DECUS
PROGRAM LIBRARY

DECUS NO.	8-118
TITLE	GENERAL LINEAR REGRESSION
AUTHOR	Ian E. Bush
COMPANY	Medical College of Virginia Richmond, Virginia
DATE	Submitted: November, 1967
FORMAT	

PROGRAM OUTLINE:

GENERAL LINEAR REGRESSION

The actual program listed here is number LR-VIII-C1. The listing and the punched symbolic tape are broken into convenient sections by leader-trailer code so that the various sections of the program can be copied, modified, or removed at will.

The heart of the program is the section called "Main Arithmetic-IX". This is also designed in modular form to allow flexible use by compiling with alternative versions of the sections of LR-VIII-C called "Control Sections". Each main section of Main Arithmetic-IX ends with the statement GO TO 400. This effectively allow these sections of program to be used as sub-routines under the control of a calling section in the main program, in this case "Control Section C2". Statement 400 in this section continuously increments a controlling variable INC4, so that the following computed GO TO statement can gain access to the appropriate section of Main Arithmetic IX. This device is useful in many programs since there is usually a definite order in which these sections of Main Arithmetic-IX need to be called upon. Repetitive access to a section in Main Arithmetic IX can be achieved by setting the controlling variable INC4 to one less than the desired number at the end of each repetitive operation. An example of this type of usage of the computed GO TO linkage is seen in the final section of the main program at statement number 431.

Main Arithmetic IX consists of four initializing statements (501 onwards); an input section (MA-1B); a weighting section (MA-1C); a section (MA-1D) which cumulates means, sums of squares, etc.; a section which calculates the relevant regression coefficients, etc. , (MA-2); and a section which calculates confidence limits as variances (MA-3). Re-iteration through the loop ending at statement 510 was designed to minimize errors while avoiding storage of the variables and was based on a suggestion by Dr. D. Calhoun of the Department of Biometry at G. D. Searle, Inc. (see also Sapega (1967) Decuscope, Vol. 6, No. 3). The statements used are very similar to those given by Sapega in his excellent article but are generalized to allow for weighting. To avoid the nuisance of error diagnostics being given by certain functions with negative values, the dependent variable is converted to its positive absolute value by Section MA-1C and squares are produced by multiplication in section MA-1D.

Section MA-2 allows for both cases of linear regression and in the computation of standard error of the intercept uses (N-2) degrees of freedom to provide a better estimate for small values of N while providing negligible differences from conventional calculation when N is large.

Section MA-3 provides a calculation of the variance of the error of the estimate of the dependent variable again using (N-2) degrees of freedom for the general case. This calculation is fully corrected for both random variance within the tested population of data and for the difference between the independent variable and the mean of the independent variable for the population of data.

The main program is designed to help the occasional or inexperienced user by asking for the required constants and data in a quasi-conversational mode. Following the conventions established for many time-sharing conversational mode computer systems, the demand for a given number is preceded by a vertical arrow. For the manipulation of large numbers of regressions previously stored on paper-tape or some other type of input, it would be advantageous to remove those FORMAT statements and TYPE statements providing for this quasi-conversational mode of operation, since it would waste time and space. For the convenience of editing these FORMATS have all been numbered in the range 901 and over, and grouped in the section called "FORMATS 2" at the start of the symbolic listing.

Included in these FORMATS are number 900 and the first four executed statements of the program. This is a useful device in many programs for the PDP-8 and consists as follows. The operator types as the first number input to the program any integer number which he does not want to use in descriptive headings, dates, etc., at the beginning of his data sheet. This allows the operator complete flexibility in the headings and dating of his data sheet and effectively keeps the computer in a non-calculating mode until the first number is typed again. When this is done, the fourth executable statement of the program transfers control to Statement 2, whereupon the program proper is entered. Subsequent operation of the program is easily understood from inspecting a typical input and data sheet which is attached. Following the execution of the three statements beginning at 401, the operator has the option of obtaining a plot of the calculated regression line together with the true envelope of confidence limits of any desired kind. On receiving the request "PLOT OR NOT? 0 or 1", the operator types a zero if he does not want the plots and a 1 if he does. Following a zero (0) at this point the computer either asks for the data of the next regression to be done, or halts with the usual exclamation mark. Following a "1" the computer asks in turn for the lower and upper limits of the independent variable over which the plot is desired ("X1, X2") and the number of intervals into which the plot should be divided. It then asks for the appropriate value of Student's T stating the correct number of degrees of freedom for which this should be looked up in a table of T, and the value of P describing the appropriate confidence limits (e.g. 95.0 for percentage fiducial limits, or 0.05). If the error of the estimates of the dependent variable is desired in the form of standard errors this is achieved by simply typing 1.0 in place of the proper value for Student's T. This part of the program is very useful for many purposes. Thus for instance if only one interval is requested two co-ordinates are rapidly given which allow a graphical plot of the true regression line to be made on graph paper. The value of P is redundant to the computation but must be typed.

If the regression is being used for a calibration curve, then this section of the program can be entered whenever the standard error or fiducial limits of the estimate of the dependent variable are desired.

This program has been extensively tested against worked examples in standard statistical textbooks using unweighted regression and has also been tested fairly extensively with certain special cases of weighted linear regression. Main Arithmetic IX is written so that the whole of the weighting section (MA-1C) can be omitted or altered without affecting the operation of the program as a whole. Transformations of the inputs of the dependent and independent variables can be inserted either at the end of section MA-1B or at the end of section MA-1C for compiling special versions of the program. There are a number of limitations and inelegancies in this program and other workers may have found other tricks to get round some of the problems which were faced. The weighting functions represent a selection of fairly useful ones but will not meet all requirements. A larger selection could be offered if the number of numbered statements were reduced by excluding some of the FORMAT statements. It is also not always easy to get agreement between statisticians as to the appropriate numbers of degrees of freedom to be used. Since much of the use of this program was intended for calibration curves and other data containing a small number of pairs of variables (N-2) degrees of freedom was adopted as the best compromise for the variance of the estimate of the independent variable and the variance of the estimates of the intercept. Accuracy is preserved by the iterative system used in Main Arithmetic IX (see Sapega above), but some of the values come out as irritating strings of six nines after the decimal point instead of the rounded whole number, an error of the FORTRAN Operating System for the PDP-8 that has been noticed by many others.

I would be glad to receive any criticisms or suggestions from readers or users of this program.

November 7, 1967
IEB/AB

EXPLANATION OF OUTPUT AND SYMBOLS

All symbols underlined were typed by the operator.

↑	=	Please type input number(s)
SUM WTS.	=	Sum of weights or no. of pairs X, Y when unweighted
SSDX	=	$\sum (x - \bar{x})^2$ with weighting
SSDY	=	$\sum (y - \bar{y})^2$ with weighting
SDXY	=	$\sum (x - \bar{x}) (y - \bar{y})$ with weighting
M	=	Slope
C	=	Intercept
S. E.	=	Standard error of the preceding parameter
SUM SQS. RESIDUALS	=	Sum of squares of deviations of dependent variable due to residual errors of observations from the regression line.
CORR. COEFF.	=	Correlation coefficient
X	=	Independent variable
Y	=	Dependent variable
EY	=	Estimate of Y from M and C
DY	=	Error of EY as S. E. or confidence limit

11/7/67
IEB/AB

```
C;  GENERAL CASE OF C NOT KNOWN TO BE ZERO
C;
 511;    EM=SDXY/SSDX
         CNST=YBAR-EM*XBAR
         SSRS=SSDY-EM*SDXY
         DUM2=SQTF(SSRS/SSDX/(EN2-1.0))
         DUM3=SQTF(SSRS/(EN2-2.0)*(1.0/SWT+XBAR*(XBAR/SSDX)))
         DUM4=SDXY/SQTF(SSDX*SSDY)
   GO TO 400
C;
C; LT CODE

C; SECTION MA-3
C;     CONFIDENCE LIMITS AS VARIANCE
C;
 504;   DUM2=DUM1-XBAR
        DUM4=(SSRS/(EN2-2.0))*(1.0+(1.0/SWT)+DUM2*(DUM2/SSDX))
     GO TO 400
C;
C; LT CODE

C; CONTROL SECN. C3
C; TYPE MEANS, WEIGHTS, ETC.
C;
 401;     TYPE 802, J, XBAR, YBAR, SWT, SSDX, SSDY, SDXY
          TYPE 803, EM, DUM2, CNST, DUM3, SSRS
          TYPE 812, DUM4
C;
C; PLOT CONFIDENCE LIMITS , ETC. IF WANTED
C;
        TYPE 907
        TYPE 920
        ACCEPT 800, IWTF
     GO TO(402), IWTF
      GO TO 600
402;    TYPE 908
        TYPE 920
       ACCEPT 801, ELM1, ELM2
       ACCEPT 300, INTC
       N2=N2-1
       TYPE 909, N2
       TYPE 920
       ACCEPT 801, STN2, DUM2
403;   EINT=INTC
       DUM5=(ELM2-ELM1)/EINT
       DUM1=ELM1-DUM5
        I=0
431; INC4=3
       DUM1=DUM1+DUM5
      GO TO 504
432;   DUM4=SGTF(DUM4)*STN2
       DUM3=DUM1*EM+CNST
       TYPE 805, DUM1, DUM3, DUM4
   I=I+1
   IF(I-INTC) 431, 431, 600
C;
C; CONTROL SECN. C7
C;
600; CONTINUE
       STOP
C;
C; LT CODE

C;
   END
```

```
C; LR-VIII-C
C;     GENERAL LINEAR REGRESSION WITHOUT STORAGE
C;      USING MA-IX.
C;
C;      VERSION 1   CONVERSATIONAL MODE
C;        NOTE;  FOR LARGE NOS. OF REGRESSIONS
C;     ETC.,ELIMINATE ALL FORMAT STATEMENTS
C;     WITH NUMBERS FROM 9Ø1 UPWARDS AND
C;      ALL TYPE STATEMENTS USING THEM
C;
C;       USES ENTRY CODE DEVICE TO ALLOW VARIABLE
C;      TITLES,DATES,ETC.AT START OF INPUT
C;     TYPE ANY NUMBER NOT NEEDED IN HEADINGS
C;      THEN TYPE HEADINGS,DATES,ETC.,
C;     THEN TYPE THE FIRST NUMBER AGAIN WHEN
C;       READY TO START COMPUTATION
C;
C; 1Ø/23/67....1Ø/27/67
C;
C; LT CODE

C; CONTROL SECN.S. PRECEDED BY FORMATS TO REDUCE
C;  FORWARD REFERENCES
C;
C;  FORMATS 2
C;
 9ØØ;    FORMAT(/,/,"ENTRY CODE ↑")
C;
9Ø1;    FORMAT(/,"NO.OF REGRESSIONS ")
9Ø2;    FORMAT(/,/,/,"NO.OF POINTS ")
9Ø3;    FORMAT(/,"INTERCEPT CASE   Ø OR 1 ")
92Ø;    FORMAT("PLEASE!",/,"↑")
9Ø4;    FORMAT(/,"POWER ")
9Ø6;    FORMAT(/,/,"X & Y ")
9Ø7;    FORMAT(/,"PLOT OR NOT?   Ø OR 1 ")
9Ø8;    FORMAT(/,"LIMITS X1,X2 & NO.OF INTERVALS ")
9Ø9;    FORMAT(/,"STUDENT'S T & P FOR ",I,"DEGREES OF FREEDOM ")
921;    FORMAT(/,"WEIGHTING  1=NONE; 2=POWER; 3=LOG; 4=EXP",/)
C;
C; LT CODE

C;
C; FORMATS
C;
8ØØ;    FORMAT(I)
8Ø1;    FORMAT(E)
8Ø2;    FORMAT(/,/,"REGRESSION NO.",I,/,"MEAN X=",E,"MEAN Y=",E,'
                "SUM WTS.=",E,/,"SSDX=",E,"  SSDY=",E,"  SDXY=",E,/,/)
8Ø3;    FORMAT("M=",E," S.E. ",E,/,"C=",E," S.E. ",E,/,'
                 "SUM SQS.RESIDUALS=",E,/,/)
8Ø5;    FORMAT(/,"X ",E,"EY ",E," DY ",E)
812;    FORMAT("CORR.COEFF.=",E,/,/)
```

```
C; SECTION MA-1B
C;      INPUT OF X AND Y.CAN INCLUDE ANY TRANSFORMATIONS
C;      IF NEEDED,E.G. Y=LOGF(Y). TO PROTECT AGAINST
C;      NEGATIVE VALUES USE SECTION SIMILAR TO
C;      START OF SECN.1C BELOW.
C;
    ACCEPT 8Ø1,X,Y
C;
C; LT CODE

C; SECTION MA-1C
C;      WEIGHTING FUNCTIONS
C;   EXITS WITH WT ONLY MODIFIED FROM 1.Ø IF WEIGHTING ASKED FOR
C;
   IF(Y)551,552,552
551;   DUM1=-Y
      GO TO 553
552;   DUM1=Y
553;  GO TO(5Ø2,554,555,556),IWTF
C;
554;    WT=DUM1**IPWR
     GO TO 5Ø2
555;    WT=LOGF(DUM1)
      GO TO 5Ø2
556;    WT=EXPF(DUM1)
C;
C; LT CODE


C; SECTION MA-1D
C;     MAIN CALCULATIONS
C;
5Ø2;   SWT=SWT+WT
      DUM1=X-XBAR
       DUM2=Y-YBAR
      DUM3=(SWT-WT)/SWT
      DUM4=WT/SWT
      SSDX=SSDX+DUM3*DUM1*DUM1
      SSDY=SSDY+DUM3*DUM2*DUM2
      SDXY=SDXY+DUM3*DUM1*DUM2
      XBAR=XBAR*DUM3+X*DUM4
      YBAR=YBAR*DUM3+Y*DUM4
51Ø;CONTINUE
C;
   GO TO 4ØØ
C;
C; LT CODE

C; SECTION MA-2
C;     REGRESSION COEFFICIENTS,ETC.
C;
C;  SUBSECTION FOR CASE OF C KNOWN TO BE ZERO
C;
5Ø3;  GO TO(511),INTC
C;
      EN2=EN2+1.Ø
     SSDX=SSDX+SWT*XBAR*XBAR
     SSDY=SSDY+SWT*YBAR*YBAR
     SDXY=SDXY+SWT*XBAR*YBAR
```

```
C;
C;
C; LT CODE
C; CONTROL SECN.C1
C;
       TYPE 9ØØ
       ACCEPT 8ØØ,IPWR
  1;   ACCEPT 8ØØ,INTC
      IF(INTC-IPWR)1,2,1
C;
  2;    TYPE 9Ø1
       TYPE 92Ø
        ACCEPT 8ØØ,N1
   ;DO 6ØØ J=1,N1
       INC4=Ø
C;
      TYPE 9Ø2
      TYPE 92Ø
      ACCEPT 8ØØ,N2
C;
     TYPE 9Ø3
     TYPE 92Ø
     ACCEPT 8ØØ,INTC
C;
     TYPE 921
      TYPE 92Ø
     ACCEPT 8ØØ,IWTF
C;
  IF(IWTF-2)4,3,4
3;     TYPE 9Ø4
       TYPE 92Ø
       ACCEPT 8ØØ,IPWR
C;
C; CONTROL-SECN.C2
C;
  4;      TYPE 9Ø6
       TYPE 92Ø
4ØØ; INC4=INC4+1
      GO TO(5Ø1,5Ø3,4Ø1,432),INC4
C;LT CODE


C; MAIN ARITHMETIC-IX
C;
C;   CUMULATIVE CALCULATION OF MEANS,SUMS OF SQUARES,ETC.
C;   WITHOUT STORAGE BASED ON D.CALHOUN'S SUGGESTION
C;   1965.SEE ALSO SAPEGA(1967,DECUSCOPE,VOL.6,NO.3)
C;   FOR UNWEIGHTED VERSION.
C;
C;   MOD...........1Ø/26/67
C; LT CODE


C; SECTION MA-1A
C;        INITIALIZING,ETC.
C;
 5Ø1;     EN2=N2
      XBAR=YBAR=SWT=SSDX=SSDY=SDXY=Ø.Ø
   ;DO 51Ø I=1,N2
        WT=1.Ø
C; LAST STATEMENT GIVES UNWEIGHTED CASE
C; LT CODE
```