CLIN. CHEM. 25/3, 432–438 (1979)

Incorrect Least-Squares Regression Coefficients in Method-Comparison Analysis

P. Joanne Cornbleet[1] and Nathan Gochman

The least-squares method is frequently used to calculate the slope and intercept of the best line through a set of data points. However, least-squares regression slopes and intercepts may be incorrect if the underlying assumptions of the least-squares model are not met. Two factors in particular that may result in incorrect least-squares regression coefficients are: (*a*) imprecision in the measurement of the independent (*x*-axis) variable and (*b*) inclusion of outliers in the data analysis. We compared the methods of Deming, Mandel, and Bartlett in estimating the known slope of a regression line when the independent variable is measured with imprecision, and found the method of Deming to be the most useful. Significant error in the least-squares slope estimation occurs when the ratio of the standard deviation of measurement of a single *x* value to the standard deviation of the *x*-data set exceeds 0.2. Errors in the least-squares coefficients attributable to outliers can be avoided by eliminating data points whose vertical distance from the regression line exceeds four times the standard error of the estimate.

Linear regression analysis is a commonly used technique in analyzing method-comparison data. If a linear relationship between the test and reference method can be defined, then the slope and intercept of this line can provide estimates of the proportional and constant error between the two methods (*1*). Furthermore, a value for the test method can be predicted from any reference method value within the range of the data set by the regression equation:[2]

$$y = a_{y \cdot x} + b_{y \cdot x}\, x$$

where x is the independent variable (reference method), y is the dependent variable (test method), $b_{y \cdot x}$ is the slope, and $a_{y \cdot x}$ is the intercept of the regression line.

The least-squares method is the most commonly used statistical technique to estimate the slope and intercept of linearly related comparison data. However, if the basic assumptions underlying the least-squares model are not met, the estimated line may be incorrect. It is the purpose of this paper to discuss three criteria for the use of least-squares regression analysis that are frequently violated in analyzing laboratory comparison data, to demonstrate the magnitude of the error in calculating the slope of the line by the least-squares method when these assumptions are not met, and to suggest alternative techniques for calculating the correct linear relationship between the two variables.

The line obtained by least-squares regression minimizes the sum of squares of the distances between the observed data points and the line in a vertical direction (Figure 1). These distances between the y values observed and those predicted by the regression line are called residuals. For the least-squares model to be valid, these residuals should be random (independent of values of x and y) and have a gaussian distribution with a mean of zero and standard deviation, $S_{y \cdot x}$. The standard deviation of the residuals (or standard error of the estimate) should be constant at every value of x; i.e., at each value of x, repeated measurements of y would have a standard deviation of $S_{y \cdot x}$. If x is a precisely measured reference method, and y an imprecise test method with a constant coefficient of variation rather than a constant measurement error at all values, then $S_{y \cdot x}$ will increase with increasing values of x, in which case a weighted regression analysis should be used (*2*). However, we will demonstrate that within the range of measurement error likely to be encountered in the laboratory (coefficient of variation up to 20%), the least-squares regression still calculates the correct line when $S_{y \cdot x}$ is proportional to x.

Spurious data points can be an important source of error in the least-squares estimate. Outlying data points generate large squared residuals, and the calculated line may be shifted toward the errant point(s). Draper and Smith (*2*) have suggested that data points that generate residuals greater than 4 $S_{y \cdot x}$ be omitted from the least-squares regression analysis. We will illustrate how analysis of the residuals about the regression line can provide a criterion for rejecting spurious values and eliminating their effect on the least-squares regression slopes.

Least-squares regression analysis is the appropriate technique to use in Model I regression problems—that is, cases in which the independent variable, x, is measured without error, and the dependent variable, y, is a random variable. Method-comparison studies in which the x variable is a precisely measured reference method, the result of which can be

Department of Pathology, University of California, San Diego, La Jolla, CA 92093; and Veterans Administration Hospital, 3350 La Jolla Village Drive, San Diego, CA 92161.

[1] Present address: Department of Pathology, Stanford University Medical Center, Stanford, CA 94305. Address to which reprint requests should be sent.

[2] Nonstandard abbreviations used: $a_{y \cdot x}$, y-intercept of the linear relationship between x and y, when x is the independent variable; $b_{y \cdot x}$, slope of the linear relationship between x and y when x is the independent variable; $b_{x \cdot y}$, slope of the linear relationship between x and y when y is the independent variable; $S_{y \cdot x}$, standard deviation of the residual error of regression (standard error of the estimate) when x is the independent variable (i.e., standard deviation of the differences between the actual y values and the $\hat{y}$ values predicted by the regression line); S_y, standard deviation of the y data set; S_x, standard deviation of the x data set; r, product moment correlation coefficient; S_{ex}, standard deviation of repeated measurement of a single x value; S_{ey}, standard deviation of repeated measurement of a single y value; λ, the ratio $S_{ex}{}^2/S_{ey}{}^2$.

Received March 16, 1978; accepted Dec. 27, 1978.

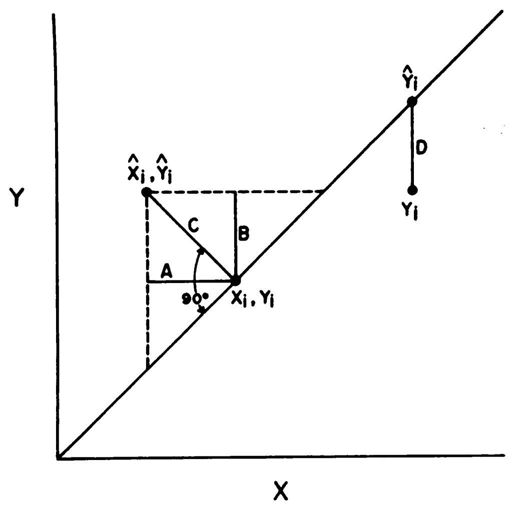

Fig. 1. Least-squares vs. Deming regression model

In the least-squares analysis, the line is chosen to minimize the residual errors in the y direction, i.e., $\Sigma_{i=1}^{N} D^2 = \Sigma_{i=1}^{N} (y_i - \hat{y}_i)^2$ for all data points is minimized. However, in the Deming regression model, the sum of the squares of both the x residual, $A^2 = (x_i - \hat{x}_i)^2$ and the y residual, $B^2 = (y_i - \hat{y}_i)^2$ is minimized. This results in choosing the line that minimizes the sum of the squares of the perpendicular distances from the data points to the line, because geometrically $C^2 = A^2 + B^2$

regarded as the "correct" value, are thus Model I regression problems. Furthermore, if the x variable can be set to preassigned values where the values recorded for it are target values (e.g., prepared concentrations of an analyte), least-squares regression can be applied (the so-called Berkson case), even though error may be present in the x-variable (*3*).

In method-comparison studies where both the x and y variable are measured with error, often there is no reason to assume that one of the two methods is the method of reference. Bias between the two methods will be indicated by the slope of the linear relationship between the absolute values measured by the two methods without error. Model II regression techniques are necessary to find the correct slope of this line. Use of the least-squares method in Model II regression cases will yield two different lines, depending on whether x or y is used as the independent variable; in fact, the line indicating the relationship between the "absolute" values of x and y lies somewhere in between.

Many statisticians have proposed solutions to Model II regression analysis. Deming (*4*) approaches the problem by minimizing the sum of the square of the residuals in both the x and y directions simultaneously. This derivation results in the best line to minimize the sum of the squares of the perpendicular distances from the data points to the line (*5*), as illustrated in Figure 1. To compute the slope by Deming's formula, one must assume gaussian error measurements of x and y with constant imprecision throughout the range of x and y values. If the ratio of the measurement errors of x and y can be estimated, the following formulas are used when x is the independent variable (*5, 6*):

$$\text{Deming } b_{y\cdot x} = U + \sqrt{U^2 + (1/\lambda)}$$

where

$$U = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2 - (1/\lambda) \sum_{i=1}^{N} (x_i - \bar{x})^2}{2 \sum (y_i - \bar{y})(x_i - \bar{x})} = \frac{S_y{}^2 - (1/\lambda) S_x{}^2}{2 r S_x S_y}$$

and

$$\lambda = \frac{S_{ex}{}^2}{S_{ey}{}^2} = \frac{\text{error variance of a single } x \text{ value}}{\text{error variance of a single } y \text{ value}}$$

As in the least-squares method, the y-intercept is calculated from the slope:

$$a_{y\cdot x} = \bar{y} - b_{y\cdot x}\bar{x}$$

The standard deviation of the residual error of regression in the y direction can be calculated and used as an indication of scatter of the points about the regression line:

$$\text{Deming } S_{y\cdot x} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \bar{y})^2 - b_{y\cdot x}(x_i - \bar{x})(y_i - \bar{y})}{N - 2}}$$

$$= \sqrt{\frac{N-1}{N-2} (S_y{}^2 - b_{y\cdot x} r \, S_x S_y)}$$

Unlike the least-squares method, Deming's method always results in one line, whether x or y is used as the independent variable.

Mandel (*6*) states that an approximate relationship between the least-squares slope and Deming's slope exists as follows:

Mandel estimate of Deming $b_{y\cdot x}$

$$= \text{least-squares } b_{y\cdot x} \left(1 + \frac{S_{ex}{}^2}{S_x{}^2 - S_{ex}{}^2}\right)$$

$$= \frac{\text{least-squares } b_{y\cdot x}}{1 - (S_{ex}{}^2/S_x{}^2)}$$

If such an estimate is valid, one may determine the need for correcting the least-squares slope a priori by noting the ratio of the variance in measuring a single x value to the variance of all the x data obtained.

Bartlett's three-group method has been suggested as a simple approach to the problem of regression when the x variable is subject to imprecision, i.e., when no knowledge of the error of measurement of x or y is required (*3*). The data are ranked by the magnitude of x and divided into thirds. The means $\bar{x}_1$ and $\bar{y}_1$ for the first third and $\bar{x}_3$ and $\bar{y}_3$ third are computed. Then:

$$\text{Bartlett } b_{y\cdot x} = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1}$$

However, Wakkers et al. (*5*) have compared predicted vs. observed regression slopes with four groups of laboratory data and concluded that Bartlett's method is not as consistent as that of Deming.

More recently, Blomqvist (*7*) has published a method of calculating the correct regression slope when the x variable is measured with error. However, his formula is applicable only when x is the initial value and y is the *change* in this initial value, and thus it cannot be used for method-comparison data.

Although these Model II regression techniques are claimed to be uninfluenced by imprecision in the measurement of the x variable, little work has been done to compare their efficacy in this regard. In this paper we use computer-simulated data and random error to study the effect of imprecision in the x variable on the least-squares $b_{y\cdot x}$, and compare the ability of the techniques of Deming, Mandel, and Bartlett to correct the resulting error. We will further investigate the effect on $b_{y\cdot x}$ estimates by these methods when proportional error (i.e., a constant coefficient of variation) exists in the measurement of both x and y.

Materials and Methods

We generated random gaussian data and calculated common statistical parameters (means, standard deviations, standard errors of estimate, correlation coefficients, regression slopes and intercepts, means of first and third groups of data, and plotting of x–y data) by use of a statistical software

Table 1. Slope of Least-Squares Line When $S_{y\cdot x}$ Is Proportional to x

	$S_{y\cdot x} = S_{ey} = 0.05\ y$[a]	$S_{y\cdot x} = S_{ey} = 0.20\ y$
Gauss (100, 25)[b]	0.901	0.904
Log gauss (100, 127)	0.893	0.896

[a] y-data measured with constant coefficient of variation.
[b] (Mean, standard deviation) of x-data, measured without error.

package called "Minitab" (*8*). One thousand data points were used for all regressions with computer-generated data to minimize differences from random error between the calculated and predicted least-squares slope. Different sets of 1000 randomly generated gaussian-distributed numbers with a mean of 0 and standard deviation of 1 were used for the x-data base, the x errors of measurement, and the y errors of measurement. Because each set was generated by the computer in a random order, all pairing of sets could be done in the order that the members of the sets were generated by the computer.

The "true" values for the x-data were generated from the x-data base as follows: gaussian x-data with a mean of 100 and standard deviations of 5 and 25 were obtained by multiplying each value by the desired standard deviation and adding the mean. A log-gaussian distribution with a mean of 100 and standard deviation of 127 was obtained by multiplying each gaussian x-data base value by 0.443 and adding 1.774 (*9*); the antilog of each number was then taken. The "true" y data were obtained from the "true" x-data by multiplying each value by 0.90 and adding 10. Thus the predicted regression equation between y and x is:

$$y = 10 + 0.90x$$

The x error base and y error base were then multiplied by the constant standard deviation desired. Standard deviations for the x error were 5, 10, and 20, and for the y error was 20 for the log-gaussian data and the gaussian data with $S_x = 25$. For the gaussian x-data with $S_x = 5$, standard deviations of the x error were 1, 2, and 4, and for the y error, 4. These random errors were added to the "true" x and y data to generate "experimental" x and y data with constant imprecision throughout the range.

Since errors of measurement are frequently not constant in the clinical laboratory, we also generated experimental data whose error of measurement was proportional to the true value (i.e., a constant coefficient of variation). The x error and y error bases were first multiplied by the desired coefficient of variation, and then by the true x or y value to which they would be added.

The regression slopes of Deming, Mandel, and Bartlett were calculated by the formulas presented earlier. The y-intercept for the line generated by any method was found by using the formula:

$$a_{y\cdot x} = \bar{y} - b_{y\cdot x}\,\bar{x}$$

When the error of measurement was proportional to the value measured, rather than constant, values for S_{ex} and S_{ey} were obtained by two methods. First, the standard deviation of all the x errors and y errors added to the true x and y values was calculated; experimentally, this could be done by measuring x and y in duplicate, where:

$$S_{\text{error}} = \sqrt{\frac{\sum_{i=1}^{N} (\text{difference between duplicates})^2}{2\,N}}$$

Second, the errors of measurement of $\bar{x}$ and $\bar{y}$ were used, i.e., the coefficient of variation times either $\bar{x}$ or $\bar{y}$.

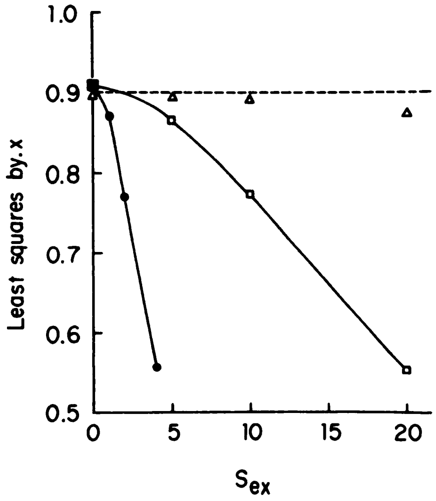

Fig. 2. Effect of increasing error of measurement of x on least-squares $b_{y\cdot x}$ when S_{ex} is constant for all x

Predicted $b_{y\cdot x}$ (- - - -); gaussian x-data, $S_x = 5$ (●); gaussian x-data, $S_x = 25$ (□); log-gaussian x-data, $S_x = 127$ (△)

Laboratory method-comparison data for sodium [continuous-flow (SMA-6) = x, manual flame photometry = y], and calcium [atomic absorption = x, continuous-flow (SMA-12) = y] were analyzed. Imprecision of measurement of x and y was estimated from repeated measurements of a control serum close to the means of the data.

Results

Characteristics of Computer-Generated Data

Means and standard deviations of the gaussian-distributed data randomly generated by the computer agree closely with the expected values. The means of the three data bases are −0.01, 0.01, and −0.01; the standard deviations are 1.03, 1.03, and 0.99. In addition, as required by the least-squares regression model, these three sets of data are independent of each other, as indicated by correlation coefficients of 0.063, 0.012, and 0.013 between the data sets.

The standard deviation of the estimate of the least-squares $b_{y\cdot x}$ is:

$$S_{b_{y\cdot x}} = \frac{S_{y\cdot x}}{\sqrt{N}\cdot S_x}$$

The largest value of $S_{y\cdot x}/S_x$ present in our experiments was 0.8, for which $S_{b_{y\cdot x}} = 0.025$. Thus, absolute differences between the observed and predicted least-squares $b_{y\cdot x}$ as great as $2\ S_{b_{y\cdot x}} = 0.05$ are significant ($p < 0.05$). It should also be noted from the above formula that low values of N in least-squares regression analysis markedly increase the uncertainty (or 95% confidence interval) of the $b_{y\cdot x}$ estimate.

$S_{y\cdot x}$ Proportional to the Value of x

If x is measured without error and y is linearly related to x, the standard error of estimate, $S_{y\cdot x}$, will be equal to S_{ey}, the standard deviation of repeated measurement of a single value of y. Thus when y is measured with a constant coefficient of variation, as is frequently the case in laboratory analysis, $S_{y\cdot x}$ will increase with increasing values of both y and x. However, as shown in Table 1, little change from the expected $b_{y\cdot x}$ of 0.90 is seen, even with a coefficient of variation of 20% in the measurement of y. Thus, although the least-squares model requires $S_{y\cdot x}$ to be constant for every value of x, the least-squares $b_{y\cdot x}$ does not appear to be greatly altered when $S_{y\cdot x}$ is proportional to x. At least at the magnitude of $S_{y\cdot x}/x$ encountered in method-comparison studies, weighted regression is not required.

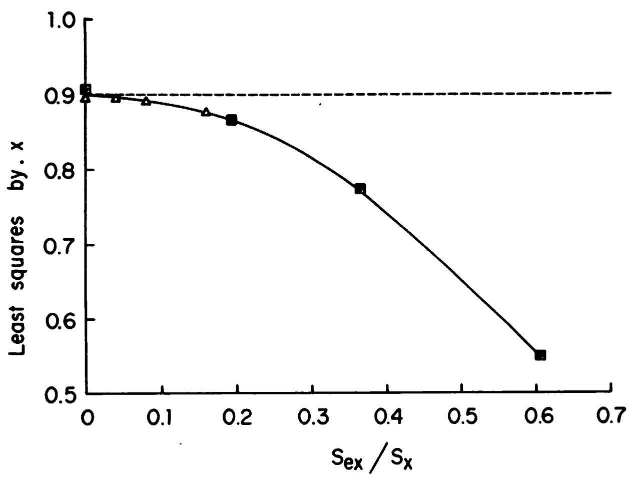

Fig. 3. The least-squares slope as a function of S_{ex}/S_x, when S_{ex} is constant for all x

Predicted $b_{y\cdot x}$ (- - - -); gaussian x-data, $S_x = 5$ (●); gaussian x-data, $S_x = 25$ (□); log-gaussian x-data, $S_x = 127$ (△)

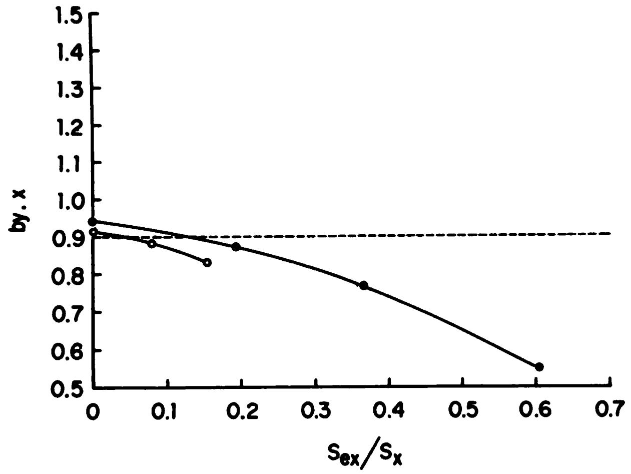

Fig. 4. Calculation of $b_{y\cdot x}$ by the method of Bartlett

Bartlett's method with gaussian x-data ($S_x = 5$ gave identical points to $S_x = 25$) (●); Bartlett's method with log-gaussian data (○)

Increasing Imprecision of *x* Measurement When Imprecision is Constant for All *x* Values

Figure 2 shows the effect of increasing the standard error of measurement of x on the least-squares regression slope for gaussian and log-gaussian x-data. The least-squares $b_{y\cdot x}$ decreases steadily from the predicted value of 0.9 with increasing S_{ex}. Because this decrease is less prominent as S_x of the "true" x-data increases, least-squares slope could perhaps be expressed as a function of the S_{ex}/S_x regardless of the distribution or dispersion of the "true" x-data. This postulate is borne out in Figure 3, where a plot of the least-squares $b_{y\cdot x}$ vs. S_{ex}/S_x yields a single curve for the combined gaussian and log-gaussian data. From this graph, significant underestimation (>0.05 from the predicted $b_{y\cdot x}$) of the absolute value of the true slope of a regression line may occur with the least-squares method when S_{ex}/S_x exceeds 0.2.

An attempt to obtain the correct slope by the method of Bartlett is presented in Figure 4. The Bartlett $b_{y\cdot x}$ decreases with increasing S_{ex}/S_x to the same extent as the least-squares slope. The relationship between S_{ex}/S_x and the Bartlett $b_{y\cdot x}$ also depends on the distribution of the "true" x-data, giving dissimilar curves for log-gaussian and gaussian data. These results suggest that Bartlett's method should not be used in estimating the slope of a line when x is measured with error.

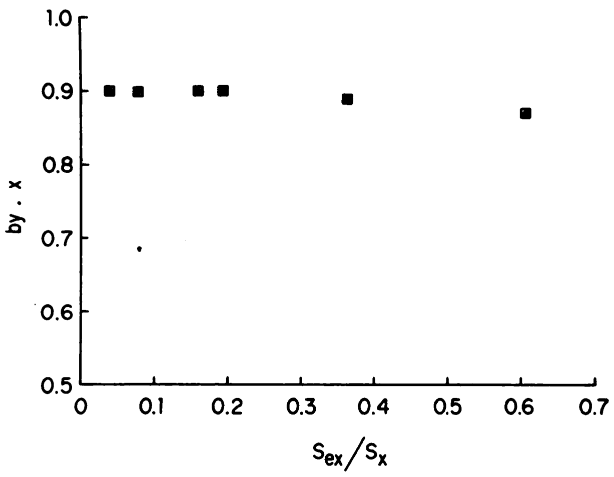

Fig. 5. Calculation of $b_{y\cdot x}$ by the methods of Deming and Mandel

The first three points to the left are from log-gaussian x-data regressions, while the three points at far right are from gaussian x-data (points with $S_x = 5$ identical to points with $S_x = 25$) regressions. Deming $b_{y\cdot x}$ (●); Mandel $b_{y\cdot x}$ (□)

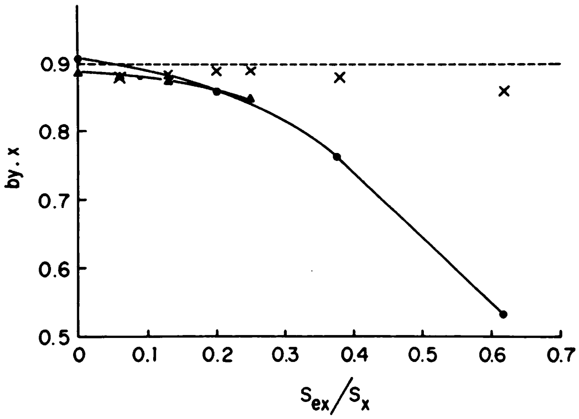

Fig. 6. Least-squares, Deming, and Mandel $b_{y\cdot x}$ as a function of S_{ex}/S_x when S_{ex} is proportional to the value of x

Predicted $b_{y\cdot x}$ (- - - -); least-squares $b_{y\cdot x}$ for gaussian x-data (either $S_x = 5$ or $S_x = 25$) (●); least-squares $b_{y\cdot x}$ for log-gaussian x-data (▲); Deming or Mandel $b_{y\cdot x}$ for the corresponding least squares $b_{y\cdot x}$ vertically below it (X)

Calculation of the regression line slope by the methods of Deming and Mandel gave identical results, as shown in Figure 5. The slopes remain close to the predicted value of 0.9 for both log-gaussian and gaussian data, even at large values of S_{ex}/S_x.

Consistency of the calculated regression coefficient can be assessed by the ability of regression with either x or y as the independent variable to produce one line.

If $x = a_{x\cdot y} + b_{x\cdot y}\,y$
then $y = -(a_{x\cdot y}/b_{x\cdot y}) + (1/b_{x\cdot y})\,x$
Since $y = a_{y\cdot x} + b_{y\cdot x}\,x$
then $b_{y\cdot x} = 1/b_{x\cdot y}$

if the reverse regression gives the same line. Results of performing the regression with y as the independent variable are shown in Table 2. While least-squares regression gives two markedly different lines at large errors of measurement, the methods of Deming and Mandel are consistent in producing one regression line.

Increasing Imprecision of *x* Measurement When Imprecision Is Proportional to Each *x* Value

The least-squares $b_{y\cdot x}$ is plotted as a function of S_{ex}/S_x in Figure 6. The calculated standard deviation of all the computer-generated proportional errors of measurement of x and

Table 2. Consistency of Least-Squares, Deming, and Mandel Regression Coefficients When S_{ex} and S_{ey} Are Constant[a]

	Gauss (100,5)[b]		Gauss (100,25)		Log gauss (100,127)	
	$b_{y \cdot x}$	$1/b_{x \cdot y}$	$b_{y \cdot x}$	$1/b_{x \cdot y}$	$b_{y \cdot x}$	$1/b_{x \cdot y}$
Least squares	0.55	1.48	0.55	1.48	0.88	0.92
Deming	0.87	0.87	0.87	0.87	0.90	0.90
Mandel	0.87	0.86	0.87	0.86	0.90	0.90

[a] For gauss (100,5), $S_{ex} = S_{ey} = 4$; for gauss (100,25) and log gauss (100, 127), $S_{ex} = S_{ey} = 20$.

[b] (Mean, standard deviation) of the "true" x-data.

Table 3. Consistency of Least-Squares, Deming, and Mandel Regression Coefficients When S_{ex} and S_{ey} Are Proportional to x and y[a]

	Gauss (100,5)[b]		Gauss (100,25)		Log gauss (100, 127)	
	$b_{y \cdot x}$	$1/b_{x \cdot y}$	$b_{y \cdot x}$	$1/b_{x \cdot y}$	$b_{y \cdot x}$	$1/b_{x \cdot y}$
Least squares	0.55	1.48	0.54	1.51	0.85	0.95
Deming	0.87	0.87	0.86	0.86	0.90	0.90
Mandel	0.87	0.86	0.87	0.85	0.90	0.89

[a] For gauss (100,5), a coefficient of variation of 4% was used to compute x and y errors; for gauss (100,25) and log gauss (100, 127), a coefficient of variation of 20% was used to compute x and y errors. The average S_{ex} and S_{ey} were used to calculate Deming and Mandel $b_{y \cdot x}$.

[b] (Mean, standard deviation) of "true" x-data.

y is used in computing this ratio, equivalent to an "average" standard deviation of measurement of x and y that would be calculated from measuring x and y in duplicate. For the gaussian x-data, the curve generated is identical to that when S_{ex} is constant (Figure 3). For the log-gaussian data, a different relationship is evident, although the least-squares $b_{y \cdot x}$ is not markedly different from that for gaussian data at the same S_{ex}/S_x. However, the methods of both Deming and Mandel give identical values of $b_{y \cdot x}$ close to 0.9 for all x-data at all S_{ex}/S_x.

Because it may not always be practical to perform measurements of x and y in duplicate to obtain the average S_{ex} and S_{ey} when S_{ex} and S_{ey} are not constant, we investigated the validity of using the standard deviation of repeated measurements of x and y at the mean of the data, i.e., the coefficient of variation for x and y times $\bar{x}$ and $\bar{y}$. Consistently lower values of S_{ex}/S_x are obtained by this calculation, particularly for the log-gaussian data. Thus the use of S_{ex} about $\bar{x}$ may give values for the Mandel $b_{y \cdot x}$ that are too low. On the other hand, the ratio of S_{ex}/S_{ey} does not change greatly when the measurement errors at the means are used; thus the Deming $b_{y \cdot x}$ can be calculated if repeated measurements of a control or patient sample are made of a specimen close to the value of $\bar{x}$ and $\bar{y}$.

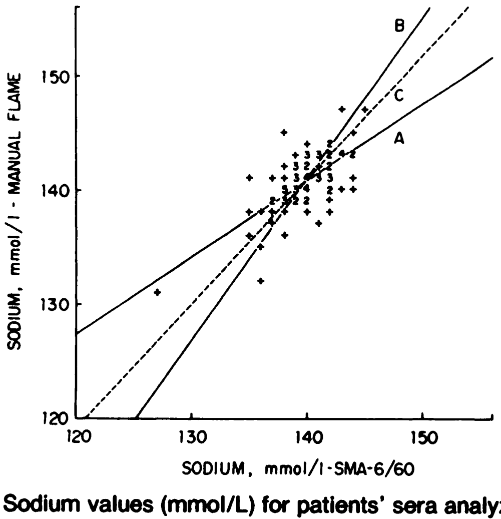

Fig. 7. Sodium values (mmol/L) for patients' sera analyzed by two flame-photometric methods

N = 87. Numerals substitute for points where more than one data point occupies a single space. *A*. Least-squares regression with x as the independent variable, $y = 47.6 + 0.667x$. *B*. Least-squares regression with y as the independent variable, $y = -59.2 + 143x$. *C*. Deming regression, using either x or y as the independent variable, $y = -11.7 + 1.09x$

Consistency of the calculated slopes at large error measurements is shown in Table 3. As in the case when the error of measurement is constant, the reverse regression yields the same line for the Deming and Mandel methods, while different lines are obtained with the least-squares method.

Regression Analysis of Laboratory Method-Comparison Data

Data with small S_x. Regression analysis of clustered method comparison data is a well-known laboratory nemesis. Figure 7 shows a plot of 87 sodium determinations by two flame-photometric methods. The "reference" method (x-axis) data has a mean of 139.8 and standard deviation of 2.67, while the "test" method (y-axis) data has a mean of 140.7 and standard deviation of 2.60. Although a line with a slope close to 1 is expected, least-squares regression gives a $b_{y \cdot x}$ of 0.66; when regression is performed with y as the independent variable, a markedly different value of $b_{y \cdot x} = 1/b_{x \cdot y} = 1.43$ results. Precision estimates from quality-control samples near the mean give approximate measurement errors of 1.09 and 0.83 for x and y, respectively. Thus $S_{ex}/S_x = 0.41$, suggesting necessity for Deming's method. As seen in Figure 7, Deming's calculation gives one line with $b_{y \cdot x} = 1.09$, regardless of whether x or y is used as the independent variable. This line is a better estimate of the relationship between x and y suggested by the data.

Data with outliers. Figure 8 shows a plot of 169 samples assayed for calcium by atomic absorption (x-axis) and SMA 12–60 (y-axis). For these data, $\bar{x} = 9.58$, $\bar{y} = 9.41$, $S_x = 0.755$, $S_y = 0.741$, $S_{ex} = 0.12$, and $S_{ey} = 0.14$. Since $S_{ex}/S_x = 0.16$, imprecision in the measurement of the x values does not greatly bias the least-squares slope estimate. Yet the least-squares regression equation is:

$$y = 1.96 \pm 0.78x$$

giving a $b_{y \cdot x}$ value substantially lower than the expected slope of 1.0. Inspection of the data in Figure 8 suggests one or more

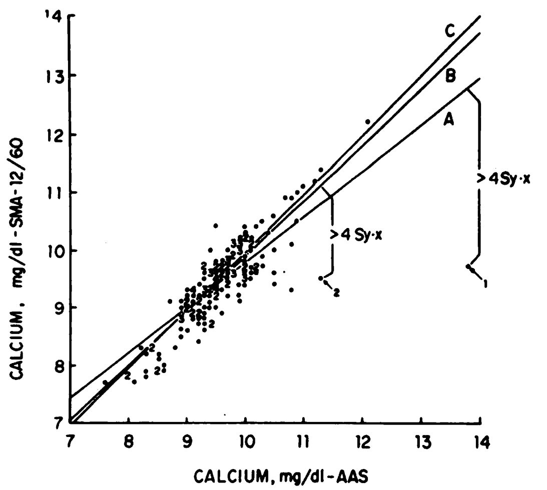

Fig. 8. Calcium values (mg/dL) for patients' sera analyzed by atomic absorption (x) and continuous-flow (y)

N = 169. Numerals substitute for points where more than one data point occupies a single space. *A*. Least-squares regression line with all data points, showing point no. 1 greater than 4 $S_{y\cdot x}$ from the line, $y = 1.96 + 0.78x$. *B*. Least-squares regression line omitting point no. 1, showing point no. 2 greater than 4 $S_{y\cdot x}$ from the line, $y = 0.39 + 0.95x$; *C*. Least-squares regression line omitting points 1 and 2, $y = 0.04 + 0.99x$

spurious data points, the most aberrant being at $x = 13.8$, $y = 9.7$. The predicted value of y by the regression equation at $x = 13.8$ is 12.8, differing from the observed value of 9.7 by 3.1. This difference exceeds four times the standard error of estimate ($S_{y\cdot x} = 0.49$, $4\ S_{y\cdot x} = 1.95$), and thus this point should be rejected from the data set. Recomputation of the regression equation yields:

$$y = 0.39 + 0.95x$$

with $S_{y\cdot x} = 0.37$. Although outlier point no. 2 (Figure 8) is not greater than 4 $S_{y\cdot x}$ in the y-direction from the initial regression line, it can be shown to be greater than 4 $S_{y\cdot x}$ in the vertical direction from the new regression line with point no. 1 omitted. Recalculation with points 1 and 2 deleted gives the least-squares regression line:

$$y = 0.04 + 0.99x$$

When this new line is used, no further data points have y values deviating by more than 4 $S_{y\cdot x}$ from the y value predicted by the regression line. Thus, even though 169 data points were used, two spurious values can still significantly change the least-squares regression slope. These values may easily be identified and omitted by noting that their residual error (observed y value − predicted $\hat{y}$ value) is greater than 4 $S_{y\cdot x}$.

Discussion

Least-squares (Model I) regression analysis may be the inappropriate regression technique to use when x is measured with imprecision. We have compared the Model II regression solutions of Deming, Mandel, and Bartlett for estimating $b_{y\cdot x}$ when x is measured with error, and find that only the methods of Deming and Mandel compute the correct slope.

The methods of both Deming and Mandel assume that the error of measurement remains constant throughout the range of values. However, clinical laboratory measurements usually increase in absolute imprecision when larger values are measured. We have approximated this situation by using a model in which the error of measurement has a constant coefficient of variation. If the "average" error is calculated by the method of duplicates, both Deming and Mandel methods will yield $b_{y\cdot x}$ close to the expected value; however, if precision is estimated by repeatedly assaying a sample close to the mean of the data, Deming's method gives better results.

On the basis of the results in this paper, the following guidelines for linear regression analysis are suggested:

- Always plot the data; perform and apply least-squares regression analysis only to the region of linearity. Although not stressed in this paper, curvilinear deviation will markedly alter the regression slope. Furthermore, suspected outliers may be identified from the data plot.
- A rough estimate of the effect of measurement errors in x can be made by looking at the ratio S_{ex}/S_x, where S_{ex} represents the precision of a single x measurement near $\bar{x}$. If this ratio exceeds 0.2, significant error occurs in the least-squares estimate of slope, and Deming's $b_{y\cdot x}$ should be computed. If the data are markedly skewed (e.g., $S_x > \bar{x}$), and the error of measurement of x is proportional to x, a ratio of 0.15 or greater may be indicative of significant error in the least-squares slope. Alternatively, more data can be selected that will increase the value of S_x, or the error of measurement of the sample by method x may be decreased by averaging N measurements, where:

$$S_{\text{error of average}} = \frac{S_{\text{error of single measurement}}}{\sqrt{N}}$$

- Calculation of the Deming $b_{y\cdot x}$ requires the ratio, S_{ex}/S_{ey}. Estimates may be obtained by precision analysis of a single sample close to the mean of the data, or, alternatively, by measuring duplicate x and y values, where:

$$S_{\text{error}} = \sqrt{\frac{\sum_{i=1}^{N} (\text{difference between duplicates})^2}{2\,N}}$$

If the average of the duplicates is used for the x and y values in the regression analysis, then it must be remembered that the standard deviation of measurement of an average of two values equals the standard deviation of measurement of a single value divided by $\sqrt{2}$.

- The standard error of regression should always be calculated. For either least-squares or Deming $b_{y\cdot x}$, this statistic may be easily computed from parameters likely to be obtained from a calculator intended for scientific use:

$$S_{y\cdot x} = \sqrt{\frac{N-1}{N-2}(S_y{}^2 - b_{y\cdot x} r S_x\ S_y)}$$

and can be interpreted as the standard deviation of the mean value expected for y for a given value of x close to $\bar{x}$. It is a measure of scatter of the points about the regression line. Although more complex and statistically exact methods are available (*10*), approximate detection of significant outliers that may bias the least-squares slope may be made by excluding any data point whose y value differs from that predicted by the regression line by more than $S_{y\cdot x}$. However, a large number of data points should not be excluded.

When measurements of x and y are both subject to error, as they are in the clinical laboratory, the least-squares regression method may give two very disparate lines, depending on whether x or y is used as the independent variable. Neither line expresses the functional relationship between the true values of x and y; both are altered by errors of measurement in the independent variant. The method of Deming can provide a solution to this dilemma, yielding one regression line between x and y that takes into account the errors of measurement of both variables.

We thank Drs. John Brimm, Lemuel Bowie, and Rupert Miller for assistance with this manuscript.

References

1. Westgard, J. O., and Hunt, M. R., Use and interpretation of common statistical tests in method-comparison studies. *Clin. Chem.* **19**, 49 (1973).

2. Draper, N. R., and Smith, H., *Applied Regression Analysis.* John Wiley and Sons, New York, NY, 1966, pp 44–103.

3. Sokal, R. R., and Rohlf, F. J., *Biometry.* W. H. Freeman and Co., San Francisco, CA, 1969, pp 481–486.

4. Deming, W. E., *Statistical Adjustment of Data.* John Wiley and Sons, New York, NY, 1943, p 184.

5. Wakkers, P. J. M., Hellendoorn, H. B. Z., Op De Weegh, G. J., and Herspink, W., Applications of statistics in clinical chemistry. A critical evaluation of regression lines. *Clin. Chim. Acta* **64,** 173 (1975).

6. Mandel, J., *The Statistical Analysis of Experimental Data.* John Wiley and Sons, New York, NY, 1964, pp 290–291.

7. Blomqvist, N., Cederblad, G., Korsan-Bengtsen, K., and Wallerstedt, S., Application of a method for correcting an observed regression between change and initial value for the bias caused by random errors in the initial value. *Clin. Chem.* **23,** 1845 (1977).

8. Ryan, T. A., Joiner, B. L., and Ryan, B. F., *Minitab Student Handbook.* Duxbury Press, North Scituate, MA, 1976.

9. Diem, K., and Lentner, C., Eds., *Documenta Geigy.* Ciba Geigy Limited, Basle, Switzerland, 1970, p 164.

10. Snedecor, G. W., and Cochran, W. G., *Statistical Methods.* Iowa State University Press, Ames, IA, 1967, p 157.