Available online at www.sciencedirect.com

ScienceDirect

International Journal of Pharmaceutics 321 (2006) 12–17

INTERNATIONAL JOURNAL OF PHARMACEUTICS

www.elsevier.com/locate/ijpharm

Minireview

On the misinterpretation of the correlation coefficient in pharmaceutical sciences

J.M. Sonnergaard*

Department of Pharmaceutics and Analytical Chemistry, The Danish University of Pharmaceutical Sciences, Universitetsparken 2, DK-2100 Copenhagen, Denmark

Received 8 May 2006; received in revised form 26 May 2006; accepted 1 June 2006
Available online 12 June 2006

Abstract

The correlation coefficient is often used and more often misused as a universal parameter expressing the quality in linear regression analysis. The popularity of this dimensionless quantity is evident as it is easy to communicate and considered to be unproblematic to comprehend. However, illustrative examples will demonstrate that the correlation coefficient is highly ineffective as a stand-alone quantity without reference to the number of observations, the pattern of the data and the slope of the regression line. Much more efficient quality methodologies are available where the correct technique depends on the purpose of the investigation. These relevant and precise methods in quality assurance of linear regression as alternative to the correlation coefficient are presented.

Keywords: Methodology; Mathematical models; Regression; Correlation; Functional relationship

1. Introduction

From time to time papers which focus on the widespread abuse or misinterpretation of the correlation coefficient, are published in pharmaceutical and related scientific journals (Hahn, 1973; Colburn and Gibaldi, 1978; Hunter, 1981; Galilea, 1995; Van Loco et al., 2002). It appears that every new generation of pharmaceutical scientists needs a brush up regarding the pitfalls in statistical treatment of paired data and in evaluation of the results. Despite the increasing use of statistical methods it is obvious that common knowledge among professional statisticians still is more or less unclear or unknown to scientists in other fields. The problems with accurate handling of regression problems are illustrated by inspection of the 23 research articles appearing in vol. 298, no.1 (2005) of Int. J. Pharm. Sci. In 8 out of the 23 papers the correlation coefficient is used in an erroneous, misleading or superfluous way.

As the name indeed strongly imply the correlation coefficient is only relevant as a statistical parameter in correlation analysis whereas it is a useless statistic in regression analysis of functional relationships. Hunter (1981) stated that: "In fitting functional models, values of r or R^2 close to +1 or -1 do provide an aura of respectability, but not much else." In a statistical textbook (Brøndum and Monrad, 1982) it is categorically established that: "in data material of type 1 (selected values of X and dependent values of Y) the concept of correlation coefficient does not exist."

The squared correlation coefficient is simply a measure of how much of the variation measured as the sum of squares of the Y variable that is accounted for by a mathematical model. The coefficient is inappropriate in evaluation of linearity; it is useless as quality measurement in estimation of parameters or constants and much better methods are available in handling calibration problems. Furthermore, caution in evaluation of the goodness of fit after transformation, especially when the Y values are transformed, should be taken.

The popularity of the correlation coefficient is obvious. It is easy to communicate and to understand for everyone and it is promptly available even on small pocket calculators.

The aim of this paper is to present pitfalls in statistical regression analysis and the frequently observed misuse of the correlation coefficient. The purpose is moreover to show alternative and more informative methods in quality assurance of regression analysis.

* Tel.: +45 35 306 271; fax: +45 35 306 031.
E-mail address: jms@dfuni.dk.

0378-5173/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpharm.2006.06.001

2. The distinction between regression and correlation

In analysis of the relationship between pair-wise variables two principally different but often mixed-up techniques may be used: the regression and the correlation analysis. Regression analysis deal with functional relationships where one or several X variables are independent, i.e. the levels are selected by the investigator, and the Y variable is the dependent or response variable. The purpose for regression analysis may vary from calibration procedures based on standard curves to estimation of parameters, e.g. describing material characteristics or to mathematical modeling of, e.g. manufacturing processes.

In correlation analysis the relationship between two dependent variables are investigated and none of the levels are chosen a priori. An example of correlation is investigation of in vivo/in vitro relationships in biopharmaceutics. The correlation coefficient is also of relevance as a statistic in analytical work when two analytical methods are compared (Chinchilli and Gruemer, 1994).

One of the causes for the many misinterpretations seems to be the fact that though correlation and regression must be interpreted fundamentally different, the calculation and handling of data based on least square techniques is the same.

3. Anscombe's quartet and linearity

Many investigators would consider a R^2 value of 0.99 as a proof of linearity or as a strong verification that a mathematical model is valid. This fundamental misconception is illustrated by the four data sets in Table 1 originally proposed by Anscombe (1973) but here presented in a modified version. The intercept and the slope of the linear regression is the same for all sets. The correlation coefficient (0.99) and the average values of the X and Y values are also identical. Judged from these estimates alone the four data sets appear to be equivalent. However, as visualised in Fig. 1 this is certainly not the case.

Table 1
Anscombe's quartet: four data sets fitted to the same linear model

	Case							
	A		B		C		D	
	X	Y	X	Y	X	Y	X	Y
	10	5.01	10	5.23	10	4.89	8	3.92
	8	3.99	8	4.23	8	3.95	8	3.75
	13	6.12	13	6.35	13	7.15	8	4.14
	9	4.76	9	4.75	9	4.42	8	4.37
	11	5.47	11	5.65	11	5.36	8	4.29
	14	6.99	14	6.62	14	6.77	8	4.01
	6	3.25	6	3.03	6	3.02	8	3.65
	4	1.85	4	1.62	4	2.08	19	9.5
	12	6.37	12	6.03	12	5.83	8	3.71
	7	3.16	7	3.65	7	3.48	8	4.18
	5	2.54	5	2.35	5	2.55	8	3.98
Mean	9.00	4.50	9.00	4.50	9.00	4.50	9.00	4.50
Slope	0.50		0.50		0.50		0.50	
Intercept	0.00		0.00		0.00		0.00	
r	0.990		0.990		0.990		0.990	

Data modified from Anscombe (1973).

Fig. 1A illustrates the ideal pattern with randomly scattered residuals around the regression line. In Fig. 1B an obvious curved relationship gives the same estimated results. This illustrates that linearity never can be verified by a correlation coefficient close to 1. Fig. 1C demonstrates how the effect of a single point as an outlier accounts for the standard deviation. The extreme pattern in Fig. 1D underlines the importance of a visual inspection of data or a more formal analysis of the residuals. These four examples emphasize the statement of Anscombe "a computer should make both calculation and graphs."

In the ICH guideline on validation of analytical procedures (ICH, 2005), it is stated that evaluation of linearity is based on the correlation coefficient, y-intercept, slope of the regression line and residual sums of squares. Furthermore, a plot of the data should be included and a plot of residuals might be helpful.

4. The correlation coefficient and the relationship to the standard deviation of regression

The ANOVA scheme of a simple linear regression model $Y = a + bX$ in Table 2 demonstrates how the total variation ($\mathrm{SS_{tot}}$) of the Y values is separated in two parts: The sum of squares explained by the model ($\mathrm{SS_{mod}}$) and the residual variation not accounted for by the model ($\mathrm{SS_{res}}$). The residual variance (S_0^2) is estimated with $(n-2)$ degrees of freedom.

From this ANOVA the squared correlation coefficient (r^2) is simply defined as

$$r^2 = \frac{\mathrm{SS_{mod}}}{\mathrm{SS_{tot}}} \tag{1}$$

The sum of squares of the model can be calculated as (Draper and Smith, 1981):

$$\mathrm{SS_{mod}} = b^2\,\mathrm{SS}_X \tag{2}$$

where SS_X is the sum of squares of the X values. The total sum of squares may be formulated as

$$\mathrm{SS_{tot}} = \mathrm{SS_{mod}} + S_0^2(n-2) \tag{3}$$

Eq. (1) may now be changed to

$$r^2 = \frac{b^2}{(b^2 + S_0^2 \times n - 2/\mathrm{SS}_X)} \tag{4}$$

From Eq. (4) it is clearly seen that the correlation coefficient is an invalid expression for the scatter around the regression line,

Table 2
ANOVA table for a straight linear regression

Source of variation	Sum of squares (SS)	Degree of freedom (d.f.)	Mean square (MS)
Regression	$\mathrm{SS_{mod}}$	1	$\mathrm{SS_{mod}}/1$
Residual	$\mathrm{SS_{res}}$	$n-2$	$\mathrm{SS_{res}}/(n-2) = S_0^2$
Total	$\mathrm{SS_{tot}}$	$n-1$	

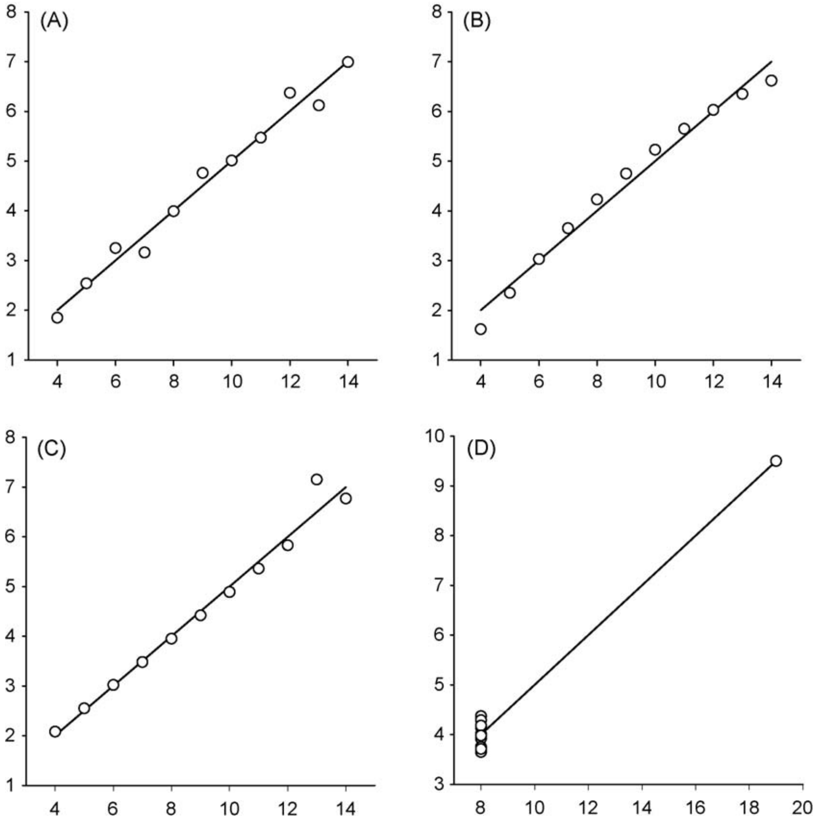

Fig. 1. Four data sets from Table 1 fitted to the same linear model. Data modified from Anscombe (1973).

expressed by S_0. Other important factors are the slope of the regression line, the number of data points and the variability of the X variable. The practical implications of Eq. (4) were confirmed by Colburn and Gibaldi (1978) who investigated the specific effect of different slopes on the estimated correlation coefficient. They concluded that only when the slopes are equal the correlation coefficient reflects the difference in goodness of fit. In a study on correlated data, Lee (1992) noticed that the r-value might be considerably affected by the standard deviation of the X variable.

5. Transformation of variables

Logarithmic and exponential transformations of variables are frequently used in mathematical models. Although informative curved plots today are easy to sketch the models are still often presented as plots in the transformed linear style. This is both unnecessary and rather old-fashioned and might in some cases be directly misleading.

When the goodness of fit is judged solely by the correlation coefficient considerable errors in decision-making may arise. Fig. 2 shows two mathematical models fitted to the same data by a spreadsheet programme (Excel, Microsoft). Apparently the exponential model ($r^2 = 0.974$) gives a better fit than the logarithmic transformation ($r^2 = 0.968$). This conclusion is however erroneous as the exponential model here is computed on the logarithmic transformed Y values ($\ln(Y) = -0.385X + \ln(18.926)$) rather than the initial form. In this way residuals calculated for large values will be underestimated. When the squared correlation coefficient is calculated on the untransformed exponential model the poorer fit to the data is expressed correctly with $r^2 = 0.9279$. It is thus concluded that transformation of in partic-

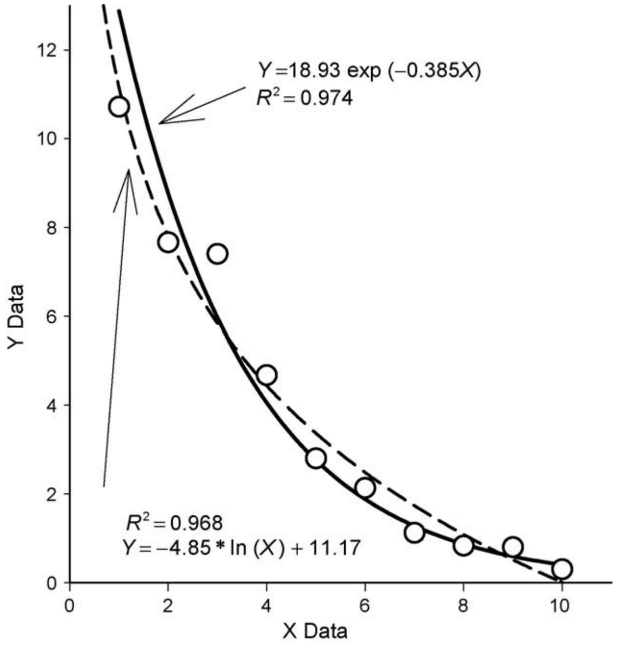

Fig. 2. Two mathematical models: logarithmic (dashed line) and exponential (solid line) fitted to the same data.

ular the *Y* variable may lead to mistakes in evaluating the quality of the fit through the correlation coefficient.

6. Multiple regression analysis

Multiple regression analysis is an extension of the linear regression where one response is related to several independent variables. It is here a common and recognised practice to use the squared correlation coefficient R^2 as a general measure of the goodness of fit to the observed data. If the R^2 value is 0.95 the model accounts for that 95% of the observed variability in the response variable *Y*. The variability term denotes here the summed squares of deviations without any consideration of the number of observations. A 95% reduction of the variance does not however imply that the residual standard deviation is reduced with a similar percentage. In many circumstances it is of interest to examine the residual standard deviation that is not explained by the multiple regression model. The ratio between the residual standard deviation S_0 and the initial total standard deviation of the *Y* variable S_{tot} is from Eqs. (2) and (4) calculated as

$$\frac{S_0}{S_{\mathrm{tot}}} = \sqrt{1 - R^2\left(\frac{n-1}{n-k-1}\right)} \qquad (5)$$

where *k* is the number of estimated coefficients. This relationship is illustrated in Fig. 3 for $n = 10$ and 100 and $k = 1$. When $R^2 = 0.95$ the standard deviation is still only reduced to around 23% and a R^2 value of 0.99 corresponds to 10% of the original standard deviation before the regression analysis was performed. The number of observations (10 and 100) is of minor importance in this situation.

A general dilemma in multiple regression analysis is the essential compromise between an optimal R^2 value and the number of coefficients in the equation. The effect of an unrealistic number of terms is depicted in Fig. 4 where seven observations with replicates are fitted perfectly with a sixth degree polynomial. The fitted curve passes through all the mean values of the observations. If the correlation coefficient is calculated on the mean values rather than on the individual data R^2 will be 1. If the calculation properly is performed on all data with the same equation the squared correlation coefficient will be 0.912. This example thus illustrates that when measurements are replicated it is impossible to obtain a R^2 higher than the value delimited by the measurement error.

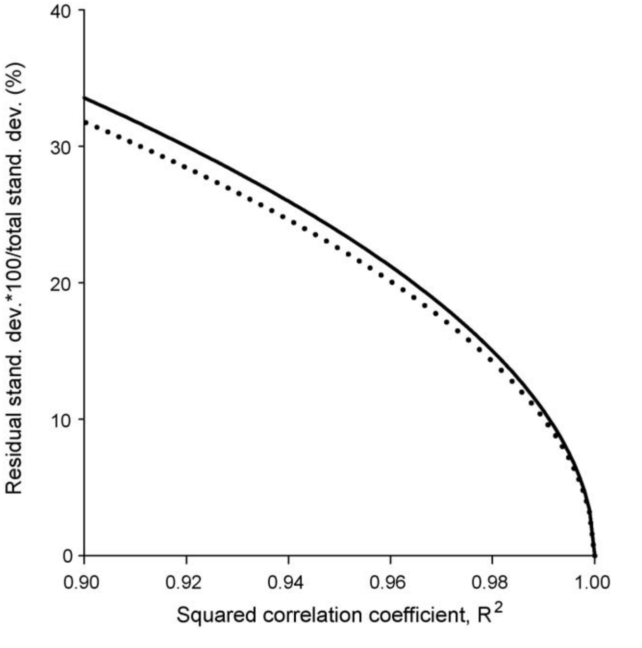

Fig. 3. The residual standard deviation in percent of the total standard deviation versus the squared correlation coefficient. Straight-line linear regression analysis with $n = 10$ (dotted line) and $n = 100$ (solid line).

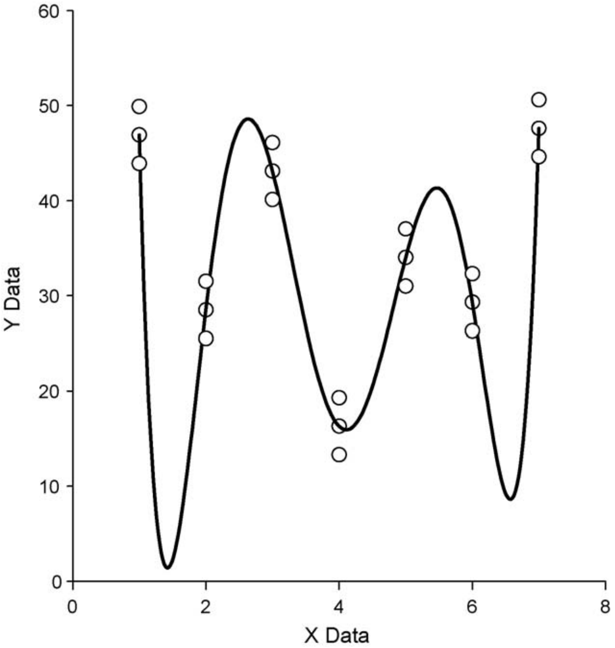

Fig. 4. Sixth degree polynomial fitted to seven points with replicates by the equation: $y = 0.4197x^6 - 10.27x^5 + 99.72x^4 - 486.4x^3 + 1240x^2 - 1538x + 731.3$.

7. Alternative and better ways to do it

As mentioned previously, the optimal way to submit the quality related to regression analysis depends on the purpose of the analysis. The four important subjects considered here shortly are:

- estimation of a characteristic constant (slope or intercept) for a material, a process or a finished product;
- calibration of an analytical method or of a technical measurement in general;
- test of linearity or trace a linear part of an intrinsic curved pattern, i.e. goodness of fit;
- establish the validity of more complicated mathematical models in multiple regression analysis.

7.1. Estimation of parameters

When linear regression is performed in order to estimate either a slope or an intercept as a characteristic of a given condition, the correlation coefficient is of little or no value. It is meaningful and gives much more information to attach the uncertainty as the standard deviation or the confidence interval of the estimated parameters. These standard deviations or their confidence intervals are readily available even in common spreadsheet programs.

The standard deviation S_b of the slope b in a linear model $Y=a+bX$ is thus

$$S_b = \frac{S_0}{\sqrt{\mathrm{SS}_x}} \tag{6}$$

and the standard deviation S_a of the intercept a is

$$S_a = S_0\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_x}} \tag{7}$$

where n is the number of observations.

A major advantage by utilizing these standard deviations is of course that simple statistical tests regarding a difference compared to a constant value or another estimate are easy to perform. If one of the variables and in particular the dependent Y value is transformed care should be taken and it is advisable to plot the data in a XY-plot in their original linear form and to examine the residuals accurately.

7.2. Calibration and standard curves

Linear regression is perhaps most frequently used in calibration procedures where a known concentration of a substance and the associated analytical response is identified. The quality of this regression is of essential importance in evaluation of the future precision utilizing the model. The exact and correct way to establish this quality is to calculate the standard deviation of the predicted value. This is relatively simply done as the standard deviation S_{x_0} of the calculated concentration value x_0 is

$$S_{x_0} = \frac{S_0}{b}\sqrt{1 + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{b^2\,\mathrm{SS}_x}} \tag{8}$$

where y_0 is a given observed response value. Eq. (8) is used as a basis for calculation of a 95% confidence interval or limits of prediction (Johnson, 1994) around the predicted x_0 value.

A somehow overlooked but promising and simple alternative to the correlation coefficient is the quality coefficient (QC) (Van Loco et al., 2002; Laborda et al., 2004). QC is a dimensionless parameter analogous to the well known relative standard deviation. The coefficient is based on the average relative deviation of the data points from the fitted equation:

$$\mathrm{QC} = 100 \times \sqrt{\frac{\sum((y_i - \hat{y}_i)/\hat{y}_i)^2}{n-2}} \tag{9}$$

where y_i is the individual observation and $\hat{y}_i$ is the signal predicted by the linear model. example data from Table 1.

7.3. Testing linearity and selecting linear parts of curved relationships

In general it is only possible to test whether a linear model is valid when repeated observations of the dependent variable is available. Otherwise a linear model may only be assumed by analysis of the residuals where a random scatter without systematic trends must be present. In several cases it is desirable to select a portion of a distinct curved profile and force linear regression model upon the data. In these cases it is preferable to investigate the goodness of fit by drawing a 95% confidence lines around the regression lines. These curves are available in more sophisticated statistical programs, but is easily constructed in a spreadsheet (Sonnergaard, 2006).

7.4. Multiple regression

The squared correlation coefficient (R^2) is an accepted standard quality parameter in multiple regression analysis where for instance data are collected from a factorial designed experiment and optimisation procedures are performed on an empirical model. Without going in detail multiple regression is executed most efficiently by use of the stepwise regression technique as described by Davies and Goldsmith (1976) and available in software packages like Statistica from StatSoft. Potential terms are entered or eliminated in the equation (forward selection and backwards elimination). The steps are repeated until a constrained compromise between a maximal explanation expressed as R^2 and a sufficient and relevant number of coefficients are obtained. A more pragmatic and straightforward approach in selecting the best subset of variables would be to compare the F-statistic of the regression with and without a specific term. This is readily done in spreadsheet programs like Excel using the Data analysis Add-Inn.

8. Conclusive remarks

Despite the many warnings listed in this paper the correlation coefficient still will undoubtedly be used by many investigators (including this author) as a coarse and easy accessible parameter. However, it should never be used in serious scientific work as a stand-alone quality parameter when functional relationships are analysed.

References

Anscombe, F.J., 1973. Graphs in statistical analysis. Am. Stat. 27, 17–21.

Brøndum, L., Monrad, J.D., 1982. Statistik II. Anvendt statistik, Den Private Ingeniørfond København, p. 459.

Chinchilli, V.M., Gruemer, H.D., 1994. The correlation coefficient in the interpretation of laboratory data. Clin. Chim. Acta 229, 1–3.

Colburn, W.A., Gibaldi, M., 1978. Correlation analysis and goodness of fit: a commentary. Can. J. Pharm. Sci. 13, 31–32.

Davies, O.L., Goldsmith, P.L., 1976. Statistical Methods in Research and Production, 4th ed. Longman Group Ltd., London, pp. 264–268.

Draper, N.R., Smith, H., 1981. Applied Regression Analysis, 2nd ed. John Wiley & Sons, New York, p. 18.

Galilea, P.A., 1995. Inappropriate use of the correlation coefficient. Int. J. Sports Med. 16, 338–339.

Hahn, G.J., 1973. The coefficient of determination exposed! ChemTech 3, 609–612.

Hunter, J.S., 1981. Calibration and the straight line: current statistical practices. J. Assoc. Off. Anal. Chem. 64, 574–583.

ICH Guideline Q2(R1), 2005. Validation of analytical procedures: text and methodology. http://www.ich.org/LOB/media/MEDIA417.pdf.

Johnson, R.A., 1994. Miller & Freund's Probability & Statistics for Engineers, 5th ed. Prentice-Hall, New Jersey, p. 342.

Laborda, F., Medrano, J., Castillo, J.R., 2004. Influence of the number of calibration points on the quality of results in inductively coupled plasma mass spectrometry. J. Anal. At. Spectrom. 11, 1434–1441.

Lee, J., 1992. A cautionary note on the use of the correlation coefficient. Br. J. Ind. Med. 49, 526–527.

Sonnergaard, J.M., 2006. Quantification of the compactibility of pharmaceutical powders. Eur. J. Pharm. Biopharm. 63, 270–277.

Van Loco, J., Elskens, M., Croux, C., Beernaert, H., 2002. Linearity of calibration curves: Use and misuse of the correlation coefficient. Accred. Qual. Assur. 7, 281–285.