Percent residual accuracy for quantifying goodness-of-fit of linear calibration curves

Brian A. Logue*, Erica Manandhar

Department of Chemistry and Biochemistry, South Dakota State University, Avera Health and Science, Box 2202, Brookings, SD 57007, USA

ABSTRACT

Linear models for calibration curves are overwhelmingly created based on minimization of least squares error, with their goodness-of-fit (GOF) quantified using the square of the correlation coefficient (R^2). Yet, R^2 has well-known disadvantages when used to quantify GOF of calibration curves stemming from its calculation based on the absolute error of the signal (i.e., calculated vs. experimental). These disadvantages are exacerbated when using a geometric series of concentrations for calibration standards (e.g., 1, 2, 5, 10, etc.) and when calibration curves span 2–3 orders of magnitude, which is typical for modern analytical techniques. While there are multiple alternative GOF measures, R^2 overwhelmingly persists in the field of Analytical Chemistry as the most reported measure of GOF. We evaluated R^2, alternative GOF measures, and multiple quantitative bias parameters, along with residual analysis, for over 60 experimental calibration curves. R^2 did a poor job of consistently and accurately quantifying the GOF over the entire calibration curve. This was especially true for situations where the low concentration calibrators were not accurately described by the calibration equation. While other GOF parameters, including the sum of the absolute percent error, mean absolute percent error, and quality coefficient, did a better job of describing GOF of calibration curves, each had significant theoretical and/or practical disadvantages. Therefore, we introduce a descriptive GOF parameter called Percent Residual Accuracy (%RA or PRA) which equally weights the accuracy of all calibrators into a single value, generally falling between 0% and 100%, with 100% representing a perfect fit and a "good" fit for calibration data producing a %RA of 90–100%. The %RA much more effectively described the GOF for the entire calibration range than R^2, and it similarly quantified GOF as compared to the other GOF parameters tested. With the performance and practical advantages of %RA, we conclude that it is the most advantageous GOF parameter and that it should be reported as a standard GOF measure for calibration curves.

1. Introduction

At the heart of quantitative Analytical Chemistry is the ability to accurately quantify the concentration of an analyte from a matrix of interest. With few exceptions, the amount/concentration of an analyte cannot be measured directly, especially for low analyte concentrations. Therefore, calibration is essential to converting the signals produced from modern analytical instrumentation into concentration [1,2]. While calibration behavior can be modeled in many ways, including polynomial [3,4], power/logarithmic functions [5] and even inverse linear [6–8], most analytical methods are calibrated using a linear equation [9–14] relating concentration to signal. The two most common methods to model a linear calibration curve are ordinary least squares (OLS) and weighted least squares (WLS). For linear calibration curves, OLS minimizes the sum of squared errors between predicted and

experimental signal data to produce a linear mathematical model (i.e., $y = mx + b$) which best fits the experimental data. WLS weights the OLS technique by a user-defined weighting parameter, typically $1/x$, $1/x^2$, $1/y$, $1/y^2$, $1/s$, or $1/s^2$. Multiple studies have been conducted to show the value of WLS over OLS for calibration of modern analytical methods [2,15–21]. For example, using advanced Monte-Carlo simulations, Tellinghusien [22] showed that calibration curves are predisposed to favor $1/y^2$ or $1/x^2$ weighting, even if the calibration data are homoscedastic. Meyer [23] extended this analysis to show that a weighted fit is theoretically necessary if the absolute standard deviation for the calibrators is heteroscedastic (i.e., the variability of the signal is unequal across the calibration range, which is overwhelmingly the case for calibrating modern analytical techniques), but only practically necessary if: 1) the signal standard deviation is not constant, 2) the calibration range is "large", 3) the calibration standards are equally

distributed, and 4) the analytical result is at the lower end of the calibration range. Requirements 1 and 2 are met for most analytical techniques and, because the concentration of most samples is unknown, Requirement 4 must be assumed to be met in most situations. Requirement 3 can be interpreted as equally distributed both arithmetically and geometrically, which is most often the case for modern analytical techniques. Therefore, the practical requirements for using WLS over OLS are certainly met for most analytical techniques. Gu et al. [17] even recommended that $1/x^2$ weighting should be used for all LCMSMS assays.

The differences between OLS and WLS for calibration curves have been extensively evaluated. Most of these studies conclude that WLS is necessary and $1/x^2$ or $1/y^2$ are the most appropriate weighting factors. Conversely, the advantages and disadvantages of available goodness-of-fit (GOF) parameters for calibration curves has received limited attention. Although the best practice is to use multiple parameters, GOF is typically reported as a single parameter selected from available GOF measures that represents how closely a model can predict the true value of interest. Overwhelmingly, the GOF of calibration curves is represented by the squared correlation coefficient (R^2). Although R^2 has advantages in determining correlations, such as an intuitive scale between 0 and 1 and a close relationship with OLS regression, Meier and Zund [24] concluded that R^2 is not overly useful in Analytical Chemistry, especially as a GOF measure for calibration curves. More strongly, the IUPAC Guidelines for calibration in Analytical Chemistry [25] stated that R^2 has no meaning in calibration because x-values (i.e., calibrator concentrations) are not random quantities, while R^2 is a measure of the relationship between two random variables. This is because the nominal concentration of the calibrators should be very close to their true concentration, and, within the linear dynamic range (LDR), the relationship between the concentration (x) and the signal (y) is not random (i.e., it is linearly correlated, producing an R^2 very close to 1).

The major disadvantage of application of R^2 to calibration is that its values are very close to 1 for almost all linear fits of concentration within the LDR, even if there is a significant departure from linearity for multiple calibration standards [16,25–38]. While these problems are present for all calibrations, they are more readily apparent when the LDR is large and the calibration data is heteroscedastic. This leads to standard deviations for higher concentrations having a much greater influence on the R^2. Therefore, if the fit is inaccurate for lower concentration calibrators, the R^2 does not reflect the inadequate fit. Because the LDR for modern analytical techniques is typically two or more orders of magnitude, a geometric series of calibration standards is used, and the signal standard deviation scales with increasing signal, R^2 does a poor job of describing the GOF for most calibration curves of modern analytical techniques.

While multiple other GOF parameters are available, they often produce conflicting results, and there is no consensus about what parameter to use to constitute a "good fit" [39–41]. Moreover, despite the disadvantages of R^2 and the availability of other GOF parameters, examination of peer-reviewed analytical method development literature confirms that no other parameter has come close to supplanting its use in quantifying GOF for calibration curves. For example, of the articles published in Journal of Chromatography A in the month of January 2018 which include linear calibration curves, only one (1) of 18 showed residuals in terms of percent concentration accuracy and one (1) other mentioned residual analysis, while reporting R^2. All other works exclusively reported R^2 or R, with multiple papers including statements that a particular R^2 value (i.e., usually > 0.99) indicated "good" linearity.

Considering the disadvantages of R^2, the objective of this study was to evaluate other GOF parameters, also comparing to R^2, to determine the most appropriate GOF measure for linear calibration curves. Percent Residual Accuracy (PRA or %RA) is introduced, evaluated, and recommended as a simple descriptive measure of GOF for modern calibration curves.

2. Theory

2.1. Correlation coefficient

The correlation coefficient is a well-known quantity which measures the correlation between two variables, and is symbolized by R [42]. R can have values of − 1 to + 1. For calibration, the correlation coefficient is typically squared (R^2), producing values between 0 and + 1. R^2 is calculated in several ways [1,2,43], but the calculation of R^2 which allows direct comparison between multiple methods of fitting is defined by Eq. 1, where the ratio of the sum-of-squares error (SS_{err}) to the total sum-of-squares (SS_{tot}) is subtracted from 1. SS_{err} is the sum of squares of the absolute difference between the predicted ($y_{i,pred}$) and experimental ($y_{i,exp}$) value of the dependent variable for each data point i. SS_{tot} is the sum of squares of the difference between the experimental value and the average experimental values (y_{avg}). N is the number of calibrators used in the calibration curve.

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{N}(y_{i,exp} - y_{i,pred})^2}{\sum_{i=1}^{N}(y_{i,exp} - y_{avg})^2} \tag{1}$$

As evident from Eq. (1), the R^2 is based on absolute error of the dependent variable (i.e., signal in the case of calibration curves), with 1 being a "perfect" fit based on each experimental and predicted y-value being equal, and 0 indicating no correlation between the model and the data.

2.2. Mean Absolute Percent Error (MAPE)

A GOF measure based on absolute error of the signal, as with R^2, has inherent disadvantages for calibration. Mean Absolute Percent Error (MAPE) is a measure of GOF based on the relative error between the experimental and predicted dependent variable [13, 14]. MAPE is calculated by Eq. (2).

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_{i,exp} - y_{i,pred}}{y_{i,exp}}\right|, \tag{2}$$

MAPE is the average absolute percent error of the predicted dependent variable as compared to the experimental value averaged over all data points used to create the model. MAPE is used more for assessing GOF for other models outside of calibration curves, but is evaluated here for potential usefulness as a GOF parameter for linear calibration curves.

Obviously, the value of MAPE decreases as the GOF becomes better, with a "perfect" fit producing a value of zero (0). In the unlikely event that the predicted signal is extremely inaccurate, two situations arise, where either: 1) $y_{i,pred} < < y_{i,exp}$, or 2) $y_{i,pred} > > y_{i,exp}$. In the first case, the absolute value term becomes 1, and MAPE becomes 100% for that calibrator. In the second situation, the quotient in the absolute value term becomes $y_{i,pred}/y_{i,exp}$. Theoretically, this could result in a value of ∞. Since lower MAPE represent better GOF, both cases would indicate extremely poor fits.

2.3. Sum of percent relative errors ($\sum RE\%$)

The most important function of the calibration curve is to accurately predict the concentration of unknown samples, and because the relative error in the signal (y) is very likely larger than the error in the concentration of the calibration standards (x) for modern analytical methods, GOF parameters based on difference between the nominal and calculated concentrations are more desirable. $\sum RE\%$ is a parameter which has been used for quantification of calibration curve GOF and is based on Eq. [3].