ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2016; 7(2);296-308

Comparison of Classical Linear Regression and Orthogonal Regression with Respect to the Sum of Squared Perpendicular Distances

Dik Uzaklıklar Kareler Toplamına Göre Klasik Doğrusal Regresyon ile Ortogonal Regresyonun Karşılaştırılması*

Taliha KELEŞ** Murat ALTUN***

Abstract
Regression analysis is a statistical technique for investigating and modeling the relationship between variables. The purpose of this study was the trivial presentation of the equation for orthogonal regression (OR) and the comparison of classical linear regression (CLR) and OR techniques with respect to the sum of squared perpendicular distances. For that purpose, the analyses were shown by an example. It was found that the sum of squared perpendicular distances of OR is smaller. Thus, it was seen that OR line has appeared to present a much better fit for the data than CLR line. Depending on those results, the OR is thought to be a regression technique to obtain more accurate results than CLR at simple linear regression studies.

Keywords: orthogonal regression equation, perpendicular distance, the sum of squared perpendicular distances

Öz
Regresyon analizi, değişkenler arasındaki ilişkiyi inceleme ve modellemede kullanan istatistiksel bir tekniktir. Bu çalışmanın amacı ortogonal regresyon (OR) eşitliğini açık bir şekilde sunmak ve dik uzaklıklar kareleri toplamına göre klasik doğrusal regresyon (KDR) ile ortogonal regresyonu karşılaştırmaktır. Bu amaçla analizler bir örnek üzerinden gösterilmiştir. Araştırma sonucunda ortogonal regresyonun dik uzaklıklar kareler toplamının daha küçük olduğu bulunmuştur. Buradan bağımlı ve bağımsız değişkenler arasındaki doğrusal ilişkiyi, ortogonal regresyon doğrusunun klasik doğrusal regresyon doğrusundan daha iyi temsil ettiği görülmüştür. Bu sonuçlara bağlı olarak basit doğrusal regresyon çalışmalarında OR tekniğinin KDR tekniğinden daha doğru sonuçlar elde etmede kullanılabilecek bir regresyon tekniği olduğu düşünülmektedir.

Anahtar Kelimeler: ortogonal regresyon eşitliği, dik uzaklıklar, dik uzaklıklar kareler toplamı

INTRODUCTION

Regression analysis is a statistical technique for investigating and modeling the relationship between dependent and independent variables (Montgomery, Peck & Vining, 2012; Sykes, 1993). Applications of regression analysis exist in almost every field, including engineering, physical and chemical sciences, education, economics, management, astronomy, medical sciences, political science, life and biological sciences, and social sciences. In fact, regression analysis may be the most widely used statistical technique (Montgomery et al., 2012).

The most fundamental aim of simple linear regression analysis is to predict the values of the dependent variable, Y, when the values of the independent variable, X, is known in order to write line, curve and surface equations which represent a cloud of the data. This prediction is made according to the regression line and the curve (Lane, 2016).

* This study was presented at The 2nd International Researches Statisticians and Young Statisticians Congress on May 4-8 2016, Ankara-Turkey.

** Dr., Provincial Directorate of National Education, Strategy Development Service, R&D Department, Bursa-Turkey, e-mail: talihak@hotmail.com
*** Prof. Dr., Uludag University, Faculty of Education, Department of Elementary Mathematics Education, Bursa-Turkey, e-mail: maltun@uludag.edu.tr

DOI: 10.21031/epod.277885
Received: 20.07.2016
Accepted: 03.11.2016

The equation of $\hat{Y} = f(X)$ is used to find the predicted values of dependent variable $\hat{Y}$ from the values of independent variable (X). The formula for linear regression line is

$$\hat{Y} = a + bX$$

Where *a* is the intercept, *b* is the slope of the line.

As illustrated in Figures 1a, classical linear regression (CLR) equation is attained by minimizing the sum of squares of vertical distances from the data points to the regression line (Ortiz, Pogliani & Besalú, 2010; Ding, Chu, Jin & Zhu, 2013; Elfessi & Hoar, 2001; Isobe, Feigelson, Akritas & Babu, 1990; Kane & Mroch, 2010; Leng, Zhang, Kleinman & Zhu, 2007; Ludbrook, 2010).

The sum of squares of vertical distances from data points to the regression line, $E = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - bx_i - a)^2$ is to be minimized, where e_i is called the i^{th} residual (Calzada & Scariano, 2003; Ortiz et al., 2010; Glaister, 2005; Scariano & Barnet, 2003).

As illustrated in Figures 1b, orthogonal regression (OR) equation is attained by minimizing the sum of the squares of perpendicular distances between the data points and the regression line (Carr, 2012; Ortiz et al., 2010; Ding et al., 2013; Elfessi & Hoar, 2001; Isobe et al., 1990; Kane & Mroch, 2010; Leng et al., 2007; Li, 1984; Ludbrook, 2010; Nievergelt, 1994; Scariano & Barnet, 2003). The sum of the squares of perpendicular distances between the data points and the regression line, $E_{\perp} = \sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} \frac{(y_i - bx_i - a)^2}{b^2 + 1}$ is to be minimized, where d_i is perpendicular distance (Calzada & Scariano, 2003; Ortiz et al., 2010; Glaister, 2005; Li, 1984; Scariano & Barnet, 2003).

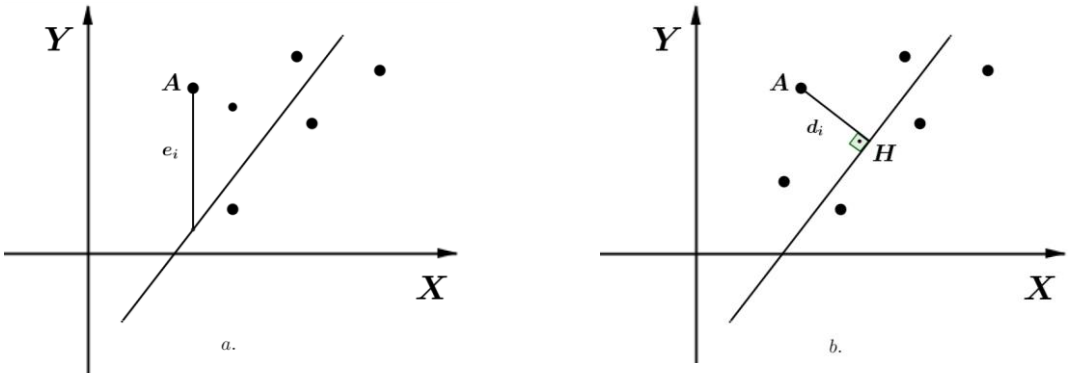

Figure 1. Classical Linear Regression (a), Orthogonal Regression (b)

OR identifies the line that minimizes the sum of squared perpendicular distances of the data points from the line (Carroll & Ruppert, 1996; Isobe et al., 1990; Kane & Mroch, 2010). OR has been known by various names; for example, it is known as the errors-in-variables (Carroll & Ruppert, 1996; Fišerová & Hron, 2010; Markovsky & Van Huffel, 2007), total least squares (Calzada & Scariano, 2003; Elfessi & Hoar, 2001; Golub & Van Loan, 1980; Markovsky & Van Huffel, 2007; Nievergelt, 1994), least squares (Adcock, 1878), major axis (Carr, 2012; Isobe et al., 1990; Kermack & Haldane, 1950; Ludbrook, 2010), measurement errors (Markovsky & Van Huffel, 2007), orthogonal error regression, and orthogonal least squares regression (Carr, 2012). In addition, OR is called a method of moments estimator in 1987 by Fuller (Carroll & Ruppert, 1996).

The method of OR was first discovered by Adcock in 1878 (Adcock, 1878) and later by Person (Isobe et al., 1990; Pearson, 1901). The method of OR has been rediscovered many times, often

independently (Ding et al., 2013). This method has been understood for well over a hundred years, but CLR technique is preferred to orthogonal regression technique in the statistical community because of its easy computation (Ludbrook, 2010). CLR is the most popular method for linear regression (Ding et al., 2013). Computational formulas for orthogonal regression line can be somewhat quite complicated (Calzada & Scariano, 2003; Carr, 2012; Carroll & Ruppert, 1996; Ortiz et al., 2010; Isobe et al., 1990; Kane & Mroch, 2010; Scariano & Barnet, 2003). Most literature on this method is necessarily brief and heavily symbolic (Calzada & Scariano, 2003; Carr, 2012; Kermack & Haldane, 1950; Nievergelt, 1994).

Simple linear regression analysis is used to define the nearest line to the data points (Akdeniz, 2013; Genç, Sertkaya & Demirtaş, 2003; Saraçlı, Doğan & Doğan, 2009). The shortest distance from a point to a line is perpendicular distance (Warton, Wright, Falster & Westoby, 2006). Thus, it is the perpendicular distance that is important, and the statistical distance in this case is the shortest distance to the line. The orthogonal regression is the line that minimizes the sum of squares of the shortest distances from the data points to the line.

One of the assumptions of classical linear regression, is that the independent variable can not include any measurement error, indicating that the only source of the error term is the dependent variable (Glaister, 2005; Scariano & Barnet, 2003). However, Deming (1943), indicated in his book "Statistical Adjustment of Data" in that independent variable may include measurement error in practice (Saraçlı, Doğan & Doğan, 2009). Besides, Glaister (2005) and Stöckl, Dewitte and Thienpont (1998) stated that both dependent and independent variables are subject to measurement error in practice. In this case, the estimates in CLR are no longer accurate (Glaister, 2005) in practice. The estimates in OR technique, when there is a measurement error in both variables (Carr, 2012; Glaister, 2005; Scariano & Barnet, 2003), are more accurate. Since the independent variable may include measurement error in measurement error models, OR technique may be more appropriate and may give better results in those situations (Calzada & Scariano, 2003; Carr, 2012; Warton et al., 2006).

A review of international literature indicated that considerable researches have been conducted analyses to compare OR with other regression techniques (Calzada & Scariano, 2003; Carr, 2012; Ding et al., 2013; Elfessi & Hoar, 2001; Glaister, 2005; Leng et al., 2007; Lolli & Gasperini, 2012; Lundbrook, 2010; Ortiz et al., 2010). Calzada and Scariano (2003) had studied on contrasting the ordinary and orthogonal (or total least squares) regression techniques using real data. They found that OR is better than ordinary least squares regression in this example. They suggested that there is no clear cut answer as to whether the ordinary least squares or orthogonal regression technique should be preferred in a given application. The method chosen by the researcher must be based on a keen understanding of the data as well as the sources of the errors present in the observations. Carr (2012) had compared three of such linear regression methods (classical linear regression, orthogonal regression, and reduced major axis) using geyser eruption data. He found that mean square error of classical linear regression is the smallest. He stated that classical linear regression has better than orthogonal regression and reduced major axis regression. Ding et al. (2013) had investigated and compared least squares and orthogonal regression in measurement error modeling for prediction of material property. They stated that orthogonal regression has better performance than classical linear regression with respect to standard deviation of residuals. Ortiz et al. (2010) had reviewed classical linear regression and orthogonal regression techniques in two variable linear regressions. They used an example with 18 data. They stated that standard deviation of the residues of orthogonal regression is smaller than the classical linear regression implying that it is a better method than classical linear regression.

The method comparison studies carried out in Turkey were analyzed. It is remarkable that very few studies have been done in this regard. Bland-Altman and OLS Bisector (Saraçlı & Çelik, 2012), Deming regression technique and ordinary least squares (Saraçlı, Doğan & Doğan, 2009), least squares and robust M estimation methods (Coşkuntuncel, 2013), least squares and least median squares estimation methods (Alma & Vupa, 2008), and simple linear regression and hierarchical linear model (Atar, 2010) techniques of method comparison studies have been conducted.

In Turkey, there are few studies about orthogonal regression in geophysics (Öztürk, 2012) and statistic (Saraçlı, 2011). Öztürk (2012) had made an analysis on the application of four different statistical regression methods for different data sets with respect to the determination of coefficient (R^2). It was showed that the representation of empirical relationships will be more suitable and reliable by Least Sum of Absolute Deviations or Robust regressions for clustered samples whereas by Least Squares or Orthogonal regressions for scattered data. Saraçlı (2011) had examined that data sets with different sample sizes simulated via Monte Carlo simulation were used to see the performances of Type II regression techniques (Ordinary least square (OLS) bisector, major axis, reduced major axis, Deming regression and passing bablok regression techniques) by mean square error. The result of his study concluded that the OLS bisector technique gave the best results in all conditions with different distribution types, different sample sizes, and the data set with or without an outlier.

In comparison studies published both in national and international literature, it has been seen that CLR (Carr, 2012) and ordinary least square (OLS) bisector (Saraçlı, 2011) demonstrated the superiority with respect to mean square error. In addition to those studies, OR technique demonstrated the superiority with respect to determination of coefficient (R^2) (Calzada & Scariano, 2003) or standard deviation of residuals (Ding et al., 2013; Ortiz et al., 2010).

Although the benefits of OR technique has been discussed, it has been seen that researches in social sciences and education field are using only the CLR technique in Turkish literature (Bağçeci, Döş & Sarıca, 2011; Bahar, 2006; Bayat, Şekercioğlu & Bakır, 2014; Doruk, Özdemir, & Kaplan, 2015; İlğan, Erdem, Yapar, Aydın & Aydemir, 2012; Kesicioğlu & Güven, 2014; Üredi & Üredi, 2005;). In Turkey, there is no study about orthogonal regression in social sciences and education. In this respect, it is expected that this study sheds light on studies that can be done in the future about this subject.

This Study Aim

The purpose of this study is the trivial presentation of the equation for orthogonal regression (OR) and the comparison of classical linear regression (CLR) and OR according to the sum of squared perpendicular distances.

METHOD

Classical linear regression and orthogonal regression were compared with respect to the sum of squared perpendicular distances. For this purpose, this analysis was shown by an example. For this example, data was taken from Akdeniz's book (Akdeniz, 2013).

We restricted our study to CLR and OR. Classical linear regression and orthogonal regression are compared with respect to the sum of squared perpendicular distances which is defined with the following formulation,

$$E_{\perp} = \sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} \frac{(y_i - bx_i - a)^2}{b^2 + 1}$$

Microsoft Excel 2010 was used for data organization and calculation of OR equation and sum of the squared perpendicular distances of OR and CLR. SPSS 17.0 was used for the calculation of CLR equation.

The equations for the slope and intercept of orthogonal regression were calculated in detail. The derivation of these equations is presented as follows.

Equation Derivation for OR Line

Herein, OR is explained in simple, straightforward way to facilitate easy understanding for equations.

From Figure 1b, we see that $|AH| = d_i = \frac{|y_i - bx_i - a|}{\sqrt{b^2+1}}$

The sum of squares of the perpendicular distances is equal to $E_\perp = \sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} \frac{(y_i - bx_i - a)^2}{b^2+1}$

where n is the sample size for the data. Derivatives with respect to a and b are calculated and equaled to zero.

$\frac{\partial E_\perp}{\partial a} = 0$ and $\frac{\partial E_\perp}{\partial b} = 0$

The first step is taking the partial derivative of $E_\perp$ with respect to the regression intercept, a,

$$\frac{\partial E_\perp}{\partial a} = \sum_{i=1}^{n} \frac{-2(y_i - bx_i - a)(b^2+1) - 0}{(b^2+1)^2} = 0$$

Multiplying both sides of this equation $(b^2+1)^2$ clears the denominator,

$$-2(b^2+1)\sum_{i=1}^{n}(y_i - bx_i - a) = 0$$

$$\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i - na = 0$$

$$\sum_{i=1}^{n} y_i = b\sum_{i=1}^{n} x_i + na \quad \text{and} \quad a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

Similarly, next step is putting this result aside and taking the partial derivative of $E_\perp$ with respect to the slope of regression, *b*,

$$\frac{\partial E_\perp}{\partial b} = \sum_{i=1}^{n} \frac{-2x_i(y_i - bx_i - a)(b^2+1) - 2b(y_i - bx_i - a)^2}{(b^2+1)^2} = 0$$

Multiplying both sides of this equation $(b^2+1)^2$ clears the denominator,

$$\sum_{i=1}^{n}\left[(b^2+1)(y_i\, x_i - bx_i^{\,2} - ax_i) + b(y_i^{\,2} + b^2x_i^{\,2} + a^2 - 2bx_iy_i - 2ay_i + 2abx_i)\right] = 0$$

$$\sum_{i=1}^{n}\left[(b^2+1-2b^2)\, x_i\, y_i + (-b^3 - b + b^3)x_i^{\,2} + a(b^2-1)x_i + by_i^{\,2} - 2aby_i + a^2b\right] = 0$$

$(-b^2+1)\sum_{i=1}^{n} x_i\, y_i - b\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} y_i^{\,2} + a(b^2-1)\sum_{i=1}^{n} x_i - 2ab\sum_{i=1}^{n} y_i + na^2b = 0$ Multiplying each side of the equation by -1

$(b^2-1)\sum_{i=1}^{n} x_i\, y_i + b\sum_{i=1}^{n} x_i^{\,2} - b\sum_{i=1}^{n} y_i^{\,2} - a(b^2-1)\sum_{i=1}^{n} x_i + 2ab\sum_{i=1}^{n} y_i - na^2b = 0$

$(b^2-1)\sum_{i=1}^{n} x_i\, y_i + b(\sum_{i=1}^{n} x_i^{\,2} - \sum_{i=1}^{n} y_i^{\,2}) - a(b^2-1)\sum_{i=1}^{n} x_i + 2ab\sum_{i=1}^{n} y_i - na^2b = 0$

$$\sum_{i=1}^{n} y_i = b\sum_{i=1}^{n} x_i + na \text{ and } \quad a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

Substituting the solution for a into the equation for above equation gives

$$(b^2-1)\sum_{i=1}^{n} x_i y_i + b(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i^2) - (b^2-1)\sum_{i=1}^{n} x_i \left[\frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}\right] + 2b\sum_{i=1}^{n} y_i \left[\frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}\right] - nb\left[\frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}\right]^2 = 0$$

Rearrangement of terms yields a quadratic function of b,

$$b^2\left[n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right] + b\left[n(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i)^2 + (\sum_{i=1}^{n} y_i)^2\right] - \left[n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right] = 0$$

$$b_{1,2} = \frac{-\left[n(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i)^2 + (\sum_{i=1}^{n} y_i)^2\right] \mp \sqrt{\left[n(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i)^2 + (\sum_{i=1}^{n} y_i)^2\right]^2 + 4\left[n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right]^2}}{2\left[n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right]}$$

$$b_{1,2} = \frac{\left[(\sum_{i=1}^{n} x_i)^2 - (\sum_{i=1}^{n} y_i)^2 - n(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i^2)\right] \mp \sqrt{\left[(\sum_{i=1}^{n} x_i)^2 - (\sum_{i=1}^{n} y_i)^2 - n(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i^2)\right]^2 + 4\left[n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right]^2}}{2\left[n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right]}$$

We can write regression equation $\hat{Y} = a + bX$. Value b is the slope of the regression equation. There are two different real roots of b. Scatter diagrams are drawn. Value b which is the same sign of slope of draft line is selected.

FINDINGS

For comparison of OR and CLR techniques, the analyses were conducted using an example.

A simple example

The relationship between the hardness and the durability of a material produced in a production center was investigated. 10 pieces of substances produced for this purpose were selected. Then, the hardness and durability test was conducted. The data obtained is as in Table 1.

Table 1. Example Data

Piece No	**X (**Hardness)	**Y** (Durability)
1	7	10
2	9	12
3	5	6
4	8	9
5	6	8
6	9	11
7	7	10
8	4	5
9	8	10
10	7	9

Table 1 displays data. Table 2 shows quantities needed for calculation of regression equation.

Table 2. Quantities Needed for Calculation of Regression Equation

Piece No	X(Hardness)	Y(Durability)	X^2	Y^2	XY
1	7	10	49	100	70
2	9	12	81	144	108
3	5	6	25	36	30
4	8	9	64	81	72
5	6	8	36	64	48
6	9	11	81	121	99
7	7	10	49	100	70
8	4	5	16	25	20
9	8	10	64	100	80
10	7	9	49	81	63
Total	70	90	514	852	660

We want to predict the values of durability from the values of hardness according to Table 2 based on the following equation.

$$\hat{Y} = a + bX$$

$$90 = 10a + 70b \text{ or } 9 = a + 7b$$

$$b_{1,2} = \frac{180 \mp \sqrt{(180)^2 + 4(300)^2}}{2(300)} = \frac{180 \mp \sqrt{392400}}{600} \qquad b_1 = 1.34 \text{ and } b_2 = -0.74$$

Table 3 shows the regression equations.

Table 3. Regression Equations

Regression Equations	Slope, b	Intercept, a	$\hat{Y} = a + bX$	Correlation Coefficient, r
OR line	$b_1 = 1.34$	$a_1 = -0.38$	$\hat{Y} = -0.38 + 1.34X$	0.945
	$b_2 = -0.74$	$a_2 = 14.18$	$\hat{Y} = 14.18 - 0.74X$	
CLR line	$b = 1.25$	$a = 0.25$	$\hat{Y} = 0,25 + 1,25X$	0.945

Table 3 displays the Pearson correlation coefficient (r=0.945, r^2=0.893) indicating a good linear fit. The Pearson correlation between Hardness and Durability is 0.945, indicating that Hardness explains about 89.3 % of the variance in Durability.

OR line has equation $\hat{Y} = -0.38 + 1.34X$, whereas CLR line has equation $\hat{Y} = 0.25 + 1.25X$.

OR line is $Durability = -0.38 + 1.34 * Hardness$ while the CLR line is $Durability = 0.25 + 1.25 * Hardness$.

OR and CLR lines are shown in Figure 2 for data.

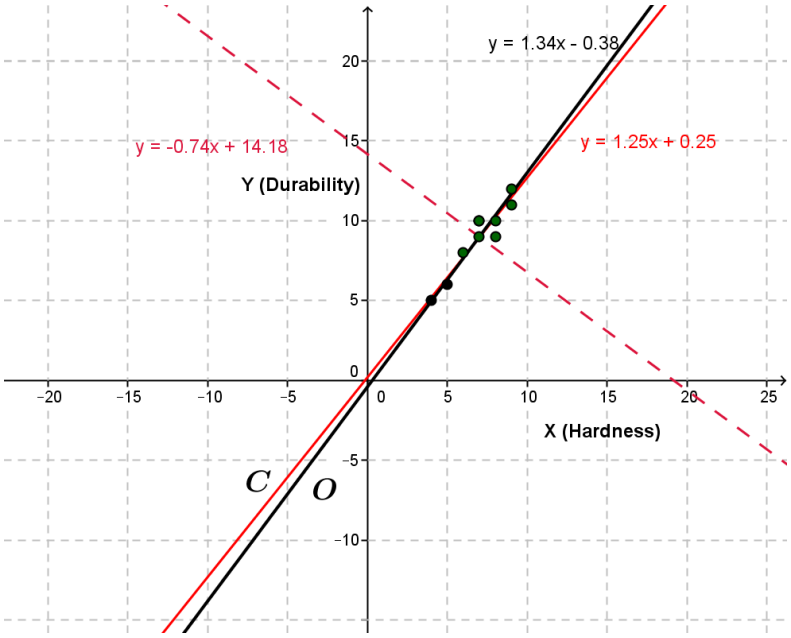

Figure 2. CLR Line (C: $\hat{Y} = 0.25 + 1.25X$) and OR Line (O: $\hat{Y} = -0.38 + 1.34X$) Applied to the Hardness and Durability of Substances Data

Table 4 shows the sum of squared perpendicular distances of regression lines.

Table 4. The Sum of Squared Perpendicular Distances of Regression Lines

Regression Lines	CLR Line C: $\hat{Y} = 0.25 + 1.25X$	OR Line O: $\hat{Y} = -0.38 + 1.34X$
$E = \sum_{i=1}^{n} (y_i - bx_i - a)^2$	4.50	4.69
Sum of Squares Perpendicular Distances $E_{\perp} = \sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} \frac{(y_i - bx_i - a)^2}{b^2 + 1}$	$\frac{4.5}{(1.25)^2 + 1} = 1.75$	$\frac{4,69}{(1.34)^2 + 1} = 1.67$

Presented in Table 4, CLR line and OR line give the values of $E_{\perp}$, 1.75 and 1.67 respectively.

We can see that OR line has smaller sum of squared perpendicular distances than CLR line which implies it has better performance than CLR in the prediction of durability from hardness of substances in this example.

RESULTS AND DISCUSSION

Regression analysis is expressed in a mathematical equation representing the relationship between variables. The effect of independent variable on dependent variable can be predicted via that mathematical equation (Büyüköztürk, Çokluk & Köklü, 2012). The fundamental purpose of simple linear regression analysis is to determine the best method that predicts the dependent variable. The

purpose of this study was the trivial presentation of the equation for OR and the comparison of OR and CLR according to the sum of squared perpendicular distances.

As Akdeniz (2013), Genç et al. (2003) and Saraçlı et al. (2009) stated, simple linear regression analysis is used to define the nearest line to the data points. As a result of comparison of the value of $E_{\perp}$ attained from two regression equations, it is found that value of $E_{\perp}$ for OR line is smaller. This result shows that OR has better performance than CLR in the prediction of durability from hardness of substances. OR line is closer to points than CLR line. OR line appears to present a much better fit for the data. This result is also similar to the finding from Calzada and Scariano (2003), Ding et al. (2013), Glaister (2005) and Ortiz et al. (2010). Although, OR is a better method than classical linear regression with respect to the sum of squared perpendicular distances, this result contradicts with the results of Carr (2012) and Saraçlı (2011).

Glaister (2005), Scariano and Barnet (2003), Deming (1943), Stöckl et al. (1998) stated that both dependent variable and independent variable are subject to measurement error in practice. Then, the result of our study indicates that it is precisely these circumstances for which the statistical distance to be minimized is the shortest distance to the line, and the appropriate regression line is OR line as shown in Table 4. This finding is also similar to the finding from Glaister (2005).

This research has revealed that OR is a better method than CLR with respect to the sum of squared perpendicular distances. It is more accurate to use OR technique, when there are measurement errors in both dependent variable and independent variable.

Depending on these results, the OR is thought to be a regression technique to obtain more accurate results than CLR at simple linear regression studies.

REFERENCES

Adcock, R.J. (1878). A Problem in least squares, *Annals of Mathematics*, 5(2), 53-54.

Akdeniz, F. (2013). *Olasılık ve İstatistik*, Nobel Kitabevi, 18. Press, Ankara, 459-492.

Alma, Ö.G., & Vupa, Ö. (2008). Regresyon Analizinde kullanılan En Küçük Kareler ve En küçük Medyan Kareler Yöntemlerinin Karşılaştırılması, *Süleyman Demirel Üniversitesi Fen Edebiyat Fakültesi Fen Dergisi*, 3(2), 219-229.

Atar, B. (2010). Basit Doğrusal Regresyon Analizi ile Hiyerarşik Doğrusal Modeller Analizinin Karşılaştırılması, *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi,* 1(2), 78-84.

Bahar, H.H. (2006). An Evaluation of KPSS Scores According to Grade Point Avarege and Gender, *Education and Science*, 31(140), 68-74.

Bağçeci, B., Döş, B., & Sarıca, R. (2011). An Analysis of Metacognitive Awareness Levels and Academic Achievement of Primary School Students, *Mustafa Kemal University Journal of Social Sciences Institute*, 8(16), 551-566.

Bayat, N., Şekercioğlu, G., & Bakır, S. (2014). Okuduğunu Anlama ve Fen Başarısı Arasındaki İlişkinin Belirlenmesi, *Eğitim ve Bilim, (tedmem),* 39(176), 457-466.

Büyüköztürk, Ş., Çokluk, Ö., & Köklü, N. (2012). *Sosyal bilimler için İstatistik.* Ankara: Pegem Akademi.

Calzada, M.E., & Scariano, S.M. (2003). Contrasting total least squares with ordinary least squares part II: Examples and Comparisons, *Mathematics and Computer Education*, 37(2), 159-174.

Carr, J.R. (2012). Orthogonal regression: A Teaching perspective, *International Journal of Mathematical Education in Science and Technology*, 43(1), 134-143.

Carroll, R.J., & Ruppert D. (1996). The use and misuse of orthogonal regression in linear errors-in-variables models, *The American Statistician*, 50 (1), 1-6.

Coşkuntuncel, O. (2013). The Use of alternative Regression Methods in Social Sciences and the Comparison of Least Squares and M Estimation Methods in Terms of the Determination of Coefficient, *Educational Sciences: Theory & Practice*, 13(4), 2139-2158.

Deming, E.W. (1943). *Statistical Adjustment of Data*, Dover Publications, Inc., New York.

Ding, G., Chu, B., Jin, Y., & Zhu, C. (2013). Comparison of orthogonal regression and least squares in measurement error modeling for prediction of material property, *Nanotechnology and Material Engineering Research. Advanced Materials Research,* 661, 166-170.

Doruk, M., Özdemir, F., & Kaplan, A. (2015). The Relationship Between Prospective Mathematics Teachers' Conceptions on Constructing Mathematical Proof and Their Self-Efficacy Beliefs Towards Mathematics, *Kastamonu Üniversitesi Kastamonu Eğitim Dergisi*, 23(2), 861-874.

Elfessi, A., & Hoar, R.H. (2001). Simulation study of a linear relationship between two variables affected by errors, *Journal of Statistical Computation and Simulation*, 71(1), 29-40.

Fišerová, E., & Hron, K. (2010). Total least squares solution for compositional data using linear models, *Journal of Applied Statistics*, 37(7), 1137-1152.

Genç, Y., Sertkaya, D., & Demirtaş, S. (2003). Klinik Araştırmalarda İki Ölçüm Tekniğinin Uyumunu İncelemede Kullanılan İstatistiksel Yöntemler. *Ankara Üniversitesi Tıp Fakültesi Mecmuası*, 56(1), 1-6.

Glaister, P. (2005). The use of orthogonal distances in generating the total least squares estimate, *Mathematics and Computer Education*, 39(1), 21-30.

Golub, G.H., & Van Loan, C.F. (1980). An Analysis of the total least squares problem, *SIAM Journal on Numerical Analysis*, 17(6), 883-893.

İlğan, A., Erdem, M., Yapar, B., Aydın. S., & Aydemir, Ş.Ş. (2012). Parents Interest and Regression Level of Primary State School Students Level Determination Exam (SBS) Score, *Journal of Educational Sciences Research*, 2(2), 1-17.

Isobe, T., Feigelson, E.D, Akritas, M.G., & Babu, G.J. (1990). Linear regression in astronomy I, *The Astrophysical Journal*, 364, 104-113.

Kane, M.T., & Mroch, A.A. (2010). Modeling group differences in OLS and Orthogonal Regression: Implications for differential validity studies, *Applied Measurement in Education*, 23, 215-241.

Kermack, K.A., & Haldane, J.B.S. (1950). Organic correlation and allometry, *Biometrika Trust*, 37(1), 30–41.

Kesicioğlu, O.S., & Güven, G. (2014). Investigation of the Correlation Between Preservice Early Childhood Teachers' Self-Efficacy Levels and Problem Solving, Empathy and Communication Skills, *Turkish Studies, International Periodical For the Languages, Literature and History of Turkish or Turkic*, 9(5), 1371-1383.

Lane, D. M. (2016). "Introduction to Linear Regression", http://onlinestatbook.com/2/regression/intro.html Accessed 21 February 2016.

Leng, L., Zhang, T, Kleinman, L., & Zhu, W. (2007). Ordinary Least Square Regression, Orthogonal Regression, Geometric Mean Regression and Their Applications in Aerosol Science, *Journal of Physics, Conference Series* 78(1), 1-5, http://iopscience.iop.org/1742-6596/78/1/012084 Accessed 17 March 2016.

Li, H.C. (1984). Generalized problem of least squares, *The American Mathematical Monthly*, 91(2), 135-137.

Lolli, B., & Gasperini, P. (2012). A comparison among general orthogonal regression methods applied to earthquake magnitude conversions, *Geophysical Journal International*, 190(2), 1135-1151.

Ludbrook, J. (2010). Linear regression analysis for comparing two measurers or methods of measurement: But which regression?, *Clinical and Experimental Pharmacology and Physiology*, 37, 692-699.

Markovsky, I., & Van Huffel, S. (2007). Overview of total least- squares methods, *Signal Processing*, 87, 2283-2302.

Montgomery D.C., Peck E.A., & Vining G.G. (2012). Introduction to Linear Regression Analysis, 5th Edition, 1-11. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470542810.html Accessed 11 February 2016.

Nievergelt, Y. (1994). Total least squares: State-of-the-art regression in numerical analysis, *SIAM Review (Society for Industrial and Applied Mathematics)*, 36(2), 258–264.

Ortiz, J.V., Pogliani, L., & Besalú, E. (2010). Two-variable linear regression: Modeling with orthogonal least-squares analysis, *Journal of Chemical Education*, 87(9), 994-995.

Öztürk, S. (2012). İstatistiksel Regresyon Yöntemlerinin Farklı Veri Gruplarına Uygulanması Üzerine Bir Analiz, *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 2(2), 55-67.Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, 2, 559-572.

Saraçlı, S. (2011). Tip II Regresyon Tekniklerinin Monte-Carlo Simülasyonu ile Karşılaştırılması, *e-Journal of New World Sciences Academy*, 6(2), 26-35.

Saraçlı, S., & Çelik, H.E. (2012). Metot Karşılaştırma Çalışmalarında Bland-Altman ve Tip II Regresyon Analizinin Karşılaştırılması, *Düzce Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi*, 2(1), 11-14.

Saraçlı, S., Doğan, İ., & Doğan, N. (2009). Medikal Metod Karşılaştırma Çalışmalarında Deming Regresyon Tekniği, *Turkiye Klinikleri J Biostat*, 1(1), 9-15.

Scariano, S.M., & Barnet, II.W. (2003). Contrasting total least squares with ordinary least squares part I: Basic ideas and result, *Mathematics and Computer Education*, 37(2), 141-158.

Sykes, A.O. (1993). An Introduction to Regression Analysis, *Coase-Sandor Institute for Law & Economics Working*, 20, 1-34.

Stöckl, 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?, *Clinical Chemistry*, 44(11), 2340–2346.

Üredi, I., & Üredi, L. (2005). İlköğrtim 8. Sınıf Öğrencilerinin Öz-düzenleme Stratejileri ve Motivasyonel İnançlarının Matematik Başarısını Yordama Gücü, *Mersin University Journal of the Faculty of Education*, 1(2), 250-260.

Warton, D.I., Wright, I.J., Falster, D.S. & Westoby, M. (2006). Bivariate line-fitting methods for allometry, *Biological Reviews*, 81(2), 259-291.

GENİŞ ÖZET

Giriş

Regresyon analizi, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi inceleme ve modellemede kullanan istatistiksel bir tekniktir (Montgomery vd., 2012; Sykes, 1993). Regresyon analizin uygulamaları tıpta, mühendislikte, ekonomide, astronomide, sosyal ve biyolojik bilimlerde, fen bilimlerinde ve eğitim bilimlerinde, kısacası hemen hemen her alanda görülmektedir. Aslında regresyon analizi en yaygın kullanılan istatistiksel tekniklerden biridir (Montgomery vd., 2012).

İstatistiğin önemli tahminleme (yordama) tekniklerinden biri olan regresyon analizinin en temel amacı; veri bulutunu temsil edebilecek bir doğru, eğri ya da yüzey denklemi yazmak için bu denklem üzerinden bağımsız değişkenin bilinmesi halinde bağımlı değişkenin ne olabileceğini tahmin etmektir. Bu tahminleme (yordama) regresyon doğrusu veya eğrisi üzerinden yapılmaktadır (Lane, 2016). Gözlenen x (bağımsız) değişkeniyle eşlenen $\hat{Y}$ (bağımlı) değerinin bulunabilmesi için $\hat{Y} = f(X)$ eşitliği kullanılmaktadır. Doğrusal regresyon doğrusu $\hat{Y} = a + bX$ ile gösterilir ki buradaki $\hat{Y}$, tahmin edilen değerdir. Klasik doğrusal regresyon (KDR) denklemi gözlenen verilerin regresyon doğrusuna olan düşey uzaklıklarının kareleri toplamı minimize edilerek elde edilir (Ortiz vd., 2010; Ding vd., 2013; Elfessi ve Hoar, 2001; Isobe vd., 1990; Kane ve Mroch, 2010; Leng vd., 2007; Ludbrook, 2010). Şekil 1a'da görüldüğü gibi verilerin regresyon doğrusuna olan düşey uzaklıklarının kareler toplamı,

$$E = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (\mathrm{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (\mathrm{y}_i - bx_i - a)^2 \text{ dir.}$$

Ortogonal regresyon (OR) ise gözlenen noktaların tahmin edilen regresyon doğrusuna olan dik uzaklıklarının kareleri toplamı minimize edilerek bulunur (Carr, 2012; Ortiz vd., 2010; Ding vd., 2013; Elfessi ve Hoar, 2001; Isobe vd., 1990; Kane ve Mroch, 2010; Leng vd., 2007; Li, 1984; Ludbrook, 2010; Nievergelt, 1994; Scariano ve Barnet, 2003).

Şekil 1b'de görüldüğü gibi verilerin regresyon doğrusuna dik uzaklıklarının kareleri toplamı,

$$E_{\perp} = \sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} \frac{(y_i - bx_i - a)^2}{b^2 + 1} \text{ dir.}$$

OR farklı isimlerle bilinmektedir. Bunlardan bazıları, toplam en küçük kareler (total least squares) (Calzada ve Scariano, 2003; Elfessi ve Hoar, 2001; Golub ve Van Loan, 1980; Markovsky ve Van Huffel, 2007; Nievergelt 1994), değişkenlerde hatalar (errors-in-variables) (Carroll ve Ruppert, 1996; Fišerová ve Hron, 2010; Markovsky ve Van Huffel, 2007) ve major eksen (major axis) (Carr, 2012; Isobe vd., 1990; Kermack ve Haldane, 1950; Ludbrook, 2010) dir.

OR metodu ilk olarak 1878 yılında Adcock tarafından keşfedilmiştir (Adcock, 1878). Bu metot bir yüzyıldır bilinmesine rağmen klasik doğrusal regresyon kullanılmaktadır. Klasik regresyonun tercih edilmesinin temel nedeni hesaplanmasının kolaylığıdır (Ludbrook, 2010).

Doğrusal regresyon analizi noktalara en yakın doğruyu tanımlamak için kullanılır (Akdeniz, 2013; Genç vd., 2003; Saraçlı vd., 2009). Bir noktadan bir doğruya en kısa uzaklık dik uzaklıktır (Warton vd., 2006). OR noktalardan doğruya en kısa uzaklıkların kareleri toplamını minimize eden doğrudur. Üstelik KDR bağımsız değişkenin (X) ölçüm hatası içermediğini, hatanın kaynağının yalnızca bağımlı değişken (Y) olduğu varsayılır. Deming, 1943 yılında yazdığı "Statistical Adjustment of Data" isimli kitabında, En Küçük Kareler (EKK) regresyon tekniğinin varsayım hatalarına dikkat çekmiş ve bağımsız değişkenin (X) de pratikte hata içerebileceğini belirtmiştir (Saraçlı vd., 2009). Bunun yanında Glaister (2005) ve Stöckl vd. (1998) pratikte her iki değişkende ölçme hatası olduğunu belirtmiştir. Bu durumda KDR'deki tahminler pratikte artık doğru değildir (Glaister, 2005). Her iki değişkende ölçme hatası içeren OR tekniğinin (Carr, 2012; Glaister, 2005; Scariano ve Barnet, 2003) tahminleri daha doğrudur. Bağımsız değişkenin de ölçme hatasını içerdiği OR tekniği, hesaplamalarda daha iyi sonuçlar verebilir (Calzada ve Scariano, 2003; Carr, 2012; Warton vd., 2006).

Uluslararası literatür incelendiğinde OR ile diğer regresyon tekniklerini karşılaştıran oldukça çok araştırmanın olduğu görülmektedir (Calzada ve Scariano, 2003; Carr, 2012; Ding vd., 2013; Elfessi ve Hoar, 2001; Glaister, 2005; Leng vd., 2007; Lolli ve Gasperini, 2012; Lundbrook, 2010; Ortiz vd., 2010). Ülkemizde yapılan metot karşılaştırma çalışmaları incelendiğinde ise bu konuda yapılan çalışmaların oldukça az olduğu dikkat çekmektedir. Bland Altman ile EKK Açıortay Type II tekniğini (Saraçlı ve Çelik, 2012), Deming regresyon tekniği ile EKK tekniğini (Saraçlı vd., 2009), EKK ve robust M tahmin yöntemini (Coşkuntuncel, 2013), EKK ve en küçük medyan kareler yöntemlerini (Alma ve Vupa, 2008) ve basit doğrusal regresyon ile hiyerarşik doğrusal tekniklerini (Atar, 2010) karşılaştıran metot karşılaştırma çalışmaları yapılmıştır. Türkiye'de OR tekniği üzerine jeofizikte (Öztürk, 2012) ve istatistikte (Saraçlı, 2011) yapılan çalışmalar da azdır.

Hem ulusal hem de uluslararası bu karşılaştırma çalışmalarında, hata kareler ortalamasına göre KDR (Carr, 2012) ve En Küçük Kareler Açıortay (Saraçlı, 2011) tekniklerinin üstün olduğu gösterilmiştir. Bunların yanında OR tekniğinin; belirleyicilik katsayısına (R^2) göre (Calzada ve Scariano, 2003) ve hataların standart sapmasına göre (Ding vd., 2013; Ortiz vd., 2010) üstünlüğünü gösteren çalışmalar da vardır.

OR tekniğinin yararlarından bahsedilmesine rağmen teknik; basit doğrusal regresyon hakkında Türk literatüründe sosyal bilimler ve eğitim alanındaki araştırmalarda yalnızca KDR tekniği ile karşılaşılmıştır (Bağçeci vd., 2011; Bahar, 2006; Bayat vd., 2014; Doruk vd., 2015; İlğan vd., 2012; Kesicioğlu ve Güven, 2014; Üredi ve Üredi, 2005). Yapılan literatür çalışmasında, OR üzerine sosyal bilimler ve eğitim alanında ülkemizde yapılmış bir çalışmaya rastlanmamıştır. Bu doğrultuda bu çalışmanın bundan sonra yapılacak çalışmalara ışık tutması beklenmektedir.

Bu araştırmanın amacı OR eşitliğini açıkça sunmak ve klasik regresyon ile ortogonal regresyonu dik uzaklıklar kareler toplamına göre karşılaştırmaktır.

Yöntem

Kolay anlaşılması için OR eşitliği yeniden literatürden bağımsız olarak sunulmuştur. Çünkü bu metot üzerine çalışmakta olan çoğu literatür kısa ve yoğun sembol içermektedir (Calzada ve Scariano, 2003; Carr, 2012; Kermack ve Haldane, 1950; Nievergelt, 1994). Klasik doğrusal regresyon ile OR'yi dik uzaklıklar kareler toplamına göre karşılaştırmak için analizler bir örnek üzerinden gösterilmiştir. Örnek, Akdeniz'in (2013) kitabından alınmıştır.

Verilerin analizinde, verilerin düzenlenmesinde, OR eşitliğinin hesaplanmasında ve OR ve KDR'nin dik uzaklıklar kareler toplamının hesaplanmasında Microsoft Excel 2010 kullanılmıştır. KDR eşitliğinin hesaplanmasında SPSS 17.0 programı kullanılmıştır.

Sonuç ve Tartışma

Regresyon analizinde değişkenler arasındaki ilişki bir matematiksel eşitlikle ifade edilir. Bağımsız değişkenin bağımlı değişken üzerindeki etkileri bu matematiksel eşitlik yoluyla tahmin edilebilir (Büyüköztürk vd., 2012).

Bu araştırmada, OR eşitliğini açıkça sunmak ve klasik regresyon ile OR tekniğini dik uzaklıklar kareler toplamına göre karşılaştırmak amaçlanmıştır. Bu karşılaştırma bir örnek üzerinden gerçekleştirilmiştir.

Akdeniz (2013), Genç vd. (2003) ve Saraçlı vd. (2009)'nin belirttiği gibi basit doğrusal regresyon analizi veri noktalarına en yakın doğruyu tanımlamak için kullanılır. KDR doğrusunun ve OR doğrusunun dik uzaklıklar kareler toplamına göre karşılaştırılması sonucunda, OR'nin dik uzaklıklar kareler toplamının daha küçük olduğu görülmüştür. Bu sonuç, maddelerin setliklerinden dayanıklılığın tahmininde OR tekniğinin performansının KDR tekniğinin performansından daha iyi olduğunu göstermiştir. Buradan OR doğrusu, verilere KDR doğrusundan daha yakındır. Bu bağlamda OR doğrusunun verileri daha iyi temsil ettiği görülmüştür. Bu sonuç Calzada ve Scariano (2003), Ding vd. (2013), Glaister (2005) ve Ortiz vd. (2010)'nin bulgularıyla benzerdir. OR tekniği dik uzaklıklar kareler toplamına göre KDR tekniğinden daha iyi bir metot olmasına karşın Carr (2012) ve Saraçlı'nın (2011) sonuçları ile çelişmektedir. Bağımlı ve bağımsız değişkenin her ikisinin de ölçme hatasını içeren ve doğruya en kısa uzaklığı minimize eden istatistik uzaklık koşulları altında OR doğrusunun Tablo 4'te görüldüğü gibi uygun bir regresyon doğrusudur.

Bu araştırma OR tekniğinin dik uzaklıklar kareler toplamına göre KDR tekniğinden daha iyi olduğunu, bağımlı ve bağımsız değişkenin her ikisinin de ölçme hatası içerdiği durumda OR tekniğinin kullanılmasının daha doğru olduğunu ortaya koymuştur.

Bu sonuçlara bağlı olarak OR tekniğinin basit doğrusal regresyon çalışmalarında KDR tekniğinden daha doğru sonuçlar elde etmede kullanılabilecek bir regresyon tekniği olduğu düşünülmektedir.