Some common mistakes in linear regression application

In analytical chemistry, we apply the concept of linear regression in our instrumental calibration by plotting a series of working standard concentrations against the instrumental responses in UV/visible/IR light absorbance, areas or peak heights under the curve, etc.

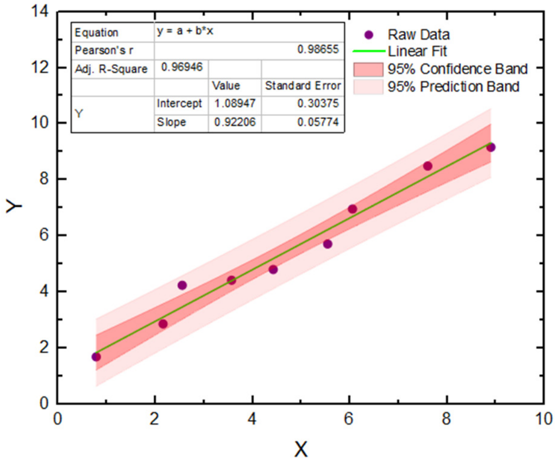
However, some mistakes are common in routine application of such linear regression that is worth for us to describe them so that our laboratory analysts can avoid them:

1. *Plotting too few standard concentration points against the instrumental responses.*

   Some analytical instruments such as inductively coupled plasma (ICP) spectrophotometer are well known to produce a very good linear range in its calibration.

   For example, measurement of copper level at wavelength 224.700nm can be safely made up to 450ppm linear range by using ICP/OES. Likewise, cobalt can be measured up to 250ppm linear range at wavelength 238.892nm with the same type of instrument.

   Hence, some laboratory analysts tend to prepare only 3 or 4 standard calibration solution for their ICP instrument calibration. Such fewer points of calibration lead to undesirable higher gradient uncertainty at 95% confidence limit, particularly at the lower and higher sections of the curve plotted as shown in the figure below.



2. *Incorrectly forcing the regression through zero.*

   Some instrument software allows a regression to be forced through zero (for example, by specifying removal of the intercept or ticking a ‘Set intercept zero’ option in Excel spreadsheet). This is valid only with good evidence to support its use, for example, if it has been previously shown that the y-intercept is not significant via a significance statistic testing.   Otherwise, interpolated values at the ends of the calibration range will be incorrect – often very seriously so near zero.

3. *Including the point (0,0) in the regression when it has not been measured.*

   Sometimes it is argued that the point (x=0, y=0) should be included in the regression, usually on the grounds that y=0 is the expected response at x=0.   This is entirely fallacious. It is simply cooking figures. To include such invented data is always bad practice in any case, it also has adverse effects on the statistical interpretation.

   Adding an invented point at (0,0) will cause the fitted line to move closer to (0,0), making the line fit more poorly near zero and also making it more likely that a real non-zero intercept will go undetected (because the calculated intercept will be smaller).

   The only circumstance in which a point (0,0) can validly be added to a regression data set is when a standard at zero concentration has been included and the observed response is either zero or is too small to detect and can reasonably be interpreted as zero.