UNITED STATES DEPARTMENT OF THE INTERIOR

GEOLOGICAL SURVEY

Algorithm and BASIC Program for Ordinary Least Squares Regression in Two and Three Dimensions

by Gary R. Olhoeft

Open-File Report 78- 876

September, 1978

Algorithm and BASIC Program for Ordinary Least Squares Regression in Two and Three Dimensions

by Gary R. Olhoeft

Equations must frequently be fitted to data to make the data more easily usable (Bevington, 1969; Searle, 1971; Daniel and Wood, 1971). Most least squares fitting algorithms are written for equations in two dimensions. A typical linear example is

$$y = f(x) = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k$$

where the data are in pairs of x,y points and the coefficients, b_i, are to be found. This type of equation, which is linear in the coefficients, may be represented in matrix form as

$$Y = XB$$

where all capital letters are matrices, and

Y is n by 1 in size (n data points).

X is n by k+1 in size (k+1 coefficients in b)

and B is k+1 by 1.

The desired solution to the above equation is the coefficient matrix B. The expectation value of the coefficient matrix is

$$\underline{B} = (X^T X)^{-1} X^T Y.$$

It must be noted that the inversion of $X^T X$ may be ill conditioned for large k unless the arithmetic precision of the computer is sufficiently high. The programs illustrated here require 48-bit arithmetic. The development of this type of ordinary least squares fit is adequately derived in Searle (1971). Table 1 summarizes the resultant findings. Program 1 illustrates a typical use: fitting electrical conductivity data of aqueous solutions to their concentration dependence using the equation

$$y = b_0 + b_1 x^{1/2} + b_2 x \ln x + b_3 x + b_4 x^{3/2}.$$

However, it is sometimes desirable to fit a three-dimensional equation, y = f(x,z), such as the electrical conductivity dependence

upon both concentration and temperature. In this case, the matrix formulation becomes

$$Y = XBZ$$

with the expectation

$$\underline{B} = (X^TX)^{-1}X^TYZ^T(ZZ^T)^{-1}.$$

The extension of the two dimensional model to three dimensions is fairly straightforward using Searle (1971). The coefficients in the B matrix must still be linear in both X and Z, and the matrices may become ill conditioned if insufficient precision is used. Program 2 illustrates the three-dimensional example.

Both programs 1 and 2 are written in Hewlett-Packard 9830 BASIC*, but may be easily adapted to any language that has the appropriate matrix functions and precision.

REFERENCES:

Bevington, P. R., 1969, Data Reduction and Error Analysis for the Physical Sciences: New York, McGraw-Hill, 336 p.

Daniel, C. and Wood, F. S., 1971, Fitting Equations to Data: New York, Wiley-Interscience, 342 p.

Searle, S. R., 1971, Linear Models: New York, Wiley, 532 p.

* The use of trade names is for illustrative purposes only and does not imply endorsement or recommendation by the U.S. Geological Survey.

TABLE 1 : Ordinary least squares regression (intercept model, Searle, 1971)

$y(x) = b_o + b_1 x + \ldots + b_k x^k$ (or other equation linear in coefficients b)

$Y = XB$ where Y is n by 1, X is n by k+1, and B is k+1 by 1

$\underline{B} = (X^T X)^{-1} X^T Y$ = expectation value of coefficient matrix B

$\text{var}(\underline{B}) = (X^T X)^{-1}\sigma^2$ = variance in expectation value of B

σ^2 = square of standard deviation of original data

$\underline{Y} = X\underline{B}$ = estimated expectation value of Y

$\text{var}(\underline{Y}) = X(X^T X)^{-1} X^T \sigma^2$

$SSE = (Y - \underline{Y})^T (Y - \underline{Y})$ = residual error sum of squares = $Y^T Y - \underline{B}^T X^T Y$

$\underline{\sigma^2} = SSE/(n-k-1)$ = square of expectation value of standard deviation for n data points and k+1 coefficients

$SST = Y^T Y$ = total sum of squares

$SSR = SST - SSE = \underline{B}^T X^T Y$ = sum of squares due to regression

$SSM = n\bar{y}^2$ = sum of squares of mean

$SSR_m = SSR - SSM$ = regression sum of squares corrected for the mean

$SST_m = SST - SSM$ = corrected sum of squares

$R^2 = SSR/SST$ = square of multiple correlation coefficient for $b_o = 0$

= fraction of total sum of squares accounted for by fitting the model

$R_m^2 = SSR_m/SST_m = R^2$ for $b_o \neq 0$

Program 1 : Two-dimensional ordinary least squares regression

Lines 10-20 Dimension arrays for equation of form Y = XA
30-40 Input a title for the printout
50-100 Enter the data in x,y pairs: this program has been abbreviated to illustrate the principle. As shown, it requires exactly 30 x,y pairs to be entered, but it may be easily modified to accept an arbitrary number.
110-150 Solution for coefficient matrix, A.
160-260 Solution for SSR (S3), SST (S2), SSE (S1), standard deviation (V2) and R*R (R2).
270-400 Printout coefficient matrix, raw data, expectation values of y, and differences.
410-480 Function y(x) to be fitted.

NOTES: Sufficient arithmetic precision must be used to prevent X^TX from becoming ill conditioned: 48-bit is recommended. To alter the program for use with a different number of x,y pairs, change the dimensions that are equal to 30 in lines 10, 20, 50, and 360. For an arbitrary number, place a test at line 71 to exit the FOR-NEXT loop at an agreed value of x,y. Set the X and Y matrices equal to zero at line 21, and set the dimension that is 30 in line 360 equal to the variable I.

To change the function, alter lines 420-460. If more or less than 5 parameters in x are desired, change the dimension that is 5 in lines 10, 20, 250, 290, 300, and 330 as appropriate.

Program 1 : Two-dimensional ordinary least squares regression

```
10 DIM Y[30,1],X[30,5],A[5,1],E[5,1],D[5,30],E[5,5],F[5,5]
20 DIM T$[72],J[1,30],S[1,1],T[1,5],H[5,30],I[30,1]
30 DISP "TITLE";
40 INPUT T$
50 FOR I=1 TO 30
60 DISP "X, Y(X)";
70 INPUT X,Y
80 Y[I,1]=Y
90 Z=FNF1
100 NEXT I
110 MAT D=TRN(X)
120 MAT E=D*X
130 MAT F=INV(E)
140 MAT H=F*D
150 MAT A=H*Y
160 MAT J=TRN(Y)
170 MAT S=J*Y
180 S2=S[1,1]
190 MAT T=TRN(A)
200 MAT J=T*D
210 MAT S=J*Y
220 S3=S[1,1]
230 S1=S2-S3
240 V2=0
250 V2=SQR(S1/(I-5))
260 R2=S3/S2
270 PRINT
280 PRINT T$
290 WRITE (15,300)A[1,1],A[2,1],A[3,1],A[4,1],A[5,1]
300 FORMAT "Y(X)=",F8.2," +",F8.2,"SQRX +",F9.2,"XLNX +",F9.2,"X +",F9.2,"X^
310 WRITE (15,320)S1,V2,R2
320 FORMAT "SSE=",F12.5,"  S.D.=",F12.5,"  R*R=",F12.5
330 PRINT I-1;"DATA POINTS;";"  5 PARAMETERS"
340 MAT I=X*A
350 PRINT "  X           Y OBS     Y CALC   Y OBS - Y CALC"
360 FOR I=1 TO 30
370 WRITE (15,380)X[I,4],Y[I,1],I[I,1],Y[I,1]-I[I,1]
380 FORMAT E10.3,3F10.3
390 NEXT I
400 END
410 DEF FNF(I)
420 X[I,1]=1
430 X[I,2]=SQRX
440 X[I,3]=X*LOGX
450 X[I,4]=X
460 X[I,5]=X^(3/2)
470 RETURN I
480 END
```

Program 2 : Three-dimensional ordinary least squares regression

lines 10-20	Dimension arrays for equation of the form S = CBT.
30-50	Input a title for the printout.
60-230	Input the data: for 5 concentrations, 16 pairs of temperatures and conductivites are entered.
240-330	Solve for coefficient matrix B.
340-350	Solve for estimated expectation values of S (given in the O matrix).
360-550	Print the results: raw data, estimated data, and the difference in percent as well as the coefficient matrix, B.

NOTES: Again, for brevity and to demonstrate the principle, much of the program has been reduced to essentials. This program, as shown, requires exactly 5 concentrations and 16 conductivity-temperature pairs at each concentration. These may be changed by appropriately modifying lines 10, 20, 60, 140, and 360-480. The functions to be fit for C are in lines 90-130 and for T in lines 170-210. These may be rewritten as functions as shown in Program 1. Both the conductivity dependence upon concentration and temperature are shown with 5 variables in this listing. If that is altered to be more or less than 5, then lines 10, 20, 360-480, and 510-540 must also be altered.

Program 2 : Three-dimensional ordinary least squares regression.

```
10 DIM B[5,5],T$[80],N[5,5],O[5,16],Q[5,16]
20 DIM C[5,5],T[5,16],S[5,16],D[5,5],E[5,5],F[16,5],G[5,5],H[5,5]
30 DISP "PRINTOUT TITLE";
40 INPUT T$
50 PRINT T$
60 FOR I=1 TO 5
70 DISP "CONCENTRATION";
80 INPUT C
90 C[I,1]=C
100 C[I,2]=C*SQRC
110 C[I,3]=C^2*LOGC
120 C[I,4]=C^2
130 C[I,5]=C*C^(3/2)
140 FOR J=1 TO 16
150 DISP C;" TEMPERATURE, CONDUCTIVITY ";
160 INPUT T,S[I,J]
170 T[1,J]=1
180 T[2,J]=1/T
190 T[3,J]=T
200 T[4,J]=T^2
210 T[5,J]=T^3
220 NEXT J
230 NEXT I
240 MAT E=TRN(C)
250 MAT F=TRN(T)
260 MAT B=E*C
270 MAT G=INV(B)
280 MAT N=G*E
290 MAT Q=N*S
300 MAT H=Q*F
310 MAT B=T*F
320 MAT G=INV(B)
330 MAT B=H*G
340 MAT D=C*B
350 MAT O=D*T
360 WRITE (15,450)0,C[1,1],C[2,1],C[3,1],C[4,1],C[5,1]
370 PRINT
380 FOR J=1 TO 16
390 WRITE (15,450)T[3,J],S[1,J],S[2,J],S[3,J],S[4,J],S[5,J]
400 WRITE (15,450)T[3,J],O[1,J],O[2,J],O[3,J],O[4,J],O[5,J]
410 FOR I=1 TO 5
420 Q[I,J]=100*(O[I,J]-S[I,J])/S[I,J]
430 NEXT I
440 WRITE (15,460)Q[1,J],Q[2,J],Q[3,J],Q[4,J],Q[5,J]
450 FORMAT F10.0,5F10.4
460 FORMAT 10X,5F10.2
470 PRINT
480 NEXT J
490 PRINT
500 PRINT "Coefficient matrix:"
510 FOR I=1 TO 5
520 WRITE (15,530)B[1,I],B[2,I],B[3,I],B[4,I],B[5,I]
530 FORMAT 5E10.3
540 NEXT I
550 END
```