analysis (MRA). We will consider some of the basic tenets of this approach to gain a firm understanding of the statistical procedures that define a QSAR. Regression analysis is a powerful means for establishing a correlation between independent variables and a dependent variable such as biological activity (56).

$$Y_i = b + aX_i + E_i \qquad (1.10)$$

Certain assumptions are made with regard to this procedure (57):

1. The independent variables, which in this case usually include the physicochemical parameters, are measured without error. Unfortunately, this is not always the case, although the error in these variables is small compared to that in the dependent variable.
2. For any given value of X, the Y values are independent and follow a normal distribution. The error term E_i possesses a normal distribution with a mean of zero.
3. The expected mean value for the variable Y, for all values of X, lies on a straight line.
4. The variance around the regression line is constant. The "best" straight line for model $Y_i = b + aZ_i + E$ is drawn through the data points, such that the sum of the squares of the vertical distances from the points to the line is minimized. Y represents the value of the observed data point and Y_{calc} is the predicted value on the line. The sum of squares $SS = \Sigma (Y_{obs} - Y_{calc})^2$.

$$Y_{obs} = aX_i + b + E_i \qquad (1.11)$$

$$Y_{calc} = aX_i + b \qquad (1.12)$$

$$E = Y_{obs} - aX_i - b \qquad (1.13)$$

$$\sum_{i=1}^{n} E_i^2 = \sum A^2 = SS = \sum (Y_{obs} - Y_{calc})^2 \qquad (1.14)$$

$$\text{Thus, } SS = \sum_{i=1}^{n} (Y_{obs} - aX_i - b)^2 \qquad (1.15)$$

Expanding Equation 1.15, we obtain

$$SS = \sum_{i=1}^{n} (Y_{obs}^2 - Y_{obs}aX_i - Y_{obs}b - Y_{obs}aX_i + a^2X_i^2 + aX_ib - bY_{obs} + abX_i + b^2) \qquad (1.16)$$

Taking the partial derivative of Equation 1.14 with respect to b and then with respect to a, results in Equations 1.17 and 1.18.

$$\frac{dSS}{db} = \sum_{i=1}^{n} -2(Y_{obs} - b - aX_i) \qquad (1.17)$$

$$\frac{dSS}{da} = \sum_{i=1}^{n} -2X_i(Y_{obs} - b - aX_i) \qquad (1.18)$$

SS can be minimized with respect to b and a and divided by -2 to yield the normal Equations 1.19 and 1.20.

$$\sum_{i=1}^{n} (Y_{obs} - b - aX_i) = 0 \qquad (1.19)$$

$$\sum_{i=1}^{n} X_i(Y_{obs} - b - aX_i) = 0 \qquad (1.20)$$

These "normal equations" can be rewritten as follows:

$$b \sum_{i=1}^{n} X_i + a \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_iY_{obs} \qquad (1.21)$$

$$b + a \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_{obs} \qquad (1.22)$$

The solution of these simultaneous equations yields a and b. More thorough analyses of these procedures have been examined in detail (19, 58–60). The following simple example, illustrated by Table 1.3, will illustrate the nuances of a linear regression analysis.