Historic, archived document

Do not assume content reflects current scientific knowledge, policies, or practices.

SCREEN(F): A FORTRAN77 Program to Identify Predictors of Dichotomous Dependent Variables

David A. Hamilton, Jr.¹

ABSTRACT

The computer algorithm described by Hamilton and Wendt (1975), which screens potential predictor variables of a dichotomous dependent variable, has been rewritten in FORTRAN77. This note describes changes introduced in the FORTRAN77 version of the program, updates the original user's guide, and describes program availability and hardware requirements for running the program.

KEYWORDS: data screening, logistic regression, modeling, computer algorithm, statistical method

Analytical procedures that effectively screen potential relationships when the dependent variable is continuous may not be appropriate or as effective when the dependent variable is dichotomous. Hamilton and Wendt (1975) described a computer algorithm specifically designed to screen potential relationships between a set of independent variables and a dichotomous dependent variable. The theory upon which this algorithm is based was described by Sterling and others (1969).

The algorithm was originally programmed as two PL/1 procedures by Malcolm Glessner in PL/1 48-character set. Hamilton and Wendt (1975) modified the procedures, discussed properties and potential uses of the algorithm, and prepared documentation and a user's guide for the procedures.

Because the procedures were written in PL/1, their use has been limited to those individuals who have access to a PL/1 compiler. Thus some potential users have been unable to use the algorithm. When I attempted to recompile the procedures on the two latest releases of the PL/1

compiler, I encountered what appear to be compiler bugs. The procedures apparently compile successfully but will not execute a test problem correctly.

These problems provided the motivation to rewrite the procedures in FORTRAN77. The new program, SCREEN(F), eliminates the problems encountered when the old procedures were compiled under the latest releases of the PL/1 compiler and the need to have access to a PL/1 compiler. This latter fact greatly increases the number of potential users who will be able to run the program on their computer system.

UPDATES OR MODIFICATIONS TO THE ALGORITHM

There are only a few minor differences between SCREEN(F) and the program documented by Hamilton and Wendt (1975). The original program permitted unlimited numbers of independent variables. When SCREEN(F) is used, the analysis is limited to 50 potential independent variables. If this limitation creates problems for a user, additional independent variables may be evaluated by making multiple runs of the program or by increasing the dimension of the variables NCTG, VN, VNS, INDEX, IPVAR, ISUB, INA, and DUMMY in the source code of the main program and subroutine GRAPH.

The PL/1 version of this program was written as two PL/1 procedures. The first procedure (SEARCH) consisted of the screening algorithm. The second procedure (GRAPH) printed the results of the screening in a "decision tree" format. The two procedures were either run independently or linked together with job control language (JCL).

SCREEN(F) is written as a single program. The main program contains what was in the PL/1 procedure SEARCH. The GRAPH procedure is now a subroutine

¹Research forester located at Intermountain Station's Forestry Sciences Laboratory, Moscow, ID.

VARIABLE	NODE									
	1	2	3	4	5	6	7	8	9	10
DATASET NUMB	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FOREST CODE	0.00	0.00	0.00	0.00	2.68	3.29	6.23	0.00	0.00	0.00
REGEN.SYSTEM	1.65	2.95	0.00	0.00	6.01	0.00	6.73	0.00	0.00	0.00
HABITAT TYPE	0.00	0.00	0.00	0.00	10.69	0.00	0.00	0.00	0.00	0.00
ASPECT	2.18	0.00	3.02	0.00	4.63	3.32	0.00	0.00	3.07	0.00
SLOPE PERCNT	1.74	0.00	0.00	0.00	2.81	0.00	2.29	0.00	0.00	1.61
TOPO.POSIT'N	0.00	0.00	3.53	0.00	4.81	0.00	0.00	0.00	0.00	0.00
SITE PREP.	0.00	0.00	3.17	0.00	0.00	3.05	2.49	0.00	2.80	2.77
SEED SOURCE	0.00	0.00	0.00	0.00	0.00	1.98	0.00	0.00	2.21	0.00
ANGLE 2 SEED	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R.BASAL AREA	0.00	3.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ELEVATION	1.96	0.00	0.00	0.00	5.26	0.00	3.72	3.62	2.08	2.10
TIME	6.28	0.00	0.00	0.00	8.46	0.00	3.07	0.00	2.34	2.06
REGEN. MODE	0.00	0.00	0.00	0.00	5.54	2.74	4.82	0.00	0.00	0.00
STOCKING	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DEFOL.BEFORE	1.38	0.00	0.00	0.00	2.54	0.00	3.49	2.85	0.00	0.00
DEFOL.AFTER	3.81	0.00	3.45	1.89	5.28	1.42	0.00	0.00	1.85	0.00

Figure 1—Example of SCREEN output showing a table of statistics used in variable selection process.

called from the main program. This eliminates the need to use JCL to link the two components of the program. Further, the duplication of control input data required in the PL/1 version of the program has been eliminated. The only input that must be supplied to the GRAPH subroutine by the user is the title that is to be used as a page header for the "decision tree" diagram. The remaining control input data required for GRAPH is now passed as arguments in the subroutine CALL statement. Details of the control input data required to run SCREEN(F) are provided in the "Updated User's Guide" section of this note.

SCREEN(F) prints one new type of output that I have found useful in interpreting the results of the algorithm. A table is printed that contains values of the test statistic used to determine which independent variable is selected at each node in each of the first three levels of the "decision tree." The statistic is an entropy statistic that is distributed approximately as chi-square. The value printed in the table has been adjusted for differences in degrees of freedom. The statistic and the adjustment procedure are described in detail in Sterling and others (1969). If the statistic is not significant at the user-specified significance level, a value of zero is printed in the table.

Although the statistics are dependent on which variables have been selected at previous nodes, at any specified node the statistics for each of the independent variables are independent. Thus they provide the user with information about which independent variable would have been selected at each node if the originally chosen variable had been eliminated from the analysis. This information is particularly useful when two or more of the variables being considered explain approximately equal amounts of the remaining variation. Previously, the only method available to evaluate what would happen when a variable was removed from the analysis was to rerun the analysis with that variable deleted. Although complete analysis of

a set of data may still require that selected variables be removed from the analysis and the program rerun, the new output provides information that is very useful in structuring successive runs of SCREEN and thus frequently reduces the number of runs that are required to complete an analysis.

Figure 1 presents an example of this new output. Each of the nodes in the first three levels of the "decision tree" are labeled with a number enclosed in brackets (fig. 2). The table reproduced in figure 1 lists the statistics used to select the most significant independent variable at each node. The statistics for node 1 are in the first column of numbers in figure 1. The largest number in this column is the value 6.28 for the variable "TIME", the variable selected at node 1. The second largest value is 3.81 for the variable "DEFOL.AFTER". Thus, if "TIME" were removed from the analysis, it would be replaced at this node by "DEFOL.AFTER". Similar interpretations should be given to the values in the other columns of the table. Note that there is only one nonzero value in the column for node 4. This implies that there are no other significant independent variables at this node and if the selected variable, "DEFOL.AFTER", is removed from the analysis, the node will be dropped from the "decision tree."

UPDATED USER'S GUIDE

Most of the instructions for using the screening algorithm remain unchanged from what was presented by Hamilton and Wendt (1975). In this update of the user's guide, only those instructions that are different will be discussed.

The input data set is now read on logical unit 2. The requirements for the format of the observations making up the input data set have not been changed. Each independent variable can occur at one of eight levels, coded 0 to 7. The dependent variable is coded as either 0 or 1. Finally,


```

TIME
[ 1] 1,2,3YRS
      4+5YEARS
      .....
      .    20.
      .    15.
      .....
      6+7YEARS
      .....
      .    30.
      .    1.
      .....

      R.BASAL AREA
      [ 2] NONE
            10SQ.FT.
            20SQ.FT.
            .....
      8+9YEARS .    33.
      ..... .    35.
      .    61. ....
      .    41. 30-40
      ..... 50-70
            80-100
            110-180
            190PLUS
            .....
            .    28.
            .    6.
            .....

      SITE PREP.
            NONE
            MECH
            .....
      TOPO.POSIT'N .    18.
      [ 3] BOTTOM .    45.
            LOWER .....
            MIDSLOPE BURN
            ..... ROADCUT
            .    35. ROADBED
            .    51. ROADFILL
            .....
            .    17.
            .    6.
            .....

```

Figure 2—Sample of SCREEN "decision tree" output showing nodes in first three levels identified by numbers enclosed in square brackets.

it is no longer required that the logical record length of the input data set be equal to the number of variables in each observation.

Data for program control are input on logical unit 5 on six types of records in the main program. These records correspond to the six card types discussed by Hamilton and Wendt (1975). Record types 2, 3, 5, and 6 are identical to card types 2, 3, 5, and 6. On record type 1, the variables LRECL and INPT have been deleted. The two remaining

variables on this record are NVAR (number of variables) and NLR (number of observations). The format of this new record type 1 is 2I6. As in the PL/1 version of the program, if the product of NVAR and NLR is greater than 32,676, the input data set must be read and processed one observation at a time from logical unit 2 for each node in the "decision tree." This is necessary to avoid exceeding the dimensions specified for the input array. If the product is less than 32,676, the input data set is read only once and stored in an array. A new feature of SCREEN(F) is that the appropriate method of data input is now determined by the program. SCREEN runs much more efficiently when the data are read only once into an array. Thus, although the user can no longer control the method of data input with the variable INPT, some care should be taken in determining the size of the data set to be analyzed. If a data set is larger than the limit that permits the data to be read only once, it may be advantageous to select a subset of the data for analysis that is smaller than the limit.

Record type 4 is similar to card type 4 with the variable NTOTAL deleted. The remaining variable names on the record have been converted to integer variable names, and the format for the record is 3I6. On several of the record types, some variable names have been changed in the source code to correspond to FORTRAN naming conventions. Variable type (for example, real or integer) for these variables remains as it was documented by Hamilton and Wendt (1975).

Because GRAPH is now a subroutine, much of what was previously read on input control cards is now passed in the subroutine argument list. All the control data that remain to be read in GRAPH is the variable TITLE, which is now read from record type 7. TITLE is declared CHARACTER*80. Thus, the format for record type 7 is A80.

Results of the data screening are passed from the main program to the GRAPH subroutine on a scratch file. The scratch file is written on logical unit 8.

To make multiple runs of the algorithm using the same input data set, it is only necessary to prepare record types 3 to 7 for the second and each succeeding run. There is an additional advantage to making multiple runs with SCREEN(F) that did not exist in the PL/1 version. If the data set is small enough to be stored in an array, it will be read only for the first run. This reduces the time the program spends doing input/output operations and thus reduces elapsed execution time.

HARDWARE REQUIREMENTS AND PROGRAM AVAILABILITY

SCREEN(F) has been run on the IBM 3090² at Washington State University, on a Data General MV 15000, and on IBM PC's. On the IBM 3090, the program was compiled using version 4.0 of the IBM FORTRAN VS compiler. On the Data General MV 15000, the program has been run under AOS/VS using the F77 FORTRAN compiler. On PC's, the program has been successfully compiled with

²The use of trade or firm names in this publication is for reader information and does not imply endorsement by the U.S. Department of Agriculture of any product or service.

Lahey FORTRAN, IBM Professional FORTRAN, Microsoft FORTRAN, and Ryan McFarland FORTRAN.

The executable code on the PC requires 260 K bytes. Although the program has only been run on an IBM PC AT and on an IBM PS/2 model 60, the program should run on either an IBM PC or XT or compatibles. Output files require a logical record length of 132. Thus, when run on a PC, the printer used must either be wide carriage or be capable of compressed printing.

Source code for the program is available on request from:

David Hamilton
Forestry Sciences Laboratory
1221 S. Main Street
Moscow, ID 83843

Requests should include either a 5 1/4-inch or a 3 1/2-inch floppy disk. Files provided on the disk will include source

code for SCREEN(F), an example test data set of 1,000 observations (each observation made up of 17 variables), control input data needed to run the example, and output generated by the example. Executable code for a PC operating under PC/DOS or MS/DOS is also available on request.

REFERENCES

- Hamilton, David A., Jr.; Wendt, Donna L. R. 1975. SCREEN: a computer program to identify predictors of dichotomous dependent variables. Gen. Tech. Rep. INT-22. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 20 p.
- Sterling, Theodor D.; Binks, Randall G.; Haberman, Shelby; Pollack, Seymour V. 1969. Robot data screening—a ubiquitous automatic search technique. In: Milton, Roy C.; Nelder, John A., eds. Statistical computation. New York: Academic Press: 319-333.

INTERMOUNTAIN RESEARCH STATION

The Intermountain Research Station provides scientific knowledge and technology to improve management, protection, and use of the forests and rangelands of the Intermountain West. Research is designed to meet the needs of National Forest managers, Federal and State agencies, industry, academic institutions, public and private organizations, and individuals. Results of research are made available through publications, symposia, workshops, training sessions, and personal contacts.

The Intermountain Research Station territory includes Montana, Idaho, Utah, Nevada, and western Wyoming. Eighty-five percent of the lands in the Station area, about 231 million acres, are classified as forest or rangeland. They include grasslands, deserts, shrublands, alpine areas, and forests. They provide fiber for forest industries, minerals and fossil fuels for energy and industrial development, water for domestic and industrial consumption, forage for livestock and wildlife, and recreation opportunities for millions of visitors.

Several Station units conduct research in additional western States, or have missions that are national or international in scope.

Station laboratories are located in:

Boise, Idaho

Bozeman, Montana (in cooperation with Montana State University)

Logan, Utah (in cooperation with Utah State University)

Missoula, Montana (in cooperation with the University of Montana)

Moscow, Idaho (in cooperation with the University of Idaho)

Ogden, Utah

Provo, Utah (in cooperation with Brigham Young University)

Reno, Nevada (in cooperation with the University of Nevada)

USDA policy prohibits discrimination because of race, color, national origin, sex, age, religion, or handicapping condition. Any person who believes he or she has been discriminated against in any USDA-related activity should immediately contact the Secretary of Agriculture, Washington, DC 20250.

Intermountain Research Station
324 25th Street
Ogden, UT 84401