Deming, Theil-Sen, and Passing-Bablock Regression

Terry Therneau
Mayo Clinic

September 23, 2014

1 Introduction

The methods in the *deming* package are concerned with the problem of comparing two assays, both of which are measured with error. Let x_i and y_i be the two measurements of a compound where the true value of the quantity is u_i and assume that both assays are linear.

$$x_i = a + bu_i + \epsilon_i \tag{1}$$
$$y_i = c + du_i + \delta_i \tag{2}$$
$$\tag{3}$$

where ϵ and δ are the errors. We would like to find the calibration equation $y = \alpha + \beta x$ that best maps between the two assays.

In this situation ordinary least squares applied to x and y is unsatisfactory since it is asymmetric. The fitted lines for $y \sim x$ and $x \sim y$ are not the same, and neither has an expected slope of 1 when $\beta = 1$.

2 Generalized Deming Regression

Least squares regression of y on x assumes that the x variate is measured without error, and minimizes the sum of squared vertical distance between the data points y and the fitted regression line. Regression of x on y minimizes the horizontal distances. Adcock [1] in 1878 suggested minimizing the sum of squared horizontal + vertical distances to the predicted values. However the idea of Adcock remained largely unnoticed for more than 50 years, until it was widely propogated in the book by Deming [2]. The latter has become so well known that the common label for the method is "Deming regression" in nearly all fields in which it is used.

There are a number of alternate ways to compute this regression line. The Deming line will be the first principle component of the centered data, the first eignevector of the matrix Z whose 2 columns are the centered x and y vectors, or the first component of a singular value decomposition or factor analysis of Z. A partial least squares (PLS) or structural equation modeling (SEM) model fit to x and y will also recover the Deming estimate of slope.

There would appear to be little need for yet another program to compute this quantity other than providing a recognizable name to search for in the R libraries. For laboratory work, however, it is the generalized Deming method that is of most interest. Returning to our original

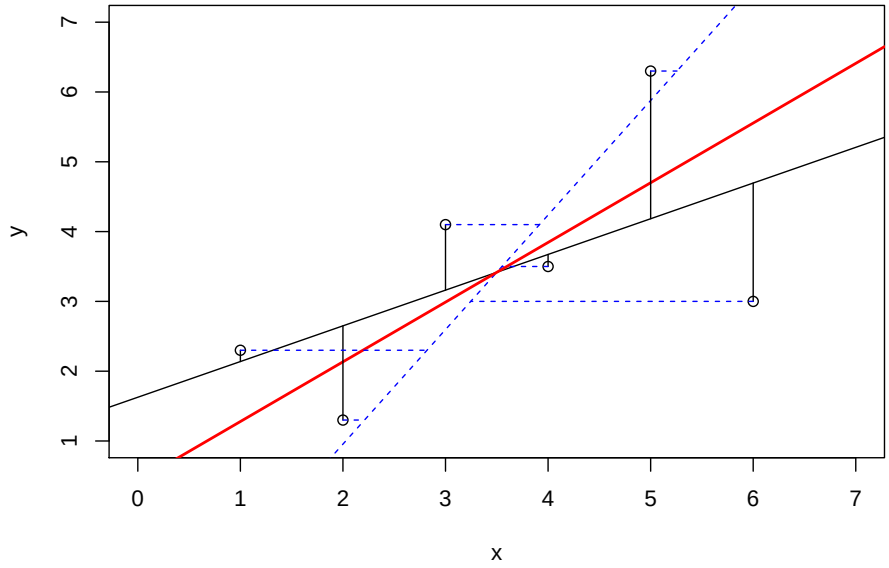

Figure 1: Example of linear and deming regression applied to a simple data set. The ordinary linear regression of y on x (black) minimizes the sum of squared vertical distances. The regression of x on y (blue, dashed) minimizes a sum of squared horizontal distances. The Deming regression (red) minimizes the sum of orthagonal distances between the points and the line.

definitions (1) and (2), ordinary Deming regression is based on the assumtion that that the assay errors ϵ and δ are equal in magnitude for the two assays and are constant across the range of u. This latter is rarely if ever true for biologic assays.

Figure 2 shows a Bland-Altman plot of paired assay results from long-term monitoring of a ferritin assay. Each time that a new lot of the principle reagent was brought into use, a subset of current samples were assayed in duplicate using both the old and new lot. If the assumptions of standard Deming regression hold we would expect to see approximately constant vertical variation across the range of the X axis of the plot. This is clearly not the case. The x-axis was plotted on a square root scale to spread out the data somewhat, but this does not change the message.

```
> f.ave <- with(ferritin, (old.lot + new.lot)/2)
> f.diff<- with(ferritin, old.lot - new.lot)
> plot(sqrt(f.ave), f.diff, xaxt='n',
       xlab="Average", ylab="Difference")
> temp <- 0:7*5
> axis(1, temp, temp^2)
```

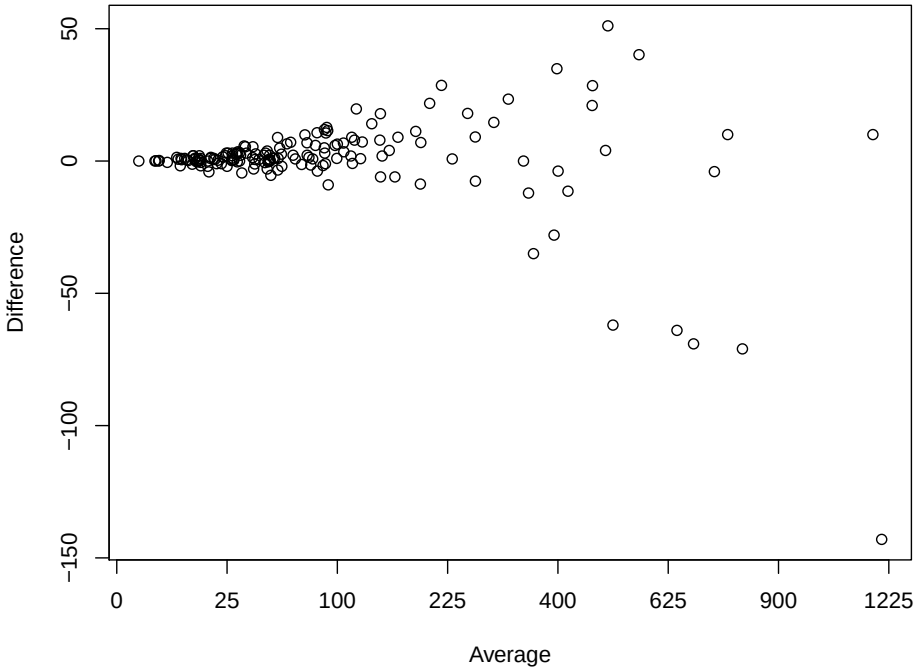

Figure 2: Bland-Altman plot of the ferritin data.

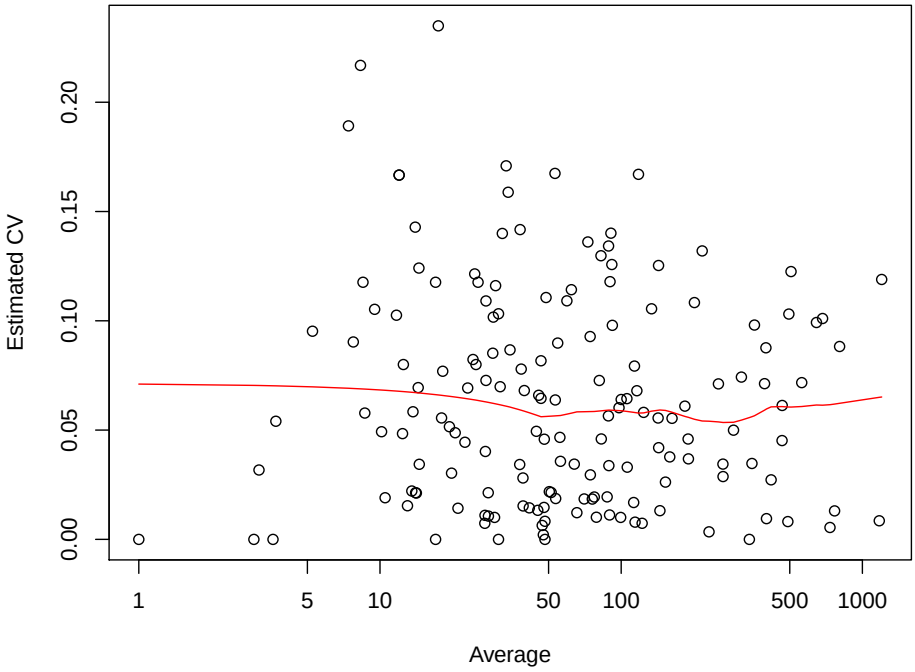

Figure 3: Revised variance plot for the ferritin data on a coefficient of variation scale.

Figure 3 shows a revised plot with average of the two assay values on the horizontal and abs(difference/mean) along the vertical axis, along with a lowess line. A horizontal trend in this plot corresponds to constant coefficient of variation, which for this data set appears to be a reasonable assumption.

```
> plot(f.ave, abs(f.diff/f.ave), log='x',
       xlab="Average", ylab="Estimated CV")
> lines(lowess(f.ave, abs(f.diff/f.ave)), col=2)
```

Linnet [3] discusses fitting regression lines in the situation of constant coefficient of variation, and gives a more complete rationale. We use an algorithm based on Ripley and Thompson [7] which includes both ordinary Deming regression and Linnet's extension within a more general framework. Referring again to equations (1) and (2), assume that x and y both estimate the common unknown quantity u, and the error terms have standard deviations

$$\mathrm{sd}(x) = \sigma[e + fu] = \mathrm{sd}\epsilon \tag{4}$$

$$\mathrm{sd}(y) = \sigma[g + hu] = \mathrm{sd}\delta \tag{5}$$

for known constants e, f, g, and h and an unknown scale factor σ, where u is again the true value. A value of $(e, f, g, h) = (1, 0, 1, 0)$ corresponds to standard Deming regression, and $(e, f, g, h) = (0, 1, 0, 1)$ corresponds to the constant proportional errors assumption of Linnet. The `cv` argument of the `deming` function chooses between these two cases, or all four constants can be supplied using the `stdpat` argument. A second alternative is for the user to directly supply values for $\mathrm{sd}(x)$ or $\mathrm{sd}(y)$ using the `xstd` and `ystd` arguments. The following produces the 7 calibration equations for each of the 7 reagent changes in the ferritin data set.

```
> cmat <- matrix(0, nrow=3, ncol=7)
> for (i in 1:7) {
      dfit <- deming(new.lot ~ old.lot, data=ferritin,
                     subset=(period==i), cv=TRUE)
      cmat[1:2,i] <- coef(dfit)
      cmat[3,i] <- coef(lm(new.lot ~ old.lot, ferritin,
                           subset= (period==i), weight=1/new.lot))[2]
  }
> dimnames(cmat) <- list(c("Intercept", "old.lot", "old.lot (LS)") , 1:7)
> round(cmat,3)
                  1      2     3     4     5      6     7
Intercept    -0.015 -0.982 2.390 0.234 0.208 -0.079 0.085
old.lot       0.986  1.015 0.962 0.948 0.913  0.981 0.971
old.lot (LS)  1.080  1.015 0.933 0.940 0.911  0.996 0.996
```

When the data has both a wide range and results near zero, it will often be necessary for the error to include both a constant and a proportional portion. The arsenate data set contains results of two different methods for assessment of arsenate(V) in river waters; the resultant estimates range from 0 to 19.25 $\mu g/l$. Constant proportional error (constant CV) is clearly untenable, since it would predict infinite precision for the smallest values. This data set contains estimates of the precision of each point, which we can use to obtain an appropriate fit.

```
> afit <- deming(aas ~ aes, arsenate, xstd=se.aes, ystd=se.aas)
> afit
Call:
deming(formula = aas ~ aes, data = arsenate, xstd = se.aes, ystd = se.aas)

n= 30
                Coef  se(coef) lower 0.95 upper 0.95
Intercept -0.1094048 0.3083245 -0.7137096  0.4949001
Slope      1.0277709 0.1705373  0.6935239  1.3620179

   Scale= 1.165495
> dfit <- deming(aas ~ aes, arsenate)
> lfit <- lm(aas ~ aes, arsenate)
> temp <- cbind(coef(afit), coef(dfit), coef(lfit))
> dimnames(temp)[[2]] <- c("Ripley", "Deming", "Linear")
> round(temp,3)
            Ripley Deming Linear
(Intercept) -0.109 -0.490 -0.299
aes          1.028  1.142  1.089
```

For values less than .3 (about 10% of the data) the constant part of the error is predominant while for those above 2 the proportional part is the largest. The calibration fits that do or do not properly account for the error differ by important amounts.

3 Theil-Sen Regression

One interesting way to characterize the slope of least squares regression line is that it is the solution of $\rho(x, r(\beta)) = 0$, where ρ is the Pearson correlation coefficient and $r(\beta)$ are the residuals from a fitted line with slope β. A non-parametric counterpoint to this is Thiel-Sen regression, which satisfies $\tau(x, r(\beta)) = 0$ where τ is Kendall's tau, a rank based alternative to the correlation coefficient. This was proposed by Theil [10], Sen [8] extended the results and added a confidence interval estimate. The approach is well known in selected fields (e.g. astronomy), and almost completely unknown in others. It has strong resistance to outliers and nearly full efficiency compared to linear regression when the errors are Gaussian.

To calculate the TS regression fit first draw a line segment between each of the $n(n-1)/2$ unique pairs of points in the data; the TS slope estimate is the median of these $n(n-1)/2$ slope values.

Figure 4 shows a plot of $x_i - x_j$ vs $y_i - y_j$ for all $8 * 7 = 56$ data pairs from a small set of 8 data points. A line from the origin to each point has identical angle to a line connecting that pair of points in a plot of 8 original (x, y) pairs. Each pair of points i, j appears twice in the paired plot, corresponding once to $y_i - y_j$ and a second time using $y_j - y_i$. The Thiel-Sen estimate of slope is that line through the origin such that quadrants 1–4 of the plot, formed by this line and the vertical axis, each have the same number of points. The solution is simply `median(atan(dy/dx))` where `dy` and `dx` are the paired y and x differences, respectively. Since $(y_i - y_j)/(x_i - x_j) = (y_j - y_i)/(x_j - x_i)$ the code only uses the $n(n-1)/2$ unique values, removing

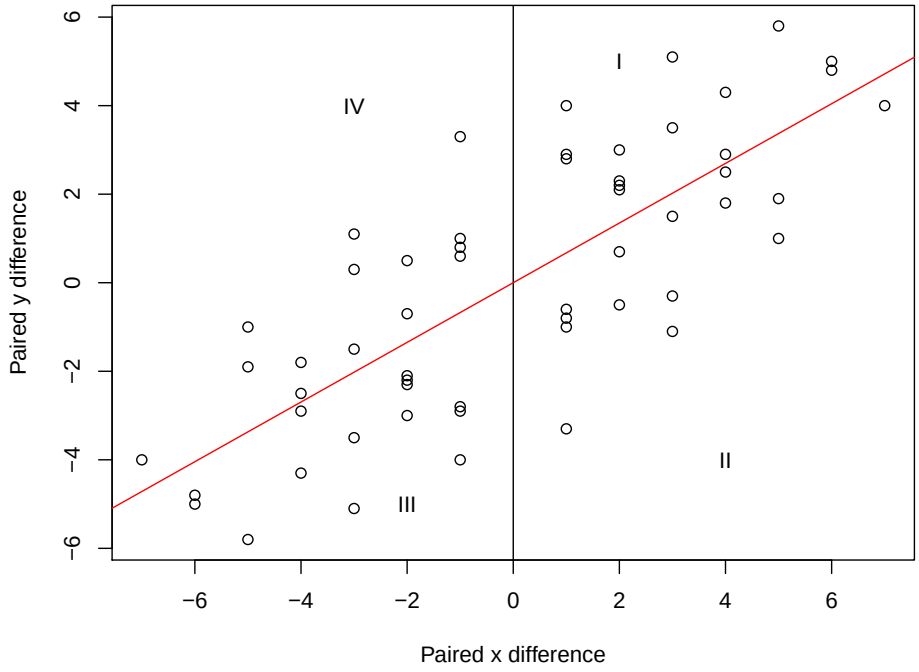

Figure 4: The geometry underlying the Thiel-Sen estimator. The set of values $x_i - x_j$ is plotted versus $y_i - y_j$ for all $i \neq j$ along with a reference line $x = 0$. The red line divides the points into four equal groups, and is the Thiel-Sen estimate of slope.

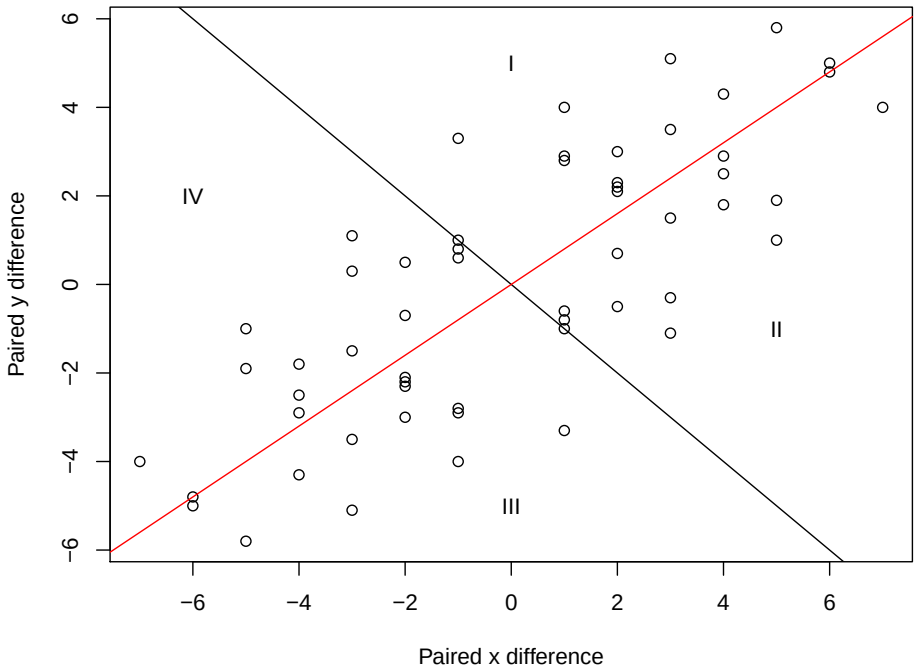

Figure 5: The geometry underlying the Passing-Bablock estimate.. The set of values $x_i - x_j$ is plotted versus $y_i - y_j$ for all $i \neq j$. The reference line with slope -1 (black) and the estimated PB slope coefficent (red) divide the points into 4 equal groups.

any which lie exactly on the vertical axis since they would count equally in two quadrants and thus cancel. Theil-Sen regression of x on y would use the horizontal axis, rather than vertical, as the reference for forming quadrants.

4 Passing-Bablock Regression

The Thiel-Sen slope, like ordinarly least squares, is biased towards zero if there is error in both x and y, and like ordinary LS it is not symmetric in x and y. Passing and Bablock proposed variations on the Thiel-Sen estimate to address these concerns. Their method is well known in the field of laboratory testing but almost unheard of outside of that domain. There are actually 3 estimators, proposed in a series of papers in 1983, 1984, and 1988.

The first Passing-Bablock method (PB1) is described in their 1983 paper [4]. It modifies the Thiel-Sen estimate so as to make the procedure symmetric about the line $y = x$ instead of about the horizontal axis. To do so replace the vertical axis of figure 4 with the line $y = -x$ as the second reference for forming the four quadrants as shown in figure 5. Computationally, it suffices to modify the arctan function so as to return angles in the range of $(-\pi/4, 3\pi/4)$ instead of the default of $(-\pi/2, \pi/2)$. The kernel of the R code is three lines:

```
theta <- atan(dy/dx)
theta <- ifelse(theta < -pi/4, theta+pi, theta)
slope <- median(theta)
```

where `dy` and `dx` are the paired differences in x and y. Points where the angle is exactly $-\pi/4$ would count equally in both quadrants so can be ignored. (Since both x and y are measured with error such values should be rare in real data.) As with the Theil-Sen estimate, the underlying R routine only evaluates and uses points in quadrants I and II.

For a two-sided confidence interval Passing and Bablock use an identical formula to that derived for Thiel-Sen regression, namely the kth angles above and below the median value where

$$k = (z_{\alpha/2}/2)\sqrt{V_n}/2$$
$$V_n = (1/18)[n(n-1)(2n+5)/18$$

In the second paper of their series [5] they show that this method has excellent power, nearly as good as Deming regression when the data has Gaussian errors, while gaining resistance to outliers.

The PB1 estimate is symmetric in x and y, but it is also biased towards a slope of 1 to the same extent as the Theil-Sen estimate is biased towards 0. It is also not scale invariant: if all of the y values for a dataset are multipled by some constant k the recomputed slope does not necessarily equal $k\beta$. In the third paper of their series [6] two further estimators PB2 and PB3 are proposed which are scale invariant, while retaining symmetry in x and y. For the PB2 estimate, first find a value m which is the median of the angles in the lower right portion of figure 5, i.e. points with $dy < 0$ and $dx > 0$. Then remap the θ values to lie in the interval $(m, m+\pi)$ before taking the median, i.e., we use the line with slope m as our second reference in defining the four quadrants.

The PB3 estimate is defined self referentially such that $\hat{\beta}$ is the median angle after mapping θ into the range $(-\hat{\beta}, -\hat{\beta}+\pi)$. Referring to figure 4 or 5, a pair of lines at angles β and $-\beta$ are opened and shut like a pair of scissors about the x-axis until they enclose 1/2 of the data points. Passing and Bablock describe an iterative estimation procedure, however it is easy to see that `median(abs(theta))` provides a direct solution.

5 Other estimates

Another approach to modifying the Thiel-Sen estimate is to directly mimic one of the definitions of the Deming estimate, namely as the solution to an estimating equation. The angle θ of the Deming regression line is that rotation of the original data set such that $\rho(x^*, y^*) = 0$ where x^*, y^* are the rotated data points

$$x^* = y\sin(\theta) + x\cos(\theta)$$
$$y^* = y\cos(\theta) - x\sin(\theta)$$

Similarly we can define a *circularly symmetric* Thiel-Sen estimate (STS) as that angle θ such that $\tau(x^*, y^*) = 0$. Equivilantly, the Deming estimate is that rotation such that a least squares slope for the rotated data would be zero, and the STS estimate is that rotation for which the Theil-Sen slope would be zero.

An iterative algorithm for finding the optimal rotation would be the following:

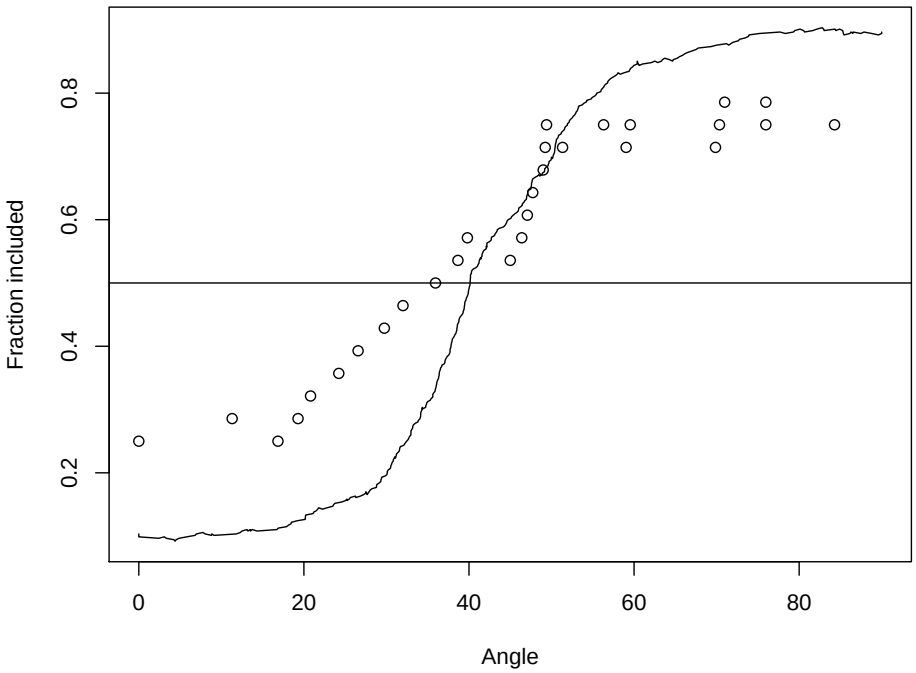

Figure 6: Solution trace of the algorithm for the simple test data set (points) and for the larger arsenate data set as a connected line.

1. Rotate by some starting estimate ϕ
2. Compute the Theil-Sen regression for the rotated data, with resulting angle θ
3. Rotate the data further by θ, if $\theta \neq 0$ return to step 2.

Passing-Bablock methods 1 and 2 can be viewed as one step approximations to the STS estimate that use $\phi = -\pi/2$ and $\phi = m$ as their starting estimates. Not surprisingly, for data that clusters tightly about a line this first step is nearly sufficient and the STS and PB1 or PB2 estimators differ only slightly.

Referring again to figures 4 and 5, the STS solution will be that pair of *orthagonal* axes that divides the data into four equal portions. It is easy to verify that rotation of the original data cloud by some angle ϕ results in the rotation of figure 4 by precisely the same amount; Kendall's τ for the rotated data will be 0 if the count of points in the 4 quadrants formed by the vertical and horizontal axes are identical. The R code efficiently enumerates all $n(n-1)/2$ possible rotations.

Figure 6 shows a plot of the fraction of points inside one quadrant as we rotate from 0 to 90 degrees in computing the circularly symmetrc estimate. The solution is that angle where the graph crosses 1/2. Unlike the other estimates considered so far the STS estimate can have multiple zeros. For a data set like the arsenate study, where the overall data clusters tightly

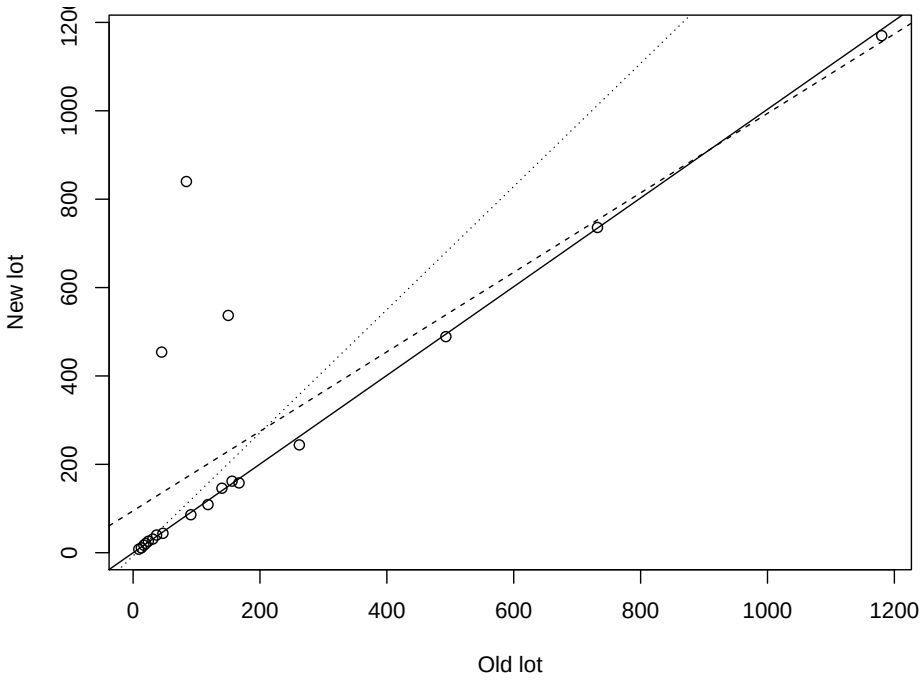

Figure 7: Ferritin data with outliers, along with OLS (dashed), Deming (dotted), and Passing-Bablock (solid) regression lines.

around a line, multiple solutions are uncommon, and when they occur normally form a small tight cluster of values. The other extreme is a set of points evenly distributed in cirle about the origin, for which there will be n solutions. When multiple solutions occur the program returns the value of that one having the smalled MAD of the residuals. The output structure also includes all of the solutions in its `angle` component.

For the PB2, PB3 and STS methods it is not clear that the Sen estimator of confidence limits is valid. Since they are not based on starting with a prespecified 1/2 of the plane a 1 to 1 mapping between the slope and Kendall's tau which forms the basis for Sen's argument no longer holds. Secondly, extending the Sen variance formula to data with case weights is far from clear. The `pbreg` and `thielsen` routines therefore also include an option for bootstrap confidence intervals, and we recommend using it whenever there are case weights or for the STS, PB2, and PB3 estimators. Due to the excessive number of ties that would be generated by ordinary bootstrap sampling the wild bootstrap method [11] is used.

6 Which method is best?

The two primary advantages of the robust methods in laboratory studies are that they give a robust estimate of the slope in the case of outliers and are less sensitive to choosing the correct

variance specification. Figure 7 shows the result on a data set with outliers: one of the two laboratory methods has had 3 assay failures. The PB regression line tracks the main body of the data, while the other two lines are pulled away.

```
> plot(new.lot ~ old.lot, data=ferritin2, subset=(period==2),
       xlab="Old lot", ylab="New lot")
> dfit <- deming(new.lot ~ old.lot, ferritin2, subset=(period==2),
                 cv=TRUE)
> lfit <- lm(new.lot ~ old.lot, ferritin2, subset=(period==2))
> pfit <- pbreg(new.lot ~ old.lot, ferritin2, subset=(period==2))
> abline(pfit, col=1)
> abline(lfit, lty=2)
> abline(dfit, lty=3)
```

A discussion by Støckl, Dewitte, and Thienpont provides a useful counterpoint. Essentially, if the data is good, all the methods will agree on that fact. If there are assay issues, outliers in particular, then the actual source of the problem needs to be investigated rather than just using a "better" regression tool. Understanding data requires more than pushing a button.

They argue further, and I think incorrectly, that ordinary least squares can suffice. The ferritin data is a counter-example. In order to provide long term calibration of the assay for the purposes of patient care, the calibration corrections used by the lab will be the cumulative product of the regression slopes. If OLS were used at each stage the downward bias, even if it is small for each given reagent change, would accumulate over time.

References

[1] Adcock, R. J. (1878). A problem in least squares. The Analyst (Annals of Mathematics) 5:53–54.

[2] Deming, W. E. (1943). Statistical adjustment of data. Wiley, NY.

[3] Linnet, K. (1990), Estimation of the linear relationship between the measurements of two methods with proportional errors. Statistics in Medicine 9:1463-1473.

[4] Passing, H. and Bablock, W. (1983). 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. J. Clin. Chem. Clin. Biochem. 21:709–720.

[5] Passing, H. and Bablock, W. (1984). Comparison of several regression procedures for method comparison studies and determination of sample size. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part II. J. Clin. Chem. Clin. Biochem. 22:431–435.

[6] Bablock, W., Passing, H., Bender, R. and Schneider, B. (1988). A general regression procedure for method transformations. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part III. J. Clin. Chem. Clin. Biochem. 26:783–790.

[7] Ripley, B. D. and Thompson, M. (1987), Regression techniques for the detection of analytical bias, Analyst 112:377-383.

[8] Sen, P.B. (1968), Estimates of the regression coefficient based on Kendall's tau, Journal of the American Statistical Association 63: 1379-1389.

[9] Støckl D., Dewitte K., Thienpont L. M. (1998). Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data? Clin Chem. 44:2340-6.

[10] Theil, H. (1950), A rank-invariant method of linear and polynomial regression analysis. I, II, III, Nederl. Akad. Wetensch., Proc. 53: 386-392, 521-525, 1397-1412.

[11] Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). Annals of Statistics, 14, 1261-1350.