A Statistical Method for Determining the Breakpoint of Two Lines

RICHARD H. JONES AND BRUCE A. MOLITORIS

Department of Preventive Medicine and Biometrics, and Department of Medicine, School of Medicine, University of Colorado Health Sciences Center and V. A. Medical Center, Denver, Colorado 80262

Received February 24, 1984

A method is presented for determining the breakpoint of a line which suddenly changes slope at some unknown point. A statistical test is given for testing whether the broken line is a significantly better fit to the data than a single straight line. An approximate confidence interval can be obtained for the position of the breakpoint.

KEY WORDS: breakpoints; statistical methods; Hill plots; confidence intervals.

The problem of finding the breakpoint of two straight lines joined at some unknown point has a long statistical history. An early paper on the subject is Sprent (1), and a recent bibliography is given by Shaban (2). A common method is to find the division of the points into two groups which gives the smallest residual sum of squares when a different line is fit to each group (3). A problem with this method is that the point where the two lines intersect may not be in the interval where the division was made, and could possibly even be outside the range of the data. The approach used in this paper is to fit two lines that join at a breakpoint that can vary continuously within the range of the data. The main purpose is to present a simple statistical method that can be carried out on a microcomputer. This method has multiple biological and nonbiological (2) uses. In the present paper the breakpoint of renal cortical alkaline phosphatase is quantitated.

It is important that the user realize the underlying assumptions. The errors on the observations are assumed to be independently distributed about the joined lines with a Gaussian distribution with constant variance. The computer program prints out the normalized residuals so a judgement can be made about whether this assumption is true. The normalized residuals are the deviations of the observations from the fitted joined lines divided by their standard deviation. Only about 5% of these normalized residuals should be greater than 2.0 in absolute value.

When a transformation is used to obtain straight lines, the original untransformed data may satisfy the above assumptions and the transformation may destroy these properties. In this case it would be better to use nonlinear regression methods on the original data. This problem is discussed by Klotz (4) and Munson and Rodbard (5).

It is possible to fit other models which approximate a broken straight line. Griffiths and Miller (6) use hyperbolic regression for two-phase piecewise linear regression with a smooth transition between regimes at the breakpoint.

THE METHOD[1]

If x_0 is the position of the unknown breakpoint, the equations of the two lines are

$$y = \beta_0 + \beta_1 x \quad x \leqslant x_0$$

and

$$y = \beta_2 + \beta_3 x \quad x > x_0.$$

Adding the following constraint forces the lines to joint at x_0,

$$\beta_0 + \beta_1 x_0 = \beta_2 + \beta_3 x_0.$$

[1] A BASIC program listing for an IBM-PC is available from the first author.

0003-2697/84 $3.00

One of the constants can be eliminated by solving this equation for β_2,

$$\beta_2 = \beta_0 + \beta_1 x_0 - \beta_3 x_0.$$

Now the two equations are

$$y = \beta_0 + \beta_1 x \quad x < x_0$$

$$y = \beta_0 + \beta_1 x_0 + \beta_3(x - x_0) \quad x \geqslant x_0.$$

The method to be used searches for the value of x_0 that minimizes the residual sum of squares (RSS) for the above three-parameter linear regression. This is actually a four-parameter regression involving three linear parameters, β_0, β_1, and β_3, and one nonlinear parameter, x_0. It is important to note that it is only necessary to search over values of the nonlinear parameter. For each value of this parameter, the breakpoint, a linear regression problem is solved and the RSS calculated. Different values of x_0 are tried until the value that minimizes RSS is found. The range of x that is searched is from the next to the smallest value to the next to the largest value. It is easy to see that if the breakpoint is within the last interval at either end of the data, the line will break and go through the end point. If there are n points, this will give the same RSS as fitting a line to the $n - 1$ points which exclude the end point. If the breakpoint is outside the range of the data, then a single straight line is the best fit to the data.

A systematic search procedure is used to find the value of x_0 that minimizes the RSS ($\mathrm{RSS_{min}}$). When this value is found, the mean square error (MSE) is calculated as

$$\mathrm{MSE} = \mathrm{RSS_{min}}/(n - 4),$$

where n is the number of observations. The degrees of freedom are $n - 4$ since three linear parameters and one nonlinear parameter (x_0) have been estimated.

Since this is nonlinear regression, statistical tests are approximate rather than exact. An approximate F test can be calculated to test whether the broken line is a significantly better fit than a single straight line. Fit a single straight line to the data, and let the residual sum of squares be $\mathrm{RSS_s}$. The F test, which has 2 and $n - 4$ degrees of freedom, based on the extra sum of squares principle (7), is

$$F_{2,n-4} = (\mathrm{RSS_s} - \mathrm{RSS_{min}})/(2\ \mathrm{MSE}).$$

A significant F indicates that the broken line is a better fit than a single straight line.

An approximate 95% confidence interval can also be found by first finding the appropriate F value with 1 and $n - 4$ degrees of freedom. In the example used here, $n = 15$, so

$$F_{1,11}(.05) = 4.84.$$

This value is multiplied by MSE to determine how much the RSS must increase for x_0 to change significantly. This can be determined graphically by plotting the values of RSS that were calculated during the search for the minimum, and is shown in Fig. 1.

Approximate confidence intervals on the estimated slopes can be calculated using the

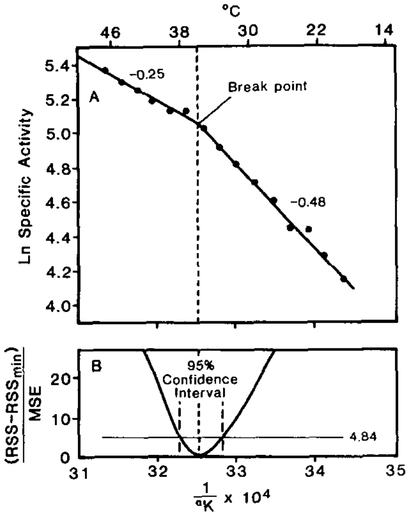

FIG. 1. The Arrhenius analysis of renal cortical brush border membrane alkaline phosphatase (A). The breakpoint (34.7°C) is indicated by the vertical dashed line. The data represent individual determinations (ln specific activity) at 2°C intervals. (B) Shows the graphic determination of the confidence interval around the breakpoint. $F_{1,11}(P = .05)$ is 4.84 shown by the horizontal line.

standard errors of the slopes calculated as in linear regression using the t distribution with $n - 4$ degrees of freedom. The approximation is due to fixing the breakpoint at its estimated value.

RESULTS

Figure 1A shows an Arrhenius plot of renal cortical alkaline phosphatase with the breakpoint (34.7°C) determined using the described microcomputer method. The F value ($F_{2,11}$) was 26.4 and, therefore, the data fit a broken line better than a straight line ($P < 0.001$). Figure 1B reveals how an approximate 95% confidence interval can be constructed. The interval is not necessarily symmetrical because of the nonlinear estimation.

Table 1 shows the data of 10 different Arrhenius plots from two different treatment groups. In each group, four out of five of the plots are better modeled by a line with a breakpoint. It is doubtful whether the plots that do not have a significant breakpoint should be used to calculate either the breakpoint or the apparent energy of activation for the group. Statistical analysis for group differences (Student's t test) are carried out using only the plots with significant breakpoints (Sig. vs. Sig.) and using all of the data (All vs. All). The results show that the locations of the breakpoints are significantly different between the two groups, and the apparent energies of activation are not different between the two groups. The apparent difference in E_{act2} is due to case 3 in group II, where a straight line is a better fit to the data so the breakpoint is not well determined.

TABLE 1

COMPUTER-ASSISTED ANALYSIS OF ARRHENIUS DATA FOR RENAL BRUSH BORDER ALKALINE PHOSPHATASE

	Breakpoint (°C)	P	Apparent energy of activation (kcal/mol) E_{act1}	E_{act2}
Group I				
1	32.8	<0.05	6.22	9.79
2	34.4	<0.01	4.53	10.66
3	34.7	<0.01	5.03	9.51
4	30.0	NS	6.54	10.66
5	35.9	<0.01	6.13	9.24
Sig (n = 4)	34.5 ± 0.6		5.5 ± 0.4	9.8 ± 0.3
All (n = 5)	33.6 ± 1.0		5.7 ± 0.4	10.0 ± 0.3
Group II				
1	31.9	<0.05	6.54	13.36
2	28.1	<0.01	5.95	9.24
3	28.1	NS	8.69	14.04
4	23.6	<0.01	7.50	12.85
5	30.4	<0.01	4.76	12.49
Sig (n = 4)	28.5 ± 1.8		6.2 ± 0.6	12.0 ± 0.9
All (n = 5)	28.4 ± 1.4		6.7 ± 0.7	12.4 ± 0.8
Sig vs Sig	<0.05		NS	NS
All vs All	<0.05		NS	<0.05

Note. Statistical analysis to determine if the data better fit a one- or two-line model was carried out as described in the text. Those plots which fit a two-line model better are designated "Sig" (n = 4). "All" (n = 5) refers to the data from all plots. Comparisons between groups was carried out using a 2-tailed t test. Values are reported as the means ± SE. These data are used as an illustration only.

DISCUSSION

The program described above has several advantages. First, data can be statistically analyzed so that significance levels can be attached to conclusions such that a broken line fits the data better than a single straight line. This has been a constant problem in the use of Arrhenius analysis (8). This program uses the variation about the individual data points to determine if the data is better fit by a one or two component line. The data in Table 1 emphasizes the importance of this determination. In a method reported previously (3), multiple determinations of individual points were needed to allow statistical evaluation. A second advantage of this method is the calculation of an approximate 95% confidence interval. The third advantage is use of the program on a microcomputer. Finally, to analyze for more than one breakpoint, the data above and below each questioned breakpoint can be analyzed separately.

REFERENCES

1. Sprent, P. (1961) *Biometrics* **17**, 634–645.
2. Shaban, S. A. (1980) Int. Stat. Rev. **48**, 83-93.
3. Cook, D. A., and Charnock, J. S. (1979) *J. Pharmacol. Methods* **2**, 13-19.
4. Klotz, I. M. (1982) *Science* **217**, 1247-1249.
5. Munson, P. J., and Rodbard, D. (1983) *Science* **220**, 979-981.
6. Griffiths, D. A., and Miller, A. J. (1973) *Commun. Statist.* **2**, 561-569.
7. Draper, N. R., and Smith, H. (1981) Applied Regression Analysis, 2nd ed., Wiley, New York.
8. McElhaney, R. N. (1982) *Curr. Top. Memb. Trans.* **17**, 317-380.