LIST,C),(OP1,OUTP1),(OP2,OUTP2) ,AID21317
          1 (SIGN,FAIL)                                                AID21318
          R                                                            AID21319
           PRINT COMMENT$0$                                            AID21320
          R                                                            AID21321
          R          STORE  BASIC  GROUP (PG)  STATISTICS.             AID21322
          R                                                            AID21323
           PG=1                                                        AID21324
           LOC(1)=1                                                    AID21325
           TWT(1)=TWT                                                  AID21326
           TN(1)=TN                                                    AID21327
           TY1(1)=TY1                                                  AID21328
           TY2(1)=TY2                                                  AID21329
           ITRMAX=ITRMAX-1                                             AID21330
          R                                                            AID21331
          R          SET  OUTPUT  OPTION  1  SWITCH.                   AID21332
          R                                                            AID21333
           WHENEVER OUTP1                                              AID21334
           Z=1                                                         AID21335
           OTHERWISE                                                   AID21336
           Z=2                                                         AID21337
           END OF CONDITIONAL                                          AID21338
          R                                                            AID21339
          R   FROM PARENT GROUP(PG),TO SELECT THE BEST PREDICTOR       AID21340
          R         AND  TO  FIND  OFF-SPRINGS.                        AID21341
          R                                                            AID21342
           NOGP=1                                                      AID21343
           ITR=0                                                       AID21344
          R                                                            AID21345
          R          INITIALIZE  START  OF  ITERATIONS.                AID21346
          R                                                            AID21347
ENTER      ITR=ITR+1                                                   AID21348
           PRINT FORMAT OUT1, ITR,PG                                   AID21349
           WHENEVER ITR .G. ITRMAX,TRANSFER TO ENDEN                   AID21350
BEGIN      TMAX=0.0                                                    AID21351
           SAVE=0                                                      AID21352
          R                                                            AID21353
          R          PARTITION  SCAN  STARTS.                          AID21354
          R                                                            AID21355
           THROUGH CHOICE, FOR JI=1,1,JI .G. NP                        AID21356
           JP=P(JI)                                                    AID21357
           JB=(JI-1)/6+2                                               AID21358
           M=MAX(JI)                                                   AID21359
           JS=(JI-JI/6*6)*6                                            AID21360
           EXECUTE ZERO.(N(0)...N(M),W(0)...W(M),Y1(0)...Y1(M),Y2(0)... AID21361
          1 Y2(M))                                                     AID21362
           X=LOC(PG)                                                   AID21363
JUMP       J=X(X) .RS. 18                                              AID21364
           WHENEVER X .E. 0,TRANSFER TO REST                           AID21365
           WHT=X(X) .A. 77777K                                         AID21366
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           Y=D(X+1)                                                     AID21367
           JC=X+JB                                                      AID21368
           P=(X(JC) .RS. JS) .A. 77K                                    AID21369
           N(P)=N(P)+1                                                  AID21370
           W(P)=W(P)+WHT                                                AID21371
           Y1(P)=Y1(P)+Y*WHT                                            AID21372
           Y2(P)=Y2(P)+Y*Y*WHT                                          AID21373
           X=J                                                          AID21374
           TRANSFER TO JUMP                                             AID21375
REST       CHECK=-1                                                     AID21376
           N(64)=0                                                      AID21377
           W(64)=0.0                                                    AID21378
           Y1(64)=0.0                                                   AID21379
           Y2(64)=0.0                                                   AID21380
           THROUGH RA,FOR K=0,1,K .G. M                                 AID21381
RA         WHENEVER N(K) .NE. 0, CHECK=CHECK+1                          AID21382
          R          TEST  IF  NON-ZERO  CATEGORIES  ARE  MORE  THAN  1.AID21383
           WHENEVER CHECK .LE. 0                                        AID21384
           PRINT FORMAT OUT2, JP,PG,ITR                                 AID21385
           TRANSFER TO CHOICE                                           AID21386
           END OF CONDITIONAL                                           AID21387
          R                                                             AID21388
          R          SQUEEZE  ZERO  CATEGORIES  AND  COMPUTE  SUMS.     AID21389
          R                                                             AID21390
           J=-1                                                         AID21391
           THROUGH RB, FOR K=0,1,K .G. M                                AID21392
           WHENEVER N(K) .E. 0, TRANSFER TO RB                          AID21393
           J=J+1                                                        AID21394
           CODE(J)=K                                                    AID21395
           C(J)=J                                                       AID21396
           YBAR(J)=Y1(K)/W(K)                                           AID21397
           N(64)=N(64)+N(K)                                             AID21398
           W(64)=W(64)+W(K)                                             AID21399
           Y1(64)=Y1(64)+Y1(K)                                          AID21400
           Y2(64)=Y2(64)+Y2(K)                                          AID21401
RB         CONTINUE                                                     AID21402
          R                                                             AID21403
          R          PREDICTOR  TYPE  SWITCH  --  FREE  OR  MONOTONE.   AID21404
          R                                                             AID21405
           TRANSFER TO SCAN(TYPE(JI))                                   AID21406
          R                                                             AID21407
          R          SORT  MEANS  IN  DESCENDING  ORDER  ON  FREE  TYPE.AID21408
          R                                                             AID21409
SCAN(0)    THROUGH RCA, FOR I=CHECK, -1, I .E. 0                        AID21410
           K=0                                                          AID21411
           THROUGH RD,FOR J=0,1,J .E. I                                 AID21412
           WHENEVER YBAR(J) .L. YBAR(J+1)                               AID21413
           Y=YBAR(J)                                                    AID21414
           YBAR(J)=YBAR(J+1)                                            AID21415
           YBAR(J+1)=Y                                                  AID21416
           X=C(J)                                                       AID21417
           C(J)=C(J+1)                                                  AID21418
           C(J+1)=X                                                     AID21419
           K=1                                                          AID21420
RD         END OF CONDITIONAL                                           AID21421
RCA        WHENEVER K .E. 0, TRANSFER TO SCAN(1)                        AID21422
          R                                                             AID21423
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          R          SWITCH  FOR  OUTPUT  OPTION 2  AND  PRINT  HEADER. AID21424
          R                                                             AID21425
SCAN(1)    WHENEVER OUTP2 .OR. ITR .E. 1                                AID21426
           PRINT FORMAT OUT3, JP, PG                                    AID21427
           Q=1                                                          AID21428
           PRINT COMMENT $0     CODE     N    SUM OF WEIGHT      SUM  OFAID21429
          1   Y    SUM  Y-SQUARE       MEAN           STD.   DEV.       AID21430
          2    B  S  S$                                                 AID21431
           OTHERWISE                                                    AID21432
           Q=2                                                          AID21433
           END OF CONDITIONAL                                           AID21434
          R                                                             AID21435
          R          SEARCH  FOR  THE  LARGEST  B S S  STARTS.          AID21436
          R                                                             AID21437
           TY1=0.0                                                      AID21438
           YA=0.0                                                       AID21439
           CI=CHECK-1                                                   AID21440
           BSS(64)=Y1(64)*Y1(64)/W(64)                                  AID21441
           BMAX=0.0                                                     AID21442
          R                                                             AID21443
           THROUGH RED,FOR K=0,1,K .G. C1                               AID21444
           L=CODE(C(K))                                                 AID21445
           YA=YA+Y1(L)                                                  AID21446
           YB=Y1(64)-YA                                                 AID21447
           TY1=TY1+W(L)                                                 AID21448
           TY2=W(64)-TY1                                                AID21449
           BSS(K)=YA*YA/TY1+YB*YB/TY2-BSS(64)                           AID21450
          R          OUTPUT  OPTION 2 SWITCH  IS ON  IF  Q=1.           AID21451
           TRANSFER TO ROSE(Q)                                          AID21452
ROSE(1)    Y=(Y2(L)-Y1(L)*YBAR(K))/W(L)                                 AID21453
           WHENEVER Y .G.0.                                             AID21454
           Y=SQRT.(Y)                                                   AID21455
           OTHERWISE                                                    AID21456
           Y=0.                                                         AID21457
           END OF CONDITIONAL                                           AID21458
           PRINT FORMAT OUT4,   L    ,N(L),W(L),Y1(L),Y2(L),YBAR(K),Y,  AID21459
          1 BSS(K)                                                      AID21460
          R                                                             AID21461
ROSE(2)    WHENEVER BSS(K) .G. BMAX                                     AID21462
           SMAX=K                                                       AID21463
           BMAX=BSS(K)                                                  AID21464
RED        END OF CONDITIONAL                                           AID21465
          R                                                             AID21466
           BSS(64)=Y2(64)-BSS(64)                                       AID21467
           L=CODE(C(K))                                                 AID21468
           YA=BMAX/BSS(64)                                              AID21469
           TRANSFER TO BARA(Q)                                          AID21470
BARA(1)    Y=(Y2(L)-Y1(L)*YBAR(K))/W(L)                                 AID21471
           WHENEVER Y .G.0.                                             AID21472
           Y=SQRT.(Y)                                                   AID21473
           OTHERWISE                                                    AID21474
           Y=0.                                                         AID21475
           END OF CONDITIONAL                                           AID21476
           PRINT FORMAT OUT4,   L    ,N(L),W(L),Y1(L),Y2(L),YBAR(K),Y,  AID21477
          1  BSS(64)                                                    AID21478
BARA(2)    PRINT FORMAT OUT1,JP,NAME1(JI),NAME2(JI),BMAX ,YA            AID21479
          R                                                             AID21480
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          R          SAVE  THE  BEST  SPLIT  INFORMATION.               AID21481
          R                                                             AID21482
           WHENEVER TMAX .L. BMAX                                       AID21483
           TMAX=BMAX                                                    AID21484
           SAVE=CHECK                                                   AID21485
           PX=JP                                                        AID21486
           PV=JI                                                        AID21487
           PMAX=SMAX                                                    AID21488
               THROUGH REMA,FOR I=0,1,I .G. SAVE                        AID21489
           J=CODE(C(I))                                                 AID21490
                   N1(I)=N(J)                                           AID21491
                   W1(I)=W(J)                                           AID21492
                   Y3(I)=Y1(J)                                          AID21493
                   Y4(I)=Y2(J)                                          AID21494
                   Y5(I)=YBAR(I)                                        AID21495
                   KODE(I)=J                                            AID21496
REMA               BSSP(I)=BSS(I)                                       AID21497
           BSSP(64)=BSS(64)                                             AID21498
           END OF CONDITIONAL                                           AID21499
CHOICE     CONTINUE                                                     AID21500
          R                                                             AID21501
          R          END  OF  PARTITION  SCAN.                          AID21502
          R          TEST  IF  SPLIT  SATISFIES  CRITERION 2.           AID21503
          R                                                             AID21504
           WHENEVER TMAX.LE. PB                                         AID21505
           SIGN(PG)=1B                                                  AID21506
           PRINT FORMAT OUT5, PG,PX,TMAX                                AID21507
           SIGN=1B                                                      AID21508
           TRANSFER TO SEARCH                                           AID21509
           END OF CONDITIONAL                                           AID21510
          R                                                             AID21511
          R          PERFORM  PARTITION - ASSIGN  SPLIT  GROUP  I. D.S. AID21512
          R                                                             AID21513
           NOGP=NOGP+2                                                  AID21514
           WHENEVER NOGP.G. 127, TRANSFER TO EXIT                       AID21515
           GA=NOGP - 1                                                  AID21516
           GB=NOGP                                                      AID21517
           N=0                                                          AID21518
           W=0.0                                                        AID21519
           Y1=0.0                                                       AID21520
           Y2=0.0                                                       AID21521
          R                                                             AID21522
          R          STORE  PARTITION  CODES  --  FIRST  GROUP.         AID21523
          R                                                             AID21524
           THROUGH ONE,FOR K=0,1,K .G. PMAX                             AID21525
           I=KODE(K)                                                    AID21526
           WHENEVER I .L. 36                                            AID21527
           CLASS(GA,1)=(1 .LS. I) .V. CLASS(GA,1)                       AID21528
           OTHERWISE                                                    AID21529
           CLASS(GA,2)=(1 .LS.(I-36 ) .V. CLASS(GA,2)                   AID21530
           END OF CONDITIONAL                                           AID21531
           LIST(I)=GA                                                   AID21532
           N=N+N1(K)                                                    AID21533
           Y1=Y1+Y3(K)                                                  AID21534
           Y2=Y2+Y4(K)                                                  AID21535
ONE        W=W+W1(K)                                                    AID21536
           N(1)=0                                                       AID21537
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           W(1)=0.0                                                     AID21538
           Y1(1)=0.0                                                    AID21539
           Y2(1)=0.0                                                    AID21540
          R                                                             AID21541
          R          STORE  PARTITION  CODES  --  SECOND  GROUP.        AID21542
          R                                                             AID21543
           THROUGH TWO,FOR J=PMAX+1,1,J .G. SAVE                        AID21544
           I=KODE(J)                                                    AID21545
           WHENEVER I .L. 36                                            AID21546
           CLASS(GB,1)=(1 .LS. I) .V. CLASS(GB,1)                       AID21547
           OTHERWISE                                                    AID21548
           CLASS(GB,2)=(1 .LS.(I -36)) .V. CLASS(GB,2)                  AID21549
           END OF CONDITIONAL                                           AID21550
           LIST(I)=GB                                                   AID21551
           N(1)=N(1)+N1(J)                                              AID21552
           W(1)=W(1)+W1(J)                                              AID21553
           Y1(1)=Y1(1)+Y3(J)                                            AID21554
TWO        Y2(1)=Y2(1)+Y4(J)                                            AID21555
          R                                                             AID21556
          R         THREADING OF GROUPING DATA.                         AID21557
          R                                                             AID21558
           SKIP=1                                                       AID21559
           JS=(PV-PV/6*6)*6                                             AID21560
           JB=(PV-1)/6+2                                                AID21561
           L=LOC(PG)                                                    AID21562
           X=(X(L+JB) .RS. JS) .A. 77K                                  AID21563
           A=LIST(X)                                                    AID21564
           LOC(A)=L                                                     AID21565
BACK       M=X(L) .RS. 18                                               AID21566
           WHENEVER M .E. 0,TRANSFER TO GETIN                           AID21567
           X=(X(M+JB) .RS. JS) .A. 77K                                  AID21568
           B=LIST(X)                                                    AID21569
           WHENEVER A .NE. B                                            AID21570
           SIGN=1B                                                      AID21571
           TRANSFER TO INTO(SKIP)                                       AID21572
INTO(1)    LOC(B)=M                                                     AID21573
           SKIP=2                                                       AID21574
           SIGN=0B                                                      AID21575
INTO(2)    LAST(A)=L                                                    AID21576
           A=B                                                          AID21577
           X(L)=X(L) .A. 77777K                                         AID21578
           WHENEVER SIGN,X(LAST(A))=X(LAST(A)).V.(M .LS. 18)            AID21579
           END OF CONDITIONAL                                           AID21580
           L=M                                                          AID21581
           TRANSFER TO BACK                                             AID21582
          R                                                             AID21583
          R         STORE SPLIT    DATA INTO  MASTER  ARRAY.            AID21584
          R                                                             AID21585
GETIN      HI(PG)=GA                                                    AID21586
           LO(PG)=GB                                                    AID21587
           SIGN(PG)=1B                                                  AID21588
           INDEX(GA)=PV                                                 AID21589
           INDEX(GB)=PV                                                 AID21590
           ID(GA)=PG                                                    AID21591
           ID(GB)=PG                                                    AID21592
           TN(GA)=N                                                     AID21593
           TN(GB)=N(1)                                                  AID21594
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           TWT(GA)=W                                                    AID21595
           TWT(GB)=W(1)                                                 AID21596
           TY1(GA)=Y1                                                   AID21597
           TY1(GB)=Y1(1)                                                AID21598
           TY2(GA)=Y2                                                   AID21599
           TY2(GB)=Y2(1)                                                AID21600
          R                                                             AID21601
          R          PRINT PARTITION  INFORMATION - HOW  IT'S BEEN DONE.AID21602
          R                                                             AID21603
           PRINT FORMAT OUT6,PG,GA,GB,PX,ITR                            AID21604
           PRINT COMMENT $0     CODE     N    SUM OF WEIGHT      SUM  OFAID21605
          1  Y     SUM  Y-SQUARE       MEAN           STD.   DEV.       AID21606
          2    B  S  S$                                                 AID21607
          R                                                             AID21608
          R          COMPUTE  AND  PRINT  PARTITIONED  STATISTICS.      AID21609
          R                                                             AID21610
           C1=SAVE-1                                                    AID21611
           THROUGH KIYOI, FOR I=0, 1, I .G. C1                          AID21612
           Y=(Y4(I)-Y3(I)*Y5(I))/W1(I)                                  AID21613
           WHENEVER Y .G. 0.                                            AID21614
           Y=SQRT.(Y)                                                   AID21615
           OTHERWISE                                                    AID21616
           Y=0.                                                         AID21617
           END OF CONDITIONAL                                           AID21618
           PRINT FORMAT OUT4,KODE(I),N1(I),W1(I),Y3(I),Y4(I),Y5(I),Y,   AID21619
          1 BSSP(I)                                                     AID21620
KIYOI      CONTINUE                                                     AID21621
          R                                                             AID21622
           Y=(Y4(I)-Y3(I)*Y5(I))/W1(I)                                  AID21623
           WHENEVER Y .G. 0.                                            AID21624
           Y=SQRT.(Y)                                                   AID21625
           OTHERWISE                                                    AID21626
           Y=0.                                                         AID21627
           END OF CONDITIONAL                                           AID21628
           PRINT FORMAT OUT4,KODE(I),N1(I),W1(I),Y3(I),Y4(I),Y5(I),Y,   AID21629
          1 BSSP(64)                                                    AID21630
           SIGN=0B                                                      AID21631
          R                                                             AID21632
          R          END  OF  PARTITION                                 AID21633
          R          SEARCH  FOR  NEW  CANDIDATE  GROUPS.               AID21634
          R                                                             AID21635
SEARCH     TSS=0.0                                                      AID21636
           J=0                                                          AID21637
           TRANSFER TO TAMA(Z)                                          AID21638
TAMA(1)    PRINT COMMENT$4    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.$     AID21639
           PRINT COMMENT $0     GROUP        N       TOTAL WEIGHT       AID21640
          2    SUM OF Y         SUM  Y-SQUARE          T   S   S$       AID21641
TAMA(2)    THROUGH SAKU, FOR I=2, 1, I .G. NOGP                         AID21642
           WHENEVER SIGN(I), TRANSFER TO SAKU                           AID21643
           WHENEVER HI(I) .NE. 0, TRANSFER TO SAKU                      AID21644
           Y=TY2(I)-TY1(I)*TY1(I)/TWT(I)                                AID21645
          R          CHECK  GROUP  SIZE  AND  TEST  CRITERION 1.        AID21646
           WHENEVER Y.L.PA .OR. TN(I) .L. MSIZE                         AID21647
           SIGN(I)=1B                                                   AID21648
           TRANSFER TO SAKU                                             AID21649
           END OF CONDITIONAL                                           AID21650
          R          OUTPUT  OPTION 1 SWITCH  IS ON  IF  Z=1.           AID21651
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           TRANSFER TO HANA(Z)                                          AID21652
HANA(1)    PRINT FORMAT OUT7, I,TN(I),TWT(I),TY1(I),TY2(I),Y            AID21653
HANA(2)    WHENEVER Y .G. TSS                                           AID21654
           J=I                                                          AID21655
           TSS=Y                                                        AID21656
           END OF CONDITIONAL                                           AID21657
SAKU       CONTINUE                                                     AID21658
          R                                                             AID21659
          R          TEST  IF  FOUND  ANY  CANDIDATE  GROUPS.           AID21660
          R                                                             AID21661
           WHENEVER J .E. 0                                             AID21662
           PRINT FORMAT OUTJ, ITR,NOGP                                  AID21663
           TRANSFER TO ENDEN                                            AID21664
           END OF CONDITIONAL                                           AID21665
          R          NEW  PARENT  GROUP  WILL BE J.                     AID21666
           PG=J                                                         AID21667
           WHENEVER SIGN, TRANSFER TO BEGIN                             AID21668
           TRANSFER TO ENTER                                            AID21669
          R                                                             AID21670
          R          END   OF  ITERATIONS.                              AID21671
          R                                                             AID21672
ENDEN      PRINT COMMENT$0  ** THIS  IS  THE  END  OF  2ND  CORE.$      AID21673
           EXECUTE WRATIM.(0)                                           AID21674
          R                                                             AID21675
           EXECUTE  SEQPGM.(0)                                          AID21676
          R                                                             AID21677
EXIT       PRINT COMMENT $0  **  WE HAVE MORE THAN 127 GROUPS.  WHY  **$AID21678
           EXECUTE ERROR.                                               AID21679
          R                                                             AID21680
          R         FORMAT  SPECIFICATIONS.     SECOND  CORE.           AID21681
          R                                                             AID21682
           VECTOR VALUES DIM=2,1,2                                      AID21683
          R                                                             AID21684
           VECTOR VALUES OUT1=$20H4**  S T E P  NO. = ,I3,S9, 15HPARENT AID21685
          1 GROUP =,I3,3H ***$                                          AID21686
          R                                                             AID21687
           VECTOR VALUES OUT2=$1H0,S4, 8HVARIABLE,I4,12H  OVER GROUP,I4,AID21688
          132H  IS A CONSTANT.    S T E P   = ,I3,2H .*$                AID21689
          R                                                             AID21690
           VECTOR VALUES OUT3=$1H0,S4,19H TRY  ON   VARIABLE,I4,12H  OVEAID21691
          1R GROUP,I4,20H .  RESULTS  FOLLOW.*$                         AID21692
          R                                                             AID21693
           VECTOR VALUES OUT4=$1H ,S5,I3,S3,I4,S2,E15.8,S2,E15.8,S2,    AID21694
          1E15.8,S2,E15.8,S2,E15.8         /S108,E15.8*$                AID21695
          R                                                             AID21696
           VECTOR VALUES OUT5=$1H0,S4,21HFAILED TO SPLIT GROUP,I4,19H  TAID21697
          1RIED ON VARIABLE,I4,15H , BUT   BSS = ,E15.8*$               AID21698
          R                                                             AID21699
           VECTOR VALUES OUT6=$1H0/1H0,S4,15HDECOMPOSE GROUP,I4,12H  INTAID21700
          10 GROUP,I4, 5H  AND,I4,14H  BY VARIABLE ,I3,14H  IN  S T E P AID21701
          2,I4,2H . *$                                                  AID21702
          R                                                             AID21703
           VECTOR VALUES OUT7=$1H ,S4,I5,S5,I5,4(S5,E15.8)*$            AID21704
          R                                                             AID21705
           VECTOR VALUES OUTJ=$1H2,S4,66HTHAT IS ALL.  NO MORE GROUPS ARAID21706
          1E AVAILABLE. FINAL  S T E P  NO. IS,I4,S2,                   AID21707
          2 18H NO. OF GROUPS ARE, I5,2H . *$                           AID21708
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          R                                                             AID21709
           VECTOR VALUES OUTI=$1H ,S4,16H*  FOR  VARIABLE,I4,3H ( ,2C6, AID21710
          12H ),11H   B S S = , E15.8,S8, 11H BSS/TSS =  F8.5*$         AID21711
          R                                                             AID21712
           END OF PROGRAM                                               AID21713
$ASSEMBLE, PUNCH OBJECT                                         SAPTIM01AID21714
       ENTRY   WRATIM                                                   AID21715
SAVE   PZE                                                              AID21716
       PZE                                                              AID21717
       PZE                                                              AID21718
SIXTY  DEC     60                                                       AID21719
HRS    PZE                                                              AID21720
MIN    PZE                                                              AID21721
SEC    PZE                                                              AID21722
FRACT  PZE                                                              AID21723
WRATIM SXD     SAVE,4                                                   AID21724
       SXD     SAVE+1,2                                                 AID21725
       SXD     SAVE+2,1                                                 AID21726
       CALL    DAYTIM                                                   AID21727
       LRS     35                                                       AID21728
       DVP     SIXTY                                                    AID21729
       STO     FRACT                                                    AID21730
       ZAC                                                              AID21731
       DVP     SIXTY                                                    AID21732
       STO     SEC                                                      AID21733
       ZAC                                                              AID21734
       DVP     SIXTY                                                    AID21735
       STO     MIN                                                      AID21736
       ZAC                                                              AID21737
       DVP     SIXTY                                                    AID21738
       STO     HRS                                                      AID21739
       PRINT   FMA,HRS,MIN,SEC,FRACT,0                                  AID21740
OUT    LXD     SAVE,4                                                   AID21741
       LXD     SAVE+1,2                                                 AID21742
       LXD     SAVE+2,1                                                 AID21743
       TRA     2,4                                                      AID21744
FMA    BCI     *,12HOTIME IS NOW,4(I3,1H.)*                             AID21745
       END                                                              AID21746
$BREAK                                                                  AID21747
$ COMPILE MAD,PRINT OBJECT,PUNCH OBJECT                         AIDM2301AID21748
          R                                                             AID21749
          R          PROGRAM  NAME  --  A I D.    THIRD CORE.           AID21750
          R                                                             AID21751
          R              WRITTEN  BY  ROBERT  W   HSIEH.                AID21752
          R            AID  -  MODEL  2  -  REWRITTEN  ON  AUGUST 1963. AID21753
          R                                                             AID21754
           DIMENSION ID(128),INDEX(128),HI(128),LO(128),TN(128),TWT(128)AID21755
          1 ,TY1(128),TY2(128),CLASS(256,DIM),LOC(128),FAIL(128),MAX(36)AID21756
          2 ,P(36),TSS(128),BSS(128),MEAN(128),N(128) ,C(72), TYPE(36), AID21757
          3 NAME1(36),NAME2(36),X(20000),D(20000),V(100) ,LAB(12)       AID21758
          R                                                             AID21759
           PROGRAM COMMON NAME1,NAME2,NP,NV,LAB,                        AID21760
          0               ID,INDEX,HI,LO,TN,TWT,TY1,TY2,CLASS,LOC,FAIL, AID21761
          1 MAX,P,NOGP,ITR,ITRMAX,PA,PB,OP1,OP2,X,MSIZE,SCFIN,SCFOUT,   AID21762
          2  KONST,AA,BB,RUN,ZWANT,ZTYPE,ZTAPE,BOB,TYPE                 AID21763
          R                                                             AID21764
           EQUIVALENCE(K,KL),(X,D),(I,L),(SCFOUT,SFB)                   AID21765
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          R                                                             AID21766
           BOOLEAN KL ,L,SFB                                            AID21767
          R                                                             AID21768
           FLOATING POINT TWT,TY1,TY2,TSS,BSS,MEAN,SQRT.,N,R1,R2,Q,PA,PBAID21769
          1 ,D,NG,FY                                                    AID21770
          R                                                             AID21771
           NORMAL MODE IS INTEGER                                       AID21772
          R                                                             AID21773
           WHENEVER NOGP .LE. 1, TRANSFER TO ADIEU                      AID21774
          R                                                             AID21775
          R         PRINT  SUMMARY  AND  BASIC  STATISTICS.             AID21776
          R                                                             AID21777
           PRINT COMMENT$1                                   *  (A)UTOMAAID21778
          1TIC  (I)NTERACTION   (D)ETECTOR    (MODEL  2)  *$             AID21779
           PRINT COMMENT$0                                              AID21780
          1*  *  *  S  U  M  M  A  R  Y  *  *  *$                       AID21781
           PRINT FORMAT OUT1,ID,NAME1,NAME2,INDEX                       AID21782
           TWT=TN(1)                                                    AID21783
           MEAN(1)=TY1(1)/TWT(1)                                        AID21784
           N=TWT(1)-TWT(1)/TWT                                          AID21785
           TSS(1)=TY1(1)*MEAN(1)                                        AID21786
           TSS=TY2(1)-TSS(1)                                            AID21787
           BSS(1)=SQRT.(TSS/TWT(1))                                     AID21788
           PRINTFORMATOUT2,TN(1),MEAN(1),TY1(1),TSS,TWT(1),BSS(1),TY2(1)AID21789
           BSS=0.0                                                      AID21790
           NG=-1.0                                                      AID21791
          R                                                             AID21792
          R         PRINT  STATISTICS  FOR  EACH  GROUP  OBTAINED.      AID21793
          R                                                             AID21794
           THROUGH PGM,FOR I=2,1,I .G. NOGP                             AID21795
           C=0                                                          AID21796
           MAX=MAX(INDEX(I))                                            AID21797
           WHENEVER MAX .L. 36                                          AID21798
           THROUGH THIS,FOR J=0,1,J .G. MAX                             AID21799
           K=(CLASS(I,1) .RS. J) .A. 1                                  AID21800
                WHENEVER KL                                             AID21801
                C=C+1                                                   AID21802
                C(C)=J                                                  AID21803
THIS            END OF CONDITIONAL                                      AID21804
           OTHERWISE                                                    AID21805
                THROUGH IS,FOR J=0,1,J .G. 35                           AID21806
                K=(CLASS(I,1) .RS. J) .A. 1                             AID21807
                WHENEVER KL                                             AID21808
                C=C+1                                                   AID21809
                C(C)=J                                                  AID21810
IS              END OF CONDITIONAL                                      AID21811
                MAX=MAX/36                                              AID21812
                THROUGH AID   ,FOR J=0,1,J .G. MAX                      AID21813
                K=(CLASS(I,2) .RS. J) .A. 1                             AID21814
                WHENEVER KL                                             AID21815
                C=C+1                                                   AID21816
                C(C)=J+36                                               AID21817
AID             END OF CONDITIONAL                                      AID21818
           END OF CONDITIONAL                                           AID21819
           MEAN(I)=TY1(I)/TWT(I)                                        AID21820
           J=INDEX(I)                                                   AID21821
           PRINT FORMAT OUT3,I,ID(I),P(J),NAME1(J), NAME2(J)            AID21822
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

           WHENEVER C .NE. 0, PRINT FORMAT OUT4, C(1)...C(C)            AID21823
           Q=TY1(I)*MEAN(I)                                             AID21824
           WHENEVER HI(I) .E. 0                                         AID21825
           NG=NG+1.0                                                    AID21826
           BSS=BSS+Q                                                    AID21827
           PRINT COMMENT$   ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF AID21828
          1 FINALS.$                                                    AID21829
           END OF CONDITIONAL                                           AID21830
           D=MEAN(I)-MEAN(1)                                            AID21831
           TSS(I)=TY2(I)-Q                                              AID21832
           BSS(I)=SQRT. (.ABS.(TSS(I)/TWT(I)))                          AID21833
           R1=TWT(I)/TWT(1)*100.0                                       AID21834
           R2=TSS(I)/TSS                                                AID21835
           PRINT FORMAT OUT5,TN(I),MEAN(I),D,TY1(I),TWT(I),BSS(I),TSS(I)AID21836
          1 ,TY2(I),R1,Q,R2                                             AID21837
PGM        CONTINUE                                                     AID21838
          R                                                             AID21839
          R          PRINT  ANALYSIS  OF VARIANCE  TABLE.               AID21840
          R                                                             AID21841
           PRINT COMMENT$4                      * * *     ANALYSIS  OF  AID21842
          1VARIANCE  TABLF     * * *$                                   AID21843
           PRINT COMMENT$0      SOURCF OF           SUM  OF             AID21844
          1DEGREE OF          MEAN$                                     AID21845
           PRINT COMMENT$       VARIATION           SQUARES             AID21846
          1 FREEDOM          SQUARE             F$                      AID21847
           BSS=BSS-TSS(1)                                               AID21848
           D=TSS-BSS                                                    AID21849
           R1=BSS/NG                                                    AID21850
           Q=N-NG                                                       AID21851
           R2=D/Q                                                       AID21852
           MEAN=R1/R2                                                   AID21853
           PRINT FORMAT ABC, TSS,N,BSS,NG,R1,MEAN,D,Q,R2                AID21854
          R                                                             AID21855
          R          COMPUTE  RESIDUALS.                                AID21856
          R                                                             AID21857
           TRANSFER TO MIS(ZWANT)                                       AID21858
          R                                                             AID21859
          R          IDENTIFY  EACH  DATUM  WITH  ITS  GROUP  NUMBER.   AID21860
          R                                                             AID21861
MIS(1)     THROUGH BBC, FOR I=2,1,I.G.NOGP                              AID21862
           WHENEVER HI(I) .E.0                                          AID21863
           X=LOC(I)                                                     AID21864
ABCD       J=X(X).RS.18                                                 AID21865
           X(X)=I                                                       AID21866
           WHENEVER J .E. 0, TRANSFER TO BBC                            AID21867
           X=J                                                          AID21868
           TRANSFER TO ABCD                                             AID21869
BBC        END OF CONDITIONAL                                           AID21870
          R                                                             AID21871
           REWIND TAPE  4                                               AID21872
           REWIND TAPE  3                                               AID21873
          R                                                             AID21874
           K=1-KONST                                                    AID21875
           NN=NV+1                                                      AID21876
           V=1                                                          AID21877
          R                                                             AID21878
          R             SET  SCALE  FACTOR  SWITCH.                     AID21879
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          R                                                             AID21880
           FY=1.                                                        AID21881
           SQ=1                                                         AID21882
           WHENEVER SFB                                                 AID21883
           SQ=0                                                         AID21884
           FY=10.0.P.SCFOUT                                             AID21885
           END OF CONDITIONAL                                           AID21886
          R                                                             AID21887
          R          GET  DATA  FROMM  TAPE.                            AID21888
          R                                                             AID21889
CBC        READ BINARY TAPE   BB , KARD,V(1)...V(NV)                    AID21890
           WHENEVER KARD.E.$E$, TRANSFER TO LAST                        AID21891
           WHENEVER KARD.E.$YES$                                        AID21892
           K=K+KONST                                                    AID21893
           P=X(K)                                                       AID21894
           Y=K+1                                                        AID21895
           N=D(Y)                                                       AID21896
           Q=MEAN(P)                                                    AID21897
           D=N - Q                                                      AID21898
           TRANSFER TO SCALE(SQ)                                        AID21899
SCALE(0)   D=D*FY                                                       AID21900
SCALE(1)   WHENEVER D .GE. 0.                                           AID21901
           I=D + .5                                                     AID21902
           OTHERWISE                                                    AID21903
           I=D - .5                                                     AID21904
           END OF CONDITIONAL                                           AID21905
           WHENEVER .NOT. L, I=0                                        AID21906
           TRANSFER TO FACT(0)                                          AID21907
FACT(0)    Q=Q*FY                                                       AID21908
FACT(1)    WHENEVER Q .GE. 0.                                           AID21909
           J=Q + .5                                                     AID21910
           OTHERWISE                                                    AID21911
           J=Q - .5                                                     AID21912
           END OF CONDITIONAL                                           AID21913
           TRANSFER TO NBC                                              AID21914
           END OF CONDITIONAL                                           AID21915
           I=-0                                                         AID21916
           J=-0                                                         AID21917
           P=-0                                                         AID21918
          R                                                             AID21919
          R          SWITCH  FOR  RESIDUAL  OUTPUT.                     AID21920
          R                                                             AID21921
NBC        TRANSFER TO OUT(ZTYPE)                                       AID21922
OUT(2)     CONTINUE                                                     AID21923
          R                                                             AID21924
          R          RESIDUAL  ON  CARD.                                AID21925
          R                                                             AID21926
OUT(0)     PUNCH FORMAT PCHOUT,LAB(0),LAB(1),V(HI),P,J,V(ID),I,V(INDEX) AID21927
           TRANSFER TO MAD(ZTAPE)                                       AID21928
OUT(1)     CONTINUE                                                     AID21929
          R                                                             AID21930
          R          RESIDUAL  ON  TAPE .                               AID21931
          R                                                             AID21932
MAD(1)    V(NN)=I                                                       AID21933
           WRITE BINARY TAPE AA, KARD, V(1)...V(NN)                     AID21934
MAD(0)     TRANSFER TO CBC                                              AID21935
          R                                                             AID21936
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

LAST       PRINT COMMENT$4         RESIDUALS  ARE  OBTAINED.$           AID21937
           EXECUTE WRATIM.(0)                                           AID21938
           TRANSFER TO MAN(ZTYPE)                                       AID21939
MAN(0)     PRINT COMMENT$0           RESULTS  ARE  ON  CARDS.$          AID21940
           TRANSFER TO LEFIN                                            AID21941
MAN(1)     PRINT COMMENT$0           RESULTS  ARE  ON  TAPE.$           AID21942
          R                                                             AID21943
MET        WRITE BINARY TAPE  AA  , KARD,V(1)...V(NN)                   AID21944
           END OF FILE TAPE  AA                                         AID21945
          R                                                             AID21946
           REWIND TAPE  4                                               AID21947
           REWIND TAPE  3                                               AID21948
          R                                                             AID21949
           TRANSFER TO LEFIN                                            AID21950
MAN(2)     PRINT COMMENT$0           RESULTS  ARE  BOTH  ON  CARD  AND  AID21951
          1TAPE.$                                                       AID21952
           TRANSFER TO MET                                              AID21953
MIS(0)     PRINT COMMENT$4         RESIDUALS  ARE  NOT  REQUESTED.$     AID21954
LEFIN      PRINT COMMENT$4                                     *  *  *  AID21955
          1    E    N    D      *  *  *$                                AID21956
          R                                                             AID21957
           EXECUTE WRATIM.(0)                                           AID21958
           TRANSFER TO EXIT                                             AID21959
          R                                                             AID21960
ADIEU      PRINT COMMENT$4      ORIGINAL  GROUP  HAS  NO  SUBGROUPS.$   AID21961
          R                                                             AID21962
EXIT       EXECUTE SELPGM.(1)                                           AID21963
          R                                                             AID21964
          R        FORMAT  SPECIFICATIONS.        THIRD  CORE           AID21965
          R                                                             AID21966
           VECTOR VALUES DIM=2,1,2                                      AID21967
          R                                                             AID21968
           VECTOR VALUES OUT1=$1H0,23H     DEPENDENT VARIABLE,I4,4H  ( ,AID21969
          1 2C6,1H)/1H0,S5,22HWEIGHTED  BY  VARIABLE, I5*$              AID21970
          R                                                             AID21971
           VECTOR VALUES OUT2=$1H0,15H** TOTAL  GROUP/1H0,S10,4HN  = ,  AID21972
          1I12,S13,6HMEAN =,E15.8,S12,7HSUM Y =,E15.8,S8,6H TSS =,E15.8/AID21973
          21H ,14H TOTAL WT SUM=,F12.0,S8,11HSTD. DEV. =,E15.8,S8,      AID21974
          311HSUM Y SQ. =,E15.8*$                                       AID21975
          R                                                             AID21976
           VECTOR VALUES OUT3=$1H0,15H *   GROUP  NO.,I4,20H  SPLIT  FROAID21977
          1M  GROUP,I4,15H  ON  VARIABLE I4, 2H (,2C6,1H)*$             AID21978
          R                                                             AID21979
           VECTOR VALUES OUT4=$1H ,S5,36HVALUES  OF  PREDICTOR  INCLUDEDAID21980
          1  ARE,18I5/1H ,S5,25I5*$                                     AID21981
          R                                                             AID21982
           VECTOR VALUES OUT5=$1H ,14H          N  =,I12,S13,6HMEAN =,  AID21983
          1E15.8,S8,17HGROUP DEVIATION =,E15.8,S8,7HSUM Y = E15.8/1H ,  AID21984
          214H  WEIGHT SUM =,F12.0,S8,11HSTD. DEV. =,E15.8,S17,8HTSS(I) AID21985
          3=,E15.8,S4,11HSUM Y SQ. =,E15.8/1H ,14HPCT OF TOTAL =,F8.1,  AID21986
          4 S8,15HWTD. MEAN SQ. =,E15.8,S8,17H(TSS(I)/TSS(T)) =,E15.8*$ AID21987
          R                                                             AID21988
           VECTOR VALUES ABC=$1H0,S6,9HT O T A L,S7,E15.8,S7,F9.0 /1H0, AID21989
          1 S6,7HBETWEEN,S9,E15.8,S7,F9.0,S7,E15.8,S5,E15.8/1H0,S6,     AID21990
          2 6HWITHIN,S10,E15.8,S7,F9.0,S7,E15.8*$                       AID21991
          R                                                             AID21992
           VECTOR VALUES PCHOUT=$3C6,I3,I8,I6,I8,I7*$                   AID21993
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

          R                                                             AID21994
           END OF PROGRAM                                               AID21995
$ASSEMBLE, PUNCH OBJECT                                         SAPTIM01AID21996
       ENTRY   WRATIM                                                   AID21997
SAVE   PZE                                                              AID21998
       PZE                                                              AID21999
       PZE                                                              AID22000
SIXTY  DEC     60                                                       AID22001
HRS    PZE                                                              AID22002
MIN    PZE                                                              AID22003
SEC    PZE                                                              AID22004
FRACT  PZE                                                              AID22005
WRATIM SXD     SAVE,4                                                   AID22006
       SXD     SAVE+1,2                                                 AID22007
       SXD     SAVE+2,1                                                 AID22008
       CALL    DAYTIM                                                   AID22009
       LRS     35                                                       AID22010
       DVP     SIXTY                                                    AID22011
       STO     FRACT                                                    AID22012
       ZAC                                                              AID22013
       DVP     SIXTY                                                    AID22014
       STO     SEC                                                      AID22015
       ZAC                                                              AID22016
       DVP     SIXTY                                                    AID22017
       STO     MIN                                                      AID22018
       ZAC                                                              AID22019
       DVP     SIXTY                                                    AID22020
       STO     HRS                                                      AID22021
       PRINT   FMA,HRS,MIN,SEC,FRACT,0                                  AID22022
OUT    LXD     SAVE,4                                                   AID22023
       LXD     SAVE+1,2                                                 AID22024
       LXD     SAVE+2,1                                                 AID22025
       TRA     2,4                                                      AID22026
FMA    BCI     *,12H0TIME IS NOW,4(I3,1H.)*                             AID22027
       END                                                              AID22028
$BREAK                                                                  AID22029
$DATA                                                                   AID22030
1   AID 3TEST   A I D  -  MODEL  2.   TEST  RUN- ON  3/8/64.
2     C    90     7
3     4     6  .02  .005  50 20      5  I N C O ME
4   1 F    A  G  E      2 M EDUCATION     3 F  R A C E       4 F OCCUPATION
  (C1, S3, 4I1, I4, I2,S60,C6*)
DATAFOLLOWS
71901111299008                                                                01
71901211401009                                                                02
71901311502011                                                                03
71901321200008                                                                04
71901321202008                                                                05
71901112301012                                                                06
71901212400011                                                                07
71901312500009                                                                08
71901113299008                                                                09
71901213401011                                                                10
71901313502009                                                                11
71901114301012                                                                12
71901214400009                                                                13
71901314500011                                                                14
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

71901115250009                                                                15
71901215300009                                                                16
71902111350010                                                                17
71902211450008                                                                18
71902311601008                                                                19
71902112350010                                                                20
71902212450009                                                                21
71902312601012                                                                22
71902322202009                                                                23
71902322200009                                                                24
71902113350010                                                                25
71902213450011                                                                26
71902313600011                                                                27
71902114350010                                                                28
71902214450012                                                                29
71902314600009                                                                30
71902115250011                                                                31
71902215300011                                                                32
71903111500010                                                                33
71903211550010                                                                34
71903311701010                                                                35
71903112500010                                                                36
71903212550010                                                                37
71903312700008                                                                38
71903113503010                                                                39
71903213550010                                                                40
71903313699010                                                                41
71903223200010                                                                42
71903223202010                                                                43
71903114503010                                                                44
71903214550010                                                                45
71903314700012                                                                46
71903115250009                                                                47
71903215300009                                                                48
71904111579008                                                                49
71904211620010                                                                50
71904311801010                                                                51
71904112580012                                                                52
71904212620010                                                                53
71904312801010                                                                54
71904113581008                                                                55
71904213630010                                                                56
71904313800010                                                                57
71904114580012                                                                58
71904214630010                                                                59
71904314800010                                                                60
71904224202011                                                                61
71904224200011                                                                62
71904115250011                                                                63
71904215300010                                                                64
71905111570010                                                                65
71905211640010                                                                66
71905311960010                                                                67
71905112580010                                                                68
71905212650010                                                                69
71905312960010                                                                70
71905113560010                                                                71
```

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890

71905213650010                                                                72
71905313950010                                                                73
71905114570010                                                                74
71905214660010                                                                75
71905314950010                                                                76
71905124200012                                                                77
71905124202012                                                                78
71905115250010                                                                79
71905215300011                                                                80
71906111101012                                                                81
71906311100012                                                                82
71906112100011                                                                83
71906212101011                                                                84
71906113101010                                                                85
71906213100010                                                                86
71906114100009                                                                87
71906224100008                                                                88
71906215101009                                                                89
71906125101008                                                                90
E
‡                  750032                DECK
  2130
```

APPENDIX I

ON TRANSFERRING AID (2) TO ANOTHER COMPUTER

The program was written for a 32k IBM 7090 with an on-line clock which is interrogated by the program. The U.M. 7090 has a core-protect device. Any transfer to another computer will have to take these factors into account. In general, there will be few problems with tape limitations on other equipment, since the program uses only five tapes as follows: BCD input, BCD output, two scratch tapes and one program segmentation (ping-pong) tape.

The program will run in its present form on any 32k IBM 709 or 7090 computer capable of accepting the University of Michigan Executive (MAD) system, September 1963 version.

The program is written in MAD (not Fortran) and uses several subroutines written in IMAP, in addition to input-output and other subroutines supplied by the Executive System.

Thus, if the potential user has access to an IBM 709 or 7090 system and if his computing center administration can operate, at least part of the time, under the U. of M. Executive System, the present MAD program may be used. Since the program is primarily written in MAD, it cannot be used in its present form on a computer which does not have a MAD translator implemented for it. This would require re-writing the source program in FORTRAN, ALGOL or some other suitable language. This can be done, but would require considerable programming skill, and a knowledge of both MAD and FORTRAN or ALGOL. It is estimated that an equivalent FORTRAN program would be somewhat larger than the MAD program.

However, complete documentation in the form of descriptions of storage allocation, flow charts, listings, etc., are provided in this document.

A potential user should:

1. Determine from his computing center whether it has the University of Michigan Executive System available for use, and if not, whether it can be obtained. (It is available from the IBM user's organization called SHARE.)

2. If this is the case, an IBM 1401-compatible tape (1/2 inch, 200 or 556 bpi density) may be shipped to:

   Data Processing Section
   Institute for Social Research
   The University of Michigan
   Box 1248
   Ann Arbor, Michigan 48106

   together with a request for AID 2. Desired tape density must be specified. A small charge will be made to cover handling and shipping costs. The symbolic program, and test data will be written on tape and shipped as unblocked 80-character BCD records in the desired density.

3. If the necessary equipment or Executive System is not available, the documentation presented here should be sufficient for conversion of the program to another computer of suitable size.

For either use, the following materials would be useful:

1. University of Michigan Executive System for the IBM 7090 Computer, University of Michigan Computing Center, September 1963.

2. Michigan Algorithmic Decoder, Bruce Arden, Bernard Galler, Robert Graham, University of Michigan Computing Center, January 1963.

Disclaimer

Although this program has been tested thoroughly by its programmer, no warranty, express or implied, is made by the programmer or the Institute for Social Research or the University of Michigan as to the accuracy and functioning of the program and related program material, and no responsibility is assumed by the programmer, the Institute or the University in connection therewith.

APPENDIX J

PROBLEMS IN THE ANALYSIS OF SURVEY DATA, AND A PROPOSAL**

JAMES N. MORGAN AND JOHN A. SONQUIST*
University of Michigan

Most of the problems of analyzing survey data have been reasonably well handled, except those revolving around the existence of interaction effects. Indeed, increased efficiency in handling multivariate analyses even with non-numerical variables, has been achieved largely by assuming additivity. An approach to survey data is proposed which imposes no restrictions on interaction effects, focuses on importance in reducing predictive error, operates sequentially, and is independent of the extent of linearity in the classifications or the order in which the explanatory factors are introduced.

A. NATURE OF THE DATA AND THE WORLD FROM WHICH THEY COME

THE increasing availability of rich data from cross section surveys calls for more efficient methods of data scanning and data reduction in the process of analysis. The purpose of this paper is to spell out some of the problems arising from the nature of the data and the nature of the theories which are being tested with the data, to show that present methods of dealing with these problems are often inadequate, and to propose a radical new method for analyzing survey data. There are seven things about the data or about the world from which they come which need to be kept in mind.

First, there is a wide variety of information about each person interviewed in a survey. This is good, because human behavior is motivated by more than one thing. But the very richness of the data creates some problems of how to handle them.

Second, we are dealing not with variables for the most part, but with classifications. These vary all the way from age, which can be thought of as a variable put into classes, to occupation or the answers to attitudinal questions, which may not even have a rank order in any meaningful sense. Even when measures seem to be continuous variables, such as age or income, there is good reason to believe that their effects are not linear. For instance, people earn their highest incomes in the middle age ranges. Expenditures do not change uniformly with changes in income at either extreme of the income scale.

Third, there are errors in all the measures, not just in the dependent variable, and there is little evidence as to the size of these errors, or as to the extent to which they are random.

Fourth, the data come from a sample and generally a complex one at that. Hence, there is sample variability piled on top of measurement error. The fact that almost all survey samples are clustered and stratified leads to problems of the proper application of statistical techniques. Statistical tests usually assume simple random samples rather than probability samples. More ap-

---

* The authors are indebted to many individuals for advice and improvements. In particular, Professor L. J. Savage noticed that some interactions would remain hidden, and Professor William Ericson proved that locating the best combination of subclasses of a single code was simple enough to incorporate into the program. A Ford Foundation grant to the Department of Economics of the University of Michigan supported the author's work on some substantive problems which led to the present focus on methods. Support from the Rockefeller Foundation is also gratefully acknowledged.

**Reprinted by permission from the Journal of the American Statistical Association, 58 (June 1963), 415-35.

propriate tests have been developed for simple statistics such as proportions, means, and a few others.

Fifth, and extremely important, there are intercorrelations between many of the explanatory factors to be used in the analysis—high income goes along with middle age, with advanced education, with being white, with not being a farmer, and so forth. This makes it difficult to assess the relative importance of different factors, since their intercorrelations get in the way. Since many of them are classifications rather than continuous variables, it is not even easy to measure the extent of the intercorrelation. Measures of association for cross classification raise notoriously difficult problems which have not really been solved in any satisfactory way.[1]

Sixth, there is the problem of interaction effects. Particularly in the social sciences, there are two powerful reasons for believing that it is a mistake to assume that the various influences are additive. In the first place, there are already many instances known of powerful interaction effects—advanced education helps a man more than it does a woman when it comes to making money; and it does a white man more good than a Negro. The effect of a decline in income on spending depends on whether the family has any liquid assets which it can use up. Women have their hospitalizations at different ages than men. Second, the measured classifications are only proxy variables for other things and are frequently proxies for more than one construct. Several of the measured factors may jointly represent a theoretical construct. We may have interaction effects not because the world is full of interactions, but because our variables have to interact to produce the theoretical constructs that really matter. The idea of a family life cycle, unless arbitrarily created out of its components in advance, is a set of interactions between age, marital status, presence, and age of children.[2] It is therefore often misleading to look at the over-all gross effects of age or level of education. Where interaction effects exist, the concept of a main effect is meaningless, and it is our belief that in human behavior there are so many interaction effects that we must change our approach to the problems of analysis.

Another example of interaction effects appeared in the attempt to build equivalent adult scales to represent the differences in living expenses of families of different types. After many years of analysis, one of the most recent studies in this field has concluded "when its size changes, families' tastes appear to change in more complicated ways than visualized by our hypothesis."[3] More

---

[1] One seemingly appropriate measure for two classifications both being used to predict the same variable is one called lambda suggested by Goodman and Kruskal. With many kinds of survey data this measure, which assumes that an absolute prediction has to be made for each individual, is too insensitive to deal with situations where each class on the predicting characteristic has the same modal class on the other characteristic that is to be predicted. An effective and properly stochastic measure would be derived by assigning a one-zero dummy variable to belonging to each class of each of the two characteristics and then computing the canonical correlation between the two sets of dummy variables.

See Leo A. Goodman and William H. Kruskal, "Measures of association for cross classifications," *Journal of the American Statistical Association*, 49 (December, 1954), 732–64.

[2] John B. Lansing and James N. Morgan, "Consumer finances over the life cycle," in *Consumer Behavior*, Volume II, L. Clark (Editor) (New York: New York University Press, 1955).

See also Leslie Kish and John B. Lansing, "Family life cycle as an independent variable," *American Sociological Review*, XXII (October, 1957), 512–9.

[3] In other words family composition had different effects on different expenditures. F. G. Forsythe, "The relationship between family size and family expenditure," *Journal of the Royal Statistical Society*, Series A, vol. 123 (1961), 367–97, quote from p. 386.

recently in analyzing factors affecting spending unit income, it has become obvious that age and education cannot operate additively with race, retired status, and whether the individual is a farmer. The attached table illustrates this with actual average incomes for a set of nonsymmetrical groups. The twenty-one groups account for two-thirds of the variance of individual spending unit incomes, whereas assuming additivity for race and labor force status even with joint age-education variables produces a regression which with 30 variables accounts for only 36 per cent of the variance. A second column in the

TABLE 1. SPENDING UNIT INCOME AND THE NUMBER IN THE UNIT WITHIN VARIOUS SUBGROUPS

Group	Spending unit average (1958) income	Number in unit	Number of cases
Nonwhite, did not finish high school	$ 2489	3.3	191
Nonwhite, did finish high school	5005	3.4	67
White, retired, did not finish high school	2217	1.7	272
White, retired, did finish high school	4520	1.7	72
White, nonretired farmers, did not finish high school	3950	3.6	87
White nonretired farmers, did finish high school	6750	3.6	24
The Remainder			
0–8 grades of school			
18–34 years old	4150	3.8	72
35–54 years old	4670	3.8	240
55 and older—not retired	4846	2.2	208
9–11 grades of school			
18–34 years old	5032	3.7	112
35–54 years old	6223	3.4	202
55 and older—not retired	4720	2.1	63
12 grades of school			
18–34 years old	5458	3.3	193
35–54 years old	7765	3.8	291
55 and older—not retired	6850	2.0	46
Some college			
18–34 years old	5378	3.0	102
35–54 years old	7930	3.8	112
55 and older—not retired	8530	2.0	36
College graduates			
18–34 years old	7520	3.8	80
35–54 years old	8866	2.9	150
55 and older—not retired	10879	1.8	34

Source: 1959 Survey of Consumer Finances.

table gives the average number of people in the unit, and it can be seen that this particular breakdown is not particularly useful for analyzing the number of people in a unit. On the other hand, if each group were to be used to analyze expenditure behavior, income, and family size are likely to operate jointly rather than additively.

In view of the fact that intercorrelation among the predictors on the one hand and interaction effects on the other are frequently confused, it seems useful to give a pictorial example indicating both the differences between them and the way in which they operate when both are present. Our concern is not with statistical tests to distinguish between them, but with the effects of ignoring their presence.

Chart I shows pictorially three cases, real but exaggerated. First, there is a case where the two explanatory factors, income and education, are correlated with one another, but do not interact. Second, a case where income and being self-employed interact with one another but are not correlated, and third, a situation where income and asset holdings are correlated with one another and also interact in their effect on saving. The ellipsoids represent the area where most of the dots on a scatter diagram would appear. In the first case, it is clear that a simple relation between income and saving would exaggerate the effect of income on saving by failing to allow for the fact that high income people have more education, and that highly educational people also save more. An ordinary multiple regression, however, using a dummy variable representing high education would adequately handle this difficulty. In the second case there is no particular correlation, we assume, between income and being self-employed, but the self-employed have a much higher marginal propensity to save than other people. Here, the simple relationship between income and saving becomes a weighted compromise between the two different effects that really exist. A multiple correlation would show no effect of being self-employed and the same compromise effect of income. Only a separate analysis for the self-employed and the others would reveal the real state of the world. In the third case, not only do the high-asset people have a higher marginal propensity to save, but they also tend to have a higher income. Multiple correlation clearly will not take care of this situation in any adequate way. It *will* produce an "income effect" which can be added to an "asset effect" to produce an estimate of saving. Here the income effect is an average of two different income effects. The estimated asset effect is likely to come out closer to zero than if income had been ignored. Of course, where interactions exist, there is little use in attempting to measure separate effects.

Finally, there are logical priorities and chains of causation in the real world. Some of the predicting characteristics are logically prior to others in the sense that they can cause them but cannot be affected by them. For instance, where a man grows up may affect how much education he gets, but his education cannot change where he grew up. We are not discussing here the quite different analysis problem where the purpose is not to explain one dependent variable but to untangle the essential connections in a network of relations.

In dealing with a single dependent variable representing some human behavior, we might end up with at least three stages in the causal process—early childhood and parental factors, actions and events during the lifetime, and current situational and attitudinal variables. If this were the end of the problem we could simply run three separate analyses. The first would analyze the effects of early childhood and parental factors. The second would take the residuals from this analysis and analyze them against events during a man's lifetime up until the present, and the third would take the residuals from the

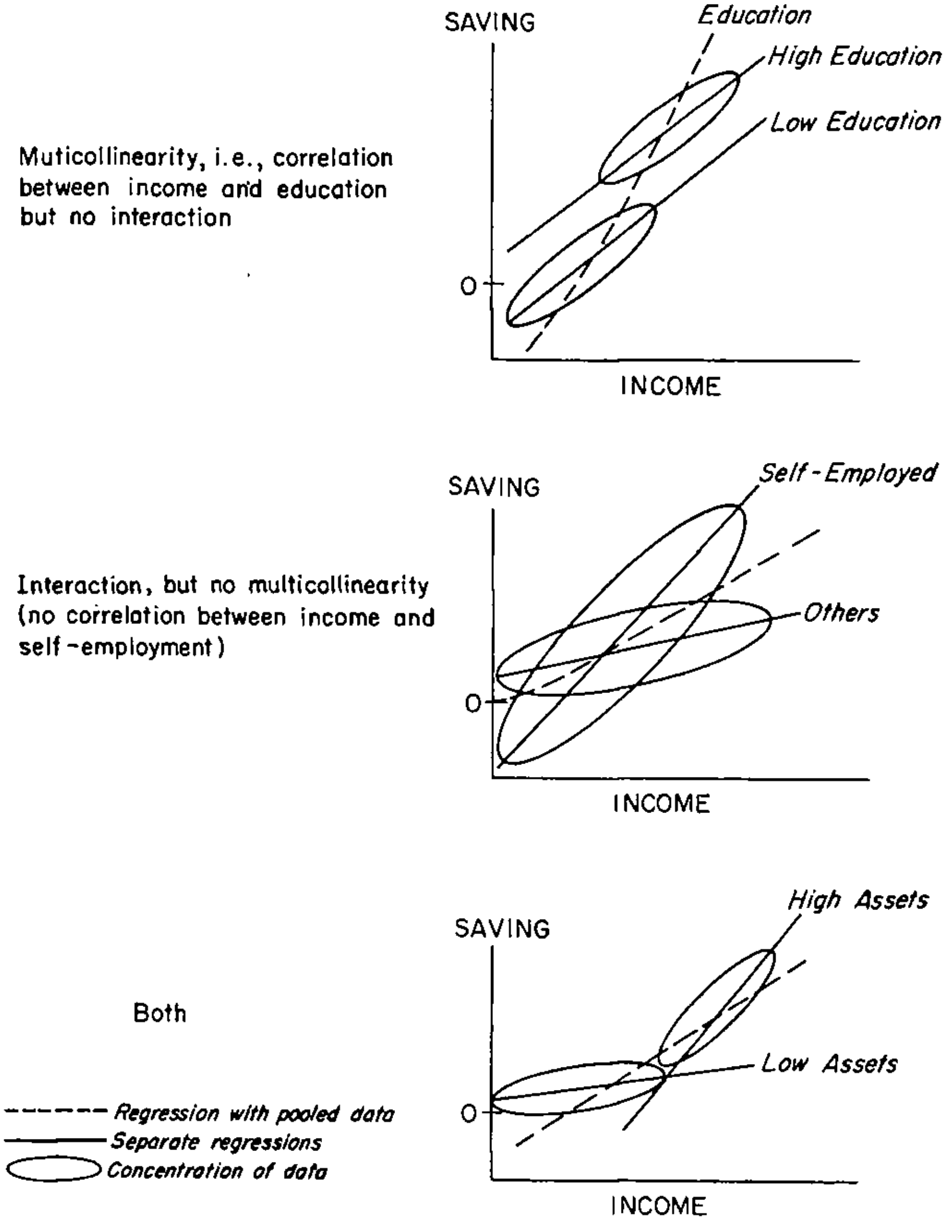


CHART I. Combinations of Multicollinearity and Interaction and Their Effects.

second analysis and analyze them against current situational and attitudinal variables. But the real world is not even that simple, because some of the same variables which are logically prior in their direct effects may also tend to mediate the effect of later variables. For instance, a man's race has a kind of logical priority to it, but at the same time it may affect the way other things such as the level of his education operate to determine his income.

This is an impressive array of problems. Before we turn to a discussion of current attempts to solve these problems and to our own suggestions, it is essential to ask first what kind of theoretical structure is being applied and what the purposes of analysis are.

B. NATURE OF THE THEORY AND PURPOSES OF ANALYSIS

Perhaps the most important thing to keep in mind about survey data in the social sciences is that the theoretical constructs in most theory are not identical with the factors we can measure in the survey. The simple economic idea of ability to pay for any particular commodity is certainly a function not only of income but of family size, other resources, expected future income, economic security, and even extended family obligations. A man's expectations about his own economic future, which we may theorize will affect his current behavior, might be measured by a battery of attitudinal and expectational questions or by looking at his education, occupation, age, and the experience of others in the same occupation and education group who are already older. The fact that the theoretical constructs in which we are interested are not the same as the factors we can measure, nor even simply related to them, should affect our analysis techniques and focus attention on creating or locating important interaction effects to represent these constructs.

Second, there are numerous hypotheses among which a selection is to be made. Even if the researcher preferred to restrict himself to a single hypothesis and test it, the intercorrelations among the various explanatory factors mean that the same result might support any one of several hypotheses.[4] Hence, comparisons of relative importance of predictors, and selecting those which reduce predictive errors most, are required.

When we remember that there are also variable errors of measurement, the problem of selecting between alternative hypotheses becomes doubly difficult, and ultimately requires the use of discretion on the part of the researcher. Better measurement of a factor might increase its revealed importance.

Finally, researchers may have different reasons why they wish to predict individual behavior. Most will want to predict behavior of individuals in the population, not just in the sample, which makes the statistical problem somewhat more complicated. But some may also want to focus on the behavior of some crucial individuals by assigning more weight to the behavior of some rather than others. Others may want to test some explanatory factors, however small their apparent effect, because they are important. They may be important because they are subject to public policy influences or because they

---

[4] For an excellent statement of the application of this problem to the economists' concern with the permanent income hypothesis versus the relative income hypothesis, see Jean Crockett, "Liquid assets and the theory of consumption" (New York: National Bureau of Economic Research, 1962) (mimeographed).

are likely to change over time, or because they are crucial to some larger theoretical edifice. The nature of these research purposes thus combines with the nature of the data and their characteristics to make up the problem of how to analyze the data.

C. THE STRATEGY CHOICE IN ANALYSIS

One can think of a series of strategies ranging from taking account of only the main effects of each explanatory classification separately or jointly, to trying to take account of all possible combinations of all the classifications at once. Even if there were enough data to allow the last, however, it would not be of much use. The essence of research strategy then consists of putting some restrictions on the process in order to make it manageable. One possibility is to cut the number of explanatory factors utilized, and another is to restrict the freedom with which we allow them to operate.[5] One might assume away most or all interaction effects, for instance, and keep a very large number of explanatory classifications. Still further reduction in the number of variables is possible, if one assumes linearity for measured variables or, what amounts to the same thing builds arbitrary scales, incestuously derived out of the same data in order to convert each classification into a numerical variable. Clearly, the more theoretical or statistical assumptions one is willing to impose on the data, the more he can reduce the complexity of the analysis. A difficulty is that restrictions imposed in advance cannot be tested. There seems some reason to argue that it would be better to use an approach which developed its restrictions as it went along. In any case keeping these problems in mind we turn now to a summary of how analysis problems in using survey data are currently being handled and some of the difficulties that present methods still leave unsolved.

D. HOW PROBLEMS IN ANALYSIS ARE CURRENTLY BEING HANDLED—AN APPRAISAL

We take the seven problems in section A in the same order in which they are presented there plus the major problem in section B, that of theoretical constructs not measured directly by the factors on which we have data. The first problem was the existence of many factors. The simplest procedure has been to look at them one at a time always keeping in mind the extent to which one factors is intercorrelated with others. Another technique, particularly with attitudes, has been to build indexes or combinations of factors either arbitrarily or with the use of some sort of factor analysis technique.[6] The difficulty is that the first of these is quite arbitrary, and the second is arbitrary in a different sense, in that most mechanical methods of combining factors are based on the intercorrelations between the factors themselves and not in the way in which they may affect the dependent variable. It is quite possible for two highly correlated factors to influence the dependent variable in opposite ways. Building a combination of the two only on the basis of their intercorrelation would create a factor which would have no correlation at all with the dependent

---

[5] For a discussion of alternative strategies made while commenting on a series of papers, see James Morgan. "Comments," in *Consumption and Saving*, Volume I, I. Friend and R. Jones (Editors.) (Philadelphia: University of Pennsylvania Press, 1960), pp. 276–84.

[6] Charles Westoff and others, *Family Planning in Metropolitan America* (Princeton: Princeton University Press, 1961).

variable. With highly correlated attitudes, however, some such reduction to a few factors may be required and meaningful.

With the advent of better computing machinery, the problem of multiple factors has frequently been handled by using multiple correlation techniques. The use of these techniques, of course, required solving the second problem, that arising from the fact that in many cases we have classifications rather than continuous variables. This has been done in two ways, first, by building arbitrary scales. For instance, one could assign the numbers one, two, three, four, five, and six to the six age groups in order. Or if age were being used to predict income, one could assign a set of numbers representing the average income of people in those age groups.[7] But unless machine capacity is extremely limited, a far more flexible method which is coming into favor is to use what have been called dummy variables.[8] The essence of this technique is to assign a dummy variable to each class of a characteristic except one. It is called a dummy variable because it takes the value one if the individual belongs in that subclass or a zero if he does not. If ordinary regression procedures are to be used, of course, dummy variables cannot be assigned to every subclass of any characteristic, since this would overdetermine the system. However, at the Survey Research Center we have developed an iterative program for the IBM 7090, the output of which consists of coefficients for each subclass of each characteristic, the set for each characteristic having a weighted mean of zero. This means that the predicting equation has the over-all mean as its constant term, and an additive adjustment for each characteristic, depending on the subclass into which the individual falls on that characteristic. This is the standard analysis of variance formulation when all interactions are assumed to be zero. Of course, the coefficients of dummy variables using a regular matrix inversion routine can easily be converted into sets of this sort. There remain two difficulties with this technique. One is the problem of interaction effects, which are either assumed away or have to be built in at the beginning in the creation of the classes. A second arises from the nature of the classifications frequently used in survey data. Even though association between, say, occupation and the incidence of unemployment faced by an individual is not terribly high, the occupation code generally includes one or two categories such as the farmers and the retired who, by definition, cannot be unemployed at all. When dummy variables are assigned to these classes, it may easily occur that there is a perfect association between a dummy variable representing one of these peculiar (not applicable) groups in one code and a dummy variable representing something else in another classification (not unemployed). If the researcher omits one of each such pair of dummy variables in a regression routine, he is all right.

A third problem, that of errors in the data, is generally handled by not re-

---

[7] For an example see Jerry Miner, "Consumer Personal Debt—An Intertemporal Analysis," in *Consumption and Saving*, Volume II, I. Friend and R. Jones (Editors) (Philadelphia: University of Pennsylvania Press, 1960), 400–61.

[8] Daniel Suits, "The Use of Dummy Variables in Regression Equations," *Journal of the American Statistical Association*, 52 (December, 1957), 548–51.

T. P. Hill, "An Analysis of the Distribution of Wages and Salaries in Great Britain," *Econometrica*, 27 (July, 1959), 355–81.

jecting hypotheses too easily and by attempting to use some judgment in the assessment of relative importance of different factors or different hypotheses keeping in mind the accuracy with which the variables have probably been measured.

The fact that the data come from a sample has frequently been ignored. As the analysis techniques become more complicated, it becomes almost impossible to keep the structure of the sample in mind too. However, there is some reason to believe that the clustering and stratification of the sample become less and less important the more complex and more multivariate the analysis being undertaken.[9]

What about intercorrelations among the predictors? The main advantage of multivariate techniques like multiple regression is that they take care of these intercorrelations among the predictors, at least in a crude sense. Indeed, if one compares an ordinary subclass mean with the multivariate coefficient of the dummy variable associated with belonging to that subclass, the difference between the two is the result of adjustments for intercorrelations. Where these differences seem likely to be the result of a few major interrelations, some statement as to the factors correlated with the one in question (and responsible for the attenuation of its effect on the multivariate analysis) are often given to the reader. It is, of course, true that where intercorrelations between two predictors are too high, no analysis can handle this problem, and it becomes necessary to remove one of them from the analysis.

Perhaps the most neglected of the problems of analysis has been the problem of interaction effects. The reason is very simple. The assumption that no interactions exist generally leads to an extremely efficient analysis procedure and a great reduction in the complexity of the computing problem. Those of us who have looked closely at the nature of survey data, however, have become increasingly impressed with the importance of interaction effects and the useful way in which allowing for interactions between measured factors gets us closer to the effects of more basic theoretical constructs. Where interaction effects have not been ignored entirely, they have been handled in a number of ways. They can be handled by building combination predictors in the first place, such as combinations of age and education or the combination of age, marital status, and children known as the family life cycle.[10] Sometimes where almost all the interactions involve the same dichotomy, two separate analyses are called for.[11] Interactions are also handled by rerunning the analysis for

---

[9] Actually there are no formulas available for sampling errors of many of the statistics from complex probability samples. Properly selected part-samples can be used to estimate them by a kind of hammer-and-tongs procedure, but this is expensive. See Leslie Kish, "Confidence intervals for clustered samples," *American Sociological Review*, 22 (April, 1957), 154–65. So long as the samples are representative of a whole population the basic statistical model is presumably the "fixed" one, see M B. Wilk and O. Kempthorne, "Fixed, mixed, and random models," *Journal of the American Statistical Association*, 50 (December, 1955), 1144–67.

See also L. Klein and J. Morgan, "Results of alternative statistical treatments of sample survey data," *Journal of the American Statistical Association*, 46 (December, 1951), 442–60.

[10] Guy Orcutt and others, *Microanalysis of Socioeconomic Systems* (New York: Harper and Brothers, 1961).

[11] For instance, hospital utilization was studied separately for men and women in Grover Wirick, Robin Barlow, and James Morgan, "Population survey: Health care and its financing," *Hospital and Medical Economics*, Volume I, Walter McNerney (Editor) (Chicago: American Hospital Association, 1962).

Participation in recreation was studied separately for those with and without paid vacations; see Eva Mueller and Gerald Gurin, *Participation in Outdoor Recreation: Factors Affecting Demand Among American Adults* (U.S. (U.S.G.P.O., ORRRC Study Report 20, 1962.)

some subgroup of the population. In a recent study of factors affecting hourly earnings, for instance, the analysis was rerun for the white, nonfarmer males only, to test the hypothesis that some of the effects like that of education were different for the non-whites, women, and farmers.[12] A difficulty with this technique, of course, is that if one merely wants to see whether the interaction biases the estimates for the whole population seriously, one reruns the analysis with the group that makes up the largest part of the sample. But if one wants to know whether there are different patterns of effects for some small subgroup, the analysis must be run for that small subgroup.

Another method of dealing with interaction effects is to look at two- and three-way tables of residuals from an additive multivariate analysis. This requires the process, often rather complicated and expensive, of creating the residuals from the multivariate analysis and then analyzing them separately.[13] Where some particular interaction is under investigation, an effective alternative is to isolate some subgroup on a combination of characteristics such as the young, white, college graduates. It is then possible to derive an estimate of the expected average of that subgroup on the dependent variable by summing the multivariate coefficients multiplied by the subgroup distributions over each of the predictors. Comparing this expected value with the actual average for that subgroup indicates whether there is something more than additive effect. It is only feasible to do this with a few interactions, just as it is possible to put in cross product terms in multiple regressions in only a few of the total possible cases. Consequently, most of these methods of dealing with interaction effects are either limited, or expensive and time-consuming.

Still another technique for finding interactions is to restrict the total number of predictors, use cell means as basic data, and use a variance analysis looking directly for interaction effects.[14] Aside from the various statistical assumptions that have to be made, this turns out to be a relatively cumbersome method of dealing with the data. It requires a good deal of judgment in the selecting of the classes to avoid getting empty cells or cells with very small numbers of cases,

---

[12] James Morgan, Martin David, Wilbur Cohen, and Harvey Brazer, *Income and Welfare in the United States* (New York: McGraw-Hill, 1962).

Malcolm R. Fisher, "Exploration in savings behavior," *Bulletin of the Oxford University Institute of Statistics*, 18 (August, 1956), 201–77.

[13] James Morgan, "An analysis of residuals from 'normal' regressions," in *Contributions of Survey Methods to Economics*, L. Klein (Editor) (New York: Columbia University Press, 1954).

[14] F. Gerald Adams, *Some Aspects of the Income Size Distribution* (unpublished Ph.D. dissertation, The University of Michigan, 1956); and a summary, "The size of individual incomes: Socio-economic variables and chance variation," *Review of Economics and Statistics*, XL (November, 1958), 394–8.

James Morgan, "Factors related to consumer savings" in *Contributions of Survey Methods to Economics*, L. Klein (Editor) (New York: Columbia University Press, 1954).

Mordechai Kreinin, "Factors associated with stock ownership," *Review of Economics and Statistics*, XLI (February, 1959), 12–23; "Analysis of liquid asset ownership," *Review of Economics and Statistics*, XLIII (February, 1961), 76–80.

M. Kreinin, J. Lansing, J. Morgan, "Analysis of life insurance premiums," *Review of Economics and Statistics*, XXXIX (February, 1957), 46–54.

Robert Ferber has pointed out that using the highest order interaction as "error" may hide significant main effects or lower-order interaction effects, and that the heteroscedasticity of means based on subcells of different sizes may make the tests nonconservative. He has made use of the more complex method of fitting constants which provides an exact test for interactions but assumes that the individual observations are all independent. Since this assumption is not correct for most multistage samples the results of this method are also nonconservative. See Robert Ferber, "Service expenditures at mid-century," in *Consumption and Saving*, Volume I, I. Friend and R. Jones (Editors) (Philadelphia: University of Pennsylvania Press, 1960), pp. 436–60.

and the unequal cell frequencies lead to heterogeneity of variances which makes the F-test nonconservative. Sometimes interaction effects are considered important only when they involve one extremely important variable. In the case of much economic behavior, current income appears to be such a variable. In this case one can rely on covariance techniques, but these techniques tend to become far too complex when a large number of other factors are involved. Also, as more and more questions arise about the meaning of current income as a measure of ability to pay, the separation of current income for special treatment becomes more doubtful.

Finally, it is also true that if we restrict the number of variables, multiple regression techniques, particularly using dummy variables, can build in almost all feasible interaction effects. One way to restrict the number of variables is to make an analysis with an initial set and run the residuals against a second set of variables. However, unless there is some logical reason why one set takes precedence over another, this is treacherous since the explanatory classifications used in the second set will have a downward bias in their coefficients if they are at all associated with the explanatory classifications used in the first set.[15]

All these methods for dealing with interaction effects require building them in somehow without knowing how many cases there are for which each interaction effect could be relevant. The more complex the interaction, the more difficult it is to tell, of course.

The problem of logical priorities in the data and chains of causation can be handled either by restricting the analysis to one level or by conducting the analysis sequentially, always keeping in mind that the logically prior variables may have to be reintroduced in later analyses on the chance that they may mediate the effects of other variables. In practice, very little analysis of survey data has paid much attention to this problem. Perhaps the reason is that only recently has anyone been able to handle the other problems so that a truly multivariate analysis was possible. And it is only when many variables begin to be used simultaneously that the problem of their position in a causal structure becomes crucial.

Finally, there is the problem remaining from section B that the constructs of theories do not have any one-to-one correspondence with the measures from the survey. Sometimes this problem is handled by building complex variables that hopefully represent the theoretical construct. The life cycle concept, for instance, has been used this way. In a recent study, a series of questions that seemed to be asking evaluations of occupations were translated into a measure which was (hopefully) an index measure of achievement motivation.[16] More commonly, the analyst has been constrained to interpret each of the measured characteristics in terms of some theoretical meaning which it hopefully has. This is often not very satisfactory. In the case of liquid assets, the amount of

---

[15] James Morgan, "Consumer investment expenditures," *American Economic Review*, XLVIII (December, 1958), 874–902, Appendix, 898–901.

Arthur S. Goldberger and D. B. Jochems, "A note on stepwise least squares," *Journal of the American Statistical Association*, 56 (March, 1961), 105–11.

[16] Morgan, David, Cohen, and Brazer, *Income and Welfare in the United States*. (New York: McGraw-Hill Book Company, Inc., 1962).

these assets a man has represents both his past propensity to save and his present ability to dissave, two effects which could be expected to operate in opposite directions. In general, the analysis of survey data has been much better than this summary of problems would indicate. Varied approaches have been ingeniously used, and cautiously interpreted.

E. PROPOSAL FOR A PROCESS FOR ANALYZING DATA

One way to focus on the problems of analyzing data is to propose a better procedure. The proposal made here is essentially a formalization of what a good researcher does slowly and ineffectively, but insightfully on an IBM sorter. With large masses of data, weighted samples, and a desire for estimates of the reduction in error, however, we need to be able to simulate this process on large scale computing equipment. The basic idea is the sequential identification and segregation of subgroups one at a time, nonsymmetrically, so as to select the set of subgroups which will reduce the error in predicting the dependent variable as much as possible relative to the number of groups. A subgroup may be defined as membership in one or more subclasses of one or more characteristics. If more than one characteristic is used, the membership is joint, not alternative.

It is assumed that where the problem of chains of causation and logical priority of one variable over another exists, that this problem will be handled by dividing the explanatory variables or predictors into sets. One then takes the pooled residuals from an analysis using the first set of predictors and analyses these residuals against the second set of predictors. The residuals from the analysis using this second set could then be run against a third set. In practice, we might easily end up with three states—early childhood or parental factors, actions and events during the lifetime, and current situational and attitudinal variables.

The possibilities of interactions between variables in different stages can be handled by reintroducing in the second or third analyses, factors whose simple effects have already been removed, but which may also mediate the effects of factors at one of the later stages, that is, nonwhites may have their income affected by education differently from whites.

Temporarily setting aside these complications, we turn now to a description of the process of analysis using the variables from any *one* stage of the causal process. Since even the best measured variable may actually have nonlinear effects on the dependent variable, we treat each of the explanatory factors as a set of classifications. As we said, our purpose is to identify and segregate a set of subgroups which are the best we can find for maximizing our ability to predict the dependent variable. We mean maximum relative to the number of groups used, since an indefinitely large number of subgroups would "explain" everything in the sample. To be more sophisticated, if we use a model based on the assumption that we want to predict back to the population, there is an optimal number of subgroups. However, as an approximation we propose that with samples of two to three thousand we arbitrarily segregate only those groups, the separation of which will reduce the total error sum of squares by at least one per cent and do not even attempt further subdivision unless the group to be divided has a residual error (within group sum of squares) of at least two per cent of the total sum of squares. This restricts us to a *maximum* of fifty-one groups. It is just as arbitrary as the use of the 5 per cent level in significance tests and perhaps should be subject to later revision on the basis of experience.

We now describe the process of analysis in the form of a series of decision rules and instructions. We think of the sample in the beginning as a single group. The first decision is what single division of the parent group into two will do the most good. A second decision has then to be made: Which of the two groups we now have has the largest remaining error sum of squares, and hence should be investigated next for possible further subdivision? Whenever a further subdivision of a group will not reduce the unexplained sum of squares by at least one per cent of the total original sum of squares, we pay no further attention to that subgroup. Whenever there is no subgroup accounting for at least two per cent of the original sum of squares, we have finished our job. We turn now to a more orderly description of this process.

1) Considering all feasible divisions of the group of observations on the basis of each explanatory factor to be included (but not combinations of factors) find the division of the classes of any characteristic such that the partitioning of this group into two subgroups on this basis provides the largest reduction in the unexplained sum of squares.

Starting with any given group, and considering the various possible ways of splitting it into two groups, it turns out that a quick examination of any possible subgroup provides a rapid estimate of how much the error variance would be reduced by segregating it:

The reduction in error sum of squares is the same size (opposite sign) as the increase in the explained sum of squares.

For the group as a whole, the sum of squares explained by the mean is

$$N\overline{X}^2 = \frac{(\sum X)^2}{N} \tag{1}$$

and the total sum of squares (unexplained by the mean) is

$$\sum (X - \overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{N} \ . \tag{2}$$

If we now divide the group into two groups of size N_1 and N_2 and means $\overline{X}_1$ and $\overline{X}_2$, what happens to the explained sum of squares?

$$\text{Explained sum of squares} = N_1\overline{X}_1^2 + N_2\overline{X}_2^2. \tag{3}$$

The division which increases this expression most over $N\overline{X}^2$ clearly does us the most good in improving our ability to predict individuals in the sample.

Fortunately we do not even need to calculate anything more than a term involving the subgroup under inspection, since N and $\sum X$ remain known and constant throughout this search process.

$$N_2 = N - N_1 \tag{4}$$

$$\sum X_2 = \sum X - \sum X_1 \tag{5}$$

$$\therefore \text{ explained sum of squares } = N_1\left(\frac{\sum X_1}{N_1}\right)^2 + (N - N_1)\left(\frac{\sum X_2}{N - N_1}\right)^2$$

$$= \frac{(\sum X_1)^2}{N_1} + \frac{(\sum X - \sum X_1)^2}{N - N_1}.$$

The number of cases (or proportion of sample) and the sum of the dependent variable for any subgroup are enough to estimate how much reduction in error sum of squares would result from separating it from the parent group.

If it seems desirable, a variance components model which takes account of the fact that we really want optimal prediction of members of the population not merely of the sample, can be used. Indeed, the expression for the estimate of the explained, or "between" component of variance in the population turns out to be

$$\hat{\sigma}_B^2 = \frac{\left[\frac{N-1}{N-2}\left[\frac{(\sum X_1)^2}{N_1} + \frac{(\sum X - \sum X_1)^2}{N - N_1}\right] - \frac{\sum X^2}{N-2}\right] - \frac{(\Sigma X)^2}{N}}{N - \frac{N_1^2 + N_2^2}{N}} \tag{7}$$

which, though it looks formidable, contains only one new element and that is a term from the total sum of squares of the original group which is constant and can be ignored in selecting the best split. The expression in the brackets is the explained sum of squares already derived. N, $\sum X$, and $\sum X^2$ are known and constant. The denominator is an adjustment developed by Ganguli for a bias arising from unequal N's. Where N_1 equals N_2, the denominator becomes equal to N_1. The more unequal the N's, the smaller the denominator, relative to an arithmetic mean of the N's. The ratio of the explained component of variance to the total is *rho*, the intraclass correlation coefficient. Hence, in using a population model, we are searching for the particular division of a group into two that will provide the largest *rho.*[17] Computing formulas for weighted data or a dummy (one or zero) dependent variable can be derived easily.

(2) Make sure that the actual reduction in error sum of squares is larger than one per cent of the total sum of squares for the whole sample, i.e., $> .01\ (\sum X^2, -N\overline{X}^2)$ (If not select the next most promising group for search for possible subdivision, etc.)

(3) Among the groups so segregated, including the parent, or bereft ones, we now select a group for a further search for another subgroup to be split off. The selection of the group to try is on the basis of the size of the unexplained

---

[17] R. L. Anderson and T. A. Bancroft, *Statistical Theory in Research* (New York: McGraw-Hill Book Company, 1952).

M. Ganguli, "A note on nested sampling," *Sankhya* 5 (1941), 449–52.

For an example of the use of rho in analysis see Leslie Kish and John Lansing, "The family life cycle as an independent variable," *American Sociological Review*, XXII (October, 1957), 512–4.

sum of squares within the group, or the heterogeneity of the group times its size, which comes to the same thing. It may well *not* be the group with the most deviant mean.

In other words, among the groups, select the one where

$$\sum X_{ij}^2 - N_i\overline{X}_i^2 \text{ is largest.}$$

If it is less than two per cent of the total sum of squares for the whole sample, stop, because no further subdivision could reduce the error sum of squares by more than two per cent. If it is more than two per cent, repeat Step 1.

Note that the process stops when no group accounts for more than two per cent of the error sum of squares. If a group being searched allows no further segregation that will account for one per cent, the next most promising group is searched, because it may still be possible that another group with a smaller sum of squares within it can be profitably subdivided.

Since only a single group is split off at a time, the order of scanning to select that one should not affect the results. Since an independent scanning is done each time, the order in which groups are selected for further investigation should not matter either, hence our criterion is a pure efficiency one.

Chart II shows how the process suggested might arrive at a set of groups approaching those given earlier in Table 1. The numbers are rough estimates from Table 1.

Note on Amount of Detail in the Codes

The search for the best single subgroup which can be split off involves a complete scanning at each stage of each of the explanatory classifications, and within each classification of all the feasible splits. This is not so difficult as it seems, for within any classification not all possible combinations of codes are feasible. If one orders the subclasses in ascending sequence according to their means (on the dependent variable), then it can be shown that the best single division—the one which maximizes the explained sum of squares—will never combine noncontiguous groups.

Hence, starting at either end of the ordered subgroups, the computer will sequentially add one subgroup after another to that side and subtract it from the other side, always recomputing the explained sum of squares. By "explained" we mean that the means of the two halves are used for predicting rather than the over-all mean. Whenever the new division has a higher explained sum of squares, it is retained, otherwise the previous division is remembered. But in any case, the process is continued until there is only one subgroup left on the other side, to allow for the possibility of "local maxima."

The machine then remembers the best split, and the explained sum of squares associated with it, and proceeds to the next explanatory characteristic. If upon repeating this procedure with the subclasses of that characteristic, a still larger explained sum of squares is discovered, the new split on the new characteristic is retained and the less adequate one dropped.

The final result will thus be the best single split, allowing any reasonable

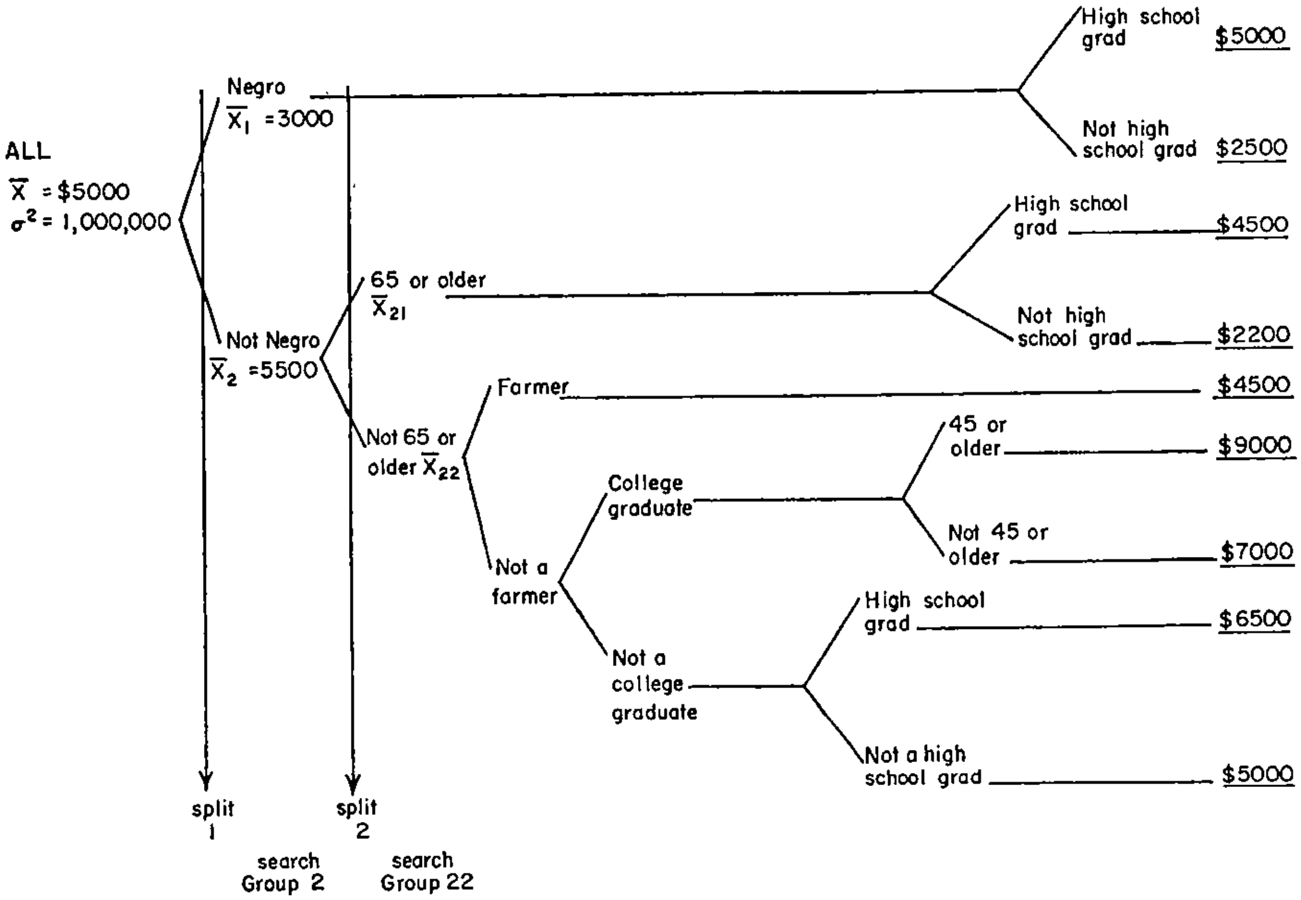


CHART II. Annual Earnings.

combination of subclasses of a single category, to maximize the explained sum of squares. It is easy to see that this choice will not depend on the order in which factors are entered, but may depend on the amount of detail with which they are coded. The number of subclasses probably should not vary too much from one factor to the next.

The authors are planning to try out such a program under a grant from the National Science Foundation. Data which have already been analyzed using dummy variable multiple regressions will be re-analyzed to see whether the new program provides new insights.

DISCUSSION

What is the theoretical model behind this process? Instead of simplifying the analysis by arbitrary or theoretical assumptions that restrict the number of variables or the way in which they operate, this process essentially restricts the complexity of the analysis by insisting that there be a large enough sample of any particular subgroup so that we can be sure it matters, and by handling problems one at a time. This is essentially what a researcher does when first investigating a sample using a sorter and his own judgment. It is assumed that the sample being used in a situation like this is a representative probability sample of a large important population. It is possible that there may be subgroups of the population whose behavior is of more importance than that of other subgroups, in which case it would be easily possible to weight the data to take account of this fact. It may be that there are certain crucial characteristics, the importance of which must be investigated. In this case, either lower admission criteria could be used or an initial arbitrary division of the sample according to this characteristic could be made before starting.

Why not take all possible subsets, in other words, all possible combinations of characteristics, and then start combining subcells where the means are close to one another? The simple reason is that there are far too many possible subsets, and since this is a sample, the means of these subsets are unstable and unreliable estimates. It is true, however, that this is the only way one would avoid all possibility of failing to discover interaction effects. Let us take a simple example of a stituation where the method we propose would fail to discover interaction effects. Suppose we have males and females, old and young, in the following proportions who go to the hospital each year, young females eight per cent, young males two per cent, old females two per cent, old males eight per cent. Assuming half the population is male and half the population is old, the old-young split would give means of five and five per cent, and the male-female split would give means of five and five per cent. Thus we would never discover that it is young females and old males who go to the hospital. One way out of this difficulty which would also vastly increase the efficiency of the machine processes would be to set up a relatively arbitrary division of the sample into perhaps ten groups to start with, groups which are known to be important and suspected to be different in their behavior. The only problem with this is that the remaining procedures will not be invariant with respect to which initial groups were selected.

One can never be sure that there does not exist previous work relevant to any "new" idea. William Belson has suggested a sequential, nonsymmetrical division of the sample which he calls "biological classification," for a different purpose, that of matching two groups on other characteristics used as controls so that they can be compared.[18] His procedure is restricted to the case where the criterion can be converted to a one-zero division, and the criterion for subdivision is the best improvement in discrimination. The method takes account of the number of cases, i.e., focuses on improvement in prediction, not on levels of significance. We have proposed this same focus. No rules are provided as to when to stop, or in what order to keep searching, though an intelligent researcher would intuitively follow the rules suggested here.

Another approach to the problem as been suggested and tried by André Danière and Elizabeth Gilboy. Their approach attempts to keep numerical variables whenever there appears to be linearity, at least within ranges, and to repool groups whenever there does not appear any substantial nonlinearity or interaction effect. The method is feasible only where the number of factors is limited. The pooling both of groups and of ranges of "variables" makes it complicated.[19] In practice, they found it useful to restrict the number of allowable interaction effects.

There are also studies going on in the selection of test items to get the best prediction with a limited set of predictors. But the prediction equation in these analyses always seems to be multiple regression without any interaction effects.[20] Group-screening methods have been suggested whereby a set of factors is lumped and tested and the individual components checked only if the group seems to have an effect. These procedures, however, require knowledge of the direction of each effect and again assume no interaction effects.[21] These group-screening methods are largely used in experimental designs and quality control procedures. It is interesting, however, that they usually end up with two-level designs, and our suggested procedure of isolating one subgroup at a time has some similarity to this search for simplicity.

The approach suggested here bears a striking resemblance to Sewall Wright's path coefficients, and to procedures informally called "pattern analysis." The justification for it, however, comes not fron any complicated statistical theory, nor from some enticing title, but from a calculated belief that for a large range of problems, the real world is such that the proposed procedure will facilitate understanding it, and foster the development of better connections between theoretical constructs and the things we can measure.

One possible outcome, for those who want precise measurement and testing,

---

[18] William A. Belson, "Matching and prediction on the principle of biological classification," *Applied Statistics*, VIII (1959), 65–75.

[19] André Danière and Elizabeth Gilboy, "The specification of empirical consumption structures, in *Consumption and Saving*, Volume I, I. Friend and R. Jones (Editors) (Philadelphia: University of Pennsylvania Press, 1960), pp. 93–136.

[20] Paul Horst and Charlotte MacEwan, "Optimal test-length for multiple prediction, the general case," *Psychometrika*, 22 (December, 1957), 311–24 and references cited therein.

[21] G. S. Watson, "A Study of the group-screening method," *Technometrics*, 3 (August, 1961), 371–88.

G. E. P. Box, "Integration of techniques for process control," *Transactions of the Eleventh Annual Convention of the American Society for Quality Control, 1958.*

is the development of new constructs, as combinations of the measured "variables," which are then created immediately in new studies and used in the analysis. The family life cycle was partly theoretical, partly empirical in its development. Other such constructs may appear from our analysis, and then acquire theoretical interpretation.

F. WHAT NEEDS TO BE DONE?

It may seem that the procedure proposed here is actually relatively simple. Each stage involves a simple search of groups defined as a subclass of any one classification and a selection of one with a maximum of a certain expression which is easily computed. It turns out, however, that the computer implications of this approach are dramatic. The approach, if it is to use the computer efficiently requires a large amount of immediate access storage which does not exist on many present-day computers. Our traditional procedures for multivariate analysis involve storing information in the computer in the form of a series of two-way tables, or cross-product moments. This throws away most of the interesting and potentially fruitful interconnectedness of survey data, and we only recapture part of it by multivariate processes which assume additivity. The implications of the proposed procedure are that we need to be able to keep track of all the relevant information about each individual in the computer as we proceed with the analysis.

Only an examination of the pedigree of the groups selected by the machine will tell whether they reveal things about the real world, or lead to intuitively meaningful theoretical constructs, which had not already come out of earlier "multivariate" analyses of the same data.

It may prove necessary to add constraints to induce more symmetry, such as giving priority to seriatim splits on the same characteristic, since this might make the interpretation easier. Or we may want to introduce an arbitrary first split, say on sex, to see whether offsetting interactions previously hidden could be uncovered in this way.

Most statistical estimates carry with them procedures for estimating their sampling variability. Sampling stability with the proposed program would mean that using a different sample, one would end up with the same complex groups segregated. No simple quantitative measure of similarity seems possible, nor any way of deriving its sampling properties. The only practical solution would seem to be to try the program out on some properly designed half-samples, taking account of the original sample stratification and controls, and to describe the extent of similarity of the pedigrees of the groups so isolated. Since the program "tries" an almost unlimited number of things, no significance tests are appropriate, and in any case the concern is with discovering a limited number of "indexes" or complex constructs which will explain more than other possible sets.

It seems clear that the procedure takes care of most of the problems discussed earlier in this paper. It takes care of any number of explanatory factors, giving them all an equal chance to come in. It uses classifications, and indeed only those sets of subclasses which it actually proves important to distinguish. The results still depend on the detail with which the original data were coded.

Differential quality of the measures used remains a problem. Sample complexities are relatively unimportant since measures of importance in reducing predictive error are involved rather than tests of significance, and one can restrict the objective to predicting the sample rather than the population. Intercorrelations among the predictors are adequately handled, and logical priorities in causation can be.

Most important, however, the interaction effects which would otherwise be ignored, or specified in advance arbitrarily from among a large possible set, are allowed to appear if they are important.

There is theory built into this apparently empiristic process, partly in the selection of the explanatory characteristics introduced, but more so in the rules of the procedures. Where there is one factor of supreme theoretical interest, it can be held back and used to explain the differences remaining within the homogeneous groups developed by the program. This is a severe test both for the effect of this factor and for possible first-order interaction effects between it and any of the other factors used in defining the groups.

Finally, where it is desired to create an index of several related measures, such as attitudinal questions in the same general area, the program can be restricted to these factors and to five or ten groups, and will create a complex index with maximal predictive power.

Reprinted from the JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
June, 1963, Vol. 58
pp. 415-434

APPENDIX K

INPUT VARIABLES
TWO-STAGE WAGE-RATE ANALYSES

(ISR Project 678, Deck 35)

Variable Number	Column Number	
1	3	Physical condition--spending unit head 0. SU head completely disabled 1. SU head severely disabled 2. SU head somewhat disabled, disabled but not limited, limitation NA 3. SU head reports no disability
2	9	E12. E13. Geographic mobility 1. lived in one state more than 100 miles from here 2. lived in two states 3. lived in three states 4. lived in four or more states 5. NA how many states lived in 0. lived in one state less than 100 miles from her; head never worked
3	10	E4-E7. Education of the head of spending unit 1. grade school (1-8 years) or less 2. some high school (9-11 years); some high school plus noncollege training; grade school plus noncollege training 3. high school (12 years) 4. high school plus noncollege training, i.e., business college, trade school, etc. 5. college, no degree 6. college, bachelor's degree or no advanced degree mentioned 7. college, advanced degree 9. NA 0. none

Variable Number	Column Number	
4	13	Immigration of head or father 0. spending unit head grew up in a foreign country 1. spending unit head grew up in the United States, father grew up in a foreign country 2. both spending unit head and father grew up in the United States
5	14	Occupation of spending unit head 1. professional, technical and kindred 2. managers and officials, nonself-employed 3. self-employed businessmen and artisans 4. clerical and kindred, sales workers 5. craftsmen, foremen, and kindred 6. operatives and kindred 7. laborers, farm and nonfarm, service workers 8. farmers and farm managers 9. government protective workers, members of the armed forces 0. housewives, widows, students, rentier, never worked, occupation NA
6	15	Supervisory responsibility of spending unit head 0. head is self-employed 1. head supervises others 2. head is neither self-employed nor supervisor

Variable Number	Column Number	
7	17	<u>Frequency of unemployment</u> 1. usual, seasonal, almost every year 2. happens occasionally; every few years (a few times, 3 or more) 3. short spells are usual, but not longer spells; unusual to be unemployed for more than a short period 4. unusual to be unemployed, work is steady, seldom unemployed 5. has <u>never</u> been unemployed 6. entered labor force recently; was self-employed until recently (any other evidence of no experience) 9. DK, NA 0. Inap., does not work for someone else Code 5 only if R says he is <u>never</u> unemployed
8	19	<u>Rank in school of spending unit heads</u> 0. head's grades above average 1. head's grades average, DK, NA, and age less grades completed is 7 or less 2. head's grades average, DK, NA, and age less grades completed is 8 or 9 3. head's grades below average and age less grades completed is 7 or less 4. head's grades below average and age less grades completed is 8 or 9 5. head's grades not above average and age less grades completed is 10 or more 6. head's grades not above average and had college training, or nonacademic training, has no education, retardation NA

Variable Number	Column Number	
9	20	Religious preference and church attendance of spending unit head 0. head is Catholic; attends two or three times a month or more, attendance NA 1. head is Catholic; attends once a month or less 2. head is Fundamentalist Protestant; attends two or three times a month or more, attendance NA 3. head is Fundamentalist Protestant; attends once a month or less 4. head is non-Fundamentalist Protestant, attends two or three times a month or more, attendance NA 5. head is non-Fundamentalist Protestant, attends once a month or less 6. head is non-Christian, religion NA
10	22	Attitude toward hard work and need-achievement index hard work is equal to or more important than luck 0. N/Ach score greater than .35 1. N/Ach score is between .15-.34 2. N/Ach score is less than .14 luck is more important than hard work 3. N/Ach score is greater than .35 4. N/Ach score is between .15-.34 5. N/Ach score is less than .14 6. N/Ach score is NA
11	25	M1. Race 1. white 2. Negro, other (Mexicans, Filipinos, Orientals, etc.)
12	26	Age of head of spending unit 1. Under 25 2. 25-34 3. 35-44 4. 45-54 5. 55-64 6. 65-74 7. 75 and over

Variable Number	Column Number	
13	32	Difference in the education of the spending unit head and his wife 0. no wife present 1. wife has two or more levels more education than head 2. wife has one level more education than head 3. wife has the same level of education as the head 4. wife has one level less education than the head 5. wife has two or more levels less education than the head 6. education of wife NA
14	33	Urban-rural migration of spending unit head head grew up on a farm: 0. lives in a rural area now 1. lives in a town 2,500-49,999 now 2. lives in a city 50,000 or over now head grew up in a small town or a city 3. lives in a rural area now 4. lives in a town or city 2,500 or over now 5. all other responses (NA where grew up, grew up in "other" or several places)
15	34	North-South migration of spending unit head head did not grow up in the South 0. moved into the South 1. does not live in the South now head grew up in the South 2. is still in the South 3. moved out of the South 4. head grew up outside the United States 5. all other responses (NA where grew up, grew up in several regions)

Variable Number	Column Number	
16	35	Family composition 0. single male head of SU, no children 1. single male head of SU, 1 or more children 2. single female head of SU, no children 3. single female head of SU, 1 or more children 4. married head of SU, no children 5. married head of SU, 1 child 6. married head of SU, 2 children 7. married head of SU, 3 or more children
17	37	Plans to help parents or children 0. no plans to help parents or children 1. plans to help parents in the future; DK, NA, or no plans for children 2. plans to send children to college; DK, NA, or no plans to help parents 3. plans both to send children to college and help parents in the future
18	39	Interviewer's assessment of head's ability to communicate 0. alert, answers easily 1. has slight difficulty in understanding or answering 2. has considerable difficulty understanding and answering 3. NA
19	44	Size of place 1. central cities of the 12 largest SMA's 2. cities 50,000 and over, exclusive of the central cities of the 12 largest PSU's 3. urban places 10,000-49,999 4. urban places 2500-9999; urbanized areas not included in above codes 5. rural, near a city 6. rural, not near a city

Variable Number   Column Number

20   45   Difference in education of head and father

0. father had 1 or more levels of education more than the head
1. father had same education as the head
2. father had 1 level less education than the head
3. father had 2 levels less than the head

For fathers, levels of education are defined as:

1) 0-8 years, NA
2) 9-12 years
3) some college or college degree

For spending unit heads, levels of education are defined as:

1) 0-11 grades
2) 12 grades
3) college

21   52-54   Head's earning rate

(The quotient of head's total wage income divided by hours worked x 100.)

xxx. Actual amount
-xx. Negative amount xx
998. Positive over the field amount (N = 4)*
-98. Negative over the field amount (N = 16)**
000. Head had no wage income (N = 451)

22   59   Sex of head of this adult unit

1. Male
2. Female

9. NA

---

*Self-employed businessmen and/or artisans, white

**Primarily White, Farmers and also several self-employed businessmen and/or artisans

Variable Number	Column Number	
23	64	<u>Religious preference of head</u> 1. Catholics 2. Fundamentalist Protestants 3. Non-Fundamentalist Protestants 4. non-Christians; not ascertained
24	65	<u>Need-achievement score of head</u> 1. under .15 2. .15-.34 3. .35 and over 4. not ascertained
25	66	<u>Background of head</u> grew up in Deep South 1. on farm 2. in small town or large city grew up outside Deep South in United States 3. on farm 4. in small town or large city 5. grew up in foreign country 6. not ascertained
26	67-68	Weights
27	69-72	Interview number

APPENDIX L

Listing of Sample Computer Input AID (2)

```
0000000001111111111222222222233333333334444444444555555555566666666667777777778  Column Heading
12345678901234567890123456789012345678901234567890123456789012345678901234567890
                                                Input File follows:
1S719 MTR 51  WAGE RATE - H AID-2 P. 678 - DECK 35.  RUN3                         Parameter Card 1
2     W  2997   027EXCLUD     IN00000000000000021                                        .        2
3    09   026.02000.00500063025   021WAGE RATE H                       TAP 27            .        3
4 001 M PHYS COND     003 M EDUCATION    008 F RANK IN SCHO 011 F RACE             Predictor Cards
4 012 M AGE           022 F SEX          023 F RELIGION     024 F NEED/ACH                 .
4 025 F BACKGROUND                                                                         .
 (C1,S1,I1,S5,2I1,S2,3I1,S1,I1,S1,2I1,S1,I1,S2,2I1,S5,4I1,S1,I1,S1,I1,             Format Information
 S4,2I1,S6,I3,S4,I1,S4,3I1,I2,C4*)                                                         .
DATAFOLLOWS                                                                        DATAFOLLOWS Card
 530031213022720400250111309710012302201003424000000000001252222314810271678 -35  Input Data Deck
 320030401112522412350011433943332700412003506700000000004111112322512481678 -35          .
 430332224622420450550011333762141602202402415100160001003122113342623731678 -35          .
 330030224022120406352311201410041220002002220000000000001222122142725021678 -35
 210031401012722401025111409610031300415004611000000000001212211144833711678 -35
 330031401121720455052011446670341552101401414000000000000141111143925821678 -35
 230030005012722240601101110151003523000210352000000000003252221262818241678 -35
 432034401112622412211011446676300601412033614621000910001141112232209821678 -35
 635036327312812450251011223222500602201053633026002420001141112234414891678 -35
 331043131012721405241512508110010221214043313025000450001242122232323541678 -35
 431033401012722401251512209110010303314043310023000530009212122232105271678 -35
         .         .
         .         .
         .         .
         .         .
         .         .
         .         .
 636037422212522441221101334640310503202033413021003550001111122234806621678 -35
 634135203412812441211311123223201502102052613726141780433132122235313121678 -35
 423044332012421401311101101310001021002052510023001260001132212235843321678 -35
 533035504212712470321211433740501722301042524023001390001132212232618891678 -35
 534035401012622403354401508610001040011332510021001720000131112235811651678 -35
 421132441112722445354311657677301400302034606721090140674112212234814911678 -35
 836437326312122476291411557886511420103432436421093141781132112234802261678 -35
 735337305412522426210001224523410522105043434322052981031111112334815361678 -35
 635036421212522421350011223313210723312033413421002860000111111233581537I678 -35
 736337442312522421220311546682201502102142616422023611171132122334620971678 -35
 625036542212621451210101334643311753304052414620002610001151222335808321678 -35
 523036421312521451210011532652111400305252415429001380001122112334832911678 -35
 534035404012622406210401301710021221001032220021001810001221212335819171678 -35
 735536402312612451110301333654211700305032413621162142001152221335312021678 -35
 536035001311512411110011433653111701301034415421003160001121111334804681678 -35
 625036402312622411410001333650200702301033614021002860001141113334806121678 -35
E                                                                                  Type E card
1S719 MTR 51  WAGE RATE - H   RESIDUALS P. 678 - DECK 35.  RUN 4                   Parameter Card 1
2     T  2997   028                                                                      .        2
3    17   026.02000.00500063025   028RESIDUALS 51                                        .        3
4 002 M GEOG MOBILIT 003 M EDUCATION    004 M IMMIGRATION  005 F OCCUPATION       Predictor Cards
4 006 F SUPR RESP    007 F FREQ OF UNEM 009 F REL X ATTEND 010 F WORK X N/ACH              .
4 011 F RACE         013 F H-W ED DIFF  014 F URB-RUR MTG  015 F N-S MIG                   .
4 016 F FAM COMP     017 F INCOME COMM  018 M ABIL TO COMM 019 M SIZE OF PLAC              .
4 020 F H-F ED DIFF                                                                        .
                                                                             End of input file
```

```
JOB NO. 006584        UNIVERSITY OF MICHIGAN EXECUTIVE SYSTEM (MODEL DV223)     WEDNESDAY, DECEMBER 18, 1963       12 06 14.6 PM

   H7479  719  51    HSIEH       S326F              014    333    000       A I 0 2*
```

```
    $EXECUTE, DUMP, BINARY                                                                                006584   12/18/63     12 06 16.3 PM

PROGRAM ON TAPE   00002, ID=  00001

 MAP
     DAYTIM 00000*       WEFTAP 00000*       REWTAP 00000*       RUNTAP 00000*       SPUNCH 00000*       WRSBIN 00000*
     CHEKIO 00000*       RDSBIN 00000*       SELRCD 00000*       SYSTEM 00000*       ERROR  00000*       SKIP6  00000*
     SPRINT 00000*       SCARDS 00000*       SPEEK  00000*       (MAIN) 10000        IRFORM 66460        EDITPM 66460
     CAS    70163        WRATIM 70205        .IOH   70257*       (SUBT) 73752        DFDP   74412*       DFMP   74412*
     .ERR   74561*       .03311 74647*       .PRINT 74664*       .READ  74750*       .PRSLT 75115*       .PCOMT 76004*
     SQRT   76023*       .EXIT  76102*       ATLOC  76157*       ZERO   76205*       SEQPGM 76241*       .IOB   76276*
     .RBIN  76425*       .WBIN  76661*       .RWT   77072*       .EFT   77117*       (PROG) 77133        (ERAS) 77741

00606 LOCS. CAN BE SAFELY USED IN EXPANDING PROG. (OCTAL)

PROGRAM ON TAPE   00002, ID=  00002

 MAP
     DAYTIM 00000*       SELRCD 00000*       SYSTEM 00000*       ERROR  00000*       SKIP6  00000*       SPRINT 00000*
     (MAIN) 10000        WRATIM 67211        .IOH   67263*       DFDP   73416*       DFMP   73416*       .ERR   73565*
     .03311 73653*       .PRINT 73670*       .PCOMT 73754*       SQRT   73773*       .EXIT  74052*       (SUBT) 74104.
     ATLOC  74127*       ZERO   74155*       SEQPGM 74211*       (PROG) 74246        (ERAS) 77741
03473 LOCS. CAN BE SAFELY USED IN EXPANDING PROG. (OCTAL)

PROGRAM ON TAPE   00002, ID=  00003

 MAP
     DAYTIM 00000*       WEFTAP 00000*       REWTAP 00000*       RUNTAP 00000*       SPUNCH 00000*       WRSBIN 00000*
     SCARDS 00000*       CHEKIO 00000*       RDSBIN 00000*       SELRCD 00000*       DPUNCH 00000*       SYSTEM 00000*
     ERROR  00000*       SKIP6  00000*       SPRINT 00000*       (MAIN) 10000        WRATIM 65450        .IOH   65522*
     DFDP   71655*       DFMP   71655*       .ERR   72024*       .03311 72112*       .PRINT 72127*       .PUNCH 72213*
     .PCOMT 72313*       SQRT   72332*       .EXIT  72411*       ATLOC  72466*       SELPGM 72514*       .IOB   72551*
     .RBIN  72700*       .WBIN  73134*       .RWT   73345*       .EFT   73372*       (PROG) 73406        (SUBT) 74004
     (ERAS) 77741
04333 LOCS. CAN BE SAFELY USED IN EXPANDING PROG. (OCTAL)
```

```
$DATA
```

```
*    (A)UTOMATIC  (I)NTERACTION  (D)ETECTOR    --    MODEL  2.    *
          WRITTEN  IN  MAD  BY  ROBERT  W.  HSIEH    -    AUGUST  1963.
```

```
S719 MTR 51  WAGE RATE - H AID-2 P. 678 - DECK 35.  RUN3




NO.  OF  INPUT  DATA                 2997
NO.  OF  VARIABLES                     27
NO.  OF  PREDICTORS                     9
WEIGHT  VARIABLE  NO.                  26
SPLIT ELIGIBILITY CRITERION         .0200
SPLIT REDUCIBILITY CRITERION        .0050
MAXIMUM  ALLOWABLE  GROUPS             63

DEPENDENT VARIABLE IS  21  (WAGE RATE H )
VALUES OF DEPENDENT VARIABLE LARGER THAN -.00000000E 00   ARE OMITTED.
  ..          ..         ..      EQUAL    TO -.00000000E 00     ..
  ..          ..         ..          ..      -.00000000E 00     ..
OUTPUT  OPTION 1  IS   1 .
OUTPUT  OPTION 2  IS   0 .

MINIMUM SIZE REQUIRED                  25

INPUT  DATA  ARE  ON            CARD

RESIDUALS  ARE        REQUESTED  AND  OUTPUT  WILL  BE TAPE .

EXCLUDE DATA WHICH LIE    INSIDE OF INTERVAL FROM      0 TO      0 ON VARIABLE     21
       WHICH LIE      SIDE OF INTERVAL FROM    -0 TO    -0 ON VARIABLE    -0
```

EXCLUDES OBSERV.
HAVING NO INCOME
FROM WAGES

```
     VARIABLE NUMBER     COLUMNS      TYPE

CARD        1
                0              1         C
                1              3         I
                2              9         I
                3             10         I
                4             13         I
                5             14         I
                6             15         I
                7             17         I
                8             19         I
                9             20         I
               10             22         I
               11             25         I
               12             26         I
               13             32         I
               14             33         I
               15             34         I
               16             35         I
               17             37         I
               18             39         I
               19             44         I
               20             45         I
               21          52-54         I
               22             59         I
               23             64         I
               24             65         I
               25             66         I
               26          67-68         I
               27          69-72         C




    INPUT-DATA FORMAT AS FOLLOWS.

  (C1,S1,I1,S5,2I1,S2,3I1,S1,I1,S1,2I1,S1,I1,S2,2I1,S5,4I1,S1,I1,S1,I1,

  S4,2I1,S6,I3,S4,I1,S4,3I1,I2,C4*)       .     .     .     .     .     .

       .     .     .     .     .     .     .     .     .     .     .     .

       .     .     .     .     .     .     .     .     .     .     .     .

       .     .     .     .     .     .     .     .     .     .     .     .

       .     .     .     .     .     .     .     .     .     .     .     .

       .     .     .     .     .     .     .     .     .     .     .     .
```

```
     .     .     .     .     .     .     .     .     .     .     .     .

     .     .     .     .     .     .     .     .     .     .     .     .

    READ  DATA  BEGINS.

TIME IS NOW 12.  6. 56. 27.

     DATA  ARE  ALL  IN.

TIME IS NOW 12.  9. 50. 46.
```

```
* *    PREDICTOR    LISTING.    * *

VARIABLE  NO.       DESCRIPTION       MAXIMUM VALUE       T Y P E
       1           PHYS COND                3               M
       3           EDUCATION                7               M
       8           RANK IN SCHO             9               F
      11           RACE                     2               F
      12           AGE                      7               M
      22           SEX                      2               F
      23           RELIGION                 4               F
      24           NEED/ACH                 4               F
      25           BACKGROUND               6               F
```

```
* STATISTICS    FOR    TOTAL.

 TOTAL  NO.  OF  DATA  READ              2997
 NO.  OF  DATA  DELETED                   451
 TOTAL  NO.  OF  DATA  USED              2546
 SUM  OF  WEIGHTS              .11676900E 06
 SUM  OF  Y                    .26937631E 08
 SUM  OF  Y-SQUARE             .86588781E 10
 MEAN     Y                    .23069163E 03
 STANDARD  DEV.  Y             .14469030E 03
 TOTAL SUM OF SQUARES (TSS)    .24445921E 10

          PA =  4.889184E 07,              PB =  1.222296E 07

TIME IS NOW 12.  9. 51.  9.
```

```
**  S T E P  NO. =   1          PARENT  GROUP =  1 **

    TRY  ON   VARIABLE   1  OVER GROUP   1 .  RESULTS  FOLLOW.

    CODE      N     SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN              STD.   DEV.              B  S  S
      0      28     .11420000E 04     .22149200E 06     .88518734E 08     .19395096E 03     .19973745E 03
                                                                                                                 .15567360E 07
      1      55     .21660000E 04     .34443600E 06     .10092698E 09     .15901939E 03     .14597552E 03
                                                                                                                 .12098368E 08
      2     272     .12006000E 05     .22782190E 07     .72836974E 09     .18975670E 03     .15703355E 03
                                                                                                                 .35643520E 08
      3    2191     .10145500E 06     .24093484E 08     .77410714E 10     .23747951E 03     .14108162E 03
                                                                                                                 .24446009E 10
  *  FOR  VARIABLE   1 ( PHYS COND     )   B S S =    .35643520E 08          BSS/TSS =    .01458

    TRY  ON   VARIABLE   3  OVER GROUP   1 .  RESULTS  FOLLOW.

    CODE      N     SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN              STD.   DEV.              B  S  S
      0      26     .79700000E 03     .85169999E 05     .15170514E 08     .10686324E 03     .87262653E 02
                                                                                                                 .12304704E 08
      1     735     .30490000E 05     .52718860E 07     .13802658E 10     .17290541E 03     .12398862E 03
                                                                                                                 .15114425E 09
      2     558     .26386000E 05     .56871080E 07     .16259950E 10     .21553505E 03     .12315860E 03
                                                                                                                 .17506989E 09
      3     408     .19431000E 05     .46180410E 07     .14350014E 10     .23766358E 03     .13178451E 03
                                                                                                                 .17241632E 09
      4     236     .11613000E 05     .29151440E 07     .94023665E 09     .25102419E 03     .13398135E 03
                                                                                                                 .16741107E 09
      5     299     .14456000E 05     .36665590E 07     .12024160E 10     .25363579E 03     .13728265E 03
                                                                                                                 .20186502E 09
      6     212     .10165000E 05     .33293010E 07     .13908226E 10     .32752592E 03     .17190527E 03
                                                                                                                 .98563839E 08
      7      72     .34310000E 04     .13644220E 07     .66901094E 09     .39767473E 03     .19195021E 03
                                                                                                                 .24446328E 10
  *  FOR  VARIABLE   3 ( EDUCATION     )   B S S =    .20186502E 09          BSS/TSS =    .08257

    TRY  ON   VARIABLE   8  OVER GROUP   1 .  RESULTS  FOLLOW.

    CODE      N     SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN              STD.   DEV.              B  S  S
      0     698     .32654000E 05     .85653540E 07     .31231447E 10     .26230642E 03     .16382581E 03
                                                                                                                 .45307584E 08
      6     535     .25087000E 05     .64594150E 07     .21928364E 10     .25748056E 03     .14530324E 03
                                                                                                                 .99524479E 08
      1     772     .36221000E 05     .77156299E 07     .21928648E 10     .21301538E 03     .12314912E 03
                                                                                                                 .61704127E 08
      3      83     .38080000E 04     .80901799E 06     .20932710E 09     .21245220E 03     .99168602E 02
                                                                                                                 .62197375E .08
      2     265     .11757000E 05     .23464610E 07     .71714412E 09     .19957991E 03     .14548218E 03
                                                                                                                 .58228160E 08
      4      56     .23870000E 04     .41333900E 06     .95960288E 08     .17316255E 03     .10107395E 03
                                                                                                                 .51935744E 08
      5     136     .48330000E 04     .62561999E 06     .12728333E 09     .12944755E 03     .97875588E 02
                                                                                                                 .23654400E 06
      9       1     .22000000E 02     .27940000E 04     .35483800E 06     .12700000E 03     .00000000E 00
                                                                                                                 .24446295E 10
  *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =    .99524479E 08          BSS/TSS =    .04071

    TRY  ON   VARIABLE  11  OVER GROUP   1 .  RESULTS  FOLLOW.

    CODE      N     SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN              STD.   DEV.              B  S  S
```

```
   1   2197    .10488800E 06    .25037432E 08    .82090384E 10    .23870635E 03    .14589064E 03
                                                                                                     .66218176E 08
   2    349    .11881000E 05    .19001990E 07    .44984763E 09    .15993595E 03    .11082991E 03
                                                                                                     .24446001E 10
 * FOR. VARIABLE  11 ( RACE           )  .B S S =    .66218176E 08          BSS/TSS =    .02709

  TRY  ON   VARIABLE  12  OVER GROUP   1 .  RESULTS  FOLLOW.

  CODE      N     SUM OF WEIGHT      SUM  OF  Y       SUM  Y-SQUARE        MEAN            STD.   DEV.             B  S  S
   1      248    .11639000E 05    .19666560E 07    .41035636E 09    .16897121E 03    .81888580E 02
                                                                                                     .49246272E 08
   2      570    .26095000E 05    .60535650E 07    .17567325E 10    .23198180E 03    .11621144E 03
                                                                                                     .18355712E 08
   3      643    .29725000E 05    .75307710E 07    .25137397E 10    .25334806E 03    .14276301E 03
                                                                                                     .44160000E 04
   4      557    .24968000E 05    .60761330E 07    .21893794E 10    .24335681E 03    .16871537E 03
                                                                                                     .48277120E 07
   5      396    .17728000E 05    .42074020E 07    .14391640E 10    .23733089E 03    .15765256E 03
                                                                                                     .28635456E 08
   6      117    .58790000E 04    .10166260E 07    .32353491E 09    .17292498E 03    .15852209E 03
                                                                                                     .94504320E 07
   7       15    .73500000E 03    .86477999E 05    .26010368E 08    .11765714E 03    .14678233E 03
                                                                                                     .24446312E 10
 * FOR  VARIABLE  12 ( AGE            )   B S S =    .49246272E 08          BSS/TSS =    .02014

  TRY  ON   VARIABLE  22  OVER GROUP   1 .  RESULTS  FOLLOW.

  CODE      N     SUM OF WEIGHT      SUM  OF  Y       SUM  Y-SQUARE        MEAN            STD.   DEV.             B  S  S
   1     2162    .99946000E 05    .24339005E 08    .81224824E 10    .24352155E 03    .14820919E 03
                                                                                                     .11419251E 09
   2      384    .16823000E 05    .25986260E 07    .53640391E 09    .15446864E 03    .89580077E 02
                                                                                                     .24446004E 10
 * FOR  VARIABLE  22 ( SEX            )   B S S =    .11419251E 09          BSS/TSS =    .04671

  TRY  ON   VARIABLE  23  OVER GROUP   1 .  RESULTS  FOLLOW.

  CODE      N     SUM OF WEIGHT      SUM  OF  Y       SUM  Y-SQUARE        MEAN            STD.   DEV.             B  S  S
   4      162    .77320000E 04    .23630970E 07    .10157016E 10    .30562558E 03    .19482398E 03
                                                                                                     .46494592E 08
   3      965    .45422000E 05    .10991533E 08    .37036491E 10    .24198699E 03    .15159472E 03
                                                                                                     .41212864E 08
   1      543    .26428000E 05    .63744669E 07    .19811031E 10    .24120126E 03    .12955397E 03
                                                                                                     .74077503E 08
   2      876    .37187000E 05    .72085340E 07    .19584594E 10    .19384554E 03    .12283758E 03
                                                                                                     .24446273E 10
 * FOR  VARIABLE  23 ( RELIGION       )   B S S =    .74077503E 08          BSS/TSS =    .03030

  TRY  ON   VARIABLE  24  OVER GROUP   1 .  RESULTS  FOLLOW.

  CODE      N     SUM OF WEIGHT      SUM  OF  Y       SUM  Y-SQUARE        MEAN            STD.   DEV.             B  S  S
   3      744    .35386000E 05    .91075809E 07    .31169442E 10    .25737808E 03    .14778575E 03
                                                                                                     .36158208E 08
   2     1140    .52648000E 05    .12108437E 08    .39050977E 10    .22998855E 03    .14587317E 03
                                                                                                     .37999488E 08
   4       90    .40420000E 04    .92735700E 06    .32673932E 09    .22943023E 03    .16792206E 03
                                                                                                     .41804608E 08
   1      572    .24693000E 05    .47942560E 07    .13101296E 10    .19415445E 03    .12393856E 03
                                                                                                     .24446248E 10
 * FOR  VARIABLE  24 ( NEED/ACH       )   B S S =    .41804608E 08          BSS/TSS =    .01710
```

```
 TRY  ON   VARIABLE  25  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE     N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN            STD.   DEV.            B  S  S
   4   1286    .62312999E 05    .16328739E 08    .55256204E 10    .26204386E 03    .14145055E 03
                                                                                                       .13134010E 09
   5    113    .55620000E 04    .14175030E 07    .51820723E 09    .25485491E 03    .16798274E 03
                                                                                                       .15340614E 09
   6     74    .33770000E 04    .82597100E 06    .27842063E 09    .24458721E 03    .15041025E 03
                                                                                                       .16411245E 09
   2    268    .12061000E 05    .25219770E 07    .72186686E 09    .20910181E 03    .12699511E 03
                                                                                                       .14721357E 09
   3    483    .21041000E 05    .39373530E 07    .11609194E 10    .18712765E 03    .14197674E 03
                                                                                                       .82709568E 08
   1    322    .12415000E 05    .19060880E 07    .45387299E 09    .15353105E 03    .11395900E 03
                                                                                                       .24446214E 10
*  FOR  VARIABLE  25 ( BACKGROUND   )  B S S =   .16411245E 09          BSS/TSS =   .06713


DECOMPOSE GROUP   1  INTO GROUP   2  AND   3  BY VARIABLE   3  IN  S T E P   1 .

 CODE     N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN            STD.   DEV.            B  S  S
   0     26    .79700000E 03    .85169999E 05    .15170514E 08    .10686324E 03    .87262653E 02
                                                                                                       .12304704E 08
   1    735    .30490000E 05    .52718860E 07    .13802658E 10    .17290541E 03    .12398862E 03
                                                                                                       .15114425E 09
   2    558    .26386000E 05    .56871080E 07    .16259950E 10    .21553505E 03    .12315860E 03
                                                                                                       .17506989E 09
   3    408    .19431000E 05    .46180410E 07    .14350014E 10    .23766358E 03    .13178451E 03
                                                                                                       .17241632E 09
   4    236    .11613000E 05    .29151440E 07    .94023665E 09    .25102419E 03    .13398135E 03
                                                                                                       .16741107E 09
   5    299    .14456000E 05    .36665590E 07    .12024160E 10    .25363579E 03    .13728265E 03
                                                                                                       .20186502E 09
   6    212    .10165000E 05    .33293010E 07    .13908226E 10    .32752592E 03    .17190527E 03
                                                                                                       .98563839E 08
   7     72    .34310000E 04    .13644220E 07    .66901094E 09    .39767473E 03    .19195021E 03
                                                                                                       .24446328E 10



CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP         N        TOTAL WEIGHT          SUM OF Y          SUM  Y-SQUARE          T   S   S
    2        2262      .10317300E 06      .22243908E 08      .65990852E 10      .18033399E 10
    3         284      .13596000E 05      .46937230E 07      .20598335E 10      .43942787E 09
```

```
**   S T E P   NO. =    2            PARENT   GROUP =   2 **
     *   FOR   VARIABLE    1 ( PHYS COND     )    B S S =    .24916480E 08            BSS/TSS =    .01382
     *   FOR   VARIABLE    3 ( EDUCATION     )    B S S =    .88422719E 08            BSS/TSS =    .04903
     *   FOR   VARIABLE    8 ( RANK IN SCHO )    B S S =    .40096064E 08            BSS/TSS =    .02223
     *   FOR   VARIABLE   11 ( RACE          )    B S S =    .49295232E 08            BSS/TSS =    .02734
     *   FOR   VARIABLE   12 ( AGE           )    B S S =    .29694784E 08            BSS/TSS =    .01647
     *   FOR   VARIABLE   22 ( SEX           )    B S S =    .96526976E 08            BSS/TSS =    .05353
     *   FOR   VARIABLE   23 ( RELIGION      )    B S S =    .49552896E 08            BSS/TSS =    .02748
     *   FOR   VARIABLE   24 ( NEED/ACH      )    B S S =    .23485888E 08            BSS/TSS =    .01302
     *   FOR   VARIABLE   25 ( BACKGROUND    )    B S S =    .10887456E 09            BSS/TSS =    .06037


     DECOMPOSE GROUP    2   INTO GROUP    4   AND    5   BY VARIABLE  25   IN   S T E P    2 .

      CODE      N      SUM OF WEIGHT        SUM  OF  Y         SUM  Y-SQUARE          MEAN              STD.   DEV.              B  S  S
        4    1091     .52898999E 05      .12987841E 08      .40625538E 10      .24552148E 03      .12852047E 03
                                                                                                                             .97205312E 08
        6      61     .27620000E 04      .62627500E 06      .13730193E 09      .22674692E 03      .14149303E 03
                                                                                                                             .10159238E 09
        5      92     .45830000E 04      .10263880E 07      .31368557E 09      .22395549E 03      .13523833E 03
                                                                                                                             .10887456E 09
        2     248     .11136000E 05      .22546760E 07      .60011289E 09      .20246731E 03      .11356247E 03
                                                                                                                             .10308211E 09
        3     458     .19837000E 05      .35591850E 07      .10058098E 10      .17942153E 03      .13605748E 03
                                                                                                                             .58765183E 08
        1     312     .11956000E 05      .17895430E 07      .41961705E 09      .14967740E 03      .11266522E 03
                                                                                                                             .18033356E 10

     CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

      GROUP          N          TOTAL WEIGHT              SUM OF Y              SUM  Y-SQUARE               T   S   S
          3         284        .13596000E 05          .46937230E 07          .20598335E 10          .43942787E 09
          4        1244        .60244000E 05          .14640504E 08          .45735412E 10          .10156043E 10
          5        1018        .42929000E 05          .76034039E 07          .20255397E 10          .67885678E 09




**   S T E P   NO. =    3            PARENT   GROUP =   4 **
     *   FOR   VARIABLE    1 ( PHYS COND     )    B S S =    .27224320E 07            BSS/TSS =    .00268
     *   FOR   VARIABLE    3 ( EDUCATION     )    B S S =    .23137408E 08            BSS/TSS =    .02278
     *   FOR   VARIABLE    8 ( RANK IN SCHO )    B S S =    .16034592E 08            BSS/TSS =    .01579
     *   FOR   VARIABLE   11 ( RACE          )    B S S =    .11641376E 08            BSS/TSS =    .01146
     *   FOR   VARIABLE   12 ( AGE           )    B S S =    .38537056E 08            BSS/TSS =    .03794
     *   FOR   VARIABLE   22 ( SEX           )    B S S =    .87423263E 08            BSS/TSS =    .08608
     *   FOR   VARIABLE   23 ( RELIGION      )    B S S =    .11854272E 08            BSS/TSS =    .01167
     *   FOR   VARIABLE   24 ( NEED/ACH      )    B S S =    .70990800E 07            BSS/TSS =    .00699
     *   FOR   VARIABLE   25 ( BACKGROUND    )    B S S =    .27146560E 07            BSS/TSS =    .00267


     DECOMPOSE GROUP    4   INTO GROUP    6   AND    7   BY VARIABLE  22   IN   S T E P    3 .
```

```
 CODE     N     SUM OF WEIGHT      SUM   OF   Y      SUM  Y-SQUARE        MEAN             STD.   DEV.           B   S   S
   1   1037     .50472999E 05     .13111926E 08     .42809713E 10     .25978099E 03     .13164685E 03
                                                                                                              .87423263E 08
   2    207     .97710000E 04     .15285780E 07     .29257008E 09     .15644028E 03     .73953599E 02
                                                                                                              .10156044E 10




CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP          N          TOTAL WEIGHT            SUM OF Y            SUM  Y-SQUARE             T   S   S
   3          284        .13596000E 05         .46937230E 07         .20598335E 10          .43942787E 09
   5         1018        .42929000E 05         .76034039E 07         .20255397E 10          .67885678E 09
   6         1037        .50472999E 05         .13111926E 08         .42809713E 10          .87474224E 09
   7          207        .97710000E 04         .15285780E 07         .29257008E 09          .53438915E 08




**  S T E P  NO. =    4           PARENT  GROUP =  6 **
    *  FOR  VARIABLE   1 ( PHYS COND     )   B S S =   .14551360E 07          BSS/TSS =   .00166
    *  FOR  VARIABLE   3 ( EDUCATION     )   B S S =   .25143776E 08          BSS/TSS =   .02874
    *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .20109600E 08          BSS/TSS =   .02299
    *  FOR  VARIABLE  11 ( RACE          )   B S S =   .76642240E 07          BSS/TSS =   .00876
    *  FOR  VARIABLE  12 ( AGE           )   B S S =   .28142176E 08          BSS/TSS =   .03217

    VARIABLE  22  OVER GROUP   6  IS A CONSTANT.    S T E P   =    4 .
    *  FOR  VARIABLE  23 ( RELIGION      )   B S S =   .12119744E 08          BSS/TSS =   .01386
    *  FOR  VARIABLE  24 ( NEED/ACH      )   B S S =   .85543679E 07          BSS/TSS =   .00978
    *  FOR  VARIABLE  25 ( BACKGROUND    )   B S S =   .36798400E 07          BSS/TSS =   .00421


    DECOMPOSE GROUP   6  INTO GROUP   8  AND   9  BY VARIABLE  12  IN  S T E P   4 .

     CODE     N     SUM OF WEIGHT      SUM   OF   Y      SUM  Y-SQUARE        MEAN             STD.   DEV.          B   S   S
       1     95     .45130000E 04     .83231900E 06     .18445733E 09     .18442699E 03     .82819856E 02
                                                                                                              .28142176E 08
       2    249     .11931000E 05     .29469950E 07     .87277833E 09     .24700318E 03     .11018884E 03
                                                                                                              .21880576E 08
       3    256     .12523000E 05     .34150060E 07     .10973253E 10     .27269871E 03     .11515298E 03
                                                                                                              .88636480E 07
       4    235     .11379000E 05     .31831330E 07     .11489660E 10     .27973749E 03     .15072966E 03
                                                                                                              .13276800E 07
       5    155     .76199999E 04     .21977400E 07     .81127215E 09     .28841732E 03     .15258313E 03
                                                                                                              .55064320E 07
       6     42     .22460000E 04     .51179100E 06     .16308227E 09     .22786776E 03     .14382759E 03
                                                                                                              .70750719E 07
       7      5     .26100000E 03     .24942000E 05     .30964860E 07     .95563218E 02     .52264733E 02
```

```
                                                                                                     .87474883E 09




CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP          N         TOTAL WEIGHT             SUM OF Y            SUM  Y-SQUARE             T    S    S
    3         284         .13596000E 05         .46937230E 07        .20598335E 10           .43942787E 09
    5        1018         .42929000E 05         .76034039E 07        .20255397E 10           .67885678E 09
    7         207         .97710000E 04         .15285780E 07        .23257008E 09           .53438915E 08
    9         942         .45960000E 05         .12279607E 08        .40965205E 10           .81565139E 09




**  S T E P  NO. =   5           PARENT  GROUP =  9 **
    *  FOR  VARIABLE   1 ( PHYS COND      )    B S S =    .21395200E 07.         BSS/TSS =   .00262
    *  FUR  VARIABLE   3 ( EDUCATION      )    B S S =    .33007552E 08          BSS/TSS =   .04047
    *  FOR  VARIABLE   8 ( RANK IN SCHD )      B S S =    .20065760E 08          BSS/TSS =   .02460
    *  FOR  VARIABLE  11 ( RACE           )    B S S =    .72751679E 07          BSS/TSS =   .00892
    *  FOR  VARIABLE  12 ( AGE            )    B S S =    .77309439E 07          BSS/TSS =   .00948

    VARIABLE  22  OVER GROUP   9  IS A CONSTANT.     S T E P   =   5
    *  FOR  VARIABLE  23 ( RELIGION       )    B S S =    .98365439E 07          BSS/TSS =   .01206
    *  FOR  VARIABLE  24 ( NEED/ACH       )    B S S =    .79952320E 07          BSS/TSS =   .00980
    *  FOR  VARIABLE  25 ( BACKGROUND     )    B S S =    .36856640E 07          BSS/TSS =   .00452


    DECOMPOSE GROUP   9  INTU GROUP  10  AND  11  BY VARIABLE   3  IN  S T E P   5 .

     CODE      N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN            STD.   DEV.            B  S  S
      0        2    .10500000E 03    .24317000E 05    .56865209E 07    .23159047E 03    .22873449E 02
                                                                                                             .13324800E 06
      1      234    .10956000E 05    .25000920E 07    .75269352E 09    .22919386E 03    .12895367E 03
                                                                                                             .22103904E 08
      2      241    .12039000E 05    .30316260E 07    .93592133E 09    .25181709E 03    .11970354E 03
                                                                                                             .33007552E 08
      3      198    .97510000E 04    .27769790E 07    .95566031E 09    .28478914E 03    .13000589E 03
                                                                                                             .21051072E 08
      4      118    .57780000E 04    .17627990E 07    .65399083E 09    .30508809E 03    .14180135E 03
                                                                                                             .82230400E 07
      5      149    .73310000E 04    .21837940E 07    .79256828E 09    .29783487E 03    .13919948E 03
                                                                                                             .81565158E 09
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N        TOTAL WEIGHT           SUM OF Y          SUM  Y-SQUARE           T   S   S
   3         284      .13596000E 05       .46937230E 07       .20598335E 10        .43942787E 09
   5        1018      .42929000E 05       .76034039E 07       .20255397E 10        .67885678E 09
   7         207      .97710000E 04       .15285780E 07       .29257008E 09        .53438915E 08
  10         477      .23100000E 05       .55560350E 07       .16943013E 10        .35795830E 09
  11         465      .22860000E 05       .67235719E 07       .24022194E 10        .42468573E 09
```

```
**  S T E P  NO. =   6          PARENT  GROUP =  5 **
   *  FOR  VARIABLE   1 ( PHYS COND    )   B S S =   .22502208E 08          BSS/TSS =   .03315
   *  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .40110176E 08          BSS/TSS =   .05908
   *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .17642544E 08          BSS/TSS =   .02599
   *  FOR  VARIABLE  11 ( RACE         )   B S S =   .18241872E 08          BSS/TSS =   .02687
   *  FOR  VARIABLE  12 ( AGE          )   B S S =   .11163296E 08          BSS/TSS =   .01644
   *  FOR  VARIABLE. 22 ( SEX          )   B S S =   .22961456E 08          BSS/TSS =   .03382
   *  FOR  VARIABLE  23 ( RELIGION     )   B S S =   .73008480E 07          BSS/TSS =   .01075
   *  FOR  VARIABLE  24 ( NEED/ACH     )   B S S =   .11644848E 08          BSS/TSS =   .01715
   *  FOR  VARIABLE  25 ( BACKGROUND   )   B S S =   .12475856E 08          BSS/TSS =   .01838
```

```
DECOMPOSE GROUP   5  INTO GROUP  12  AND  13  BY VARIABLE   3  IN  S T E P   6 .

 CODE      N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN            STD.   DEV.              B  S  S
   0      24   .69199999E 03    .60853000E 05    .94839930E 07    .87937860E 02    .77279524E 02
                                                                                                        .55934399E 07
   1     453   .17315000E 05    .24809350E 07    .57951651E 09    .14328241E 03    .11375058E 03
                                                                                                        .40110176E 08
   2     245   .11210000E 05    .21764080E 07    .59791747E 09    .19414879E 03    .12507642E 03
                                                                                                        .22339792E 08
   3     136   .61200000E 04    .12719470E 07    .36613916E 09    .20783447E 03    .12893314E 03
                                                                                                        .11544368E 08
   4      72   .34810000E 04    .70088299E 06    .18572372E 09    .20134530E 03    .11319727E 03
                                                                                                        .91328960E 07
   5      88   .41110000E 04    .91237800E 06    .28675917E 09    .22193578E 03    .14317340E 03
                                                                                                        .67885708E 09
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N        TOTAL WEIGHT           SUM OF Y          SUM  Y-SQUARE           T   S   S
   3         284      .13596000E 05       .46937230E 07       .20598335E 10        .43942787E 09
   7         207      .97710000E 04       .15285780E 07       .29257008E 09        .53438915E 08
  10         477      .23100000E 05       .55560350E 07       .16943013E 10        .35795830E 09
  11         465      .22860000E 05       .67235719E 07       .24022194E 10        .42468573E 09
  12         477      .18007000E 05       .25417880E 07       .58900050E 09        .23021302E 09
  13         541      .24922000E 05       .50616159E 07       .14365395E 10        .40853389E 09
```

```
**  S T E P  NO. =   7          PARENT  GROUP =  3 **
    *  FOR  VARIABLE   1 ( PHYS COND     )   B S S =   .50755200E 06          BSS/TSS =   .00116
    *  FOR  VARIABLE   3 ( EDUCATION     )   B S S =   .12622848E 08          BSS/TSS =   .02873
    *  FOR  VARIABLE   8 ( RANK IN SCHD )   B S S =   .54205760E 07          BSS/TSS =   .01234
    *  FOR  VARIABLE  11 ( RACE          )   B S S =   .17315200E 07          BSS/TSS =   .00394
    *  FOR  VARIABLE  12 ( AGE           )   B S S =   .21661456E 08          BSS/TSS =   .04929
    *  FOR  VARIABLE  22 ( SEX           )   B S S =   .91439839E 07          BSS/TSS =   .02081
    *  FOR  VARIABLE  23 ( RELIGION      )   B S S =   .91994879E 07          BSS/TSS =   .02094
    *  FOR  VARIABLE  24 ( NEED/ACH      )   B S S =   .76377919E 07          BSS/TSS =   .01738
    *  FOR  VARIABLE  25 ( BACKGROUND    )   B S S =   .84756960E 07          BSS/TSS =   .01929


    DECOMPOSE GROUP   3  INTO GROUP  14  AND  15  BY VARIABLE  12  IN  S T E P   7 .

     CODE      N     SUM OF WEIGHT      SUM  OF  Y       SUM  Y-SQUARE        MEAN            STD.  DEV.           B  S  S
       1      17     .79700000E 03    .16590800E 06    .42608904E 08    .20816562E 03    .10064137E 03
                                                                                                              .15904912E 08
       2      80     .38610000E 04    .11846170E 07    .42983531E 09    .30681611E 03    .13111572E 03
                                                                                                              .21661456E 08
       3      82     .39040000E 04    .14706010E 07    .69837360E 09    .37669082E 03    .19232967E 03
                                                                                                              .57249760E 07
       4      53     .25590000E 04    .10573660E 07    .54333258E 09    .41319500E 03    .20394147E 03
                                                                                                              .75937600E 06
       5      38     .17890000E 04    .64831199E 06    .29527441E 09    .36238792E 03    .18364360E 03
                                                                                                              .75025439E 07
       6      13     .63299999E 03    .14550700E 06    .41758485E 08    .22986888E 03    .11458390E 03
                                                                                                              .18377600E 06
       7       1     .52999999E 02    .21412000E 05    .86504479E 07    .40400000E 03    .00000000E 00
                                                                                                              .43942806E 09


    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

     GROUP          N          TOTAL WEIGHT           SUM OF Y           SUM  Y-SQUARE           T   S   S
        7         207         .97710000E 04        .15285780E 07        .29257008E 09        .53438915E 08
       10         477         .23100000E 05        .55560350E 07        .16943013E 10        .35795830E 09
       11         465         .22860000E 05        .67235719E 07        .24022194E 10        .42468573E 09
       12         477         .18007000E 05        .25417880E 07        .58900050E 09        .23021302E 09
       13         541         .24922000E 05        .50616159E 07        .14365395E 10        .40853389E 09
       14          97         .46580000E 04        .13505250E 07        .47244421E 09        .80877499E 08
       15         187         .89380000E 04        .33431980E 07        .15873895E 10        .33688910E 09
```

```
**  S T E P  NO. =   8          PARENT  GROUP = 11 **
    *  FOR  VARIABLE   1 ( PHYS COND    )   B S S =   .34972800E 07         BSS/TSS =   .00824
    *  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .14802880E 07         BSS/TSS =   .00349
    *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .53848640E 07         BSS/TSS =   .01268
    *  FOR  VARIABLE  11 ( RACE         )   B S S =   .17148320E 07         BSS/TSS =   .00404
    *  FOR  VARIABLE  12 ( AGE          )   B S S =   .14189136E 08         BSS/TSS =   .03341

    VARIABLE  22  OVER GROUP  11  IS A CONSTANT.   S T E P   =   8 .
    *  FOR  VARIABLE  23 ( RELIGION     )   B S S =   .53548639E 07         BSS/TSS =   .01261
    *  FOR  VARIABLE  24 ( NEED/ACH     )   B S S =   .35178400E 07         BSS/TSS =   .00828
    *  FOR  VARIABLE  25 ( BACKGROUND   )   B S S =   .13281600E 06         BSS/TSS =   .00031


    DECOMPOSE GROUP  11  INTO GROUP  16  AND  17  BY VARIABLE  12  IN  S T E P   8 .

     CODE     N    SUM OF WEIGHT     SUM  OF  Y     SUM  Y-SQUARE     MEAN           STD.  DEV.              B  S  S
       2    154   .74760000E 04   .19316540E 07   .58248862E 09   .25838068E 03   .10561202E 03
                                                                                                       .14189136E 08
       3    145   .71300000E 04   .21110840E 07   .72999275E 09   .29608471E 03   .12131414E 03
                                                                                                       .12153728E 08
       4    112   .55310000E 04   .17918200E 07   .74468955E 09   .32395950E 03   .17230632E 03
                                                                                                       .32377760E 07
       5     46   .23090000E 04   .80108199E 06   .31990484E 09   .34693893E 03   .13483433E 03
                                                                                                       .28159680E 07
       6      7   .36100000E 03   .82632000E 05   .24613734E 08   .22889751E 03   .12565039E 03
                                                                                                       .20017920E 07
       7      1   .52999999E 02   .52999999E 04   .52999999E 06   .10000000E 03   .00000000E 00
                                                                                                       .42468576E 09

    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

     GROUP        N       TOTAL WEIGHT        SUM OF Y          SUM  Y-SQUARE         T   S   S
       7        207     .97710000E 04     .15285780E 07     .29257008E 09     .53438915E 08
      10        477     .23100000E 05     .55560350E 07     .16943013E 10     .35795830E 09
      12        477     .18007000E 05     .25417880E 07     .58900050E 09     .23021302E 09
      13        541     .24922000E 05     .50616159E 07     .14365395E 10     .40853389E 09
      14         97     .46580000E 04     .13505250E 07     .47244421E 09     .80877499E 08
      15        187     .89380000E 04     .33431980E 07     .15873895E 10     .33688910E 09
      16        154     .74760000E 04     .19316540E 07     .58248862E 09     .83386548E 08
      17        311     .15384000E 05     .47919180E 07     .18197308E 10     .32711009E 09


**  S T E P  NO. =   9          PARENT  GROUP = 13 **
    *  FOR  VARIABLE   1 ( PHYS COND    )   B S S =   .69921680E 07         BSS/TSS =   .01712
    *  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .17469520E 07         BSS/TSS =   .00428
    *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .92403200E 06         BSS/TSS =   .00226
    *  FOR  VARIABLE  11 ( RACE         )   B S S =   .81641520E 07         BSS/TSS =   .01998
    *  FOR  VARIABLE  12 ( AGE          )   B S S =   .78632879E 07         BSS/TSS =   .01925
    *  FOR  VARIABLE  22 ( SEX          )   B S S =   .13678712E 08         BSS/TSS =   .03348
    *  FOR  VARIABLE  23 ( RELIGION     )   B S S =   .30364720E 07         BSS/TSS =   .00743
    *  FOR  VARIABLE  24 ( NEED/ACH     )   B S S =   .62068079E 07         BSS/TSS =   .01519
    *  FOR  VARIABLE  25 ( BACKGROUND   )   B S S =   .45307120E 07         BSS/TSS =   .01109


    DECOMPOSE GROUP  13  INTO GROUP  18  AND  19  BY VARIABLE  22  IN  S T E P   9 .
```

```
CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN            STD.   DEV.                B  S  S
  1     468    .21819000E 05    .46241720E 07    .13532411E 10    .21193327E 03    .13078805E 03
                                                                                                        .13678712E 08
  2      73    .31030000E 04    .43744400E 06    .83297822E 08    .14097454E 03    .83489316E 02
                                                                                                        .40853326E 09




CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

GROUP         N          TOTAL WEIGHT            SUM OF Y            SUM  Y-SQUARE            T   S   S
   7         207        .97710000E 04         .15285780E 07        .29257008E 09        .53438915E 08
  10         477        .23100000E 05         .55560350E 07        .16943013E 10        .35795830E 09
  12         477        .18007000E 05         .25417880E 07        .58900050E 09        .23021302E 09
  14          97        .46580000E 04         .13505250E 07        .47244421E 09        .80877499E 08
  15         187        .89380000E 04         .33431980E 07        .15873895E 10        .33688910E 09
  16         154        .74760000E 04         .19316540E 07        .58248862E 09        .83386548E 08
  17         311        .15384000E 05         .47919180E 07        .18197308E 10        .32711009E 09
  18         468        .21819000E 05         .46241720E 07        .13532411E 10        .37322520E 09


**  S T E P  NO. =  10         PARENT  GROUP = 18 **
    *  FOR  VARIABLE   1 ( PHYS COND    )    B S S =    .56836080E 07         BSS/TSS =   .01523
    *  FOR  VARIABLE   3 ( EDUCATION    )    B S S =    .20820800E 07         BSS/TSS =   .00558
    *  FOR  VARIABLE   8 ( RANK IN SCHO )    B S S =    .13005840E 07         BSS/TSS =   .00348
    *  FOR  VARIABLE  11 ( RACE         )    B S S =    .40706080E 07         BSS/TSS =   .01091
    *  FOR  VARIABLE  12 ( AGE          )    B S S =    .74525440E 07         BSS/TSS =   .01997

    VARIABLE  22  OVER GROUP  18  IS A CONSTANT.    S T E P   =  10 .
    *  FOR  VARIABLE  23 ( RELIGION     )    B S S =    .40517040E 07         BSS/TSS =   .01086
    *  FOR  VARIABLE  24 ( NEED/ACH     )    B S S =    .61689679E 07         BSS/TSS =   .01653
    *  FOR  VARIABLE  25 ( BACKGROUND   )    B S S =    .75317120E 07         BSS/TSS =   .02018

    FAILED TO SPLIT GROUP  18  TRIED ON VARIABLE  25 , BUT   BSS =   .75317120E 07


    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

    GROUP         N          TOTAL WEIGHT            SUM OF Y            SUM  Y-SQUARE            T   S   S
       7         207        .97710000E 04         .15285780E 07        .29257008E 09        .53438915E 08
      10         477        .23100000E 05         .55560350E 07        .16943013E 10        .35795830E 09
      12         477        .18007000E 05         .25417880E 07        .58900050E 09        .23021302E 09
      14          97        .46580000E 04         .13505250E 07        .47244421E 09        .80877499E 08
      15         187        .89380000E 04         .33431980E 07        .15873895E 10        .33688910E 09
      16         154        .74760000E 04         .19316540E 07        .58248862E 09        .83386548E 08
      17         311        .15384000E 05         .47919180E 07        .18197308E 10        .32711009E 09
    *  FOR  VARIABLE   1 ( PHYS COND    )    B S S =    .16469280E 07         BSS/TSS =   .00460
    *  FOR  VARIABLE   3 ( EDUCATION    )    B S S =    .32082240E 07         BSS/TSS =   .00896
    *  FOR  VARIABLE   8 ( RANK IN SCHO )    B S S =    .68684640E 07         BSS/TSS =   .01919
    *  FOR  VARIABLE  11 ( RACE         )    B S S =    .62075040E 07         BSS/TSS =   .01734
    *  FOR  VARIABLE  12 ( AGE          )    B S S =    .44794080E 07         BSS/TSS =   .01251
```

```
VARIABLE  22  OVER GROUP  10  IS A CONSTANT.    S T E P   =  10 .
*  FOR  VARIABLE  23 ( RELIGION     )   B S S =   .51176160E 07          BSS/TSS =   .01430
*  FOR  VARIABLE  24 ( NEED/ACH     )   B S S =   .29413440E 07          BSS/TSS =   .00822
*  FOR  VARIABLE  25 ( BACKGROUND   )   B S S =   .15112480E 07          BSS/TSS =   .00422

FAILED TO SPLIT GROUP  10  TRIED ON VARIABLE   8 , BUT   BSS =   .68684640E 07




CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N        TOTAL WEIGHT          SUM OF Y          SUM  Y-SQUARE          T   S   S
   7        207       .97710000E 04      .15285780E 07      .29257008E 09      .53438915E 08
  12        477       .18007000E 05      .25417880E 07      .58900050E 09      .23021302E 09
  14         97       .46580000E 04      .13505250E 07      .47244421E 09      .80877499E 08
  15        187       .89380000E 04      .33431980E 07      .15873895E 10      .33688910E 09
  16        154       .74760000E 04      .19316540E 07      .58248862E 09      .83386548E 08
  17        311       .15384000E 05      .47919180E 07      .18197308E 10      .32711009E 09
*  FOR  VARIABLE   1 ( PHYS COND    )   B S S =   .81836800E 06          BSS/TSS =   .00243
*  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .78705920E 07          BSS/TSS =   .02336
*  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .88359039E 07          BSS/TSS =   .02623
*  FOR  VARIABLE  11 ( RACE         )   B S S =   .53159039E 07          BSS/TSS =   .01578
*  FOR  VARIABLE  12 ( AGE          )   B S S =   .12696848E 08          BSS/TSS =   .03769
*  FOR  VARIABLE  22 ( SEX          )   B S S =   .14603024E 08          BSS/TSS =   .04335
*  FOR  VARIABLE  23 ( RELIGION     )   B S S =   .69697919E 07          BSS/TSS =   .02069
*  FOR  VARIABLE  24 ( NEED/ACH     )   B S S =   .16021632E 08          BSS/TSS =   .04756
*  FOR  VARIABLE  25 ( BACKGROUND   )   B S S =   .85631040E 07          BSS/TSS =   .02542


DECOMPOSE GROUP  15  INTO GROUP  20  AND  21  BY VARIABLE  24  IN  S T E P  10 .

 CODE     N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN          STD.   DEV.          B  S  S
   3     74    .35600000E 04    .14580270E 07    .70555584E 09    .40955814E 03    .17450494E 03
                                                                                                   .74626400E 07
   2     74    .35100000E 04    .13400920E 07    .67603652E 09    .38179259E 03    .21641953E 03
                                                                                                   .15974128E 08
   4     14    .65600000E 03    .22129600E 06    .10496774E 09    .33734146E 03    .21497101E 03
                                                                                                   .16021632E 08
   1     25    .12120000E 04    .32378300E 06    .10082943E 09    .26714768E 03    .10874150E 03
                                                                                                   .33688912E 09
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N        TOTAL WEIGHT          SUM OF Y          SUM  Y-SQUARE          T    S    S
    7        207       .97710000E 04      .15285780E 07      .29257008E 09      .53438915E 08
   12        477       .16007000E 05      .25417880E 07      .58900050E 09      .23021302E 09
   14         97       .46580000E 04      .13505250E 07      .47244421E 09      .80877499E 08
   16        154       .74760000E 04      .19316540E 07      .58248862E 09      .83386548E 08
   17        311       .15384000E 05      .47919180E 07      .18197308E 10      .32711009E 09
   20        162       .77260000E 04      .30194150E 07      .14865601E 10      .30653592E 09




**  S T E P  NO. =  11          PARENT  GROUP = 17 **
 *  FOR  VARIABLE   1 ( PHYS COND     )   B S S =   .26587360E 07          BSS/TSS =   .00813
 *  FOR  VARIABLE   3 ( EDUCATION     )   B S S =   .66772800E 06          BSS/TSS =   .00204
 *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .37866720E 07          BSS/TSS =   .01158
 *  FOR  VARIABLE  11 ( RACE          ).  B S S =   .86832000E 05          BSS/TSS =   .00027
 *  FOR  VARIABLE  12 ( AGE           )   B S S =   .41774880E 07          BSS/TSS =   .01277

 VARIABLE  22  OVER GROUP  17  IS A CONSTANT.    S T E P   =  11
 *  FOR  VARIABLE  23 ( RELIGION      )   B S S =   .15104320E 07          BSS/TSS =   .00482
 *  FOR  VARIABLE  24 ( NEED/ACH      )   B S S =   .20643520E 07          BSS/TSS =   .00631
 *  FOR  VARIABLE  25 ( BACKGROUND    )   B S S =   .92491200E 06          BSS/TSS =   .00283

 FAILED TO SPLIT GROUP  17  TRIED ON VARIABLE  12 , BUT   BSS =   .41774880E 07

 CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N        TOTAL WEIGHT          SUM OF Y          SUM  Y-SQUARE          T    S    S
    7        207       .97710000E 04      .15285780E 07      .29257008E 09      .53438915E 08
   12        477       .16007000E 05      .25417880E 07      .58900050E 09      .23021302E 09
   14         97       .46580000E 04      .13505250E 07      .47244421E 09      .80877499E 08
   16        154       .74760000E 04      .19316540E 07      .58248862E 09      .83386548E 08
   20        162       .77260000E 04      .30194150E 07      .14865601E 10      .30653592E 09
 *  FOR  VARIABLE   1 ( PHYS COND     )   B S S =   .28601600E 06          BSS/TSS =   .00093
 *  FOR  VARIABLE   3 ( EDUCATION     )   B S S =   .49199840E 07          BSS/TSS =   .01605
 *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .43687520E 07          BSS/TSS =   .01425
 *  FOR  VARIABLE  11 ( RACE          )   B S S =   .70982080E 07          BSS/TSS =   .02316
 *  FOR  VARIABLE  12 ( AGE           )   B S S =   .11654080E 08          BSS/TSS =   .03802
 *  FOR  VARIABLE  22 ( SEX           )   B S S =   .12848800E 08          BSS/TSS =   .04192
 *  FOR  VARIABLE  23 ( RELIGION      )   B S S =   .77663840E 07          BSS/TSS =   .02534
 *  FOR  VARIABLE  24 ( NEED/ACH      )   B S S =   .23200800E 07          BSS/TSS =   .00757
 *  FOR  VARIABLE  25 ( BACKGROUND    )   B S S =   .67075840E 07          BSS/TSS =   .02188


 DECOMPOSE GROUP  20  INTO GROUP  22  AND  23  BY VARIABLE  22  IN  S T E P   11 .

 CODE     N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN          STD.  DEV.            B  S  S
  1     142    .67309999E 04    .27360940E 07    .13906026E 10    .40649145E 03    .20337511E 03
                                                                                                  .12848800E 08
  2      20    .99500000E 03    .28332100E 06    .95957363E 08    .28474472E 03    .12393549E 03
                                                                                                  .306535B4E 09
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N         TOTAL WEIGHT           SUM OF Y             SUM  Y-SQUARE            T    S    S
    7        207        .97710000E 04        .15285780E 07        .29257008E 09        .53438915E 08
   12        477        .18007000E 05        .25417880E 07        .58900050E 09        .23021302E 09
   14         97        .46580000E 04        .13505250E 07        .47244421E 09        .80877499E 08
   16        154        .74760000E 04        .19316540E 07        .58248862E 09        .83386548E 08
   22        142        .67309999E 04        .27360940E 07        .13906026E 10        .27840382E 09




**  S T E P  NO. =  12          PARENT  GROUP = 22 **
    *  FOR  VARIABLE   1 ( PHYS COND     )   B S S =   .11724800E 06         BSS/TSS =   .00042
    *  FOR  VARIABLE   3 ( EDUCATION     )   B S S =   .41148640E 07         BSS/TSS =   .01478
    *  FOR  VARIABLE   8 ( RANK IN SCHO )    B S S =   .24187680E 07         BSS/TSS =   .00869
    *  FOR  VARIABLE  11 ( RACE          )   B S S =   .80292159E 07         BSS/TSS =   .02884
    *  FOR  VARIABLE  12 ( AGE           )   B S S =   .12913440E 08         BSS/TSS =   .04638

    VARIABLE  22  OVER GROUP  22  IS A CONSTANT.    S T E P   =  12 .
    *  FOR  VARIABLE  23 ( RELIGION      )   B S S =   .65362559E 07         BSS/TSS =   .02348
    *  FOR  VARIABLE  24 ( NEED/ACH      )   B S S =   .19868160E 07         BSS/TSS =   .00714
    *  FOR  VARIABLE  25 ( BACKGROUND    )   B S S =   .65365759E 07         BSS/TSS =   .02348


    DECOMPOSE GROUP  22  INTO GROUP  24  AND  25  BY VARIABLE  12  IN  S T E P  12 .

     CODE     N     SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN            STD.   DEV.              B  S  S
       3     66    .31380000E 04    .12481810E 07    .62852672E 09    .39776322E 03    .20513348E 03
                                                                                                              .44784000E 06
       4     41    .19500000E 04    .90403200E 06    .50267273E 09    .46360615E 03    .20700296E 03
                                                                                                              .56792640E 07
       5     24    .11030000E 04    .44446200E 06    .21448933E 09    .40295738E 03    .17912371E 03
                                                                                                              .12913440E 08
       6     10    .48700000E 03    .11800700E 06    .36263515E 08    .24231417E 03    .12548671E 03
                                                                                                              .32000000E 03
       7      1    .52999999E 02    .21412000E 05    .86504479E 07    .40400000E 03    .00000000E 00
                                                                                                              .27840390E 09


    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

     GROUP        N         TOTAL WEIGHT           SUM OF Y             SUM  Y-SQUARE            T    S    S
        7        207        .97710000E 04        .15285780E 07        .29257008E 09        .53438915E 08
       12        477        .18007000E 05        .25417880E 07        .58900050E 09        .23021302E 09
       14         97        .46590000E 04        .13505250E 07        .47244421E 09        .80877499E 08
       16        154        .74760000E 04        .19316540E 07        .58248862E 09        .83396548E 08
       24        131        .61910000E 04        .25966750E 07        .13456888E 10        .25657216E 09
```

```
**  S T E P  NO. =  13          PARENT  GROUP = 24 **
    *  FOR  VARIABLE   1 ( PHYS COND     )  B S S =   .23603200E 06          BSS/TSS =   .00092
    *  FOR  VARIABLE   3 ( EDUCATION     )  B S S =   .31312960E 07          BSS/TSS =   .01220
    *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .12647840E 07          BSS/TSS =   .00493
    *  FOR  VARIABLE  11 ( RACE          )  B S S =   .66031040E 07          BSS/TSS =   .02574
    *  FOR  VARIABLE  12 ( AGE           )  B S S =   .29865440E 07          BSS/TSS =   .01164

    VARIABLE  22  OVER GROUP  24  IS A CONSTANT.    S T E P  =  13
    *  FOR  VARIABLE  23 ( RELIGION      )  B S S =   .35699360E 07          BSS/TSS =   .01391
    *  FOR  VARIABLE  24 ( NEED/ACH      )  B S S =   .30571680E 07          BSS/TSS =   .01192
    *  FOR  VARIABLE  25 ( BACKGROUND    )  B S S =   .28208640E 07          BSS/TSS =   .01099

    FAILED TO SPLIT GROUP  24  TRIED ON VARIABLE  11 , BUT   BSS =   .66031040E 07


    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

     GROUP        N        TOTAL WEIGHT           SUM OF Y          SUM  Y-SQUARE            T    S    S
        7        207       .97710000E 04        .15285780E 07       .29257008E 09        .53438915E 08
       12        477       .18007000E 05        .25417880E 07       .58900050E 09        .23021302E 09
       14         97       .46580000E 04        .13505250E 07       .47244421E 09        .80877499E 08
       16        154       .74760000E 04        .19316540E 07       .58248862E 09        .83386548E 08
    *  FOR  VARIABLE   1 ( PHYS COND     )  B S S =   .60333119E 07          BSS/TSS =   .02621
    *  FOR  VARIABLE   3 ( EDUCATION     )  B S S =   .20381520E 07          BSS/TSS =   .00885
    *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .63719360E 07          BSS/TSS =   .02768
    *  FOR  VARIABLE  11 ( RACE          )  B S S =   .32285720E 07          BSS/TSS =   .01402
    *  FOR  VARIABLE  12 ( AGE           )  B S S =   .38996560E 07          BSS/TSS =   .01694
    *  FOR  VARIABLE  22 ( SEX           )  B S S =   .86842400E 07          BSS/TSS =   .03772
    *  FOR  VARIABLE  23 ( RELIGION      )  B S S =   .11372520E 07          BSS/TSS =   .00494
    *  FOR  VARIABLE  24 ( NEED/ACH      )  B S S =   .24419160E 07          BSS/TSS =   .01061
    *  FOR  VARIABLE  25 ( BACKGROUND    )  B S S =   .82242399E 06          BSS/TSS =   .00357

    FAILED TO SPLIT GROUP  12  TRIED ON VARIABLE  22 , BUT   BSS =   .86842400E 07




    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

     GROUP        N        TOTAL WEIGHT           SUM OF Y          SUM  Y-SQUARE            T    S    S
        7        207       .97710000E 04        .15285780E 07       .29257008E 09        .53438915E 08
       14         97       .46580000E 04        .13505250E 07       .47244421E 09        .80877499E 08
       16        154       .74760000E 04        .19316540E 07       .58248862E 09        .83386548E 08
    *  FOR  VARIABLE   1 ( PHYS COND     )  B S S =   .61586799E 06          BSS/TSS =   .00739
    *  FOR  VARIABLE   3 ( EDUCATION     )  B S S =   .22267080E 07          BSS/TSS =   .02670
    *  FOR  VARIABLE   8 ( RANK IN SCHO )   B S S =   .26983760E 07          BSS/TSS =   .03236
    *  FOR  VARIABLE  11 ( RACE          )  B S S =   .71381199E 06          BSS/TSS =   .00856

    VARIABLE  12  OVER GROUP  16  IS A CONSTANT.    S T E P  =  13 .

    VARIABLE  22  OVER GROUP  16  IS A CONSTANT.    S T E P  =  13 .
    *  FOR  VARIABLE  23 ( RELIGION      )  B S S =   .19237120E 07          BSS/TSS =   .02307
```

```
•  FOR   VARIABLE   24 ( NEED/ACH       )   B S S =    .39101440E 07           BSS/TSS =    .04689
•  FOR   VARIABLE   25 ( BACKGROUND     )   B S S =    .13403490E 07           BSS/TSS =    .01607

FAILED TO SPLIT GROUP  16  TRIED ON VARIABLE  24 , BUT   BSS =   .39101440E 07
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP         N        TOTAL WEIGHT          SUM OF Y          SUM  Y-SQUARE           T   S   S
   7          207      .97710000E 04       .15285780E 07       .29257008E 09       .53438915E 08
  14           97      .46580000E 04       .13505250E 07       .47244421E 09       .80877499E 08
•  FOR   VARIABLE    1 ( PHYS COND      )   B S S =    .19826000E 06           BSS/TSS =    .00245
•  FOR   VARIABLE    3 ( EDUCATION      )   B S S =    .13117280E 07           BSS/TSS =    .01622
•  FOR   VARIABLE    8 ( RANK IN SCHO   )   B S S =    .98332000E 05           BSS/TSS =    .00122
•  FOR   VARIABLE   11 ( RACE           )   B S S =    .91453200E 06           BSS/TSS =    .01131
•  FOR   VARIABLE   12 ( AGE            )   B S S =    .64291999E 07           BSS/TSS =    .07949
•  FOR   VARIABLE   22 ( SEX            )   B S S =    .24188000E 05           BSS/TSS =    .00030
•  FOR   VARIABLE   23 ( RELIGION       )   B S S =    .24038800E 07           BSS/TSS =    .02972
•  FOR   VARIABLE   24 ( NEED/ACH       )   B S S =    .12576320E 07           BSS/TSS =    .01555
•  FOR   VARIABLE   25 ( BACKGROUND     )   B S S =    .37923640E 07           BSS/TSS =    .04689

FAILED TO SPLIT GROUP  14. TRIED ON VARIABLE  12 , BUT   BSS =   .64291999E 07
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP         N        TOTAL WEIGHT          SUM OF Y          SUM  Y-SQUARE           T   S   S
   7          207      .97710000E 04       .15285780E 07       .29257008E 09       .53438915E 08
•  FOR   VARIABLE    1 ( PHYS COND      )   B S S =    .14900700E 07           BSS/TSS =    .02788
•  FOR   VARIABLE    3 ( EDUCATION      )   B S S =    .49192800E 07           BSS/TSS =    .09205
•  FOR   VARIABLE    8 ( RANK IN SCHO   )   B S S =    .40700920E 07           BSS/TSS =    .07616
•  FOR   VARIABLE   11 ( RACE           )   B S S =    .90757400E 06           BSS/TSS =    .01698
•  FOR   VARIABLE   12 ( AGE            )   B S S =    .88510799E 06           BSS/TSS =    .01656

VARIABLE  22  OVER GROUP   7  IS A CONSTANT.    S T E P   =  13 .
•  FOR   VARIABLE   23 ( RELIGION       )   B S S =    .37652240E 07           BSS/TSS =    .07046
•  FOR   VARIABLE   24 ( NEED/ACH       )   B S S =    .19460940E 07           BSS/TSS =    .03642
•  FOR   VARIABLE   25 ( BACKGROUND     )   B S S =    .32286200E 06           BSS/TSS =    .00604
```

```
    FAILED TO SPLIT GROUP    7  TRIED ON VARIABLE    3 , BUT   BSS =   .49192800E 07




    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.
     GROUP         N        TOTAL WEIGHT            SUM OF Y          SUM  Y-SQUARE          T   S   S




    THAT IS ALL.  NO MORE GRUUPS ARE AVAILABLE. FINAL  S T E P  NO. IS  13   NO. OF GROUPS ARE   25 .
  ** THIS  IS  THE  END  OF  2ND  CORE.
TIME IS NOW 12. 10. 44.  7.
```

```
                                        *  (A)UTOMATIC  (I)NTERACTION  (D)ETECTOR    (MODEL  2)  *

                                                  *  *  *  S  U  M  M  A  R  Y  *  *  *

        DEPENDENT VARIABLE  21  ( WAGE RATE H )

        WEIGHTED  BY  VARIABLE   26

    ** TOTAL  GROUP

            N  =          2546              MEAN =  .23069163E 03            SUM Y =  .26937631E 08         TSS =  .24445921E 10
     TOTAL WT SUM=      116769         STD. DEV. =  .14469030E 03        SUM Y SQ. =  .86588781E 10

     *   GROUP  NO.   2  SPLIT  FROM. GROUP   1  ON  VARIABLE    3 (EDUCATION    )
         VALUES  OF  PREDICTOR  INCLUDED  ARE    0    1.   2    3    4    5
            N  =          2262              MEAN =  .21559815E 03        GROUP DEVIATION = -.15093479E 02        SUM Y =  .22243908E 08
      WEIGHT SUM =      103173         STD. DEV. =  .13220740E 03                 TSS(I) =  .18033398E 10    SUM Y SQ. =  .65990852E 10
    PCT OF TOTAL =     88.4         WTD. MEAN SQ. =  .47957454E 10        (TSS(I)/TSS(T)) =  .73768536E 00

     *   GROUP  NO.   3  SPLIT  FROM  GROUP   1  ON  VARIABLE    3 (EDUCATION    )
         VALUES  OF  PREDICTOR  INCLUDED  ARE    6    7
            N  =           284              MEAN =  .34522822E 03        GROUP DEVIATION =  .11453660E 03        SUM Y =  .46937230E 07
      WEIGHT SUM =       13596         STD. DEV. =  .17977870E 03                 TSS(I) =  .43942789E 09    SUM Y SQ. =  .20598335E 10
    PCT OF TOTAL =     11.6         WTD. MEAN SQ. =  .16204056E 10        (TSS(I)/TSS(T)) =  .17975509E 00

     *   GROUP  NO.   4  SPLIT  FROM  GROUP   2  ON ·VARIABLE   25 (BACKGROUND  )
         VALUES  OF  PREDICTOR  INCLUDED  ARE    4    5    6
            N  =          1244              MEAN =  .24302012E 03        GROUP DEVIATION =  .12328487E 02        SUM Y =  .14640504E 08
      WEIGHT SUM =       60244         STD. DEV. =  .12983906E 03                 TSS(I) =  .10156043E 10    SUM Y SQ. =  .45735412E 10
    PCT OF TOTAL =     51.6         WTD. MEAN SQ. =  .35579370E 10        (TSS(I)/TSS(T)) =  .41544937E 00

     *   GROUP  NO.   5  SPLIT  FROM  GROUP   2  ON  VARIABLE   25 (BACKGROUND  )
         VALUES  OF  PREDICTOR  INCLUDED  ARE    1    2    3
            N  =          1018              MEAN =  .17711579E 03        GROUP DEVIATION = -.53575834E 02        SUM Y =  .76034039E 07
      WEIGHT SUM =       42929         STD. DEV. =  .12575165E 03                 TSS(I) =  .67885677E 09    SUM Y SQ. =  .20255397E 10
    PCT OF TOTAL =     36.8         WTD. MEAN SQ. =  .13466829E 10        (TSS(I)/TSS(T)) =  .27769735E 00

     *   GROUP  NO.   6  SPLIT  FROM  GROUP   4  ON  VARIABLE   22 (SEX         )
         VALUES  OF  PREDICTOR  INCLUDED  ARE    1
            N  =          1037              MEAN =  .25978099E 03        GROUP DEVIATION =  .29089361E 02        SUM Y =  .13111926E 08
      WEIGHT SUM =       50473         STD. DEV. =  .13164685E 03                 TSS(I) =  .87474220E 09    SUM Y SQ. =  .42809713E 10
    PCT OF TOTAL =     43.2         WTD. MEAN SQ. =  .34062291E 10        (TSS(I)/TSS(T)) =  .35782747E 00

     *   GROUP  NO.   7  SPLIT  FROM  GROUP   4  ON  VARIABLE   22 (SEX         )
         VALUES  OF  PREDICTOR  INCLUDED  ARE    2
      ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
            N  =           207              MEAN =  .15644028E 03        GROUP DEVIATION = -.74251348E 02        SUM Y =  .15285780E 07
      WEIGHT SUM =        9771         STD. DEV. =  .73953599E 02                 TSS(I) =  .53438918E 08    SUM Y SQ. =  .29257008E 09
    PCT OF TOTAL =      8.4         WTD. MEAN SQ. =  .23913117E 09        (TSS(I)/TSS(T)) =  .21860055E-01

     *   GROUP  NO.   8  SPLIT  FROM  GROUP   6  ON  VARIABLE   12 (AGE         )
         VALUES  OF  PREDICTOR  INCLUDED  ARE    1
      ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
            N  =            95              MEAN =  .18442699E 03        GROUP DEVIATION = -.46264641E 02        SUM Y =  .83231900E 06
      WEIGHT SUM =        4513         STD. DEV. =  .82819856E 02                 TSS(I) =  .30955248E 08    SUM Y SQ. =  .18445733E 09
    PCT OF TOTAL =      3.9         WTD. MEAN SQ. =  .15350208E 09        (TSS(I)/TSS(T)) =  .12662745E-01

     *   GROUP  NO.   9  SPLIT  FROM  GROUP   6  ON  VARIABLE   12 (AGE         )
         VALUES  OF  PREDICTOR  INCLUDED  ARE    2    3    4    5    6    7
            N  =           942              MEAN =  .26718030E 03        GROUP DEVIATION =  .36488676E 02        SUM Y =  .12279607E 08
```

```
  WEIGHT SUM =        45960         STD. DEV. =  .13321781E 03                      TSS(I) =  .81565145E 09    SUM Y SQ. =  .40965205E 10
PCT OF TOTAL =     39.4         WTD. MEAN SQ. =  .32808691E 10           (TSS(I)/TSS(T)) =  .33365544E 00

 *   GROUP  NO.  10  SPLIT  FROM  GROUP   9  ON  VARIABLE    3 (EDUCATION    )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     0    1    2
  ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
           N  =           477              MEAN =  .24052099E 03          GROUP DEVIATION =  .98293647E 01         SUM Y =  .55560350E 07
  WEIGHT SUM =        23100         STD. DEV. =  .12448305E 03                      TSS(I) =  .35795830E 09    SUM Y SQ. =  .16943013E 10
PCT OF TOTAL =     19.8         WTD. MEAN SQ. =  .13363430E 10           (TSS(I)/TSS(T)) =  .14642864E 00

 *   GROUP  NO.  11  SPLIT  FROM  GROUP   9  ON  VARIABLE    3 (EDUCATION    )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     3    4    5
           N  =           465              MEAN =  .29411950E 03          GROUP DEVIATION =  .63427877E 02         SUM Y =  .67235719E 07
  WEIGHT SUM =        22860         STD. DEV. =  .13629996E 03                      TSS(I) =  .42468574E 09    SUM Y SQ. =  .24022194E 10
PCT OF TOTAL =     19.6         WTD. MEAN SQ. =  .19775337E 10           (TSS(I)/TSS(T)) =  .17372458E 00

 *   GROUP  NO.  12  SPLIT  FROM  GROUP   5  ON  VARIABLE    3 (EDUCATION    )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     0    1
  ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
           N  =           477              MEAN =  .14115555E 03          GROUP DEVIATION = -.89536079E 02         SUM Y =  .25417880E 07
  WEIGHT SUM =        18007         STD. DEV. =  .11306919E 03                      TSS(I) =  .23021302E 09    SUM Y SQ. =  .58900050E 09
PCT OF TOTAL =     15.4         WTD. MEAN SQ. =  .35878748E 09           (TSS(I)/TSS(T)) =  .94172364E-01

 *   GROUP  NO.  13  SPLIT  FROM  GROUP   5  ON  VARIABLE    3 (EDUCATION    )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     2    3    4    5
           N  =           541              MEAN =  .20309831E 03          GROUP DEVIATION = -.27593322E 02         SUM Y =  .50616159E 07
  WEIGHT SUM =        24922         STD. DEV. =  .12803320E 03                      TSS(I) =  .40853389E 09    SUM Y SQ. =  .14365395E 10
PCT OF TOTAL =     21.3         WTD. MEAN SQ. =  .10280056E 10           (TSS(I)/TSS(T)) =  .16711740E 00

 *   GROUP  NO.  14  SPLIT  FROM  GROUP   3  ON  VARIABLE   12 (AGE          )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     1    2
  ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
           N  =            97              MEAN =  .28993666E 03          GROUP DEVIATION =  .59245035E 02         SUM Y =  .13505250E 07
  WEIGHT SUM =         4658         STD. DEV. =  .13176926E 03                      TSS(I) =  .80877504E 08    SUM Y SQ. =  .47244421E 09
PCT OF TOTAL =      4.0         WTD. MEAN SQ. =  .39156671E 09           (TSS(I)/TSS(T)) =  .33084252E-01

 *   GROUP  NO.  15  SPLIT  FROM  GROUP   3  ON  VARIABLE   12 (AGE          )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     3    4    5    6    7
           N  =           187              MEAN =  .37404319E 03          GROUP DEVIATION =  .14335156E 03         SUM Y =  .33431980E 07
  WEIGHT SUM =         8938         STD. DEV. =  .19414370E 03                      TSS(I) =  .33688910E 09    SUM Y SQ. =  .15873895E 10
PCT OF TOTAL =      7.7         WTD. MEAN SQ. =  .12505004E 10           (TSS(I)/TSS(T)) =  .13780994E 00

 *   GROUP  NO.  16  SPLIT  FROM  GROUP  11  ON  VARIABLE   12 (AGE          )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     2
  ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
           N  =           154              MEAN =  .25838068E 03          GROUP DEVIATION =  .27689054E 02         SUM Y =  .19316540E 07
  WEIGHT SUM =         7476         STD. DEV. =  .10561202E 03                      TSS(I) =  .83386543E 08    SUM Y SQ. =  .58248862E 09
PCT OF TOTAL =      6.4         WTD. MEAN SQ. =  .49910208E 09           (TSS(I)/TSS(T)) =  .34110616E-01

 *   GROUP  NO.  17  SPLIT  FROM  GROUP  11  ON  VARIABLE   12 (AGE          )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     3    4    5    6    7
  ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
           N  =           311              MEAN =  .31148713E 03          GROUP DEVIATION =  .80795500E 02         SUM Y =  .47919180E 07
  WEIGHT SUM =        15384         STD. DEV. =  .14581840E 03                      TSS(I) =  .32711009E 09    SUM Y SQ. =  .18197308E 10
PCT OF TOTAL =     13.2         WTD. MEAN SQ. =  .14926208E 10           (TSS(I)/TSS(T)) =  .13380968E 00

 *   GROUP  NO.  18  SPLIT  FROM  GROUP  13  ON  VARIABLE   22 (SEX          )
    VALUES  OF  PREDICTOR  INCLUDED  ARE     1
  ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
           N  =           468              MEAN =  .21193327E 03          GROUP DEVIATION = -.18758362E 02         SUM Y =  .46241720E 07
```

```
  WEIGHT SUM =       21819        STD. DEV. =  .13078805E 03              TSS(I) =  .37322520E 09    SUM Y SQ. =  .13532411E 10
PCT OF TOTAL =    18.7        WTD. MEAN SQ. =  .98001587E 09      (TSS(I)/TSS(T)) =  .15267381E 00

 *   GROUP  NO.  19  SPLIT  FROM  GROUP  13  ON  VARIABLE   22 (SEX          )
     VALUES  OF  PREDICTOR  INCLUDED  ARE     2
 ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
         N  =            73               MEAN =  .14097454E 03      GROUP DEVIATION = -.89717089E 02        SUM Y =  .43744400E 06
  WEIGHT SUM =        3103        STD. DEV. =  .83489316E 02              TSS(I) =  .21629355E 08    SUM Y SQ. =  .83297822E 08
PCT OF TOTAL =     2.7        WTD. MEAN SQ. =  .61668466E 08      (TSS(I)/TSS(T)) =  .88478381E-02

 *   GROUP  NO.  20  SPLIT  FROM  GROUP  15  ON  VARIABLE   24 (NEED/ACH     )
     VALUES  OF  PREDICTOR  INCLUDED  ARE     2    3    4
         N  =           162               MEAN =  .39081219E 03      GROUP DEVIATION =  .16012056E 03        SUM Y =  .30194150E 07
  WEIGHT SUM =        7726        STD. DEV. =  .19918807E 03              TSS(I) =  .30653592E 09    SUM Y SQ. =  .14865601E 10
PCT OF TOTAL =     6.6        WTD. MEAN SQ. =  .11800242E 10      (TSS(I)/TSS(T)) =  .12539348E 00

 *   GROUP  NO.  21  SPLIT  FROM  GROUP  15  ON  VARIABLE   24 (NEED/ACH     )
     VALUES  OF  PREDICTOR  INCLUDED  ARE     1
 ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
         N  =            25               MEAN =  .26714768E 03      GROUP DEVIATION =  .36456056E 02        SUM Y =  .32378300E 06
  WEIGHT SUM =        1212        STD. DEV. =  .10874150E 03              TSS(I) =  .14331554E 08    SUM Y SQ. =  .10082943E 09
PCT OF TOTAL =     1.0        WTD. MEAN SQ. =  .86497878E 08      (TSS(I)/TSS(T)) =  .58625542E-02

 *   GROUP  NO.  22  SPLIT  FROM  GROUP  20  ON  VARIABLE   22 (SEX          )
     VALUES  OF  PREDICTOR  INCLUDED  ARE     1
         N  =           142               MEAN =  .40649145E 03      GROUP DEVIATION =  .17579982E 03        SUM Y =  .27360940E 07
  WEIGHT SUM =        6731        STD. DEV. =  .20337511E 03              TSS(I) =  .27840384E 09    SUM Y SQ. =  .13906026E 10
PCT OF TOTAL =     5.8        WTD. MEAN SQ. =  .11121988E 10      (TSS(I)/TSS(T)) =  .11388560E 00

 *   GROUP  NO.  23  SPLIT  FROM  GROUP  20  ON  VARIABLE   22 (SEX          )
     VALUES  OF  PREDICTOR  INCLUDED  ARE     2
 ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
         N  =            20               MEAN =  .28474472E 03      GROUP DEVIATION =  .54053091E 02        SUM Y =  .28332100E 06
  WEIGHT SUM =         995        STD. DEV. =  .12393549E 03              TSS(I) =  .15283205E 08    SUM Y SQ. =  .95957363E 08
PCT OF TOTAL =      .9        WTD. MEAN SQ. =  .80674157E 08      (TSS(I)/TSS(T)) =  .62518425E-02

 *   GROUP  NO.  24  SPLIT  FROM  GROUP  22  ON  VARIABLE   12 (AGE          )
     VALUES  OF  PREDICTOR  INCLUDED  ARE     3    4    5
 ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
         N  =           131               MEAN =  .41942739E 03      GROUP DEVIATION =  .18873576E 03        SUM Y =  .25966750E 07
  WEIGHT SUM =        6191        STD. DEV. =  .20357497E 03              TSS(I) =  .25657218E 09    SUM Y SQ. =  .13456888E 10
PCT OF TOTAL =     5.3        WTD. MEAN SQ. =  .10891166E 10      (TSS(I)/TSS(T)) =  .10495500E 00

 *   GROUP  NO.  25  SPLIT  FROM  GROUP  22  ON  VARIABLE   12 (AGE          )
     VALUES  OF  PREDICTOR  INCLUDED  ARE     6    7
 ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
         N  =            11               MEAN =  .25818333E 03      GROUP DEVIATION =  .27421701E 02        SUM Y =  .13941900E 06
  WEIGHT SUM =         540        STD. DEV. =  .12851215E 03              TSS(I) =  .89183015E 07    SUM Y SQ. =  .44913963E 08
PCT OF TOTAL =      .5        WTD. MEAN SQ. =  .35995661E 08      (TSS(I)/TSS(T)) =  .36481756E-02
```

```
                    * * *     ANALYSIS  OF  VARIANCE  TABLE     * * *

     SOURCE OF              SUM  OF               DEGREE OF         MEAN
     VARIATION              SQUARES                FREEDOM         SQUARE              F

     T O T A L          .24445921E 10           116723

     BETWEEN            .59073587E 09               12          .49227989E 08      .30991909E 04

     WITHIN             .18538562E 10           116711          .15884142E 05



        RESIDUALS  ARE  OBTAINED.

  TIME IS NOW 12. 12. 32. 14.

           RESULTS  ARE  ON  TAPE.




                            * * *     E    N    D     * * *

  TIME IS NOW 12. 12. 32. 20.
```

```
                     S719 MTR 51  WAGE RATE - H   RESIDUALS P. 678 - DECK 35.  RUN 4




    NO.  OF  INPUT  DATA             2997
    NO.  OF  VARIABLES                 28
    NO.  OF  PREDICTORS                17
    WEIGHT  VARIABLE  NO.              26
    SPLIT ELIGIBILITY CRITERION     .0200
    SPLIT REDUCIBILITY CRITERION    .0050
    MAXIMUM  ALLOWABLE  GROUPS         63

    DEPENDENT VARIABLE IS  28  (RESIDUALS 51)
    VALUES OF DEPENDENT VARIABLE LARGER THAN -.00000000E 00   ARE OMITTED.
      ..         ..        ..      EQUAL   TO -.00000000E 00     ..
      ..         ..        ..         ..      -.00000000E 00     ..
    OUTPUT  OPTION 1  IS   1 .
    OUTPUT  OPTION 2  IS   0 .

    MINIMUM SIZE REQUIRED              25

    INPUT  DATA  ARE  ON           TAPE

    RESIDUALS  ARE  NOT  REQUESTED  AND  OUTPUT  WILL  BE NONE .

    NO  FILTERS.

    READ  DATA  BEGINS.

TIME IS NOW 12. 12. 36. 17.

     DATA  ARE  ALL  IN.

TIME IS NOW 12. 14. 21. 13.
```

```
* *    PREDICTOR    LISTING.    * *

VARIABLE  NO.       DESCRIPTION      MAXIMUM VALUE      T Y P E
       2            GEOG MOBILITY          5               M
       3            EDUCATION              7               M
       4            IMMIGRATION            2               M
       5            OCCUPATION             9               F
       6            SUPR RESP              2               F
       7            FREQ OF UNEM           9               F
       9            REL X ATTEND           6               F
      10            WORK X N/ACH           6               F
      11            RACE                   2               F
      13            H-W ED DIFF            6               F
      14            URB-RUR MTG            5               F
      15            N-S MIG                5               F
      16            FAM COMP               7               F
      17            INCOME COMM            3               F
      18            ABIL TO COMM           3               M
      19            SIZE OF PLAC           6               M
      20            H-F ED DIFF            3               F




* STATISTICS    FOR    TOTAL.

 TOTAL  NO.  OF  DATA  READ              2997
 NO.  OF  DATA  DELETED                   451
 TOTAL  NO.  OF  DATA  USED              2546
 SUM  OF  WEIGHTS               .11676900E 06
 SUM  OF  Y                     .91450000E 04
 SUM  OF  Y-SQUARE              .18539100E 10
 MEAN     Y                     .78317018E-01
 STANDARD  DEV.  Y              .12600287E 03
 TOTAL SUM OF SQUARES (TSS)     .18539092E 10

          PA =  3.707818E 07,              PB =  9.269546E 06

TIME IS NOW 12. 14. 21. 43.
```

```
**  S T E P  NO. =   1          PARENT  GROUP =  1 **

     TRY  ON   VARIABLE   2  OVER GROUP   1 .  RESULTS  FOLLOW.

     CODE     N    SUM OF WEIGHT     SUM  OF  Y     SUM  Y-SQUARE      MEAN          STD.   DEV.          B  S  S
      0    1358    .62668999E 05   -.37907100E 06   .93288603E 09  -.60487801E 01   .12185786E 03
                                                                                                        .50779986E 07
      1     286    .13128000E 05   -.21871600E 06   .19175844E 09  -.16660268E 02   .11970490E 03
                                                                                                        .13704528E 08
      2     515    .23342000E 05    .28944700E 06   .35306777E 09   .12400266E 02   .12236050E 03
                                                                                                        .66756172E 07
      3     180    .83629999E 04    .71092000E 05   .17207425E 09   .85007771E 01   .14319007E 03
                                                                                                        .70740207E 07
      4     157    .72120000E 04    .22456700E 06   .15037059E 09   .31137964E 02   .14039816E 03
                                                                                                        .23249790E 06
      5      50    .20550000E 04    .21826000E 05   .53757374E 08   .10620924E 02   .16138929E 03
                                                                                                        .18539137E 10
   *  FOR  VARIABLE   2 ( GEOG MOBILIT )   B S S =   .13704528E 08          BSS/TSS =   .00739

     TRY  ON   VARIABLE   3  OVER GROUP   1 .  RESULTS  FOLLOW.

     CODE     N    SUM OF WEIGHT     SUM  OF  Y     SUM  Y-SQUARE      MEAN          STD.   DEV.          B  S  S
      0      26    .79700000E 03   -.37707000E 05   .61453310E 07  -.47311167E 02   .73974535E 02
                                                                                                        .18021739E 07
      1     735    .30490000E 05   -.16334100E 06   .41991119E 09  -.53571990E 01   .11723223E 03
                                                                                                        .18080505E 07
      2     558    .26386000E 05   -.31696000E 05   .36308498E 09  -.12012431E 01   .11729907E 03
                                                                                                        .19286296E 07
      3     408    .19431000E 05   -.10491500E 06   .27635619E 09  -.53993618E 01   .11913557E 03
                                                                                                        .45102028E 07
      4     236    .11613000E 05    .13853600E 06   .17570759E 09   .11929389E 02   .12242524E 03
                                                                                                        .19924657E 07
      5     299    .14456000E 05    .20589300E 06   .23672578E 09   .14242736E 02   .12717213E 03
                                                                                                        .14289819E 03
      6     212    .10165000E 05   -.11011100E 06   .26055941E 09  -.10832366E 02   .15973621E 03
                                                                                                        .37813848E 07
      7      72    .34310000E 04    .11248600E 06   .11542463E 09   .32785193E 02   .18046279E 03
                                                                                                        .18539144E 10
   *  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .45102028E 07          BSS/TSS =   .00243

     TRY  ON   VARIABLE   4  OVER GROUP   1 .  RESULTS  FOLLOW.

     CODE     N    SUM OF WEIGHT     SUM  OF  Y     SUM  Y-SQUARE      MEAN          STD.   DEV.          B  S  S
      0     113    .55620000E 04   -.36174000E 05   .12004495E 09  -.65037756E 01   .14676769E 03
                                                                                                        .25301973E 06
      1     367    .17867000E 05    .15258700E 06   .31361849E 09   .85401578E 01   .13221199E 03
                                                                                                        .70098596E 06
      2    2066    .93340000E 05   -.10726800E 06   .14202484E 10  -.11492179E 01   .12334723E 03
                                                                                                        .18539112E 10
   *  FOR  VARIABLE   4 ( IMMIGRATION  )   B S S =   .70098596E 06          BSS/TSS =   .00038

     TRY  ON   VARIABLE   5  OVER GROUP   1 .  RESULTS  FOLLOW.

     CODE     N    SUM OF WEIGHT     SUM  OF  Y     SUM  Y-SQUARE      MEAN          STD.   DEV.          B  S  S
      1     274    .13469000E 05    .49634800E 06   .32130288E 09   .36851139E 02   .14998995E 03
                                                                                                        .20588112E 08
      2     136    .66810000E 04    .23380200E 06   .16224433E 09   .34995060E 02   .15185448E 03
                                                                                                        .31837196E 08
      5     409    .20177000E 05    .64484300E 06   .22747824E 09   .31959310E 02   .10125581E 03
                                                                                                        .71285823E 08
```

```
   3    184    .85829999E 04    .14372200E 06    .36052743E 09    .16744961E 02    .20426559E 03
                                                                                                       .80738650E 08
   6    482    .23047000E 05    .22640600E 06    .17100535E 09    .98236647E 01    .85576564E 02
                                                                                                       .10957256E 09
   4    336    .16961000E 05    .81376000E 05    .18108928E 09    .47978303E 01    .10321717E 03
                                                                                                       .15610508E 09
   0     53    .22170000E 04   -.10322700E 06    .32247831E 08   -.46561569E 02    .11125523E 03
                                                                                                       .14720672E 09
   7    421    .15936000E 05   -.77697100E 06    .17407982E 09   -.48755710E 02    .92447630E 02
                                                                                                       .98923416E 08
   9     54    .25750000E 04   -.14685000E 06    .25351742E 08   -.57029126E 02    .81197385E 02
                                                                                                       .93513232E 08
   8    197    .71230000E 04   -.79030400E 06    .19858848E 09   -.11095100E 03    .12477888E 03
                                                                                                       .18539146E 10
 *  FOR  VARIABLE   5 ( OCCUPATION    )   B S S =   .15610508E 09          BSS/TSS =   .08420

  TRY  ON   VARIABLE   6  OVER GROUP   1 .  RESULTS  FOLLOW.

  CODE     N    SUM OF WEIGHT      SUM  OF  Y     SUM  Y-SQUARE       MEAN            STD.   DEV.          B  S  S
   1    534    .26419000E 05    .80225600E 06    .45152199E 09    .30366630E 02    .12715610E 03
                                                                                                       .31323187E 08
   2   1572    .72095000E 05   -.23004300E 06    .78052735E 09   -.31908315E 01    .10400092E 03
                                                                                                       .20690552E 08
   0    440    .18255000E 05   -.56306800E 06    .62186485E 09   -.30844590E 02    .18197270E 03
                                                                                                       .18539135E 10
 *  FOR  VARIABLE   6 ( SUPR RESP     )   B S S =   .31323187E 08          BSS/TSS =   .01690

  TRY  ON   VARIABLE   7  OVER GROUP   1 .  RESULTS  FOLLOW.

  CODE     N    SUM OF WEIGHT      SUM  OF  Y     SUM  Y-SQUARE       MEAN            STD.   DEV.          B  S  S
   3     44    .20550000E 04    .52172000E 05    .24341138E 08    .25387834E 02    .10583144E 03
                                                                                                       .13399565E 07
   5    540    .25747000E 05    .38930000E 06    .33849911E 09    .15120208E 02    .11365961E 03
                                                                                                       .91103352E 07
   2     49    .22310000E 04    .29698000E 05    .20891052E 08    .13311519E 02    .95847739E 02
                                                                                                       .98523045E 07
   4    913    .43944000E 05    .44658400E 06    .46345082E 09    .10162570E 02    .10219157E 03
                                                                                                       .30677513E 08
   9    217    .98360000E 04    .82740000E 05    .15538102E 09    .84119561E 01    .12540500E 03
                                                                                                       .41763153E 08
   1    160    .63950000E 04   -.10692600E 06    .93087398E 08   -.16720250E 02    .11948519E 03
                                                                                                       .38299990E 08
   0    578    .24347000E 05   -.76388900E 06    .74235622E 09   -.31375077E 02    .17177389E 03
                                                                                                       .67081633E 07
   6     45    .22140000E 04   -.12053400E 06    .15908064E 08   -.54441734E 02    .64971621E 02
                                                                                                       .18539141E 10
 *  FOR  VARIABLE   7 ( FREQ OF UNEM )   B S S =   .41763153E 08          BSS/TSS =   .02253

  TRY  ON   VARIABLE   9  OVER GROUP   1 .  RESULTS  FOLLOW.

  CODE     N    SUM OF WEIGHT      SUM  OF  Y     SUM  Y-SQUARE       MEAN            STD.   DEV.          B  S  S
   6    162    .77320000E 04    .31032700E 06    .24378307E 09    .40135411E 02    .17296895E 03
                                                                                                       .13286310E 08
   1    112    .54579999E 04    .11789000E 06    .10976898E 09    .21599487E 02    .14016076E 03
                                                                                                       .15596997E 08
   4    455    .21122000E 05    .12677300E 06    .41072353E 09    .60019410E 01    .13931716E 03
                                                                                                       .12589499E 08
   5    510    .24300000E 05   -.19152000E 05    .40071176E 09   -.78814814E 00    .12841174E 03
                                                                                                       .96679214E 07
   0    431    .20970000E 05   -.10262500E 06    .24647000E 09   -.48938960E 01    .10830285E 03
```

```
                                                                                                              .71934526E 07
  2    499    .20835000E 05   -.16926600E 06    .27341738E 09   -.81241179E 01    .11426716E 03
                                                                                                              .46634755E 07
  3    377    .16352000E 05   -.25480200E 06    .16904052E 09   -.15582314E 02    .10047286E 03
                                                                                                              .18539145E 10
*  FOR  VARIABLE   9 ( REL X ATTEND )   B S S =   .15596997E 08          BSS/TSS =   .00841

 TRY  ON   VARIABLE  10  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN           STD.   DEV.            B  S  S
  0    624    .30109000E 05    .44443600E 06    .51722389E 09    .14760902E 02    .13023247E 03
                                                                                                              .87460156E 07
  3    120    .52770000E 04    .25273000E 05    .65230231E 08    .47892741E 01    .11107789E 03
                                                                                                              .88405690E 07
  1    913    .42631000E 05    .65183999E 05    .69837156E 09    .15290282E 01    .12798218E 03
                                                                                                              .10799386E 08
  4    227    .10017000E 05   -.30291000E 05    .16640484E 09   -.30239592E 01    .12885301E 03
                                                                                                              .11434421E 08
  6     90    .40420000E 04   -.45915000E 05    .10507942E 09   -.11359475E 02    .16083485E 03
                                                                                                              .10468305E 08
  5    138    .54929999E 04   -.93246999E 05    .68130314E 08   -.16975605E 02    .11006791E 03
                                                                                                              .79797991E 07
  2    434    .19200000E 05   -.35629500E 06    .23347454E 09   -.18557031E 02    .10870036E 03
                                                                                                              .18539140E 10
*  FOR  VARIABLE  10 ( WURK X N/ACH )   B.S S =   .11434421E 08          BSS/TSS =   .00617

 TRY  ON   VARIABLE  11  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN           STD.   DEV.            B  S  S
  1   2197    .10488800E 06    .42918500E 06    .17250888E 10    .40918408E 01    .12818041E 03
                                                                                                              .16605503E 08
  2    349    .11881000E 05   -.42004000E 06    .12882250E 09   -.35353926E 02    .97943003E 02
                                                                                                              .18539105E 10
*  FOR  VARIABLE  11 ( RACE          )   B S S =   .16605503E 08          BSS/TSS =   .00896

 TRY  ON   VARIABLE  13  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN           STD.   DEV.            B  S  S
  6     13    .54700000E 03    .14810000E 05    .14700200E 08    .27074954E 02    .16168231E 03
                                                                                                              .40053998E 06
  1    212    .99620000E 04    .26039300E 06    .18686364E 09    .26138626E 02    .13444112E 03
                                                                                                              .78722911E 07
  2    341    .15410000E 05    .22254400E 06    .24107806E 09    .14441531E 02    .12424050E 03
                                                                                                              .12185748E 08
  4    319    .15447000E 05    .14432000E 06    .27967078E 09    .93429144E 01    .13423075E 03
                                                                                                              .15277855E 08
  3    726    .32705000E 05    .12973600E 06    .59017299E 09    .39668552E 01    .13427438E 03
                                                                                                              .21663648E 08
  5    285    .13715000E 05   -.68979999E 05    .26662249E 09   -.50295296E 01    .13933741E 03
                                                                                                              .22228613E 08
  0    650    .28983000E 05   -.69367800E 06    .27480699E 09   -.23933961E 02    .94386581E 02
                                                                                                              .18539144E 10
*  FOR  VARIABLE  13 ( H-W ED DIFF  )   B S S =   .22228613E 08          BSS/TSS =   .01199

 TRY  ON   VARIABLE  14  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE       MEAN           STD.   DEV.            B  S  S
  2    134    .63559999E 04    .32991900E 06    .95259336E 08    .51906702E 02    .11087382E 03
                                                                                                              .18056209E 08
  3    444    .20438000E 05    .19536900E 06    .36917398E 09    .95591055E 01    .13405872E 03
```

```
                                                                                                  .13258253E 08
  1   207   .89940000E 04   .63586000E 05   .12390663E 09   .70698243E 01   .11716060E 03
                                                                                                  .13839096E 08
  4  1215   .59123000E 05   .66297000E 05   .88747330E 09   .11213402E 01   .12251273E 03
                                                                                                  .23615608E 08
  5    54   .24730000E 04  -.34561000E 05   .42873229E 08  -.13975333E 02   .13092447E 03
                                                                                                  .23241880E 08
  0   492   .19385000E 05  -.61146500E 06   .33522815E 09  -.31543203E 02   .12766440E 03
                                                                                                  .18539139E 10
* FOR  VARIABLE  14 ( URB-RUR MTG  )   B S S =   .23615608E 08        BSS/TSS =   .01274

 TRY  ON   VARIABLE  15  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE     N    SUM OF WEIGHT      SUM  OF  Y     SUM  Y-SQUARE       MEAN          STD.   DEV.       B  S  S
   3    106   .48720000E 04   .23514900E 06   .61983485E 08   .48265393E 02   .10194529E 03
                                                                                                  .11805314E 08
   0    353   .15317000E 05   .30458000E 05   .30394021E 09   .19885095E 01   .14085230E 03
                                                                                                  .41746346E 07
   1   1447   .69533999E 05   .86921000E 05   .10662791E 10   .12500503E 01   .12382681E 03
                                                                                                  .57440703E 07
   4    113   .55620000E 04  -.36174000E 05   .12004495E 09  -.65037756E 01   .14676769E 03
                                                                                                  .54425190E 07
   5     35   .15510000E 04  -.16947000E 05   .35239977E 08  -.10926499E 02   .15033769E 03
                                                                                                  .51517805E 07
   2    492   .19933000E 05  -.29026200E 06   .26642671E 09  -.14561882E 02   .11469117E 03
                                                                                                  .18539136E 10
* FOR  VARIABLE  15 ( N-S MIG      )   B S S =   .11805314E 08        BSS/TSS =   .00637

 TRY  ON   VARIABLE  16  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE     N    SUM OF WEIGHT      SUM  OF  Y     SUM  Y-SQUARE       MEAN          STD.   DEV.       B  S  S
   7    517   .22176000E 05   .26947000E 06   .34072566E 09   .12151425E 02   .12335702E 03
                                                                                                  .39901565E 07
   4    616   .29457000E 05   .30150500E 06   .64838538E 09   .10235428E 02   .14800839E 03
                                                                                                  .11159376E 08
   6    397   .18499000E 05   .18069700E 06   .34170790E 09   .97679333E 01   .13555915E 03
                                                                                                  .19877754E 08
   5    366   .17654000E 05  -.48849000E 05   .24828910E 09  -.27670216E 01   .11856023E 03
                                                                                                  .22228613E 08
   2    268   .12196000E 05  -.39800000E 05   .84859411E 08  -.32633650E 01   .83350591E 02
                                                                                                  .29865566E 08
   3    116   .46270000E 04  -.10576400E 06   .22806848E 08  -.22858007E 02   .66382153E 02
                                                                                                  .27674167E 08
   1     16   .64200000E 03  -.19460000E 05   .83070999E 07  -.30311526E 02   .10963858E 03
                                                                                                  .27011464E 08
   0    250   .11518000E 05  -.52865399E 06   .15883367E 09  -.45898072E 02   .10808981E 03
                                                                                                  .18539143E 10
* FOR  VARIABLE  16 ( FAM COMP     )   B S S =   .29865566E 08        BSS/TSS =   .01611

 TRY  ON   VARIABLE  17  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE     N    SUM OF WEIGHT      SUM  OF  Y     SUM  Y-SQUARE       MEAN          STD.   DEV.       B  S  S
   2    518   .24330000E 05   .46236200E 06   .36813924E 09   .19003781E 02   .12153164E 03
                                                                                                  .11007977E 08
   3    327   .15699000E 05   .27440900E 06   .27255267E 09   .17479393E 02   .13059716E 03
                                                                                                  .20459374E 08
   1    461   .21089000E 05  -.33520000E 04   .31129267E 09  -.15894542E 00   .12149435E 03
                                                                                                  .18226464E 08
   0   1240   .55650999E 05  -.72427399E 06   .90192975E 09  -.13014573E 02   .12663930E 03
                                                                                                  .18539136E 10
```

```
*  FOR  VARIABLE  17 ( INCOME COMM  )   B S S =   .20459374E 08         BSS/TSS =   .01104

 TRY  ON   VARIABLE  18  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE        MEAN           STD.   DEV.               B  S  S
   0    2024    .95421000E 05     .50798300E 06    .15294659E 10    .53235975E 01     .12649217E 03
                                                                                                             .14359911E 08
   1     406    .16913000E 05    -.37307300E 06    .25502674E 09   -.22058357E 02     .12079805E 03
                                                                                                             .37276733E 07
   2      92    .34920000E 04    -.14131500E 06    .39060813E 08   -.40468213E 02     .97714500E 02
                                                                                                             .25605632E 06
   3      24    .94300000E 03     .15550000E 05    .30358278E 08    .16489926E 02     .17866555E 03
                                                                                                             .18539110E 10
*  FOR  VARIABLE  18 ( ABIL TO COMM )   B S S =   .14359911E 08         BSS/TSS =   .00775

 TRY  ON   VARIABLE  19  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE        MEAN           STD.   DEV.               B  S  S
   1     348    .17082000E 05     .13970800E 06    .24201426E 09    .81786675E 01     .11874722E 03
                                                                                                             .13129112E 07
   2     462    .22693000E 05     .12220700E 06    .28641351E 09    .53852288E 01     .11221509E 03
                                                                                                             .25538114E 07
   3     310    .14160000E 05     .17119100E 06    .28577450E 09    .12089760E 02     .14154735E 03
                                                                                                             .63377886E 07
   4     471    .22191000E 05     .22851000E 05    .32492352E 09    .10297418E 01     .12100029E 03
                                                                                                             .76423027E 07
   5     375    .17819000E 05     .54461300E 06    .28852334E 09    .30563612E 02     .12352230E 03
                                                                                                             .53721296E 08
   6     580    .22824000E 05    -.99142499E 06    .42626606E 09   -.43437828E 02     .12957382E 03
                                                                                                             .18539144E 10
*  FOR  VARIABLE  19 ( SIZE OF PLAC )   B S S =   .53721296E 08         BSS/TSS =   .02898

 TRY  ON   VARIABLE  20  OVER GROUP   1 .  RESULTS  FOLLOW.

 CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE        MEAN           STD.   DEV.               B  S  S
   2     652    .31367000E 05     .29240700E 06    .56337605E 09    .93221218E 01     .13369325E 03
                                                                                                             .36646630E 07
   0     161    .76679999E 04     .48185000E 05    .10153274E 09    .62839071E 01     .11489826E 03
                                                                                                             .43842943E 07
   3     287    .13702000E 05    -.56995000E 05    .27367156E 09   -.41596117E 01     .14126503E 03
                                                                                                             .27006947E 07
   1    1446    .64032000E 05    -.27445200E 06    .91533370E 09   -.42861694E 01     .11948460E 03
                                                                                                             .18539133E 10
*  FOR  VARIABLE  20 ( H-F ED DIFF  )   B S S =   .43842943E 07         BSS/TSS =   .00236


DECOMPOSE GROUP   1  INTO GROUP   2  AND   3  BY VARIABLE   5  IN  S T E P   1 .

 CODE      N    SUM OF WEIGHT       SUM  OF  Y      SUM  Y-SQUARE        MEAN           STD.   DEV.               B  S  S
   1     274    .13469000E 05     .49634800E 06    .32130288E 09    .36851139E 02     .14998995E 03
                                                                                                             .20588112E 08
   2     136    .66810000E 04     .23380200E 06    .16224433E 09    .34995060E 02     .15185448E 03
                                                                                                             .31837196E 08
   5     409    .20177000E 05     .64484300E 06    .22747824E 09    .31959310E 02     .10125581E 03
                                                                                                             .71285823E 08
   3     184    .85829999E 04     .14372200E 06    .36052743E 09    .16744961E 02     .20426559E 03
                                                                                                             .80738650E 08
   6     482    .23047000E 05     .22640600E 06    .17100535E 09    .98236647E 01     .85576564E 02
                                                                                                             .10957256E 09
```

```
   4     336    .16961000E 05    .81376000E 05    .18108928E 09    .47978303E 01    .10321717E 03
                                                                                                      .15610508E 09
   0      53    .22170000E 04   -.10322700E 06    .32247831E 08   -.46561569E 02    .11125523E 03
                                                                                                      .14720672E 09
   7     421    .15936000E 05   -.77697100E 06    .17407982E 09   -.48755710E 02    .92447630E 02
                                                                                                      .98923416E 08
   9      54    .25750000E 04   -.14685000E 06    .25351742E 08   -.57029126E 02    .81197385E 02
                                                                                                      .93513232E 08
   8     197    .71230000E 04   -.79030400E 06    .19858848E 09   -.11095100E 03    .12477888E 03
                                                                                                      .18539146E 10


 CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

  GROUP        N         TOTAL WEIGHT           SUM OF Y           SUM  Y-SQUARE            T   S   S
    2       1821        .86918000E 05       .18264970E 07       .14236475E 10        .13861287E 10
    3        725        .27851000E 05      -.18173520E 07       .43026788E 09        .31168081E 09




**   S T E P  NO. =    2          PARENT  GROUP =  2 **
     *  FOR  VARIABLE   2 ( GEOG MOBILIT )   B S S =   .90724754E 07         BSS/TSS =   .00655
     *  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .18968550E 07         BSS/TSS =   .00137
     *  FOR  VARIABLE   4 ( IMMIGRATION  )   B S S =   .51775650E 06         BSS/TSS =   .00037
     *  FOR  VARIABLE   5 ( OCCUPATION   )   B S S =   .13558486E 08         BSS/TSS =   .00978
     *  FOR  VARIABLE   6 ( SUPR RESP    )   B S S =   .10490296E 08         BSS/TSS =   .00757
     *  FOR  VARIABLE   7 ( FREQ OF UNEM )   B S S =   .93321775E 07         BSS/TSS =   .00673
     *  FOR  VARIABLE   9 ( REL X ATTEND )   B S S =   .85667425E 07         BSS/TSS =   .00618
     *  FOR  VARIABLE  10 ( WORK X N/ACH )   B S S =   .28859635E 07         BSS/TSS =   .00208
     *  FOR  VARIABLE  11 ( RACE         )   B S S =   .38651645E 07         BSS/TSS =   .00279
     *  FOR  VARIABLE  13 ( H-W ED DIFF  )   B S S =   .26311622E 08         BSS/TSS =   .01898
     *  FOR  VARIABLE  14 ( URB-RUR MTG  )   B S S =   .10759083E 08         BSS/TSS =   .00776
     *  FOR  VARIABLE  15 ( N-S MIG      )   B S S =   .70148360E 07         BSS/TSS =   .00506
     *  FOR  VARIABLE  16 ( FAM COMP     )   B S S =   .23450021E 08         BSS/TSS =   .01692
     *  FOR  VARIABLE  17 ( INCOME COMM  )   B S S =   .77834205E 07         BSS/TSS =   .00562
     *  FOR  VARIABLE  18 ( ABIL TO COMM )   B S S =   .28619235E 07         BSS/TSS =   .00206
     *  FOR  VARIABLE  19 ( SIZE OF PLAC )   B S S =   .95129550E 07         BSS/TSS =   .00686
     *  FOR  VARIABLE  20 ( H-F ED DIFF  )   B S S =   .12748730E 07         BSS/TSS =   .00092


 DECOMPOSE GROUP   2  INTO GROUP   4  AND   5  BY VARIABLE  13   IN  S T E P    2 .

  CODE      N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN            STD.   DEV.           B   S   S
    6       9    .40200000E 03    .24680000E 05    .12316098E 08    .61393034E 02    .16391447E 03
                                                                                                      .67392850E 06
    1     149    .72770000E 04    .40890700E 06    .15007534E 09    .56191699E 02    .13215799E 03
                                                                                                      .10845910E 08
    2     244    .11643000E 05    .45909900E 06    .19921152E 09    .39431332E 02    .12472029E 03
                                                                                                      .16253296E 08
```

```
   3    496    .23911000E 05    .64921000E 06    .39837207E 09    .27151102E 02    .12618810E 03
                                                                                                  .19245623E 08
   4    264    .13057000E 05    .35158500E 06    .24668512E 09    .26926935E 02    .13478828E 03
                                                                                                  .26311622E 08
   5    251    .12260000E 05    .49329000E 05    .24150190E 09    .40235725E 01    .14029316E 03
                                                                                                  .18207629E 08
   0    408    .20368000E 05   -.11631300E 06    .17548545E 09   -.57105754E 01    .92645194E 02
                                                                                                  .13861287E 10



CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP          N          TOTAL WEIGHT           SUM OF Y           SUM. Y-SQUARE          T   S   S
   3          725        .27851000E 05       -.18173520E 07        .43026788E 09       .31168081E 09
   4         1162        .56290000E 05        .18934810E 07        .10066601E 10       .94296728E 09
   5          659        .32628000E 05       -.66983999E 05        .41698735E 09       .41684983E 09



** S T E P  NO. =   3          PARENT  GROUP =  4 **
    *  FOR  VARIABLE   2 ( GEOG MOBILIT )   B S S =   .71377525E 07        BSS/TSS =   .00757
    *  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .60929205E 07        BSS/TSS =   .00646
    *  FOR  VARIABLE   4 ( IMMIGRATION  )   B S S =   .13292300E 06        BSS/TSS =   .00014
    *  FOR  VARIABLE   5 ( OCCUPATION   )   B S S =   .98246095E 07        BSS/TSS =   .01042
    *  FOR  VARIABLE   6 ( SUPR RESP    )   B S S =   .66356485E 07        BSS/TSS =   .00704
    *  FOR  VARIABLE   7 ( FREQ OF UNEM )   B S S =   .17483480E 07        BSS/TSS =   .00185
    *  FOR  VARIABLE   9 ( REL X ATTEND )   B S S =   .80314264E 07        BSS/TSS =   .00852
    *  FOR  VARIABLE  10 ( WORK X N/ACH )   B S S =   .16629035E 07        BSS/TSS =   .00176
    *  FOR  VARIABLE  11 ( RACE         )   B S S =   .20889165E 07        BSS/TSS =   .00222
    *  FOR  VARIABLE  13 ( H-W ED DIFF  )   B S S =   .46431395E 07        BSS/TSS =   .00492
    *  FOR  VARIABLE  14 ( URB-RUR MTG  )   B S S =   .65482695E 07        BSS/TSS =   .00694
    *  FOR  VARIABLE  15 ( N-S MIG      )   B S S =   .64175565E 07        BSS/TSS =   .00681
    *  FOR  VARIABLE  16 ( FAM COMP     )   B S S =   .42144695E 07        BSS/TSS =   .00447
    *  FOR  VARIABLE  17 ( INCOME COMM  )   B S S =   .87210700E 06        BSS/TSS =   .00092
    *  FOR  VARIABLE  18 ( ABIL TO COMM )   B S S =   .49424185E 07        BSS/TSS =   .00524
    *  FOR  VARIABLE  19 ( SIZE OF PLAC )   B S S =   .61822464E 07        BSS/TSS =   .00656
    *  FOR  VARIABLE  20 ( H-F ED DIFF  )   B S S =   .26351700E 06        BSS/TSS =   .00028


DECOMPOSE GROUP   4  INTO GROUP   6  AND   7  BY VARIABLE   5  IN  S T E P   3 .

 CODE      N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN.           STD.  DEV.             B  S  S
   1     121    .59379999E 04    .33054200E 06    .16191511E 09    .55665543E 02    .15546370E 03
                                                                                                  .32209820E 07
   5     304    .14888000E 05    .65173300E 06    .17365376E 09    .43775725E 02    .98730412E 02
                                                                                                  .60493144E 07
   2      97    .47650000E 04    .20443900E 06    .11459013E 09    .42904302E 02    .14902186E 03
                                                                                                  .76093595E 07
```

```
  3    139    .63670000E 04    .25668900E 06    .30022628E 09    .40315533E 02    .21337326E 03
                                                                                                 .98246095E 07
  4    146    .73249999E 04    .14575300E 06    .11996997E 09    .19898020E 02    .12642081E 03
                                                                                                 .60405545E 07
  6    355    .17007000E 05    .30432500E 06    .13630499E 09    .17894102E 02    .87717967E 02
                                                                                                 .94296738E 09
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N        TOTAL WEIGHT          SUM OF Y          SUM  Y-SQUARE          T   S   S
   3         725      .27851000E 05     -.18173520E 07      .43026788E 09      .31168081E 09
   5         659      .32628000E 05     -.66983999E 05      .41698735E 09      .41684983E 09
   6         661      .31958000E 05      .14434030E 07      .75038527E 09      .68519307E 09
   7         501      .24332000E 05      .45007800E 06      .25627496E 09      .24794970E 09
```

```
** S T E P  NO. =   4        PARENT  GROUP =  6 **
   *  FOR  VARIABLE   2 ( GEOG MOBILIT )   B S S =   .77766420E 07        BSS/TSS =   .01135
   *  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .43297340E 07        BSS/TSS =   .00632
   *  FOR  VARIABLE   4 ( IMMIGRATION  )   B S S =   .23003670E 07        BSS/TSS =   .00336
   *  FOR  VARIABLE   5 ( OCCUPATION   )   B S S =   .80405249E 06        BSS/TSS =   .00117
   *  FOR  VARIABLE   6 ( SUPR RESP    )   B S S =   .45044550E 07        BSS/TSS =   .00657
   *  FOR  VARIABLE   7 ( FREQ OF UNEM )   B S S =   .21521910E 07        BSS/TSS =   .00314
   *  FOR  VARIABLE   9 ( REL X ATTEND )   B S S =   .79437280E 07        BSS/TSS =   .01159
   *  FOR  VARIABLE  10 ( WORK X N/ACH )   B S S =   .89745500E 06        BSS/TSS =   .00131
   *  FOR  VARIABLE  11 ( RACE         )   B S S =   .12451900E 06        BSS/TSS =   .00018
   *  FOR  VARIABLE  13 ( H-W ED DIFF  )   B S S =   .58364060E 07        BSS/TSS =   .00852
   *  FOR  VARIABLE  14 ( URB-RUR MTG  )   B S S =   .36815830E 07        BSS/TSS =   .00537
   *  FOR  VARIABLE  15 ( N-S MIG      )   B S S =   .70502640E 07        BSS/TSS =   .01029
   *  FOR  VARIABLE  16 ( FAM COMP     )   B S S =   .46620330E 07        BSS/TSS =   .00680
   *  FOR  VARIABLE  17 ( INCOME COMM  )   B S S =   .83204750E 06        BSS/TSS =   .00121
   *  FOR  VARIABLE  18 ( ABIL TO COMM )   B S S =   .40979280E 07        BSS/TSS =   .00598
   *  FOR  VARIABLE  19 ( SIZE OF PLAC )   B S S =   .89391670E 07        BSS/TSS =   .01305
   *  FOR  VARIABLE  20 ( H-F ED DIFF  )   B S S =   .24707560E 07        BSS/TSS =   .00361

   FAILED TO SPLIT GROUP   6  TRIED ON VARIABLE  19 , BUT   BSS =   .89391670E 07
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N       TOTAL WEIGHT            SUM OF Y           SUM  Y-SQUARE           T   S   S
   3        725       .27851000E 05       -.18173520E 07       .43026788E 09       .31168081E 09
   5        659       .32628000E 05       -.66983999E 05       .41698735E 09       .41684983E 09
   7        501       .24332000E 05        .45007800E 06       .25627496E 09       .24794970E 09
*  FOR  VARIABLE   2 ( GEOG MOBILIT )  B S S =    .18834347E 07         BSS/TSS =   .00452
*  FOR  VARIABLE   3 ( EDUCATION    )  B S S =    .32008271E 07         BSS/TSS =   .00768
*  FOR  VARIABLE   4 ( IMMIGRATION  )  B S S =    .60591387E 06         BSS/TSS =   .00145
*  FOR  VARIABLE   5 ( OCCUPATION   )  B S S =    .68592293E 07         BSS/TSS =   .01645
*  FOR  VARIABLE   6 ( SUPR RESP    )  B S S =    .28988592E 07         BSS/TSS =   .00695
*  FOR  VARIABLE   7 ( FREQ OF UNEM )  B S S =    .89513274E 07         BSS/TSS =   .02147
*  FOR  VARIABLE   9 ( REL X ATTEND )  B S S =    .54986674E 07         BSS/TSS =   .01319
*  FOR  VARIABLE  10 ( WORK X N/ACH )  B S S =    .39300873E 07         BSS/TSS =   .00943
*  FOR  VARIABLE  11 ( RACE         )  B S S =    .21088705E 07         BSS/TSS =   .00506
*  FOR  VARIABLE  13 ( H-W ED DIFF  )  B S S =    .72517744E 06         BSS/TSS =   .00174
*  FOR  VARIABLE  14 ( URB-RUR MTG  )  B S S =    .44318099E 07         BSS/TSS =   .01063
*  FOR  VARIABLE  15 ( N-S MIG      )  B S S =    .17585135E 07         BSS/TSS =   .00422
*  FOR  VARIABLE  16 ( FAM COMP     )  B S S =    .81519191E 07         BSS/TSS =   .01956
*  FOR  VARIABLE  17 ( INCOME COMM  )  B S S =    .37142788E 07         BSS/TSS =   .00891
*  FOR  VARIABLE  18 ( ABIL TO COMM )  B S S =    .83621801E 06         BSS/TSS =   .00201
*  FOR  VARIABLE  19 ( SIZE OF PLAC )  B S S =    .77507474E 07         BSS/TSS =   .01859
*  FOR  VARIABLE  20 ( H-F ED DIFF  )  B S S =    .44177350E 07         BSS/TSS =   .01060

FAILED TO SPLIT GROUP   5  TRIED ON VARIABLE   7 , BUT   BSS =   .89513274E 07




CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N       TOTAL WEIGHT            SUM OF Y           SUM  Y-SQUARE           T   S   S
   3        725       .27851000E 05       -.18173520E 07       .43026788E 09       .31168081E 09
   7        501       .24332000E 05        .45007800E 06       .25627496E 09       .24794970E 09
*  FOR  VARIABLE   2 ( GEOG MOBILIT )  B S S =    .14394610E 07         BSS/TSS =   .00462
*  FOR  VARIABLE   3 ( EDUCATION    )  B S S =    .21420190E 07         BSS/TSS =   .00687
*  FOR  VARIABLE   4 ( IMMIGRATION  )  B S S =    .11012900E 06         BSS/TSS =   .00035
*  FOR  VARIABLE   5 ( OCCUPATION   )  B S S =    .19986978E 08         BSS/TSS =   .06413
*  FOR  VARIABLE   6 ( SUPR RESP    )  B S S =    .12670401E 08         BSS/TSS =   .04065
*  FOR  VARIABLE   7 ( FREQ OF UNEM )  B S S =    .72915199E 07         BSS/TSS =   .02339
*  FOR  VARIABLE   9 ( REL X ATTEND )  B S S =    .20821010E 07         BSS/TSS =   .00668
*  FOR  VARIABLE  10 ( WORK X N/ACH )  B S S =    .26076620E 07         BSS/TSS =   .00837
*  FOR  VARIABLE  11 ( RACE         )  B S S =    .29192000E 05         BSS/TSS =   .00009
*  FOR  VARIABLE  13 ( H-W ED DIFF  )  B S S =    .16826890E 07         BSS/TSS =   .00540
*  FOR  VARIABLE  14 ( URB-RUR MTG  )  B S S =    .14023486E 08         BSS/TSS =   .04499
*  FOR  VARIABLE  15 ( N-S MIG      )  B S S =    .77869079E 07         BSS/TSS =   .02498
*  FOR  VARIABLE  16 ( FAM COMP     )  B S S =    .25444590E 07         BSS/TSS =   .00816
*  FOR  VARIABLE  17 ( INCOME COMM  )  B S S =    .19726020E 07         BSS/TSS =   .00633
*  FOR  VARIABLE  18 ( ABIL TO COMM )  B S S =    .20171800E 06         BSS/TSS =   .00065
*  FOR  VARIABLE  19 ( SIZE OF PLAC )  B S S =    .13019936E 08         BSS/TSS =   .04177
*  FOR  VARIABLE  20 ( H-F ED DIFF  )  B S S =    .53018200E 06         BSS/TSS =   .00170


DECOMPOSE GROUP   3  INTO GROUP   8  AND   9  BY VARIABLE   5  IN  S T E P   4 .

 CODE     N    SUM OF WEIGHT      SUM  OF  Y      SUM  Y-SQUARE       MEAN           STD.  DEV.            B  S  S
   0     53    .22170000E 04   -.10322700E 06   .32247831E 08   -.46561569E 02    .11125523E 03
                                                                                                  .84151100E 06
```

```
      7     421     .15936000E 05    -.77697100E 06    .17407982E 09   -.48755710E 02    .92447630E 02
                                                                                                                .14652446E 08
      9      54     .25750000E 04    -.14685000E 06    .25351742E 08   -.57029126E 02    .81197385E 02
                                                                                                                .19986978E 08
      8     197     .71230000E 04    -.79030400E 06    .19858848E 09   -.11095100E 03    .12477888E 03
                                                                                                                .31168081E 09




    CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

     GROUP          N        TOTAL WEIGHT          SUM OF Y          SUM  Y-SQUARE           T   S   S
       7          501       .24332000E 05      .45007800E 06       .25627496E 09      .24794970E 09
       8          528       .20728000E 05     -.10270480E 07       .23167940E 09      .18079038E 09
       9          197       .71230000E 04     -.79030400E 06       .19858848E 09      .11090346E 09




 ** S T E P  NO. =    5          PARENT  GROUP =  7 **
    *  FOR  VARIABLE    2 ( GEOG MOBILIT )   B S S =   .91587362E 06         BSS/TSS =   .00369
    *  FOR  VARIABLE    3 ( EDUCATION    )   B S S =   .92468287E 06         BSS/TSS =   .00373
    *  FOR  VARIABLE    4 ( IMMIGRATION  )   B S S =   .18653316E 07         BSS/TSS =   .00752
    *  FOR  VARIABLE    5 ( OCCUPATION   )   B S S =   .20559625E 05         BSS/TSS =   .00008
    *  FOR  VARIABLE    6 ( SUPR RESP    )   B S S =   .38007750E 06         BSS/TSS =   .00153
    *  FOR  VARIABLE    7 ( FREQ OF UNEM )   B S S =   .18213852E 07         BSS/TSS =   .00735
    *  FOR  VARIABLE    9 ( REL X ATTEND )   B S S =   .39891300E 07         BSS/TSS =   .01609
    *  FOR  VARIABLE   10 ( WORK X N/ACH )   B S S =   .66312012E 06         BSS/TSS =   .00267
    *  FOR  VARIABLE   11 ( RACE         )   B S S =   .11681592E 07         BSS/TSS =   .00471
    *  FOR  VARIABLE   13 ( H-W ED DIFF  )   B S S =   .13573470E 07         BSS/TSS =   .00547
    *  FOR  VARIABLE   14 ( URB-RUR MTG  )   B S S =   .64697335E 07         BSS/TSS =   .02609
    *  FOR  VARIABLE   15 ( N-S MIG      )   B S S =   .43346047E 07         BSS/TSS =   .01748
    *  FOR  VARIABLE   16 ( FAM COMP     )   B S S =   .80035762E 06         BSS/TSS =   .00323
    *  FOR  VARIABLE   17 ( INCOME COMM  )   B S S =   .10340661E 07         BSS/TSS =   .00417
    *  FOR  VARIABLE   18 ( ABIL TO COMM )   B S S =   .84908037E 06         BSS/TSS =   .00342
    *  FOR  VARIABLE   19 ( SIZE OF PLAC )   B S S =   .62111025E 06         BSS/TSS =   .00250
    *  FOR  VARIABLE   20 ( H-F ED DIFF  )   B S S =   .25363111E 07         BSS/TSS =   .01023

    FAILED TO SPLIT GROUP    7  TRIED ON VARIABLE  14 , BUT   BSS =   .64697335E 07
```

```
CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N        TOTAL WEIGHT          SUM OF Y         SUM  Y-SQUARE          T   S   S
    8        528       .20728000E 05     -.10270480E 07      .23167940E 09       .18079038E 09
    9        197       .71230000E 04     -.79030400E 06      .19858848E 09       .11090346E 09
*  FOR  VARIABLE   2 ( GEOG MOBILIT )   B S S =   .90803350E 06          BSS/TSS =   .00502
*  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .31286315E 07          BSS/TSS =   .01731
*  FOR  VARIABLE   4 ( IMMIGRATION  )   B S S =   .82016999E 05          BSS/TSS =   .00045
*  FOR  VARIABLE   5 ( OCCUPATION   )   B S S =   .16452200E 06          BSS/TSS =   .00091
*  FOR  VARIABLE   6 ( SUPR RESP    )   B S S =   .52654450E 06          BSS/TSS =   .00291
*  FOR  VARIABLE   7 ( FREQ OF UNEM )   B S S =   .10977140E 07          BSS/TSS =   .00607
*  FOR  VARIABLE   9 ( REL X ATTEND )   B S S =   .20888350E 07          BSS/TSS =   .01155
*  FOR  VARIABLE  10 ( WORK X N/ACH )   B S S =   .21217685E 07          BSS/TSS =   .01174
*  FOR  VARIABLE  11 ( RACE         )   B S S =   .54616400E 06          BSS/TSS =   .00302
*  FOR  VARIABLE  13 ( H-W ED DIFF  )   B S S =   .40359760E 07          BSS/TSS =   .02232
*  FOR  VARIABLE  14 ( URB-RUR MTG  )   B S S =   .66580199E 07          BSS/TSS =   .03683
*  FOR  VARIABLE  15 ( N-S MIG      )   B S S =   .44962625E 07          BSS/TSS =   .02487
*  FOR  VARIABLE  16 ( FAM COMP     )   B S S =   .26090160E 07          BSS/TSS =   .01443
*  FOR  VARIABLE  17 ( INCOME COMM  )   B S S =   .41310300E 06          BSS/TSS =   .00228
*  FOR  VARIABLE  18 ( ABIL TO COMM )   B S S =   .46914350E 06          BSS/TSS =   .00259
*  FOR  VARIABLE  19 ( SIZE OF PLAC )   B S S =   .23942800E 07          BSS/TSS =   .01324
*  FOR  VARIABLE  20 ( H-F ED DIFF  )   B S S =   .11693665E 07          BSS/TSS =   .00647

FAILED TO SPLIT GROUP   8  TRIED ON VARIABLE  14 ; BUT   BSS =   .66580199E 07




CANDIDATE  GROUPS  ARE  AS  FOLLOWS.

 GROUP        N        TOTAL WEIGHT          SUM OF Y         SUM  Y-SQUARE          T   S   S
    9        197       .71230000E 04     -.79030400E 06      .19858848E 09       .11090346E 09
*  FOR  VARIABLE   2 ( GEOG MOBILIT )   B S S =   .67627899E 06          BSS/TSS =   .00610
*  FOR  VARIABLE   3 ( EDUCATION    )   B S S =   .10352130E 07          BSS/TSS =   .00933
*  FOR  VARIABLE   4 ( IMMIGRATION  )   B S S =   .67561299E 06          BSS/TSS =   .00609

VARIABLE   5  OVER GROUP   9  IS A CONSTANT.    S T E P   =   5 .
*  FOR  VARIABLE   6 ( SUPR RESP    )   B S S =   .12858340E 07          BSS/TSS =   .01159
*  FOR  VARIABLE   7 ( FREQ OF UNEM )   B S S =   .49016400E 06          BSS/TSS =   .00442
*  FOR  VARIABLE   9 ( REL X ATTEND )   B S S =   .17928190E 07          BSS/TSS =   .01617
*  FOR  VARIABLE  10 ( WORK X N/ACH )   B S S =   .11293520E 07          BSS/TSS =   .01018
*  FOR  VARIABLE  11 ( RACE         )   B S S =   .19863000E 05          BSS/TSS =   .00018
*  FOR  VARIABLE  13 ( H-W ED DIFF  )   B S S =   .32641120E 07          BSS/TSS =   .02943
*  FOR  VARIABLE  14 ( URB-RUR MTG  )   B S S =   .11210810E 07          BSS/TSS =   .01011
*  FOR  VARIABLE  15 ( N-S MIG      )   B S S =   .56207870E 07          BSS/TSS =   .05068
*  FOR  VARIABLE  16 ( FAM COMP     )   B S S =   .13236100E 07          BSS/TSS =   .01193
*  FOR  VARIABLE  17 ( INCOME COMM  )   B S S =   .32449830E 07          BSS/TSS =   .02926
*  FOR  VARIABLE  18 ( ABIL TO COMM )   B S S =   .11247760E 07          BSS/TSS =   .01014
*  FOR  VARIABLE  19 ( SIZE OF PLAC )   B S S =   .64530200E 06          BSS/TSS =   .00582
*  FOR  VARIABLE  20 ( H-F ED DIFF  )   B S S =   .96574000E 06          BSS/TSS =   .00871

FAILED TO SPLIT GROUP   9  TRIED ON VARIABLE  15 , BUT   BSS =   .56207870E 07
```

```
      CANDIDATE  GROUPS  ARE  AS  FOLLOWS.
       GROUP       N      TOTAL WEIGHT        SUM OF Y        SUM  Y-SQUARE        T   S   S






      THAT IS ALL.  NO MORE GROUPS ARE AVAILABLE. FINAL  S T E P  NO. IS   5   NO. OF GROUPS ARE   9 .
     ** THIS  IS  THE  END  OF  2ND  CORE.
    TIME IS NOW 12. 15. 21. 18.
```

```
                         *  (A)UTOMATIC  (I)NTERACTION  (D)ETECTOR    (MODEL  2)  *

                                        *  *  *  S  U  M  M  A  R  Y  *  *  *

      DEPENDENT VARIABLE  28  ( RESIDUALS 51)

      WEIGHTED  BY  VARIABLE   26

  ** TOTAL  GROUP

          N  =         2546              MEAN =  .78317018E-01            SUM Y =  .91450000E 04          TSS =  .18539092E 10
   TOTAL WT SUM=       116769         STD. DEV. =  .12600287E 03        SUM Y SQ. =  .18539100E 10

   *   GROUP  NO.   2  SPLIT  FROM  GROUP   1  ON  VARIABLE    5 (OCCUPATION  )
       VALUES  OF  PREDICTOR  INCLUDED  ARE    1    2    3    4    5    6
          N  =         1821              MEAN =  .20541364E 02         GROUP DEVIATION =  .20463046E 02        SUM Y =  .18264970E 07
    WEIGHT SUM =       88918         STD. DEV. =  .12485529E 03                 TSS(I) =  .13861287E 10    SUM Y SQ. =  .14236475E 10
  PCT OF TOTAL =    76.1         WTD. MEAN SQ. =  .37518739E 08         (TSS(I)/TSS(T)) =  .74767885E 00

   *   GROUP  NO.   3  SPLIT  FROM  GROUP   1  ON  VARIABLE    5 (OCCUPATION  )
       VALUES  OF  PREDICTOR  INCLUDED  ARE    0    7    8    9
          N  =          725              MEAN = -.65252665E 02         GROUP DEVIATION = -.65330982E 02        SUM Y = -.18173520E 07
    WEIGHT SUM =       27851         STD. DEV. =  .10578757E 03                 TSS(I) =  .31168081E 09    SUM Y SQ. =  .43026788E 09
  PCT OF TOTAL =    23.9         WTD. MEAN SQ. =  .11858706E 09         (TSS(I)/TSS(T)) =  .16812086E 00

   *   GROUP  NO.   4  SPLIT  FROM  GROUP   2  ON  VARIABLE   13 (H-W ED DIFF )
       VALUES  OF  PREDICTOR  INCLUDED  ARE    1    2    3    4    6
          N  =         1162              MEAN =  .33637964E 02         GROUP DEVIATION =  .33559646E 02        SUM Y =  .18934810E 07
    WEIGHT SUM =       56290         STD. DEV. =  .12942932E 03                 TSS(I) =  .94296728E 09    SUM Y SQ. =  .10066601E 10
  PCT OF TOTAL =    48.2         WTD. MEAN SQ. =  .63692845E 08         (TSS(I)/TSS(T)) =  .50863724E 00

   *   GROUP  NO.   5  SPLIT  FROM  GROUP   2  ON  VARIABLE   13 (H-W ED DIFF )
       VALUES  OF  PREDICTOR  INCLUDED  ARE    0    5
    ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
          N  =          659              MEAN = -.20529606E 01         GROUP DEVIATION = -.21312776E 01        SUM Y = -.66983999E 05
    WEIGHT SUM =       32628         STD. DEV. =  .11303022E 03                 TSS(I) =  .41684983E 09    SUM Y SQ. =  .41698735E 09
  PCT OF TOTAL =    27.9         WTD. MEAN SQ. =  .13751551E 06         (TSS(I)/TSS(T)) =  .22484910E 00

   *   GROUP  NO.   6  SPLIT  FROM  GROUP   4  ON  VARIABLE    5 (OCCUPATION  )
       VALUES  OF  PREDICTOR  INCLUDED  ARE    1    2    3    5
    ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
          N  =          661              MEAN =  .45165623E 02         GROUP DEVIATION =  .45087306E 02        SUM Y =  .14434030E 07
    WEIGHT SUM =       31958         STD. DEV. =  .14642549E 03                 TSS(I) =  .68519307E 09    SUM Y SQ. =  .75038527E 09
  PCT OF TOTAL =    27.4         WTD. MEAN SQ. =  .65192196E 08         (TSS(I)/TSS(T)) =  .36959364E 00

   *   GROUP  NO.   7  SPLIT  FROM  GROUP   4  ON  VARIABLE    5 (OCCUPATION  )
       VALUES  OF  PREDICTOR  INCLUDED  ARE    4    6
    ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
          N  =          501              MEAN =  .18497369E 02         GROUP DEVIATION =  .18419052E 02        SUM Y =  .45007800E 06
    WEIGHT SUM =       24332         STD. DEV. =  .10094688E 03                 TSS(I) =  .24794970E 09    SUM Y SQ. =  .25627496E 09
  PCT OF TOTAL =    20.8         WTD. MEAN SQ. =  .83252590E 07         (TSS(I)/TSS(T)) =  .13374425E 00

   *   GROUP  NO.   8  SPLIT  FROM  GROUP   3  ON  VARIABLE    5 (OCCUPATION  )
       VALUES  OF  PREDICTOR  INCLUDED  ARE    0    7    9
    ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
          N  =          528              MEAN = -.49548822E 02         GROUP DEVIATION = -.49627139E 02        SUM Y = -.10270480E 07
    WEIGHT SUM =       20728         STD. DEV. =  .93391845E 02                 TSS(I) =  .18079038E 09    SUM Y SQ. =  .23167940E 09
  PCT OF TOTAL =    17.8         WTD. MEAN SQ. =  .50889018E 08         (TSS(I)/TSS(T)) =  .97518460E-01

   *   GROUP  NO.   9  SPLIT  FROM  GROUP   3  ON  VARIABLE    5 (OCCUPATION  )
```

```
      VALUES  OF  PREDICTOR  INCLUDED  ARE    8
  ***  THIS  GROUP  IS  RETAINED  AS  ONE  OF  FINALS.
           N  =          197              MEAN = -.11095100E 03        GROUP DEVIATION = -.11102932E 03        SUM Y = -.79030400E 06
   WEIGHT SUM =         7123          STD. DEV. =  .12477888E 03                 TSS(I) =  .11090346E 09   SUM Y SQ. =  .19858848E 09
 PCT OF TOTAL =      6.1         WTD. MEAN SQ. =  .87685020E 08        (TSS(I)/TSS(T)) =  .59821408E-01




                         * * *    ANALYSIS  OF  VARIANCE  TABLE    * * *

       SOURCE OF              SUM  OF             DEGREE OF          MEAN
       VARIATION              SQUARES              FREEDOM           SQUARE                  F

       T O T A L          .18539092E 10           116723

       BETWEEN            .21222829E 09                4          .53057072E 08        .37722163E 04

       WITHIN             .16416809E 10           116719          .14065225E 05



          RESIDUALS  ARE  NOT  REQUESTED.




                                                   * * *      E    N    D      * * *
 TIME IS NOW 12. 15. 23. 15.

 **** ALL INPUT DATA HAVE BEEN PROCESSED.
       AT LOC 75077
```

PUBLICATIONS AND RESEARCH REPORTS MAKING USE OF THE AID PROGRAM OR REFERRING TO *THE DETECTION OF INTERACTION EFFECTS*, SRC MONOGRAPH NO. 35 OR JASA REPRINT 58, (JUNE, 1963) *PROBLEMS IN THE ANALYSIS OF SURVEY DATA, AND A PROPOSAL*

Books, Reports, and Articles

Anderson, P.R. "An Application of AID-2 to Canadian Discretionary Saving." Unpublished manuscript, Bank of Canada, September 20, 1966.

Caplan, N., G. Suttles, J. Deshaies, H. Mattick, *The Nature of Treatment Intervention Among Delinquent Gang Youth and Factors Affecting Its Outcome*, Presented at the American Sociological Association Convention, Los Angeles, 1963.

Caplan, N., R. Lippitt, M. Gold, H. Mattick, G. Suttles, and D. Deshaies, *The Outcome of the Chicago Youth Development Project*, (A four hour presentation) Presented at the Boys Clubs of America Convention, New York, 1966.

Caplan, N., *The Success and Failures of Street Gang Work*, Presented at the Fourth Annual Conference of Community Resources Workers, University of Chicago, 1966.

Deshaies, D., and N. Caplan, *Social Intervention and a View of Man*, Presented at the International Congress of Criminology, Montreal.

Gensemer, Bruce L.,Jane A. Lean, and William B. Neenan, "Awareness of Marginal Income Tax Rates Among High-Income Taxpayers," *National Tax Journal*, XVIII (September, 1965), 258-267.

Katona, George, *Private Pensions and Individual Saving*, (SRC Monograph No. 40.) Ann Arbor: University of Michigan, Institute for Social Research, 1965, Chapter 13, pp. 81-89.

Katona, George, Eva Mueller, Jay Schmiedeskamp, and John A. Sonquist, *Survey of Consumer Finances, 1965*. (SRC Monograph No. 42.) Ann Arbor: University of Michigan, Institute for Social Research, 1966, pp. 14, 111-112.

Morgan, James N., "The Achievement Motive and Economic Behavior," *Economic Development and Cultural Change*, XII, No. 3 (April, 1964), 243-267.

______, "Housing and Ability to Pay," *Econometrica*, XXXIII, No. 2 (April, 1965), 289-306.

______, "Time, Work, and Welfare." In *Patterns of Market Behavior*, Michael J. Brennan, ed. Providence: Brown University Press, 1965.

Morgan, James N., Robin Barlow and Harvey Brazer. *The Economic Behavior of the Affluent*. Washington: Brookings Institution, 1966.

Morgan, James N., Ismail Sirageldin and Nancy Baerwaldt. *The Productive Americans*. (SRC Monograph No. 43.) Ann Arbor: University of Michigan, Institute for Social Research, 1966.

Ross, J., and S. Bang., "Predicting the Adoption of Family Planning," *Studies in Family Planning*, No. 9 (January, 1966).

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

Snowbarger, Marvin and D.B. Suits., "Consumer Expenditures for Durable Goods." Paper delivered at the National Bureau of Economic Research Conference in Madison, Wisconsin, June, 1965. To be published by the National Bureau of Economic Research in a forthcoming book on investment behavior.

Citations

Emmett, B.P., "The Design of Investigations into the Effects of Radio and Television Programmes and Other Mass Communications," Journal of the Royal Statistical Society, Series A, CXXIX, Part 1, 1966, 26-59.

Selvin, H.C., and A. Stuart. "Data Dredging Procedures in Survey Analysis," The American Statistician, XX, No. 3 (June, 1966), 20-23.

Strümpel, Burhard. Steuermoral und Steuerwiderstand der deutschen Selbständigen, (Forschungsstelle für empirische Sozialökonomik, No. 1682.) Köln: West-deutscher Verlag, 1966. p. 60.

Other Versions of the Program

Aptakin, Peter., "Automatic Interaction Detector." New York: Service Bureau Corporation, Computing Sciences Division, 635 Madison Ave., 1965.

Campbell, Robert H., "Modifications to the Automatic Interaction Detector" COMCOM/Simulation Memo No. 26, March 1, 1965. Program operational on the MIT-FMS system, Massachusetts Institute of Technology, Harvard Computation Center.

Kay, Kevin., "Automatic Interaction Detector: 'AID' Translated into CDC 3600 Fortran and Compass," Technical Report 46. East Lansing: Michigan State University, Computer Institute for Social Science Research, April 11, 1966.