JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 1980, 34, 199-206 NUMBER 2 (SEPTEMBER)

TECHNICAL NOTE

A NOTE ON FITTING HERRNSTEIN'S EQUATION

Herrnstein (1970, 1974) has described an equation that relates response strength to reinforcement rate. The equation contains two free parameters, and Cohen (1973) has described a simple method for estimating the value of these parameters. Using Cohen's method we have obtained some parameter values that are far from the best fit of Herrnstein's equation to the data. This note examines some problems surrounding the use of Cohen's method and describes two alternative methods.

Herrnstein's equation is a hyperbolic function of the form

$$R = \frac{kr}{r + r_e} \quad (1)$$

where R denotes the measure of response strength, r denotes the rate of reinforcement of the measured response, r_e denotes all sources of reinforcement other than r, and k denotes the maximal value of R when r_e approaches zero with r greater than zero.

Cohen's Double-Reciprocal Method

Cohen's method for determining the value of the parameters k and r_e involves taking the reciprocal of both sides of Equation 1, thereby reducing the problem to a linear least-squares fit in $1/R$ and $1/r$:

$$\frac{1}{R} = \frac{1}{k} + \frac{r_e}{k}\left(\frac{1}{r}\right). \quad (2)$$

Once $1/k$ and r_e/k are known, it is a simple matter to find values for k and r_e.

We have applied Cohen's double-reciprocal method to numerous sets of data obtained from measuring schedule-induced drinking as a function of the rate of food delivery (see Wetherington, 1979). Despite the fact that the obtained data points yielded negatively accelerated functions, many of the functions produced by the double-reciprocal method were actually positively accelerated. In general, the fits yielded very high parameter estimates and accounted for very little of the data variance. Panel A in Figure 1 illustrates three representative instances in which Cohen's method failed to produce satisfactory fits to the data. All data points were obtained from a single rat engaging in schedule-induced drinking (Rat J-7, Wetherington, 1979). The three graphs, respectively, show ingestion rate (R), lick rate (R), and relative time spent drinking (R) as functions of rate of food presentation (r). The smooth curves were obtained from Equation 1 by deriving the values of k and r_e from the linear least-squares solution of Equation 2. The three numbers near each curve represent the values of k, r_e, and the percentage of the variance (v) accounted for by Equation 1. For each data set the first four points fall on or near the curve. However, note the poorness of the fit for the last data point. This aberrant point contributes heavily to the failure of the functions to account for a significant portion of the data variance.

The parameter estimates for the hyperbolic fits are also unacceptable for theoretical reasons. First, the k value for relative time is 4.8 whereas theoretically this value should not exceed 1.0. Second, for ingestion rate and lick rate, the values of k and r_e are negative which makes them difficult to interpret theoretically. Moreover, these negative values result in a positively accelerated hyperbolic function though visual inspection of the data points clearly suggests a negatively accelerated function.

Panel B shows the data from Panel A replotted in the reciprocals of R and r. Here, the straight line is indeed the optimum least-squares fit to the reciprocal data. For all three measures the double-reciprocal function provides a reasonably good fit to the data, accounting for 99.3%, 96.8%, and 95.7%, respectively, of the data variance. Thus, a good fit of Equation 2 can yield an extremely poor fit of Equation 1.

As noted both by Dowd and Riggs (1965) and Cohen (1973), the direct use of Equation 2 tends to assign an inordinately important role to the low rates of response and/or reinforcement in determining the values of k and r_e. This phenomenon can be explained in the following manner. In the straight line fit of Equation 2, the values of the reciprocals of R and r are highly correlated (ideally, they are on a straight line with positive slope) so that the data points corresponding to large R are clustered near the origin whereas the data points for small R are spread out (see Panel B). Assuming that all values of R contain at least some minimal measurement error, then clearly measurement errors for small numbers have relatively larger effects on their reciprocals than measurement error for large numbers. Stated differently, perturbations in low values of R produce much greater effects on $1/R$ than do perturbations in high values of R. Thus, the spread out values of $1/R$ contain more measurement error than the clustered values of $1/R$, thereby serving to

We thank Aaron Brownstein, Richard Shull, and Nilo Niccolai for their helpful comments. We also thank Harriet Saxton who skillfully prepared the graphics. Reprints may be requested from Cora Lee Wetherington, Department of Psychology, University of North Carolina at Charlotte, Charlotte, North Carolina 28223.

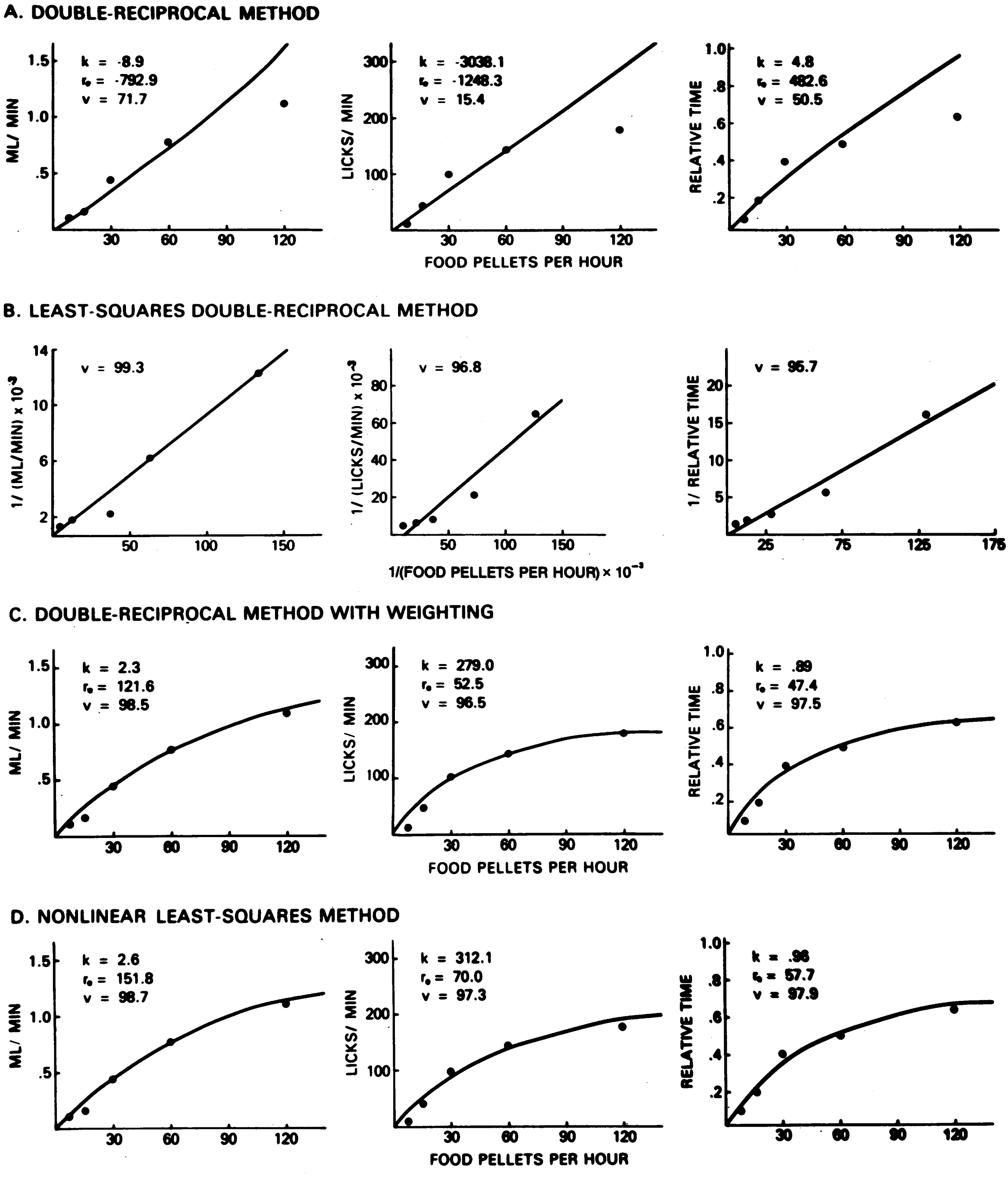

Fig. 1. *Panel A:* Ingestion rate (first column), lick rate (second column), and relative time spent drinking (third column) as functions of the rate of food delivery for Rat J-7 from Wetherington (1979). The smooth curves indicate the hyperbolic fit to the data provided by the double-reciprocal method. *Panel B:* The data from Panel A are plotted as reciprocals. The straight lines indicate the least-squares fit to the data. *Panel C:* Data points are the same as in Panel A. The smooth curves indicate the hyperbolic fit to the data provided by the double-reciprocal method with weighting. *Panel D:* Data points are the same as in Panel A. The smooth curves indicate the hyperbolic fit to the data provided by the nonlinear least-squares method.

make the estimates of k and r_e overly dependent on low rates of response. For example, if the lowest values of R are badly underestimated, their reciprocals will be highly overestimated which effectively rotates the least-squares line counter clockwise, possibly leading to physically meaningless negative values of k and r_e as seen in Panel A. Likewise, if the lowest value of R is badly overestimated, its reciprocal will be highly underestimated which effectively rotates the least-squares line clockwise, generating parameter estimates which produce a hyperbolic curve that underestimates the high values of R.

Cohen (1973) pointed out that in enzyme reaction research a hyperbolic function has been used to relate the initial velocity of an enzyme reaction to the concentration of the substrate (Dowd & Riggs, 1965). Dowd and Riggs reported that the standard procedure for estimating the value of the two free parameters had been to use any of three linear transformations of the hyperbolic function, the most popular being the double-reciprocal transformation. They compared the parameter estimates provided by the three linear transformation methods and found that the double-reciprocal method provided parameter estimates which were "excessively large, or even negative" (p. 865) and were the least reliable. Dowd and Riggs concluded that the double-reciprocal method had received "undeserved popularity" and "should be abandoned" (p. 869).

The Double-Reciprocal Method with Weighting

Dowd and Riggs (1965) pointed out that proper weighting of the data points could partially correct for the effects of the double-reciprocal transformation on measurement errors. They suggested weighting each data point by the measured value of the dependent variable raised to the second power if the measurement error is a proportionate one and raising it to the fourth power if the measurement error is a constant one. In practice, of course, one often does not know the nature of the measurement errors. Using the three data sets in Panel A, we have applied the double-reciprocal method using various integer powers and found the fourth power to yield the best fit to Equation 1. The resulting curves are shown in Panel C. In all cases the curves are negatively accelerated, provide excellent fits to the data, and yield theoretically acceptable parameter estimates. These outcomes stand in marked contrast to those produced by the double-reciprocal method without weighting (Panel A). It is worthwhile noting that for other rats from the same experiment other powers yielded the best fit. We now describe an exact approach which removes all of the difficulties of the above approximate methods.

The Nonlinear Least-Squares Method as Applied to Herrnstein's Equation

Since any method that transforms Equation 1 will give free parameter estimates for the transformed equation instead of Equation 1, a method that fits Equation 1 directly is desirable. The great advantage of a linear transformation method, with or without weights, is that it results in a linear least-squares problem which simply leads to solving two linear algebraic equations. Equation 1 leads to a nonlinear least-squares problem: find k and r_e so as to minimize the expression

$$\sum_{i=1}^{n}\left(R_i - \frac{kr_i}{r_i + r_e}\right)^2 . \tag{3}$$

This expression leads (by computing partial derivatives with respect to k and r_e) to a system of two nonlinear equations in k and r_e:

$$\sum \frac{R_i r_i}{r_i + r_e} - k \sum \frac{r_i^2}{(r_i + r_e)^2} = 0 \tag{4}$$

and

$$\sum \frac{R_i r_i}{(r_i + r_e)^2} - k \sum \frac{r_i^2}{(r_i + r_e)^3} = 0 \tag{5}$$

which may be written as

$$\frac{\sum \frac{R_i r_i}{r_i + r_e}}{\sum \frac{R_i r_i}{(r_i + r_e)^2}} = \frac{k \sum \frac{r_i^2}{(r_i + r_e)^2}}{k \sum \frac{r_i^2}{(r_i + r_e)^3}} . \tag{6}$$

Thus, cancelling the k's and cross-multiplying gives a single nonlinear equation in the remaining parameter r_e:

$$\sum \frac{R_i r_i}{r_i + r_e} \sum \frac{r_i^2}{(r_i + r_e)^3} = \sum \frac{R_i r_i}{(r_i + r_e)^2} \sum \frac{r_i^2}{(r_i + r_e)^2} . \tag{7}$$

Now an exact fit of Equaion 1 can be found by solving Equation 7 for r_e and then solving for k by using

$$k = \frac{\sum \frac{R_i r_i}{r_i + r_e}}{\sum \frac{r_i}{(r_i + r_e)^2}} \tag{8}$$

derived from Equation 4.

We have used an effective variant of the Newton-Ralphson method known as the secant method (Conte & deBoor, 1972, p. 59) in our FORTRAN code for solving the nonlinear Equation 7 to near machine accuracy (see Appendix). For starting values for r_e we use zero and the negative of the minimum of r_i divided by two. If the iterative procedure fails to converge, we restart with values of r_e slightly less than the negative of the maximum of r_i. This procedure includes the case of a positively accelerated curve, although this is not anticipated.

The resulting hyperbolic functions fitted to the data sets in Panel A are shown in Panel D. A comparison of these data with those in Panel C reveals that in all cases slightly more of the data variance is accounted for by this method. It should be stressed that this approach gives the exact fit of Equation 1 in the least-squares sense, whereas all other methods, including the weighted double-reciprocal transformation, merely estimate the optimum parameter values by fitting exactly a different equation.

Figure 2 further illustrates comparison of the double-reciprocal methods and the nonlinear method by showing rate of keypecking by pigeons exposed to variable-interval (VI) schedules of food reinforcement. Data in the first column were obtained from Bird 117 in an experiment by Catania (1963) in which reinforcement rate on two concurrent VI schedules was varied from 0 to 40 reinforcements per hour while holding the over-

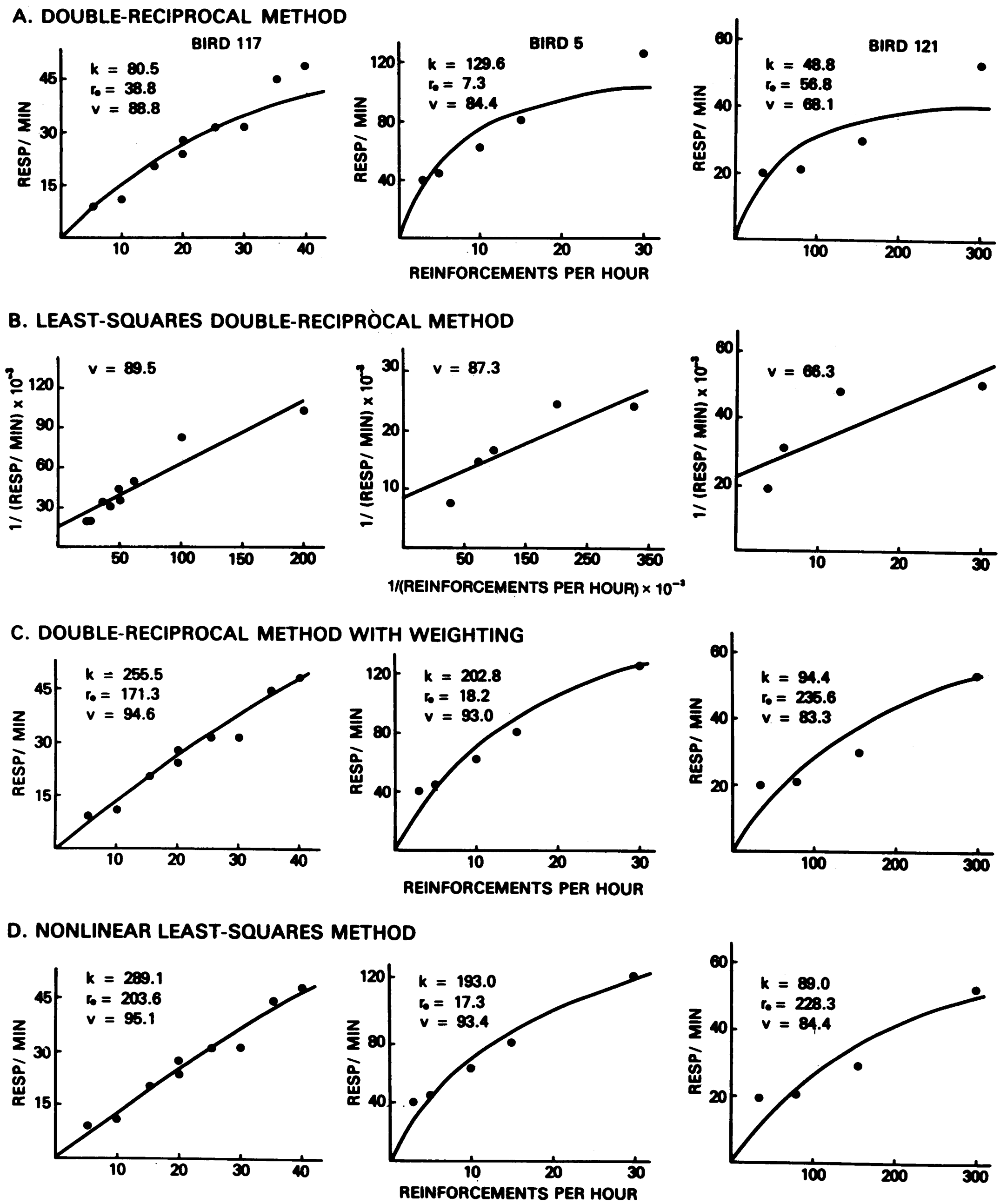

Fig. 2. *Panel A:* Key pecking rate as a function of reinforcement rate for Bird 117 from Catania (1963), first column; Bird 5 from Findley (1958), second column; Bird 121 from Catania and Reynolds (1968), third column. The smooth curves indicate the hyperbolic fit to the data provided by the double-reciprocal method. *Panel B:* The data from Panel A are plotted as reciprocals. The straight lines indicate the least-squares fit to the data. *Panel C:* Data points are the same as in Panel A. The smooth curves indicate the hyperbolic fit to the data provided by the double-reciprocal method with weighting. *Panel D:* Data points are the same as in Panel A. The smooth curves indicate the hyperbolic fit to the data provided by the nonlinear least-squares method.

all rate constant at 40 reinforcements per hour. Data points are from both schedules. The second column shows data for Bird 5 from an experiment by Findley (1958) in which reinforcement rate on one of two concurrent VI schedules was varied from 3 to 30 reinforcements per hour while the reinforcement rate provided by the other schedule remained at 6 reinforcements per hour. Data points are from the variable schedule. The third column shows data for Bird 121 from an experiment by Catania and Reynolds (1968) in which VI schedules provided reinforcement rates ranging from 33.3 to 300 per hour. As in Figure 1, Panel A of Figure 2 shows the hyperbolic curve generated via the double-reciprocal method, Panel B shows the double-reciprocal plot of the data points, Panel C shows the hyperbolic curve generated via the double-reciprocal method with weighting, and Panel D shows the hyperbolic curve generated via the nonlinear least-squares method.

In all three cases the curves produced via the double-reciprocal method (Panel A) underestimated the highest data point. In fact, in 15 of 18 data sets in the 3 studies, the double-reciprocal parameter estimates were less than the nonlinear least-squares estimates. The data sets in Figure 2 represent the case from each study in which the discrepancy was the greatest. The underestimates provided by the double-reciprocal plots indicate the existence of measurement errors which inflate lower values of R thus leading to values of $1/R$ that are too small. This is clearly indicated in both Panel A where the R values at the lowest reinforcement rate appear high relative to other values and in Panel B where the corresponding $1/R$ values at the highest $1/r$ value appear accordingly low. These $1/R$ values have the effect of reducing the slope of the regression line thereby reducing the parameter estimates and producing a curve that is lower than the highest R value. Note that in Panel C, the use of the double-reciprocal method with weighting produces curves quite similar to those produced by the nonlinear least-squares method (Panel D). The powers used in the double-reciprocal method with weighting were five, six, and nine, for the three birds, respectively. While we are illustrating the procedure of systematically determining the power of the weighting function, this is not an approach we advocate, as it has no theoretical basis and is not so satisfactory as the nonlinear least-squares approach.

In summary, we have examined three methods for estimating the free parameters of Herrnstein's equation. One of these methods, the popular double-reciprocal method, although recommended by Cohen (1973), has been previously evaluated by Dowd and Riggs (1965) and found to be inadequate. Our findings are in agreement with theirs. We found that a weighting provided a better fit to the data as was also reported by Dowd and Riggs. We now have described a third method, the nonlinear least-squares method, and have developed a simple method of implementing it. When used with a computer, it can provide the best fit of Herrnstein's equation to almost machine accuracy. If one wishes to find the best fit of Herrnstein's equation to a data set and has access to a computer, the nonlinear least-squares method is clearly the preferred method.

Cora Lee Wetherington
Thomas R. Lucas
University of North Carolina at Charlotte

REFERENCES

Catania, A. C. Concurrent performance: Reinforcement interaction and response independence. *Journal of the Experimental Analysis of Behavior*, 1963, **6**, 253-263.

Catania, A. C., & Reynolds, G. S. A quantitative analysis of the responding maintained by interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior*, 1968, **11**, 327-383.

Cohen, I. L. A note on Herrnstein's equation. *Journal of the Experimental Analysis of Behavior*, 1973, **19**, 527-528.

Conte, S. D., & deBoor, C. *Elementary numerical analysis*. New York: McGraw-Hill, 1972.

Dowd, J. E., & Riggs, D. S. A comparison of estimates of Michaelis-Menten kinetic constants from various linear transformations. *The Journal of Biological Chemistry*, 1965, **240**, 863-869.

Findley, J. D. Preference and switching under concurrent scheduling. *Journal of the Experimental Analysis of Behavior*, 1958, **1**, 123-144.

Herrnstein, R. J. On the law of effect. *Journal of the Experimental Analysis of Behavior*, 1970, **13**, 243-266.

Herrnstein, R. J. Formal properties of the matching law. *Journal of the Experimental Analysis of Behavior*, 1974, **21**, 159-164.

Wetherington, C. L. Schedule-induced drinking: Rate of food delivery and Herrnstein's equation. *Journal of the Experimental Analysis of Behavior*, 1979, **32**, 323-333.

Received March 29, 1979
Final acceptance March 31, 1980

APPENDIX

A FORTRAN CODE FOR THE NONLINEAR LEAST-SQUARES METHOD

Below is a 140 statement FORTRAN program implementing the nonlinear least squares method described above. The code is designed for conversational use over a terminal, in which data may be entered, corrected, and analysed, after which data pairs may be added or deleted and then reanalysed. The program may also be used in batch mode. The program includes a feature where it determines a convergence criterion appropriate to the particular computer on which the code is being used. It is required that all data pairs be positive numbers. The program gives instructions for entering data as it is used. It is recommended that all read statements be modified to the local version of free format. For checkout purposes the data (7.5,15.42), (15.,43.13),(30.,106.10),(60.,150.31) and (120.,192.29) gives RE = 69.9713971, K = 312.0993807, and 97.30% of the variance is explained. Total run time was .6 seconds on a Burroughs 6700.

```
C-    DRIVER FOR THE NONLINEAR LEAST-SQUARES METHOD
      COMMON N,R(100),R1(100),XK
      DIMENSION R1R(100),RR(100),W(100)
  500 FORMAT(I2)
  510 FORMAT(2F10.0)
  520 FORMAT(I1)
  530 FORMAT(A1,T1,2F10.0)
  600 FORMAT (" OPTION LIST:"//" 1 RECALL OPTION LIST"/
     *" 2 ENTER OR ADD DATA PAIRS"/" 3 CORRECT DATA PAIRS"/
     *" 4 DELETE DATA PAIRS"/" 5 LIST DATA PAIRS"/
     *" 6 STOP"/" 7 SOLVE FOR RE AND K"/)
  610 FORMAT (" ENTER DESIRED OPTION IN COL 1"/)
  620 FORMAT(" GIVE DATA POINTS X AND Y IN PAIRS. (COL 1-10,11-20)"/
     *" ENTER + IN COL 1 WHEN FINISHED."/)
  630 FORMAT(2G20.10/)
  640 FORMAT(" DO YOU WANT TO CORRECT ANY DATA PAIRS? ENTER 0 FOR NO,"
     * /" OR DATA PAIR NUMBER IN COL 1-2 FOR A CORRECTION."/)
  650 FORMAT( " ENTER NEW VALUES FOR ",I2,"TH DATA PAIR."/)
  660 FORMAT(I5,2G15.5)
  680 FORMAT(/" RESULTS BY THE NONLINEAR LEAST SQUARES METHOD (PASS"
     * ,I2,"):"/"     RE = ",G20.10," K = ",G20.10)
  690 FORMAT(" ANALYSIS OF RESULTS:"// 4X,
     * "UNEXPLAINED VARIANCE = ",G20.10/"     TOTAL VARIANCE = ",G20.10/
     * "     VARIANCE EXPLAINED BY HERRNSTEIN°S EQUATION = ",F6.2"%")
  700 FORMAT(" SUSPICIOUS VALUES:  RESTART ACTIVATED"//)
  740 FORMAT(" METHOD EITHER FAILED TO CONVERGE, OR FIT REMOVES"/
     * " LESS THEN 10% OF THE VARIANCE.  RUN TERMINATED.")
  750 FORMAT(4X,"#",9X,"R1",14X,"R")
  760 FORMAT(" ENTER NUMBER OF DATA PAIR TO BE DELETED IN COL 1-2."/)
  770 FORMAT(" DATA POINTS "I3" DELETED."I4" DATA POINTS REMAINING.")
  780 FORMAT(" INDEX = "I3" IS OUT OF RANGE:  REENTER!")
C-    DATA INPUT SECTION.  ALL READS MAY BE CHANGED TO FREE FORMAT.
      N = 0
  110 WRITE(6,600)
   50 WRITE(6,610)
      READ(5,520) IOP
      GO TO (110,120,130,140,150,160,170),IOP
  120 WRITE(6,620)
  122 NP1 = N + 1
      READ(5,530) AA,R1(NP1),R(NP1)
      IF (AA .EQ. "+") GO TO 50
  125 N = NP1
      WRITE(6,630) R1(N),R(N)
      GO TO 122
  130 WRITE(6,640)
      READ (5,500) NCORR
      IF (NCORR) 50,50,135
  135 WRITE (6,650) NCORR
      READ (5,510) R1(NCORR),R(NCORR)
      WRITE (6,630) R1(NCORR),R(NCORR)
      GO TO 130
  140 WRITE (6,760)
      READ(5,500) NDEL
      IF (NDEL .LE. N)  GO TO 145
      WRITE (6,780) NDEL
      GO TO 140
  145 R1(NDEL) = R1(N)
      R(NDEL) = R(N)
      N = N - 1
      WRITE (6,770) NDEL,N
      GO TO 50
  150 WRITE (6,750)
      DO 155 I = 1,N
```

```
155      WRITE (6,660) I,R1(I),R(I)
    GO TO 50
157 WRITE (6,740)
160 STOP
170 XMIN = 1.E20
    DO 180 I = 1,N
180    XMIN = AMIN1(XMIN,R1(I))
    XR = 0.
    XL = -XMIN/2.
    IGO = 1
    CALL SECANT(XL,XR,RE,IFLAG)
    IF (IFLAG .EQ. 0) GO TO 800
350 XMAX = 0.
    DO 360 I = 1,N
360    XMAX = AMAX1(XMAX,R1(I))
    XR = -XMAX -.01
    XL = XR - .01
    IGO = 2
    CALL SECANT (XL,XR,RE,IFLAG)
    IF (IFLAG .NE. 0) GO TO 157
800 WRITE (6,680) IGO,RE,XK
    SUM = 0.
    DO 810 I = 1,N
810    SUM = SUM +(R(I)-XK*R1(I)/(R1(I)+RE))**2
    SY = 0.
    SY2 = 0.
    DO 820 I = 1,N
       SY = SY + R(I)
820    SY2 = SY2 + R(I)*R(I)
    VAR = SY2-SY*SY/FLOAT(N)
    ETA2 = (VAR-SUM)/VAR
    ETAPER = ETA2 * 100.
    WRITE (6,690) SUM,VAR,ETAPER
    IF (XMIN .LT. -RE .AND. -RE .LT. XMAX) GO TO 830
    IF (ETA2 .GE. .1) GO TO 50
830 WRITE (6,700)
    GO TO (350,157) IGO
    END
    SUBROUTINE SECANT(XL,XR,X,IFLAG)
    IFLAG = 0
C-  COMPUTE A MACHINE DEPENDENT STOPPING TOLERANCE, EPSILON.
    EPS = 1.
 10 EPS = EPS/2.
    IF (1.+EPS .NE. 1.) GO TO 10
    EPS = EPS*100.
    FL = F(XL)
    FR = F(XR)
    DO 20 I = 1,60
       IF (I .EQ. 30 .AND. ABS(XR-XL) .GT. 100.) GO TO 30
       X = XL - FL * (XR-XL)/(FR-FL)
       FR = FL
       FL = F(X)
       IF(ABS(X-XL) .LE. EPS * (1.+ABS(X)) .AND. ABS(FL).LE. .5)RETURN
       XR = XL
 20    XL = X
 30 IFLAG = -1
    RETURN
    END
    FUNCTION F(RE)
    COMMON N,R(100),R1(100),XK
```

```
      DOUBLE PRECISION S1,S2,S3,S4,SUM,T1,T2
      S1=0.
      S2=0.
      S3=0.
      S4=0.
      DO 100 I=1,N
         SUM = DBLE(R1(I)) + DBLE(RE)
         T1 = DBLE(R(I)) * DBLE(R1(I)) / SUM
         T2 = DBLE(R1(I))**2 / SUM**2
         S1 = S1 + T1
         S2 = S2 + T2 / SUM
         S3 = S3 + T2
 100     S4 = S4 + T1 / SUM
      F = S4 - S1 * S2 / S3
      XK = S1 / S3
      RETURN
      END
```