Int. J. Metrol. Qual. Eng. **8**, 28 (2017)

DOI: 10.1051/ijmqe/2017021

International Journal of Metrology and Quality Engineering

Available online at:
www.metrology-journal.org

RESEARCH ARTICLE OPEN ACCESS

Reversed inverse regression for the univariate linear calibration and its statistical properties derived using a new methodology

Pilsang Kang*, Changhoi Koo, and Hokyu Roh

Quality Management Center, KEPCO NF, 242, Daedeok-daero 989 beon-gil, Yuseong-gu, Daejeon, 34057, Korea

Received: 12 March 2017 / Accepted: 28 September 2017

Abstract. Since simple linear regression theory was established at the beginning of the 1900s, it has been used in a variety of fields. Unfortunately, it cannot be used directly for calibration. In practical calibrations, the observed measurements (the inputs) are subject to errors, and hence they vary, thus violating the assumption that the inputs are fixed. Therefore, in the case of calibration, the regression line fitted using the method of least squares is not consistent with the statistical properties of simple linear regression as already established based on this assumption. To resolve this problem, "classical regression" and "inverse regression" have been proposed. However, they do not completely resolve the problem. As a fundamental solution, we introduce "reversed inverse regression" along with a new methodology for deriving its statistical properties. In this study, the statistical properties of this regression are derived using the "error propagation rule" and the "method of simultaneous error equations" and are compared with those of the existing regression approaches. The accuracy of the statistical properties thus derived is investigated in a simulation study. We conclude that the newly proposed regression and methodology constitute the complete regression approach for univariate linear calibrations.

Keywords: bias / classical regression / error propagation / mean-data-point-based variance / population-regression-line-based variance / reversed inverse regression / simultaneous error equations / Taylor approximation

1 Introduction

Simple linear regression is a model with a single independent variable in which a regression line is fitted through n data points such that the sum of squared errors (SSE), i.e., the vertical distances between the data points and the fitted line, is as small as possible. The statistical properties of this model have been established as theorems and are presented in many statistics textbooks, e.g., the textbook written by Walpole and Myers [1]. In this model, a regression line of y on x is fitted based on the assumption that x is fixed but y varies according to a normal distribution. This model is called "basic regression" throughout the remainder of this study. Unfortunately, when calibrating an instrument such as a chemical analyzer using basic regression, a problem arises. In practical calibrations, the observed measurements (the x values) are subject to errors, and hence they vary, thus violating the assumption of fixed inputs. As a result, in the case of calibration, the regression line fitted using the method of least squares is not consistent with the statistical properties of basic regression as already established based on this assumption.

Two approaches have been considered as possible solutions for this problem. In the first approach [2], called classical regression, the "standards" (the x values) are treated as the inputs, and the observed measurements (the y values) are treated as the response; these values are used to fit a regression line of y on x. This regression approach is consistent with the assumption that x is fixed. The problem with this approach is that estimating the x value for a new observed measurement involves the reciprocal of the estimated slope. Williams [3] demonstrated that the reciprocal of the slope has an infinite variance, which indicates that classical regression has an infinite variance and, hence, an infinite mean squared error. Nevertheless, Parker et al. [4] obtained an asymptotic approximation of the variance of the prediction interval using a formula derived by Casella and Berger [5] using the Delta Method. However, Parker et al.'s approach still has limitations. Even if we rely on this approximation, we cannot determine a prediction interval with a given confidence level because the approximation cannot be used to express the prediction interval as a t_{n-2} distribution.

* Corresponding author: pskang@knfc.co.kr

In the second approach [6], called inverse regression, the standards (the x values) are treated as the response, the observed measurements (the y values) are treated as the inputs, and these values are used to fit a regression line of x on y. This regression approach is inconsistent with the assumption that the inputs are fixed. Shukla and Datta [7] and Oman [8] derived expressions for the mean and mean squared error of predicted x value based on multiple measurements taken during the prediction stage of the calibration process. Fuller [9] made a similar suggestion regarding the derivation of both the predicted x value and the prediction interval. Fuller's approach requires that the variance of the observed measurements is known. In his approach, it is necessary to measure a standard multiple times independently to estimate the variance. Parker et al. [4] derived the bias in prediction using a formula established by Pham-Gia et al. [10] with the aid of the Delta Method. Parker et al. [4] also showed through several simulation studies that inverse regression is preferable to classical regression in terms of bias and mean squared error. However, to derive the statistical properties of inverse regression, Parker et al. were obliged to borrow their estimate for the variance of the slope from "reversed basic regression" because of technical difficulties, which devalues their approach. (Reversed basic regression is basic regression in which the roles of x and y have merely been reversed.)

As a fundamental solution for the calibration problem, which has not yet been resolved completely, the current study introduces "reversed inverse regression" along with a new methodology for deriving its statistical properties. (Simply put, "fundamental solution for the univariate linear calibration problem"="reversed inverse regression"+"new methodology for deriving the statistical properties of the regression".) In the proposed regression approach, the observed measurements (the x values) are treated as the inputs, and the standards (the y values) are treated as the response; these values are used to fit a regression line of y on x. The statistical properties of this regression are derived using the "error propagation rule" and the "method of simultaneous error equations". In this regression approach, it is not necessary to measure any standards multiple times independently. We present an example of practical calibration. Each of three types of regression (i.e., classical regression, inverse regression and reversed inverse regression) is applied to the calibration example, and the corresponding calibration results, including the subsequently calculated estimates for the variance of the prediction interval, are compared. In addition, the accuracy of the statistical properties derived using the new methodology is investigated in a Monte Carlo simulation study.

2 Regression and methodology

If the roles of x and y are reversed, then inverse regression becomes reversed inverse regression. Reversed inverse regression is more convenient to use for calibration than inverse regression because the reversed roles are consistent with the convention that the variable x represents the inputs, whereas the variable y represents the response. This regression approach also violates the assumption that the inputs are fixed. It is modeled as follows. (It may be desirable to use some other term than "reversed inverse regression", e.g., "pseudo-basic regression", to eliminate potential confusion in terminology.)

- There is a linear relationship between x and y.
- The observed measurements (the x values) are treated as the inputs, the standards (the y values) are treated as the response, and these values are used to fit a regression line of y on x.
- For the fitting of the regression line, n data points of the form $(x_i,\ y_i)$ $(i=1,\ ...,\ n)$ are used. The x_i value varies according to a normal distribution, whereas the y_i value is fixed; $y_i=\alpha+\beta x_i+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma^2)$.
- The x_i's (i.e., $x_1,\ ...,\ x_n$) are treated as variables. The variables x_i and x_j $(i\neq j)$ are independent of each other: $\mathrm{cov}[x_i,\ x_j]=0,\ i\neq j$.
- The regression line $\hat{y}=\hat{\alpha}+\hat{\beta}x$ is fitted such that SSE is minimized.
  - $SSE=\sum(y_i-\hat{\alpha}-\hat{\beta}x_i)^2$, $\hat{\beta}=S_{xy}/S_{xx}$, $S_{xx}=\sum(x_i-\overline{x})^2$, $S_{xy}=\sum(x_i-\overline{x})(y_i-\overline{y})$
- The variance of x_i is uniform for all i $(i=1,\ ...,\ n)$. In other words, the variance of the observed measurements is equal over the entire calibration range of interest.
  - σ^2_{xi} denotes the variance of the variable x_i; $\sigma^2_{x1}=\cdots=\sigma^2_{xn}(=\sigma^2_x)$.
- The population regression line $y=\alpha+\beta x$ is defined as follows:
  - $\beta=\sum(x_{i0}-\overline{x}_0)(y_i-\overline{y})/\sum(x_{i0}-\overline{x}_0)^2$, $\alpha=\overline{y}-\beta\overline{x}_0$, and $\sigma^2_x\beta^2=\sigma^2$.
  - $\overline{x}=(\sum x_i)/n$, $\overline{y}=(\sum y_i)/n$, x_{i0} is the mean of x_i, and $\overline{x}_0=(\sum x_{i0})/n$.
  - All points $(x_{i0},\ y_i)$ $(i=1,\ ...,\ n)$ lie on the population regression line. In this study, we call these points the "mean data points".

($\sum$ denotes summation from $i=1$ to n throughout this study.)

In reversed inverse regression, the assumption that the observed measurements (the x values), despite being the inputs, vary according to normal distributions is very important. Suppose that the regression line fitting is repeated an infinite number of times using a "new set of n different standards (or reference solutions)" each time. Here, this "new set of n different standards" refers to newly prepared standards whose nominal y values (or target y values) and confidence levels are identical to those of the previous set of standards. In this case, the x_i's (i.e., $x_1,\ ...,\ x_n$) will be observed to vary according to normal distributions. The standards are subject to errors that may arise when preparing or manufacturing them. However, such errors will appear as variations in the x_i's after being combined with random measurement errors. If the "same set of n different standards" is measured repeatedly, we will only observe the variance associated with the random measurement errors; the errors of the standards themselves will not be reflected. Such a variance should not be treated as the variance needed to derive the statistical properties of linear regression. In this respect, Fuller [9] is incorrect, because his approach requires a standard to be independently measured multiple times to

estimate the variance. As previously mentioned, reversed inverse regression does not require any such separate prior measurements.

The slope of the regression line that is fitted on the basis of reversed inverse regression is:

$$\hat{\beta} = S_{xy}/S_{xx} = \sum(x_i - \overline{x})(y_i - \overline{y})/\sum(x_i - \overline{x})^2.$$

Unfortunately, it is technically difficult to derive the variance of the slope directly from the definition of the variance, i.e., $\mathrm{var}[f(x_1, ..., x_n)] = \mathrm{E}[\{f(x_1, ..., x_n) - \mathrm{E}[f(x_1, ..., x_n)]\}^2]$, because $\hat{\beta}$ is a fractional expression that contains "$\sum(x_i - \overline{x})^2$" in the denominator and the x_i's vary rather than being fixed. Because of this difficulty, we directly treat the x_i's as variables and derive the variance of the slope based on the first-order Taylor approximation as follows:

$$\begin{aligned} f(x_1, ..., x_n) &= f(x_{10}, ..., x_{n0}) + \sum(x_i - x_{i0})[\partial f/\partial x_i]^* \\ &\quad + \text{Remainder}, \\ var[f(x_1, ..., x_n)] &= \mathrm{E}[\{f(x_1, ..., x_n) - \mathrm{E}[f(x_1, ..., x_n)]\}^2] \\ &\approx \mathrm{E}[\{f(x_1, ..., x_n) - f(x_{10}, ..., x_{n0})\}^2] \\ &\approx \{\sum(\partial f/\partial x_i)^2 {\sigma_{xi}}^2 \\ &\quad + 2\sum\sum(\partial f/\partial x_i)(\partial f/\partial x_j)cov[x_i, x_j]\}^*. \\ \text{Note}: \mathrm{E}[\{f(x_1, ..., x_n) &- f(x_{10}, ..., x_{n0})\}^2] \\ &= E[\{f(x_1, ..., x_n) - E[f(x_1, ..., x_n)]\}^2] \\ &\quad + \{\text{bias in} f(x_1, ..., x_n)\}^2, \end{aligned}$$

where the notation $[\ ]^*$ or $\{\ \}^*$ indicates that the value of the function contained within the bracket is determined using the mean values of the variables, i.e., $x_{10}, ..., x_{n0}$ [11]. Even in the case of derivation of expectations, this notation is often used for the same purpose. In particular, we define the expectation $\mathrm{E}[\{f(x_1, ..., x_n) - f(x_{10}, ..., x_{n0})\}^2]$ as the "mean-data-point-based variance". The approximation method for deriving the variance described herein is commonly referred to as the "error propagation rule", and only the first-order partial derivatives are included in its derivation. To derive the variance of the slope, $\mathrm{var}[\hat{\beta}]$, after the partial differentiation of $\hat{\beta}$ with respect to the x_i's, the variances of the x_i's, including the covariances of x_i and x_j $(j > i)$, are combined in accordance with the error propagation rule. The final result obtained from this combination process is the approximate variance of the slope. The same method can be used to derive the variance of the intercept and the variance of the predicted y value. All other statistical properties of reversed inverse regression, such as the expectation and bias of the slope and the expectation of the mean squared error, are derived by utilizing another special method, called the "method of simultaneous error equations" in this study, in combination with the error propagation rule. When we need to derive another statistical property from the primary expressions already obtained using the error propagation rule, the first-order Taylor approximation is mainly used. Error terms of orders higher than $(\sigma_x/A)^2$ are discarded during or after the approximation calculations. For example, $(\sigma_x/A)^4$ $(=1/10^8)$ is very small and can be neglected in comparison with $(\sigma_x/A)^2$ $(=1/10^4)$.

The Delta Method is also an asymptotic approximation method based on Taylor approximation [12]. Parker et al. [4] used the Delta Method to derive the variance of the prediction interval for classical regression. When the Delta Method is applied to the inverted equation $x = -\hat{\alpha}'/\hat{\beta}' + (1/\hat{\beta}')y$, the x_i's and y_i's are not directly treated as variables. Instead, U $(=-\hat{\alpha}' + y_0 - \varepsilon_0)$ and V $(=\hat{\beta}')$ are treated as the variables [4,5,10]. This is the most notable difference between the Delta Method and the approximation method used in this study.

3 Statistical properties of reversed inverse regression

The variance and bias of the slope and the expectation of the mean squared error are the statistical properties that are primarily required in linear regression because other properties, such as the variance and bias of the intercept and the variance of the prediction interval, depend on them. Therefore, the variance of the slope, $\mathrm{var}[\hat{\beta}]$, is first derived using the error propagation rule as follows (see supplementary material):

$$var[\hat{\beta}] = \{\sigma(\hat{\beta})\}^2 \approx \sum(\partial\hat{\beta}/\partial x_i)^2 {\partial x_i}^2 = [S_{yy}/S_{xy}^2]^*\sigma^2. \quad (1)$$

To investigate the accuracy of the variance obtained using equation (1), we should consider two factors. One is that error terms of orders higher than σ_x^2 are not included in the derivation. The other is that because equation (1) represents the population-regression-line-based variance, the bias in $\hat{\beta}$ is not reflected in the calculation of $[S_{yy}/S_{xy}^2]^*\sigma^2$ $(=[S_{yy}/S_{xy}^2]^*\sigma_x^2\beta^2)$. The bias in $\hat{\beta}$ depends on σ_x^2 and n. The details of the effects of these two factors are explained based on the simulation results in Section 5. For reference, the variance of $\hat{\beta}$ for basic regression is $[1/S_{xx}]^*\sigma^2$, and this variance is not an approximation but an exact expression. The relationship between the estimates of $\mathrm{var}[\hat{\beta}]_{\text{reversed inverse}}$ and $\mathrm{var}[\hat{\beta}]_{\text{basic}}$ for a given set of data points is as follows:

$$\begin{aligned} &\text{Estimate for } \mathrm{var}[\hat{\beta}]_{\text{revrsed inverse}} \\ &\quad = \{1/r^2(x, y)\}(\text{Estimate for } \mathrm{var}[\hat{\beta}]_{\text{basic}}), \end{aligned}$$

where $r(x, y)$ is the estimated correlation coefficient between x and y, i.e., $r(x, y) = S_{xy}/(S_{xx}S_{yy})^{1/2}$, and $r^2(x, y)$ is typically very close to 1 in linear calibrations.

The variance of the intercept, $\mathrm{var}[\hat{\alpha}]$, is also derived using the error propagation rule as follows (see supplementary material):

$$\begin{aligned} \hat{\alpha} &= (\sum y_i)/n - \{\sum(x_i - \overline{x})(y_i - \overline{y})/\sum(x_i - \overline{x})^2\} \\ &\quad \times (\sum x_i)/n, \\ var[\hat{\alpha}] &= \{\sigma(\hat{\alpha})\}^2 \approx \sum(\partial\hat{\alpha}/\partial x_i)^2\sigma_{xi}^2 = [1/n + \overline{x}^2 \\ &\quad \times (S_{yy}/S_{xy}^2)]^*\sigma^2. \end{aligned} \quad (2)$$

Separately from the previous derivation process, another equation for deriving $\mathrm{var}[\hat{\alpha}]$ can be obtained by applying the error propagation rule to $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$:

$$\begin{aligned} var\,[\hat{\alpha}] &= \{\sigma(\hat{\alpha})\}^2 \approx (\partial\hat{\alpha}/\partial\hat{\beta})^2\{\sigma(\hat{\beta})\}^2 + (\partial\hat{\alpha}/\partial\overline{x})^2\{\sigma(\overline{x})\}^2 \\ &\quad + 2(\partial\hat{\alpha}/\partial\hat{\beta})(\partial\hat{\alpha}/\partial\overline{x})\{\sigma(\hat{\beta})\}\{\sigma(\overline{x})\}r(\hat{\beta},\overline{x}) \\ &\approx [1/n + \overline{x}^2(S_{yy}/S_{xy}^2)]\sigma^2 \\ &\quad +2(-x)(-\hat{\beta})\{(S_{yy}/S_{xy}^2)\sigma_x^2\hat{\beta}^2\}^{1/2}(\sigma_x^2/n)^{1/2}r(\hat{\beta},\overline{x}). \end{aligned} \tag{3}$$

From equations (2) and (3), we can see that $r(\hat{\beta}, \overline{x}) \approx 0$, and hence, $\hat{\beta}$ and $\overline{x}$ are nearly independent of each other. In equation (2), var[$\hat{\alpha}$] is derived by treating $\hat{\alpha}$ as a function of x_i's ($i = 1, ..., n$), whereas in equation (3), var[$\hat{\alpha}$] is derived by treating $\hat{\alpha}$ as a function of $\hat{\beta}$ and $\overline{x}$. In this way, by formulating two separate equations to obtain the variance of a statistic using the error propagation rule, we can derive the covariance or correlation coefficient between any two statistics. This method is called the "method of simultaneous error equations" in this study. Nearly all of the covariances (or correlation coefficients) in a linear regression problem can be derived using this method. In addition, the derived covariances can be further used to derive other statistical properties. However, we should note that the covariances thus derived are typically approximations, not exact expressions.

A predicted y value is the y value of a point (x, y) on the fitted regression line and is determined by substituting x into $\hat{y} = \hat{\alpha} + \hat{\beta}x$. The variance of such a predicted y value, var[$\hat{y}$], is derived using the error propagation rule as follows:

$$\begin{aligned} \hat{y} &= \overline{y} - \{\sum(x_i - \overline{x})y_i / \sum(x_i - \overline{x})^2\}\{\sum x_i/n\} \\ &\quad + \{\sum(x_i - \overline{x})y_i / \sum(x_i - \overline{x})^2\}x, \\ var[\hat{y}] &= \{\sigma(\hat{y})\}^2 \approx \sum(\partial\hat{y}/\partial x_i)^2\sigma_{xi}^2 \\ &= [1/n + (x - \overline{x})^2 \times (S_{yy}/S_{xy}^2)]^*\sigma^2. \end{aligned} \tag{4}$$

Separately from equation (4), another equation for deriving var[$\hat{y}$] can be obtained by applying the error propagation rule to $\hat{y} = \hat{\alpha} + \hat{\beta}x$:

$$\begin{aligned} var[\hat{y}] &\approx (\partial\hat{y}/\partial\hat{\alpha})^2\{\sigma(\hat{\alpha})\}^2 + (\partial\hat{y}/\partial\hat{\beta})^2\{\sigma(\hat{\beta})\}^2 \\ &\quad + 2(\partial\hat{y}/\partial\hat{\alpha})(\partial\hat{y}/\partial\hat{\beta})\{\sigma(\hat{\alpha})\}\{\sigma(\hat{\beta})\}r(\hat{\alpha},\hat{\beta}) \\ &\approx [1/n + \overline{x}^2(S_{yy}/S_{xy}^2)]\beta^2\sigma_x^2 + x^2(S_{yy}/S_{xy}^2)\beta^2\sigma_x^2 \\ &\quad + 2x(S_{yy}/S_{xy}^2)^{1/2}\beta\sigma_x[1/n + (\overline{x})^2(S_{yy}/S_{xy}^2)]^{1/2} \\ &\quad \times \hat{\beta}\sigma_x r(\hat{\alpha},\hat{\beta}). \end{aligned} \tag{5}$$

From equations (4) and (5), the correlation coefficient r ($\hat{\alpha}$, $\hat{\beta}$) can be determined as follows:

$$r(\hat{\alpha},\hat{\beta}) \approx -\overline{x}S_{yy}^{1/2}/S_{xy}[1/n + \overline{x}^2(S_{yy}/S_{xy}^2)]^{1/2}.$$

As the next step, we derive the expectations of $\hat{\beta}$ and $\hat{\alpha}$, and the biases in $\hat{\beta}$, $\hat{\alpha}$ and $\hat{y}$. For this purpose, the following statistical properties are derived in advance using the method of simultaneous error equations (see supplementary material):

$$\begin{aligned} \mathrm{E}[\textstyle\sum(x_i - \overline{x})^2] &= \sum(x_{i0} - \overline{x}_0)^2 + (n-1)\sigma_x^2, \\ cov[\textstyle\sum(x_i - \overline{x})^2, 1/\sum(x_i - \overline{x})^2] &\approx -[4/S_{xx}]\sigma_x^2, \\ cov[\textstyle\sum(x_i - \overline{x})y_i, 1/\sum(x_i - \overline{x})^2] &\approx -[2/S_{xy}]\beta^2\sigma_x^2. \end{aligned}$$

$\mathrm{E}[1] = \mathrm{E}[\sum(x_i-\overline{x})^2/\sum(x_i-\overline{x})^2] = \mathrm{E}[\sum(x_i-\overline{x})^2]\cdot\mathrm{E}[1/\sum(x_i-\overline{x})^2] + \mathrm{cov}[\sum(x_i-\overline{x})^2, 1/\sum(x_i-\overline{x})^2]$, and hence, $\mathrm{E}[1/\sum(x_i-\overline{x})^2] \approx \{1 + [4/S_{xx}]\sigma_x^2\}/\{\sum(x_{i0}-\overline{x}_0)^2 + (n-1)\sigma_x^2\}$. Therefore, the expectation of the slope, β_E, can be derived as follows (see supplementary material for more details):

$$\begin{aligned} \beta_E &= \mathrm{E}[\textstyle\sum(x_i - \overline{x})y_i / \sum(x_i - \overline{x})^2] \\ &= \mathrm{E}[\textstyle\sum(x_i - \overline{x})y_i]\mathrm{E}[1/\sum(x_i - \overline{x})^2] + cov[\sum(x_i \\ &\quad - \overline{x})y_i, 1/\textstyle\sum(x_i - \overline{x})^2] \approx S_{xy}\{1 \\ &\quad + [4/S_{xx}]\sigma_x^2\}/\{\textstyle\sum(x_{i0} - \overline{x}_0)^2 + (n-1)\sigma_x^2\} \\ &\quad - [2/S_{xy}]\beta^2\sigma_x^2. \end{aligned}$$

If we apply the first-order Taylor approximation to simplify the expression $S_{xy}\{1 + [4/S_{xx}]\sigma_x^2\}/\{\sum(x_{i0} - \overline{x}_0)^2 + (n-1)\sigma_x^2\}$, we obtain the following expressions for β_E and α_E:

$$\begin{aligned} \beta_E = \mathrm{E}[\hat{\beta}] &\approx \beta - \beta[1/S_{xx}]^*(n-3)\sigma_x^2, \\ \alpha_E = \mathrm{E}[\hat{\alpha}] = \mathrm{E}[\overline{y} - \hat{\beta}\overline{x}] &\approx \alpha + \overline{x}_0\beta[1/S_{xx}]^*(n-3)\sigma_x^2. \end{aligned}$$

Accordingly, the biases in $\hat{\beta}$, $\hat{\alpha}$ and $\hat{y}$ are as follows:

$$\mathrm{bias}[\hat{\beta}] \approx -\beta[1/S_{xx}]^*(n-3)\sigma_x{}^2, \tag{6}$$

$$\begin{aligned} \mathrm{bias}[\hat{\alpha}] &\approx +\overline{x}_0\beta[1/S_{xx}]^*(n-3)\sigma_x{}^2, \\ \mathrm{bias}[\hat{y}] &= \mathrm{E}[\hat{\alpha} + \hat{\beta}x] - (\alpha + \beta x) \\ &= \mathrm{E}[\hat{\alpha}] + x\mathrm{E}[\hat{\beta}] - (\alpha + \beta x) \\ &\approx -(n-3)(x - \overline{x}_0)[1/S_{xy}]^*\sigma^2. \end{aligned} \tag{7}$$

Based on these biases, we can see that β and α are not the mean, median, or mode of the $\hat{\beta}$ and $\hat{\alpha}$ distributions. However, we can say that $\hat{\beta}$ and $\hat{\alpha}$, despite being slightly skewed, follow approximately normal distributions centered at β and α respectively, because the terms $\beta[1/S_{xx}]^*(n-3)\sigma_x^2$ and $\overline{x}_0\beta[1/S_{xx}]^*(n-3)\sigma_x^2$ are each very small in magnitude in practical calibrations. (When n is 3, β coincides with β_E. The same can be said of α and α_E.)

To show that the slope, intercept and predicted y value in reversed inverse regression can be expressed as t_{n-2} distributions, it is necessary to know the statistical properties of the mean squared error (*MSE*). The expectation of *MSE* is first derived (see supplementary material for more details):

$$\begin{aligned} cov[\hat{\beta}, \overline{x}] &\approx 0, \\ \textstyle\sum cov[\hat{\beta}, x_i] &\approx 0, \\ cov[\hat{\beta}^2, \textstyle\sum(x_i - \overline{x})^2] &\approx -4\hat{\beta}^2\sigma_x^2, \\ cov[\hat{\beta}, \textstyle\sum(x_i - \overline{x})y_i] &\approx (S_{yy}S_{xx}/S_{xy}^2 - 2)\hat{\beta}^2\sigma_x^2, \\ \textstyle\sum var[x_i - \overline{x}] &= (n-1)\sigma_x^2, \end{aligned}$$

$$\begin{aligned} \mathrm{E}^2[\hat{\beta}] &\approx \{\{\beta - \beta[1/S_{xx}]^*(n-3)\sigma_x^2\}^2\} \\ &\approx \beta^2 - 2\beta^2[1/S_{xx}]^*(n-3)\sigma_x^2, \end{aligned}$$

$$SSE = \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \sum \{\{y_i - (\overline{y} - \hat{\beta}\overline{x}) - \hat{\beta}x_i\}\}^2$$
$$= \sum \{\{(y_i - \overline{y}) - (\hat{\beta}x_i - \hat{\beta}\overline{x})\}\}^2$$
$$= \sum (y_i - \overline{y})^2 - 2\sum (y_i - \overline{y})\hat{\beta}(x_i - \overline{x})$$
$$+ \sum \hat{\beta}^2 (x_i - \overline{x})^2 \{\text{Typically}, y_i - \overline{y} \neq \hat{\beta}(x_i - \overline{x})\},$$

$$\mathrm{E}[SSE] = \mathrm{E}[\sum (y_i - \overline{y})^2] - 2\{\mathrm{E}[\hat{\beta}]\mathrm{E}[\sum (x_i - \overline{x})y_i]$$
$$+ cov[\hat{\beta}, \sum (x_i - \overline{x})y_i] - \overline{y}\mathrm{E}[\hat{\beta}]\mathrm{E}[\sum (x_i - \overline{x})]\}$$
$$+ \{var[\hat{\beta}] + \mathrm{E}^2[\hat{\beta}]\}\{\sum var[x_i - \overline{x}] + \sum \mathrm{E}^2[x_i - \overline{x}]\}$$
$$+ cov[\hat{\beta}^2, \sum (x_i - \overline{x})^2] \approx \{(n-1) - [S_{yy}S_{xx}/S_{xy}^2]^*\}\sigma_x^2\beta^2$$
$$= (n-2)\sigma^2.$$

$$MSE = \sum (y_i - \hat{\alpha} - \hat{\beta}\, x_i)^2/(n-2),$$
$$\therefore \mathrm{E}[MSE] \approx \sigma^2 (= \sigma_x^2\beta^2). \quad (8)$$

To investigate the accuracy of the expectation of *MSE* obtained using equation (8), we should consider the same factors taken into account in the case of the variance of $\hat{\beta}$. The accuracy of the derived E[*MSE*] is discussed in detail based on simulation results in Section 5.

The correlation coefficient between the slope and the mean squared error, $r(\hat{\beta}, MSE)$, is derived using the method of simultaneous error equations. Let $K = \hat{\beta}\sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = (S_{xx}S_{yy}S_{xy} - S_{xy}^3)/S_{xx}^2$, $A = \hat{\beta} = S_{xy}/S_{xx}$, and $F = \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = (S_{xx}S_{yy} - S_{xy}^2)/S_{xx}$. Then, two separate equations for deriving the variance of K can be established. The correlation coefficient $r(\hat{\beta}, MSE)$ is obtained from these two equations:

$$K = \hat{\beta} \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = (S_{xx}S_{yy}S_{xy} - S_{xy}^3)/S_{xx}^2,$$
$$K = AF,$$

$$\sigma_K^2 \approx \sum (\partial K/\partial x_i)^2 \sigma_{xi}^2,$$
$$\sigma_K^2 \approx (\partial K/\partial A)^2\sigma_A^2 + (\partial K/\partial F)^2\sigma_F^2 + 2(\partial K/\partial A)(\partial K/\partial F)$$
$$\times \sigma_A\sigma_F r(A, F),$$

$$r(A, F) = r(\hat{\beta}, \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2) = r(\hat{\beta}, SSE)$$
$$\approx -\{(S_{xx}S_{yy}S_{xy}$$
$$- S_{xy}^3)/S_{xx}\}\{S_{xx}/(S_{xx}S_{yy}^2S_{xy}^2 - S_{yy}S_{xy}^4)\}^{1/2}$$
$$= -\{1 - r^2(x, y)\}^{1/2} \approx 0,$$
$$\therefore r(\hat{\beta}, MSE) \approx 0.$$

Additionally, $\hat{\beta}$ and $\overline{x}$ are independent of each other and $\overline{x}$ and *MSE* are also independent of each other, then $r(\hat{\alpha}, MSE) = r(\overline{y} - \hat{\beta}\overline{x}, MSE) \approx 0$.

In the expression $\sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2/(n-2)$, the y_i's are constant, $\hat{\beta}$ and $\hat{\alpha}$ follow approximately normal distributions, and the x_i's also follow normal distributions. Therefore, $(n-2)MSE/\sigma^2$ approximately follows a χ^2 distribution with $n-2$ degrees of freedom. In addition, both $\hat{\beta}$ and $\hat{\alpha}$ are nearly independent of *MSE*. Based on these facts, the following expressions can be obtained (see equations (1), (2), (4) and (8)):

$$T_1 = (\hat{\beta} - \beta)/[S_{yy}/S_{xy}^2]^{1/2}\hat{\sigma} \sim t_{n-2},$$
$$T_2 = (\hat{\alpha} - \alpha)/[1/n + \overline{x}^2(S_{yy}/S_{xy}^2)]^{1/2}\hat{\sigma} \sim t_{n-2},$$
$$T_3 = \{\hat{y} - (\alpha + \beta x)\}/[1/n + (x - \overline{x})^2$$
$$(S_{yy}/S_{xy}^2)]^{1/2}\hat{\sigma} \sim t_{n-2}, T_4 = \{y_0 - (\hat{\alpha} + \hat{\beta}x)\}/$$
$$[1 + 1/n + (x - \overline{x})^2(S_{yy}/S_{xy}^2)]^{1/2}\hat{\sigma} \sim t_{n-2},$$

where $\hat{\sigma}$ is the square root of *MSE* and y_0 is the nominal y value of a newly prepared standard. The T's are all approximate t_{n-2} distributions. Although $\overline{x}^2$, $(x - \overline{x})^2$ and S_{yy}/S_{xy}^2, which appear in the T's, are functions of x_i ($i = 1$, ..., n), the t_{n-2} distributions are not greatly deformed by these functions because the fluctuations of S_{yy}/S_{xy}^2 (or $[1/n + \overline{x}^2\,(S_{yy}/S_{xy}^2)]$) corresponding to the variations of the x_i's are typically very small compared with the magnitude of S_{yy}/S_{xy}^2 (or $[1/n + \overline{x}^2\,(S_{yy}/S_{xy}^2)]$) itself. Based on these t_{n-2} distributions, we can evaluate the uncertainty (or confidence interval) of a measurement value determined based on the fitted regression line.

4 Comparison of regression approaches

Krutchkoff [6,13] compared classical regression and inverse regression using Monte Carlo simulations and recommended inverse regression based on the mean squared error. However, Berkson [14] and Halpern [15] presented significant criticisms of Krutchkoff's work. Parker et al. [4] also conducted several simulation studies and concluded that inverse regression performs better than classical regression. It seems that such debates arise because the existing regression approaches and accompanying methodologies are theoretically incomplete. Unusually, we compare different linear regression approaches using a practical calibration example. Each of three types of regression (classical, inverse and reversed inverse) is applied to the calibration scenario. In practical calibrations, the variance of the prediction interval is one of the most important statistical properties. Therefore, we identify the differences among the three regressions based on a comparison of the variances of the prediction interval estimated using the three regression approaches. For the fitting of a regression line as an example of practical calibration, we use a set of data points collected by Suh [16] while evaluating the uncertainty in the measurements recorded by an absorption spectrometer. The spectrometer determines the chemical concentrations (ppm) in a sample by measuring the absorbances (%) due to the corresponding chemical elements. Suh measured five different Cd (cadmium) standards. The data points collected by Suh and the calibration results are as follows:

$$(0.1\text{ppm}, 0.028\%), (0.3\text{ppm}, 0.084\%), (0.5\text{ppm}, 0.135\%),$$
$$(0.7\text{ppm}, 0.180\%), (0.9\text{ppm}, 0.215\%).$$

4.1 Classical regression

– x: Cd concentration (ppm), y: absorbance (%).
– $\overline{x} = 0.5$, $\overline{y} = 0.1284$, $S_{xx} = 0.4$, $S_{yy} = 0.02225$, $S_{xy} = 0.094$, $r(x, y) = S_{xy}/(S_{xx}S_{yy})^{1/2} = 0.9964$.

– $MSE(\hat{\sigma}^2) = \sum (y_i - \hat{\alpha}' - \hat{\beta}' x_i)^2/(5-2) = 0.000056$, $\hat{\beta}' = S_{xy}/S_{xx} = 0.235$, $\hat{\alpha}' = 0.0109$.
– Regression line: $\hat{x} = -\hat{\alpha}'/\hat{\beta}' + (1/\hat{\beta}')y$.
– Estimator for the variance of the prediction interval (EV_C): $[1 + 1/n + (x - \overline{x})^2/S_{xx})]\hat{\sigma}^2(1/\beta')^2$.
  • $\hat{x} = -0.04638 + 4.25532y$. (Measurement equation)
  • $EV_C = \{1 + 1/5 + (0.8685 - 0.5)^2/0.4\}0.000056(1/0.235)^2 = 0.0015611$ (at $x = 0.8685$ ppm).

Note: $-0.04638 + 4.25532 \times 0.215(\%) = 0.8685$ (ppm).

4.2 Inverse regression

– x: Cd concentration (ppm), y: absorbance (%).
– $\overline{x} = 0.5, \overline{y} = 0.1284, S_{xx} = 0.4, S_{yy} = 0.02225, S_{xy} = 0.094, r(x, y) = S_{xy}/(S_{xx}S_{yy})^{1/2} = 0.9964$.
– $MSE(\hat{\sigma}^2) = \sum(x_i - \gamma'_0 - \gamma'_1 y_i)^2/(5-2) = 0.001$, $\gamma'_1 = S_{xy}/S_{yy} = 4.22472$, $\gamma'_0 = -0.04245$.
– Regression line: $\hat{x} = \gamma'_0 + \gamma'_1 y$.
– Estimator for the variance of the prediction interval (EV_I): $[1 + 1/n + (y - \overline{y})^2/S_{yy}]\hat{\sigma}^2$.
  • $\hat{x} = -0.04245 + 4.22472y$. (Measurement equation)
  • $EV_I = \{1 + 1/5 + (0.215 - 0.1284)^2/0.02225\} \times 0.001 = 0.0015371$ (at $y = 0.215\%$).

4.3 Reversed inverse regression

– x: absorbance (%), y: Cd concentration (ppm).
– $\overline{x} = 0.1284, \overline{y} = 0.5, S_{xx} = 0.02225, S_{yy} = 0.4, S_{xy} = 0.094, r(x, y) = S_{xy}/(S_{xx}S_{yy})^{1/2} = 0.9964$.
– $MSE(\hat{\sigma}^2) = \sum(y_i - \hat{\alpha} - \hat{\beta}x_i)^2/(5-2) = 0.001$, $\hat{\beta} = S_{xy}/S_{xx} = 4.22472$, $\hat{\alpha} = -0.04245$.
– Regression line: $\hat{y} = \hat{\alpha} + \hat{\beta}x$.
– Estimator for the variance of the prediction interval (EV_{RI}): $[1 + 1/n + (x - \overline{x})^2(S_{yy}/S_{xy}^2)]\hat{\sigma}^2$.
  • $\hat{y} = -0.04245 + 4.22472x$. (Measurement equation)
  • $EV_{RI} = \{1 + 1/5 + (0.215 - 0.1284)^2(0.4/0.094^2)\} \times 0.001 = 0.0015395$ (at $x = 0.215\%$).

The estimate EV_{RI} derived via reversed inverse regression at $x = 0.215\%$ (the upper end of the calibration range) is compared with the estimate EV_C derived via classical regression at $x = 0.8685$ ppm and with the estimate EV_I derived via inverse regression at $y = 0.215\%$. All three estimates are different from one another. Classical regression yields the largest estimate, and inverse regression yields the smallest one. This can be explained by rewriting and comparing the following three estimators. (Both EV_C and EV_I are those derived by Parker et al. [4].) When rewriting EV_C and EV_I, the roles of x and y were reversed to facilitate comparison. In addition, $\hat{\sigma}_y^2(1/\hat{\beta}')^2$ in the expression for classical regression was changed to $\hat{\sigma}_x^2(1/\hat{\beta}')^2$.

$$EV_C = [1 + 1/n + (\hat{y} - \overline{y})^2/S_{yy}]\hat{\sigma}_x^2(1/\hat{\beta}')^2,$$
$$EV_I = [1 + 1/n + (x - \overline{x})^2/S_{xx}]\hat{\sigma}_2,$$
$$EV_{RI} = [1 + 1/n + (x - \overline{x})^2(S_{yy}/S_{xy}^2)]\hat{\sigma}_2$$

The correlation coefficient $r(x, y)$ $\{= S_{xy}/(S_{xx}S_{yy})^{1/2}\}$ is very close to, but always smaller than, 1 in linear calibrations. In addition, $Syy/S_{xy}^2 = (1/S_{xx})1/r^2(x, y)$ and $(\hat{y} - \overline{y})^2/S_{yy} = (x - \overline{x})^2(S_{yy}/S_{xy}^2)$. The term $\hat{\sigma}_x^2(1/\hat{\beta}')^2$ is greater than $\hat{\sigma}^2$. Therefore, the estimates can be arranged in order of increasing magnitude as follows: "inverse," "reversed inverse" and then "classical". This ordering holds for all linear calibrations. The differences among the three estimates depend on $r(x, y)$. In Suh's measurement experiment, $r(x, y)$ is 0.9964 ($n = 5$), the estimate derived via classical regression at the upper end of the calibration range is approximately 1.5% greater than that derived via inverse regression, and the estimate derived via reversed inverse regression is approximately 0.15% greater than that derived via inverse regression. If Suh had repeated this measurement experiment, the results would have been similar to those of this calibration. Regarding these calibration results, we should remind ourselves that although we rely on the estimate derived via classical regression, we cannot determine the prediction interval with a given confidence level because the estimate cannot be used to express the prediction interval as a t_{n-2} distribution. In addition, we should remind ourselves that the estimate derived via inverse regression is not a theoretically correct one.

5 Simulation study

We conducted a Monte Carlo simulation study to investigate the accuracy of the statistical properties derived using the error propagation rule and the method of simultaneous error equations based on the first-order Taylor approximation. var[$\hat{\beta}$], bias[$\hat{\beta}$] and E[MSE] were the main targets of investigation because the accuracy of other properties, such as var[$\hat{\alpha}$], bias[$\hat{\alpha}$], var[$\hat{y}$], bias[$\hat{y}$] and var [prediction interval], depends on the accuracy of these three properties. We designed a simulation of regression line fitting using five data points based on reversed inverse regression. We first created five intended mean data points (x_{i0}, y_i) ($i = 1, ..., 5$) that were needed for the simulation as follows:

$$(x_{1,0} = 412, y_1 = 10), (x_{2,0} = 812, y_2 = 20), (x_{3,0} = 1212, y_3 = 30), (x_{4,0} = 1612, y_4 = 40), (x_{5,0} = 2012, y_5 = 50).$$

- $\overline{x}_0 = 1212, \overline{y} = 30$.
- Intended population regression line: $y = -0.3 + 0.025x$ ($\beta = 0.025$, $\alpha = -0.3$).

Depending on the intended variance σ_x^2, the simulation study was organized into five simulation groups, SG1, SG2, SG3, SG4 and SG5, and the intended variances assigned to the five groups were 90^2, 60^2, 24^2, 12^2 and 6^2, respectively. Five simulations per group were conducted (25 simulations in total). In every simulation, the regression line fitting was repeated 50 000 times using independent random numbers generated from normal distributions using the program "Minitab 15". The results of the conducted simulations are presented along with the corresponding theoretically derived properties in Tables 1 and 2. (Even if different parameters, such as a different number of data points, a different ratio of $\overline{y}$ to $\overline{x}_0$, or non-equal distances between

Table 1. Simulation results and theoretically derived properties.

Group	Simulation results			Derived properties		
	Svar$[\hat{\beta}]$[a]	Sbias$[\hat{\beta}]$[b]	SE[*MSE*][c]	Dvar$[\hat{\beta}]$[d]	Dbias$[\hat{\beta}]$[e]	DE[*MSE*][f]
SG1-1	$(0.0017625)^2$	−0.0002535	5.013210	$(0.0017788)^2$	−0.0002531	5.0625
SG1-2	$(0.0017582)^2$	−0.0002515	5.012360	$(0.0017788)^2$	−0.0002531	5.0625
SG1-3	$(0.0017689)^2$	−0.0002360	4.978110	$(0.0017788)^2$	−0.0002531	5.0625
SG1-4	$(0.0017622)^2$	−0.0002648	5.030650	$(0.0017788)^2$	−0.0002531	5.0625
SG1-5	$(0.0017532)^2$	−0.0002541	5.017950	$(0.0017788)^2$	−0.0002531	5.0625
SG2-1	$(0.0011780)^2$	−0.0001035	2.233260	$(0.0011859)^2$	−0.0001125	2.2500
SG2-2	$(0.0011789)^2$	−0.0001180	2.236960	$(0.0011859)^2$	−0.0001125	2.2500
SG2-3	$(0.0011828)^2$	−0.0001127	2.245950	$(0.0011859)^2$	−0.0001125	2.2500
SG2-4	$(0.0011804)^2$	−0.0001179	2.239320	$(0.0011859)^2$	−0.0001125	2.2500
SG2-5	$(0.0011828)^2$	−0.0001073	2.221680	$(0.0011859)^2$	−0.0001125	2.2500
SG3-1	$(0.0004773)^2$	−0.0000176	0.358494	$(0.0004734)^2$	−0.0000180	0.3600
SG3-2	$(0.0004726)^2$	−0.0000162	0.359327	$(0.0004734)^2$	−0.0000180	0.3600
SG3-3	$(0.0004717)^2$	−0.0000197	0.359700	$(0.0004734)^2$	−0.0000180	0.3600
SG3-4	$(0.0004740)^2$	−0.0000182	0.358222	$(0.0004734)^2$	−0.0000180	0.3600
SG3-5	$(0.0004752)^2$	−0.0000162	0.360404	$(0.0004734)^2$	−0.0000180	0.3600
SG4-1	$(0.0002386)^2$	−0.0000035	0.090080	$(0.0002372)^2$	−0.0000035	0.0900
SG4-2	$(0.0002387)^2$	−0.0000042	0.090021	$(0.0002372)^2$	−0.0000035	0.0900
SG4-3	$(0.0002385)^2$	−0.0000042	0.090132	$(0.0002372)^2$	−0.0000035	0.0900
SG4-4	$(0.0002369)^2$	−0.0000052	0.089894	$(0.0002372)^2$	−0.0000035	0.0900
SG4-5	$(0.0002374)^2$	−0.0000052	0.090035	$(0.0002372)^2$	−0.0000035	0.0900
SG5-1	$(0.0001188)^2$	−0.0000012	0.022492	$(0.0001186)^2$	−0.0000011	0.0225
SG5-2	$(0.0001186)^2$	−0.0000011	0.022493	$(0.0001186)^2$	−0.0000011	0.0225
SG5-3	$(0.0001182)^2$	−0.0000008	0.022422	$(0.0001186)^2$	−0.0000011	0.0225
SG5-4	$(0.0001189)^2$	−0.0000010	0.022408	$(0.0001186)^2$	−0.0000011	0.0225
SG5-5	$(0.0001190)^2$	−0.0000019	0.022542	$(0.0001186)^2$	−0.0000011	0.0225

[a] Svar$[\hat{\beta}]$ is the variance of $\hat{\beta}$ observed through each simulation. The square of the standard deviation of the 50 000 slopes obtained from the 50 000 regression line fittings is treated as Svar$[\hat{\beta}]$.
[b] Sbias$[\hat{\beta}]$ is the mean of the 50 000 slopes subtracted by the slope of the population regression line.
[c] SE[*MSE*] is the mean of the 50 000 *MSE*s obtained from the 50 000 regression line fittings.
[d] Dvar$[\hat{\beta}]$ is the variance of $\hat{\beta}$ derived using equation (1) based on the intended mean data points and the intended variance σ_x^2.
[e] Dbias$[\hat{\beta}]$ is the bias in $\hat{\beta}$ derived using equation (6).
[f] DE[*MSE*] is the expectation of *MSE* derived using equation (8).

the x_{i0}'s, were applied in a simulation study, such a simulation study would yield conclusions essentially similar to those of this study.)

In Tables 1 and 2, the ratio of Svar$[\hat{\beta}]$ to Dvar$[\hat{\beta}]$ ranges from 0.971 to 1.017 and the ratio of SE[*MSE*] to DE[*MSE*] ranges from 0.983 to 1.002. In addition, the two derived variances *Dvar$[\hat{\beta}]$ and Dvar$[\hat{\beta}]$ are very close to each other. Therefore, we can conclude that the variance of the slope and the expectation of the mean squared error derived using the error propagation rule and the method of simultaneous error equations largely coincide with the simulation results.

According to Table 1, when σ_x^2 is 6^2, the ratio of bias$[\hat{\beta}]$ to $\{\text{var}[\hat{\beta}]\}^{1/2}$ is approximately −0.01, and when σ_x^2 is 90^2, the ratio is approximately −0.14. These two ratios are very different from each other in magnitude. In the case of either simulation or derivation, as the variance σ_x^2 increases, both the absolute value of the bias in $\hat{\beta}$ and the variance of $\hat{\beta}$ increase. The rate of increase of the absolute value of the bias in $\hat{\beta}$ is equal to the rate of increase of σ_x^2 (see equation (6)), whereas the rate of increase of $\{\text{var}[\hat{\beta}]\}^{1/2}$ is the square root of the rate of increase of σ_x^2 (see equation (1)). This indicates that as σ_x^2 increases, the $\hat{\beta}$ distribution becomes more skewed. In Tables 1 and 2, the derived values of the bias in $\hat{\beta}$ largely coincide with

Table 2. Ratios of the simulation results to the corresponding derived properties.

Group	Ratios			$^{*}\mathrm{Dvar}[\hat{\beta}]$[a]	$\mathrm{Dvar}[\hat{\beta}]$	$\mathrm{SMean}[\hat{\beta}]$[b]	σ_x^2
	$\mathrm{Svar}[\hat{\beta}]/\mathrm{Dvar}[\hat{\beta}]$	$\mathrm{Sbias}[\hat{\beta}]/\mathrm{Dbias}[\hat{\beta}]$	$\mathrm{SE}[MSE]/\mathrm{DE}[MSE]$				
SG1-1	0.982	1.002	0.990	$(0.0018022)^2$	$(0.0017788)^2$	0.0247465	90^2
SG1-2	0.977	0.994	0.990	$(0.0018017)^2$	$(0.0017788)^2$	0.0247485	90^2
SG1-3	0.989	0.932	0.983	$(0.0017958)^2$	$(0.0017788)^2$	0.0247640	90^2
SG1-4	0.981	1.046	0.994	$(0.0018048)^2$	$(0.0017788)^2$	0.0247352	90^2
SG1-5	0.971	1.004	0.991	$(0.0018030)^2$	$(0.0017788)^2$	0.0247459	90^2
SG2-1	0.987	0.920	0.993	$(0.0011914)^2$	$(0.0011859)^2$	0.0248965	60^2
SG2-2	0.989	1.049	0.994	$(0.0011915)^2$	$(0.0011859)^2$	0.0248820	60^2
SG2-3	0.995	1.002	0.998	$(0.0011943)^2$	$(0.0011859)^2$	0.0248873	60^2
SG2-4	0.991	1.048	0.995	$(0.0011922)^2$	$(0.0011859)^2$	0.0248821	60^2
SG2-5	0.995	0.954	0.987	$(0.0011877)^2$	$(0.0011859)^2$	0.0248927	60^2
SG3-1	1.017	0.978	0.996	$(0.0004739)^2$	$(0.0004734)^2$	0.0249824	24^2
SG3-2	0.997	0.900	0.998	$(0.0004745)^2$	$(0.0004734)^2$	0.0249838	24^2
SG3-3	0.993	1.094	0.999	$(0.0004746)^2$	$(0.0004734)^2$	0.0249803	24^2
SG3-4	1.003	1.011	0.995	$(0.0004738)^2$	$(0.0004734)^2$	0.0249818	24^2
SG3-5	1.008	0.900	1.001	$(0.0004753)^2$	$(0.0004734)^2$	0.0249838	24^2
SG4-1	1.012	1.000	1.001	$(0.0002373)^2$	$(0.0002372)^2$	0.0249965	12^2
SG4-2	1.013	1.200	1.000	$(0.0002372)^2$	$(0.0002372)^2$	0.0249958	12^2
SG4-3	1.011	1.200	1.001	$(0.0002374)^2$	$(0.0002372)^2$	0.0249958	12^2
SG4-4	0.997	1.486	0.999	$(0.0002371)^2$	$(0.0002372)^2$	0.0249948	12^2
SG4-5	1.002	1.486	1.000	$(0.0002373)^2$	$(0.0002372)^2$	0.0249948	12^2
SG5-1	1.003	1.091	1.000	$(0.0001186)^2$	$(0.0001186)^2$	0.0249988	6^2
SG5-2	1.000	1.000	1.000	$(0.0001186)^2$	$(0.0001186)^2$	0.0249989	6^2
SG5-3	0.993	0.727	0.997	$(0.0001184)^2$	$(0.0001186)^2$	0.0249992	6^2
SG5-4	1.005	0.909	0.996	$(0.0001183)^2$	$(0.0001186)^2$	0.0249990	6^2
SG5-5	1.007	1.727	1.002	$(0.0001187)^2$	$(0.0001186)^2$	0.0249981	6^2

[a] In every simulation, the estimate for the variance of $\hat{\beta}$, i.e., $(S_{yy}/S_{xy}^2)\hat{\sigma}^2$, was calculated for each regression line. $^{*}\mathrm{Dvar}[\hat{\beta}]$ is the mean of the 50 000 estimates thus calculated.

[b] $\mathrm{SMean}[\hat{\beta}]$ is the mean of the 50 000 slopes obtained from the 50 000 regression line fittings.

the simulation results regardless of σ_x^2. This indicates that although the first-order Taylor approximation is used to derive the bias in $\hat{\beta}$, the derived bias does not greatly differ from the simulation result. The bias in $\hat{\beta}$ plays an important role in analyzing the accuracy of other derived statistical properties.

When σ_x^2 is small, the derived variance of $\hat{\beta}$ exactly coincides with the simulation result; however, when σ_x^2 is large, the derived variance of $\hat{\beta}$ is generally slightly greater than the simulation result. When the variance of $\hat{\beta}$ (i.e., $var[\hat{\beta}] = [S_{yy}/S_{xy}^2]^{*}\sigma_x^2\beta^2$) is derived using the error propagation rule, the partial derivatives of orders higher than the first are not included in the derivation, and the approximation $\mathrm{var}[f(x_1, \ldots, x_n)] \approx \mathrm{E}[\{f(x_1, \ldots, x_n) - f(x_{10}, \ldots, x_{n0})\}^2]$ is used instead of the exact definition $\mathrm{var}[f(x_1, \ldots, x_n)] = \mathrm{E}[\{f(x_1, \ldots, x_n) - \mathrm{E}[\{f(x_1, \ldots, x_n)]\}^2]$ to derive the variance. This results in two phenomena. The first phenomenon is that error terms of orders higher than σ_x^2 are excluded from the derivation, and the second phenomenon is that the bias in $\hat{\beta}$ is not reflected in the derivation. The bias in $\hat{\beta}$ depends on σ_x^2 and n (see equation (6)). In this simulation study, n is 5. The first phenomenon typically causes the derived variance of $\hat{\beta}$ (i.e., $\mathrm{Dvar}[\hat{\beta}]$) to decrease, whereas the second phenomenon tends to cause it to increase. If σ_x^2 is small, both effects are trivial, and $\mathrm{Dvar}[\hat{\beta}]$ is nearly equal to $\mathrm{Svar}[\hat{\beta}]$. If σ_x^2 is large, both of these effects are also large. However, the effect of the second phenomenon is much greater than that of the first. As a result, if σ_x^2 is large, then $\mathrm{Dvar}[\hat{\beta}]$ is greater than $\mathrm{Svar}[\hat{\beta}]$. If we substitute β_E ($\mathrm{SMean}[\hat{\beta}]$ in Table 2) into equation (1) in place of β, we can obtain a variance of $\hat{\beta}$ that is much closer to the simulation result. For example, for SG1-1, we can obtain $(1000/40\,000^2) \times 90^2 \times 0.0247465^2 = 0.0017607^2$ by substituting β_E ($= 0.0247465$) into $[S_{yy}/S_{xy}^2]^{*}\sigma_x^2\beta_{\mathrm{E}}^2$. (The difference between $[S_{yy}/S_{xy}^2]^{*}\sigma_x^2\beta^2$ and $[S_{yy}/S_{xy}^2]^{*}\sigma_x^2\beta_{\mathrm{E}}^2$ is

approximately equal to the square of bias$[\hat{\beta}]$.) This value is very close to the simulation result. The difference that still remains can be regarded as the effect of the first phenomenon.

With regard to the expectation of the mean squared error, a similar explanation is possible. Even in this case, the effect of the second phenomenon is greater than that of the first phenomenon, and hence, DE[*MSE*] is generally greater than SE[*MSE*]. In particular, let us attempt to approximately calculate the effect of the first phenomenon using another expression for the expectation of *MSE*. Namely, in the equation $\mathrm{E}[MSE] = E[(S_{xx}S_{yy} - S_{xy}^2)/(n-2)S_{xx}] \approx \mathrm{E}[\sigma_x^2\hat{\beta}^2] = \sigma_x^2\mathrm{E}^2[\hat{\beta}] + var[\hat{\beta}] \approx \sigma_x^2\beta_{\mathrm{E}}^2 + \beta^2[S_{yy}/S_{xy}^2]$ $^*\sigma_x^4$, the last term on the right-hand side reflects the effect of the first phenomenon to a certain extent. This equation helps us understand the two phenomena.

In Table 2, if σ_x^2 is large, then $^*\mathrm{Dvar}[\hat{\beta}]$ is generally greater than $\mathrm{Dvar}[\hat{\beta}]$. In every simulation, the estimate for the variance of the slope, i.e., $(S_{yy}/S_{xy}^2)\hat{\sigma}^2$, was calculated for each regression line. $^*\mathrm{Dvar}[\hat{\beta}]$ is the mean of the 50,000 estimates thus calculated. We can also obtain $^*\mathrm{Dvar}[\hat{\beta}]$ using another method, as follows:

$$
\begin{aligned}
^*\mathrm{Dvar}[\hat{\beta}] &= \mathrm{E}[(S_{yy}/S_{xy}^2)(S_{xx}S_{yy} - S_{xy}^2)/S_{xx}(n-2)] \\
&= \mathrm{E}[(S_{yy}/S_{xy}^2)\hat{\sigma}^2] \approx \mathrm{E}[(S_{yy}/S_{xy}^2)\sigma_x^2\hat{\beta}^2] \\
&= \mathrm{E}[(S_{yy}/S_{xx}^2)\sigma_x^2] = S_{yy}\sigma_x^2\mathrm{E}[1/S_{xx}^2] \\
&= S_{yy}\sigma_x^2\{\mathrm{E}^2[1/S_{xx}] + var[1/S_{xx}]\} \\
&\approx \{S_{yy}\sigma_x^2/S_{xx}^2 + 2(7-\mathrm{n})S_{yy}\sigma_x^4/S_{xx}^3\}^*.
\end{aligned}
$$

The term $2(7-n)S_{yy}\sigma_x^4/S_{xx}^{3\,*}$ reflects the difference between $^*\mathrm{Dvar}[\hat{\beta}]$ and $\mathrm{Dvar}[\hat{\beta}]$. The difference depends on σ_x^4 and n.

In this section, we investigated the accuracy of the statistical properties of reversed inverse regression as derived using the error propagation rule and the method of simultaneous error equations through comparisons with simulation results. However, it should be noted that the main target that calibration experts wish to obtain (or approach) by means of regression line fitting is the population regression line $y = \alpha + \beta x$, not the average regression line $y = \alpha_E + \beta_E x$. In this respect, it is recommended that after the physical or chemical value of a sample is determined based on the fitted regression line, the determined value be corrected taking into account the bias in predicted y value (see equation (7)); such a bias correction will lead us closer to the true value.

6 Conclusion

From Osborne [17], it can be seen that considerable effort has been made to resolve the linear calibration problem since the 1930s. Most representatively, Eisenhart [2] suggested classical regression as a solution for the problem, and Krutchkoff [6] suggested inverse regression as another solution. Later, Parker et al. [4] derived the variances of the prediction interval and the biases in $\hat{x}$ for these two types of regression using the Delta Method. However, it can be said that the problem has not yet been resolved completely. As a fundamental solution for this problem, the current study introduced reversed inverse regression along with a methodology for deriving its statistical properties. In this study, the statistical properties of reversed inverse regression, such as the variance and bias of the slope, the expectation of the mean squared error, and the variance of the predicted y value, were derived using the error propagation rule and the method of simultaneous error equations. The method of simultaneous error equations, which was introduced for the first time in this study, is a useful tool for deriving the covariance of any two statistics. As another example of its use, all of the statistical properties of basic regression can be derived much more easily with the aid of this method. Even in the case of weighted linear regression, this method can be used to derive its statistical properties.

We presented an example of practical calibration. Each of the three types of regression (i.e., classical, inverse and reversed inverse) was applied to this calibration example. As a result, we found that the estimates of the variance of the prediction interval can be arranged in order of increasing magnitude as follows: "inverse," "reversed inverse" and then "classical". This ordering holds for all linear calibrations. The differences among the three estimates depend on $r(x, y)$. As the next step, to investigate the accuracy of the three derived statistical properties of reversed inverse regression, i.e., $\mathrm{Dvar}[\hat{\beta}]$, $\mathrm{Dbias}[\hat{\beta}]$ and DE[*MSE*], a Monte Carlo simulation study was conducted. Through this simulation study, we found that when the variance of the observed measurements, i.e., σ_x^2, is small, the theoretically derived variance and bias of the slope as well as the theoretically derived expectation of the mean squared error coincide with the simulation results. However, when σ_x^2 is large, there are small differences between the derived properties and the simulation results. Such differences are caused by two phenomena. The first phenomenon is that error terms of orders higher than σ_x^2 are excluded from the derivation, and the second phenomenon is that the bias in $\hat{\beta}$ is not reflected in the derivation. The first phenomenon typically causes the derived statistical properties to decrease, whereas the second phenomenon tends to cause them to increase (when n is greater than 3). The effect of the second phenomenon is larger than that of the first phenomenon, and hence, the values of the derived properties are typically slightly greater than the simulation results. In this way, after performing simulations we could investigate and analyze the differences between the derived statistical properties and the simulation results. This is another benefit of the new methodology used to derive the statistical properties of reversed inverse regression.

7 Implications and influences

Lwin and Maritz [18] suggested that regression models do not require the assumption of fixed inputs. In other words, regardless of whether the regression model of interest is consistent with this assumption, the method of least squares can be applied to fit a regression line. In that sense,

it is meaningless to identify whether the line fitted using one regression approach is preferable to that fitted using another regression approach. However, it is nevertheless essential to know the statistical properties of the type of regression used for fitting. Unfortunately, the known statistical properties of the existing regression approaches are not without flaw. By contrast, all of the statistical properties of reversed inverse regression can be derived using the newly proposed methodology, and the statistical properties derived in this manner are theoretically correct and sufficiently accurate. In this respect, we claim that reversed inverse regression and the new methodology for deriving its statistical properties together serve as a fundamental solution for the univariate linear calibration problem, which had not previously been completely resolved. Finally, we expect this new methodology to be widely used in the field of calibration.

Supplementary Material

Derivations of the statistical properties of reversed inverse regression. The Supplementary Material is available at https://www.metrology-journal.org/10.1051/ijmqe/2017021/olm.

The study reported in this paper was conducted as part of a plan to improve the quality assurance and control system of KEPCO Nuclear Fuel. The authors would like to express their thanks for the support from their company, without which the study could not have been successfully completed. In particular, the authors would like to express special thanks to President & CEO, Jaehee Lee; Executive Vice President & Chief Production Officer, Sundoo Kim; and Ex-Executive Vice President & Chief Production Officer, Chuljoo Park, who cordially supported and encouraged the authors in their study on the statistical theory and development of a new calibration approach using a regression model.

References

1. R.E. Walpole, R.H. Myers, *Probability and Statistics for Engineers and Scientists*, 5th edn. (Macmillan Publishing Company, London, 1993)
2. C. Eisenhart, The interpretation of certain regression methods and their use in biological and industrial research, Ann. Math. Stat. **10**, 162–186 (1939)
3. E.J. Williams, A note on regression methods in calibration, Technometrics **11**, 189–192 (1969)
4. P.A. Parker, G.G. Vining, S.R. Wilson, J.L. Szarka III, N.G. Johnson, The prediction properties of inverse and reverse regression for the simple linear calibration problem, J. Qual. Technol. **42**, 332–347 (2010)
5. G. Casella, R.L. Berger, *Statistical Inference*, 2nd edn. (Duxbury, Pacific Grove, 2002)
6. R.G. Krutchkoff, Classical and inverse regression methods, Technometrics **9**, 425–439 (1967)
7. G.K. Shukla, P. Datta, Comparison of the inverse estimator with the classical estimator subject to a preliminary test in linear calibration, J. Stat. Plan. Inference **12**, 93–102 (1985)
8. S.D. Oman, An exact formula for the M.S.E. of the inverse estimator in the linear calibration problem, J. Stat. Plan. Inference **11**, 189–196 (1985)
9. W. Fuller, *Measurement Error Models* (John Wiley & Sons, Hoboken, 1987)
10. T. Pham-Gia, N. Turkkan, E. Marchand, Density of the ratio of two normal random variables and applications, Commun. Stat. Theory Methods **35**, 1569–1591 (2006)
11. N. Tsoulfanidis, *Measurement and Detection of Radiation* (Hemisphere Publishing Corporation, Washington, 1983), 332 p.
12. A. Papanicolaou, *Taylor Approximation and the Delta Method* (coursehero, Stanford, 2009), 103 p.
13. R.G. Krutchkoff, Classical and inverse regression methods in extrapolation, Technometrics **11**, 605–608 (1969)
14. J. Berkson, Estimation of a linear function for a calibration line; consideration of a recent proposal, Technometrics **11**, 647–660 (1969)
15. M. Halpern, On inverse estimation in linear regression, Technometrics **12**, 727–736 (1970)
16. M.Y. Suh, *Methods for the Calculation of Uncertainty in Analytical Chemistry, KAERI/TR1602/2000 Korean Language* (Korea Atomic Energy Research Institute, Daejeon, 2000)
17. C. Osborne, Statistical calibration: a review, Int. Stat. Rev. **59**, 309–336 (1991)
18. T. Lwin, J.S. Maritz, An analysis of the linear calibration controversy from the perspective of compound estimation, Technometrics **24**, 235–242 (1982)

Cite this article as: Pilsang Kang, Changhoi Koo, Hokyu Roh, Reversed inverse regression for the univariate linear calibration and its statistical properties derived using a new methodology, Int. J. Metrol. Qual. Eng. **8**, 28 (2017)