

APPENDIX III: SOME MATHEMATICS USEFUL IN GEOCHEMISTRY

LINEAR REGRESSION

Fitting a line to a series of data is generally done with a statistical technique called *least squares regression*. Real data are not likely to fall exactly on a straight line; each point will deviate from the line somewhat. The idea of least squares regression is to find the best line fitting the data by minimizing the *squares of the deviations* from the regression line. The quantity to be minimized is:

$$\sum_{i=1}^n e^2 = \sum_{i=1}^n (y - a - bx)^2 \quad (1)$$

This is known as the *sum of the squares of the deviations* from the line $y = a + bx$. The use of the squares of the deviations means that large deviations will affect the calculated slope more than small deviations. By differentiating equation (1), it can be shown that the minimum value for the left side occurs when the slope is:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (2)$$

where \bar{x} and \bar{y} are the means of x and y respectively and x_i and y_i are the i^{th} pair of observations of x and y respectively. We can see from 7.23 that the regression slope is the product of the deviations of x and y from the mean divided by the square of the deviations of x from the mean. A more convenient computational form of (2) is:

$$b = \frac{\sum (x_i y_i) - \bar{y}\bar{x}n}{\sum x_i^2 - \bar{x}^2n} \quad (3)$$

The intercept is then given by: $a = \bar{y} - b\bar{x}$ (4)

Because real data never fit a line exactly, it is of interest to know the error on the estimate of slope and intercept. The error on the slope is given by:

$$\sigma_b = \sqrt{\left[\sum y_i^2 - \bar{y}^2n - \frac{(\sum (x_i y_i) - \bar{y}\bar{x}n)^2}{\sum x_i^2 - \bar{x}^2n} \right] \left[\frac{1}{(n-2)(\sum x_i^2 - \bar{x}^2n)} \right]} \quad (5)$$

The error on the intercept is:

$$\sigma_a = \sqrt{\left[\sum y_i^2 - \bar{y}^2n - \frac{(\sum (x_i y_i) - \bar{y}\bar{x}n)^2}{\sum x_i^2 - \bar{x}^2n} \right] \left[\frac{1}{n} + \frac{\bar{x}^2}{(\sum x_i^2 - \bar{x}^2n)} \right] \left[\frac{1}{n-2} \right]} \quad (6)$$

Statistics books generally give an equation for linear least squares regression in terms of one dependent and one independent variable. The independent variable is assumed to be known absolutely. With geochemical data, both x and y are often measured parameters and have some error associated with them. These must be taken into account for a proper estimate of the slope and the errors associated with it. In some cases, the errors in measurement of x and y can be correlated, and this must also be taken into account. The so-called *two-error regression* algorithm does this. This is, however, considerably less straight-forward than the above. The approach is to weight each observation according to the measurement error (the weighting factor will be inversely proportional to the analytical error). A solution, written in the context of geochronology, has been published by York (1969).