Problems with linear regression as applied to TL data

H. M. Rendell
Geography Laboratory
University of Sussex
Falmer, Brighton BN1 9QN
Sussex

The purpose of this note is to compare various techniques that are commonly used by workers within TL to fit the 'best' straight line to a set of data points. Despite the fact that the use of some of these techniques has become routine, a comparison is instructive, revealing differences of up to 6% in the intercept on the dose axis.

Discussion will be confined to the 'best fit' rather than to the error terms, and concerns fitting a regression to TL signals vs radiation dose for the simplest of cases i.e. Nat and Nat + beta, in order to determine the intercept on the dose axis when signal = zero or I_0. The nature of the TL data places certain constraints on the regression techniques. First, the pairs of values of signal and dose are not drawn at random, the dose values are selected in the laboratory, usually at regular intervals, whereas the values of TL signal are allowed to vary. Second, the data points may be clustered, with several values of signal for each dose value. Lastly, although both variables will be subject to errors, the errors in signal will tend to be much larger than those for dose.

In order to fit the equation:

$$y = a + bx$$

values of b (slope) and a (intercept on y axis) need to be determined. Owing to the apparently trivial nature of the problem, one is tempted to use a standard regression package without considering its suitability. Details of 5 different techniques for determining the slope of the regression line are given below, in all cases except equation 4, the intercept, a, is given by:

$$a = \overline{Y} - b\overline{X} \qquad \text{where: } \overline{Y} = \Sigma y/n \quad \overline{X} = \Sigma x/n$$

n = no. pairs data points

Details of equations.

Equation 1 : least squares regression of y (signal) on (x) dose

The sum of the squares of the vertical distances from the regression line is a minimum (Davis, 1973, pp. 192-200).

$$b = \frac{\Sigma xy - \{(\Sigma x . \Sigma y)/n\}}{\Sigma(x^2) - [(\Sigma x)^2/n\}}$$

Equation 2 : least squares regression of x (dose) on y (signal)

The sum of the squares of the horizontal distances from the regression line is a minimum. Since errors in dose are required for ED estimates this technique may seem the obvious choice. Problems stem from the fact that dose values are not selected at random and Williams (1985) considers that in such a case "the regression of x on y may be so greatly changed as to be meaningless".

$$b = \frac{\Sigma(y^2) - \{(\Sigma y)^2/n\}}{\Sigma xy - \{(\Sigma x\ \Sigma y)/n\}}$$

Equation 3 : major axis regression

The sum of squares of the perpendicular distances from the points to the regression line is a minimum (York, 1966, p. 1079, eqn. 1). This technique represents an attempt to take account of the fact that both the dependent and independent variables are subject to error.

$$b = \frac{\Sigma V^2 - \Sigma U^2 + \sqrt{\{(\Sigma V^2 - \Sigma U^2)^2 + 4(\Sigma UV)^2\}}}{2\ \Sigma UV}$$

$U = x - \overline{x}$, $\quad V = y - \overline{y}$.

Equation 4 : weighted least squares regression of y (signal) on x (dose)

This approach is used by Debenham (pers. comm.) and employs a weighting term based onthe assumption that the errors in signal are a fixed percentage of the signal (5% in this case).

$$b = \frac{\Sigma w\ \Sigma wxy - \Sigma wx\ \Sigma wy}{\Sigma w\ \Sigma wx^2 - \Sigma wx\ \Sigma wx}$$

$$w = 1/\ (\sqrt{0.0025\ y^2})^2 \qquad a = \overline{Y} - b\overline{X} \qquad \overline{Y} = \frac{\Sigma wy}{\Sigma w} \quad \overline{X} = \frac{\Sigma wx}{\Sigma w}$$

Equation 5 : least squares cubic fit of x (signal) on y (dose)

This weighted regression of x on y is employed by Berger (Wintle pers. comm.) and uses a special solution of the least squares cubic given by York (1966, p. 1082) that assumes no error in y, but that x is subject to error. Unconventionally, Berger treats dose as his dependent, rather than independent variable.

$$b = \frac{\Sigma(wV^2)}{\Sigma(wUV)}$$

$$U = \Sigma\{x - \frac{\Sigma(wx)}{\Sigma w}\} \qquad V = \Sigma\{y - \frac{\Sigma(wy)}{\Sigma w}\}$$

$$w = 1/x^2$$

In order to compare these equations a real data set of N, N+beta values (n=8) was used. The data have a correlation coefficient of +0.983. The results are given in Table 1. The raw data are given in the appendix.

Table 1 : Results for different linear regression techniques

Equation	b (slope)	a (intercept)	'ED' dose intercept (y = 0)
1	0.467	5.435	11.639
2	0.483	5.316	11.005
3	0.470	5.413	11.517
4	0.482	5.283	10.956
5*	n/a	n/a	10.992

* slope, intercept not comparable with 1-4

A comparison between equations 1 and 2 using an artificial data set (with r = +0.991) gave values for dose intercept differing by 2.9% (for n=10) and 1.3% (for n=20). The extent of disparity between the different techniques will vary as a function of the correlation coefficient and the number of points in the data set.

Discussion

It is apparent from the example given above that there is a significant variation in intercept values depending on which regression strategy is used. The use of weighting procedures allows the investigator to take account of known analytical errors of the co-ordinates of data points, and such a technique appears particularly valuable in the case of isochron data (Faure, 1977, pp. 89-90). However, if we can assume that the errors in the TL signal are large compared with those for dose, the simplest technique is to use Equation 1 and regress y on x minimising the squares of the vertical deviation about the regression line. Alternatively, if we can assume that the errors in signal are a fixed percentage of the signal then a weighting procedure may be more appropriate. Comments from other workers as to which techniques they use, and why, will be welcomed.

References

Davis, J. C. (1973) Statistics and data analysis in geology 550pp.

Faure, G. (1977) Principles of isotope geology 464pp.

Williams, R. B. G. (1985) Intermediate statistics for earth scientists.

York, D. (1966) Least-squares fitting of a straight line Can. J. Phys., 44, 1079-1086.

Reviewed by S. G. E. Bowman, N. C. Debenham and M. Leese.

Appendix: Regression Data

TL signal x 10^3	Dose (min)
4.922	0
5.536	0
7.569	5
8.608	5
10.279	10
9.888	10
12.973	15
11.729	15