World Applied Sciences Journal 12 (10): 1860-1865, 2011
ISSN 1818-4952

Performance of Ols-Bisector Regression in Method Comparison Studies

[1]*Sinan Saraçlı and* [2]*H. Eray Çelik*

[1]Department of Statistics, Faculty of Science and Literature,
AfyonKocatepeUniversity, ANS Campus, 03200 Afyonkarahisar,Turkey
[2]Department of Statistics, Faculty of Science, Van Yüzüncü Yıl University, Van -Turkey

Abstract: Regression Analysis is a widely used technique in methodcomparison studies. There are several studies about estimating and testing interactions in a linear regression model but It's very important to use correct regression technique in comparison study to obtain correct results. In method comparison studies, whenboth ofthe dependent and the independent variables include some measurements errors, Type II Regression techniques must be used to calculate the correct parameters. The aim of this study is to discuss eight different regression techniques (OLS, OLS-Bisector, Major Axis (MA), Reduced Major Axis (RMA), Deming, Passing-Bablok, York and Theil) that may be used in method comparison studies and which are the alternatives of Ordinary Least Squares (OLS) regression analysis when the assumptions of OLS are not met and to suggest alternative techniques for calculating the correct linear relationship between the two methods. In simulation part of this study there has been generated different types of data and in all conditions, OLS-Bisector method, which bisects the OLS (Y|X) and OLS (X|Y), estimated the parameters near to real values and, show the best performance then all other techniques.

Key words: Type II Regression · Measurement Error · Method Comparison · Simulation Study

INTRODUCTION

The comparison of analytical methods using regression analysis began in the fifties when Mandel and Linnig[1] first applied the joint confidence interval test for the intercept and the slope to chemical problems. However, applying this test to the regression parameters derived from the least squares method assumes that the results in the x-axis (often the reference method) are error-free, or that the errors assigned to the reference method are negligible with respect to those given by the new method (y-axis). This is not always true since the precision of both methods must often be taken into account. These precisions can be considered using the different existing approaches for calculating regression coefficients and related statistical parameters that consider errors in both axes [2].

Linear Measurement error models arise when the independent variable in a regression analysis is measured with error. It is well known that this random measurement error artificially inflates the dispersion of the observations on the independent variable and biased least squares estimates of slope towards zero [3]. 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[4]. OLS assumes an error-free x variable and a constant analytical imprecision (s_a)of the y variable (also called "homoscedastic" variance), both of which are seldom met in practice [5].

Most of the statistical models used in method comparison studies are designed for normally distributed data. However, some systems in clinical diagnostic are based on counting of certain particles rather than measuring a substance. In some of these cases, particularly in hematology where counting of cell types is of primary importance, the assumption of normal distribution is not always appropriate. Thus, other distributions beside normal distribution need to be considered [6].

Correspoding Author: Dr. Sinan Saraçlı, Department of Statistics, Faculty of Science and Literature, AfyonKocatepeUniversity, ANS Campus, 03200 Afyonkarahisar, Turkey. Tel: + 90-272-2281311, E-mail: ssaracli@aku.edu.tr.

In considering the types of statistical model we might wish to apply, a starting point is to consider the types of question we might which to address. Considering first the comparison of two methods of measurement x and y, with x being a standard measurement, we may wish to substitute y for x so that y and x are interchangeable without need for information regarding the method used. The motivation for this may be that y is cheaper or less invasive to the patient than x, or one may suspect that y is more reliable/precise than x. For this comparison x may be a gold standard for the characteristic being measured. Alternatively there may be a gold standard z against which both x and y are to be compared. There may be a trade-off between the precision of each measure and accuracy/validity relative to z. The scales x and y may be expressed in the same units. For reasons of convention, it may be important for y to have the same measurement scale as x [7].

MATERIALS AND METHODS

A linear relationship between the target values of the two methods is assumed as; [8]

$$Y_i = \alpha\ \beta X_i$$

The measured value is likely to deviate from the target value by some small "random" amount (ε or δ). For a given sample measured by two clinical chemistry methods, the following relations exist [8].

$$x_i = X_i + \varepsilon_i$$
$$y_i = Y_i + \delta_i$$

Under these conditions, in this study the real model is planned as:

$$Yi = X_i + u_i$$

This means that the real intercept is equal to zero (0) and the real slope is equal to one (1). Then;OLS,OLS-Bisector, Major Axis (MA), Reduced Major Axis (RMA), Deming, Passing-Bablok (Pas-Bab.) York and TheilRegression techniques are applied to the data sets which are simulated as shown above by MATLAB 7.02. In simulation,these methods are comparedin estimating the known slope and known intercept of a regression line when the independent variable is measured with error and also MSE criteria is considered whether which technique gives the best fit to the real data. Simulation numbers for each sample is made as 100000/n and the MSE is calculated as;

$$MSE = \frac{\sum \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{n-k}$$

Here k is the number of the parameters. It's equal to 2in this study because all of the techniquesare simple linear regression techniques.

The computations of the parameters of OLS,OLS-Bisector, MA, RMA, Deming, Pas-Bab.York and Theil Regression techniquesare given inAppendix.

Table 1. shows the Mean Square Errors of the Regression Techniqueswith Student distribution at 30 degrees of freedom and including outliers in different sample sizes. In this table it's clear that the OLS-Bisector Regression's MSE value is smaller than all other techniques and this technique fits the data sets better than the others.

The performance of the OLS-Bisector Regression for 4, 10 and 30 degrees of freedom of Student distribution and either including or not any outliers can be seen In Table 2. MSE value of OLS-Bisector Regression technique is again smaller than all other techniques.

Table 1: Mean Square Errors of the Regression Techniqueswith Student distribution at 30 degrees of freedom and including outliers in different sample sizes

Regression Technique	Sample Size n=50	n=100	n=200
OLS	2.8172	2.459	2.3171
OLS-Bisector	*1.1504*	*1.0952*	*1.0789*
MA	1.2782	1.177	1.1486
RMA	1.4962	1.3232	1.2786
Deming	1.27	1.1774	1.1488
Pas-Bab.	1.3539	1.241	1.1991
York	1.4486	1.3209	1.2353
Theil	1.3919	1.3404	1.313

Table 2: Mean Square Errors of the Regression Techniqueswith Student distributions at different degrees of freedoms and either including outliers or not, in sample size 200

Regression Technique	Distribution Type T~4	T~4-O	T~10	T~10-O	T~30	T~30-O
OLS	2.9682	2.1475	4.3383	2.1573	5.3682	2.3171
OLS-Bisector	*1.0401*	*1.0777*	*1.0324*	*1.0393*	*1.0577*	*1.0789*
MA	1.1793	1.1335	1.3287	1.0946	1.5184	1.1486
RMA	1.2644	1.2763	1.2916	1.2053	1.3712	1.2786
Deming	1.1828	1.1303	1.3399	1.0953	1.5278	1.1488
Pas-Bab.	1.1601	1.1732	1.1394	1.1287	1.1873	1.1991
York	1.427	1.1851	1.4921	1.2643	1.3169	1.2353
Theil	1.3563	1.2816	1.2633	1.2869	1.2837	1.313

• T~4: Student distribution at 4 degrees of freedom.
• T~4-O: Student distribution at 4 degrees of freedom and including outliers

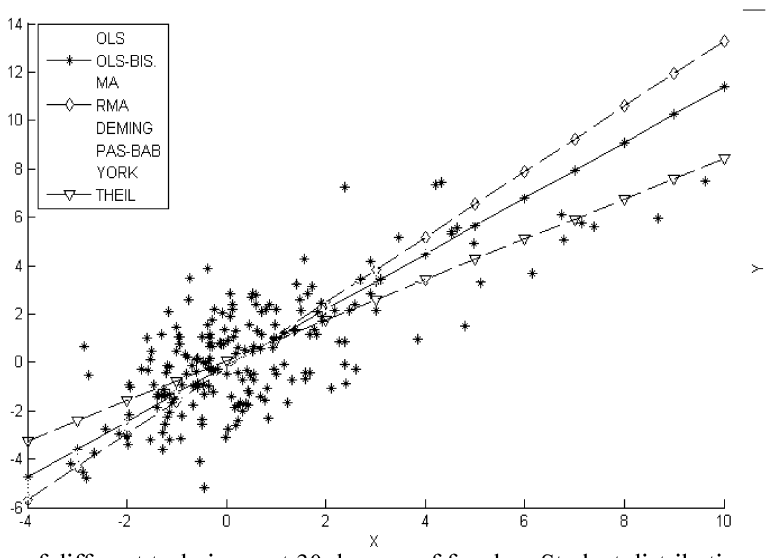

Fig. 3: Regression lines of different techniques at 30 degrees of freedom Student distribution and including outliers in sample size 200.

Table 3: Estimated regression parameters andMean Square Errors of the data sets of Student distribution at 30 different degrees of freedoms and including outliers in sample size 200

Regression Technique	$\hat{\beta}_0$	$\hat{\beta}_1$	*MSE*
OLS	-0.2976	1.6063	2.3171
OLS-Bisector	*-0.0433*	*1.1077*	*1.0789*
MA	-0.0723	1.1712	1.1486
RMA	-0.109	1.2477	1.2786
Deming	-0.0719	1.1708	1.1488
Pas-Bab.	-0.0985	1.2037	1.1991
York	-0.0743	1.1752	1.2353
Theil	0.1038	0.7136	1.313

In Figure 3., lines of all eight regression techniques are drawn at 30 degrees of freedom Student distribution and including outliers in sample size 200. This figure is also related with Table 3. and all the regression lines are drawn by the coefficients in Table 3.

In real model the regression coefficients was planned as 0 and 1 respectively. In Table 3. It can easily be seen that the coefficients of OLS-Bisector Regression is near to these values and MSE value of this technique is smaller than the others, which means that this is the best technique.

RESULTS AND DISCUSSION

In this study, performances of both Type I and Type II linear regression techniques, which are commonly used in method comparison studies, are compared via simulation study. The MSE criteria is took into consideration to determinate the best regression technique. β_0 and β_1 are considered as 0 and 1 respectively in the real model. The performances of the eight regression techniques are compared indifferent sample sizes (n=50, 100 and 200) and in different distribution types (student distribution at 4, 10 and 30 degrees of freedom) and either including or not any outliers.

As a result the OLS-Bisector regression technique, which bisects the OLS(Y|X) and OLS(X|Y), estimated the parameters near to real values than all other Type I and Type II regression techniques. MSE of this technique is also smaller than all other techniques.

In other studies about method comparison studies, it can be easily seen that the Deming regression is widespread used,and in some studies Passing-Bablok regression gives the best results. In this study the performance of OLS-Bisectional regression technique is tried to put forward and the findings showed that this technique is better than the others, in conditions of thisstudy.

When the conditions change this technique may or may not give the best performance but in other studies if the researchers think on OLS-Bisector regression, it will be very useful for them to obtain the correct results.

REFERENCES

1. Mandel, J. and F.J. Linnig, 1957. Study of Accuracy in Chemical Analysis Using Linear Calibration Curves. Anal. Chem., (29): 743-49.
2. Riu J. and F.X. Rius, 1997. Method comparison using regression with uncertainties in both axes.Trends in analytical chemistry, 16(4): 211-16.

3. Edland, S., 1996. Bias in Slope Estimates for the linear Errors in Variables Model by the Variance Ratio Method. Biometrics, (52): 243-48.
4. Cornbleet, P.J. and N. Gochman, 1979. Incorrect Least-squares Regression Coefficients in Method-Comparison Analysis. Clinical Chemistry, 25(3): 432-438.
5. Stöckl D., Dewitte K and ThienpontLM. 1998. Validityoflinearregression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data?.Clinical Chemistry, (44): 2340-46.
6. Magari, R.T., 2004. Bias Estimation in Method Comparison Studies. J. Biopharmaceutical Statistics, (14): 881-92.
7. Dunn, G. and C. Roberts, 1999. Modelling method comparison data. Statistical Methods in Medical Res., 8(2): 161-79.
8. Linnet, K., 1998. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. Clinical Chemistry, 44(5): 1024-31.
9. Saylor, R.D., E.S. Edgerton and B.E. Hartsell, 2006. Linear regression techniques for use in the EC tracer method of secondary organic aerosol estimation. Atmospheric Environment, (40): 7546-56.
10. Saraçlı, S., 2008. Comparison of Linear Regression Techniques in Measurement Error Models Monte-Carlo Simulation Study, Doctoral Dissertation, Eskişehir Osmangazi University.
11. Amman, L. and J.V. Ness, 1988. A Routine for Converting Regression Algorithms in to corresponding Orthogonal Regression Algorithms. ACM Transactions on Mathematical Software, 14(1): 76-87.
12. Isobe, T., E.D. Feigelson, M.G. Akritas and G.J. Babu, 1990. Linear Regression in Astronomy I. The Astrophysical J., (364): 104-113.
13. Cornbleet, P.J. and N. Gochman, 1979. IncorrectLeast-squares Regression Coefficients in Method-Comparison Analysis, Clinical Chemistry, 25/3: 432-438.
14. Magari, R.T., 2002. Statistics for Laboratory Method Comparison Studies, BioPharm,28-32, http:// www.biopharminternational.com/ biopharm/ article/ articleDetail.jsp?id=7276.
15. Cbstat. 2008. V.5.10, Help Menu, By Kristian Linnet.
16. York, D., 1969. Least squares fitting of a straight line with correlated errors. Earth and Planetary Lett., 5: 320-324.
17. Saylor, R.D., E.S. Edgerton and B.E. Hartsell, 2006. Linear regression techniques for use in the EC tracer method of secondary organic aerosol estimation., Atmospheric Environment, 40: 7546 -7556.
18. Theil, H., 1950. A rank-invariant method of linear and polynomial regression analysis. Indagationes Mathematica, (12): 85-1.
19. Sprent, P., 1993. Applied Nonparametric Statistical Methods. London; New York: Chapman and Hall.
20. Hussain, S.S. and P. Sprent, 1983. Nonparametric Regression. J. the Royal Statistical Society, series A, 146: 182-191.
21. Nevitt, T. and H.P. Tam, 1998. A comparison of robust and nonparametric estimators under the simple linear regression model. Multiple Linear Regression Viewpoints, 25: 54-69.
22. Mutan, O.C., 2004. Comparison of Regression Techniques via Monte Carlo Simulation, A Thesis Submitted to the Graduate School of Natural and Applied Sciences of Middle East Technical University.

Appendix

OLS-bisector Regression Technique: The OLS-Bisector regression Technique simply defines the line that mathematically bisects the OLSYX and the OLSXY lines [9]. The slope coefficient of OLS-Bisector Technique can be calculated as in (1).

$$\hat{\beta}_{Bis} = \left(\hat{\beta}_1 + \hat{\beta}_2\right)^{-1}\left[\hat{\beta}_1\hat{\beta}_2 - 1 + \sqrt{\left(1+\hat{\beta}_1^2\right)\left(1+\hat{\beta}_2^2\right)}\right] \quad (1)$$

Here $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ is the slope of OLS(X|Y) regression and $\hat{\beta}_2 = \frac{S_{yy}}{S_{xy}}$ is the slope of OLS(Y|X) regression [10].

Major Axis (MA) Regression Technique: Unlike standard regression, the MA line does not depend on which variable is called "independent" and which "dependent." It always lies between the regression line of y on x and the regression line of x on y [11].

The calculation of the slope and the variance of this slope which want to be estimated by this technique are given in (2).

$$\hat{\beta}_{MA} = \frac{1}{2}\left[\left(\hat{\beta}_2 - \hat{\beta}_1^{-1}\right) + Sign\left(S_{xy}\right)\sqrt{4+\left(\hat{\beta}_2 - \hat{\beta}_1^{-1}\right)^2}\right] \quad (2)$$

Reduced Major Axis (RMA) Regression Technique: The reduced major axis regression was proposed to alleviate the scale dependency of orthogonal regression [12].

The calculation of the slope and the variance of this slope which want to be estimated by this technique are given in (3).

$$\hat{\beta}_{RMA} = Sign\left(S_{xy}\right)\left(\hat{\beta}_1\hat{\beta}_2\right)^{1/2} \tag{3}$$

Demingregression Technique: Deming 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 [13].

To estimate the regression line in Deming regression, the ë value, given in (4), must be calculated first:

$$\lambda = \frac{S_{ex}^2}{S_{ey}^2} \tag{4}$$

Here; S_{ex} and S_{ey} are the error variances of x and y values respectively.

The calculation of the slope in Deming regression is given in (5). The terms u,p and q are also given in (6).

$$\hat{\beta}_{DEM} = \frac{\left(\lambda q - u\right) + \sqrt{\left(u - \lambda q\right)^2 + 4\lambda p^2}}{2\lambda p} \tag{5}$$

$$\begin{aligned} u &= \sum\left(x_i - \bar{x}\right)^2 \\ q &= \sum\left(y_i - \bar{y}\right)^2 \\ p &= \sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \end{aligned} \tag{6}$$

Passing-Bablok Regression Technique: Passing and Bablok have proposed a linear regression procedure with no special assumptions regarding the distribution of the data. This nonparametric method is based on ranking the observations so it is computationally intensive. The result is independent of the assignment of the reference method as *X* (the independent variable) and the test method as *Y* (the dependent variable) [14].

The computation of the slope and the intercept are given in (7), (8) and (9).

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j} \qquad 1 \le i < j \le n \tag{7}$$

$$\hat{\beta}_{PB} = \begin{cases} b_{\left(\frac{N+1}{2}+k\right)} & \text{,if N is odd} \\ \frac{1}{2}\left(b_{\left(\frac{N}{2}+K\right)} + b_{\left(\frac{N}{2}+1+K\right)}\right) & \text{,if N is even.} \end{cases} \tag{8}$$

Here N is the sample size and K is the number of the values of b_{ij} with b_{ij}<-1.

$$a = med\{y_i - bx_i\} \tag{9}$$

The method takes measurement errors for both x and y into account, but the method presumes that the ratio between analytical standard deviations is related to the slope in a fixed manner Otherwise, a biased slope estimate arises. The method is not as efficient as the corresponding parametric procedures, i.e. Deming procedure [15].

York Regression Technique: As York [16] stated in his journal, this regression technique considers the errors in both variables. The slope in York regression which requires an iterative solution is given in (10), (11), (12) and (13).

$$b = \frac{\sum_{i=1}^{n} W_i\beta_i\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} W_i\beta_i\left(x_i - \bar{x}\right)} \tag{10}$$

Here;

$$W_i = \frac{w(x_i)w(y_i)}{w(x_i) + b^2 w(y_i) - 2br_i\sqrt{w(x_i)w(y_i)}} \tag{11}$$

$$\beta_i = W_i\left[\frac{x_i - \bar{x}}{w(y_i)} + \frac{b(y_i - \bar{y})}{w(x_i)} - \left(b(x_i - \bar{x}) + (y_i - \bar{y})\right)\frac{r_i}{\sqrt{w(x_i)w(y_i)}}\right] \tag{12}$$

and

$$\bar{x} = \frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i} \text{ and } \bar{y} = \frac{\sum_{i=1}^{n} W_i y_i}{\sum_{i=1}^{n} W_i} \tag{13}$$

Since W_i and β_i are functions of *b*, Eq.(10) must be solved iteratively. Given a set of weights $w(X_i)$ and $w(Y_i)$ and error correlation r_i for each data point, choose an

initial guess for b (possibly from either the OLSXY or orthogonal technique). Iterate through the following steps until successive values of b are within a predefined tolerance:

- Using b, $w(x_i)$, $w(y_i)$ and r_i, calculate W_i for each data point from (11)
- Using the observed points (x_i, y_i) and W_i, calculate $\bar{x}$ and $\bar{y}$ from (13)
- Calculate β_i for each data point from (12).
- Calculate a new estimate for b from (10)and return to step (1).
- The intercept, a, is then found by $a=\bar{y}-b\bar{x}$.

The York regression technique is thus very straightforward to implement and in our experience seldom requires more than 10 iterations (and usually much less) for convergence [17].

For all the regression techniques given above, the intercept term (except Passing-Bablok regression) can be calculated as in (14).

$$\hat{\alpha}_j = \bar{y} - \hat{\beta}_j \bar{x} \tag{14}$$

Here, $\bar{y}$ and $\bar{x}$ are the means of the y_i and x_i values respectively.

Theil Regression Technique: Theil's regression is a nonparametric method which is used as analternative to robust methods for data sets with outliers. Although thenonparametric procedures perform reasonably well for almost any possibledistribution of errors and they lead to robust regression lines, they require a lot ofcomputation. This method is suggested by Theil [18] and it is proved to beuseful when outliers are suspected, but when there are more than few variables, the application becomes difficult.

Sprent [19] states that for a simple linear regression model to obtain theslope of a line that fits the data points, the set of all slopes of lines joining pairs ofdata points.

(xi,yi), (xj,yj), xj≠xi, for $1 \le i < j \le n$ should be calculated as

Hussain and Sprent [20] say that no generality is losen if we take$1 \le i < j \le n$ assuming that the x_i are arranged in ascending order. Note that$b_{ij}=b_{ji}$. According to these results the theTheil's slope estimator is $\hat{\theta}_1 = med\{b_{ij} \mid x_j \neq x_i\}$ where $x_1 \le x_2 \le \ldots \le x_n$. It is known that median estimators are less affected compared to the meanestimators.

Therefore, these estimators are resistant to outliers in the sample data.Nevitt and Tam [21] state that there are several methods for computingthe y-intercept. One of these methods is to calculate

$$a_{ij} = \frac{x_j y_i - x_i y_j}{x_j - x_i} \quad i < j, \; x_i \neq x_j,$$

and taking the median of these a_{ij} values will give us the y-intercept [22].