Calibration and Linear Regression Analysis: A Self-Guided Tutorial

Part 2 – The Calibration Curve, Correlation Coefficient and Confidence Limits

---

CHM314 Instrumental Analysis

Department of Chemistry, University of Toronto

Dr. D. Stone (*prepared by J. Ellis*)

1 The Calibration Curve and Correlation Coefficient

Every instrument used in chemical analysis can be characterised by a specific response function, that is an equation relating the instrument output signal (S) to the analyte concentration (C). This response function may be linear, logarithmic, exponential, or any other appropriate mathematical form, depending on the nature of the behaviour of the system being measured, and the measurement process itself. While this may be known theoretically, various factors (such as the specific analyte being measured, interference effects caused by other components of the sample matrix, or random experimental errors) require that we calibrate each instrument for the specific analyte and measurement conditions to be used in a particular experiment.

A calibration curve is an empirical equation that relates the response of a specific instrument to the concentration of a specific analyte in a specific sample matrix (the chemical background of the sample). As with the instrument response function, the calibration curve can have a number of mathematical forms, depending on the type of measurement being performed. Some common examples are listed below:

Type	Equation
Linear (zero intercept)	$S = bC$
Linear (non-zero intercept)	$S = bC + a$
Logarithmic	$S = a + b \ln C$ or $S = a + 2.303b \log C$

The calibration curve is obtained by fitting an appropriate equation to a set of experimental data (calibration data) consisting of the measured responses to known concentrations of analyte. For example, in molecular absorption spectroscopy, we expect the instrument response to follow the Beer-Lambert equation, $A = \varepsilon bC$, and so we would fit a linear equation with zero intercept to the data. On the other hand, if we were measuring electrochemical cell potentials (i.e. potentiometry) we would expect the response to be given by the Nernst equation, which is logarithmic in form. We would therefore either fit a logarithmic equation to the calibration data, or linearise the data by calculating the signal response S as 10^E (where E is the cell potential).

The most common response function encountered in instrumental analytical chemistry is linear, so we require some means of determining and qualifying the best-fit straight line through our calibration data. Before discussing this in detail, however, a word of caution: even when we expect a linear instrument response function, we should **not** assume that the calibration data must always be linear. In fact, a moment of reflection reveals that we already know that this cannot be true. For example, stray light and polychromatic radiation cause non-linear deviations from Beer's law at higher concentrations; quenching and self-absorption can cause fluorescence intensities to start decreasing with increasing concentration; and column- or detector-overload can cause non-linearities in chromatography.

1.1 The Correlation Coefficient

In Part 1 of the tutorial, we saw how to use the trendline feature in Excel to fit a straight line through calibration data and obtain both the equation of the best-fit straight line and the correlation coefficient, R (sometimes displayed as R^2). There are in fact various correlation coefficients, but the one we are interested in here is the Pearson or product-moment correlation coefficient (often simply referred to as the "correlation coefficient"). The Pearson R value provides a measure of the degree to which the values of x and y are linearly correlated. We can assess this visually using a scatter plot (Figure 1), in which we also mark the centroid of the data, $\{\bar{x},\bar{y}\}$.

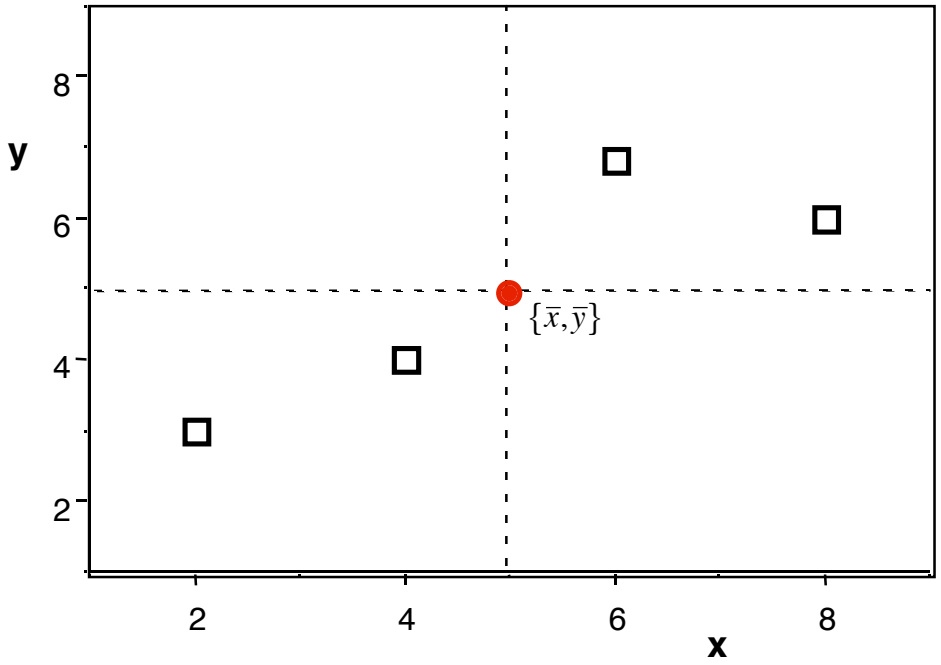

Figure 1 – *XY scatter plot showing the centroid of the data*

If x and y were linearly correlated, we would expect all the points to fall on a straight line passing through the centroid. As a result, we would expect all x values to be uniformly distributed either side of $\bar{x}$; similarly, all the y values should be uniformly distributed about $\bar{y}$. The Pearson R is calculated using the formula

$$R = \frac{\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\left[\sum_i (x_i - \bar{x})^2\right]\left[\sum_i (y_i - \bar{y})^2\right]}}$$

It follows that if x and y are perfectly correlated in a linear fashion, we would expect the value of R to be either +1 or -1, depending on whether y increases (positive slope) or decreases (negative slope) with x.

To demonstrate how to calculate this formula in Excel, we return to our previous example of fluorescence intensity data from Part 1. Then,

1. Set up a spreadsheet with the x_i and y_i values in columns

2. In the adjacent cells, set up expressions for $(x_i - \bar{x})$, $(y_i - \bar{y})$, their squares, and their product. For instance, the formula for $(x_2 - \bar{x})$ may look like "=B3-AVERAGE(B$2:B$8)," depending on the location of your cells in the spreadsheets.
3. Determine the sums of squares $\sum_i (x_i - \bar{x})^2$ and $\sum_i (y_i - \bar{y})^2$, and the sum of products $\sum_i [(x_i - \bar{x})(y_i - \bar{y})]$ in Excel and insert these values in the formula for R.
4. To calculate the square root in the denominator, use the SQRT function.

The easiest way to calculate *R* in Excel is by setting up a table to calculate the required values, as shown below. As you can see this, yields a correlation coefficient $R^2 = 0.9978$, so the data are well-correlated and the best-fit line describes the data.

Book1

	A	B	C	D	E	F	G	H
1	Fluorescence Instensities	Concentration (pg/ml)		$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
2	2.1	0		-6	-11	36	121	66
3	5	2		-4	-8.1	16	65.61	32.4
4	9	4		-2	-4.1	4	16.81	8.2
5	12.6	6		0	-0.5	0	0.25	0
6	17.3	8		2	4.2	4	17.64	8.4
7	21	10		4	7.9	16	62.41	31.6
8	24.7	12		6	11.6	36	134.56	69.6
9								
10					Sum	112	418.28	216.2
11								
12					R =	=H10/SQRT(F10*G10)		
13					R^2 =	0.9977604		

A few points to mention regarding the correlation coefficient:

- It is essential to retain a large number of significant figures in the numerator and denominator during the calculation, otherwise a misleading value of *R* may be obtained.
- Even a high *R* value of, say, 0.9991 does not necessarily indicate that the data fits to a straight line. The trendline should always be plotted and inspected visually. R^2 is more discriminating in this respect, although it no longer indicates the slope of the regression line. This, however, is evident by inspection.
- Any curvature in the data will result in erroneous conclusions about the correlation. *R* values are only applicable to linear correlations. Nonlinear correlations are possible, but involve a different measure than *R*, and *R* values will not necessarily be close to 1.
- The statistical significance of *R* depends on the number of samples in the data set *n*.

1.2 The Regression Line

Calculation of the regression line is straightforward. The equation will have the form $y = bx + a$, where *b* is the slope of the line and *a* is the *y*-intercept. The slope is given by the formula

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

and the intercept is

$$a = \bar{y} - b\bar{x},$$

both of which can be easily calculated in Excel with the table of data used in the previous section. The method is similar to that in the previous section. The AVERAGE function can be used to calculate $\bar{x}$ and $\bar{y}$. Using the fluorescence data, the equation of the line is $y = 1.930x + 1.518$.

Figure 2 shows an example of a regression line with the calibration data, centroid and *y*-residuals displayed. Note that, as is commonly the case, it is assumed that any error in the data lies solely in the *y*-values. Technically, the best-fit straight line shown is termed the 'line of regression of *y* on *x*'. This method for linear regression assumes that the errors are normally distributed. Other methods exist that do not make this type of assumption.

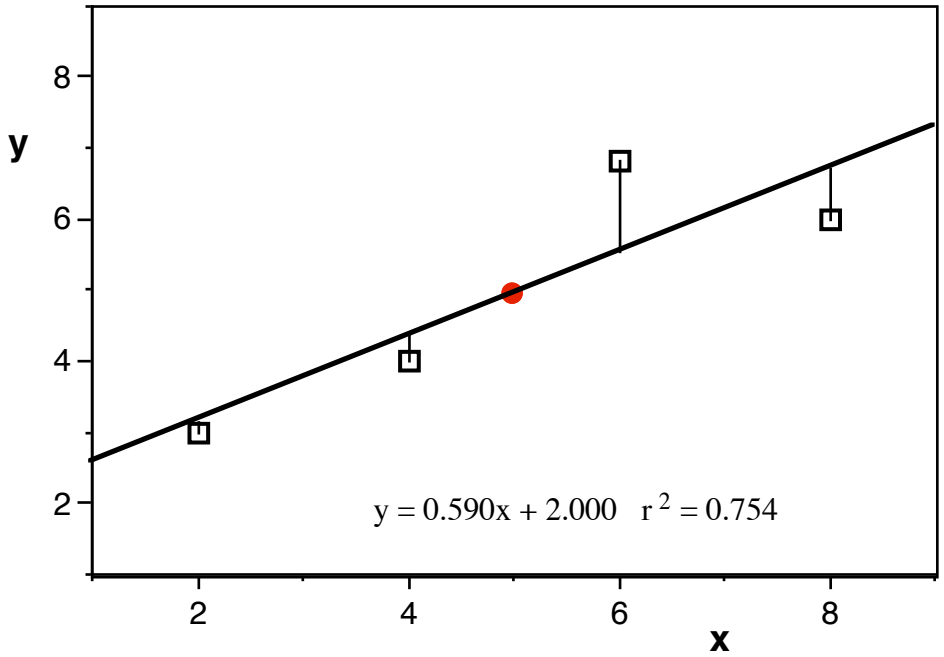

Figure 2 – *XY scatter plot showing the centroid (red circle), regression line, and y-residuals*.

Finally, it should be noted that errors in *y* values for large *x* values tend to distort or skew the best-fit line. This can be taken into account using either a weighted or robust regression technique. However, this is beyond the scope of the present tutorial.

2 Errors and Confidence Limits

In any area of measurement science, there is always some error in any signal. The error can arise from many sources, and can normally be accounted for using statistical techniques. However, because there is always some randomness associated with measurement error, it contributes some degree of uncertainty into the measurement, which corresponds to a certain confidence limit, within which we can be certain about the accuracy of our measurement. This leads to the way in which results are normally reported, where a measurement is reported with the error, such as $C = 51.2 \pm 0.05$ μg/ml. The ±0.05 is the standard error.

When preparing a calibration curve, there is always some degree of uncertainty in the calibration equation. To calculate the standard errors of the slope and the *y*-intercept, we require the residuals. The residual is the difference between the measured *y*-value and the *y*-value calculated from the calibration curve,

for a given observation. The calculated y-value is easily determined from the calibration equation and denoted $\hat{y}_i$, so the residual would be $(y_i - \hat{y}_i)$.

Once the residuals are known, we can calculate the standard deviation in the *y*-direction, which estimates the random errors in the *y*-direction.

$$s_{y/x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

This standard deviation can be used to calculate the standard deviations of the slop and the *y*-intercept using the formulas

$$s_b = \frac{s_{y/x}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

$$s_a = s_{y/x} \sqrt{\frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}}$$

where s_b is the standard deviation of the slope and s_a is the standard deviation of the *y*-intercept. The confidence limits can then be calculated from the *t*-statistic for $n - 2$ degrees of freedom. Tables of t-statistics are available in any undergraduate statistics textbook, and are also included in the lab manual. Note that some table give values of *t* for different values of *n*, while others give them for values of $\nu = n - 1$. Check carefully so that you use the appropriate value.

The confidence limits for the slope are then $b \pm t_{n-2} s_b$ and for the *y*-intercept $a \pm t_{n-2} s_a$. For a large number of samples with a 99% confidence interval, we can use $t_{n-2} = 2.58$. For the fluorescence data, the standard deviation of the slope is $s_b = 0.0409$, so the slope with confidence interval $b = 1.93 \pm (2.58 \times 0.0409) = 1.93 \pm 0.11$. The *y*-intercept with confidence interval is $a = 1.52 \pm 0.76$.

2.1 Random Error and Calculation of Concentration from the Calibration Curve: No Replication, Interpolated Value

Once we know the equation of the regression line, we can easily calculate the concentration x_0 from a given signal y_0. However, because we are now going from a *y*-value to an *x*-value (instead of the other way around), we need to find the error in *x*. This can be done with the standard deviation in x_0

$$s_{x_0} = \frac{s_{y/x}}{b} \sqrt{1 + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{b^2 \sum_i (x_i - \bar{x})^2}}$$

Here, y_0 is the experimental signal from the instrument for which x_0 is to be determined, and *n* is the number of samples. This formula only applies if there is no replication of each measurement. To calculate the concentration of a sample where the fluorescence intensity is 2.9,

1. Use the calibration equation determined previously, $y = 1.930x + 1.518$, with $y_0 = 2.9$, giving $x_0 = 0.72$ pg·ml^{-1}.
2. Calculate the standard deviation s_{x_0} using the equation above. For $n = 7$, $s_{y/x} = 0.4329$, and $b = 1.93$, we obtain $s_{x_0} = 0.26$, where the uncertainty is expressed as s_{x_0}.

3. Obtain a 95% confidence interval in the interpolated concentration by determining the two-tailed *t*-statistic for *n*-2 degrees of freedom. It is important to note that a two-tailed test is required for the interpolated results (*n*-2 d.o.f.), compared to the one-tailed test for the mean. From table of *t*-values, for $\nu = n - 2 = 5$, $t_5 = 2.57$. The interpolated concentration with 95% confidence interval is then reported as $C = x_0 \pm t_\nu s_{x_0} = 0.72 \pm 0.26$ pg·ml^{-1}.

2.2 Random Error and Calculation of Concentration from the Calibration Curve: With Replication, Interpolated Value

When you perform a sample measurement, you would normally perform more than one measurement of each sample, which is called replication. Replication is important in the statistical determination of your answer, in order to reduce the uncertainty and improve the accuracy of your measurement. Random fluctuations, which occur in any system, can lead to small errors in each measurement. By performing replications at each measurement, some or most of the error due to random fluctuations can be averaged out.

If replications are performed, the formula in the previous section must be modified to account for the extra degrees of freedom, as a result of the extra measurements. The formula for the standard deviation in x_0 with *m* replications is

$$s_{x_0,R} = \frac{s_{y/x}}{b}\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{b^2 \sum_i (x_i - \bar{x})^2}}$$

where the variable are the same as before. When working with a calibration curve with *n* measurements and a sample measurement y_0, the concentration with error as read from the calibration curve is $x_0 \pm s_{x_0,R}$.

2.3 Random Error and Calculation of Concentration from the Calibration Curve: No Replication, Extrapolated Value

In some cases, the measurement value for the sample will be outside the measured range of you calibration curve. While this situation is not desirable, due to the possibility of nonlinear effects outside the measurement range, it is sometimes unavoidable, and the results can still be used! All this requires is knowledge of a different way to calculate the standard deviation for extrapolation,

$$s_{x_E} = \frac{s_{y/x}}{b}\sqrt{\frac{1}{n} + \frac{\bar{y}^2}{b^2 \sum_i (x_i - \bar{x})^2}}$$

where *n* is the number of calibration values. The differences between this equation and the previous ones is that replications are not taken into account, and $y_0 = 0$, which is shown as part of the numerator in the square root. y_0 is shifted to the *x*-axis, and all calibration values are calculated from there. The reported sample concentration is then $x_E \pm s_{x_E}$.

2.4 Limits of Detection

As mentioned above, there is always some error associated with any instrumental measurement. This also applies to the baseline (or background or blank) measurement, i.e. the signal obtained when no analyte is present. One very important determination that must therefore be made is how large a signal needs to be before it can be distinguished from the background noise associated with the instrumental measurement. Various criteria have been applied to this determination, however the generally accepted rule in analytical chemistry is that the signal must be at least three times greater than the backgound noise.

Formally, then, the **limit-of-detection** (*lod*) is defined as the concentration of analyte required to give a signal equal to the background (blank) plus three times the standard deviation of the blank. That is, we first calculate the instrument response obtained with no analyte:

$$y_{lod} = y_{blank} + 3s_{blank}$$

and convert that value into the limit-of-detection by interpolation using the calibration equation. Where no blank has been measured, we can use the calibration data and regression statistics instead. In this case, we would use the y-intercept and standard deviation of the regression:

$$y_{lod} = a + 3s_{y/}$$

Again, the *actual* limit-of-detection is the concentration of analyte giving rise to this value. We can therefore obtain the confidence interval for the limit-of-detection in the same way as for any interpolated value as shown above.

When performing a calibration, you should always determine and report the *lod* from your calibration data, in addition to the regression statistics outlined above. The *lod* represents the level below which we cannot be confident whether or not the analyte is actually present. It follows from this that no analytical method can ever conclusively prove that a particular chemical substance is *not* present in a sample, only that it cannot be detected. In other words, there is no such thing as a zero concentration!