*Journal of Analytical Toxicology*, 2022, **46**, 443–448
DOI: https://doi.org/10.1093/jat/bkab036
Advance Access Publication Date: 13 April 2021
**Technical Note**

# Squaring Things Up with $R^2$: What It Is and What It Can (and Cannot) Tell You

**Félix Camirand Lemyre[1,2,3,§], Kevin Chalifoux[1,4], Brigitte Desharnais[4,*] and Pascal Mireault[4]**

[1]Department of Mathematics, Université de Sherbrooke, 2500 Boulevard de l'Université, Sherbrooke, QC J1K 2R1, Canada
[2]School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia
[3]S-POP Axis, Centre de recherche du Centre hospitalier universitaire de Sherbrooke, 12th Avenue North, Sherbrooke, QC J1H 5N4, Canada
[4]Department of Toxicology, Laboratoire de sciences judiciaires et de médecine légale, 1701 Parthenais Street, Montréal, QC H2K 3S7, Canada

[§]Authors are listed alphabetically and order is not indicative of contribution.
*Author to whom correspondence should be addressed. Email: brigitte.desharnais@msp.gouv.qc.ca



## Abstract

The coefficient of correlation ($r$) and the coefficient of determination ($R^2$ or $r^2$) have long been used in analytical chemistry, bioanalysis and forensic toxicology as figures demonstrating linearity of the calibration data in method validation. We clarify here what these two figures are and why they should not be used for this purpose in the context of model fitting for prediction. $R^2$ evaluates whether the data are better explained by the regression model used than by no model at all (i.e., a flat line of slope = 0 and intercept $\bar{y}$), and to what degree. Hopefully, in the context of calibration curves, the fact that a linear regression better explains the data than no model at all should not be a point of contention. Upon closer examination, a series of restrictions appear in the interpretation of these coefficients. They cannot indicate whether the dataset at hand is linear or not, because they assume that the regression model used is an adequate model for the data. For the same reason, they cannot disprove the existence of another functional relationship in the data. By definition, they are influenced by the variability of the data. The slope of the calibration curve will also change their value. Finally, when heteroscedastic data are analyzed, the coefficients will be influenced by calibration levels spacing within the dynamic range, unless a weighted version of the equations is used. With these considerations in mind, we suggest to stop using $r$ and $R^2$ as figures of merit to demonstrate linearity of calibration curves in method validations. Of course, this does not preclude their use in other contexts. Alternative paths for evaluation of linearity and calibration model validity are summarily presented.

## Introduction

Statistical tools allow for a better understanding of the underlying structure of the data and their correct interpretation in any quantitative study. The coefficient of correlation ($r$) and the coefficient of determination ($R^2$ or $r^2$) have long been used as indicators of linear dependency between two variables (1, 2). Given that analytical chemistry, bioanalytical and forensic toxicology method validation guidelines typically include a requirement to demonstrate the linearity of the data or confirm the adequacy of the calibration model used (3–5), it seemed natural to use the $r$ or $R^2$ indicators for this purpose. However, some authors have warned for quite some time about the problems of deducing a linear relationship from this indicator in the context of model fitting (1, 2, 6, 7), while some guidelines explicitly discourage this practice (4, 5). Somewhat confusingly, the American Academy of Forensic Sciences (AAFS) Standards Board's 'Standard Practices for Method Validation in Forensic Toxicology' state that the 'calibration model shall not be evaluated simply via its correlation coefficient (r)', but that 'assessment of coefficient of determination ($r^2$) for linear models' is an 'appropriate alternative' among others (also cited are the residuals plot, analysis of variance (ANOVA) lack-of-fit test and significance of the second-order term of quadratic models) (3). The use of $r$ or $R^2$ as a figure demonstrating linearity of the data in method validation has been widespread and persists to this day, see for example a sampling of recent publications (8–18).

In what follows, we hope to clarify what $r$ and $R^2$ are and what they mean with regard to a regression. In combination with practical examples, it should become clear to the reader why their use as figures of merit in an analytical method validation comes with several limitations.

## The Basics: A Definition of $r$ and $R^2$

The correlation coefficient, $r$ or $r_{xy}$, is defined by the following equation (6):

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

Where $S_{XY}$, $S_{XX}$ and $S_{YY}$ are sum of squares about the mean; $x_i$ and $y_i$ are the $x$ and $y$ values for the $i^{\text{th}}$ sample; $\bar{x}$ and $\bar{y}$ are the averages over all samples.

The correlation coefficient can be understood as the ratio of the degree to which $x$ and $y$ vary together over the degree to which $x$ and $y$ vary separately (6). $r$ is bounded, varying

between −1 (decreasing or negative slope) and +1 (increasing or positive slope).

The coefficient of determination, $R^2$ or $r^2$, as its nomenclature suggests, is the square of the coefficient of correlation. It can be defined as (6):

$$R^2 = \frac{SS_{Reg}}{S_{YY}} = 1 - \frac{SSE}{S_{YY}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (2)$$

Where $SS_{Reg}$ is the sum of squares explained by the regression, $SSE$ is the residual sum of squares, $\hat{y}_i$ is the $y$ value predicted by the model for the $i^{th}$ sample and $S_{YY}$, $y_i$ and $\bar{y}$ are as defined above.

The coefficient of determination is also bounded, with a value between 0 and 1—which follows logically from the fact that it is the square of a value between −1 and +1. Furthermore, $R^2$ will always be lower than the absolute value of the coefficient of correlation, $|r|$, for a given dataset. Intuitively, $R^2$ is the ratio of the variation of $Y$ explained by the regression over the total variation of the variable.

In what follows, for the sake of simplicity we will refer to the coefficient of determination ($R^2$) only. However, remember that this figure is simply the square of the coefficient of correlation ($r$) and as such, the same caveats apply to this indicator.

## What Does It Mean?

Essentially, $R^2$ evaluates whether the data are better explained by a linear regression than by no model at all (i.e., a flat line of slope = 0 and intercept $\bar{y}$), and to what degree. In the context of a calibration curve however, hopefully it is clear from the get-go that the dataset is much better explained by the model than none at all!

$R^2$ assumes that the regression used is the correct model to represent the data and evaluates the strength of association between the two variables under this model (6). If a linear regression is used, $R^2$ assumes that this is the adequate model; it therefore cannot indicate whether the relationship between the variables is linear. Nor can it alone disprove that there is a relationship between the two variables: an $R^2$ of zero can be obtained with data where a functional relationship clearly exists (1, 6). Figure 1 shows one such example: a mathematical function describes the data (quadratic equation generating a parabola), yet the $R^2$ is null.

## Why It Is Problematic in the Model Fitting Context

In the following section, we will present some practical examples illustrating why the use of $r$ and $R^2$ are problematic in a model fitting context (such as evaluating calibration curves in method validation). Modeling and plotting were carried out in RStudio (R version 4.0.0, RStudio version 1.2.5019). Details about the models simulated as well as the R scripts are available in Supplementary Data 1.

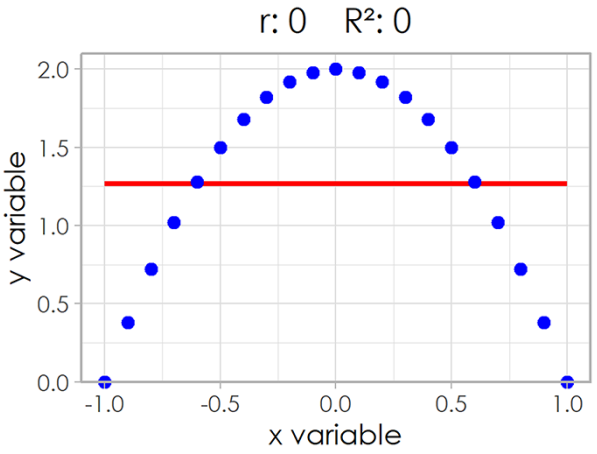


**Figure 1.** The relationship between the two variables x and y is clearly described by a mathematical model (i.e., a functional relationship exists). Nonetheless, r = 0 and $R^2 = 0$, demonstrating that these values cannot prove the absence of a relationship between variables.

### $R^2$ does not indicate if your dataset is linear; it assumes linearity

$R^2$ is often presented as a means to assess linearity in method validation, i.e., to evaluate whether the data produced by the method under evaluation are linear. While different thresholds are quoted for 'acceptable linearity', an $R^2$ above 0.9 is often cited as a criterion (2, 9, 14–16). Unfortunately, $R^2$ evaluates the strength of the association between the $x$ and $y$ variables *assuming* that the regression model used is the adequate one. Insofar as a linear regression is used with the dataset, $R^2$ will evaluate the strength of the association assuming linearity. To demonstrate that this is the case, one needs only to look at the cases presented in Figure 2. In Figure 2A, a quadratic bond between variables $x$ and $y$ is displayed by the curvature in the data points. Yet the coefficient of determination for a linear regression is >0.99. In Figure 2B, the linear relationship breaks down at higher concentration levels due to saturation of the detector. Yet, the coefficient of determination is 0.99. These examples show that high coefficient of determination values do not demonstrate the linearity of the data. In these cases, opting for a linear model on the basis of the $R^2$ result would mean using an erroneous model to represent the data (and quantify concentration in unknown samples). This is not simply a theoretical distinction; using a linear model instead of the correct quadratic model for example can have important impacts on the quantification results (2, 19, 20).

### $R^2$ is influenced by the variability of the data

By definition, an increase in the variance of the data will cause a decrease in $R^2$. Indeed, as volatility in the dataset increases, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ will increase in Equation (2), causing the $R^2$ value to drop. Hence, datasets generated with identical models, but with different variance, will yield different $R^2$ (Figure 3). The mathematical definition of $R^2$ means such behavior is expected; nonetheless, it means that divergent $R^2$ values for two datasets might be attributable to differences in precision rather than the targeted characteristics of data linearity or regression model adequacy. In a method

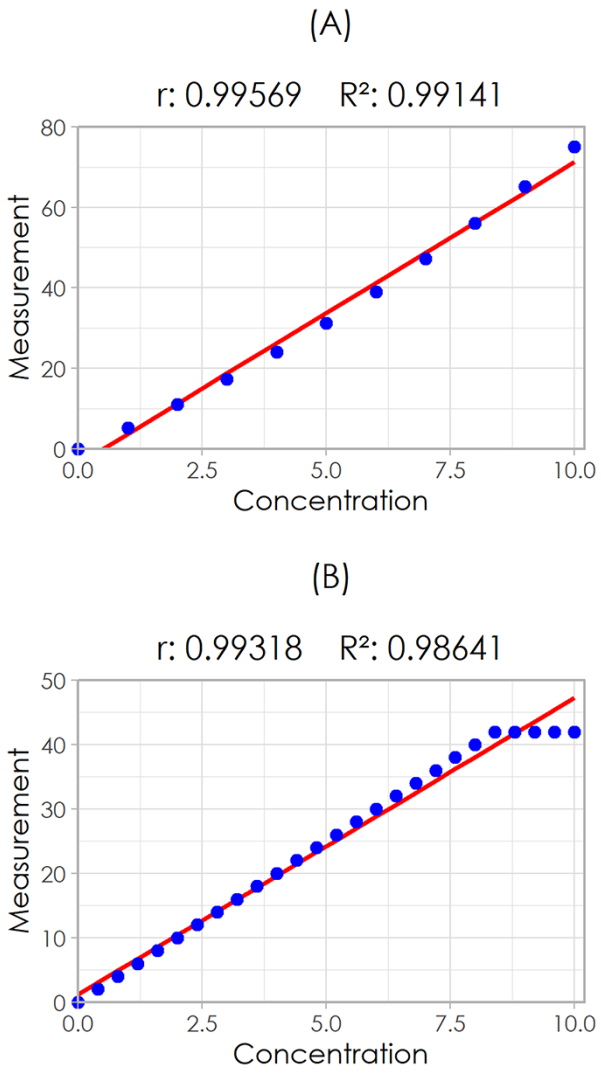


**Figure 2.** (A) A quadratic bond exists between the variables x and y, yet the coefficient of determination is high (>0.99). (B) Linearity breaks down at higher concentrations due to saturation of the detector, yet the coefficient of determination is high (0.99).

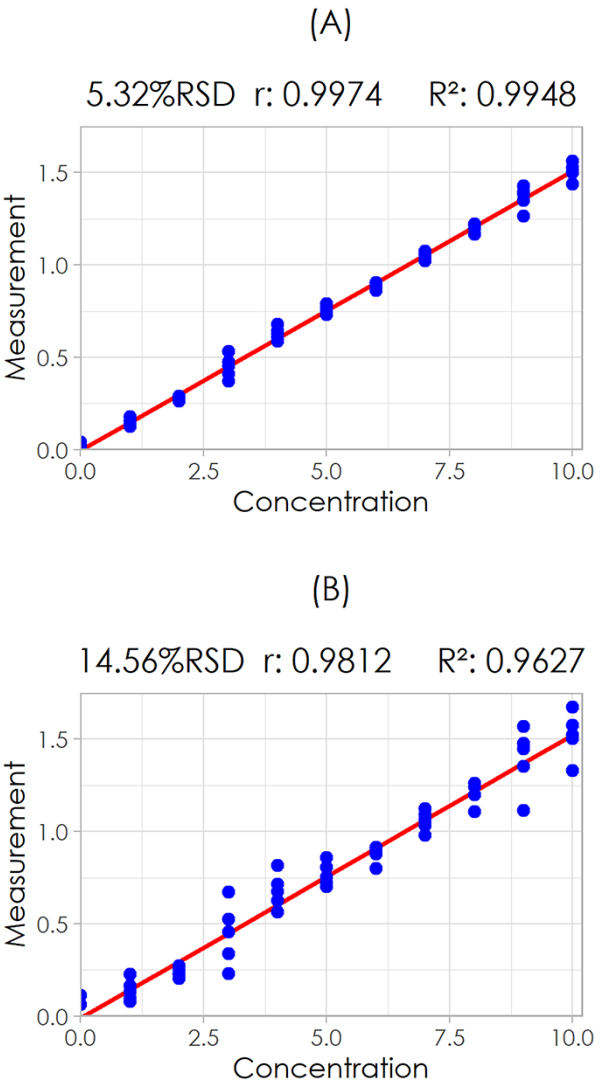


**Figure 3.** Two calibration datasets generated from the exact same linear model ($y = 0.15x$) will yield different coefficients of determination if their precision is different. (A) %RSD ≈ 5%; $R^2 = 0.9948$. (B) %RSD ≈ 15%; $R^2 = 0.9627$, lowered by the increased variability in the data.

validation setting, intra- and inter-day precision is better evaluated via the complete experiments designed for this specific purpose.

## $R^2$ is influenced by the slope of the regression

The slope of the linear regression also has an impact on the $R^2$ result. All other things being equal, the steeper the calibration curve, the higher the $R^2$ will be (7). This is shown by Figure 4, where increasing the slope by 150% yields an increase in $R^2$ from 0.8948 to 0.9807. This is why some references label $R^2$ as 'both a measure of goodness of fit and of steepness of the regression surface' (in a two-variable case, steepness of the regression line) (7). While again, this is an expected behavior of this indicator, it means that divergent $R^2$ values for two datasets might be attributable to differences in slope steepness. How then can a universal threshold be set for all analytical chemistry methods in validation guidelines? Two otherwise identical analytical methods would pass and fail such a criterion purely based on their slopes.

## If data are heteroscedastic, $R^2$ varies with the standards spacing unless a weighted version of the equation is used

Calibration datasets are frequently heteroscedastic, meaning the absolute precision of the measurements (standard deviation) changes across the concentration values (19, 21). If this characteristic of the data is not properly taken into account in calculating the coefficients of correlation and determination, further distortion of the $R^2$ value can occur in relation to the experimental setup (standards spacing within the dynamic range). In Figure 5, two heteroscedastic calibration datasets with seven concentration levels are presented. In Figure 5(A), standards are equidistant, whereas in Figure 5(B), more standards are present at the low end of the calibration range—where the precision is greater for heteroscedastic data and arguably where most cases in a therapeutic range would fall. The coefficient of determination calculated is higher in the second case (uneven placement of standards) than in the first case (equidistant standards). In fact, running a simulation for

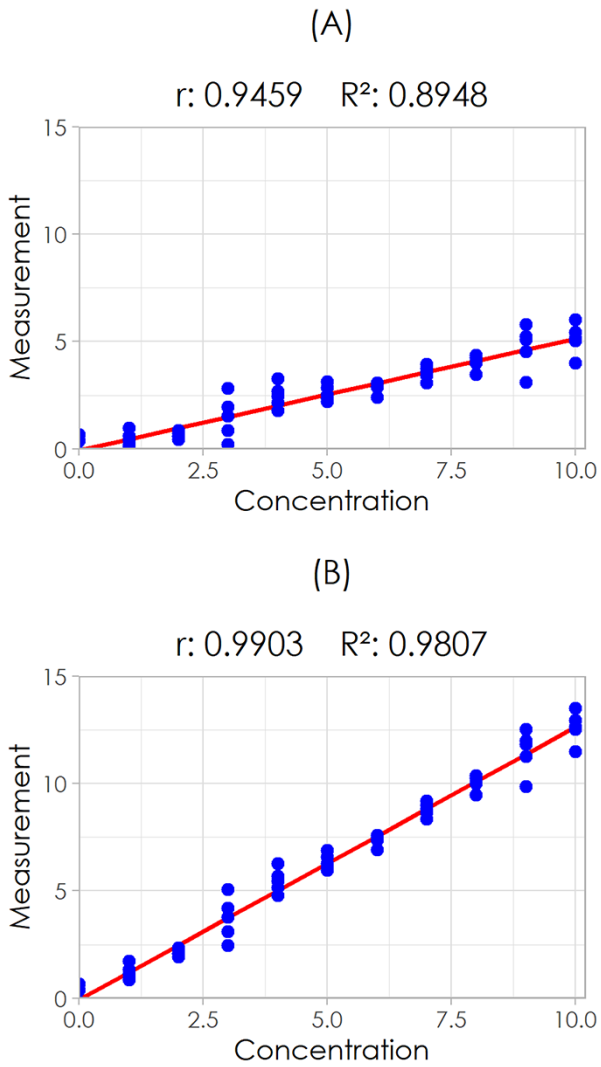


**Figure 4.** Two calibration datasets differing only by their slope, (A) $y = 0.5x$ and (B) $y = 1.25x$, will yield different coefficients of determination. The steeper the slope, the higher the $R^2$ will be.

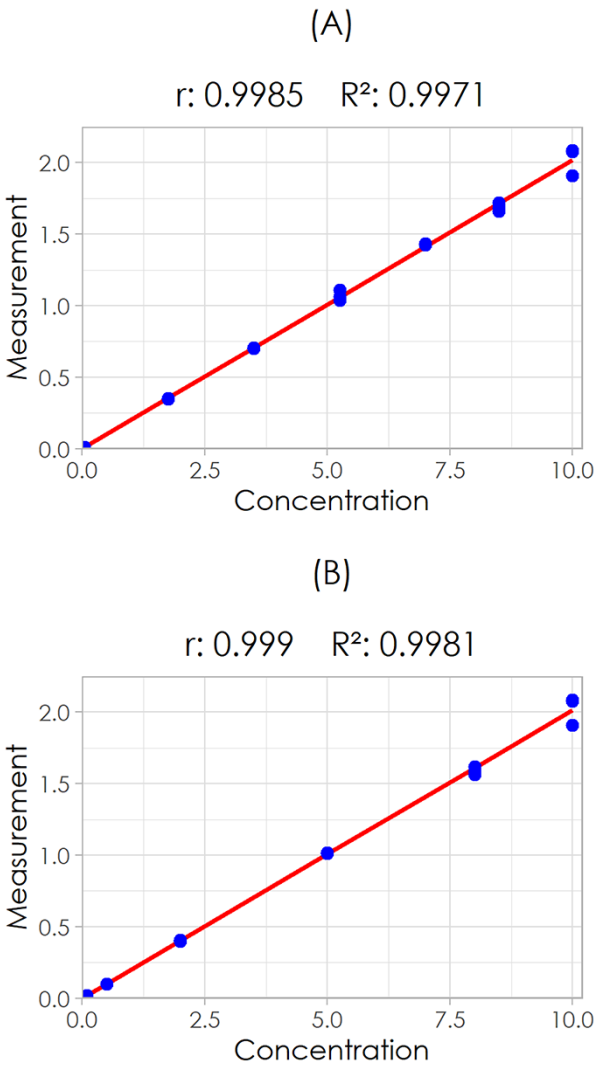


**Figure 5.** Two calibration datasets generated by the same model, but with (A) equidistant and (B) uneven calibration levels distribution. Using standard formulas, the $R^2$ value is higher in the second case, where more standards are placed in the lower variance region (low end of the calibration curve).

data generated using identical models with different standard placements shows the hypothesis that the $R^2$ values are identical can be rejected ($P = 0.0032$) (22). Details of this simulation can be found in Supplementary Data 1. If the previously stated definitions for $r$ Equation (1) and $R^2$ Equation (2) are used on heteroscedastic data, then calibration levels spacing within the dynamic range has an impact on the coefficients of correlation and determination values. A calibration curve with more points in the lower variance region (typically the low end of the curve) will result in a higher $R^2$. A calibration curve with more points in the higher variance region (typically the high end of the curve) will result in a lower $R^2$.

The problem here does not lie with the fact that a setup other than equidistant standards spacing is used—such setups are perfectly valid. Rather, the intersection of heteroscedastic data with a non-weighted formula generates this issue. The weighted versions of the calculations are (6):

$$r_w = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} w_i (x_i - \bar{x})^2 \sum_{i=1}^{n} w_i (y_i - \bar{y})^2}} \quad (3)$$

$$R^2_w = 1 - \frac{SSE}{S_{YY}} = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \quad (4)$$

where $w_i$ is the weight associated with calibration level $i$—typical weights for calibration curves in analytical chemistry include $w_i = 1/x_i$ and $w_i = 1/x_i^2$. If these weighted versions are used, then the simulation of data generated using identical models with different standard placement cannot reject the hypothesis that the $R^2$ values are identical ($P = 0.2819$).

Should one insist on calculating a coefficient of correlation or determination for a heteroscedastic calibration dataset, weighted versions of the calculations should be used to avoid the issue of standards repartition influence. Unfortunately, whereas heteroscedastic datasets are more the norm than the exception in analytical chemistry, the use of weighted $R^2$ is not systematically available (nor clearly marked) in commonly used software; some are even relying exclusively on the regular, unweighted version of the calculation.

## Conclusions

The $r$ and $R^2$ evaluate whether the data are better explained by the regression model used than by no model at all, and to what degree. They describe the strength of the association between the $x$ and $y$ variables under the regression model used.

As we have seen, heteroscedastic calibration data, which is typical in analytical chemistry, will generate $R^2$ values that vary with standards spacing, unless a weighted version of the equation is used. However, even using the proper, weighted version of the equations for these datasets, throes of other problems remain with the coefficients of correlation and determination. All other things being equal, $R^2$ can be influenced by the precision of the data and the steepness of the calibration slope. Furthermore, a high value, greater than 0.9, does not constitute a proof that the data stem from a linear association between the variables. This begs the question of usefulness in the context of calibration curves as part of analytical methods. How can we even determine an acceptability threshold when the $R^2$ result depends this much on experimental parameters?

Unfortunately, the answer is that neither $r$ nor $R^2$ are useful to satisfy method validation guidelines' requirement to demonstrate the linearity of the data or confirm the adequacy of the calibration model used (2, 4–6)—which does not preclude their use in other contexts. We suggest to stop using $r$ and $R^2$ as figures of merit for this particular purpose, especially in formulations such as 'data linearity was demonstrated, with $R^2 = 0.9$' or 'data were linear ($R^2 = 0.9$)'. These formulations imply that the $R^2$ value can prove linearity of the calibration data. Rightly so, a number of method validation guidelines simply do not mention their use in such circumstances (23–25)—and even explicitly discourage it (4, 5).

Which statistics or methods should then be used for the intended purpose of testing linearity of the data or confirming regression model adequacy? While a full review of such methods is outside the scope of this technical note, several options are detailed in a previous publication by the authors (20). The simplest (yet subjective) solution is the residual plot, as suggested by several authors and guidelines (2–4, 6, 26), which should display a random residuals pattern. Statistical testing can also be used. ANOVA-Lack-of-Fit is a popular option (2, 3, 5), although it is known to be sensitive to experimental design (20). Another option is Mandel's fitting test, which compares the sum of squares of the residuals to the mean squares of residuals (27). A close cousin is the partial $F$-test, comparing the sum of squares of the regression to the mean squares of the residuals (4, 28). The preference of the authors lay with the latest option (19, 20), but it is by no means the only available path. In any event, linearity of the data should always be evaluated after the homoscedasticity of the data has been assessed, with weight adjusted formulas if the dataset has been found to be heteroscedastic.

## Supplementary data

Supplementary data is available at *Journal of Analytical Toxicology* online.

## Data availability

All data are incorporated into the article and its online supplementary material.

## References


1. Miller, J.N., Miller, J.C. Calibration methods in instrumental analysis: regression and correlation. In: *Statistics and Chemometrics for Analytical Chemistry*. 6th edition. Pearson/Prentice Hall: Essex, England, 2010; pp 110–153.
2. Van Loco, J., Elskens, M., Croux, C., Beernaert, H. (2002) Linearity of calibration curves: use and misuse of the correlation coefficient. *Accreditation and Quality Assurance*, **7**, 281–285.
3. AAFS Standards Board. *Standard Practices for Method Validation in Forensic Toxicology*. AAFS Standards Board: Colorado Springs, CO, 2019. http://www.asbstandardsboard.org/wp-content/uploads/2019/11/036_Std_e1.pdf (Accessed Jul 22, 2020).
4. Wille, S.M.R., Coucke, W., De Baere, T., Peters, F.T. (2017) Update of standard practices for new method validation in forensic toxicology. *Current Pharmaceutical Design*, **23**, 5442–5454.
5. Peters, F.T., Drummer, O.H., Musshoff, F. (2007) Validation of new methods. *Forensic Science International*, **165**, 216–224.
6. Asuero, A.G., Sayago, A., González, A.G. (2006) The correlation coefficient: an overview. *Critical Reviews in Analytical Chemistry*, **36**, 41–59.
7. Barrett, J.P. (1974) The coefficient of determination—some limitations. *The American Statistician*, **28**, 19–20.
8. Deeb, S., Wylie, F.M., Torrance, H.J., Scott, K.S. (2020) An insight into gabapentin and pregabalin in Scottish prisoners. *Journal of Analytical Toxicology*, **44**, 504–513.
9. Cao, J.-J., Yang, K., Huang, C.-Y., Li, Y.-J., Yu, H., Wu, Y. et al. (2020) Pharmacokinetic study of multiple components of Gelsemium elegans in goats by ultra-performance liquid chromatography coupled to tandem mass spectrometry. *Journal of Analytical Toxicology*, **44**, 378–390.
10. Hubbard, J.A., Navarrete, A.L., Fitzgerald, R.L., McIntyre, I.M. (2021) Acidic drug concentrations in postmortem vitreous humor and peripheral blood. *Journal of Analytical Toxicology*, **45**, 69–75.
11. Velasco-Bejarano, B., Bautista, J., Rodríguez, M.E., López-Arellano, R., Arreguín-Espinosa, R., Carrillo, R.V. (2020) Quantification and stereochemical composition of R-(–) and S-(+)-clenbuterol enantiomers in bovine urine by liquid chromatography–tandem mass spectrometry. *Journal of Analytical Toxicology*, **44**, 237–244.
12. Nanco, C.R., Poklis, J.L., Hiler, M.M., Breland, A.B., Eissenberg, T., Wolf, C.E. (2019) An ultra-high-pressure liquid chromatographic tandem mass spectrometry method for the analysis of benzoyl ester derivatized glycols and glycerol. *Journal of Analytical Toxicology*, **43**, 720–725.
13. Kriikku, P., Pelander, A., Rasanen, I., Ojanperä, I. (2019) Toxic lifespan of the synthetic opioid U-47,700 in Finland verified by reanalysis of UPLC-TOF-MS data. *Forensic Science International*, **300**, 85–88.
14. Ramírez Fernández, M.D.M., Wille, S.M.R., Di Fazio, V., Samyn, N. (2019) Influence of bleaching and thermal straightening on endogenous GHB concentrations in hair: an in vitro experiment. *Forensic Science International*, **297**, 277–283.
15. Misailidi, N., Athanaselis, S., Nikolaou, P., Katselou, M., Dotsikas, Y., Spiliopoulou, C. et al. (2019) A GC–MS method for the determination of furanylfentanyl and ocfentanil in whole blood with full validation. *Forensic Toxicology*, **37**, 238–244.

16. Lowry, J., Truver, M.T., Swortwood, M.J. (2019) Quantification of seven novel synthetic opioids in blood using LC–MS/MS. *Forensic Toxicology*, **37**, 215–223.
17. Lelievre, B., Triau, S., Codron, P., Mariau, Y., Papin-Lefebvre, F., Collin, A. et al. (2020) A chasing dead-end case report: a fatal lead intoxication following an attempted homicide. *Forensic Toxicology*, **38**, 505–510.
18. Papoutsis, I., Mendonis, M., Nikolaou, P., Athanaselis, S., Pistos, C., Maravelias, C. et al. (2012) Development and validation of a simple GC–MS method for the simultaneous determination of 11 anticholinesterase pesticides in blood—clinical and forensic toxicology applications. *Journal of Forensic Sciences*, **57**, 806–812.
19. Desharnais, B., Camirand-Lemyre, F., Mireault, P., Skinner, C.D. (2017) Procedure for the selection and validation of a calibration model I—description and application. *Journal of Analytical Toxicology*, **41**, 261–268.
20. Desharnais, B., Camirand-Lemyre, F., Mireault, P., Skinner, C.D. (2017) Procedure for the selection and validation of a calibration model II—theoretical basis. *Journal of Analytical Toxicology*, **41**, 269–276.
21. Gu, H., Liu, G., Wang, J., Aubry, A.-F., Arnold, M.E. (2014) Selecting the correct weighting factors for linear and quadratic calibration curves with least-squares regression algorithm in bioanalytical LC-MS/MS assays and impacts of using incorrect weighting factors on curve stability, data quality, and assay performance. *Analytical Chemistry*, **86**, 8959–8966.
22. Soper, D.S. *A-Priori Sample Size Calculator for Student t-Tests*. *Free Statistics Calculators (Version 4.0)*. https://www.danielsoper.com/statcalc/calculator.aspx?id=47 (Accessed Jul 30, 2020).
23. Gesellschaft für Toxikologische und Forensische Chemie. *Guidelines for Quality Assurance in Forensic-Toxicological Analyses - Appendix B: Requirements for the Validation of Analytical Methods*. Gesellschaft für Toxikologische und Forensische Chemie: Germany, 2009. https://www.gtfch.org/cms/images/stories/files/Appendix%20B%20GTFCh%2020090601.pdf (Accessed Jul 22, 2020).
24. European Medicines Agency. *Guideline on Bioanalytical Method Validation*. European Medicines Agency: London, 2011. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-bioanalytical-method-validation_en.pdf (Accessed Jul 22, 2020).
25. Food and Drug Administration. *Bioanalytical Method Validation - Guidance for Industry*. Food and Drug Administration: Silver Spring, MD, 2018. https://www.fda.gov/files/drugs/published/Bioanalytical-Method-Validation-Guidance-for-Industry.pdf (Accessed Jul 22, 2020).
26. Wille, S.M.R., Peters, F.T., Di Fazio, V., Samyn, N. (2011) Practical aspects concerning validation and quality control for forensic and clinical bioanalytical quantitative methods. *Accreditation and Quality Assurance*, **16**, 279.
27. Mandel, J. Testing the statistical model. In: *The Statistical Analysis of Experimental Data*. Dover Publications: New York, NY, 1964; pp 160–193.
28. Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J. Multiple and polynomial regression. In: *Handbook of Chemometrics and Qualimetrics: Part A*. Vol. 20A. Elsevier: Amsterdam, 1997; pp 263–303.