Clinical Chemistry 46:1
100–104 (2000)

General Deming Regression for Estimating Systematic Bias and Its Confidence Interval in Method-Comparison Studies

Robert F. Martin

Background: **Various forms of least-squares regression analyses are used to estimate average systematic error (bias) and its confidence interval in method-comparison studies. When assumptions that underlie a particular regression method are inappropriate for the data, errors in estimated statistics result. In this report, I present an improved method for regression analysis that is free from the usual simplifying assumptions and is generally applicable to linearly related method-comparison data.**

Methods: **Theoretical equations based on the Deming approach, further developed by physicists and extended herein, were applied to method-comparison data analysis. Monte Carlo simulations were used to demonstrate the validity of the new procedure and to compare its performance to ordinary linear regression (OLR) and simple Deming regression (SDR) procedures.**

Results: **Simulation studies included three types of data commonly encountered in method-comparison studies: (*a*) constant within-method SDs for both methods, (*b*) constant within-method CVs for both methods, and (*c*) neither SDs nor CVs constant for both methods. For all cases examined, OLR produced unreliable confidence intervals of the estimated bias. However, OLR point estimates of systematic bias were reliable when the correlation coefficient was >0.975. SDR produced reliable estimates of systematic bias for all cases studied, but the confidence intervals of systematic bias were unreliable when SDs of methods varied as a function of analyte concentration.**

Conclusion: **Only iteratively reweighted general Deming regression produced statistically unbiased estimates of systematic bias and reliable confidence intervals of bias for all cases.**

The objective of method-comparison studies for quantitative assays in laboratory medicine is to estimate average systematic bias and the confidence interval (CI)[1] for estimated bias at medical decision levels. These estimates are then compared with a manufacturer's claims or internal criteria to judge acceptability of the method under evaluation. When the test and comparative methods are linearly related in accordance with the linear statistical model ($y = a + bx$), linear regression analysis is commonly used to estimate the average bias and its CI. The n test method results (y_i) and corresponding comparative method results (x_i) throughout the data range are used to estimate parameters of the model (*a*, the intercept; and *b*, the slope) and their standard errors, SE(*a*) and SE(*b*). The estimated bias ($\hat{B}_C$) at medical decision level, X_C, is given by:

$$\hat{B}_C = a + X_C(b - 1)$$

The CI of the estimated bias is given by:

$$\text{CI} = \hat{B}_C \pm t_{(1 - \alpha/2;n - 2)}\sqrt{Var(\hat{B}_C)}$$

where $t_{(1 - \alpha/2;\text{n} - 2)}$ is the Student *t*-statistic at the desired confidence level ($1 - \alpha$) with ($\text{n} - 2$) degrees of freedom. The variance of the bias estimate is given by:

$$Var(\hat{B}_C) = Var(a) + X_C \cdot Var(b) \cdot (X_C - 2\bar{x}_w)$$

where $Var(a)$ and $Var(b)$ are the variances of the estimated intercept and slope, respectively. Note that $Var(a) = [\text{SE}(a)]^2$ and $Var(b) = [\text{SE}(b)]^2$. The calculation of $\bar{x}_w$, the weighted mean of the comparative method values, is described below.

The reliability of the estimated bias and its CI depend on the appropriateness of the regression procedure for analysis of the particular set of experimental data. In

MarChem Associates, Inc., 325 College Rd., Concord, MA 01742. Fax 978-371-9055; e-mail bobmartin@marchem.com.

Received in revised form October 18, 1999; accepted October 18, 1999.

[1] Nonstandard abbreviations: CI, confidence interval; OLR, ordinary linear regression; SDR, simple Deming regression; and IRGDR, iteratively reweighted general Deming regression.

current practice, regression procedures are selected based on what assumptions can justifiably be made about the data. For example, in ordinary linear regression (OLR), the most commonly used regression procedure for method-comparison calculations, it is assumed that comparative method values are without random error and that test method random error is constant throughout the range of the data. Although these assumptions are never strictly justified, results of OLR are of acceptable accuracy and precision when the random error of the comparative method is small compared with the range of the data and when the test method data are not "significantly" heteroscedastic. When OLR cannot be used because of substantial violations of its assumptions, an appropriate form of Deming regression may be selected.

Deming regression is the term used in laboratory medicine to refer to linear regression analysis in which the random error of both the comparative and test methods is taken into account. Although Deming's approach to a generalized regression procedure was basically sound, he oversimplified the problem by expanding the straight line function in a Taylor series about assumed values of slope, intercept, and adjusted points. Because squared and higher terms in the expansion were neglected, Deming's original general exposition can lead to significant errors in some instances, as he recognized.

Nevertheless, Deming presented an exact solution for the particular case in which both x and y are subject to random error but in such a way that the ratio $\lambda = Var(x)/Var(y)$ is constant and not infinite or zero throughout the data range *(1)*. With this constraint, he derived equations for the slope and intercept for a weighted least-squares regression model. When the variance of x is constant throughout the data range, the variance of y must be constant, and the equations for the Deming slope and intercept reduce to the well-known formulae for simple Deming regression (SDR) called to our collective attention by Cornbleet and Gochman *(2)*. More recently, Linnet *(3)* independently rederived the cited formulae for the Deming slope and intercept and focused on a special case in which random errors in both x and y are proportional to the overall average value of the test and comparative results for each sample. For convenience, we refer to the Linnet procedure as "constant CV Deming regression", although it is not precisely so. When assumptions about constant SDs or proportional SDs do not apply to the data, CIs of the bias are unreliable for the cited Deming methods.

We present here a generally applicable statistical procedure for Deming regression without constraints on random error of the test or comparative method. We then use Monte Carlo simulations to demonstrate the validity of the new procedure and to compare its performance to other regression procedures frequently used in current practice.

Materials and Methods

THEORY

Guided by Deming's basic concepts, York *(4)* developed the foundations of an exact general treatment of the problem. York's procedure, which contained errors in equations for SEs of slope and intercept, was used for certain method-comparison calculations by Gerbet et al. *(5)*. Williamson *(6)* and, later, Reed *(7, 8)* corrected and refined York's work to derive equations for the linear regression parameters and their SEs. To our knowledge, these corrected results have not been used for method-comparison calculations; therefore, the complete equations are reproduced here. As presented by Williamson *(6)*, the slope and intercept are given by:

$$b_D = \frac{\sum_{i=1}^{n} w_i z_i y'_i}{\sum_{i=1}^{n} w_i z_i x'_i}, \qquad a_D = \bar{y}_w - b_D \bar{x}_w$$

where:

$$u_i = Var(x_i),\ v_i = Var(y_i),\ w_i = [v_i + (b_D)^2 u_i]^{-1}$$

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},\ \bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i},\ x'_i = x_i - \bar{x}_w,\ y'_i = y_i - \bar{y}_w$$

$$z_i = w_i(v_i x'_i + b_D u_i y'_i)$$

Because z_i, w_i, $\bar{x}_w$, and $\bar{y}_w$ are functions of b_D, an iterative calculation procedure is required.

Unbiased estimates of a_D and b_D are obtained with these equations when the true weights of the observed points (x_i, y_i) are known. In method-comparison work where weights are typically some function of concentration, weights corresponding to observed points are not optimal because the observed points are subject to random error of the method. We therefore extend the procedure described above by estimating improved weights based on the adjusted values ($\hat{X}_i$, $\hat{Y}_i$), which are those points through which the least-squares line is drawn and which represent our best estimates of the true values (X_i, Y_i). Linnet *(3)* used a similar approach for weighting observed points in his development of constant CV Deming regression. The relationships between observed and adjusted points were given by York *(4)*:

$$\hat{X}_i = x_i - \delta_i u_i b_D$$

$$\hat{Y}_i = y_i - \delta_i v_i$$

$$\delta_i = w_i(a_D + b_D x_i - y_i)$$

Weights for each observed point are calculated iteratively. After an initial estimate of a_D and b_D based on weights derived from observed points, revised weights are computed using adjusted points, which in turn are used to calculate new values for a_D and b_D. The process is repeated using updated estimates of adjusted values and weights until the correction to b_D is <0.0001. In our experience, four or fewer iterations are required for convergence, even for extremely imprecise methods. We refer to this overall procedure for obtaining the unbiased slope and intercept as iteratively reweighted general Deming regression (IRGDR).

Williamson *(6)* derived the variances of the estimated slope and intercept from first order derivatives of b_D and a_D with respect to the observed points (x_i and y_i):

$$Var(b_D) = Q^2 \sum_{i=1}^{n} w_i^2[(x_i')^2 v_i + (y_i')^2 u_i]$$

$$Var(a_D) = \left(\sum_{i=1}^{n} w_i \right)^{-1} + 2(\bar{x}_w + 2\bar{z}_w Q)\bar{z}_w Q$$

$$+ (\bar{x}_w + 2\bar{z}_w)^2 Var(b_D)$$

where:

$$Q^{-1} = \sum_{i=1}^{n} w_i[x_i' y_i'/b_D + 4z_i'(z_i - x_i')]$$

$$z_i' = z_i - \bar{z}_w, \; \bar{z}_w = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i}$$

The derivatives used to calculate $Var(b_D)$ and $Var(a_D)$ may also be evaluated at adjusted points. For well-correlated data typically encountered in method-comparison studies, the difference in variances estimated by the two procedures is small, with values based on observed points being slightly larger.

SIMULATION STUDIES

We compared the performance characteristics of the IRGDR procedures to those of OLR and SDR using Monte Carlo simulations. For each simulation run, we set the true slope at 1.0, the true intercept at 0.0, and n = 50 samples with duplicate values for test and comparative methods at each point. The random error for each simulated result had a gaussian distribution. Regression calculations were performed by each procedure using only the first replicate of each analytical method to estimate the average bias and 95% CI of the bias at medical decision levels. For SDR calculations, duplicates of the test and comparative method results for each sample were used to estimate SDs, and SEs of the SDR slope and intercept were computed using the general Deming regression relationships with the constant SDs. Computations were performed with an adaptation of a Windows® application developed by the author (EP_Suite 9A, a module in EP_Suite for Windows^{TM}) that contains components for each of the regression procedures.

Results

The results of three representative simulations are summarized in Table 1: Table 1A presents a sodium evaluation with data in the range 132–155 mmol/L and constant SDs; Table 1B presents an albumin evaluation with data in the range 15–50 g/L and constant CVs; and Table 1C presents a glucose evaluation with data in the range 2.2–27.8 mmol/L (40–500 mg/dL) with neither SDs nor CVs held constant. In the last case, SDs varied linearly from 0.055 at 2.2 mmol/L to 0.166 at 27.8 mmol/L for the comparative method, whereas test method SDs ranged linearly from 0.111 to 0.555 mmol/L over the same concentration interval.

For each of 5000 simulation runs for each case, the slope, intercept, their respective SEs (based on observed

Table 1. Performance characteristics of three regression procedures for simulated method comparisons.[a]

Statistic	OLR	SDR	IRGDR
A. Sodium simulation (constant SDs): SD(*x*) = 1.0; SD(*y*) = 2.0			
Average slope	0.9782[b]	1.0005	1.0003
SD of slopes	0.0483	0.0505	0.0499
$\overline{SE}(b)$	0.0486	0.0501	0.0501
Average intercept	3.1343[c]	−0.0787	−0.0464
SD of intercepts	6.9612	7.2634	7.1759
$\overline{SE}(a)$	6.9900	7.2073	7.2107
$\hat{\alpha}$ (at 130 mmol/L)	0.069	0.052	0.050
$\hat{\alpha}$ (at 150 mmol/L)	0.063	0.052	0.048
B. Albumin simulation (constant CVs): CV(*x*) = 2.5%; CV(*y*) = 5.0%			
Average slope	0.9947[b]	1.0002	1.0003
SD of slopes	0.0246	0.0254	0.0216
$\overline{SE}(b)$	0.0232	0.0234	0.0214
Average intercept	0.1572[c]	−0.0007	−0.0095
SD of intercepts	0.6356	0.6546	0.5307
$\overline{SE}(a)$	0.7586	0.7648	0.5262
$\hat{\alpha}$ (at 20 g/L)	0.004	0.003	0.049
$\hat{\alpha}$ (at 35 g/L)	0.102	0.106	0.049
C. Glucose simulation (variable SDs and CVs): SD(*x*) = 0.06–0.17; SD(*y*) = 0.11–0.55			
Average slope	0.9998	1.0001	1.0000
SD of slopes	0.0075	0.0076	0.0064
$\overline{SE}(b)$	0.0057	0.0057	0.0064
Average intercept	0.0015	−0.0016	−0.0003
SD of intercepts	0.0637	0.0651	0.0500
$\overline{SE}(a)$	0.0801	0.0803	0.0494
$\hat{\alpha}$ (at 2.78 mmol/L)	0.007	0.006	0.053
$\hat{\alpha}$ (at 6.99 mmol/L)	0.003	0.003	0.054

[a] 5000 simulation runs for each case.

[b] Significantly different from 1.0 (*P* <0.001).

[c] Significantly different from 0.0 (*P* <0.001).

values), and the 95% CI of the systematic bias were computed. The means and SDs of the 5000 slopes (and intercepts) are listed as the "average slope" (intercept) and SD of slopes (intercepts). The root-mean-square of the 5000 computed SEs [$\overline{\mathrm{SE}}(a)$ and $\overline{\mathrm{SE}}(b)$] are tabulated in their respective rows. The proportion of calculated CIs of systematic bias that did not include the true value of bias (0.0) at each medical decision level is represented by $\hat{\alpha}$. Thus $\hat{\alpha}$ is the confidence coefficient estimated from the simulation runs.

An adequate regression procedure must provide (*a*) statistically unbiased estimates of slope and intercept to compute unbiased point estimates of systematic bias at each medical decision level, and (*b*) an estimated confidence coefficient ($\hat{\alpha}$) equal to the preestablished value (0.05 in our study), thus indicating reliable CIs of the bias at each level. Review of the data presented in Table 1 leads to the following conclusions regarding the adequacy of the different regression procedures for the various cases:

1. Only IRGDR produced unbiased estimates of systematic bias and reliable CIs of bias for all cases.
2. SDR produced unbiased estimates of systematic bias in all cases, but its CIs of the bias were unreliable when data were heteroscedastic (cases B and C). (Note: Results shown in Table 1 for SDR are based on method SDs calculated from method duplicates. When method variances are constant, as in Table 1A, equivalent results are obtained whether true or estimated SDs are used for the calculations.)
3. OLR led to statistically biased estimates of systematic bias and marginally reliable CIs of the bias for the constant SD simulation (case A).
4. When random error of the comparative method (*x*) is small compared with the analytical range of the data (cases B and C), OLR yielded unbiased (or nearly so) estimates of systematic bias, but CIs of bias were unreliable.

Relative to point 4, we evaluated the use of the correlation coefficient (*r*) as a criterion for assessing whether the range of data is adequate for use of the OLR procedure. Among others, Westgard *(9)* and NCCLS *(10)* have indicated that if the correlation coefficient is ≥0.975, OLR may be used to estimate systematic bias. In our studies, *r* was <0.975 in 99.8% of simulation runs for case A, correctly indicating that OLR should not be used. For cases B and C, *r* was ≥0.975 in 99.8% and 100.0% of cases, respectively, indicating that OLR should be adequate under these conditions. Thus, our data support the usual correlation coefficient criteria for using OLR to estimate average systematic error. However, when the range of data is very broad, heteroscedasticity is likely and the CI of the bias based on OLR should be considered suspect, as revealed in cases B and C.

Discussion

We conclude from these simulations that IRGDR yields unbiased estimates of the regression parameters and their SEs without constraints on random errors of test or comparative methods. The procedure will therefore be generally useful in estimating systematic difference (bias) and its CI in method-comparison studies whenever a linear relationship exists between the test and comparative methods. The main advantage of the new IRGDR technique over SDR is the reliability of the CI for test method bias. With reliable CIs for bias, we can now depend on conclusions regarding the probability of acceptability of test method bias.

From a practical point of view, the improved reliability of general Deming regression comes at the cost of knowing (or determining) imprecision profiles for both test and comparative methods. In SDR, the constant imprecisions calculated from sample duplicates or from external imprecision studies serve the purpose well. When imprecision of one or both methods varies across the concentration interval, the task is less straightforward. Although several procedures exist for determining weighting functions directly from the method-comparison data *(11, 12)*, such procedures may be risky in view of the not uncommon paucity of imprecision information at the extremes of the data. Furthermore, when weights are estimated directly from the data, calculated results are somewhat less reliable because estimating weights introduces another source of variability.

We have, therefore, preferred to use imprecision profiles from data external to the method-comparison experiment (e.g., imprecision studies, reportable range studies, quality control, or manufacturer's documentation). There are two primary requirements on such external imprecision data: (*a*) data must accurately reflect imprecision on authentic patient samples, and (*b*) the imprecision profile must span the entire range of data collected in the method-comparison experiment. With imprecisions on three to seven levels, we then use cubic spline calculations to create a continuous imprecision profile for each method.

As with any weighted regression procedure, performance depends on the adequacy of the imprecision profiles. To validate the IRGDR procedure, the simulation studies presented here used known (true) imprecision profiles, which in practice are rarely available. Thus, the simulation results presented may give a somewhat optimistic impression of performance of the method.

One of our Windows programs that performs IRGDR calculations is available as a data supplement from the *Clinical Chemistry* Web site. The file can be accessed by a link from the on-line Table of Contents (http://www.clinchem.org/content/vol46/issue1/). Executing the downloaded file named "Deming.exe" will create a Windows program group that includes the statistical program (GDR), instruction manuals (in Adobe® Acrobat® pdf format), and a Readme text file that contains important information about the system.

References

1. Deming WE. Statistical adjustment of data. New York: John Wiley & Sons, 1943. Reprinted: New York: Dover Publications, Inc., 1964:184.
2. Cornbleet PJ, Gochman N. Incorrect least-squares regression coefficients in method-comparison analysis. Clin Chem 1979;25: 432–8.
3. Linnet K. Estimation of the linear relationship between the measurements of two methods with proportional errors. Stat Med 1990;9:1463–73.
4. York D. Least-squares fitting of a straight line. Can J Phys 1966;44:1079–86.
5. Gerbet D, Richardot P, Auget J, Maccario J, Cazalet C, Raichvarg D, et al. New statistical approach in biochemical method-comparison studies using Westlake's procedure, and its application to continuous-flow, centrifugal analysis, and multilayer film analysis techniques. Clin Chem 1983;29:1131–6.
6. Williamson JH. Least-squares fitting of a straight line. Can J Phys 1968;46:1845–7.
7. Reed BC. Linear least-squares fits with errors in both coordinates. Am J Phys 1989;57:642–6.
8. Reed BC. Linear least-squares fits with errors in both coordinates. II. Comments on parameter variances. Am J Phys 1992;60:59–62.
9. Westgard JO. Points of care in using statistics in method comparison studies [Editorial]. Clin Chem 1998;44:2240–2.
10. National Committee for Clinical Laboratory Standards. Method comparison and bias estimation using patient samples; approved guideline. NCCLS document EP9-A. Wayne, PA: NCCLS, 1995: 36pp.
11. Carroll RJ, Ruppert D. Transformation and weighting in regression. New York: Chapman Hall, 1988:9–114.
12. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied linear statistical models, 4th ed. Chicago: Richard D. Irwin, 1996: 400–9.