AUTOMATIC INTERACTION DETECTOR-VERSION 4
(AID)-4
REFERENCE MANUAL ADDENDUM 1

By
C. Deene Gott
Janos B. Koplyay

COMPUTATIONAL SCIENCES DIVISION
Brooks Air Force Base, Texas 78235

July 1977
Final Report for Period October 1973 – April 1977

Approved for public release; distribution unlimited.
NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Computational Sciences Division, under project 6323, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235. Dr. Janos B. Koplyay, Statistical and Computer Technology Branch, was the project monitor.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

ROBERT A. BOTTENBERG, Chief
Computational Sciences Division

DAN D. FULGHAM, Colonel, USAF
Commander
**REPORT DOCUMENTATION PAGE**

**1. REPORT NUMBER**
AFHRL-TR-77-30

**2. GOVT ACCESSION NO.**

**3. RECIPIENT'S CATALOG NUMBER**

**4. TITLE (and Subtitle)**

**5. TYPE OF REPORT & PERIOD COVERED**
Final report
October 1973 - April 1977

**6. PERFORMING ORG. REPORT NUMBER**

**7. AUTHOR(S)**
C. Deene Gott
Janos B. Koplyay

**8. CONTRACT OR GRANT NUMBER(S)**

**9. PERFORMING ORGANIZATION NAME AND ADDRESS**
Computational Sciences Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

**10. PROGRAM ELEMENT, PROJECT, TASK
AREA & WORK UNIT NUMBERS**
62203F 1845
63230506

**11. CONTROLLING OFFICE NAME AND ADDRESS**
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

**12. REPORT DATE**
July 1977

**13. NUMBER OF PAGES**
16

**14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)**

**15. SECURITY CLASS. (of this report)**
Unclassified

**16. DISTRIBUTION STATEMENT (of this Report)**
Approved for public release; distribution unlimited.

**17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)**

**18. SUPPLEMENTARY NOTES**

**19. KEY WORDS (Continue on reverse side if necessary and identify by block number)**
interaction detection in prediction systems
regression model building
model building
multiple linear regression
pattern identification
prediction system

**20. ABSTRACT (Continue on reverse side if necessary and identify by block number)**
PREFACE

This research was completed under project 6323, Personnel Data Analysis; task 632305, Development of Analytic Methodology for Air Force Personnel Research Data. Special credit is given to Mr. Larry K. Whitehead who made the computer program conversion from the IBM 7040 to the UNIVAC 1108, and to Mrs. Doris E. Black for editing this report for technical accuracy.
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>I.</th>
<th>Introduction</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>II.</td>
<td>Change of Limitations of the AID-4 Computer Algorithm Adapted to the UNIVAC 1108 (Reference AFHRL-TR-73-17)</td>
<td>5</td>
</tr>
<tr>
<td>III.</td>
<td>AID-4 Random Selection Process Correction and Technical Description</td>
<td>6</td>
</tr>
<tr>
<td>IV.</td>
<td>Supplemental Information for Running AID-4 on the UNIVAC 1108</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Reference</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>13</td>
</tr>
</tbody>
</table>
AUTOMATIC INTERACTION DETECTOR-VERSION 4 (AID)-4
REFERENCE MANUAL ADDENDUM I

I. INTRODUCTION

AID-4, an acronym for automatic interaction detector-version 4, is a computer-based regression model building algorithm which was originally made operational on an IBM 7040 computer system at the Computational Sciences Division, Air Force Human Resources Laboratory (AFHRL), Lackland Air Force Base, Texas. The background, basic algorithm, and technical details required for use of AID-4 are described by Koplyay, Gott, and Elton (1973).

Subsequent acquisition of a UNIVAC 1108 computer system by the Computational Sciences Division, allowed for the expansion of AID-4 capabilities. This report serves as an addendum to the original reference manual mentioned previously. This addendum contains three sections which describe changes to the IBM 7040 version of the program, technical notes on certain aspects of the program not mentioned before, and information to assist users on the UNIVAC 1108 system. Specifications provided should also prove useful for implementing AID-4 on other computer systems.

II. CHANGE OF LIMITATIONS OF THE AID-4 COMPUTER ALGORITHM ADAPTED TO THE UNIVAC 1108 (Reference AFHRL-TR-73-17)

1. All references in AFHRL-TR-73-17 to the upper range of the recode categories should be changed from (0-39) to (0-49); i.e., the total number of categories per variable changed from 40 to 50. The following pages are affected:
   - Page 16 Last 3 lines
   - Page 19 4th line from bottom of page
   - Page 22 Middle of page (3d and 4th line from para III)
   - Page 23 4th and 5th line from bottom of page
   - Page 28 Limitation No. 8
   - Page 29 Error No. 30
   - Page 30 In the formulas in Adoption Note No. 8 change In 40 to In 50 in the denominators and the results; i.e., replace 6.58 = 6, 5.82 = 5, and 11.08 = 11 with 6.20 = 6, 5.49 = 5, and 10.45 = 10, respectively.

2. Change all references in AFHRL-TR-73-17 to the maximum number of input variables from 80 to 300. The following pages are affected:
   - Page 14 Under Card Columns 27-29 and 30-32, change 83 to 303
   - Page 28 Under Limitations change 83 to 303 (para 1) and 80 to 300 (para 2)
   - Page 29 Error No. 38: Change 83 to 303

3. The references in AFHRL-TR-73-17 to the total number of categories of a given problem require the following changes:
   - Page 28 Limitations No. 3, change 700 to 2500
   - Page 29 Errors No. 26, 27 and 31, change 700 to 2500

4. There are numerous changes to the FORTRAN listing appearing in AFHRL-TR-73-17 which were necessary to implement AID-4 on the UNIVAC 1108. These changes do not affect the user and will
III. AID-4 RANDOM SELECTION PROCESS CORRECTION AND TECHNICAL DESCRIPTION

1. The program listing of AID-4 given in AFHRL-TR-73-17 has an error on line AID06480 (page 51 of AFHRL-TR-73-17). Line AID064480 reads

   PROI = FLOAT(NSAB—NSEL)/FLOAT(NC—NIN+1).

   The above statement should be changed to

   PROI = FLOAT(NSAB—NSEL—NOUT)/FLOAT(NC—NIN+1).

2. The random selection process used in AID-4 when a user specifies IRUN = 2 or 3 (on the title card) employs the same concepts as the following “balls from an urn” analogy, where \( N, N_A, \) and \( N_B \) in the urn problem are analogous to

   \[\begin{align*}
   N & \quad \text{number of cases in the original input file}, \\
   N_A & \quad \text{number of cases requested by the user to be randomly selected for assignment to sample A}, \\
   N_B & \quad \text{number of cases requested by the user to be randomly selected for assignment to sample B}.
   \end{align*}\]

   Select balls one at a time without replacement from an urn which initially contains \( N \) balls of which \( N_A \) are labeled A, \( N_B \) are labeled B and the remainder, if any, are unlabeled. If \( i \) balls have been selected where \( n_A \) were labeled A, \( n_B \) were labeled B, and \( i = n_A + n_B \), then the probability that the next ball selected is labeled A or B is

   \[\begin{align*}
   p(A \text{ or } B) & = \frac{n_A + n_B - n_A - n_B}{N - i}.
   \end{align*}\]

   The probability that the next ball selected is labeled A is

   \[\begin{align*}
   p(A) & = \frac{n_A - n_A}{N - i}.
   \end{align*}\]

   AID-4 makes its assignments to sample A and sample B by generating a uniform random deviate \( r \) in the interval \((0,1)\) and by applying it to the following logic sequence:

   (a) Assign case to sample A if \( r \leq p(A) \) and case does not contain out-of-range data; therefore, if the case is assigned to sample A, \( n_A = n_A + 1 \).

   (b) Assign case to sample B if \( p(A) < r \leq p(A \text{ or } B) \) and case does not contain out-of-range data; therefore, if the case is assigned to sample B, \( n_B = n_B + 1 \).

   (c) Case is not assigned to a sample if \( r > p(A \text{ or } B) \).

   NOTE: If the case is not assigned to a sample because it contains out-of-range data, the following computation is performed to approximate the number of cases in sample A and sample B.

   \[\begin{align*}
   n_A & = n_A + N_A/(N_A + N_B) \\
   n_B & = n_B + N_B/(N_A + N_B).
   \end{align*}\]
IV. SUPPLEMENTAL INFORMATION FOR RUNNING AID-4 ON THE UNIVAC 1108

1. Runstreams for the UNIVAC 1108.

The following card sequence is required to use the AID-4 program as it is operational on the AFHRL UNIVAC 1108 computer. (Other computer systems will require different systems-runstreams.) The files involved in the run may be either tape of mass storage. The cataloging options (as required by tape or mass storage) will be supplied by the user.

Note: Several types of runs may be performed by AID-4 depending upon the value of “IRUN” in column 50 of the title/parameter card. These runs may be briefly described as follows:

IRUN=0 Use every case in the original input file for a normal AID-4 run; no forced splitting.

IRUN=1 Select a random sample A from the original input file, and use only the cases that belong to A for a normal AID-4 run.

IRUN=2 Select a random sample A and a random sample B from the original input file, and use only those cases that belong to A for a normal AID-4 run. Then force those cases that belong to B to make the same splits as taken by A; i.e., single cross-validation.

IRUN=3 Select a random sample A and a random sample B from the original input file, and use only the cases that belong to A for a normal AID-4 run. Then force those cases that belong to B to make the same splits as taken by A. Then use only those cases that belong to B for a normal AID-4 run, and force those cases that belong to A to make the same split as taken by B; i.e., double cross-validation.

IRUN=4 Given sample A and sample B (no random selection by the program), use only the cases that belong to A for a normal AID-4 run. Then force those cases that belong to B to make the same splits as taken by A; i.e., single cross-validation. Note that double cross-validation can be accomplished by submitting a second job with samples A and B switched.

1.1 IRUN=0, 1, 2, or 3.

Order Type

1. @RUN RUN-ID, Job,Section
2. @ASG,A DATA
3. USE 10,DATA
4. @ASG,T TEMP-1
5. USE 11,TEMP-1
6. @ASG,T TEMP-2
7. USE 12,TEMP-2
8. @ASG,T TEMP-3
9. USE 13,TEMP-3
10. @ASG,T TEMP-4
11. USE 14,TEMP-4
12. @ASG,T TEMP-5
13. USE 15,TEMP-5
14. @ASG,T TEMP-6
15. USE 18,TEMP-6
16. @XQT T*T.AID-4
17. Title (Parameters) Card
18. Data Format Card(s)
19. Description Card
20. Predictor Card(s): at least one per predictor
21. Criterion Card
22. Data (if on cards, otherwise on Ul)
23. End-of-job Card
24. @FIN

1.2 IRUN=4.

Order Type
1. @RUN RUN-ID:Job,Section
2. @ASG,A SAMP-A ) Sample A
3. @USE 10, SAMP-A ) Data file
4. @ASG,T TEMP-1,... ) Scratch file
5. @USE 11, TEMP-1
6. @ASG,T TEMP-2,... ) Scratch file
7. @USE 12, TEMP-2
8. @ASG,T TEMP-3,... ) Scratch file
9. @USE 13, TEMP-3
10. @ASG,T TEMP-4,... ) Scratch file
11. @USE 14, TEMP-4
12. @ASG,T TEMP-5,... ) TREE PLOT
13. @USE 15, TEMP-5 ) Information
14. @ASG,A SAMP-B ) Sample B
15. @USE 18, SAMP-B ) Data file
16. @XQT T*T:AID-4
17. Title (Parameters) Card
18. Data Format Card(s)
19. Description Card
20. Predictor Card(s): at least one per predictor
21. @FIN
22. End-of-job Card
23. @FIN

2. File Flowcharts.

Note: All unmentioned files are "scratch" (temporary) files used in either "split" or "forced split."

2.1 IRUN=0 and IRUN=1.
IRUN = 0 (Use all cases)

IRUN = 1 (Select random sample A)

U10 data in

Data Phase

U13 Total S

SPLIT

U15 TEMP-5 S

U18 Resid 0

TREE

U10 data in

Data Phase

U13 SAMP-A S

SPLIT

U15 TEMP-5 S

U18 Resid 0

TREE
2.2 IRUN=2. (Single cross-validation)
2.3 IRUN=3. (Double cross-validation)
2.4 IRUN=4. (Single cross-validation given samples A and B)

REFERENCE

BIBLIOGRAPHY


