Linear regression – Testing for non-linearity

In analytical chemistry, linear regression is commonly used in the construction of calibration functions required for analytical techniques such as gas chromatography, atomic absorption spectrometry and UV-visible spectrometry where a linear relationship is expected between the working standard concentration of the analyte (independent variable) plotted on the x-axis of the scatter plot, and the instrument response (dependent variable) plotted on the y-axis..

The general equation which describes a fitted straight line can be written as:

$$y = a + bx$$

where b is the gradient of the line and a, its intercept with the y-axis.

The method of least-squares linear regression is used to establish the values of a and b. The 'best fit' line obtained from least-squares linear regression is the line which minimizes the sum of the squared differences between the observed (or experimental) and fitted values for y. The signed difference between an observed value (y_i) and a fitted value ($\hat{y}_i$) is known as residual. (*Note*: A fitted y value is obtained by inserting the x-value to the linear equation of the plot.) Hence, i^{th} residual of $y = y_i - \hat{y}_i$ and the total sum of squares of residual SSR is:

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

It may be noted that the most common form of regression is of y on x. This assumes that the x values are known exactly and the only error occurs in the measurement of y.

However, there are a number of assumptions for a simple least-squares linear regression of y on x, such as:

1. The errors of the x-axis should be negligible;
2. For estimating confidence intervals and drawing inferences, the error associated with the y-axis must be normally distributed. If there is any doubt about the normality, a few replicates of the y-values can be averaged as mean value tends to be normally distributed even where individual results are not;
3. The variance of the error in the y-values should be constant across the

range of interest. *i.e.* the standard deviation should be constant. Simple least-squares regression gives equal weight to all points; this will not be appropriate if some points are much less precise than others;

4. Both the x- and y-data must be continuous valued and not restricted to integers, truncated or categorized (for example, sample numbers, days of the week).

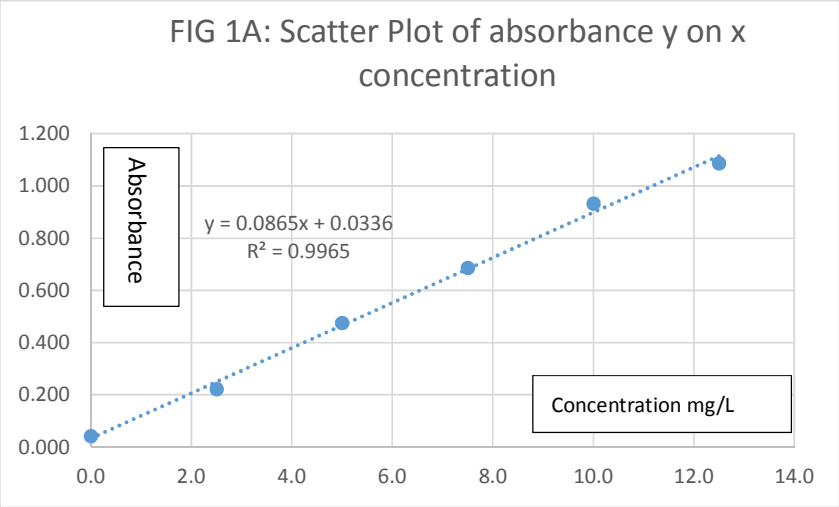
There are a few ways to test if the least-squares regression is truly linear.

a. Visual examination of regression data through their residual plots

Plotting the residuals can help to identify problems with poor or incorrect curve fitting.

If there is a good fit between the data and the regression model, the residuals should be distributed approximately randomly around zero. There is no trend in the spread of residuals with concentration on x-axis as shown in the Figure 1B example. However, the Figure 2B illustrates a typical plot of residuals that is obtained when a straight line is fitted through data that follow a non-linear trend. It can be noted that the plot shows a certain curve trend instead of randomly scattered around zero.

In another scenario, a residual plot can exhibit a straight line trend if the standard deviation of the y-values increases with analyte concentration. Such plot can be made when we have replicated results for each y-value.



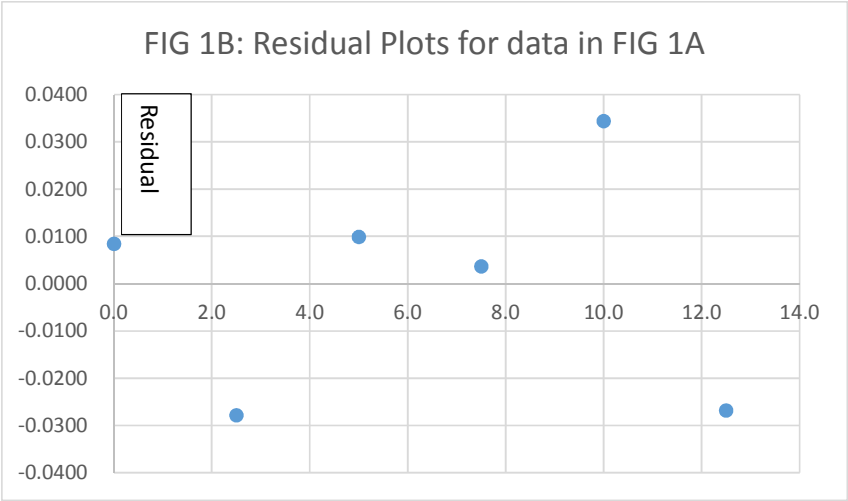


FIG 1B: Residual Plots for data in FIG 1A
Residual
0.0400
0.0300
0.0200
0.0100
0.0000
-0.0100
-0.0200
-0.0300
-0.0400
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0

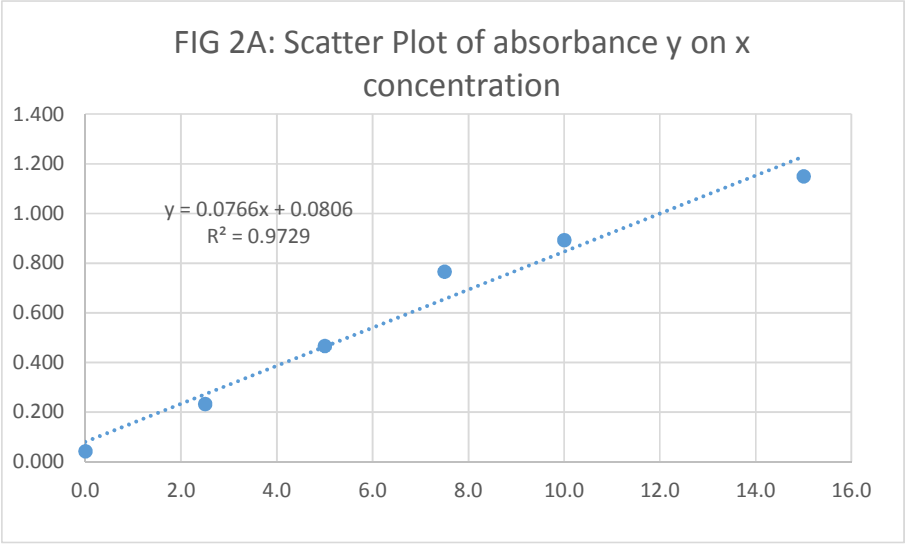


FIG 2A: Scatter Plot of absorbance y on x concentration
y = 0.0766x + 0.0806
R² = 0.9729
1.400
1.200
1.000
0.800
0.600
0.400
0.200
0.000
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
16.0

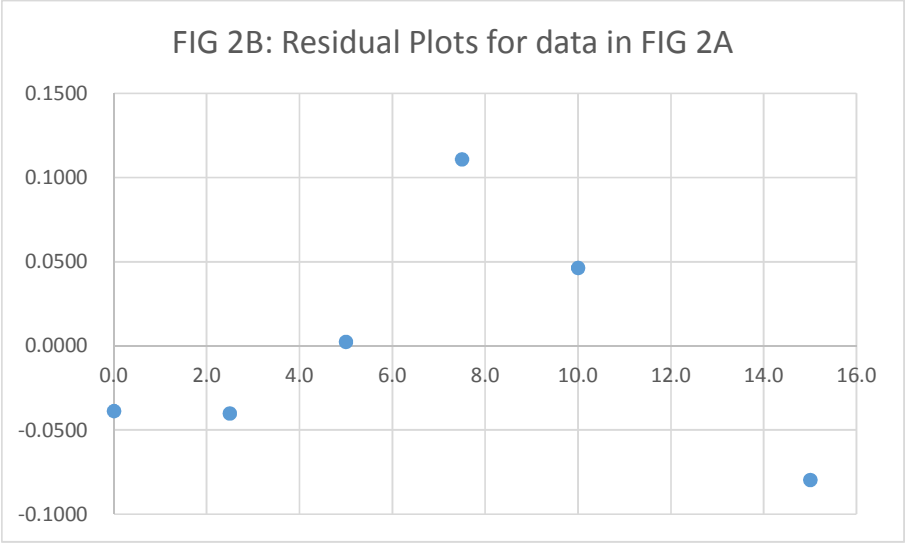


FIG 2B: Residual Plots for data in FIG 2A
0.1500
0.1000
0.0500
0.0000
-0.0500
-0.1000
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
16.0

b. *F*-Test for residual standard deviation against repeatability standard deviation

The residual standard deviation can be compared with an independent estimate of the repeatability of the y-values at a single x-value using an F-test as shown in the following equation:

$$F = \frac{s_{y/x}^{\ 2}}{s_r^{\ 2}}$$

where:

$$s_{y/x}^{\ 2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

and s_r is the stated repeatability of the method in question, which normally can be found in its standardized method..

A significance testing can be carried out. The null hypothesis for the test is H_o: $s_{y/x} = s_r$ and the alternative hypothesis is H_1: $s_{y/x} > s_r$. The test therefore is a 1-tailed test and the $s_{y/x}$ estimate has n-2 degrees of freedom where n is the number of pairs of data in the regression data set.

At the 95% confidence level, the appropriate F critical value is obtained from tables for $\alpha = 0.05$, v_1 = degrees of freedom for $s_{y/x}$ and $v2$ = degrees of freedom for s_r. If the calculated F exceeds the critical value, the null hypothesis is rejected, i.e. H_1 is true.

The inference is that the residuals are more widely dispersed than can be accounted for by random error alone. This could be evidence of non-linearity but a significant result could also occur if, for example, one or two observations were biased by other factors. Hence, studying the scatter plot and plot of the residuals will help to decide between the two.

3. ANOVA applied to residuals

If experimental observations are replicated at each value of x, applying one-way ANOVA to the residuals obtained using the x-values as the grouping factor can warn of non-linearity. A significant F value between group mean square indicates that the group means deviate from the line more than would be expected from the repeatability alone as represented by the within-group

mean square. This may point to significant non-linearity. Visual inspection of the residuals is till advisable, however, because a variety of effects can cause a significant between group effect in the residuals, such as volumetric errors or in the case of CRM, matrix effect. We shall show a worked example to illustrate the use of ANOVA applied to residuals in another paper.

4. Testing for significant higher order terms

Another practical approach to evaluating non-linear data is to fit a more complex (higher order) equation to the data, such as a quadratic equation with second order, x^2, and determine whether the new equation is a better representation of the data. We shall also illustrate this point in future communications.