Table 1.3 Antibacterial Activity of *N'*-(*R*-phenyl)sulfanilamides

Compound	$\sigma(X)$	Observed BA (Y)
1. 4-CH_3	−0.17	4.66
2. 4-H	0	4.80
3. 4-Cl	0.23	4.89
4. 2-Cl	0.23	5.55
5. 2-NO_2	0.78	6.00
6. 4-NO_2	0.78	6.00

k = no. of variables = 1
n = no. of data points = 6
$\Sigma X = 1.85$
$\Sigma Y = 31.90$
$\Sigma X^2 = 1.352$
$\Sigma Y^2 = 171.45$
$\Sigma XY = 10.968$

For linear regression analysis, Y = $ax + b$

$$a = (n \cdot \sum xy - \sum x \cdot \sum y)/n \cdot \sum x^2 - (\sum x)^2 = 1.45$$

$$b = (\sum y - a \cdot \sum x)/n = 4.869$$

$$r^2 = \sum xy - \sum x \cdot \sum y/n)^2/(\sum x^2 - (\sum x)^2/n) \cdot (\sum y^2 - (\sum y)^2/n) = 0.875 \quad \therefore r = 0.935$$

$$s^2 = (1 - r^2) \times (\sum y^2 - (\sum y)^2/n)/(n - k - 1) = 0.058 \quad \therefore s = 0.240$$

$$F = r^2 \cdot (n - k - 1)/k(1 - r^2) = 28.52$$

The correlation coefficient r, the total variance SS_T, the unexplained variance SSQ, and the standard deviation, are defined as follows:

$$r^2 = 1 - \frac{\Sigma \Delta^2}{SS_T} \tag{1.23}$$

$$SS_T = \sum (Y_{obs} - Y_{mean})^2 = \sum y^2 - (\sum y)^2/n \tag{1.24}$$

$$\sum A^2 = SSQ = \sum (Y_{obs} - Y_{calc})^2 \tag{1.25}$$

$$s = \sqrt{\frac{\Sigma \Delta^2}{n - k - 1}} = \sqrt{\frac{SSQ}{n - k - 1}} \tag{1.26}$$

The correlation coefficient r is a measure of quality of fit of the model. It constitutes the variance in the data. In an ideal situation one would want the correlation coefficient to be equal to or approach 1, but in reality because of the complexity of biological data, any value above 0.90 is adequate. The standard deviation is an absolute measure of the quality of fit. Ideally s should approach zero, but in experimental situations, this is not so. It should be small but it cannot have a value lower than the standard deviation of the experimental data. The magnitude of s may be attributed to some experimental error in the data as well as imperfections in the biological model. A larger data set and a smaller number of variables generally lead to lower values of s. The F value is often used as a measure of the level of statistical significance of the regression model. It is defined as denoted in Equation 1.27.

$$F_{k_2-k_1,n-k_2} = \frac{(SS_1 - SS_2)}{SS_2} \frac{(n - k_2 - 1)}{k_2 - k_1} \tag{1.27}$$

A larger value of F implies a more significant correlation has been reached. The confidence intervals of the coefficients in the equation reveal the significance of each regression term in the equation.

To obtain a statistically sound QSAR, it is important that certain caveats be kept in mind. One needs to be cognizant about collinearity between variables and chance correlations. Use of a correlation matrix ensures that variables of significance and/or interest are orthogonal to each other. With the rapid proliferation of parameters, caution must be exercised in amassing too many variables for a QSAR analysis. Topliss has elegantly demonstrated that there is a high risk of ending up with a chance correlation when too many variables are tested (62).

Outliers in QSAR model generation present their own problems. If they are badly fit by the model (off by more than 2 standard deviations), they should be dropped from the data set, although their elimination should be noted and addressed. Their aberrant behavior