S. Afr. J. Chem., 2016, **69**, 166–173,
<http://journals.sabinet.co.za/sajchem/>.

Regression Analysis in Analytical Chemistry. Determination and Validation of Linear and Quadratic Regression Dependencies

Rafał I. Rawski[a], Przemysław T. Sanecki[b,*], Klaudia M. Kijowska[b], Piotr M. Skitał[b] and Dorota E. Saletnik[b]

[a]*Faculty of Biology and Agriculture, University of Rzeszow, 35-601 Rzeszow, Poland.*

[b]*Faculty of Chemistry, Rzeszow University of Technology, 35-959 Rzeszow, Poland.*

Received 25 February 2016, revised 24 June 2016, accepted 29 June 2016.

ABSTRACT

The theory and practice of the extended statistical evaluation for linear and quadratic regression models used for calibration were presented. Two complete examples, solved step by step were presented as a short guide. The validation of regression dependences was based on classic F-Snedecor, Lack of Fit, F_{IUPAC} and Mandel tests.

KEYWORDS

Correlation, regression, Lack of Fit test, Mandel test, calibration.

1. Introduction

In analytical chemistry practice, signal-concentration plots are often used where a concentration of analyzed samples is proportional to a respective analytical device's signal, e.g. absorbance, current, potential, peak height and peak surface.

To obtain reliable results, a calibration curve, i.e. a dependence of signal strength *vs.* concentration of substance in measured sample, must be drawn up first. The calibration curve is not a function dependence in mathematical sense, where one independent variable x corresponds to one and only one value of dependent variable y. It is a regression dependence, a 'slightly worse' than a functional dependence where x values may correspond to several values of y. In case of function dependence all points lie on the model curve and are indistinguishable from it. In practice, we have a set of data in the form of a matrix consisting of one x column and one or more y columns. The question, whether there is a relationship (correlation) between y and x variables has to be answered. Quantitative determination of such correlation's force is called the regression analysis.

The aim of this study is to give the reader a clear tool for an extended statistical evaluation of linear and quadratic regression models used for calibration. Therefore, in addition to the necessary minimum of theory, a number of specific examples are given. The correct choice between linear and quadratic regression models used for calibration is crucial in many biochemical tests such as Bradford or Smith protein determination methods.

Editorial offices of many analytical journals no longer accept the simple statistical analysis conducted during a calibration curve determination (straight-line equation $ax + b$, s_a, s_b or Δa, Δb and r^2). Using regression coefficient to verify the quality of correlation is not enough and can even lead to wrong conclusions as its value close to 1 can also be obtained for clearly curved dependences.[1,2] For this reason, publications must present a full statistical evaluation including classic Fisher-Snedecor test, Mandel's *F*-test and Lack of Fit test for possible curvilinearity presence. This work pays special attention to these tests, especially the second and third one mainly because they are poorly represented in literature. The existing literature on Mandel and Lack of Fit tests refers to a number of procedures, without giving clear examples solved step by step, with a clear final conclusion. Instead, reference works often contain references to statistical software, unavailable or unclear to the reader.[3,4] In our approach of the problem, standard programs like Origin, Excel or Libre Office will suffice.

2. Calibration Curves as an Example of Regression Dependence. Function Dependencies *vs.* Regression Dependencies

Regression analysis aims to create a model describing a set of experimental x and y data and to predict unknown x values using created model. This generates some estimation errors. For a given set of data there may be a large number of regression models, but in order to obtain reliable results, the model showing the smallest deviation from experimental data should be chosen. This can be done intuitively by leading a line or curve between points or strictly mathematically.

For a good understanding of the statistical data processing idea and the terms such as modelling, estimation, calibration and calibration curves, the concept of functional dependence and regression dependence must be firstly distinguished.[5,6] In case of function dependence $y = f(x)$ there is a situation when every variable x_i has assigned exactly one value of y_i. In the case of regression dependence points do not lie exactly on the line and one independent variable x_i can have assigned several values of y_i. A comparison of function and regression dependences is shown in Fig. 1.

For the slope (a) and intercept (b) of the straight line equation $y = ax + b$, the confidence intervals Δa and Δb can be determined with the use of parameters generated by the Origin software. Unfortunately, the program outputs symbols which are inverse to those adopted in mathematics, namely $y = B \cdot x + A$ instead of $y = a \cdot x + b$ (Table 1). Moreover, the nature of the parameter 'Error' for A and B has to be clarified: they are just standard deviations s_a and s_b. Origin program outputs also the standard devia-

* To whom correspondence should be addressed. E-mail: psanecki@prz.edu.pl

ISSN 0379-4350 Online / ©2016 South African Chemical Institute / http://saci.co.za/journal
DOI: http://dx.doi.org/10.17159/0379-4350/2016/v69a20

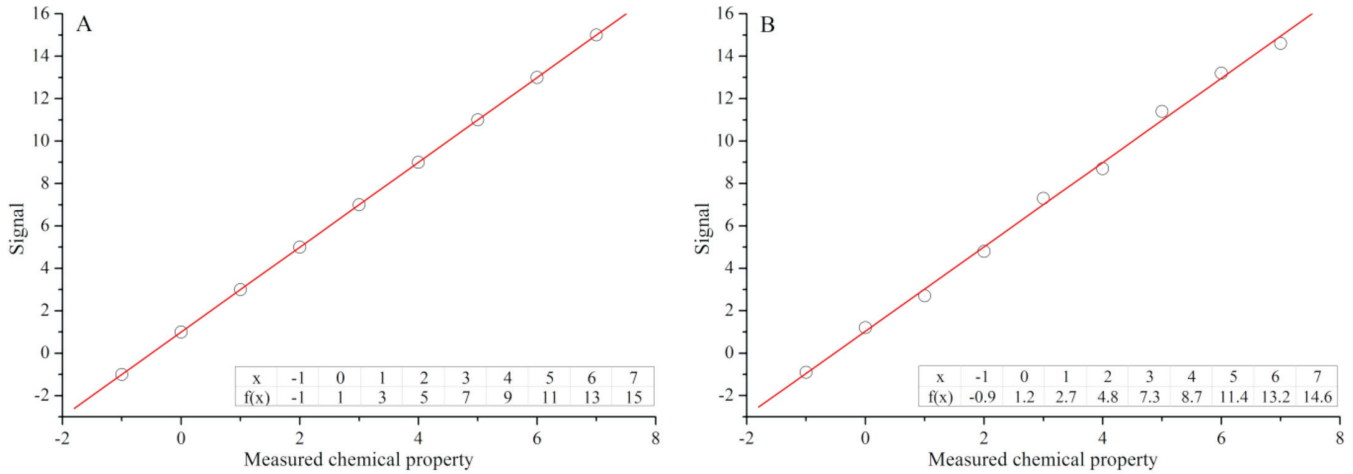

Figure 1 Comparison and distinction of function and regression dependences. (**A**) Function dependence: $y = 2x + 1$; $\Delta a = 0$; $\Delta b = 0$; $r^2 = 1$. (**B**) Regression dependence: $y = 1.988x + 1.035$; $\Delta a = 0.004$; $\Delta b = 0.0628$; $r^2 = 0.9970$.

Table 1 Explanation of Origin symbols concerning regression analysis.

Origin $Y = A + BX$	Statistics $y = ax + b$	Equation
B	slope a	(9)
A	intercept b	(9)
SD	s_0	(1)
Error B	s_a	(2)
Error A	s_b	(2)

tion of the fit s_0, marked as SD (Table 1). For our example of regression dependence, Origin's $s_a = 0.04117$; $s_b = 0.1630$; $s_0 = 0.3189$.

3. Confidence Interval and Determination Coefficient

Determination of the regression equation must be followed by the confidence intervals determination for a and b coefficients. This can be done with the use of respective standard deviations and $t_{Student}$ coefficient, $t(\alpha,df)$.[5] Standard deviations are described by Equations (1) and (2) and can be calculated or obtained from Origin.

$$s_0^2 = \frac{\sum_{i=1}^{n} y_i^2 - b\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i y_i}{n-2} \quad (1)$$

$$s_a^2 = \frac{n s_0^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}; \quad s_b^2 = \frac{s_a^2}{n}\sum_{i=1}^{n} x_i^2 \quad (2)$$

The $t_{Student}$ coefficient is given for the assumed level of significance α and the number of degrees of freedom $n-2$, where n is the number of x values.

The confidence intervals of a and b coefficients are determined by Equation (3)

$$\Delta a = t(\alpha, df) \cdot s_a^2; \quad \Delta b = t(\alpha, df) \cdot s_b^2 \quad (3)$$

To evaluate the quality (strength) of the regression, correlation coefficient r or determination coefficient r^2 is used.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

The r^2 coefficient ranges between 0 and 1. The closer the value is to 1, the better applied model describes a given set of experimental points. However, it is not a decisive criterion.

Standard deviation s_a and s_b alone, obtained using Origin are also not enough. Many people mistake s_a and s_b for confidence intervals, thus considering the former as a measure of a and b uncertainty. To obtain the actual confidence intervals, s_a^2 and s_b^2 have to be multiplied by $t_{Student}$ coefficient (e.g. 2.365) for the assumed level of significance (e.g. $\alpha = 0.05$) and $n-2$ degrees of freedom. This leads to full linear regression equation shown in Fig. 1, namely $y = ax + b$; $a = 1.988 \pm 0.004$; $b = 1.035 \pm 0.06280$; $r^2 = 0.9970$.

The confidence interval, as a uncertainty measure of determined equation, depends significantly on the number of measurement points *via* $t_{Student}$ parameter. It pays off to have 8 or even 9 of them, because then the confidence interval is significantly narrower. However, starting from $n = 10$, $t_{Student}$ begins to decrease only slightly with and therefore there is no real need to further increase the number of measurement points.

Based on the basic regression equation y_1, and changing its a and b values by respective Δa and Δb according to the rule: $(a\uparrow)(b\uparrow)$, $(a\uparrow)(b\downarrow)$, $(a\downarrow)(b\uparrow)$, $(a\downarrow)(b\downarrow)$ four boundary regression equations (5) were determined:

$$y_1 = 1.988x + 1.035;\ y_2 = 2.085x + 1.420;$$
$$y_3 = 2.085x + 0.6495;\ y_4 = 1.891x + 1.420; \quad (5)$$
$$y_5 = 1.891x + 0.6495$$

The resulting y_2 and y_5 boundary equations determine the widest range of error, i.e. confidence interval in graphic form for the whole model dependence. Thus, the value of x acquired from the model dependence also has its confidence interval Δx, which should also be taken into account. The measured value of y and the basic regression equation with its confidence intervals (Fig. 2.) makes it possible to determine an unknown x value (e.g. concentration).

3.1. Summary of Example from Fig. 2

1. The basic regression equation were determined *via* Origin.
2. The confidence intervals for a and b were determined with the use of s_a^2, s_b^2 and $t_{Student}$ values.
3. The four border equations were determined. Basic equation and two external equations giving the widest confidence interval, were used.
4. For the measured value y_{exp}, the three values of $x = 2.548$; 2.245; 2.882 were determined from regression equation. This ultimately gives a range of $x \in \langle 2.245, 2.882 \rangle$.

The overall form of a linear regression is given by Equation (6):

$$\hat{y} = (a \pm \Delta a)x + b \pm \Delta b \quad (6)$$

where Δa and Δb are respective confidence intervals (the ^ sym-

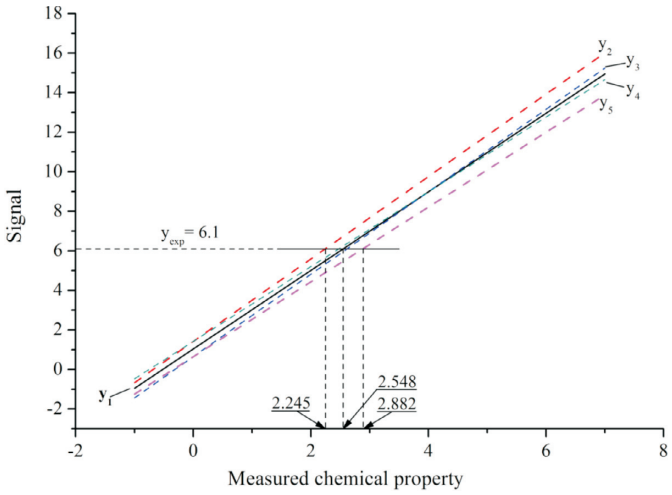

Figure 2 Confidence interval in graphic form for the linear model curve. The basic regression equation y_1 is overlapped with boundary regression equations y_2 and y_5. The latter equations are the result of determined Δa and Δb confidence intervals. Numerical data were taken from Fig. 1 and regression equations (5).

bol indicates an estimated value read from the regression equation). If $\Delta a = 0$ and $\Delta b = 0$, we recognize a function dependence $y = ax + b$, which confirms that the function dependence is a special case of regression dependence, when all confidence intervals are zero.

The value read from the regression equation for given x, is always an estimated value labelled as $\hat{y}$. The $y_i - \hat{y}_i$ difference, or precisely its square, is essential for the least squares method.

4. Least Squares Method

The above analysis was based on the Origin program without specifying the fact that it uses the least squares method to determine the regression curves (Figs. 1 and 2). This method is an optimization of a model dependence, where the criterion for optimization is to minimize the sum of model's deviations from experimental points, which can be visualized as a sum of square areas[6] marked by dotted lines in Fig. 3, a solid line denotes model-experiment difference. Please note that the least squares method applied in analytical chemistry assumes no error in the concentration of the standards (x) and that the only variable is the signal (y).

If a function $y = f(x)$ is drawn through a set of measurement points x_i, y_i, then each x value will be matched with two values: measured (y) and estimated $\hat{y}$. The model, i.e. the regression equation, may be linear or curvilinear and expressed by an appropriate function.

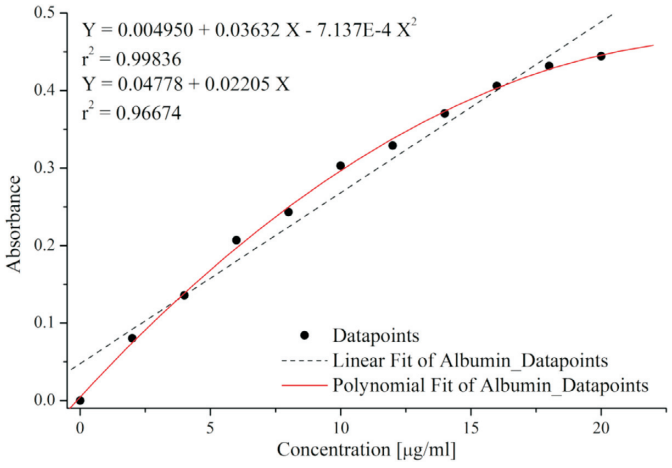

Figure 3 The essence of a linear model optimization for experimental data in Least Squares Method. The model-experiment differences are indicated by solid lines.

For the linear regression $y = ax + b$, optimization criterion is to minimize the sum $S(a,b)$:

$$S(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b - ax_i)^2 = minimum \qquad (7)$$

where $\hat{y}_i$ represents the estimated value based on the regression equation, and n represents the number of measured x points (in the case of multiple y_i measurements for each point x, average values $\bar{y}$ should be used).

Equation (7) is a function of two variables, a and b. Determination of such function's minimum is possible by finding a point at which the partial derivatives for all its variables are equal to zero. Hence, the system of equations (8) and its rearranged form (9):

$$\begin{cases} \dfrac{\partial S(a,b)}{\partial a} = 2\sum_{i=1}^{n}(y_i - b - ax_i)(-x_i) = 0 \\ \dfrac{\partial S(a,b)}{\partial b} = 2\sum_{i=1}^{n}(y_i - b - ax_i)(-1) = 0 \end{cases} \qquad (8)$$

$$\begin{cases} b\sum_{i=1}^{n} x_i + a\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \\ bn + a\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \end{cases} \qquad (9)$$

From the system (9) stem formulas for the a and b coefficients of a linear equation:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (10)$$

The method used is sensitive to presence of significantly deviating points in the dataset. Each of these will deviate outlined function toward its value, weakening the fit. Therefore, before using this method, any gross errors must be eliminated from the dataset, with the use of appropriate mathematical criterion.[6]

4.1. Least Squares Method – Example

Spectrophotometric method was used to measure samples containing known concentration of albumin (Bradford method) and glycine (ninhydrin method). The experiment was performed three times and the obtained experimental data are given in Table 2.

From formulas (1), (2) and (10) or faster using Origin, for albumin data we get:

$$a = \frac{11\cdot 39.21 - 110\cdot 2.951}{11\cdot 1540 - (110)^2} = 0.02205;$$

$$b = \frac{2.951\cdot 1540 - 110\cdot 39.21}{11\cdot 1540 - (110)^2} = 0.04778;$$

$$s_0^2 = \frac{1.013 - 0.04778\cdot 2.951 - 0.02205\cdot 39.21}{11-2} = 8.174\cdot 10^{-4};$$

$$s_a^2 = \frac{11\cdot 8.174\cdot 10^{-4}}{11\cdot 1540 - (110)^2} = 1.850\cdot 10^{-6};$$

$$s_b^2 = \frac{1.850\cdot 10^{-6}}{11}\cdot 1540 = 2.602\cdot 10^{-4}.$$

Table 2 Results of spectrophotometric measurement of albumin and glycine concentration. Both experiments were repeated three times.

Albumin standard /μg mL^{-1}	Peak height /$\times 10^{-3}$	Average peak height /$\times 10^{-3}$	Glycine standard /μg mL^{-1}	Peak height /$\times 10^{-3}$	Average peak height /$\times 10^{-3}$
0	0, 0, 0	0	0	0, 0, 0	0
2	89, 82, 70	80.33	0.00250	175, 221, 222	206.0
4	137, 142, 128	135.7	0.00500	368, 435, 458	420.3
6	215, 201, 205	207.0	0.00750	627, 643, 635	635.0
8	245, 242, 243	243.3	0.01000	861, 799, 858	839.3
10	302, 317, 290	303.0	0.01125	952, 915, 961	942.7
12	335, 325, 327	329.0	0.01250	1087, 1009, 1015	1037
14	366, 375, 370	370.3	0.01375	1189, 1097, 1180	1155
16	413, 405, 400	406.0	0.01500	1301, 1174, 1248	1241
18	434, 441, 420	431.7			
20	442, 450, 441	444.3			

Thus, the calibration curve equation is: $y = 0.02205x + 0.0477$.

The $t(\alpha,df)$ coefficient for the level of significance $\alpha = 0.95$ and 9 degrees of freedom is 2.262 and the confidence intervals are as follows:

$\Delta a = 2.262 \cdot 1.850 \cdot 10^{-6} = 4.185 \cdot 10^{-6}$; $\Delta b = 2.262 \cdot 2.602 \cdot 10^{-4} = 5.886 \cdot 10^{-4}$. Therefore the final, rounded, linear regression equation looks like this:

$$y = (2.205 \cdot 10^{-2} \pm 4.185 \cdot 10^{-6})x + (4.778 \cdot 10^{-2} \pm 5.886 \cdot 10^{-4}) \quad (11)$$

The r^2 coefficient of Equation (11) equals 0.9667 which is a rather good but not the best result, as indicated by the second digit after the decimal point. This suggests the need to check other models.

On the other hand for glycine data, $a = 83.26$; b = 0.002650; $s_0^2 = 4.733 \cdot 10^{-5}$; $s_a^2 = 0.2179$; $s_b^2 = 2.144 \cdot 10^{-5}$, which gives an equation in the form of:

$y = 83.26x + 2.650 \cdot 10^{-3}$.

For $\alpha = 0.95$ and df = 7, $t(\alpha,df) = 2.365$, thus calculated confidence intervals are consecutively $\Delta a = 0.5153$; $\Delta b = 5.070 \cdot 10^{-5}$, which gives us the final equation:

$$y = (83.26 \pm 0.5153)x + 2.650 \cdot 10^{-3} \pm 5.070 \cdot 10^{-5} \quad (12)$$

For both considered examples, the values s_0^2, s_a^2, s_b^2 were calculated using Origin. Values of Δa and Δb, computationally simple, were calculated using an ordinary calculator.

For Equation (12), r^2 coefficient is equal to 0.9998. Its value being so close to 1 suggests a nearly perfect regression equation. However, in light of the current requirements of analytical journals, even almost perfect r^2 value does not constitute enough evidence of the suitability of the adopted model. It is still required to provide the results of additional statistical tests.

5. Lack of Fit Test[7,8]

As was stated before, the r^2 coefficient cannot reliably indicate whether the adopted model is appropriate or not. To decide, statistical analysis should be performed on the adopted model. One of the statistical tests used for this purpose is Lack of Fit test, which can detect mismatches between the data and the adopted regression model.

Now, the regression model $y_{ij} = ax_i + b + \varepsilon_{ij}$; $\{(x_i, y_{ij}): i = 1,...,n; j = 1,...,c\}$ is tested for the case where each of x_i matches with at least $c_i = 3$ y values (Table 2) for the data, where n is the number of measurements as previously used and c is the number of repetitions of the measurement for a given x. To consider that linear model fits the data, a null hypothesis H_0: $y_{ij} = ax_i + b$ must be maintained. To test it one has to calculate the sum of squares due to errors for the full (13) and reduced (14) models and then subtract them from each other (15).

$$SSE_{full} = \sum_{i=1}^{n}\sum_{j=1}^{c_i}\left(y_{ij} - \bar{y}_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{c_i} y_{ij}^2 - \sum_{i=1}^{n} n_i \bar{y}_i^2 \quad (13)$$

$$SSE_{red} = \sum_{i=1}^{n}\sum_{j=1}^{c_j}\left(y_{ij} - b - ax_i\right)^2 \quad (14)$$

$$SSLF = SSE_{red} - SSE_{full} \quad (15)$$

Results obtained from Equations (13) and (15) should then be divided by their respective degrees of freedom (16), (17) which leads to Equations (18) and (19). The final result of this test is the F_{LoF} value (20).

$$df_{SSEfull} = c - n; \mathrm{c} = \sum_{i=1}^{n} c_i \quad (16)$$

$$df_{SSLF} = n - 2 \quad (17)$$

$$MSE_{full} = \frac{SSE_{full}}{df_{SSEfull}}; \quad (18)$$

$$MSLF = \frac{SSLF}{df_{SSLF}} \quad (19)$$

$$F_{LoF} = \frac{MSLF}{MSE_{full}} \quad (20)$$

To decide whether to accept or reject the null hypothesis posed at the beginning of the test, the calculated F_{LoF} value (20) should be compared with the critical value $F_{LoF}^{\#}(\alpha; df_{SSLF}; df_{SSEfull})$ taken from F distribution tables. If for the assumed level of significance α, the inequality $F_{LoF} > F_{LoF}^{\#}$ is true, the linearity hypothesis H_0: $y_{ij} = a_{xi} + b$ must be rejected.

5.1. Lack of Fit – Example

Lack of Fit test requires a minimum of three y values for each x. Therefore, in the example the original experimental values were used instead of the average ones (Table 2). The data were introduced consecutively into the formulas (13), (14), (15), next to (16), (17), (18), (19) and (20) giving results:

Albumin

$$SSE_{full} = 3.039 - 3.038 = 0.001220;$$
$$SSE_{red} = 0.02329;$$
$$SSLF = 0.02329 - 0.001220 = 0.02207;$$
$$df_{SSEfull} = 33 - 11 = 22;$$
$$df_{SSLF} = 11 - 2 = 9;$$
$$MSE_{full} = \frac{0{,}001220}{22} = 5.545 \cdot 10^{-5};$$
$$MSLF = \frac{0.02207}{9} = 2.452 \cdot 10^{-3};$$
$$F_{LoF} = \frac{2.452 \cdot 10^{-3}}{5.545 \cdot 10^{-3}} = 44.21;$$
$$F^{\#}_{LoF}(0.95; 9; 22) = 2.420$$

Glycine

$$SSE_{full} = 0.02663;$$
$$SSE_{red} = 0.02764;$$
$$SSLF = 1.009 \cdot 10^{-3};$$
$$df_{SSLF} = 7;\ df_{SSEfull} = 18;$$
$$MSE_{full} = 1.479 \cdot 10^{-3};$$
$$MSLF = 1.441 \cdot 10^{-4};$$
$$F_{LoF} = 0.09742;$$
$$F^{\#}_{LoF}(0.95; 7; 18) = 2.577$$

As one can see, the inequality $F_{LoF} > F^{\#}_{LoF}$ is true for albumin data. Thus, the null hypothesis of good linear model fit has to be rejected, and the alternative hypothesis stating that the linear model for the albumin data exhibits a mismatch should be accepted.

In the case of glycine data an opposite situation can be observed. $F_{LoF} < F^{\#}_{LoF}$, thus there is no basis to reject the null hypothesis of linear model fitting the experimental data.

6. Classic Fisher-Snedecor Test[9,10]

This test involves separate determination and comparison of the linear and quadratic effect's significance. The linear effect E_{lin}, which means that part of total variability of y, that can be described by the accepted model $\hat{y} = ax + b$, results from the expression (21)

$$E_{\text{lin}} = r^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{21}$$

$$R^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{22}$$

where $\bar{y}$ means the average value of the whole population of y. However, if the linear model is replaced with the quadratic one, i.e. $\hat{y} = ax^2 + bx + c$, then the part of the total variability of the dependent variable, described by the parabola, results from expression (22), which is almost identical to Equation (21), with only one element different.

The quadratic effect E_q of the dependent variable's total variation is the part that can be further explained by replacing the linear model with the quadratic model. In order to obtain a formula for this effect, expression (21) has to be subtracted from expression (22) resulting in Equation (23):

$$E_q = R^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 - r^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 = (R^2 - r^2) \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{23}$$

This test defines the significance of the both effects as a ratio of said effect to its residual variance. The value of the residual variance describes this part of the dependent variable's variability that remains after deducting the variability described by the independent variable. It can be calculated by dividing the sum of deviations squares by the number of degrees of freedom. The number of degrees of freedom is equal to the number of independent variables minus the polynomial degree plus 1: df = n – (k + 1).

$$s^2_{\hat{y}} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - k - 1} \tag{24}$$

To ease the calculation, Equation (24) can be simplified by inserting an average value $\bar{y}$.

$$s^2_{\hat{y},lin} = \frac{(1 - r^2) \sum_{i=1}^{n} (y_i - \bar{y})^2}{n - k - 1};$$
$$s^2_{\hat{y},q} = \frac{(1 - R^2) \sum_{i=1}^{n} (y_i - \bar{y})^2}{n - k - 1} \tag{25}$$

Having defined the concept of effects (21), (23) and residual variance (24), (25), the final test formulas for linear (26) and quadratic (27) effect significance can be obtained.

$$F_{lin} = \frac{r^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}{s^2_{\hat{y},lin}}; \tag{26}$$

$$F_q = \frac{(R^2 - r^2) \sum_{i=1}^{n} (y_i - \bar{y})^2}{s^2_{\hat{y},q}} \tag{27}$$

Lastly, the obtained linear and quadratic effect significances have to be compared with critical values $F^{\#}$ (α; 1; $n - k - 1$) taken from F-distribution tables. If the significance of the effect is less than corresponding critical value, given effect should be considered as not significant. The more the F-value is greater than the $F^{\#}$, the more significant the model is. Preferably F should exceed $F^{\#}$ by one or more orders of magnitude.

This test allows the possibility of having positive results for all tested models. Unlike other tests that give 'zero-one' like answers, the classic F-test results shows quantitative measure of each model's contribution in the data description. Therefore, despite its limited precision, it is a useful test in the evaluation of regression models.

6.1. Fisher-Snedecor Test – Example

The classic F-Snedecor test does not require multiple measurements for each independent variable. Moreover, in the case of multiple y data, y_{ij} values should be averaged beforehand. Therefore, both albumin's linear and quadratic regression models and their determination coefficients were obtained using the averaged data from Table 2.

At first, linear and quadratic effects and residual variance for both models were calculated. Afterward, the significance of both effects was determined.

$$E_{lin} = 0.9668 \cdot 0.2212 = 0.2139;$$
$$E_q = (0.9984 - 0.9668) \cdot 0.2212 = 6.994 \cdot 10^{-3};$$
$$s^2_{\hat{y},lin} = \frac{(1 - 0.9668) \cdot 0.2212}{11 - 1 - 1} = 8.173 \cdot 10^{-4};$$

$$s^2_{\hat{y},q} = \frac{(1-0.9984)\cdot 0.2212}{11-2-1} = 4.520\cdot 10^{-5}$$

$$F_{lin} = \frac{0.9668\cdot 0.2212}{8.173\cdot 10^{-4}} = 2.617\cdot 10^{2};$$

$$F_q = \frac{(0.9984-0.9668)\cdot 0.2212}{4.520\cdot 10^{-5}} = 1.547\cdot 10^{2};$$

$$F^{\#}_{lin} = 5.117;\ F^{\#}_q = 5.318.$$

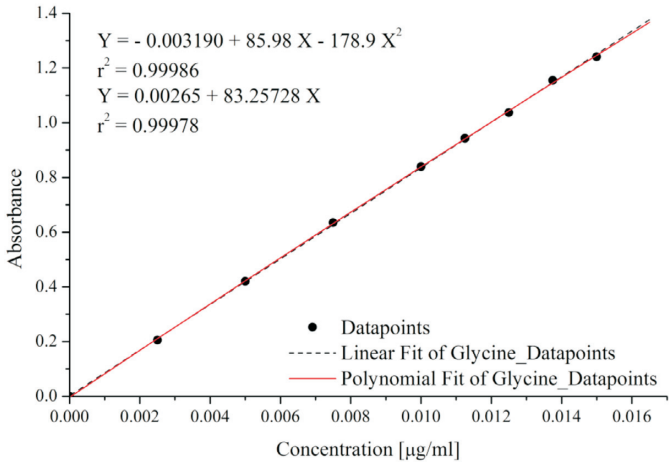

Figure 4 Linear and quadratic regression models for albumin calibration data points (Table 2), determined by Bradford method.

By comparing values of F_{lin} and $F^{\#}_{lin}$ it can be noticed that F_{lin} exceeds the critical value by two orders of magnitude. This indicates high importance of this effect and could be a basis for accepting the linear model. However, the value of quadratic model's significance exceeds its critical value by two degrees of magnitude as well. Hence the conclusion that the quadratic effect is an important contribution and cannot be ignored. This situation is a case where positive results are obtained for both effects.

For comparison, the same test was performed for glycine data (Fig. 5).

All the values were calculated same way as for albumin data.

$$E_{lin} = 1.505;\ E_q = 1.137\cdot 10^{-4};$$

$$s^2_{\hat{y},lin} = 4.8\cdot 10^{-5};\ s^2_{\hat{y},q} = 3.71\cdot 10^{-5}$$

$$F_{lin} = 3.13\cdot 10^{4};\ F_q = 3.07;$$

$$F^{\#}_{lin} = 5.591;\ F^{\#}_q = 5.987.$$

Significance of linear model F_{lin} exceeds the critical value $F^{\#}_{lin}$ by four degrees of magnitude, which states greatly in favour of linear model. Moreover, the significance of quadratic model F_q is lower than its critical value $F^{\#}_q$. Such results provide a strong basis for completely rejecting the significance of the quadratic model and using the linear regression model instead.

7. F-Snedecor Test Modification by IUPAC and Mandel

To prevent the possibility of an inconclusive the classical F-Snedecor test results, the International Union of Pure and Applied Chemistry (IUPAC) adopted its modified version based only on residual variances of both linear and quadratic model.[10] Modified test checks whether the variance explained by the quadratic model is greater than the variance explained by the linear model. The null hypothesis H_0: *'the variance explained by the additional term is not different from the residual variance'* is formulated. Thus, it is assumed that the quadratic model is not significant. Test equation has the form of Equation (28):

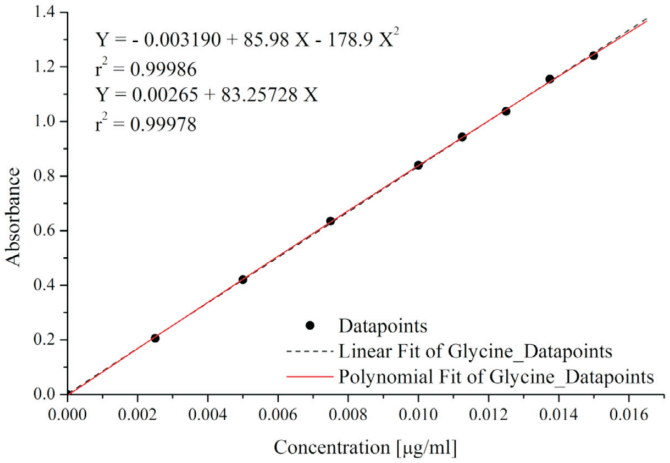

Figure 5 Linear and quadratic regression models for glycine calibration data points (Table 2), determined by Smith method.

$$F_{IUPAC} = \frac{s^2_{\hat{y},lin} - s^2_{\hat{y},q}}{s^2_{\hat{y},q}} \tag{28}$$

The modified test involves comparing the obtained value F_{IUPAC} with the critical value $F^{\#}_{IUPAC}$ (α; 1; $n-3$). If the obtained value is greater than the critical value (exactly opposite to classic *F*-test), the null hypothesis has to be rejected in favour of the alternative hypothesis: the variance explained by the quadratic term is larger than the residual variance. That result would imply the need for a quadratic model.

The F_{IUPAC} test is simple and easy to understand. It does not consider directly the effect of degrees of freedom, which may call into question its usefulness in most cases where the number of calibration points is relatively low. For this reason, in 1964 a chemist and statistician, J. Mandel, suggested an improved version of the test. It was summarized as a comparison of the residual standard deviation of the linear model with that of the nonlinear model.[11] Unlike IUPAC, Mandel defined the *F*-test not as a subtraction of the variances of the two models, but as a subtraction of the sum of squares of the linear and quadratic fits, divided by the difference of their degrees of freedom.

$$F_M = \frac{SS_{lin} - SS_q}{(n-2)-(n-3)}\cdot\frac{1}{s^2_{\hat{y},q}} = \frac{SS_{lin} - SS_q}{s^2_{\hat{y},q}};$$
$$SS = df\cdot s^2_{\hat{y}} \tag{29}$$

Equation (29), can be simplified to a form similar to one specified by IUPAC (30):

$$F_M = \frac{df_{lin}\cdot s^2_{\hat{y},lin} - df_q\cdot s^2_{\hat{y},q}}{s^2_{\hat{y},q}} \tag{30}$$

As can be seen from (28) and (30), the only difference between the IUPAC and Mandel approach is that the latter one includes respective degrees of freedom. This results in a more detailed outcome, particularly in the smaller amount of data available, wherein the differences between both tests are the most visible.

Mandel test, as an IUPAC modification, compares the test to the same critical value $F^{\#}_{IUPAC} = F^{\#}_M$ (α; 1; $n-3$) and has the same condition for accepting or rejecting the null hypothesis.

Due to the use of degrees of freedom in Mandel's equation, it so happens that the result of the test differs sufficiently from IUPAC version. As a consequence, when both results are aligned with the same critical value, the obtained conclusions can be opposite. In such a situation one should accept the Mandel's *F*-test result as a more accurate one.

7.1. IUPAC and Mandel Tests – Example

To concretize the inconclusive results of the F-Snedecor test, both datasets (Table 2) were analyzed using F_{IUPAC} and F_{Mandel} tests. The results of all F_{IUPAC} calculation steps are shown below.

Albumin

$$s^2_{\hat{y},lin} = 8.17 \cdot 10^{-4};\ s^2_{\hat{y},q} = 4.54 \cdot 10^{-5};$$

$$F_{IUPAC} = \frac{8.17 \cdot 10^{-4} - 4.54 \cdot 10^{-5}}{4.54 \cdot 10^{-5}} = 17.075;$$

$$F^{\#}_{IUPAC} = 5.318$$

Glycine

$$s^2_{\hat{y},lin} = 4.80 \cdot 10^{-5};\ s^2_{\hat{y},q} = 3.71 \cdot 10^{-5};$$

$$F_{IUPAC} = 0.2951;\ F^{\#}_{IUPAC} = 5.987$$

As it can be seen for albumin data, the F_{IUPAC} value is three times bigger than its critical value $F^{\#}_{IUPAC}$. This result indicates the need of quadratic model for albumin data.

In turn, test for glycine data showed no basis for quadratic model usage since $F_{IUPAC} \ll F^{\#}_{IUPAC}$. At this point one would say that the situation looks straightforward. However, F_{IUPAC} test, by not including the respective degrees of freedom in its equation, can give misleading results for some datasets[12] This is why even though the situation may look clear, a wise move is to conduct the more accurate form of this test, namely Mandel test. For the same set of data, Mandel test results are shown below:

Albumin

$$s^2_{\hat{y},lin} = 8.17 \cdot 10^{-4};\ s^2_{\hat{y},q} = 4.54 \cdot 10^{-5};$$

$$df_{lin} = 9;\ df_q = 8;$$

$$F_M = \frac{9 \cdot 8.17 \cdot 10^{-4} - 8 \cdot 4.54 \cdot 10^{-5}}{4.54 \cdot 10^{-5}} = 154.673;$$

$$F^{\#}_M = 5.318$$

Glycine

$$s^2_{\hat{y},lin} = 4.80 \cdot 10^{-5};\ s^2_{\hat{y},q} = 3.71 \cdot 10^{-5};$$

$$df_{lin} = 7;\ df_q = 6;$$

$$F_M = 3.066;\ F^{\#}_M = 5.987$$

Since for albumin data $F_M > F^{\#}_M$ and for glycine data $F_M < F^{\#}_M$, the conclusions of Mandel and F_{IUPAC} tests are alike. This confirms that albumin data requires quadratic model to be used while there is no need to do that for glycine data.

8. A Summary of Respective Statistical Tests

For clarity, the results of the four tests application were gathered in Table 3.

As shown in Table 3, for glycine data the linear model was consistently confirmed by all four test results and there is no justification for using the quadratic model.

Albumin data do not give such consistent results. Only three of four conducted tests, namely Lack of Fit, F_{IUPAC} and Mandel tests, stated in favour of quadratic model while classic F-Snedecor test showed a draw. Even though this is enough to indicate the need for quadratic model, as the linear one is not sufficient in case of this dataset.

As it can be seen above, the F_M value for albumin is considerably higher than F_{IUPAC} for the same data. This shows how big of a difference makes including degrees of freedom, mainly for the cases with fewer data points. In such cases, where values given by both tests are close to the test's critical values, one has to evaluate the numerical values obtained from the tests, taking their force into account. Also, one could consider redoing the calibration procedure and adding more experimental points.

Now that the statistical tests are done and the quadratic model was chosen, one could question how greatly the choice of the model impacts the outcome of an experiment that utilizes it to get results. To show that we will once more use albumin data (Table 2, Fig. 4), but this time we will be using the model rather than creating it. Albumin data points were obtained experimentally, hence their (x, y) values are certain. Therefore, one can compare those values with the x values obtained from both models to see how much a calculated value can deviate from an actual, measured value. The results are shown in Table 4.

One can easily see that the values attained from linear model deviate in general from the actual values by almost a full unit of concentration. Therefore using this model could lead to a significant error in the calculated results. On the other hand, the values obtained from the quadratic model deviate in general just by a quarter of a full unit of concentration, which means that this model is about four times less likely to cause result errors.

As much as the quadratic model is more precise in this situation it also comes with some downsides which cannot be forgotten. To properly use this regression model, one has to remember that its sensitivity is not constant along the curve as it is for linear model. Its value is equal to regression model's first derivative $dAbs/dC = -14.27 \cdot 10^{-4} x + 0.03632$. This can be easily shown by testing the model for small absorbance changes. For small absorbance range, e.g. from 0 to 0.1, a little change, $\Delta Abs = 0.05$, corresponds to an equally small change, $\Delta C = 1.408$ in calculated concentration, whilst the same change $\Delta Abs = 0.05$ for final absorbance range from 0.400 to 0.500 corresponds to a much bigger change, $\Delta C = 4.806$. It is visible that at the end of the calibration curve the model is four times less sensitive on measured factor (Abs) than at its beginning. As an effect, expected C result, obtained from final range of the model, is loaded with much greater error. If possible, one should work in the part of quadratic model where the sensitivity is still fairly high what can be achieved, e.g. by respective dilution. It can be assumed that said region is maintained up to the point where the respective derivative is greater than the linear dependence model's slope, i.e. 0.02205. The critical point obtained by simple calculation is at $C = 9.948 \approx 10\ \mu g\ mL^{-1}$, which remains in good agreement with an intuitive observation (Fig. 4). It means, that in considered case one should use the region up to $C = 10\ \mu g\ mL^{-1}$ to obtain reliable results.

Table 3 Comparison of results of four statistical tests applied for albumin and glycine regression models.

Test	Albumin	Glycine
Lack of Fit	Linear Model gives weak fit	Linear Model gives strong fit
Fisher-Snedecor	Linear model is significant Quadratic model is significant	Linear model is significant Quadratic model is not significant
IUPAC	Quadratic model is valid	Linear model is valid
Mandel	Quadratic model is valid	Linear model is valid

Table 4 Output comparison of linear and quadratic models. The reference material was taken from Table 2 for albumin calibration points. The x values for quadratic model were calculated with the use of Wolfram|Alpha software; link https://wolframalpha.com/.

x_{exp}	y_{exp}	Linear Model		Quadratic model	
		x_{calc}	$\Delta x_{lin} = \lvert x_{exp} - x_{calc} \rvert$	x_{calc}	$\Delta x_q = \lvert x_{exp} - x_{calc} \rvert$
0	0	–2.167	2.167	–0.1359	0.1359
2	0.08033	1.476	0.5238	2.168	0.1678
4	0.1357	3.987	0.01270	3.899	0.1014
6	0.2070	7.221	1.221	6.357	0.3572
8	0.2433	8.867	0.8671	7.740	0.2604
10	0.3030	11.57	1.575	10.28	0.2848
12	0.3290	12.75	0.7537	11.54	0.4619
14	0.3703	14.63	0.6267	13.80	0.1969
16	0.4060	16.25	0.2458	16.20	0.1977
18	0.4317	17.41	0.5887	18.41	0.4092
20	0.4443	17.98	2.017	19.80	0.1984
Mean			0.9635		0.2520

In conclusion it can be said that the presented and applied four statistical tests are useful and necessary. They also, through positive feedback, lead to an improved experimental procedures as they oblige researchers to pay more attention to the layout of their experiments. The validation of the adopted regression model adequacy should be based on several rather than one of the provided statistical tests which give a complete picture of reality. All of the discussed tests can be easily done using only basic spreadsheet programs or even just by using a calculator. Therefore, the presented statistical analysis of linear and quadratic regression models is commonly available and should be performed whenever it is possible.

Abbreviations

Symbol	*Meaning*
a	Slope of linear dependence
b	Intercept of linear dependence
s_0	Standard deviation
s_a	Standard deviation of a
s_b	Standard deviation of b
s_y^2	Residual variance
B	Slope in ORIGIN software output
A	Intercept in ORIGIN software output
α	Level of significance
$t(\alpha, df)$	$t_{Student}$ coefficient
df	Degrees of freedom
n	The number of x values
c_i	The number of y_i repetitions for given x
E_{lin}, E_q	Linear and quadratic effects
F_{lin}, F_q	Fisher-Snedecor test values
E_{Lof}	Lack of Fit F-test value
F_{IUPAC}	IUPAC F-test value
F_M	Mandel's F-test value
SSE_{red}	Sum of squares due to reduced model errors
SSE_{full}	Sum of squares due to full model errors
$SSLF$	Sum of squares due to lack of fit
MSE_{full}	Mean squares due to full model errors
$MSLF$	Mean squares due to lack of fit
r^2	Determination coefficient of linear model
R^2	Determination coefficient of quadratic model
$\hat{}$	Symbol of estimated value
$\bar{}$	Symbol of mean (average) value
$^{\#}$	Symbol of critical value

References

1 J. Van Loco, M. Elskens, C. Croux and H. Beernaert, Linearity of calibration curves: use and misuse of the correlation coefficient, *Accred. Qual. Assur.*, 2002, **7**, 281–285.
2 Analytical Methods Committee, Is my calibration linear?, *Analyst*, 1994, **119**, 2363–66.
3 L. Brüggemann, W. Quapp and R. Wennrich, Test for non-linearity concerning linear calibrated chemical measurements, *Accred. Qual. Assur.*, 2006, **11**, 625–631.
4 J.J. Faraway, Practical Regression and Anova using R, Johns Hopkins Bloomberg School of Public Health, 2002, http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch3.pdf
5 J.R. Taylor and A. McGuire, ed., *An Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements*, 2nd edn., University Science Books, Sausalito California, USA, 1997.
6 K. Doerffel, *Statistik in der Analytischen Chemie*, VEB Deutscher Verlag für Grundstoffindustrie Leipzig, 1984, 181–182.
7 J. Wei and D. Paul, *StatWiki* 2015, UC Davis GeoWiki by University of California http://statwiki.ucdavis.edu/Regression_Analysis/Simple_linear_regression/Test_for_Lack_of_Fit
8 L.C. Rodriguez, A.M.G. Campana and J.M.B. Sendra, Statistical estimation of linear calibration range, *Anal. Lett.*, 1996, **29**(7), 1231–1239.
9 J. Czermiński, A. Iwasiewicz, Z. Paszek, A. Sikorski, *Statistical Methods for Experimental Data Processing (Metody Statystyczne w Doświadczalnictwie Chemicznym)*, 2nd. edn., Polish Scientific Publishing House (PWN), Warsaw, Poland, 1974 (in Polish).
10 K. Danzer and L.A. Currie, Guidelines for calibration in analytical chemistry. Part I. Fundamentals and single component calibration (IUPAC Recommendations 1998), *Pure Appl. Chem.*, 1998, **70**(4), 993–1014.
11 J.M. Andrade and M.P. Gomez-Carracedo, Notes on the use of Mandel's test to check for nonlinearity in laboratory calibrations, *Anal. Methods*, 2013, **5**, 1145–1149.
12 M. Zaręba, P.T. Sanecki and R.I. Rawski, Simultaneous determination of thimerosal and aluminum in vaccines and pharmaceuticals with the use of HPLC method, *Acta Chromatographica*, 2015, DOI: 10.1556/1326.2016.28.3.2, to be published in issue no. 3 (September) / 2016.
13 B. George and E.P. Papadopoulos, Heterocycles from N-ethoxycarbonylthioamides and dinucleophilic reagents, *J. Org. Chem.*, 1977, **42**, 441–443

Appendix

Additional Materials

Excel spreadsheets used for all of the applied tests are available on request by e-mail.

Software Alternatives

There is a possibility to substitute both Origin Pro and MS Excel with free open source programs, receiving the same output values. Instead of Origin Pro and MS Excel one can easily apply following programs:

- SciDaVis – https://sourceforge.net/projects/scidavis/
- Libre Office – http://libreoffice.org/download/libreoffice-fresh/