HOW TO TEST FOR LINEARITY

by
Howard Mark
Mark Electronics
21 Terrace Avenue
Sufffern, NY 10901

Prepared for the FDA Process Analytical Advisory Committee
February 20, 2002

This report describes a new method of testing analytical data for linearity. This method overcomes the shortcomings of the current and proposed recommendations for linearity testing.

Let us begin by discussing what we want to test. The FDA/ICH guidelines, starting from a univariate perspective, considers the relationship between the actual analyte concentration and what generically called the “test result”. This terminology therefore holds good for every analytical methodology from manual wet chemistry to the latest high-tech instrument. In the end, even the latest instrumental methods have to produce a number, representing the final answer for that instrument’s quantitative assessment of the concentration, and that is the test result from that instrument. This is a univariate concept to be sure, but the same concept that applies to all other analytical methods. This is currently the way all analytical results are reported and evaluated. So the question to be answered, for any given method of analysis, is: ”Is the relationship between the instrument readings (test results) and the actual concentration linear?”

The FDA/ICH guidelines provide variant descriptions of the meaning of the term linearity. One definition is: “... ability (within a given range) to obtain test results which are directly proportional to the concentration (amount) of analyte in the sample.”(1) This is an extremely strict definition, one which is unattainable in practice when noise and error are taken into account. Figure 1 illustrates the problem. Here we have a plot of a set of hypothetical data that most of us would agree represents a linear relationship between the test result and the analyte concentration. While there is a line that meets the criterion that “test results are directly proportional to the concentration of analyte in the sample”, none of the data points fall on that line, therefore in the strict sense of the phrase, none of the data representing the test results can be said to be proportional to the analyte concentration. In the face of non-linearity of response, there are systematic departures form the line as well as random departures, but in neither case is any data point strictly proportional to the concentration.

Less strict descriptions of linearity are also provided. One recommendation is visual examination of a plot (unspecified, but presumably also of the method response versus the analyte concentration). This method works fairly well, but is subjective and open to different interpretations. It also is not amenable to application of computerized automated screening methods.

Another recommendation is to use “statistical methods”: calculation of a linear regression line is advised. If regression is performed, the correlation coefficient, slope, y-intercept and residual sum of squares are to be reported. These requirements are all in keeping

with their background of being applied to univariate methods of analysis. There is no indication given, however, as to how these quantities are to be related to linearity, and as Anscombe shows(2) they are not. Figure 3 presents two of the data sets from the Anscombe paper. One is a data set which is substantially linear, the other is a data set which is obviously very non-linear. The key point about those two sets of data is that when linear regression is performed, as recommended by the guidelines, all the regression statistics are identical for the two sets of data. Therefore it is immediately obvious that the regression results cannot distinguish between the two cases, since the regression results are the same for both of them.

In fact, the recommendations in the official guidelines for assessing linearity, while well-intended, are themselves not suitable for their intended purpose in this regard, not even for univariate methods of analysis. Therefore, let us propose a definition, that can at least serve as a basis for our own discussions. Let us define linearity as the characteristic of data such that a straight line provides as good a fit (using the least-squares criterion) as any other mathematical function, as a description of the relationship between the method response and the concentration of the analyte.

As a test for linearity, the Durbin -Watson statistic(3-5) has been proposed(6, 7) and is a step in the right direction, but it also has several shortcomings, including the fact that it can be fooled by data that had just the right (or wrong!) characteristics. For a relationship between test results and analyte concentration that is linear but contains random, independent, Normally distributed errors, the expected value of DW is 2. But for the data sequence:

0, 1, 0, -1, 0, 1, 0, -1, 0, 1, 0, -1, 0, ...

the value of the Durbin-Watson statistic for this set is also two, even though the sequence is neither random nor Normally distributed but is definitely non-linear.

DW also requires a relatively large number of samples, much larger than is needed for the linearity test to be described here. The method we now present is mathematically sound, more subject to statistical validity testing, based on well-known mathematical principals, is sensitive to smaller numbers of samples and can distinguish different types of non-linearity from each other.

This new method of determining non-linearity can be viewed from a number of different perspectives, and can be considered as coming from several sources. One way to view it is as having a pedigree as a method of numerical analysis(8).

The new method of determining non-linearity (or showing linearity) can also be related to a discussion of derivatives, particularly when using the Savitzky-Golay method of convolution functions. This is not very surprising, once you consider that the Savitzky-Golay convolution functions are also (ultimately) derived from similar considerations of numerical analysis.

In some ways it also bears a resemblance to the current method of assessing linearity that the FDA and ICH guidelines recommend, that of fitting a straight line to the data, and assessing the goodness of the fit. However, based on the work of Anscombe(2), we see that the currently recommended method for assessing linearity is faulty because it

cannot distinguish linear from non-linear data, nor can it distinguish between non-linearity and other types of defects in the data.

But an extension of that method can.

Having defined “linearity” as the characteristic of data such that a straight line provides as good a fit (using the least-squares criterion) as any other mathematical function, as a description of the relationship between the method response and the concentration of the analyte, we see that this seems to be almost the same as the FDA/ICH approach, which we just discredited. But there is a difference. The difference is the question of fitting other possible functions to the data; the FDA/ICH guidelines only specify trying to fit a straight line to the data. This is also more in line with out own proposed definition of linearity. We can try to fit functions other than a straight line to the data, and if we cannot obtain an improved fit, we can conclude that the data is linear

It is possible to fit other functions to a set of data, using least-squares mathematics. In fact, this is what the familiar Savitzky-Golay method does. The Savitzky-Golay algorithm, however, does a whole bunch of things, and lumps all those things together in a single set of convolution coefficients: it includes smoothing, differentiation, curve-fitting of polynomials of various degrees, least-squares calculations and finally combines all those operations into a single set of numbers that you can multiply your measured data by to directly get the desired final answer.

For our purposes, though, we don’t want to lump all those operations together. Rather, we want to separate them and retain only those operations that are useful for our own purposes. For starters, we discard the smoothing, derivatives and performing a successive (running) fit over different portions of the data set, and keep only the curve-fitting. Texts dealing with numerical analysis tell us what to do and how to do it. Many texts exist dealing with this subject, but we will follow the presentation of Arden(8). Arden points out and discusses in detail many applications of numerical analysis. These methods are all based on using a Taylor series to form an approximation to a function describing a set of data. The nature of the data, and the nature of the approximation considered differs from what we are used to thinking about, however. The data is assumed to be univariate (which is why this is of interest to us here) and to follow the form of some mathematical function, although we may not know what the function is. Using a Taylor series implies that the approximating function that we wind up with will be a polynomial, and perhaps one of very high degree (the “degree” of a polynomial being the highest power to which the variable is raised in that polynomial.)

The concepts of interest to us are contained in Arden’s book in a chapter entitled “Approximation”. This chapter takes a slightly different tack than the rest of the discussion in the book, but one that goes in exactly the direction that we want to go. In this chapter, the scenario described above is changed very slightly. There is still the assumption that there is a single (univariate) mathematical system (corresponding to our desired “analyte concentration” and “test reading”), and that there is a functional relationship between the two variables of interest although again, the nature of the relationship may be unknown. The difference, however, is the recognition that data may have error, therefore we no longer impose the condition that the function we arrive at must pass through every data point. We replace that criterion with a different criterion, and the criterion we use is one that will allow us to say that the function we use to describe the data “follows” the data in

some sense. The most common criterion used for this purpose is the “least squares” principle: to find parameters for any given mathematical function that minimizes the sum of the squares of the differences between the data and a corresponding point of the function.

Similarly, many different types of functions can be used. Arden discusses, for example, the use of Chebyshev polynomials, which are based on trigonometric functions (sines and cosines). There are other types of polynomials that could also be used, such as Legendre polynomials, Jacobi polynomials, and others. But by far the simplest to deal with, and therefore the most widely used approximating functions are simple polynomials; they are also convenient in that they are the direct result of applying Taylor’s theorem, since Taylor’s theorem produces a description of a polynomial that estimates the function being reproduced:

$$Y = a_0 + a_1X + a_2X^2 + a_3X^3 + ... + a_nX^n \quad \text{(Equation 1)}$$

For some data a polynomial can provide a better fit to that data than can a straight line. We present an example of that result as figure 2, for ease of reference. Higher order polynomials may provide an even better fit, if the data requires it.

The mathematics of fitting a polynomial by least squares are relatively straightforward, and we present a derivation here, one that follows Arden, but is rather generic, as we shall see: Starting from equation 1, we want to find coefficients (the a_i) that minimize the sum-squared difference between the data and the function’s estimate of that data, given a set of values of X. Therefore we first form the differences:

$$D = a_0 + a_1X + a_2X^2 + a_3X^3 + ... + a_nX^n - Y \quad \text{(Equation 2)}$$

Then we square those differences and sum those squares over all the sets of data (corresponding to the samples used to generate the data):

$$\sum_i D^2 = \sum_i (a_0 + a_1X + a_2X^2 + a_3X^3 + ... + a_nX^n - Y)^2 \quad \text{(Equation 3)}$$

The problem now is to find a set of values for the a_i that minimizes ΣD^2 with respect to each a_i. We do this by using the usual procedure of taking the derivative of ΣD^2 with respect to each a_i and setting each of those derivatives equal to zero. Note that since there are n + 1 different a_i. we wind up with n + 1 equations, although we only show the first three of the set:

$$\partial\,(\textstyle\sum_i D^2)\,/\,\partial a_0 = \partial\,(\sum\,(a_0 + a_1X + a_2X^2 + a_3X^3 + ... + a_nX^n\ - Y)^2\,)\,/\,\partial a_0 = 0 \quad \text{(Equation 4a)}$$

$$\partial\,(\textstyle\sum_i D^2)\,/\,\partial a_1 = \partial\,(\sum\,(a_0 + a_1X + a_2X^2 + a_3X^3 + ... + a_nX^n\ - Y)^2\,)\,/\,\partial a_1 = 0 \quad \text{(Equation 4b)}$$

$$\partial\,(\textstyle\sum_i D^2)\,/\,\partial a_2 = \partial\,(\sum\,(a_0 + a_1X + a_2X^2 + a_3X^3 + ... + a_nX^n\ - Y)^2\,)\,/\,\partial a_2 = 0 \quad \text{(Equation 4c)}$$

etc.

Now we actually take the indicated derivative of each term and separate the summations. Noting that $\partial\,(\sum_i F^2) = 2\sum_i F\,\partial F$ (where F is the inner summation of the a_i X):

$$2\,a_0\,\textstyle\sum(1) + 2a_1\,\sum_i X + 2a_2\,\sum_i X^2 + 2a_3\,\sum_i X^3 + ... + 2a_n\,\sum X^n - 2\sum_i Y = 0 \quad \text{(Equation 5a)}$$

$$2a_0\,\textstyle\sum_i X + 2a_1\,\sum_i X^2 + 2a_2\,\sum_i X^3 + 2a_3\,\sum_i X^4 + ... + 2a_n\,\sum_i X^{n+1} - 2\,\sum_i XY = 0 \quad \text{(Equation 5b)}$$

$$2a_0\,\textstyle\sum_i X^2 + 2a_1\,\sum_i X^3 + 2a_2\,\sum_i X^4 + 2a_3\,\sum_i X^5 + ... + 2a_n\,\sum_i X^{n+2} - 2\,\sum_i X^2Y = 0 \quad \text{(Equation 5c)}$$

etc.

Dividing both sides of equations 5 by two eliminates the constant term and subtracting the term involving Y from each side of the resulting equations puts the equations in their final form:

$$a_0\,\textstyle\sum(1) + a_1\,\sum_i X + a_2\,\sum_i X^2 + a_3\,\sum_i X^3 + ... + a_n\,\sum_i X^n = \sum_i Y \quad \text{(Equation 6a)}$$

$$a_0\,\textstyle\sum_i X + a_1\,\sum_i X^2 + a_2\,\sum_i X^3 + a_3\,\sum_i X^4 + ... + a_n\,\sum_i X^{n+1} = \sum_i XY \quad \text{(Equation 6b)}$$

$$a_0\,\textstyle\sum_i X^2 + a_1\,\sum_i X^3 + a_2\,\sum_i X^4 + a_3\,\sum_i X^5 + ... + a_n\,\sum_i X^{n+2} = \sum_i X^2Y \quad \text{(Equation 6c)}$$

etc.

The values of X and Y are known, since they constitute the data. Therefore equations 6 comprise a set of n + 1 equations in n + 1 unknowns, the unknowns being the various values of the a_i since the summations, once evaluated, are constants. Therefore, solving equations 6 for the a_i as simultaneous equations results in the calculation of the coefficients that describe the polynomial (of degree n) that best fits the data in the least squares sense.

In principle, the relationships described by equations 6 could be used directly to construct a function that relates test results to sample concentrations. In practice there is an important consideration that must be taken into account. This consideration is the possibility of correlation between the various powers of X. We find, for example, that the correlation coefficient of the integers from 1 to10 with their squares is 0.974, a rather high value.

Correlation effects are of concern to us. Our goal, recall, is to formulate a method of testing linearity in such a way that the results can be justified statistically. Ultimately we

will want to perform statistical testing on the coefficients of the fitting function that we use. In fact, we will want to use a t-test to see whether any given coefficient is statistically significant, compared to the standard error of that coefficient. We do not need to solve the general problem, however, just as we do not need to create the general solution implied by equation 1. In the broadest sense, equation 1 is the basis for computing the best-fitting function to a given set of data, but that is not our goal. Our goal is only to determine whether the data represent a linear function or not. To this end it suffices only to ascertain whether the data can be fitted better by *any* polynomial of degree greater than 1, than it can by a straight line (which is itself a polynomial of degree 1). To this end we need to test a polynomial of any higher degree. While in some cases, the use of more terms may be warranted, in the limit we need test only the ability to fit the data using only one term of degree greater than one. Hence, while in general we may wish to try fitting equations of degrees 2, 3, ... m (where m is some upper limit less than n), we can begin by using polynomials of degree 2, i.e., quadratic fits.

A complication arises. We learn from considerations of multiple regression analysis, that when two (or more) variables are correlated, the standard error of both variables is increased over what would be obtained if equivalent but uncorrelated variables are used. This is discussed by Daniel and Wood (see page 55 in(9)), who show that the variance of the estimates of coefficients (their standard errors) is increased by a factor of:

$$VIF = 1 / (1 - R^2) \quad \text{(Equation 7)}$$

when there is correlation between the variables, where R represents the correlation coefficient between the variables and we use the term VIF, as is sometimes done, to mean Variance Inflation Factor. Thus we would like to use uncorrelated variables. Arden describes a general method for removing the correlation between the various powers of X in a polynomial, based on the use of Chebyshev or other types of orthogonal polynomials, as we briefly mentioned above. But this method is unnecessarily complicated for our current purposes. In addition, Chebyshev polynomials (along with the Legendre and other types of polynomials) have a limitation: they require that the data be uniformly (or at least symmetrically) spaced along the X-axis, a requirement that real data may not always be able to meet. Since, as we shall see, we do not need to deal with the general case, we can use a simpler method to orthogonalize the variables, based on Daniel and Wood, who showed how a variable can be transformed so that the square of that variable is uncorrelated with the variable. This is a matter of creating a new variable by simply calculating a quantity Z and subtracting that from each of the original values of X. Z is calculated using the expression (see page 121 in(9)):

$$Z = \frac{\sum_{j=1}^{N} X_j^2 (X_j - \overline{X})}{2\sum_{j=1}^{N} (X_j - \overline{X})^2} \quad \text{(Equation 8)}$$

where the summations are taken over all the samples. Then the set of values $(X - Z)^2$ will be uncorrelated with X, and estimates of the coefficients will have the minimum possible variance, making them suitable for statistical testing. This calculation also has the inherent characteristic of not imposing special requirements on the sample distribution.

In his discussion of using these approximating polynomials, Arden presents a computationally efficient method of setting up and solving the pertinent equations. But we are less concerned with abstract concepts of efficiency than we are with achieving our goal of determining linearity. To this end, we point out that the equations 6, and indeed the whole derivation of them starting from equation 1, is familiar to us, although in a different context. We are all familiar with using a relationship similar to equation 1; in using spectroscopy to do quantitative analysis, one of the representations of the equation involved is:

$$C = b_0 + b_1X_1 + b_2X_2 + ... \qquad \text{(Equation 9)}$$

which is the form we commonly use to represent the equations needed for doing quantitative spectroscopic analysis using the MLR (Multiple Linear Regression) algorithm. The various X_i in equation 9 represent entirely different variables. Nevertheless, starting from equation 9, we can derive the set of equations for calculating the MLR calibration coefficients, in exactly the same way we derived equations 6 from equation 1. This derivation is presented in (5) and in (10).

Because of this parallelism between the situations we can set up the equivalencies:

$$a_0 = b_0$$
$$a_1 = b_1 \qquad X_1 = X$$
$$a_2 = b_2 \qquad X_2 = X^2$$
$$a_3 = b_3 \qquad X_3 = X^3$$
$$\text{etc.}$$

and we see that by replacing our usual MLR-oriented variables X_1, X_2, X_3, etc. with X, X^2, X^3, etc., respectively, we can use our common and well-understood mathematical methods (and computer programs) to perform the necessary calculations. Furthermore, along with the values of the coefficients, we can obtain all the usual statistical estimates of variances, standard errors, goodness of fit, etc. that MLR programs produce for us. Of special interest is the fact that MLR programs compute estimates of the standard errors of the coefficients, as described by Draper and Smith (see, for example, page 129 in (5)). This allows testing the statistical significance of each of the coefficients, which, as we recall, are now the coefficients of the various powers of X that comprise the polynomial we are fitting to the data.

This is the basis of our tests for non-linearity. We need not use polynomials of high degree since our goal is not necessarily to fit the data as well as possible. Especially since we expect that well-behaved methods of chemical analysis will produce results that are already close to linearly related to the analyte concentrations, we expect non-linear terms to decrease as the degree of the fitting equation used increases. Thus we need only fit a quadratic, or at most a cubic equation to our data to test for linearity, although there is nothing to stop us from using equations of higher degree if we choose. Data well-described by a linear equation will produce a set of coefficients with a statistically-significant value for the X^1 term and non-significant values for the coefficients of X^2 and (if used) X^3 or higher.

One recipe for performing the calculations can be expressed as follows:

1) We start with a set of test results (X) and corresponding analyte concentrations (Y)

2) Compute Z according to equation 8

3) From the set of test results (X) create a set of numbers from the values of X by calculating $(X - Z)^2$. Consider this set of values as a new variable, X_2.

4) Perform Multiple Linear Regression calculations of X and X_2 against Y, the analyte concentrations. The program should calculate the coefficients of X and X_2, regression statistics to evaluate the overall fit of the model to the data, calculate the standard error of the coefficients and from those the t-values for the coefficients.

As an example, these concepts are applied to the Anscombe data(2). Figure 3 reproduces two of the plots from Anscombe's paper, the ones showing a "normal", linear relationship and the one showing non-linearity. We recall that the linear regression statistics for both sets of data were essentially identical. Table 1 shows the results of applying this to both the "normal" data (Anscombe's X1, Y1 set) and the data showing non-linearity. We computed the nature of the fit using both a straight-line (linear) fit only as was done originally by Anscombe, and we also fitted a polynomial using the quadratic term a well. It is interesting to compare results for the four cases.

We find that in all four cases, the coefficient of the linear term is 0.5. In Anscombe's original paper, this is all he did, and obtained the same result, but this was by design. Using the polynomial, the fact that we obtained the same coefficient demonstrates that the quadratic term was indeed uncorrelated to the linear term.

The improvement in the fit from the quadratic polynomial indicated that the square term was indeed an important factor in fitting the data. The coefficient obtained for the quadratic term is comparable in magnitude to the linear term, as we might expect from the amount of curvature of the line we see in Anscombe's plot(2). The coefficient of the quadratic term for the "normal" data is much smaller than for the linear term.

As we expected, for the "normal", linear relationship, the t-value for the quadratic term for the linear data is not statistically significant. This demonstrates our contention that this method of testing linearity is indeed capable of distinguishing the two cases, in a manner that is statistically justifiable.

The performance statistics, the SEE and the correlation coefficient, show that including the square term in the fitting function for Anscombe's non-linear data set, gives essentially an almost perfect fit. Indeed, the fit is so good that it is fair to say that we have probably reproduced his generating function: the values of the coefficients obtained are probably the ones he used to create the data in the first place. The small SEE and very large t-values of the coefficients are indicative of the fact that we are near to having only computer round-off error as operative in the difference between the data he provided and the values calculated from the polynomial that included the second-degree term.

This is the basis for our new test of linearity. It has all the advantages we described: it gives an unambiguous determination of whether any non-linearity is affecting the relationship between the test results and analyte concentration. It provides a means of

distinguishing between different types of non-linearity, if they are present, since only those that have statistically-significant coefficients are active.

CONCLUSION

This new linearity test provides all the statistical tests that the current FDA/ICH test procedure recommends. and it also provides information as to whether, and how well, the analytical method gives a good fit of the test results to the actual concentration values. It can distinguish between different types of non-linearities, if necessary, while simultaneously evaluating the overall goodness of the fitting function. As the results from applying it to the Anscombe data show, it is eminently suited to evaluating the linearity characteristics of small data set as well al large ones.

This new test also provides all the statistical tests that the current FDA/ICH test procedure recommends. and therefore also provides information as to whether, and how well, the analytical method gives a good fit of the test results to the actual concentration values.

Table 1 - the results of applying the new method of detecting non-linearity to Anscombe's data sets, both the linear and non-linear, as described in the text.

Parameter	Coefficient when using only linear term	t-values for the coefficients using only linear term	Coefficient including square term	t-values for the coefficients including the square term
Results for non-linear data (figure 3B)				
Constant	3.000		4.268	
Linear term	0.500	4.24	0.5000	3135.5
Square term	----------	----------	-0.1267	-2219.2
S.E.E	1.237		0.0017	
R	0.816		1.0000	
Results for linear data (figure 3A)				
Constant	3.000		3.316	
Linear term	0.500	4.24	0.500	4.1
Square term	----------	----------	-0.0316	-0.729
S.E.E	1.237		1.27	
R	0.816		0.8291	

FIGURES

Figure 1 - A representation of linear data

Figure 2 - A quadratic polynomial can provide a better fit to a nonlinear function over a given region than a straight line can; in this case the second derivative of a Normal absorbance band

Figure 3 - Anscombe's plots reproduced, as described in the text. Figure 1A: the data showing a linear relationship. Figure 1B: the data showing a non-linear relationship.

FIGURE 1

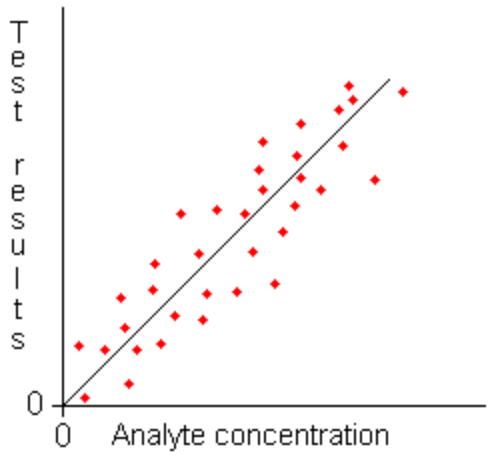


FIGURE 2

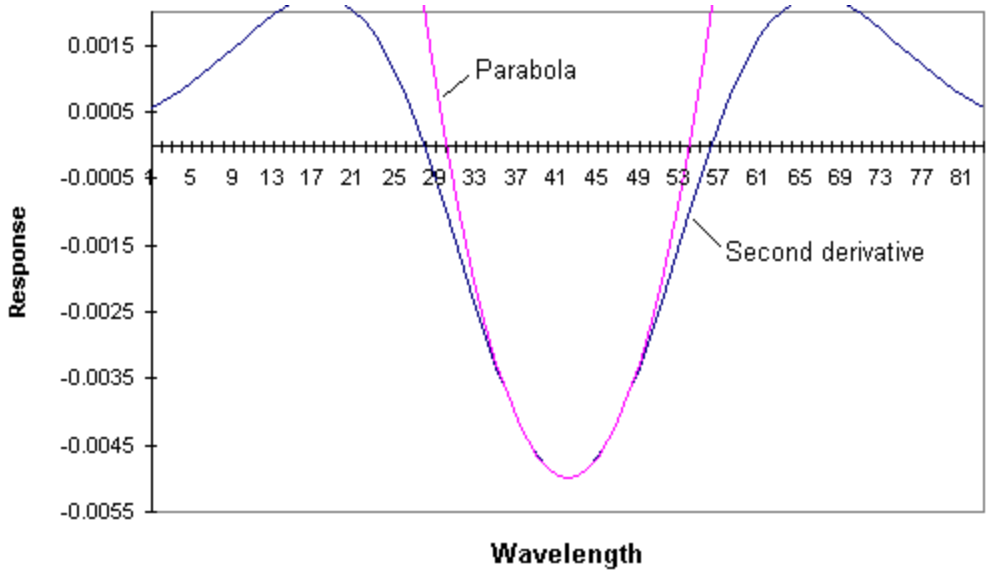

0.0015
0.0005
-0.0005
-0.0015
-0.0025
-0.0035
-0.0045
-0.0055
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81
Parabola
Second derivative
Response
Wavelength

FIGURE 3A

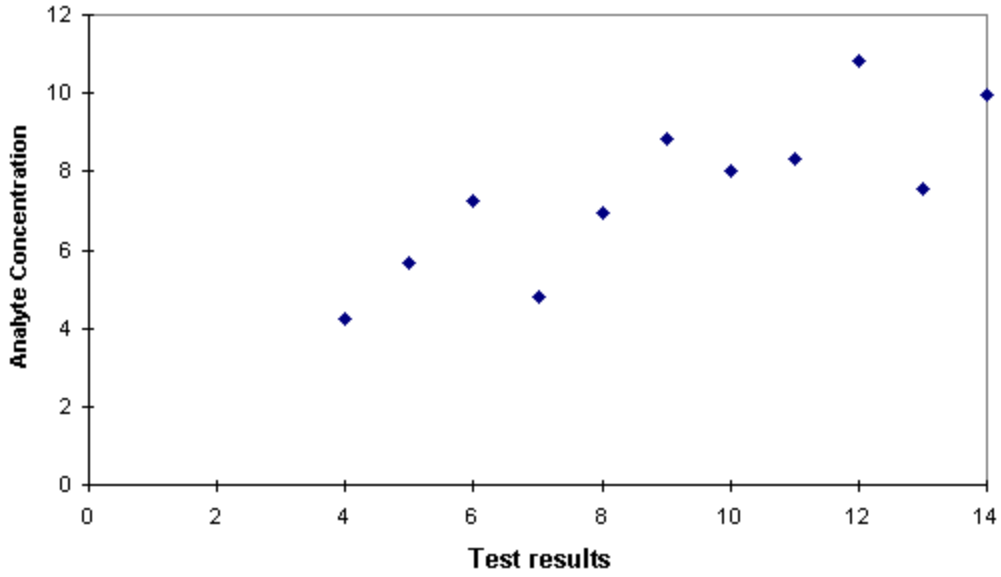
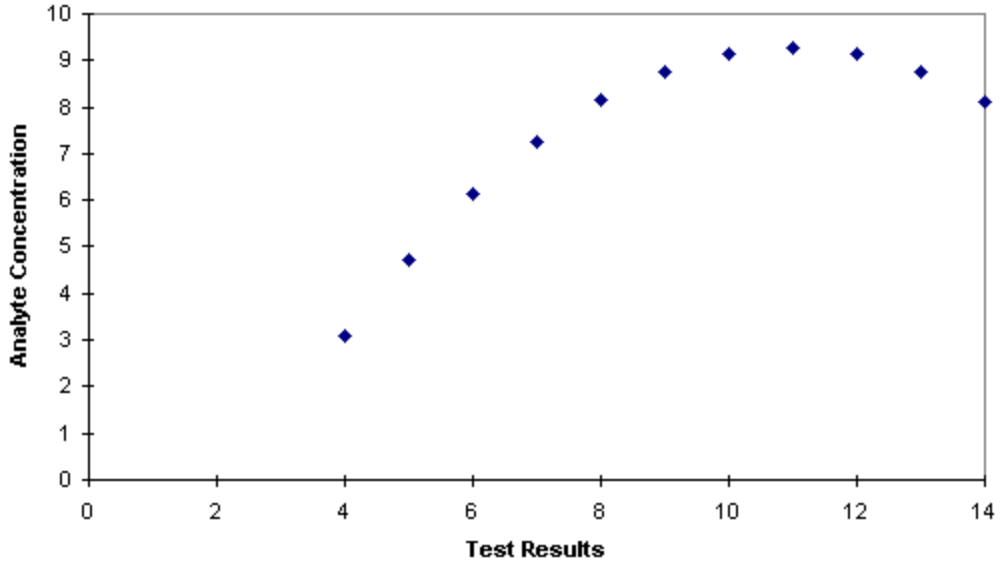


FIGURE 3B



REFERENCES

1. ICH-Q2A; "Guideline for Industry: Text on Validation of Analytical Procedures" (1995)

2. Anscombe, F. J.; Amer. Stat.; 27 p.17-21 (1973)

3. Durbin, J., Watson, G. S.; Biometrika; 37 p.409-428 (1950)

4. Durbin, J., Watson, G. S.; Biometrika; 38 p.159-178 (1951)

5. Draper, N., Smith, H.; "Applied Regression Analysis"; 3 ed; John Wiley & Sons; New York (1998)

6. Ritchie, G. E., Roller, R. W., Ciurczak, E. W., Mark, H., Tso, C., MacDonald, S. A.; Journal of Pharmaceutical and Biomedical Analysis; (Part A: in press) (2002)

7. Ritchie, G. E., Roller, R. W., Ciurczak, E. W., Mark, H., Tso, C., MacDonald, S. A.; Journal of Pharmaceutical and Biomedical Analysis; (Part B: in press) (2002)

8. Arden, B. W.; "An Introduction to Digital Computing"; 1st ed; Addison-Wesley Publishing Co., Inc.; Reading, Mass. (1963)

9. Daniel, C., Wood, F.; "Fitting Equations to Data - Computer Analysis of Multifactor Data for Scientists and Engineers"; 1 ed; John Wiley & Sons; (1971)

10. Mark, H.; "Principles and Practice of Spectroscopic Calibration"; John Wiley & Sons; New York (1991)