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

The authors continue their discussion of the correlation coefficient in developing a calibration for quantitative analysis.

Jerome Workman Jr. and Howard Mark

This column is a continuation of our previous column describing the use of goodness of fit statistical parameters (1). When developing a calibration for quantitative analysis one must select the analyte range over which the calibration is performed. For a given standard error of analysis the size of the range will have a direct effect on the magnitude of the correlation coefficient. The standard deviation of Y also has a direct effect, demonstrated by noting the computation for correlation between X and Y, in matrix notation, denoted as

\[ r = \frac{\text{covar}(X, Y)}{\text{stdev}(X) \cdot \text{stdev}(Y)} \]  

[1]

Note for this example that covar(X,Y) represents the covariance of (X,Y), stdev(X) is the standard deviation of the X data and stdev(Y) is the standard deviation of the Y data.

For the MathCad program (MathCad 2001i software, MathSoft Engineering & Education, Inc., Cambridge, MA), the stdev(X) is represented by the variable symbol Sr, which can be thought of as the set of many possible standard deviations for a set of data X.

Thus, a comparison of the correlation coefficient between two or more sets of X, Y data pairs cannot be performed adequately unless the standard deviations of the two data sets are nearly identical or unless the correlation coefficient confidence limits for the data sets are compared.

In summary, if Set A of X, Y paired data has a correlation of 0.95 this does not necessarily indicate that it is more highly correlated than Set B of X, Y paired data with a correlation of 0.90. The meaning of this will be described in greater detail later.

Let us look at seven slightly different equations (\( r_1 \) through \( r_7 \), or equations 6–12) for calculating correlation between X (known concentration or analyte data for a set of standards) and Y (instrument measured data for those standards) using MathCad function or summation notation nomenclature. First we must define the calculation of the standard error of performance, also termed the standard error of prediction (SEP), and the calculations for the slope (\( K_1 \)) and the intercept (\( K_0 \)) for the linear regression line between X and Y. The regression line for estimating the concentration denoted by (PredX or \( \hat{x} \)) is given as

\[ \text{PredX} = \hat{x} = K_1 Y + K_0 \]  

[2]

The standard error of performance, which represents an estimate of the prediction error (1 sigma) for a regression line, is given as

\[ \text{SEP} = \sqrt{\frac{\sum (\hat{x} - x)^2}{n}} \]  

[3]
The slope ($K_1$) and intercept ($K_0$) of the line for this regression line is given as

$$K_1 = \frac{n \cdot \sum (Y \cdot X) - \sum Y \cdot \sum X}{n \cdot \sum Y^2 - (\sum Y)^2} \quad [4]$$

$$K_0 = \frac{(\sum Y^2) \cdot \sum X - \sum Y \cdot \sum (Y \cdot X)}{n \cdot (\sum Y^2) - (\sum Y)^2} \quad [5]$$

The seven ways ($r_1$ through $r_7$) for calculating correlation as the square root of the ratio of the explained variation over the total variation between $X$ (concentration of analyte data) and $Y$ (measured data) are described using many notational forms. For example, many software packages provide built-in functions capable of calculating the coefficient of correlation directly from a pair of $X$ and $Y$ vectors as given by $r_1$ (equation 6). (This is the built-in MathCad correlation function.)

$$r_1 = \text{corr}(X,Y) \quad [6]$$

Several software packages contain simple command lines for performing matrix computations directly and thus are capable of conveniently computing the correlation coefficient, as shown in $r_2$ (equation 7).

$$r_2 = \frac{\text{covar}(X,Y)}{\text{stdev}(X) \cdot \text{stdev}(Y)} \quad [7]$$

Equation 7 denotes the ratio of the covariance of $X$ on $Y$ to the standard deviation of $X$ times the standard deviation of $Y$, where $X$ and $Y$ are vectors.

If the software is capable of using summation notation, then one can use this algebraic form for calculating the correlation as in $r_3$ and $r_4$ (equations 8 and 9, respectively).

$$r_3 = \frac{\sum (X - \bar{X})^2}{\sum (X - \bar{X})^2} \quad [8]$$

Equation 8 is the square root of the ratio comprising the sum of the squared differences between each predicted $X$ and the mean of all $X$, to the sum of the squared differences between all individual $X$ values and the mean of all $X$.
Equation 9 denotes the square root of one minus the ratio comprising the sum of the squared differences between each predicted $X$ and its corresponding $X$, to the sum of the squared differences between all individual $X$ values and the mean of all $X$.

And if the software allows you to assign variable names as needed for specific computations, such as standard error of performance or standard deviations, then you can proceed to use computational descriptions such as $r_5$ and $r_6$ (equations 10 and 11, respectively) to compute the correlation.

$$r_5 = \sqrt{1 - \frac{(\text{SEP}^2)}{(\text{stdevX})^2}} \quad [10]$$

Equation 10 indicates that the correlation coefficient is represented by the square root of one minus the ratio comprising the square of the standard error of performance, to the square of the standard deviation of all $X$.

$$r_6 = \sqrt{1 - \frac{(\text{SEP}^2)}{(\text{stdevX})^2}} \quad [11]$$

Equation 11 is simply the algebraic equivalent of the equation found above.

Other computational methods for correlation are given in reference 2, page 105 (as shown in equation 12).

$$r_7 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\left[ \left( \sum (x_i - \overline{x})^2 \right) \left( \sum (y_i - \overline{y})^2 \right) \right]^{1/2}} \quad [12]$$

You might be surprised that for our example data from reference 2, page 106, the correlation coefficient calculated using any of these methods of computation for the $r$-value is 0.9988795634852. When we evaluate the correlation computation we see that given a relatively equivalent prediction error represented by the standard error of performance, the standard deviation of the data set ($X$) determines the magnitude of the correlation coefficient. This is illustrated using Figures 1a and 1b. These graphics allow the correlation coefficient to be displayed for any specified standard error of prediction, also occasionally denoted as the standard error of estimate (SEE). It should be obvious that for any statistical study one must compare the actual computational...
recipes used to make a calculation, rather than rely on nonstandard terminology and assume that the computations are what one expected.

For a graphical comparison of the correlation ($r[Sr]$) and the standard deviation of the samples used for calibration ($Sr$), a value is entered for the standard error of performance for a specified analyte range as indicated through the standard deviation of that range. The resultant graphic displays $Sr$ (as the abscissa) versus $r$ (as the ordinate). From this graphic it can be seen how the correlation coefficient increases with a constant standard error of performance as the standard deviation of the data increases. Thus when comparing correlation results for analytical methods, one must consider carefully the standard deviation of the analyte values for the samples used in order to make a fair comparison. For the example shown, the standard error of estimate is set to 0.10, while the correlation is scaled from 0.0 to 1.0 for $Sr$ values from 0.10 to 4.0.

Figure 1b demonstrates the correlation range above 0.99 for the figure in Figure 1a. Note that the correlation begins to flatten when $Sr$ is over an order of magnitude times the standard error of the estimate.

Note from Figure 1c that at a certain value for standard deviation of $X$ (denoted as $Sr$), a small change in $Sr$ results in a large apparent change in the correlation. For example, in this case where the standard error of the estimate is set to 0.10, the correlation changes from 0.86 to 0.95 when $Sr$ is changed only from 0.20 to 0.32. As is the general case, using correlation to compare analytical methods requires identical sample analyte standard deviations, or comparison of the confidence limits for the correlation coefficients to interpret the significance of the different correlation values.

For a graphical comparison of the coefficient of determination ($R^2$) and $Sr$, a value is entered for the standard error of estimate for a specified range of $Sr$. The resultant graphic (Figure 2) displays $Sr$ (abscissa) versus $R^2$ (ordinate). From this graph it can be seen how the coefficient of determination increases as the standard deviation of the data. The standard error of estimate is set at 0.10 as in the examples shown in Figures 1a and 1b. Note that the same recommendation holds whether using $r$ or $R^2$—that relative comparisons for this statistic should not be used unless the standard deviations of the comparative data sets are identical.

Figure 3 shows the relative ratio of the range ($Sr$) to the standard error of estimate (abscissa) as compared with the correlation coefficient $r$ as the
ordinate. This graph shows that the correlation coefficient continues to increase as the ratio of $Sr$/SEE even when the ratio approaches more than 60. Note that when the ratio is greater than 10 there is not much improvement in the correlation.

A graphical comparison of $r$ versus the standard error of estimate is shown in Figure 4. This graphic clearly shows that when $Sr$ is held constant ($Sr = 4$) the correlation decreases as the standard error of estimate increases.

Figure 5 shows the relationship between correlation and the ratio of $SEE/Sr$, as the standard error of estimate increases relative to $Sr$ the correlation decreases rapidly.

We have introduced several common methods for calculating the correlation coefficient between a set of paired X and Y data. During this discussion we have shown that the absolute values for correlation are obviously quite dependent upon the standard deviation of the ranges for these data. Likewise, the magnitude of the standard error of performance (or standard error of estimate) is also important for correlation, which affects the correlation when its magnitude changes relative to the standard deviation (or range) of the data. Thus it is important that the data ranges be equivalent when simply comparing absolute values for correlation. In future columns, we will calculate confidence limits for comparing these statistical parameters, including considerations for varying sample size.

References


Note:
The authors have received some error notices regarding the recent series of columns that discussed derivatives; the results presented should therefore not be used without verification. Corrections will be published as the errors are verified and as the publishing schedule permits. The authors apologize for any inconvenience.