PROGRAM ABSTRACTS/ALGORITHMS

APL programs for interactive data analysis: Correlation and data entry

SELBY EVANS, FRED H. GAGE,
and JERRY D. NEIDOFFER

Texas Christian University, Fort Worth, Texas 76129

The purpose of this report is to present functions for correlational analysis to meet the needs of users who have enough experience in APL to work in the desk calculator mode and to use file operations, but who do not have time to develop such functions for themselves. These functions, and others planned for later reports, are designed to be readily used in the desk calculator mode and equally suitable for use in functions written by users to meet their own needs. The functions have been developed through active use in our local APL system and incorporate features found desirable in exploratory analysis of psychological data.

Since correlation studies typically involve more data than can be conveniently entered and corrected in the desk calculator mode, a function to handle data entry is also provided. Correlation tables of modest size exceed the line width on terminals (especially CRTs). They also represent valuable output that may need special handling to produce clean copy and storage for reuse. A function offering these amenities is provided.

Data Entry

TABLE. TABLE is designed to support data entry in tabular form. Its use requires a minimum of APL skills, ensures the right number of entries in each row, and provides a simple means of correcting errors. TABLE and its supporting functions, BFR and REPLACES, are shown in Figure 1.

When invoked, TABLE will ask for the number of entries per row. It will then offer a menu: ENTER, FIX, or END. ENTER permits entry of data, one row at a time, with end of row indicated by carriage return. If the number of entries is not correct, TABLE will report the error and go into the FIX mode. It will return to the ENTER mode only when the user has made changes that result in a correct length for the row.

The FIX option permits the user to enter one or more numbers that include the place at which the error is to be corrected. The string of numbers must be unique within the data set to be corrected. (During data entry, the data set is the line just entered. When the FIX mode is entered from the menu, the data set is the entire string of values entered up to that point.) The user is asked to enter a second string of numbers in which the error has been corrected. The first string is then replaced by the second.

```
TABLE BFR REPLACES 1 24 80
V TABLR:TXT:KND:Q:R:T:F
[1] KND←DATA←10
[2] 'HOW MANY ENTRIES PER ROW:'
[3] R←0
[4] MENU:'ENTER, FIX, END:'
[5] ~0,ND,0oDATA←DATA
[6] ENTER:'ENTER ITEMS, TO QUIT TYPE END AT 0:'
[7] ST←MENU←0oTXT←0
[8] 'OK:CA←P:0oTXT'
[9] 'ERROR -':0oTXT:'ENTRIES - REPLACE WHAT:'
[10] 0←0
[11] 'BY WHAT:'
[12] ~OK,0J REPLACES T
[13] CA←ST,0DATA←DATA,R+TXT
[14] FIX:'TO QUIT TYPE END'
[15] TXT←DATA
[16] FX←REPLAC WHAT:'
[17] ~MENU←0oQ←0
[18] 'BY WHAT:'
[19] ~FX←0oQ←((0Q)≠0T←0)/'ENTRIES MUST BE EQUAL LENGTH'
[20] T REPLACES Q
[21] ~FX,0DATA←TXT
[22] ND:DATA←(QF,0oDATA)+F0DATA
[23] 0←1:'Y'×1+0,0←'SHOW?'
[24] (2 4o'IS P8.2')(1+V0X11)DATA;1oPNT DATA
[25] ~MENU←1:'Y'×1+0,0←'END?'
[26] 'RESULT LEFT IN DATA'
[27] NOTE THAT 0PNT IS CALLED 0PNT IN APLSV
V
R←T BFR K:S
[1] ~LX←2oP8,0oR←R/1oR←S(1)←(1-05+K)+T
[2] ~LX←2oP8,0oR←R/(T(9+1)←S(2))/R
[3] R←(T(R←1+1 1-05)A←24S)/R
[4] LI←0+((1o0oK)≠1oP8←R-1
[5] 0R:'INSTANCES OF ':S
V
D←P REPLACES B:R
[1] D←'NO CHANGE'
[2] ~((1 0oP8←R((0P←P)-0R)×1+1oR←TXT BFR R←B)/RP,0
[3] 0←1:'Y'×1+0P,0←REPLAC ALLT'
[4] RP:TXT←(R(1)0oTXT),P,(R(1)0oB)+TXT
[5] ~RP×1oR←1+R
[6] D←'DONE'
V
```

Figure 1. APL code for functions supporting data entry. Documentation is given in the text.

The END option results in an offer to display the table. The display, if requested, presents a table in integer form (five column fields) or decimal form (eight column fields with two decimal places), depending on whether the data are integers or not. Following the display, the user is offered the option to END or to return to the menu for correction of errors that may have been identified in the display.

When the function ends, it leaves the final result as a two-dimensional array named DATA.

R←T BFR K. BFR (before) is a general-purpose support function designed to locate in the vector T all occasions of the sequence given in K. The result, R, represents the location of each occasion by giving index numbers of the last character before each occasion. (The reason for this reference point can be seen in the way the result is used in REPLACES.)

If the number of occasions found is other than one, BFR reports the number. For some uses of BFR, this output may be inconvenient; it can be suppressed by making K a two-dimensional array.

BFR operates on either character or numerical data.

D←P REPLACES B. REPLACES operates much like the REPLACE function found in many APL libraries, except that it will work with either character or numerical data. It operates on the global vector TXT. The effect is that the string P replaces the string B, provided one instance of that string is found in TXT. If more than one instance is found, the user is given the option of replacing all instances or none.

The result, D, is a character vector intended for output to the terminal. The contents of D indicate whether a change was made or not.

Output

BOP O. BOP (big output) is a general-purpose output function. It accepts a character vector in O, attaches a label provided by the calling function, adjusts the width of the output lines to conform to the specification given in the global variable OWD (output width), files the result, pauses for the user to adjust the terminal to receive output, and delivers or withholds output at the user's request. BOP and its supporting function, FAPE, are presented in Figure 2.

BOP assumes the following conditions to be provided by the user or the calling function:

TB and STR are character vectors. Normally, TB would carry a standard label, such as the study title, to be put at the top of each result. STR would carry other key information, such as sample size, to be put on each result. Following these conventions insures that each unit of output gets the same label.

CTR is a numerical value of probable importance in the interpretation of the output. For example, it may specify a threshold below which no values are reported. BOP guarantees that a report of its value will appear on every unit of output.

OWD is a numerical variable specifying the number of characters per line of output. For a typewriter terminal, OWD might be set to 120. For a CRT, OWD might be set to 80 or 72. For typed output in manuscript form, OWD might be set to 64. OWD does not override the)WIDTH setting of the work space; this value must be equal to or greater than OWD for proper arrangement of output.

In conjunction with user adherence to a formatting convention, OWD permits properly arranged output of tables when the line width exceeds the width of the terminal. When this control is left to the)WIDTH setting, the remainder of the line is presented as the next line, destroying the tabular structure. BOP breaks up a table into two or more vertical segments of width less than or equal to OWD. It presents each separately. If they are typed out, the separate parts can be put side by side to reform the table.

```

BOP FAPE 1 24 80
  V BOP O;S;Q;P
[1] P←1+S÷S/(O+S+1),O←Q×A
[2] ←FX×12÷/(OWD+1)×P-1+S
[3] B0←O←Q÷O×P=0
[4] ST←0,((OWD(S-1+P;Δ)+P),Δ
[5] Q←Q,(OWD+S+P),Δ
[6] ←SP+10×P-1+S+P
[7] FI(0=TB,STR,('CTR=', 'P4.2'ΔPMT CTR),Δ,0,2ΔΔ)FAPE 2
[8] ←FX×10×O÷(TYPE+1)/'READY'
[9] EN←B0×1-Δ/Δ=0
[10] EN←B0×1-Δ/Δ=0
  V
  V P←X,FAPE Y
[1] X←PAPPEND Y
[2] P←1+1+PΔIN Y
  V

```

Figure 2. APL code for a general-purpose output function. Documentation is given in the text.

For good appearance, OWD needs to be set so as to segment the table between columns. We find it convenient to follow the convention that all tables have columns eight characters wide. (Most results in psychological research can be formatted in fields of F8.2.) With this convention, any multiple of eight can be chosen for OWD.

If the output vector contains only one line greater than OWD, segmenting will not occur.

TYPE is a numerical value set to one if the user wants the output sent to the terminal. Under that setting, BOP will type "READY" and wait for a carriage return before delivering the output. If a table is being segmented under OWD control, this action will occur before each segment. The pause allows the user to make adjustments at the terminal before accepting output.

The user can suppress terminal output by typing NO in response to the READY signal. This will set TYPE to zero and cause BOP to suppress terminal output until TYPE is reset.

An output file tied to the number 2 is assumed to be available. BOP stores each result in this file (after segmenting the table if necessary). This permits the results to be sent to the line printer (on some systems), to be retyped for better appearance, or to be available for review if the work is being done on a CRT.

BOP calls the special file function FAPE. The standard file function, FAPPEND may be substituted without affecting BOP.

O←X FAPE Y. FAPE is a slightly modified version of FAPPEND, the standard file function normally available in FILES (APL 1). It appends the contents of the variable X as the next component in the file tied to Y, as does FAPPEND. In addition, FAPE delivers in O the number of the component in which X was put. This report makes it much more convenient to reaccess the filed data.

Correlations

V←R MVST X. MVST is designed to provide commonly needed results for many multivariate and univariate applications. X may be a vector or a two-dimensional array with columns corresponding to variables and rows corresponding to subjects (more generally, rows represent the dimension over which summation is to be done). R is a control parameter that specifies the desired result:

R = 1 delivers a correlation matrix as the result in V. The means, standard deviations, and variances of the columns are left in the global variables MX, SD, and VR, respectively.

R = 2 delivers in V the variance-covariance matrix. MX, SD, and VR are formed as above.

R = 3 delivers in V the sum of squares and cross-products matrix. MX is formed; SD and VR are not formed.

X = 0 delivers in V the values of X restated as

line and maintain the variables as global variables in a work space. In that way they can more easily be reset according to the user's needs. The control parameters, OWD and TYPE, are explained in the discussion of BOP. CTR sets a threshold for printing correlations. Only correlations exceeding this threshold are displayed. The others are replaced by blanks. CTR can thus be used to suppress correlations the user considers to be scientifically or statistically insignificant.

General Considerations and Use

CORRELATE will operate in a 32K work space with 24 variables and 50 subjects, assuming the data array is in integer form. From the user's standpoint it produces no appreciable delay. It is operating on a Honeywell Sigma-9, but it uses functions and conven-

tions generally accepted by the APL community. Potential differences in function names are noted in program comments.

This set of functions may suggest the need for functions to produce scatterplots or principal components analysis. Reports are presently being prepared on functions compatible with those described here and designed to meet such needs.

REFERENCE

- EVANS, S. H., & GAGE, F. H. APL programs for interactive data analysis: Basic statistics and histograms. *Behavior Research Methods & Instrumentation*, 1979, 11, 605-606.

(Accepted for publication April 18, 1980.)