In marketing research, AID (Automatic Interactive Detector) analysis developed by Survey Research Center of Michigan University is widely used to identify clusters of observations having similar attributes under interaction. We have developed an AID algorithm together with SAS interface (AID/SAS).

PROC AID starts:

```
PROC AID; MODEL Y = X1 X2 X3; RUN;
```

Y: dependent variable
X1, X2: independent variables

Written in FORTRAN IV with up-to-date function, AID/SAS is compact and portable. Currently it runs on an IBM compatible machine Facom M160 in our institute.

Introduction

AID (Automatic Interactive Detector) is a data analysis tool for sample survey. Data analysis consists of two important steps: the first is searching the best model with theory and examination of data in the process and the second one is assessment of the best model by statistical inference. AID is powerful tool in the first step.

AID finds a relation between quantitative dependent variables and qualitative independent ones, and the relation is described as a binary tree structure of factors. In the case of income and its factors such as age, sex, hometown, job status and others that were classified prior to analysis are arranged in order of dependency to target variable.

In AID procedure, samples are divided into two categorical groups by the most dependable variable. Every observation is a member of exactly one of these subgroups. They are chosen so that at each step in the procedure, the two new means account for more of the total sum of squares than the means of any other pair of subgroups.

AID analysis has been developed by Survey Research Center of Michigan University and we developed an AID program as a part of our survey data analysis package (SORIS) in 1977 for the purpose of various social survey projects. Later, SAS system is installed in our institute in October 1982 for economic analysis, statistical database and other purpose. Then users of SAS in the institute felt a strong necessity of AID procedure that works with SAS database. So we developed AID/SAS interface in PL/I, by which the AID program is linked to SAS. This SAS expansion procedure is named PROC AID and its command sequence is similar to PROC GLM or PROC REG.

Mathematical description of AID procedure is as follows.

1. Choose an unsplit group L (the "parent group") which has the largest sum of squares. In the first step, group L has all observations in the sample survey, and it is split into two subgroups by the BSS indicator (between-group sum of squares).

   \[ SS_L = \sum_{a=1}^{N_L} (Y_{aL} - \bar{Y}_L)^2 \]
   
   \[ = BSS + WSS \]

   \[ BSS = \frac{N_L}{N_L^L} \sum_{i=1}^{L1} (Y_{iL1} - \bar{Y}_{L1})^2 + \frac{N_L}{N_L^L} \sum_{i=1}^{L2} (Y_{iL2} - \bar{Y}_{L2})^2 - N_L \bar{Y}_L^2 \]

   \[ WSS = \sum_{a=1}^{N_{L1}} (Y_{aL1} - \bar{Y}_{L1})^2 + \sum_{a=1}^{N_{L2}} (Y_{aL2} - \bar{Y}_{L2})^2 \]

   \[ SS : \text{Sum of Squares} \]

   \[ BSS : \text{Between-group Sum of Squares} \]

   \[ WSS : \text{Within-group Sum of Squares} \]

   \[ \bar{Y} : \text{Mean of dependent variables} \]

   \[ N : \text{Observations in group} \]

2. Select an independent variable and its split pattern of the highest BSS, and if BSS \( = pBSS \) (the parameter pe is an eligibility criterion) or observations of subgroups are less than the minimum limit, then group L is deemed a final group and is not splitted.

3. Calculate SS of each subgroup, and return to the step (1).

Theory of AID procedure

The procedure AID is implemented in PL/I and written in FORTRAN IV with up-to-date function. Currently it runs on an IBM compatible machine Facom M160 in our institute.
Program structure of PROC AID

PROC AID is composed of three subsystems as follows.

1. Command parser subprogram
   This subprogram is written in assembler macro, and its function is parsing SAS command similar to PROC REG.

2. AID/SAS interface subprogram
   This subprogram is written in PL/I. It reads SAS data set and creates temporary files which include selected variables for AID subprogram for the maximum category size definition and printing labels.

3. AID subprogram
   This subprogram is written in FORTRAN IV and executes AID analysis.
   This algorithm was developed in our institute, and its major benefits are quickness and compact core size thanks to bit processing on binary tree handling.

Figure 1 shows general flowchart of this algorithm.

PROC AID STATEMENT
PROC AID;
MODEL Y = X1 X2 X3 X4 / TERM = 0.001 MINGRP = 30; RUN;

The constraint of variables
Maximum number of independent variables that could be used in analysis is 30, and each variable must have categories under 20.

Y : Dependent variable
   Only numeric variable is permitted.

X1-X4: Independent variables
   Only character variables are permitted.

TERM: Terminate condition
   The parameter when BSS/SS is less than this group is deemed final group.
   The default value of this parameter is 0.001.

MINGRP: Minimum group size
   When divided subgroup's size is less than MINGRP this group is deemed final group.
   The default value of this parameter is 30.
Example

AID/SAS puts out two results. The first output is the proof list of every examination, and the second one is the final structure of factors, which is printed like a binary tree. Figure 2 is an example of this binary tree. This case is assumed to be an analysis of household yearly income in Japan. Table 1 shows variables and their categories used in the analysis.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types of Tenure of Dwelling</td>
<td>Occupation</td>
</tr>
<tr>
<td>Owend houses</td>
<td>Regular worker</td>
</tr>
<tr>
<td>Rented houses(Private)</td>
<td>Temporary worker</td>
</tr>
<tr>
<td>Rented houses(Public)</td>
<td>Private office-worker</td>
</tr>
<tr>
<td>Dwelling provided by employer</td>
<td>Official office-worker</td>
</tr>
<tr>
<td>Rented rooms</td>
<td>Merchants, artisans</td>
</tr>
<tr>
<td></td>
<td>Private administrators</td>
</tr>
<tr>
<td></td>
<td>Corporative administrators</td>
</tr>
<tr>
<td></td>
<td>Professional services</td>
</tr>
<tr>
<td></td>
<td>Others</td>
</tr>
<tr>
<td></td>
<td>No occupation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Age group</th>
<th>Size of enterprise</th>
<th>Household members</th>
<th>Earners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-24 years</td>
<td>1 - 29 persons</td>
<td>1 person</td>
<td>1 person</td>
<td></td>
</tr>
<tr>
<td>25-29 years</td>
<td>30 - 499 persons</td>
<td>2 persons</td>
<td>2 persons</td>
<td></td>
</tr>
<tr>
<td>30-34 years</td>
<td>500 - 999 persons</td>
<td>3 persons</td>
<td>3 persons</td>
<td></td>
</tr>
<tr>
<td>35-39 years</td>
<td>1000 persons over</td>
<td>4 persons</td>
<td>4 persons over</td>
<td></td>
</tr>
<tr>
<td>40-44 years</td>
<td>Government</td>
<td>5 persons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45-49 years</td>
<td></td>
<td>6 persons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50-54 years</td>
<td></td>
<td>7 persons over</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55-59 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60-64 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64 years over</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable Yearly income (unit is 1,000 yen)
Figure 2 Example of an AID result
Other SAS expansion procedure in MKI

We developed other SAS expansion procedures as tools of studies, such as economic analysis, social survey, area studies, and other scientific research. We report the typical procedure of them, as examples of a SAS user's activity in Japan.

(1) SAS/KANJI

SAS/KANJI is a tool which adds power to SAS system users who wish to use the Japanese language. "KANJI" is a set of Chinese symbolic characters used in the Japanese language, and capability of a software in treating KANJI is very important in Japanese software business. SAS/KANJI has two Japanese language dictionaries. One is SAS keyword dictionary and another is user dictionary. The keyword dictionary stores statistical and other terms used in SAS, as well as their corresponding Japanese terms. And in user dictionary users can register variable names and titles which they use in SAS, both in English and in Japanese.

SAS/KANJI checks a original SAS output and translates it into the Japanese language referring to the dictionaries. So in debugging users can use SAS in English, and if necessary they can easily put out messages in Japanese in well trimmed form.

(2) Data dictionary system on SAS data set

SAS data set is powerful database for statistical analysis, but SAS data set has only a few informations about variables. We developed selfdescriptive database with data dictionary on SAS data set. This database has three dictionaries. They are dataset dictionary, attribute dictionary, and domain dictionary, and the database has interactive query functions with menu screen.

This system is written in SAS language with two SAS expansion procedures, one is SAS/IPF interface, (IPF is a screen control utility of FACOM machine.) and another is a macro procedure that can use parameters and global variables. (This macro procedure's function is a subset of the macro procedure in SAS/82)

This system has been developed for experimental construction of large statistical database in Economic Planning Agency of Japanese government.

References


Research Institute of MKI 1978: SORIS USER'S GUIDE