Regression and Calibration

Shaun Burke, RHM Technology Ltd, High Wycombe, Buckinghamshire, UK.

One of the most frequently used statistical methods in calibration is linear regression. This third paper in our statistics refresher series concentrates on the practical applications of linear regression and the interpretation of the regression statistics.

Calibration is fundamental to achieving consistency of measurement. Often calibration involves establishing the relationship between an instrument response and one or more reference values. Linear regression is one of the most frequently used statistical methods in calibration. Once the relationship between the input value and the response value (assumed to be represented by a straight line) is established, the calibration model is used in reverse; that is, to predict a value from an instrument response. In general, regression methods are also useful for establishing relationships of all kinds, not just linear relationships. This paper concentrates on the practical applications of linear regression and the interpretation of the regression statistics. For those of you who want to know about the theory of regression there are some excellent references (1–6).

For anyone intending to apply linear least-squares regression to their own data, it is recommended that a statistics/graphics package is used. This will speed up the production of the graphs needed to confirm the validity of the regression statistics. The built-in functions of a spreadsheet can also be used if the routines have been validated for accuracy (e.g., using standard data sets (7)).

What is regression?

In statistics, the term regression is used to describe a group of methods that summarize the degree of association between one variable (or set of variables) and another variable (or set of variables). The most common statistical method used to do this is least-squares regression, which works by finding the "best curve" through the data that minimizes the sums of squares of the residuals. The important term here is the "best curve", not the method by which this is achieved. There are a number of least-squares regression models, for example, linear (the most common type), logarithmic, exponential and power. As already stated, this paper will concentrate on linear least-squares regression.

[You should also be aware that there are other regression methods, such as ranked regression, multiple linear regression, non-linear regression, principal-component regression, partial least-squares regression, etc., which are useful for analysing instrument or chemically derived data, but are beyond the scope of this introductory text.]

What do the linear least-squares regression statistics mean?

Correlation coefficient: Whether you use a calculator's built-in functions, a spreadsheet or a statistics package, the first statistic most chemists look at when performing this analysis is the correlation coefficient (r). The correlation coefficient ranges from –1, a perfect negative relationship, through zero (no relationship), to +1, a perfect positive relationship (Figures 1(a–c)). The correlation coefficient is, therefore, a measure of the degree of linear relationship between two sets of data. However, the r value is open to misinterpretation (8) (Figures 1(d) and (e), show instances in which the r values alone would give the wrong impression of the underlying relationship). Indeed, it is possible for several different data sets to yield identical regression statistics (r value, residual sum of squares, slope and intercept), but still not satisfy the linear assumption in all cases (9). It, therefore, remains essential to plot the data in order to check that linear least-squares statistics are appropriate.

As in the *t*-tests discussed in the first paper (10) in this series, the statistical significance of the correlation coefficient is dependent on the number of data points. To test if a particular r value indicates a statistically significant relationship we can use the Pearson's correlation coefficient test (Table 1). Thus, if we only have four points (for which the number of degrees of freedom is 2) a linear least-squares correlation coefficient of –0.94 will not be significant at the 95% confidence level. However, if there are more than 60 points an r value of just 0.26 (r^2 = 0.0676) would indicate a significant, but not very strong, positive linear relationship. In other words, a relationship can be statistically significant but of no practical value. Note that the test used here simply shows whether two sets are linearly related; it does not "prove" linearity or adequacy of fit.

It is also important to note that a significant correlation between one variable and another should not be taken as an indication of causality. For example, there is a negative correlation between time (measured in months) and catalyst performance in car exhaust systems. However, time is not the cause of the deterioration, it is the build up of sulfur and phosphorous compounds that gradually poisons the catalyst. Causality is,

in fact, very difficult to prove unless the chemist can vary systematically and independently all critical parameters, while measuring the response for each change.

Slope and intercept

In linear regression the relationship between the X and Y data is assumed to be represented by a straight line, Y = a + bX (see Figure 2), where Y is the estimated response/dependent variable, b is the slope (gradient) of the regression line and a is the intercept (Y value when X = 0). This straight-line model is only appropriate if the data approximately fits the assumption of linearity. This can be tested for by plotting the data and looking for curvature (e.g., Figure 1(d)) or by plotting the residuals against the predicted Y values or X values (see Figure 3).

Although the relationship may be known to be non-linear (i.e., follow a different functional form, such as an exponential curve), it can sometimes be made to fit the linear assumption by transforming the data in line with the function, for example, by taking logarithms or squaring the Y and/or X data. Note that if such transformations are performed, weighted regression (discussed later) should be used to obtain an accurate model. Weighting is required because of changes in the residual/error structure of the regression model. Using non-linear regression may, however, be a better alternative to transforming the data when this option is available in the statistical packages you are using.

Residuals and residual standard error

A residual value is calculated by taking the difference between the predicted value and the actual value (see Figure 2). When the residuals are plotted against the predicted (or actual) data values the plot becomes a powerful diagnostic tool, enabling patterns and curvature in the data to be recognized (Figure 3). It can also be used to highlight points of influence (see Bias, leverage and outliers overleaf).

The residual standard error (RSE, also known as the residual standard deviation, RSD) is a statistical measure of the average residual. In other words, it is an estimate of the average error (or deviation) about the regression line. The RSE is used to calculate many useful regression statistics including confidence intervals and outlier test values.

$$RSE = s_{(y)}\sqrt{\frac{(n-1)}{(n-2)}(1-r^2)}$$

where $s_{(y)}$ is the standard deviation of the y values in the calibration, n is the number of data pairs and r is the least-squares regression correlation coefficient.

Confidence intervals

As with most statistics, the slope (b) and intercept (a) are estimates based on a finite sample, so there is some uncertainty in the values. *(Note: Strictly, the uncertainty arises from random variability between sets of data. There may be other uncertainties, such as measurement bias, but these are outside the scope of this article.)* This uncertainty is quantified in most statistical routines by displaying the confidence limits and other statistics, such as the standard error and p values. Examples of these statistics are given in Table 2.

Degrees of freedom	Confidence level	
(n-2)	95% (α = 0.05)	99% (α = 0.01)
2	0.950	0.990
3	0.878	0.959
4	0.811	0.917
5	0.754	0.875
6	0.707	0.834
7	0.666	0.798
8	0.632	0.765
9	0.602	0.735
10	0.576	0.708
11	0.553	0.684
12	0.532	0.661
13	0.514	0.641
14	0.497	0.623
15	0.482	0.606
20	0.423	0.537
30	0.349	0.449
40	0.304	0.393
60	0.250	0.325

Significant correlation when |r| ≥ table value

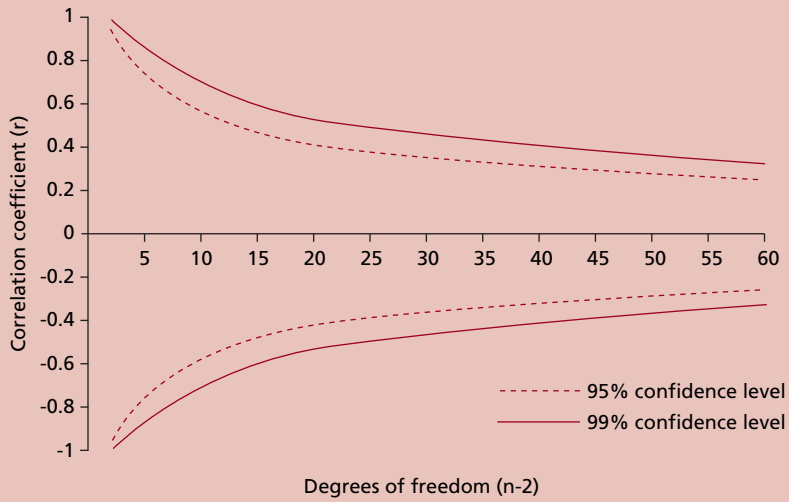

table 1 *Pearson's correlation coefficient test.*

The p value is the probability that a value could arise by chance if the true value was zero. By convention a p value of less than 0.05 indicates a significant non-zero statistic. Thus, examining the spreadsheet's results, we can see that there is no reason to reject the hypothesis that the intercept is zero, but there is a significant non-zero positive gradient/relationship. The confidence intervals for the regression line can be plotted for all points along the x-axis and is dumbbell in shape (Figure 2). In practice, this means that the model is more certain in the middle than at the extremes, which in turn has important consequences for extrapolating relationships.

When regression is used to construct a calibration model, the calibration graph is used in reverse (i.e., we predict the X value from the instrument response [Y-value]). This prediction has an associated uncertainty (expressed as a confidence interval)

$$X_{predicted} = \left(\frac{(\bar{Y} - a)}{b}\right)$$

Conf. interval for the prediction is: $X_{predicted} \pm \left(\frac{t(RSE)}{b}\right)\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{Y}-\bar{y})^2}{b^2(n-1)\,s_{(x)}^2}}$

where a is the intercept and b is the slope obtained from the regression equation. $\bar{Y}$ is the mean value of the response (e.g., instrument readings) for m replicates (replicates are repeat measurements made at the same level).
$\bar{y}$ is the mean of the y data for the n points in the calibration. *t* is the critical value obtained from *t*-tables for n–2 degrees of freedom. $s_{(x)}$ is the standard deviation for the x data for the n points in the calibration. RSE is the residual standard error for the calibration.

If we want, therefore, to reduce the size of the confidence interval of the prediction there are several things that can be done.

1. Make sure that the unknown determinations of interest are close to the centre of the calibration (i.e., close to the values $\bar{x},\bar{y}$ [the centroid point]). This suggests that if we want a small confidence interval at low values of x then the standards/reference samples used in the calibration should be concentrated around this region. For example, in analytical chemistry, a typical pattern of standard concentrations might be 0.05, 0.1, 0.2, 0.4, 0.8, 1.6

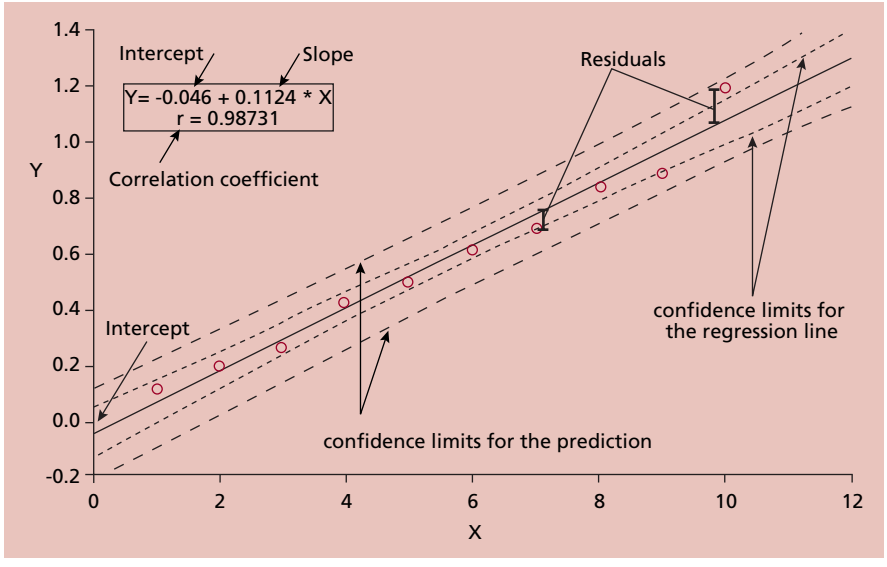

figure 2 *Calibration graph.*

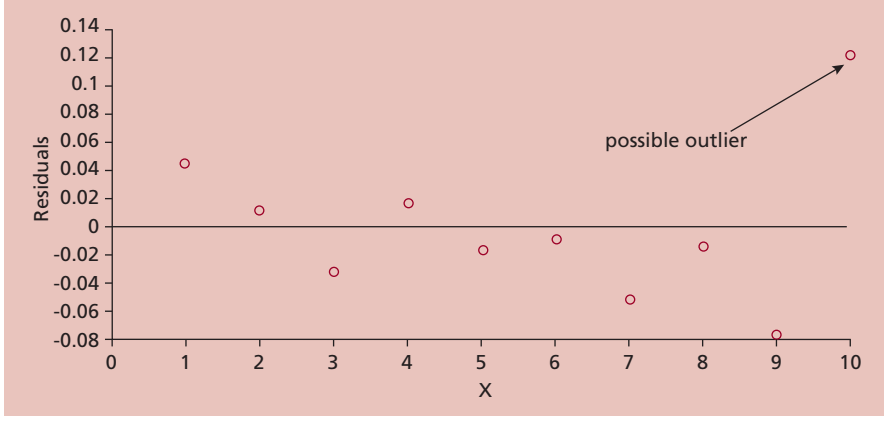

figure 3 *Residuals plot.*

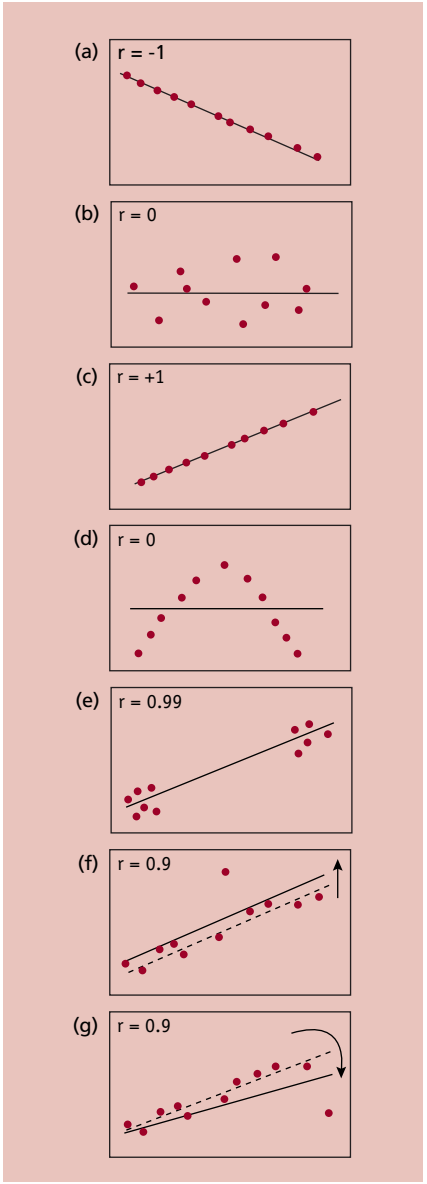

figure 1 *Correlation coefficients and goodness of fit.*

(i.e., only one or two standards are used at higher concentrations). While this will lead to a smaller confidence interval at lower concentrations the calibration model will be prone to leverage errors (see below).
2. Increase the number of points in the calibration (n). There is, however, little improvement to be gained by going above 10 calibration points unless standard preparation and analysis is rapid and cheap.
3. Increase the number of replicate determinations for estimating the unknown (m). Once again there is a law of diminishing returns, so the number of replicates should typically be in the range 2 to 5.
4. The range of the calibration can be extended, providing the calibration is still linear.

Bias, leverage and outliers

Points of influence, which may or may not be outliers, can have a significant effect on the regression model and therefore, on its predictive ability. If a point is in the middle of the model (i.e., close to $\bar{x}$) but outlying on the Y axis, its effect will be to move the regression line up or down. The point is then said to have influence because it introduces an offset (or bias) in the predicted values (see Figure 1(f)). If the point is towards one of the extreme ends of the plot its effect will be to tilt the regression line. The point is then said to have high leverage because it acts as a lever and changes the slope of the regression model (see Figure 1(g)). Leverage can be a major problem if one or two data points are a long way from all the other points along the X axis.

A leverage statistic (ranging between $\frac{1}{n}$ and 1) can be calculated for each value of x. There is no set value above which this leverage statistic indicates a point of influence. A value of 0.9 is, however, used by some statistical software packages.

$$\text{Leverage}_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

where x_j is the x value for which the leverage statistic is to be calculated, n is the number of points in the calibration and $\bar{x}$ is the mean of all the x values in the calibration.

To test if a data point (x_i, y_i) is an outlier (relative to the regression model) the following outlier test can be applied.

$$\text{Test value} = \frac{|\text{residual}_{max}|}{\text{RSE}\sqrt{1 + \frac{1}{n} + \frac{(Y_i - \bar{y})^2}{(n-1)\,s_y^2}}}$$

where RSE is the residual standard error, s_y is the standard deviation of the Y values, Y_i is the y value, n is the number of points, $\bar{y}$ is the mean of all the y values in the calibration and residual_{max} is the largest residual value.

For example, the test value for the suspected outlier in Figure 3 is 1.78 and the critical value is 2.37 (Table 3 for 10 data points). Although the point appears extreme, it could reasonably be expected to arise by chance within the data set.

Extrapolation and interpolation

We have already mentioned that the regression line is subject to some uncertainty and that this uncertainty becomes greater at the extremes of the line. If we, therefore, try to extrapolate much beyond the point where we have real data (±10%) there may be relatively large errors associated with the predicted value. Conversely, interpolation near the middle of the calibration will minimize the prediction uncertainty. It follows, therefore, that when constructing a calibration graph, the standards should cover a larger range of concentrations than the analyst is interested in. Alternatively, several calibration graphs covering smaller, overlapping, concentration ranges can be constructed.

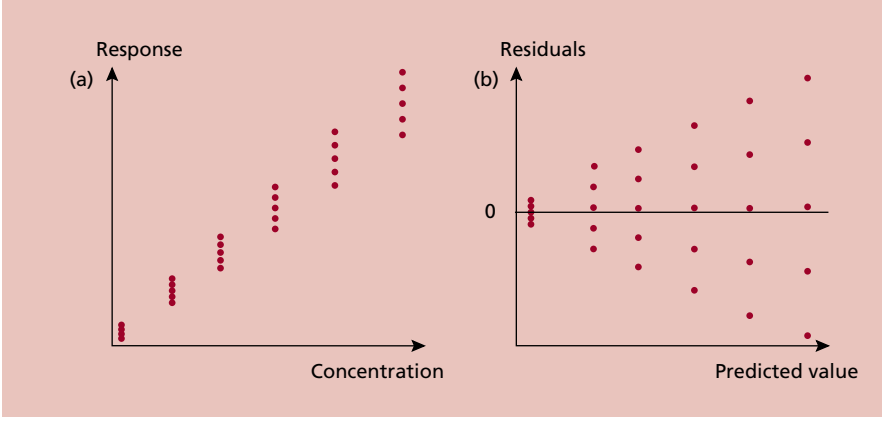

figure 4 *Plots of typical instrument response versus concentration.*

	Coefficients	Standard Error	*t* Stat	p value	Lower 95%	Upper 95%
Intercept	-0.046000012	0.039648848	-1.160185324	0.279423552	-0.137430479	0.045430455
Slope	0.112363638	0.00638999	17.58432015	1.11755E-07	0.097628284	0.127098992

**Note the large number of significant figures. In fact none of the values above warrant more than 3 significant figures!*

table 2 *Statistics obtained using Excel 5.0 regression analysis function from the data used to generate the calibration graph in Figure 2.*

Weighted linear regression and calibration

In analytical science we often find that the precision changes with concentration. In particular, the standard deviation of the data is proportional to the magnitude of the value being measured, (see Figure 4(a)). A residuals plot will tend to show this relationship even more clearly (Figure 4(b)). When this relationship is observed (or if the data has been transformed before regression analysis), weighted linear regression should be used for obtaining the calibration curve (3). The following description shows how the weighted regression works. Don't be put off by the equations as most modern statistical software packages will perform the calculations for you. They are only included in the text for completeness.

Weighted regression works by giving points known to have a better precision a higher weighting than those with lower precision. During method validation the way the standard deviation varies with concentration should have been investigated. This relationship can then be used to calculate the initial weightings

Sample size	Confidence table-value	
(n)	95%	99%
5	1.74	1.75
6	1.93	1.98
7	2.08	2.17
8	2.20	2.23
9	2.29	2.44
10	2.37	2.55
12	2.49	2.70
14	2.58	2.82
16	2.66	2.92
18	2.72	3.00
20	2.77	3.06
25	2.88	3.25
30	2.96	3.36
35	3.02	3.40
40	3.08	3.43
45	3.12	3.47
50	3.16	3.51
60	3.23	3.57
70	3.29	3.62
80	3.33	3.68
90	3.37	3.73
100	3.41	3.78

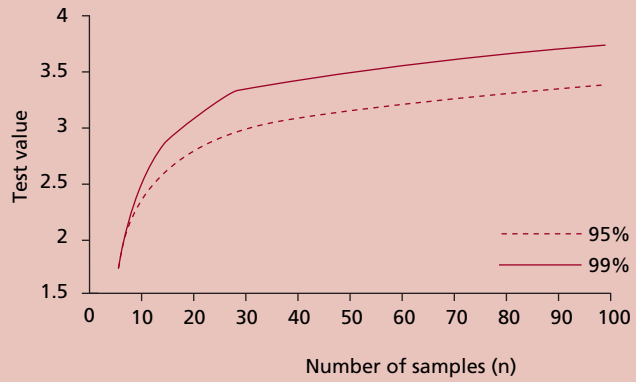

table 3 *Outlier test for simple linear least-squares regression.*

$(w_i = \frac{1}{s_i^2})$ at each of the n concentrations in the calibration. These initial weightings can then be standardized by multiplying by the number of calibration points divided by the sum of all the weights to give the final weights (W_i).

$$W_i = w_i\left(\frac{n}{\sum_{j=1}^{n} w_j}\right)$$

The regression model generated will be similar to that for non-weighted linear regression. The prediction confidence intervals will, however, be different.

The weighted prediction (x_w) for a given instrument reading (y) for the regression model forcing the line through the origin (y = bx) is:

$$X_{(w)predicted} = \left(\frac{\bar{Y}}{b_{(w)}}\right)$$

with

$$b_{(w)} = \frac{\sum_{i=1}^{n} W_i x_i y_i}{\sum_{i=1}^{n} W_i x_i^2}$$

where $\bar{Y}$ is the mean value of the response (e.g., instrument readings) for m replicates and x_i and y_i are the data pair for the *i*th point.

By assuming the regression line goes through the origin a better estimate of the slope is obtained, providing that the assumption of a zero intercept is correct. This may be a reasonable assumption in some instrument calibrations. However, in most cases, the regression line will no longer represent the least-squares "best line" through the data.

The associated uncertainty for the weighted prediction, expressed as a confidence interval is then:

Conf. interval for the prediction is

$$X_{(w)_{predicted}} \pm \left[\frac{t(RSE_{(w)})}{b_{(w)}}\right] \sqrt{\frac{1}{mW_i} + \frac{\bar{Y}^2}{b^2_{(w)} \sum_{j=1}^{n} W_j x_j^2}}$$

where t is the critical value obtained from t tables for n–2 degrees of freedom at a stated significance level (typically a = 0.05), W_i is the weighted standard deviation for the x data for the ith point in the calibration, m is the number of replicates and the weighted residual.

Standard error for the calibration $$RSE_{(w)} = \sqrt{\frac{\sum_{j=1}^{n} W_j y_j^2 - b^2_{(w)} \sum_{j=1}^{n} W_j x_j^2}{n-1}}$$

Conclusions

- Always plot the data. Don't rely on the regression statistics to indicate a linear relationship. For example, the correlation coefficient is not a reliable measure of goodness-of-fit.
- Always examine the residuals plot. This is a valuable diagnostic tool.
- Remove points of influence (leverage, bias and outlying points) only if a reason can be found for their aberrant behaviour.
- Be aware that a regression line is an estimate of the "best line" through the data and that there is some uncertainty associated with it. The uncertainty, in the form of a confidence interval, should be reported with the interpolated result obtained from any linear regression calibrations.

Acknowledgement

The preparation of this paper was supported under a contract with the Department of Trade and Industry as part of the National Measurement System Valid Analytical Measurement Programme (VAM) (11).

References

(1) G.W. Snedecor and W.G. Cochran, *Statistical Methods*, The Iowa State University Press, USA, 6th edition (1967).
(2) N. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons Inc., New York, USA, 2nd edition (1981).
(3) BS ISO 11095: Linear Calibration Using Reference Materials (1996).
(4) J.C. Miller and J.N. Miller, *Statistics for Analytical Chemistry*, Ellis Harwood PTR Prentice Hall, London, UK.
(5) A.R. Hoshmand, *Statistical Methods for Environmental and Agricultural Sciences*, 2nd edition, CRC Press (ISBN 0-8493-3152-8) (1998).
(6) T.J. Farrant, *Practical Statistics for the Analytical Scientist, A Bench Guide*, Royal Society of Chemistry, London, UK (ISBN 0 85404 4226) (1997).
(7) Statistical Software Qualification: Reference Data Sets, Eds. B.P. Butler, M.G. Cox, S.L.R. Ellison and W.A. Hardcastle, Royal Society of Chemistry, London, UK (ISBN 0-85404-422-1) (1996).
(8) H. Sahai and R.P. Singh, *Virginia J. Sci.*, **40**(1), 5–9, (1989).
(9) F.J. Anscombe, *Graphs in Statistical Analysis, American Statistician*, **27**, 17–21, February 1973.
(10) S. Burke, *Scientific Data Management*, **1**(1), 32–38, September 1997.
(11) M. Sargent, VAM Bulletin, Issue 13, 4–5, Laboratory of the Government Chemist (Autumn 1995).

Shaun Burke currently works in the Food Technology Department of RHM Technology Ltd, High Wycombe, Buckinghamshire, UK. However, these articles were produced while he was working at LGC, Teddington, Middlesex, UK (http://www.lgc.co.uk).