Notes on a linear regression calibration curve

The ability of instrumental analysis techniques to handle a wide range of analyte concentrations against the instrument responses indicates that the results of the sample concentration can be calculated, and their random errors evaluated, in a particular manner that differ from that used when a single measurement is repeated several times.

One may prepare a series of prepared standard solutions (normally at least 5 or 6, and possibly several more) in which the concentrations of the analyte are known. These calibration standards are measured in the analytical instrument under the same conditions as those subsequently used for the test (i.e. unknown) samples. A calibration graph is thus established between the known concentrations and the signals of the instrument.

In this respect, we can ask several statistical questions:

a) Is the calibration graph *linear*? If it is a curve, what is the form of the curve? In fact, although we would like to have a linearity in our calibration curve in order to minimize errors, we should always ask if our calibration plot obtained is *really* linear? In other words, is there any *lack-of-fit* of the linearity?

b) Bearing in mind that each of the points on the calibration graph is subjected to errors, what is the best straight line (or curve) through these points?

c) Assuming that the calibration plot is actually linear, what are the estimated errors and *confidence limits* for the slope (gradient) and the y-intercept of the line?

d) When the calibration plot is used for the analysis of a prepared test sample solution, what are the errors and confidence limits for the determined concentration?

e) What is the *limit of detection* of the method? That is, what is the least concentration of the analyte that can be detected with a predetermined level of confidence?

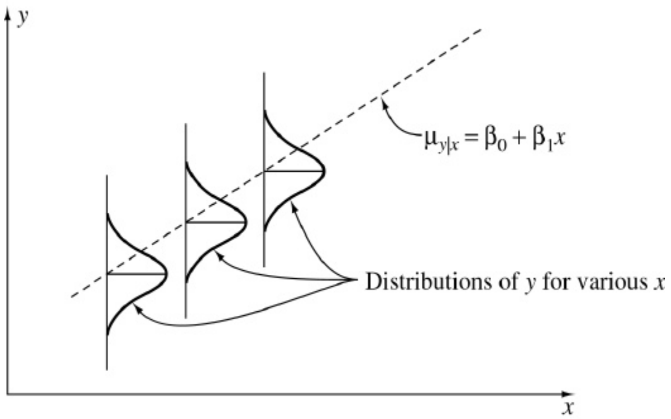
Before tackling these questions in detail, we must consider a number of aspects of plotting calibration graphs:

a) it is essential that the calibration standards cover the whole range of concentrations required in the subsequent analyses. With the ***important exception*** of the ‘standard addition method’, which is treated separately, concentrations of test samples are normally determined by **interpolation** and *not* by extrapolation.

b) it is critically important to include the value for a ‘reagent blank’ sample in the calibration curve. The blank contains no deliberately added analyte, but does contain the same solvent, reagents used for the analysis, etc. This blank must be treated as a test sample and subjected to exactly the same sequence of analytical procedures. The instrument signal given by the reagent blank sample will often not be zero. It is of course subject to errors, like all the other points on the calibration plot.

   Hence, it is wrong in principle to subtract the blank value from the other standard signal values before plotting the calibration graph.

c) The calibration curve is always plotted with the instrumental response on the vertical (y-) axis and the standard concentrations on the horizontal (x-) axis. This is because many of the procedures assume that all the errors are in the y-values and the standard concentrations (x-values) are having comparatively much smaller errors that can be *ignored*. Hence, any result uncertainty is deemed to have come only from the normal distribution of the y-values obtained from the instrument concerned.


y
$\mu_{y|x} = \beta_0 + \beta_1 x$
Distributions of y for various x
x

A straight-line plot takes the algebraic form:

$$y \quad = \quad a + bx \qquad \ldots [1]$$

where b = gradient of the slope
and a = intercept point at y-axis (i.e. when x = 0),

The individual points on the line are referred to as (x_1, y_1) (normally a set of the ‘blank’ values), (x_2, y_2), (x_3, y_3), …., (x_n, y_n), i.e. there are n points altogether plotted on a calibration curve.

The line of regression of y on x is normally prepared by the **least square method**. To do so, we first assume that there is a linear relationship between the analytical signals (y_i) and the various working standard concentrations (x_i). Let’s see how we can calculate the ‘*best*’ straight line through the calibration graph points, each of which is subject to experimental error.

In here, we are making an important assumption that all the errors are in y only (i.e. instrument responses) and so, we are seeking the line that minimises the deviations of the y-direction between the experimental points and the calculated line. As some of these deviations (technically known as the y-*residuals*) will be positive and some negative, it is sensible to seek to *minimise* the **sum of the squares of the residuals (or errors).**

Hence, for a given value of x_i, we can have two types of y values:

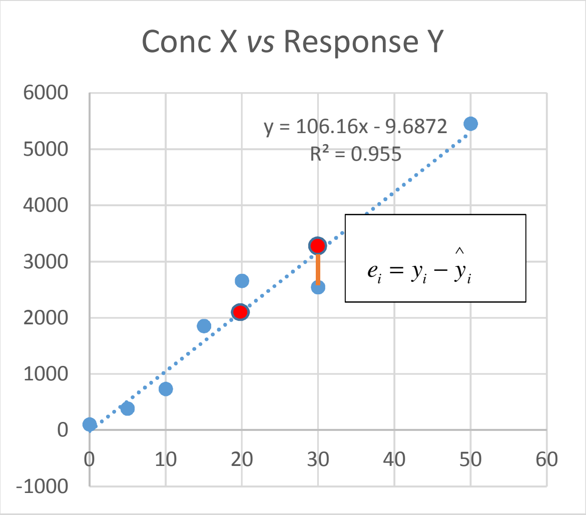
y_i (observed or experimental value of y at i th point), and

$\hat{y}_i$ (calculated value of y_i from the equation $y = a + bx$)

If we define a residual or error as the distance between the experimental value of y_i and the calculated value of y_i, i.e.

$$e_i = y_i - \hat{y}_i \qquad \text{... [2]}$$

we notice that the sum of all the e_i, $\Sigma e_i = 0$, indicating that the straight line has the least deviation from all the plot points considered by moving in between them. In the following graph with exaggerated spread of the points on the curve, we can see that all the e_i's will be either positive or negative and the total sum should be zero if it is the best fit line.



The ***sum of squares of the residuals or errors,*** SSE is :

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \qquad \text{... [3]}$$

Now, the ***mean-square error,*** MSE is given by the SSE divided by its number of degrees of freedom (n-2), i.e. :

$$MSE = \frac{SSE}{n-2} \qquad \text{… [4]}$$

The ***standard error of the y-estimate*** *SE(y)* or standard deviation of the regression is defined as **the square root of** *MSE*:

$$SE(y) = \sqrt{\frac{SSE}{n-2}} \qquad \text{… [5]}$$

In fact, *SE(y)* is an important formula in the study of error for linear regression. This can be seen in the following series of statistical equations in various forms:

1. The Pearson's linear correlation coefficient, *r* is given by:

$$r = b\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{\sum_{i=1}^{n}(y_i - y)}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

2. The gradient of the curve or slope is :

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

with standard deviation

$$SD(b) = \frac{SE(y)}{\sqrt{\sum x_i^{\,2} - \left[\left(\sum x_i\right)^2 / n\right]}}$$

We report a 95% confidence level of the gradient as $b \pm t_{(n-2)}\, SD(b)$ with a *df* of *(n-2)*.

3. *a*, the *y*-intercept of the line is given by:

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \overline{y} - b\overline{x}$$

with standard deviation *SD(a)* as:

$$SD(a) = SE(y)\sqrt{\frac{\sum x_i^{\,2}}{n\sum x_i^{\,2} - \left(\sum x_i\right)^2}}$$

4. Commonly in the calibration curve, we use the measured y-value to estimate an x-value. Once the slope (gradient) and intercept of a straight line calibration curve has been established, it is easy to calculate an x-value (e.g. observed value) from a measured y-value (e.g. instrument response value).

The 95% confidence limits of the estimated x-value are given by the following equation:

$$x_{meas} \pm t_{(n-2)} \frac{1}{b} SE(y)\sqrt{\frac{1}{R} + \frac{1}{n} + \frac{(y_{meas} - \bar{y})^2}{b^2 \sum (x_i - \bar{x})^2}}$$

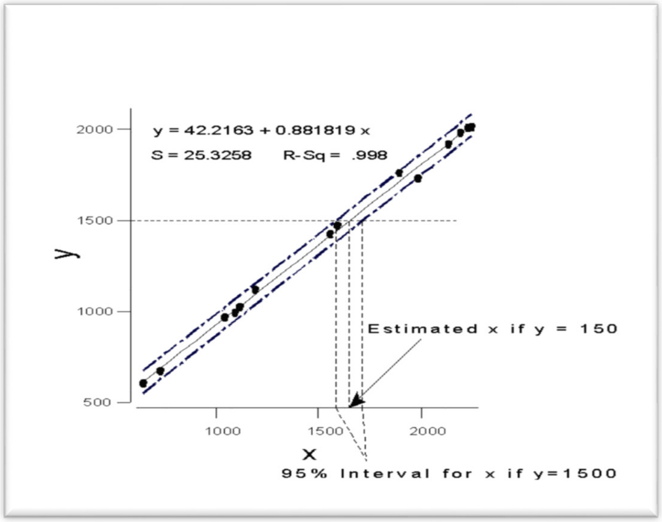
where R is the number of repeated measurements of y. If there is only one measurement, we then have

$$x_{meas} \pm t_{(n-2)} \frac{1}{b} SE(y)\sqrt{1 + \frac{1}{n} + \frac{(y_{meas} - \bar{y})^2}{b^2 \sum (x_i - \bar{x})^2}}$$

It is to be noted that the following assumptions have been made for this linear correlation, namely:

- Normality of error (i.e. random error, following the Normal distribution function)
- Homoscedasticity (which requires that the variation around the line of regression be constant for all values of x, i.e. the error varies by the same amount when x is a low value as when x is a high value.)
- Independence of errors for x.

The graph below illustrates the confidence limits on an example of calibration curve:



The next article will show why we need to prepare a suitable range of working standard solutions for a calibration curve in order to minimize the error of the output, i.e. the test results of our unknown sample solution. For example, if we were to have prepared a wide range of standard metal concentrations from 10 µg/L to 500 µg/L for a metal analyte measurement of only 20 µg/L in a water sample solution by the ICP technique, the uncertainty of this measurement is going to be significantly large.