Chemometrics in Spectroscopy
Comparison of Goodness of Fit Statistics for Linear Regression, Part IV

Part IV of the series describes the use of confidence limits in data analysis for calculating slope and intercept.

Jerome Workman Jr. and Howard Mark

In this column we continue to describe the use of confidence limits for comparison of *X*, *Y* data pairs. This is a continuation of a series of articles describing this subject (1–3). A MathCad Worksheet (MathSoft Engineering & Education, Inc., Cambridge, MA) provides the computations for interested readers. This will be covered in a subsequent column or can be obtained in MathCad format by contacting the authors with your e-mail address. The worksheet allows the direct calculation of the *t*-statistic by entering the desired confidence levels. In addition, the confidence limits for the calculated slope and intercept are computed from the original data table. The lower limits for the slope and intercept are displayed using two different sets of equations (and are identical). The intercept confidence limits also are calculated and displayed.

For calculations of slope and intercept two sets of equations will be shown, one as a summation notation set useful for application in MathCad software, and a second set as shown from Miller and Miller (4). For these formulas, *X* represents the concentration and *Y* represents the instrument response. This is to demonstrate that the two computational formula sets yield the same precise answer.

To begin, the following summation notation can be used to calculate the slope (k_1) of a linear regression line given a set of *X*, *Y* paired data (equation 1).

$$k_1 = \frac{n \cdot \sum(X \cdot Y) - \sum X \cdot \sum Y}{n \cdot \left(\sum X^2\right) - \left(\sum X\right)^2} \quad [1]$$

The summation notation formula for calculating the intercept (k_0) of a linear regression line given a set of *X*, *Y* paired data is as equation 2.

$$k_0 = \frac{\left(\sum X^2\right) \cdot \sum Y - \sum X \cdot \sum (X \cdot Y)}{n \cdot \left(\sum X^2\right) - \left(\sum X\right)^2} \quad [2]$$

Miller and Miller (4) use the following for the slope (b) calculation (equation 3).

$$b = \frac{\sum_i \{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum_i (x_i - \bar{x})^2} \quad [3]$$

The intercept (a) is given by the same authors (4) as equation 4.

$$a = \bar{y} - b\bar{x} \quad [4]$$

The reader might be surprised to learn that for the selected data the slope using either method computes to a value of 1.93035714285714, while the intercept for both methods of computation has values of 1.51785714285715 (summation notation method) versus 1.51785714285714 for the Miller and Miller cited method (however, this is the probable result of computational round-off error).

The confidence limits for the slope and intercept can be calculated using the Student's *t* statistic, noting equations 5–8 below. The slope (k_1) confidence limits are computed as equation 5 or equations 6–8.

$$\text{Limits} = k_1 \pm \left(\frac{t}{\sqrt{n-2}} \cdot \sqrt{\frac{\sum (Y - \hat{Y})^2}{\sum (X - \bar{X})^2}} \right) \quad [5]$$

Jerome Workman Jr. serves on the Editorial Advisory Board of *Spectroscopy* and is chief technical officer and vice president of research and engineering for Argose, Inc. (Waltham, MA). He can be reached by e-mail at workmans@rcn.com.

Howard Mark serves on the Editorial Advisory Board of *Spectroscopy* and runs a consulting service, Mark Electronics (69 Jamie Court, Suffern, NY 10901). He can be reached via e-mail at hlmark@prodigy.net.

Miller and Miller (4) cite equations 6–8 for calculation of the slope (b) confidence limits.

$$S_{y/x} = \left\{ \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} \right\}^{\frac{1}{2}} \quad [6]$$

$$s_b = \frac{s_{y/x}}{\left\{ \sum_i (x_i - \bar{x})^2 \right\}^{\frac{1}{2}}} \quad [7]$$

$$\text{Limits} = b \pm t \cdot s_b \quad [8]$$

As the reader might suspect by now, these methods of computation yield precisely the same answer as LL= 1.82521966597124; and UL = 2.03549461974305.

The intercept (k_0) confidence limits are computed as equation 9.

$$\text{Limits} = k_0 \pm t \sqrt{\frac{\sum (Y - \hat{Y})^2 \sum X^2}{(n-2)\left[n \sum (X - \bar{X})^2\right]}} \quad [9]$$

Miller and Miller (4) cite equations 10–12 for calculation of the intercept (a) confidence limits.

Again, the methods of computation shown yield precisely the same values for LL= 0.759700015087087; and

$$s_{y/x} = \left\{ \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} \right\}^{\frac{1}{2}} \quad [10]$$

$$s_a = s_{y/x} \left\{ \frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2} \right\}^{\frac{1}{2}} \quad [11]$$

$$\text{Limits} = a \pm t \cdot s_a \quad [12]$$

UL = 2.27601427062721. We will be discussing a more detailed interpretation for the slope and intercept confidence limits in later columns. However, the reader will note that the regression line for any X, Y paired data rotates at the epicenter point designated by the mean X and mean Y data point. Thus, the farther from the mean of X and Y a data point along a line occurs, the less the overall confidence in the relative position of the line. A more detailed description of the confidence limits surrounding any regression line will be discussed in a subsequent installment of "Chemometrics in Spectroscopy" using the F-distribution.

References

1. J. Workman and H. Mark, *Spectroscopy* **19**(4), 38–41 (2004).
2. J. Workman and H. Mark, *Spectroscopy* **19**(6) 29–33 (2004).
3. J. Workman and H. Mark, *Spectroscopy* **19**(7) 31–33 (2004).
4. J.C. Miller and J.N. Miller, *Statistics for Analytical Chemistry*, 2nd edition (Ellis Horwood, New York, 1992), pp. 100–111. ■