Qeios

Deming Regression: Least-squares Analysis with Errors in both x and y Data, and a Simple Spreadsheet Implementation

Leslie Glasser*

Curtin Institute for Computation, Discipline of Chemistry

Curtin University, GPO Box U1987, Perth, WA 6845, Australia

*Corresponding author:

L. Glasser: Email: l.glasser@curtin.edu.au
ORCID iD: 0000-0002-8883-0564

https://doi.org/10.32388/6GMLFI

Abstract

Almost every system or device capable of computation contains a linear least-squares parametric analysis program by which to fit a set of dependent y-data against a set of independent x-data to a straight line. Such programs are used in fields as diverse as finance and sport and are of especial significance in science. Although widely – and appropriately – applied in many circumstances it is not always recognised that it is required that the x-data be without error (or, at least, of small error) and that the fitted values are sensitive to outliers. If this is not the case, the slope and intercept parameters fitted to the line may have significant error. We briefly mention "median of medians" non-parametric procedures by which effects of outliers may be considerably reduced.

Errors-in-Variables (EIV) methods, of which Deming Regression is the most popular, provide a remedy by recognising that both x- and y-data may contain significant error and so yielding more reliable fitted parameters. Unfortunately, software for Deming Regression is not generally available except by purchase or subscription.

Here, we describe the statistics of Deming Regression and provide an Excel spreadsheet which applies the program to user-supplied data. The Excel spreadsheet is compatible with the free Open Office spreadsheet, and may also be used with Google Online Sheets but with some loss of functionality.

Introduction

Most modern calculators and communication systems, smartphones, tablets, browsers, laptops, computers, have access to a linear least-squares analysis program which provides a best fit approximation of a straight line, y = a + bx, using the dependent y data referred to the reference x-data. These statistical programs are used for a myriad of activities including finance, social purposes, sports statistics, medical tests, and so on. They are also regularly used in every field of scientific analysis.

The conditions for these programs to return reliable values of the linear parameter estimates, the slope, b, and intercept, a, are well-known: "that the standard errors (SE) of the parameter estimates are exactly predictable if the error structure of the data is known; the estimates are normally distributed if the data error is normal, and even if it is not, in the limit of a large number of points, where the central limit theorem ensures normality."[1] This implies constant variance, also termed homoscedasticity. In addition, the independent x values are assumed to be with little or no error.

In many circumstances, however, it may not be possible to rely on error-free x values. For example, in analytical and clinical chemistry there is a need to compare the efficacies of various analytical methods against one another – say, a new instrumental method against a current standard method[2] - or a correlation of one thermodynamic property against another.[3] In circumstances such as these linear least-squares regression needs to be substituted by a more general type of analysis, often termed "errors in variables" (EIV) regression.[4-5] The most-frequently applied of these methods is one popularised by W. E. Deming (from the 1940's),[6-7] often termed Deming Regression.[8] Programs for this regression are available from most statistical software systems, but generally only for the more sophisticated purchased or subscribed versions. Even the free trial version of XLSTAT requires purchase to access Deming Regression.[9] The free R statistical software does have a Deming Regression but requires special loading.[10]

Deming Regression is sensitive to outliers because of its least-squares basis. An alternative to Deming Regression which is robust to outliers and available with an Excel application[11] uses a "median of medians" regression[12-15] rather than least squares.

The purpose of the current publication is to introduce Deming Regression together with a simple spreadsheet version in widely-available statistical software such as Excel, Google Sheets Online, and even the free Open Office suite. In 1992, Ward and Cornish[16] published some details of their now-archaic Lotus 1-2-3 program.

Deming Regression

As implied by the term "least-squares", the objective of these linear regression programs is to adjust the position of the representative line so as to minimise the sum of the squared distances between the data points and the line. Note that the use of distances which are squared enhances the vagaries of any data outliers,[11] which thus need to be eliminated before analysis.[17] In the standard least-squares regression program, where it is assumed that the independent x values are without error, it is the sum of the squared *vertical* distances between the data points, (x_i, y_i), and the fitted line BC, which is minimised, as may be seen in Fig. 1.

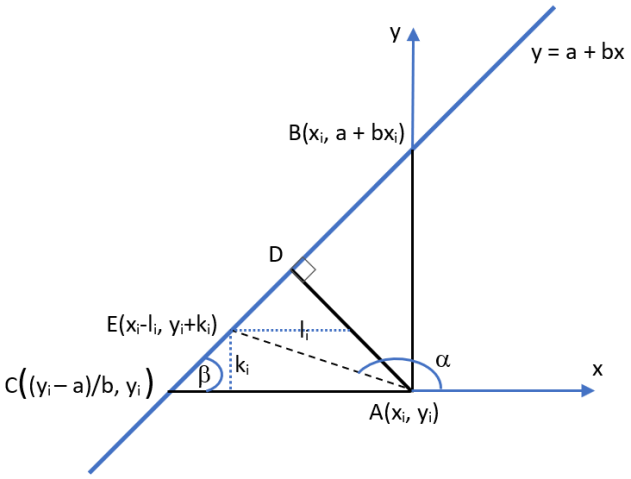

Figure 1: Distance from a data point (x, y) to the estimated regressed line y = a + bx. Figure adapted and extended from Wakkers, et al.[18], and Xu.[5] Special cases[5] are as follows: If α = $\pi/2$ or $3\pi/2$ then the distance to the line is the vertical |AB| = $|(y_i - \bar{y}) - b(x_i - \bar{x})|$; if $\alpha = 0$ or π then the distance to the line is the horizontal |AC| = $|\frac{1}{b}| \times |(y_i - \bar{y}) - b(x_i - \bar{x})|$; if $\alpha = \beta \pm \pi/2$ then the distance to the line is the orthogonal |AD| = |AB||AC|/|BC| = $|\frac{1}{\sqrt{1+b^2}}| \times |(y_i - \bar{y}) - b(x_i - \bar{x})|$; while when $\tan(\alpha) = -sign(S_{xy})\sqrt{\frac{S_{yy}}{S_{xx}}}$ then the adjusted distance to the line is |AE|= $\frac{\sqrt{S_{xx} + S_{yy}}}{|b|\sqrt{S_{xx}} + \sqrt{S_{yy}}} |(y_i - \bar{y}) - b(x_i - \bar{x})|$. (See text below for introduction of the weighted statistical summation terms S_{xx}, S_{xy} and S_{yy}.)

In Deming Regression. if the errors are the same in both the x and y data, then the distances to be minimised are those perpendicular to the fitted line, in a process that is termed orthogonal least-squares. When the errors in x and y are different, so that the data is heteroscedastic, then the distances to be minimised will lie at an angle α to the horizontal, as seen in line AE in Fig. 1. If the errors in the methods are defined by the variance ratio $\lambda = \sigma_y^2 / \sigma_x^2$ then the random errors in x and y have relative magnitudes $\sqrt{\lambda}$ and 1, respectively.

In summary,[5] in the regression:

> if the vertical differences (residuals) are minimised so that $\lambda = \infty$ then the slope is b_x
> if the horizontal differences (residuals) are minimised so that $\lambda = 0$ then the slope is b_y
> if the orthogonal differences (residuals) are minimised so that $\lambda = 1$ then the slope is b_0
> if the weighted differences (residuals) are minimised so that $\lambda = \sigma_y^2 / \sigma_x^2$ then the slope is b

We now follow the development of the Deming least-squares analysis according to Linnet.[19-20]

If (x_i, y_i) are the measured observations of the values (X_i, Y_i) on the regression line with independent standard errors e and d (where "e" and "d" represent "exploratory" and "dependent" respectively) such that

$$y_i = Y_i + d_i \text{ and } x_i = X_i + e_i$$

with variance ratio $\lambda = \sigma_d^2 / \sigma_e^2$, then we seek the line of "best fit" Y = a + bX by finding the minimised weighted sum of squared residuals[20]

$$S_w = \sum_{i=1}^{n} [w_i (x_i - \hat{X}_i)^2 + \lambda (y_i - \hat{Y}_i)^2]$$

where the circumflex denotes the linear-fitted estimate of the true values, that is $\hat{Y}_i = a + bx_i$.

The following weighted statistical quantities, each summed from 1 to n, are required:

$$\text{Weighted Sample Means: } \bar{x} = \Sigma w_i x_i / \Sigma w \qquad \bar{y} = \Sigma y_i / \Sigma w$$
$$\text{Weighted Covariance: } S_{xy} = \Sigma w_i (x_i - \bar{x})(y_i - \bar{y})$$
$$\text{Weighted Variances: } S_{xx} = \Sigma w_i (x_i - \bar{x})^2$$
$$S_{yy} = \Sigma w_i (y_i - \bar{y})^2$$

Finally, the least-squares estimates of the model's parameters will be

$$b = \frac{\lambda S_{xx} - S_{yy} + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}}$$

$$a = \bar{y} - b\bar{x}$$

$$X_i = x_i + \frac{\lambda b}{\lambda b^2 + 1}(y_i - a - bx_i) \qquad Y_i = y_i - \frac{(y_i - a - bx_i)}{\lambda b^2 + 1}$$

The net effect of the difference in variance between the x and y data sets is encapsulated in the variance ratio parameter $\lambda = \sigma_y^2 / \sigma_x^2$ leading to a corresponding change in angle of the distances to be minimised (cf. Fig. 1). If the variances are equal, $\lambda = 1$ and $\alpha = 45°$, while if the variance of x is zero, then $\lambda = \infty$ which is the case for the standard linear least-squares regression.

The details of these analyses may be found in standard statistical textbooks, especially Deming's "Statistical Adjustment of Data",[7] together with publications by Cornbleet and Gochman,[21-22]

Glaister (who does not reference Deming),[23] Linnet,[19-20] and York, et al.[24] supplemented by important analyses by Tellinghuisen[1, 25-27] and online resources such as NCSS Software.[8]

Implementations

As noted above, most statistical software includes a Deming Regression program (although surprisingly not for SPSS) though often at the price of a purchase or subscription. Billo provides such a program on the CD accompanying his book[28] while Dr. Jon Peltier has implemented this program as a free Excel add-in.[29] Both require an even number of (x_i,y_i) data pairs. Tellinghuisen has a brief demonstration in Excel and in KaleidaGraph.[1] Zaiontz provides an Excel description in his "Real Statistics" online blog.[30] It should be noted that the Billo/Peltier program includes a fixed weighting of the input data by using the means of successive pairs for the input data, in a procedure noted by Linnet[20] - this will influence the values of the fitted parameters.

In conjunction with the present publication, we provide a simple Excel spreadsheet which may be compatible with other spreadsheet software (except for the Billo/Peltier Excel add-in). The layout is inspired by the Zaiontz blog[30] but is extended to allow user-weighting of the data, and includes regression of both y on x together with x on y. This spreadsheet includes a demonstration that these two regressions are symmetrically related – that an estimate of y from x gives exactly the same value as an estimate of x from y.

Table 1 compares the various parameters yielded from the regression analyses for the given set of data.

Table 1: Slopes, b, intercepts, a, and values of λ for both the y vs x regression and the x vs y regression for the data on the accompanying spreadsheet for each of the three procedures. LINEST is a straightforward Excel linear least-squares regression assuming no error in the independent variable, the Current Excel (or Deming) regression assumes weighting of the x and y variables by their inverse variances, and the Billo/Peltier regression is weighted by using the mean of paired (x, y) values for weighting. $\lambda = \sigma_x^2 / \sigma_y^2$.

	y vs x regression			x vs y regression		
	slope, b	**intercept, a**	$\lambda_{y\|x}$	**slope, b**	**intercept, a**	$\lambda_{x\|y}$
LINEST	1188.52	23.97	∞	7.82E-04	-1.26E-02	0
Current Excel	1232.47	20.14	6.58E-07	8.11E-04	-1.63E-02	1.52E+06
Billo/Peltier	1277.46	16.21	-	7.83E-04	-1.27E-02	-

Format of Spreadsheet

The accompanying spreadsheet is divided into two columnar areas: on the left the data is provided in three columns, x, y , and w (weight) in columns B, C and D respectively; on the right the same data is copied as y, x, and w, in columns N, O, and P, respectively. The data is manipulated within the upper-left light-green block to provide statistical results, principally λ, slope a and intercept b for the (x, y) data set. The identical calculations are provided in the upper light-yellow block on the right-hand side for the (y, x) data. Fig. 2 shows the formulae of the calculations as involved for the (x, y) data.

Deming		Weighted	Unweighted
	xbar	=SUM(D6:D215*B6:B215)/G12	=SUM(B$6:B$215)/H12
	ybar	=SUM(D6:D215*C6:C215)/G12	=SUM(C$6:C$215)/H12
	Sx^2	=SUM((D6:D215)*(B6:B215-G7)*(B6:B215-G7))/G12	=DEVSQ(B$6:B$215)/H12
	Sy^2	=SUM((D6:D215)*(C6:C215-G8)*(C6:C215-G8))/G12	=DEVSQ(C$6:C$215)/H12
	Sxy	=SUM((D6:D215)*(B6:B215-G7)*(C6:C215-G8))/G12	=SUM((B6:B215-H7)*(C6:C215-H8))/H12
		=D2	=COUNT(D6:D215)
λ =	lambda	=G9/G10	=H9/H10
b =	**slope**	**=SQRT(G10/G9)**	**=(H13*H10-H9+SQRT((H13*H10-H9)^2+4*H13*H11^2))/(2*H13*H11)**
a =	**intercept**	**=G8-G7*G14**	**=H8-H7*H14**
	R =	=(G12*SUM(B6:B215*C6:C215)-B2*C2)/SQRT((G12*SUM(B6:B215*B6:B215)-B2^2)*(G12*SUM(C6:C215*C6:C215)-C2^2))	=PEARSON(B6:B215,C6:C215)
	R^2 =	=G16^2	=H16^2
	Billo/Peltier =	**=Deming(C6:C215,B6:B215)**	

Figure 2: The Excel formulae used to generate the statistical results required. The x, y and w data are listed in columns B, C, and D. Alterations to the values in column D will alter the results in the Weighted column but not in the Unweighted column. The upper light-yellow block on the right of

the spreadsheet contains the identical calculations but referred to columns N, O and P for the y, x, w data.

The coloured blocks below the upper calculation blocks provide Excel LINEST results for the data to their left. It will be noted that the slope and intercept data (and their standard errors) do not exactly match their corresponding Excel Deming values in the upper block because the built-in LINEST weighting does not correspond with the weighting listed in the spreadsheet's w column. The results from the Billo/Peltier procedure in the base of the upper block differ similarly.

The lower "Symmetry" light-blue block on the left actively demonstrates that the slopes and intercepts of the (x, y) and the (y, x) regressions are symmetric – that is, an x-value input into the (x, y) regression generates a y-value which, when itself entered into the (y, x) regression, yields the identical initial x-value. Correspondingly, the pair of λ values are reciprocals so that their product is one exactly.

Finally, a chart on the right of the spreadsheet (here copied as Fig. 3) plots the input data as blue dots as well as the trendlines of the Billo/Peltier calculation and of the current Excel Deming analysis.

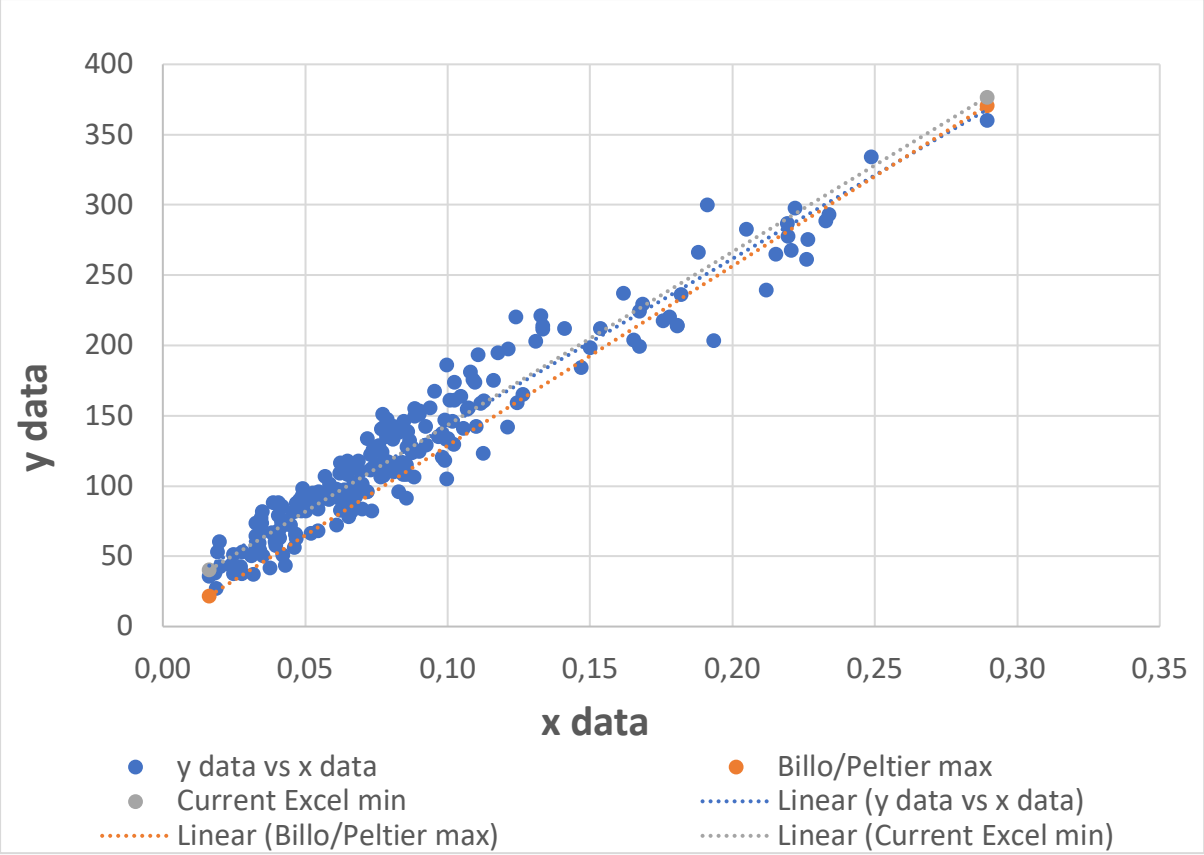

Figure 3: A chart showing the fitted lines generated by the Excel least-squares linear regression, the Billo/Peltier and the current Excel Deming regressions. The chart on the spreadsheet is active and will respond to changes in the input data and its weighting.

Acknowledgements

The advice of Dr. Jon Peltier in understanding the weighting procedure in the Billo algorithm is gratefully acknowledged as well as the provision of an updated Deming Regression utility. I also acknowledge Charles Zaiontz's blog as the inspiration for the layout of the accompanying Excel spreadsheet.

Supplementary Files

To ensure compatibility with Open Office, which usually uses semi-colons as separators while Excel uses commas, the Excel function “SUMPRODUCT(range1, range2)” is substituted by the equivalent “SUM(range1*range2)” and unfamiliar functions are avoided while enhanced functions such as “VAR.S” use the simpler form “VAR”. Open Office does not use the Billo/Peltier add-in function. The data set used in the spreadsheet relate ambient entropies of ionic solids to their formula volumes[3] and was chosen to reflect the effects of differences in scale between x and y data on the variances and the consequent difference of λ from the value 1 in the Excel Deming Regression analysis.

The free Deming utility is available from Jon Peltier’s website https://peltiertech.com/deming-regression/.

References

1. Tellinghuisen, J., Least-Squares Analysis of Data with Uncertainty in y and x: Algorithms in Excel and KaleidaGraph. *J. Chem. Educ.* **2018,** *95*, 970-977.
2. Analyse-it Method comparison / Agreement. https://analyse-it.com/docs/user-guide/method-comparison/method-comparison (accessed December, 2023).
3. Glasser, L., Thermodynamic estimation: Ionic materials. *J. Solid State Chem.* **2013,** *206*, 139-144.
4. Wikipedia - Errors-in-variables models. https://en.wikipedia.org/wiki/Errors-in-variables_models (accessed December, 2023).
5. Xu, S., A Property of Geometric Mean Regression. *The American Statistician* **2014,** *68*, 277-281.
6. Deming, W. E., Statistical Adjustment Of Data Wiley, New York: 1943. https://archive.org/details/in.ernet.dli.2015.18293/mode/2up.
7. Deming, W. E., *Statistical Adjustment of Data*. Dover Publications: 2011.
8. NCSS Statistical Software Deming Regression. https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Deming_Regression.pdf

https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/PASS/Deming_Regression.pdf (accessed December, 2023).
9. XLSTAT Deming Regression. https://www.xlstat.com/en/solutions/features/deming-regression.
10. Diagnostics, R. mcr: Method Comparison Regression. https://cran.r-project.org/package=mcr (accessed Septembner, 2024).
11. Glasser, L., Dealing with Outliers: Robust, Resistant Regression. *J. Chem. Educ.* **2007,** *84*, 533.
12. Siegel, A. F., ROBUST REGRESSION USING REPEATED MEDIANS. *Biometrika* **1982,** *69*, 242-244.
13. H., P.; W., B., A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part I. *Journal of Clinical Chemistry and Clinical Biochemistry* **1983,** *21*, 709–20.
14. Passing, H.; W Bablok, W., Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part II. *J Clin Chem Clin Biochem.* **1984,** *22*, 431-45.
15. Holmes, D. T. Deming and Passing Bablok Regression in R. https://www.r-bloggers.com/2015/09/deming-and-passing-bablok-regression-in-r/ (accessed September, 2024).
16. Ward, L. C.; Cornish, B., Use of a spreadsheet program for Deming's linear regression analysis. *Computer Methods and Programs in Biomedicine* **1992,** *37*, 101-105.
17. Twomey, P. J.; Kroll, M. H., How to use linear regression and correlation in quantitative method comparison studies. *International Journal of Clinical Practice* **2008,** *62*, 529-538.
18. Wakkers, P. J. M.; Hellendoorn, H. B. A.; De Weegh, G. J. O.; Heerspink, W., Applications of statistics in clinical chemistry: A critical evaluation of regression lines. *Clin. Chim. Acta* **1975,** *64*, 173-184.
19. Linnet, K., Evaluation of regression procedures for methods comparison studies. *Clin. Chem.* **1993,** *39*, 424-432.
20. Linnet, K., Estimation of the linear relationship between the measurements of two methods with proportional errors. *Statistics in Medicine* **1990,** *9*, 1463-1473.
21. Cornbleet, P. J.; Gochman, N., Incorrect least-squares regression coefficients in method-comparison analysis. *Clin. Chem.* **1979,** *25*, 432-438.
22. Cornbleet, P. J.; Gochman, N., When Linear Regression Gets Out of Line: Finding the Fix. *Clin. Chem.* **2020,** *66*, 1238-1239.

23. Glaister, P., Least Squares Revisited. *The Mathematical Gazette* **2001,** *85*, 104-107.
24. York, D.; Evensen, N. M.; Martínez, M. L.; De Basabe Delgado, J., Unified equations for the slope, intercept, and standard errors of the best straight line. *Am. J. Phys.* **2004,** *72*, 367-375.
25. Tellinghuisen, J., Least-squares analysis of data with uncertainty in x and y: A Monte Carlo methods comparison. *Chemometrics and Intelligent Laboratory Systems* **2010,** *103*, 160-169.
26. Tellinghuisen, J., Using Least Squares for Error Propagation. *J. Chem. Educ.* **2015,** *92*, 864-870.
27. Tellinghuisen, J., Using Least Squares To Solve Systems of Equations. *J. Chem. Educ.* **2016,** *93*, 1061-1067.
28. Billo, E. J., *Excel for Chemists*. 3rd ed.; Wiley: Hoboken, New Jersey, 2011.
29. Peltier, J. Deming Regression. https://peltiertech.com/deming-regression/ (accessed December, 2023).
30. Zaiontz, C. Real Statistics Using Excel: Deming Regression Basic Concepts. https://real-statistics.com/regression/deming-regression/deming-regression-tool/deming-regression-2/ (accessed December, 2023).