Accred Qual Assur (2002) 7:281–285
DOI 10.1007/s00769-002-0487-6

PRACTITIONER'S REPORT

Joris Van Loco
Marc Elskens
Christophe Croux
Hedwig Beernaert

Linearity of calibration curves: use and misuse of the correlation coefficient

Received: 15 January 2002
Accepted: 18 April 2002

J. Van Loco (✉) · H. Beernaert
Scientific Institute of Public Health,
J. Wytsmanstraat 14, 1050 Brussels,
Belgium
e-mail: Joris.vanloco@iph.fgov.be
Tel.: +32–2-6425356
Fax: +32–2-6425691

M. Elskens
Vrije Universiteit Brussel,
Laboratory of Analytical Chemistry,
Pleinlaan 2, 1050 Brussels, Belgium

C. Croux
Department of Applied Economics,
K.U. Leuven, Naamsestraat 69,
3000 Leuven, Belgium

Abstract The correlation coefficient is commonly used to evaluate the degree of linear association between two variables. However, it can be shown that a correlation coefficient very close to one might also be obtained for a clear curved relationship. Other statistical tests, like the Lack-of-fit and Mandel's fitting test thus appear more suitable for the validation of the linear calibration model. A number of cadmium calibration curves from atomic absorption spectroscopy were assessed for their linearity. All the investigated calibration curves were characterized by a high correlation coefficient (r >0.997) and low quality coefficient (QC <5%), but the straight-line model was systematically rejected at the 95% confidence level on the basis of the Lack-of-fit and Mandel's fitting test. Furthermore, significantly different results were achieved between a linear regression model (LRM) and a quadratic regression (QRM) model in forecasting values for mid-scale calibration standards. The results obtained with the QRM did not differ significantly from the theoretically expected value, while those obtained with the LRM were systematically biased. It was concluded that a straight-line model with a high correlation coefficient, but with a lack-of-fit, yields significantly less accurate results than its curvilinear alternative.

Keywords Linearity · Goodness of fit · Correlation coefficient · Lack-of-fit · Calibration

Introduction

The linear range of most analytical instruments is known to be limited. Therefore, during method validation the linearity of the calibration curve should be assessed and the working range of the calibration curve should be determined. [1, 2]. The correlation coefficient (r) is commonly used for this purpose, and curves with $r \geq 0.995$ are usually considered to be linear. Nevertheless, several investigators focussed on the fact that r might not be a useful indicator of linearity [2, 3], and other statistical tests or quality parameters have been suggested to ascertain the goodness of fit of the calibration curve [2, 4, 5].

On the contrary, a calibration curve with $r \geq 0.995$ can be considered nearly linear. Furthermore, from an inference point of view, linear regression models (LRMs) are easy to implement, compared to curvilinear or non-linear regressions models [6]. Therefore, a straight-line calibration curve should always be preferred over curvilinear or non-linear calibration models if equivalent results can be gained. A prerequisite, however, is that one should be able to assess this equivalence. In other words, is there any evidence for a systematic difference between the results of the two models at a given confidence level?

In this paper, cadmium calibration curves from atomic absorption spectroscopy (AAS) were tested for their linearity. Alternative curvilinear regressions were proposed when linearity was rejected. Predictions made on the basis of the fitted curve for both linear and curvilinear models were compared.

Table 1 The *F*-value of the Lack-of-fit (LOF) test ($F_{crit,95\%}$ = 4.53) and Mandel's fitting test ($F_{crit,95\%}$= 5.12) are compared with the quality coefficient and the correlation coefficient for several linear calibration lines of Cd. For the quadratic regression model, the *F*-value of the Lack-of-fit test ($F_{crit,95\%}$=4.76) and the *P*-value for testing significance of the second order coefficient for the quadratic regression model are represented. The significant values at the 95% confidence level are underlined

Linear regression model				Quadratic regression model	
LOF	Mandel's test value	QC (%)	r	LOF	P-value on second-order coefficient
11.08	51.46	3.93	0.9982	0.63	0.0000
19.42	56.84	4.23	0.9978	1.58	0.0000
7.13	26.29	3.67	0.9985	0.94	0.0006
6.99	37.73	3.79	0.9984	0.18	0.0002
11.43	58.21	4.03	0.9981	0.31	0.0000
29.91	53.02	3.53	0.9986	4.08	0.0000
49.80	71.07	3.76	0.9984	5.69	0.0000
23.77	73.86	3.19	0.9989	1.66	0.0000
31.95	63.37	3.24	0.9988	3.55	0.0000
7.49	33.50	2.92	0.9991	0.54	0.0003
9.99	55.19	3.95	0.9983	0.15	0.0000
10.71	28.65	4.70	0.9975	1.89	0.0005
25.21	79.60	3.34	0.9987	1.62	0.0000
13.16	35.74	3.37	0.9987	1.93	0.0002

Assessing the linearity of calibration curves

Graphite furnace atomic absorption spectroscopy (GF-AAS) is known to have a limited linear calibration range. In order to assess the linearity of the calibration process, several calibration lines for cadmium were constructed over a period of 4 months. These calibrations were performed using standard solutions prepared from the corresponding high purity metal Baker Cd Atomic Absorption Standard of 1000 µg/ml (National Institute for Standards and Technology – NIST traceable). The GF-AAS was programmed to produce a calibration curve with the following concentrations: 0, 0.8, 1.6, 2.4, 3.2 and 4.0 ng/ml. The solutions were injected in duplicate.

The linearity of the calibration process was investigated by means of the Lack-of-fit test [2], Mandel's fitting test value [5], the quality coefficient (QC) [2,4] and *r* [3,2]. The results are summarized in Table 1.

The Lack-of-fit test and Mandel's fitting test are commonly used to ascertain whether the chosen regression model adequately fits the data. The test values for these two statistical tests follows an *F*-distribution and the significance of the test values can be calculated. On the contrary, the QC and *r* are used to arbitrary accept or reject the LRM. The equations of the QC and *r* for LRMs are given below:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (1)$$

$$QC(\%) = 100 \sqrt{\frac{\sum \left(\frac{Y_i - \hat{Y}_i}{\bar{Y}} \right)^2}{n-1}} \quad (2)$$

with Y_i the measured response and $\hat{Y}_i$ the response predicted by the model.

The results in Table 1 shows that for the Lack-of-fit test the LRM must systematically be rejected at the 95% confidence level ($F_{crit,95\%}$ = 4.53), and for Mandel's fitting test even rejected at the 99% confidence level ($F_{crit,99\%}$ = 10.56). Thus, despite the fact that *r* and *QC* are greater than 0.997 and lower than 5%, respectively, the linearity of the calibration lines were rejected on the basis of the before mentioned *F*-tests. This corroborates the statements of the Analytical Methods Committee [3] that *r* should be used with care when evaluating the linearity of calibration lines. Moreover, questions arise regarding the significance of *QC*, for which an upper limit of 5% was proposed in assessing the suitability of a calibration process [4]. Here, even with a *QC*-value less than 3%, the LRM is rejected at the 95% confidence level (Table 1).

Alternatively, the residual plots give useful information to validate the chosen regression model. The residual plot can be used to check whether the underlying assumptions, like normality of the residuals and homoscedasticity, are met as for evaluating the goodness of fit of the regression model [2]. Figure 1a shows a residual plot for an LRM. The U-shaped residual plot indicates that a curvilinear regression model should be preferred over an LRM.

Several authors [2, 6, 8–11] recommended alternative calibration functions when linearity of the calibration curve has to be rejected. In order to correct the non-linearity, a quadratic curvilinear function ($f(x) = a + bx + cx^2$) was chosen. The Lack-of-fit tests for the quadratic regression model (QRM) are summarized in Table 1. The test for Lack-of-fit reveals that this QRM adequately fits the calibration data at 99% (highly significant) confidence level and at the 95% (significant) confidence level in all cases except one. In determining whether the order of the polynomial regression model is appropriate, the

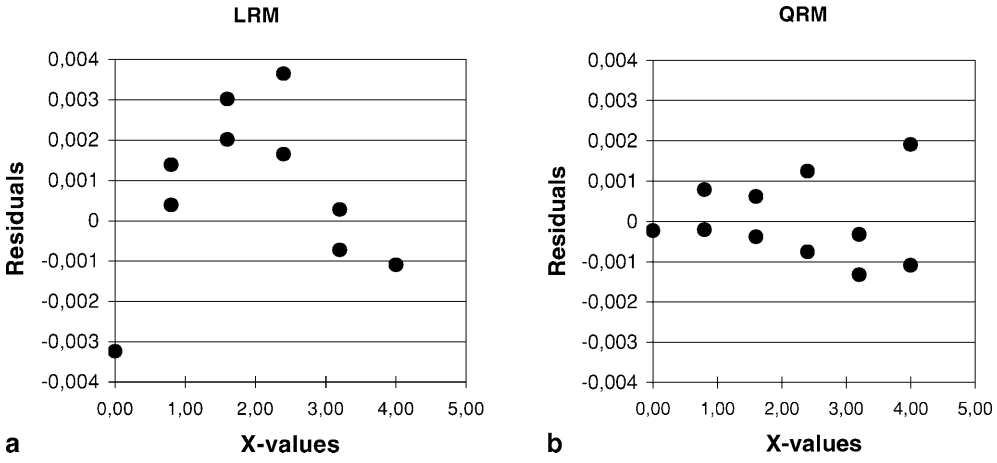

Fig. 1 Plots of residuals for (**a**) the linear regression model (LRM) and (**b**) the quadratic regression model (QRM) versus predicted values

significance of the second order coefficient is estimated. The *P*-value on the second order coefficient, shown in Table 1, is systematically smaller than 1%. Consequently, a lower order model should not be considered. In addition, residual plots (Fig. 1b) were constructed for this QRM. The residuals were randomly scattered within a horizontal band around the centre line. Therefore, the QRM was chosen as the reference model. It is noted that an increase of the variance is observed at higher concentrations.

Predictions made on the basis of the fitted curve for linear (LRM) and quadratic (QRM) regression models

To gauge the agreement/disagreement between predicted concentrations calculated from the LRM and the QRM, a mid-scale calibration standard (2 ng/ml) was systematically injected in duplicate. The instrument signal corresponds to a point close to the centroid of the data cloud, where the confidence limits for the regression line of LRM is the narrowest.

To compare the outcome of both regression models, the predicted concentration of the mid-scale standard was expressed both as a recovery rate and as a relative deviation. Hence, the following equations were used:

$$\text{Recovery (\%)} = \frac{\text{determined concentration}}{\text{nominal concentration}} \times 100\% \qquad (3)$$

$$\text{Relative deviation (\%)} = \left| 100\text{-recovery (\%)} \right| \qquad (4)$$

In order to investigate possible effects of time, several calibration lines were produced over a period of almost 4 months. The mid-scale standard was determined twice at the beginning and at the end of an analysis. The recovery rate and relative deviation of the results are summarized in Table 2.

If both curves yield equivalent results and are not biased, the recovery rate should be around 100%. Figure 2

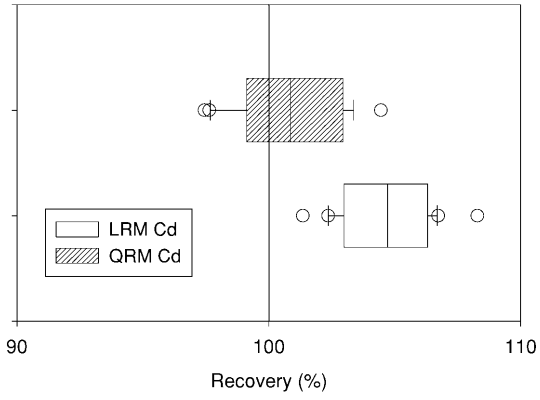

Fig. 2 Recovery results for the mid-scale standard calculated with a second-order calibration curve (QRM) and a linear calibration curve (LRM)

shows the recovery rates for the mid-scale standard calculated with the LRM and the QRM. It clearly appears that the median from the LRM differs from the theoretical value of 100%. In general, the recovery rates are overestimated when calculated with the LRM. A systematic error of about 4% was found.

This result is supported by the Wilcoxon one-sample test [12]. The null hypothesis (median = 100%) is rejected for the results determined with the LRM ($P = 0.0004$), but not for those derived from the QRM ($P = 0.1788$). The one-sample student *t*-test gave similar outcomes (LRM: $P = 8.3.10^{-8}$, QRM: $P = 0.177$).

Furthermore, to gauge whether the results calculated with the QRM were more accurate than those obtained from the LRM the relative deviation of the data were compared. The non-parametric paired “sign” test was chosen for this evaluation because the relative deviations were not symmetrically distributed. The “sign” test is an alternative to the paired *t*-test, when the underlying assumptions of the student *t*-test are violated. The “sign” test makes use of positive and negative signs depending

Table 2 Recovery and relative deviations for the linear (LRM) and quadratic (QRM) calibration curves for cadmium.

	LRM		QRM	
	Recovery	Relative deviation	Recovery	Relative deviation
	104.8	4.8	100.8	0.87
	103.1	3.1	99.1	0.86
	103.6	3.6	99.8	0.22
	108.3	8.3	104.5	4.5
	104.7	4.7	101.1	1.1
	106.2	6.2	102.7	2.7
	105.0	5.0	101.5	1.5
	106.7	6.7	103.2	3.2
	104.1	4.1	99.8	0.18
	102.4	2.4	98.1	1.9
	102.4	2.4	97.4	2.6
	105.9	5.9	101.0	0.97
	106.4	6.4	103.4	3.4
	106.4	6.4	103.4	3.4
	101.4	1.4	97.6	2.4
	102.8	2.8	99.1	0.90
Mean	104.6	4.6	100.8	1.9
Standard deviation	1.9	1.9	2.2	1.3
Median	104.7	4.7	100.9	1.7
1st Quartile	103.1	3.0	99.1	0.88
3rd Quartile	106.3	6.3	102.8	2.9

on the difference between the values in conditions 1 and 2. The null-hypothesis is rejected if the number of positive and negative signs is statistically different. The null-hypothesis stating equivalent results from both models must be rejected in this case (n- = 14, n+ = 2, P <0.006). Therefore, our results indicate a systematic bias on the forecast concentration when applying the LRM. The relative deviation for the QRM is significantly smaller than the one obtained from the LRM. This means that the bias for the results obtained with LRM is larger than the bias of the results from the QRM.

On the contrary, there is no evidence for a difference in the spread or dispersion of the results between both models, which has been confirmed with the Levene test for homogeneity of variances. (F_{Levene} = 2.70; P = 0.11) [13].

Conclusion

As claimed by several investigators [2, 3], this paper corroborates the fact that the correlation coefficient is not a useful indicator of linearity in the calibration model, even for r-values >0.997. In addition, the present results raise the question about the relevance of the QC in assessing the process calibration. Other statistical tests like the Lack-of-fit and Mandel's fitting test seems more appropriate for evaluating the linearity of the calibration curve during method validation. Preferably, the Lack-of-fit and Mandel's fitting test should be used in conjunction with an evaluation of the residual plot.

Furthermore, it is shown that a straight-line model with r >0.997 and QC <5% but with Lack-of-fit, yielded forecast values for a mid-scale calibration standard that significantly differ from the nominal ones. In general, the recovery tests were overestimated, while the precision on the result was comparable in both the LRM and QRM. The bias can be considered as significant since the repeatability of the injection of standard solutions is usually less than 2% relative standard deviation (RSD). Furthermore, the situation would be even worse if the comparison had been carried out with either high- or low-range calibration standards.

In conclusion, the results in this paper indicate that the correlation coefficient is not suitable for assessing the linearity of calibration curves. Statistical tests like the Lack-of-fit and Mandel's fitting test should be systematically applied during full method validation. It was shown that in this application cadmium concentrations calculated with the LRM were constantly overestimated by about 4%.

References

1. European Commission 90/515/EEC (1990): Laying down the reference methods for detecting residues of heavy metals and arsenic. EEC Official Journal L286: 33–39
2. Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J (1997) Handbook of chemometrics and qualimetrics: Part A. Elsevier, Amsterdam
3. Analytical Methods Committee (1988) Analyst 113: 1469–1471
4. Vankeerberghen P, Smeyers-Verbeke J (1992) Chemometrics Intell Lab Syst 15: 195–202

5. Mandel J (1964) The statistical analysis of experimental data. Wiley, New York
6. Ratkowsky DA (1990) Handbook of nonlinear regression models. Marcel Dekker, New York
7. ISO 11843–2: 2000 (2000) Capability of detection – Part 2: Methodology in the linear calibration case. International Organization for Standardization (ISO), Geneva
8. ISO 8466–2: 2001 (2001) Water quality – Calibration and evaluation of analytical methods and estimation of performance characteristics, Part 2: Calibration strategy for nonlinear second-order calibration functions. ISO, Geneva
9. MacTaggart DL, Farwell SO (1992) J AOAC Int 75: 594–608
10. Wang X, Smeyers-Verbeke J, Massart DL (1992) Analusis 20: 209–215
11. Funk W, Dammann V, Donnevert G (1995) Quality assurance in analytical chemistry. VCH, Weinheim
12. Sheskin DJ (2000) Handbook of parametric and non-parametric statistical procedures, 2nd edn. Chapman and Hall/CRC Press, Boca Raton
13. Levene H (1960) In: Olkin I. et al. (eds) Contributions to probability and statistics: Essays in honor of Harold Hotelling. Stanford University Press, Stanford, pp 278–292