STABILITY SYSTEM

ScienTek Software, Inc.

This short discussion is an excerpt from "A Crash Course on Chemical Kinetics and Statistical Data Treatment" that is part of the documentation for STABILITY SYSTEM program from ScienTek Software, Inc.

We have shown in the previous chapter how important kinetic parameters can be deduced from various equations. Most of these equations have a common characteristic that they are the equations of straight lines. Two such examples are the integrated rate equations for the zero-order reaction and the Arrhenius equation:

$$C = C_o - kt \tag{1}$$

$$\ln k = \ln A - \left(\frac{E}{R}\right)\left(\frac{1}{T}\right) \tag{2}$$

EQ(1) is a straight line relationship between C and t with a slope of $-k$ and an intercept of C_o. EQ(2) is a straight line relationship between $\ln k$ and $\frac{1}{T}$ with $-\frac{E}{R}$ as the slope and $\ln A$ as the intercept. Important kinetic parameters can be determined from the slope and/or intercept. In the first case, the rate constant can be calculated from the slope. In the second example, the activation energy and the pre-exponential coefficient can be calculated from the slope and the intercept respectively.

Regression Analysis

The statistics provides us a regression method to estimate the values of these parameters. A regression model is required to show how a variable varies in a systematic manner with another variable(s). In our discussion, the regression model is often provided by a particular theory. For example, EQ(1) is a mathematical expression derived from a zero-order rate law and shows how the concentration C varies with time t. In this case, C is a dependent variable, while time t is an independent variable.

The regression model for a straight line relationship is:

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, \ldots, n \tag{3}$$

where y_i and x_i are the values of the dependent variable and the independent variable in the ith trail, β_o and β_1 are the parameters (i.e., intercept and slope) to be estimated, and ε_i is a random error term. In our discussion, x_i's are known constants. For example, in EQ(1) the variable t is known without any variability. However, the dependent variable C is to be measured in some way and is therefore subjected to variation of some sort. This variation is reflected in the normally distributed random error term ε_i which has a mean zero and a constant variance of s^2 in the regression model (EQ(3)).

Least Squares Method

To find good estimates for the parameters β_o and β_1, we use the method of *least squares*. Specifically, we want to find estimators to be able to minimize the quantity Q:

$$Q = \sum \big(y_i - (b_o + b_1 x_i)\big)^2 \tag{4}$$

where b_o and b_1 are the *least squares estimators* for β_o and β_1 respectively. This is achieved by setting the partial derivatives of Q with respect to b_o and b_1 to zero:

$$\frac{\partial Q}{\partial b_o} = -2 \sum (y_i - b_o - b_1 x_i) = 0 \tag{5}$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum x_i (y_i - b_o - b_1 x_i) = 0 \tag{6}$$

EQ(5) and (6) can be rearranged to:

$$n b_o - b_1 \sum x_i = \sum y_i \tag{7}$$

$$b_o \sum x_i - b_1 \sum {x_i}^2 = \sum x_i y_i \tag{8}$$

These two equations are called the *normal equations.* Solving for b_o and b_1, we obtain:

$$b_o = \bar{y} - b_1 \bar{x} \tag{9}$$

where $\bar{x} = \dfrac{\sum x_i}{n}$ and $\bar{y} = \dfrac{\sum y_i}{n}$

$$b_1 = \frac{\sum x_i y_i - \dfrac{\sum x_i \sum y_i}{n}}{\sum {x_i}^2 - \dfrac{\sum {x_i}^2}{n}} \tag{10}$$

Therefore the estimated regression function is:

$$\hat{y}_i = b_o + b_1 x_i = \bar{y} + b_1 (x_i - \bar{x}) \tag{11}$$

For a specific x_i, $\hat{y}_i = b_o + b_1 x_i$ is a fitted value and y_i is the observed value.

The following equations are relationships important for the statistical treatment to be discussed:

residual:

$$e_i = y_i - \hat{y}_i \tag{12}$$

total sum of squares:

$$SSTO = \sum (y_i - \bar{y})^2 \tag{13}$$

regression sum of squares:

$$SSR = \sum (\hat{y}_i - \bar{y})^2 \tag{14}$$

error sum of squares:

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{15}$$

It can be shown that:

$$SSTO = SSR + SSE \tag{16}$$

The residuals measure the scatter of the observed y values around the regression line. When $e_i = 0$ for all i's, $\hat{y}_i = y_i$ and all observations fall on the fitted regression line, $SSE = 0$. Therefore SSE can be used as a measure of the variation of the observed y's around the regression line.

The variance (s^2) of the random error term is estimated by the mean square of error (MSE) which is calculated by:

$$MSE = \frac{SSE}{n-2} \tag{17}$$

where $n - 2$ is the *degree of freedom* associated with SSE.

Although in its most general sense, a regression function does not necessarily suggest a cause-and-effect relationship between y and x, it happens that all the situations discussed here do have such relationship supported by a theory. Therefore it is particularly important that we have some knowledge concerning the degree of association between y and x. Such association is measured by the *coefficient of determination* r^2:

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{18}$$

So defined, r^2 means the portion of the total variation explainable by the functional dependence of y on x.

From EQ(16), it is obvious that $0 \leq SSR \leq SSTO$. It follows that:

$$0 \leq r^2 \leq 1 \tag{19}$$

The two extreme cases are:

1. When $r^2 = 1$, $SSE = 0$. As we mentioned earlier, all observations fall on the fitted line and all variation in the y_i's is accounted for by x.

2. When $r^2 = 0$, $SSE = SSTO$ or $SSR = 0$. Combining EQs (11) and (14), we obtain:

$$SSR = {b_1}^2 \sum (x_i - \bar{x})^2 \tag{20}$$

Since $\sum (x_i - \bar{x})^2$ is positive, SSR is zero only when $b_1 = 0$. The term b_1 is the slope of the regression line. A zero slope means that we have a horizontal line and there is no relationship between y and x as modeled in EQ(3).

The use of the *correlation coefficient* r is quite popular:

$$r = \pm\sqrt{r^2} \tag{21}$$

where the sign is determined by the sign of the slope. A negative slope dictates a negative r value, while a positive slope indicates a positive r value. It is noted that since $|r| > r^2$, the use of r can be misleading.