Comparison of Classical Least Squares and Orthogonal Regression in Measurement Error Models

Taliha KELES


The aim of this study is to investigate the effect of measurement errors on the estimation of simple linear regression and compare the performance of the classical least squares (CLS) regression method and orthogonal regression according to the standard deviation of the residuals. For this aim, analyses were performed on three data sets of different size. The first set of data was consisting 10 data for hardness and durability of substances. The second set of data was collected from 157 eighth-grade students' non-routine problem-solving test scores and mathematics scores on Examination for Transition to Basic Secondary Education. The third set of data was collected from 956 eighth-grade students' PISA Mathematics Literacy Test scores and Level Determination Examination scores. As a result of comparisons on three data sets, it was seen that orthogonal regression gave the smaller standard deviation of residuals than CLS regression method. Depending on these findings, it was determined that orthogonal regression has better performance than CLS regression method in estimating the linear relationship between two variables. The results of this study can be guiding for making important contributions about measurement error modeling to researchers working in social sciences.

ARTICLE INFO

Article History:
Received 01.02.2018
Received in revised form 12.05.2018
Accepted 02.06.2018
Available online 11.06.2018

ABSTRACT

The aim of this study is to investigate the effect of measurement errors on the estimation of simple linear regression and compare the performance of the classical least squares (CLS) regression method and orthogonal regression according to the standard deviation of the residuals. For this aim, analyses were performed on three data sets of different size. The first set of data was consisting 10 data for hardness and durability of substances. The second set of data was collected from 157 eighth-grade students’ non-routine problem-solving test scores and mathematics scores on Examination for Transition to Basic Secondary Education. The third set of data was collected from 956 eighth-grade students’ PISA Mathematics Literacy Test scores and Level Determination Examination scores. As a result of comparisons on three data sets, it was seen that orthogonal regression gave the smaller standard deviation of residuals than CLS regression method. Depending on these findings, it was determined that orthogonal regression has better performance than CLS regression method in estimating the linear relationship between two variables. The results of this study can be guiding for making important contributions about measurement error modeling to researchers working in social sciences.

Keywords:
classical least squares, measurement error, orthogonal regression, regression analysis, standard deviation of the residuals

Introduction

Regression analysis is a statistical technique for investigating and mathematical modeling the relationship between two or more variables (Akdeniz, 2013; Montgomery, Peck & Vining, 2001; Sykes, 1993; Yan & Su, 2009). This relationship is typically approximated by a straight line and regression analysis is used to define the best fitting line (Kane & Mroch, 2010; Scariano & Barnet, 2003). The relationship between variables in regression analysis is expressed by a mathematical equation (Büyüköztürk, Çokluk & Köklü, 2012; Can, 2013; Montgomery et al., 2001; Yan & Su, 2009). Through this mathematical equation, the effects of independent variables on the dependent variable can be estimated (Yan & Su, 2009).

Regression analysis has an important role in scientific research projects because it allows a researcher to examine relationship between two variables or to predict the future, which is the most important goals of science (Büyüköztürk, 2012; Büyüköztürk et al., 2012; Coşkuntuncel, 2013; Weisberg, 2005). Applications of regression analysis are seen in almost every field, including statistics, engineering, physical and chemical sciences, life and biological sciences, social sciences, education, economics, management, astronomy, medical research, political science, agriculture, biology, sociology, geology, meteorology, and many other areas of academic and applied science (Keleş & Altun, 2016; Montgomery et al., 2001; von Eye & Schuster, 1998; Yan & Su, 2009).
Researchers are often investigated in the relationships between one variable and several other variables (Yan & Su, 2009). It is used to answer questions such as Does smoking cause lung cancer? Do mathematics attitudes affect on mathematics achievement of students? Does school absenteeism affect the academic success of students, and so on (Weisberg, 2005; Yan & Su, 2009). All of them fall within the scope of regression analysis.

Many methods have been developed to determine various parametric relationships between dependent variable and independent variables. For example, linear regression, logistic regression, Poisson regression, and probit regression, etc. (Yan & Su, 2009). Linear regression requires that model is linear in regression parameters. There are three types of regression: simple linear regression, multiple linear regression and nonlinear regression or curvilinear regression (Büyüköztürk, et al., 2012; Montgomery et al., 2001; Yan & Su, 2009). The simple linear regression is for modeling the linear relationship between one dependent variable and one independent variable (Büyüköztürk et al., 2012; Yan & Su, 2009).

In scientific research projects, linear regression analysis applications have been frequently seen (Büyüköztürk et al., 2012; Küçüksille, 2016; Weisberg, 2005). The fundamental purpose of simple linear regression analysis is to determine the best model in order to predict the dependent variable (Yan & Su, 2009). It is estimated according to the regression line and the curve (Freund & Wilson, 1998). One of the most obvious advantages of the regression is to predict unknown by using the known variant if one of the variables is not known.

The equation of $\hat{y} = f(x)$ is used to find the predicted values of the dependent variable from the values of the independent variable $x$. Where $\hat{y}$ is the estimated value corresponding to $x$ value (Montgomery et al., 2001). The simple linear regression model can be written in the form $y = bx + a$ (Freund & Wilson, 1998; von Eye & Schuster, 1998). Where $y$ is the dependent variable, $a$ is intercept coefficient, $b$ is the slope coefficient of the simple linear regression line, $x$ is the independent variable. These coefficients are usually called regression coefficients (Freund & Wilson, 1998; Montgomery et al., 2001). The dependent variable is also called response variable, explained variable, or predicted variable and the independent variable is called explanatory, or predictor variable (Yan & Su, 2009).

The formula for linear regression line is $\hat{y} = bx + a$ where $\hat{y}$ represents that value, $y$, predicted from $x$ (Carr, 2012; Yan & Su, 2009). Regression models are attained by minimizing of measurement errors that may arise from variable or variables (Isobe, Feigelson, Akritas & Babu, 1990).

Measurement error is the difference between the quantity observed and the true quantity (Armağan, 1983; Ding, Chu, Jin & Zhu, 2013; Erkan & Kan, 2006; Saraçlı, Doğan & Doğan, 2009a). Unavoidably, measurement errors always exist during the measurement (Ding et al., 2013; Erkan & Kan, 2006). Linear regression methods with measurement errors called as measurement error models, or errors-in-variables models, have a variety mathematical optimum solutions (Deming, 1948; Ding et al., 2013; Fuller, 1987; Isobe et al., 1990; Stefanski, 2000). Although the increasing use of measurement error models in the statistics applications practice over the last 30-35 years, many researchers don’t know exactly measurement error problems (Stefanski, 2000).

Measurement error models are used for the linear regression techniques. The most widely known techniques for measurement errors are the classical least squares (CLS) regression method (von Eye & Schuster, 1998; Yan & Su, 2009) and orthogonal regression (Carroll & Ruppert, 1996; Isobe et al., 1990; Stefanski, 2000). CLS method is also known as ordinary least squares (Ding et al., 2013; Isobe et al., 1990; Kane & Mroch, 2010; von Eye & Schuster, 1998).

Ruppert, 1996; Markovsky & Van Huffel, 2007), and method of moments estimator (Fuller, 1987). In additional Carr (2012), Carroll and Ruppert (1996), Cheng and van Ness (1992), Isobe et al. (1990), Kane and Mroch (2010), Keleş and Altun (2016), and McCartin (2003) use the name orthogonal regression.

CLS regression method is assumed that there isn’t any measurement error in the independent variable, but there is only in the dependent variable (Glaister, 2005; Kane & Mroch, 2010; Leng, Zhang, Kleinman & Zhu, 2007; Scariano & Barnet, 2003). The CLS regression method minimizes the sum of squares vertical distances from data points to the regression line (Ding et al., 2013; Draper & Smith, 1981; Elfessi & Hoar, 2001; Isobe et al., 1990; Kane & Mroch, 2010; Leng et al., 2007; Ludbrook, 2010; Markovsky & Van Huffel, 2007; Ortiz, Pogliani & Besalú, 2010; Sen & Srivastava, 1990). Herein, The sum of squares of the errors is to be minimized as:

$$E_{CLS} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - bx_i - a)^2$$

where $\epsilon_i$ is called the residual (Calzada & Scariano, 2003; Glaister, 2005; Ortiz et al., 2010; Scariano & Barnet, 2003; Sen & Srivastava 1990).

$\epsilon_i = y_i - \hat{y}_i = y_i - (bx_i + a)$ is the vertical deviation from the point to the line as shown in Figure 1a. In the CLS regression, there is no error in the abscissa values, there is only in the ordinate values (Glaister, 2005). $y_i = \hat{y}_i + \epsilon_i$, where $\epsilon_i$ is the additional error, as shown in Figure 1a.

![Figure 1](image)

(a) Classical Least Squares Regression (a), Orthogonal Regression (b).

While CLS regression supposes that the only source of the measurement error is the dependent variable, the orthogonal regression method takes into account measurement error both dependent and independent variables (Carr, 2012; Deming, 1948; Gazeloğlu & Saracılı, 2013; Glaister, 2005; Kane & Mroch, 2010; Scariano & Barnet, 2003). Adcock (1878) first discovered orthogonal regression method under these assumptions more than 130 years ago.

This is different from the CLS regression method which measures error parallel to the y axis. $y_i = Y_i + \epsilon_i$ as in CLS regression and $x_i = X_i + \delta_i$, where $\delta_i$ is the additional error, as shown in Figure 1b. For each $i = 1, 2, 3, ..., n$, the pair $(x_i, y_i) = (X_i + \delta_i, Y_i + \epsilon_i)$ represents the coordinates of the $i^{th}$ measured data point (Elfessi & Hoar, 2001).

As illustrated in Figure 1b, because the abscissa values and the ordinate values are both subject to measurement error, orthogonal regression equation, is attained by minimizing the sum of the squares of perpendicular (or orthogonal) distances between the data and line (Carr, 2012; Ding et al., 2013; Elfessi & Hoar, 2001; Isobe et al., 1990; Kane & Mroch, 2010; Keleş & Altun, 2016; Leng et al., 2007; Li, 1984; Ludbrook, 2010; Nievergelt, 1994; Ortiz et al., 2010; Scariano & Barnet, 2003). The residual is defined as:
\[ d_i = \frac{|y_i - bx_i - a|}{\sqrt{b^2 + 1}} \]

Where \( d_i \) is deviation measured perpendicularly from the point to the line (Calzada & Scariano, 2003; Glaister, 2005; Li, 1984; Ortiz et al., 2010; Scariano & Barnet, 2003) as shown in Figure 1b. The sum of the squares of errors between the data points and regression line is to be minimized as:

\[ E_{OR} = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left( \frac{(y_i - bx_i - a)^2}{b^2 + 1} \right) \]

Thus, both CLS regression method and orthogonal regression are the least squares method, in that they minimize the variance (the sum of squares of the residuals) about the line, using different definitions for residual (Warton, Wright, Falster & Westoby, 2006).

The fact that there is always no clear distinction between the independent and dependent variable in practice (Calzada & Scariano, 2003, Glaister, 2005; Kane & Mroch, 2010; Saraçlı et al., 2009a). However, both dependent and independent variables include measurement error in the measurement (Deming, 1948; Glaister, 2005; Golub & Loan, 1980; Kane & Mroch, 2010; Saraçlı, 2011; Saraçlı, Yılmaz & Doğan, 2009b; Stöckl, Dewitte & Thienpont, 1998). In fact, the assumption of CLS regression is that the independent variable is known without error (Kane & Mroch, 2010). In this case, if both variables contain error, it would be wrong to expect that the CLS regression analysis approach would give a valid result (Glaister, 2005; Saraçlı et al., 2009b). Therefore, the orthogonal regression technique, which takes into account both variant errors and is called type II regression techniques in the literature, will give more accurate results (Calzada & Scariano, 2003; Carr, 2012; Kane & Mroch, 2010; Keleş & Altun, 2016; Saraçlı et al., 2009b; Warton et al., 2006).

CLS regression is one of the most frequently used methods of simple linear regression (Ding et al., 2013; von Eye & Schuster, 1998). The most fundamental reason for the widespread use of CLS regression is that it is easy to calculate (Coşkuntuncel, 2013; Ludbrook, 2010; von Eye & Schuster, 1998) and it is part of statistical software packages (Ortiz et al., 2010; von Eye & Schuster, 1998). When the goal is to predict an unknown dependent variable, CLS regression method is used in the social sciences. However, when the goal is to examine the relationship between two variables, which both contain error, orthogonal regression method is used (Calzada & Scariano, 2003; Carr, 2012; Kane & Mroch, 2010).

Although the orthogonal regression method has a long history in statistics, chemometrics and economics, the orthogonal regression method has not been adequately addressed in social sciences (Calzada & Scariano, 2003; Carr, 2012; Markovsky & Van Huffel, 2007). The most important reason is that the computation of orthogonal regression is complex and difficult (Adcock, 1878; Carrol & Ruppert, 1996; Carr, 2012; Kane & Mroch, 2010; Ortiz et al., 2010).

A review of literature indicated that few studies have been conducted analysis to compare orthogonal regression with other regression techniques according to mean square error (Carr, 2012; Elfessi & Hoar, 2001; Saraçlı, 2011), or according to determination of coefficient (\( R^2 \)) (Calzada & Scariano, 2003; Öztürk, 2012), or according to the sum of squared perpendicular distances (Keleş & Altun, 2016), or according to standard deviation of the residuals (Ding et al., 2013; Ortiz et al., 2010).

The aim of this study is to investigate the effect of measurement errors on the estimation of simple linear regression and compare the performance of the CLS regression method and orthogonal regression. In this context, CLS regression method and orthogonal regression method are compared according to the standard deviation of the residuals on three data sets. This study will give insight into measurement error modeling as well as guidelines on its use in estimation methods in social sciences for current and future researchers.

**Method**

In this study, CLS regression method and orthogonal regression method are compared of according to the standard deviation of the residuals on three data sets of different size.
The Research Data

In the study, three data sets were studied with small, medium and large volume samples. The first data set was taken from the book of Akdeniz (2013) and contains 10 data. The relationship between the hardness and the durability of 10 pieces of substances a material produced in a production center was investigated. The values of durability were used as dependent (y) variable, the values of hardness were used as independent (x) variable in the study.

The second set of data consisted of 157 eighth-grade students in Nilüfer district of Bursa during the spring semester of the 2014-2015 academic year. The relationship between students’ success of non-routine problem-solving and their mathematics scores on Examination for Transition to Basic Secondary Education (TEOG) was investigated. To measure non-routine problem solving skills of students, six non-routine open-ended problems were used. Students’ TEOG scores were obtained using the e-school system. The TEOG scores were used as dependent (y) variable, problem-solving test (PST) scores were used as independent (x) variable in the study.

The third set of data consisted of 956 eighth-grade students attending six different secondary schools in Osmangazi, Yıldırım and Nilüfer district of Bursa during the spring semester of the 2010-2011 academic year. The relationship between students’ success of PISA Mathematics Literacy Test (PMLT) and their scores on Level Determination Examination (SBS) was investigated. In the study, 25 mathematical literacy problems, that 8 of the questions were multiple-choice and the 17 of the questions were open-ended ones, selected from among those made optional in PISA 2003 and 2006 examinations. Students' SBS scores were obtained using the e-school system. The SBS scores were used as dependent (y) variable, PMLT scores were used as independent (x) variable in the study.

Data Analysis

In order to be able to make a simple linear regression analysis, before starting the analysis, it is necessary to test the hypothesis of if there is normal distribution among variables, whether there is a linear relationship. It is usually assumed that measurement errors are the normal distribution with zero mean the simple linear regression (Büyüköztürk, 2012; Yan & Su, 2009). In this study, three sets of data of different sizes were studied. For each data set, firstly, the assumptions of normality and linearity of variables were examined with graphs between standardized estimated values and standardized residual values (Büyüköztürk, 2012). Then, CLS regression and orthogonal regression equations were computed and the scatter diagram of CLS regression and orthogonal regression lines were obtained. Finally, the standard deviation of the residuals for CLS regression and orthogonal regression method were calculated and compared.

The slope and the intercept coefficient of CLS regression method are calculated according to defined formulas by Akdeniz (2013) and Montgomery et al. (2001), and Sen and Srivastava (1990). Formulas for coefficients of CLS regression method are given as:

\[
\begin{align*}
a &= \frac{\sum_{i=1}^{n} y_i}{n} - b \frac{\sum_{i=1}^{n} x_i}{n} \\
b &= \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}
\end{align*}
\]

Where \(a\) is the intercept coefficient of CLS regression and \(b\) is the slope coefficient of CLS regression method.

The slope and the intercept coefficient of orthogonal regression method are calculated according to defined formulas by Keleş and Altun (2016). Formulas for coefficients orthogonal regression method are given as:
\[
a = \frac{\sum_{i=1}^{n} y_i}{n} - b \frac{\sum_{i=1}^{n} x_i}{n}
\]

\[
b = \left( \frac{\sum_{i=1}^{n} x_i^2 \sum_{j=1}^{m} y_j^2 - m \sum_{i=1}^{n} x_i^2 \sum_{j=1}^{m} y_j^2}{\left[ (\sum_{i=1}^{n} x_i^2)^2 - m (\sum_{i=1}^{n} x_i^2)^2 - (\sum_{i=1}^{n} y_i^2)^2 - m (\sum_{i=1}^{n} y_i^2)^2 \right]} + 4 \left[ n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{m} y_i \right]^2 \right) \times \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{m} y_i}{2 n \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}
\]

Where \( a \) is the intercept coefficient of orthogonal regression and \( b \) is the slope coefficient of orthogonal regression method.

The standard deviation of the residuals for CLS regression method is calculated according to defined by Akdeniz (2013), Ding et al. (2013), and Lolli and Gasperini (2012). The formula for the standard deviation of the residuals for CLS regression method is given as:

\[
S_{CLS} = \sqrt{\frac{n}{n-2} \left( \sum_{i=1}^{n} (y_i - bx_i - a)^2 \right)}
\]

Where \( S_{CLS} \) is the standard deviation of the residuals for CLS regression and \( n \) is the number of observed data.

The standard deviation of the residuals for orthogonal regression method is calculated according to defined by Ortiz et al. (2010). The formula for the standard deviation of the residuals for orthogonal regression method is given as:

\[
S_{OR} = \sqrt{\frac{n}{(1+b^2)(n-2)} \left( \sum_{i=1}^{n} (y_i - bx_i - a)^2 \right)}
\]

Where \( S_{OR} \) is the standard deviation of the residuals for orthogonal regression and \( n \) is the number of observed data.

Microsoft Excel 2010 was used for data organization and calculation of orthogonal regression equation and standard deviation of the residuals, SPSS 23.0 was used for the CLS regression equation.

**Results**

According to the data sets in this section, the findings of testing the assumptions, the scatter diagram of CLS and the orthogonal regression lines and, the comparative results are given with respect to the CLS regression method and orthogonal regression method.

**Durability Level Data**

For this small volume data, graphs for examining the assumptions of normality and linearity are given in Figure 2.
There is a positive and linear relationship between the variables according to Figure 2a. In Figure 2b, it is seen that the histogram and normal distribution curves generated for standardized predicted values have the normal distribution. Scatter diagram of CLS and orthogonal regression lines for durability level data is shown in Figure 3.

As illustrated in Figure 3, the CLS regression of durability level data has slope smaller than that of the orthogonal regression line, causing the line to be closer to being horizontal than the orthogonal regression line.

Table 1 shows the regression equations, standard deviations of the residuals and correlation, for this data under the two regression methods.
The last column in Table 1 displays the Pearson correlation coefficient (r=0.945, r²=0.893) which indicates that there is a positive and very strong linear relationship between the variables. The Pearson correlation between hardness and durability of substances is 0.945, indicating that hardness of substances explains about 89.3% of the variance in the durability of substances.

The second column displays regression equations for the two regressions methods. When the CLS regression and orthogonal regression method are applied to estimate the coefficient between hardness and durability of substances, CLS regression line is \( \hat{y} = 1.25x + 0.25 \) and the orthogonal regression line is \( \hat{y} = 1.34x - 0.38 \).

The third column in Table 1 reports the standard deviation of the residuals for CLS regression and orthogonal regression method. It is seen that orthogonal regression estimator gives less standard deviation of the residuals than CLS regression estimator which presents that it has better performance than CLS regression estimator in the estimating the linear relationship between hardness and durability of substances in this example. In addition, to estimate the linear relationship between hardness and durability of substances the orthogonal regression method is more appropriate than CLS regression method.

**TEOG Scores Data**

For this medium volume, graphs for examining the assumptions of normality and linearity are given in Figure 4.

![Normal P-P Plot of Regression Standardized Residual](image)

(a)

![Dependent Variable: TEOG](image)

(b)

**Figure 4.** Linearity (a), Normal Distribution Curve (b) for TEOG Scores Data.

There is a positive and linear relationship between the variables according to Figure 4a. In Figure 4b, it is seen that the histogram and normal distribution curves generated for standardized predicted values have the normal distribution.

Scatter diagram of CLS and orthogonal regression lines for TEOG scores data is shown in Figure 5.

---

**Table 1. Comparison of Results Obtained with CLS and Orthogonal Regression Methods for Durability Level Data**

<table>
<thead>
<tr>
<th>Regression Method</th>
<th>( \hat{y} = bx + a )</th>
<th>Standart Deviation of the Residuals</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLS Regression</td>
<td>( \hat{y} = 1.25x + 0.25 )</td>
<td>( S_{CLS} = 0.75 )</td>
<td>0.945</td>
</tr>
<tr>
<td>Orthogonal Regression</td>
<td>( \hat{y} = 1.34x - 0.38 )</td>
<td>( S_{OR} = 0.45 )</td>
<td></td>
</tr>
</tbody>
</table>
As illustrated in Figure 5, the CLS regression of TEOG scores data has the slope smaller than that of the orthogonal regression line, causing the line to be closer to being horizontal than the orthogonal regression line. Table 2 shows the regression equations, standard deviations of the residuals and correlation, for this data under the two regression methods.

<table>
<thead>
<tr>
<th>Regression Method</th>
<th>( \hat{y} = bx + a )</th>
<th>Standard Deviation of the Residuals</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLS Regression</td>
<td>( \hat{y} = 1.12x + 46.11 )</td>
<td>( S_{\text{CLS}} = 20.57 )</td>
<td>0.522</td>
</tr>
<tr>
<td>Orthogonal Regression</td>
<td>( \hat{y} = 3.54x + 0.04 )</td>
<td>( S_{\text{OR}} = 9.21 )</td>
<td></td>
</tr>
</tbody>
</table>

The last column in Table 2 displays the Pearson correlation coefficient (r=0.522, \( r^2 = 0.272 \)) which indicates that there is a moderate linear relationship between the variables. The Pearson correlation between PST scores and TEOG scores is 0.522, indicating that PST score explains about 27.2% of the variance in TEOG score.

The second column displays regression equations for the two regressions methods. When the CLS regression and orthogonal regression method are applied to estimate the coefficient between PST scores and TEOG scores, CLS regression line is \( TEOG = 1.12xPST + 46.11 \) and the orthogonal regression line is \( TEOG = 3.54xPST + 0.04 \).

The third column in Table 2 reports the standard deviation of the residuals for CLS regression and orthogonal regression method. It is seen that orthogonal regression estimator gives less standard deviation of the residuals than CLS regression estimator which presents that it has better performance than CLS regression estimator in the estimating the linear relationship between PST scores and TEOG scores in this example. In addition, to estimate the linear relationship between PST scores and TEOG scores the orthogonal regression method is more appropriate than CLS regression method.

**SBS Scores Data**

For this large volume data, graphs for examining the assumptions of normality and linearity are given in Figure 6.
Figure 6. Linearity (a), normal distribution curve (b) for SBS scores data.

There is a positive and linear relationship between the variables according to Figure 6a. In Figure 6b, it is seen that the histogram and normal distribution curves generated for standardized predicted values have the normal distribution.

Scatter diagram of CLS and orthogonal regression lines for SBS scores data is shown in Figure 7.

Figure 7. Scatter Diagram of CLS ($C: \hat{y} = 7.09x + 192.41$) and Orthogonal Regression Lines ($O: \hat{y} = 12.52x + 81.44$) for SBS Scores Data.

As illustrated in Figure 7, the CLS regression of SBS scores data has slope smaller than that of the orthogonal regression line, causing the line to be closer to being horizontal than the orthogonal regression line.

Table 3 shows the regression equations, standard deviations of the residuals and correlation, for this data under the two regression methods.
Table 3. Comparison of Results Obtained with CLS and Orthogonal Regression Methods for SBS Score Data

<table>
<thead>
<tr>
<th>Regression Method</th>
<th>$\hat{y} = bx + a$</th>
<th>Standard Deviation of the Residuals</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLS Regression</td>
<td>$\hat{y} = 7.09x + 192.41$</td>
<td>$S_{CLS} = 63.07$</td>
<td>0.751</td>
</tr>
<tr>
<td>Orthogonal Regression</td>
<td>$\hat{y} = 12.52x + 81.44$</td>
<td>$S_{OR} = 6.65$</td>
<td></td>
</tr>
</tbody>
</table>

The last column in Table 3 displays the Pearson correlation coefficient ($r=0.751$, $r^2=0.564$) which indicates that there is a strong linear relationship between the variables. The Pearson correlation between PMLT scores and SBS scores is 0.751, indicating that PMLT score explains about 56.4% of the variance in SBS score.

The second column displays regression equations for the two regression methods. When the CLS regression and orthogonal regression method are applied to estimate the coefficient between PMLT scores and SBS scores, CLS regression line is $SBS = 7.09xPMLT + 192.41$ and the orthogonal regression line is $SBS = 12.52xPMLT + 81.44$.

The third column in Table 3 reports the standard deviation of the residuals for CLS regression and orthogonal regression method. It is seen that orthogonal regression estimator gives less standard deviation of the residuals than CLS regression estimator which presents that it has better performance than CLS regression estimator in the estimating the linear relationship between PMLT scores and SBS scores in this example. In addition, to estimate the linear relationship between PMLT scores and SBS scores the orthogonal regression method is more appropriate than the CLS regression method.

Discussion and Conclusion

In this study, the effect of measurement errors on the estimation of simple linear regression which is frequently used in social sciences is investigated. For this purpose, CLS regression method and orthogonal regression are compared according to the standard deviation of the residuals on small, medium and large volume samples. The goal of these analyses for three data sets should guide the choice of method for the regression equations.

According to the findings of this study, it is seen that orthogonal regression gives less standard deviation of the residuals than CLS regression on three data sets of different size (Table 1, Table 2, Table 3). The results obtained for durability level data and TEOG scores data are important to examine the performance of CLS regression and orthogonal regression method over small and medium-sized samples. Findings presented that orthogonal regression provides more accurate conclusions in examining the relationship between variables than CLS regression (Table 1, Table 2). SBS scores data is important in terms of large sample performance in data. As a result, the orthogonal regression provides the more accurate conclusion in examining the relationship between two variables than CLS regression (Table 3).

These findings reveal that orthogonal regression estimator is the more appropriate than the CLS estimator in these examples. Thus, the findings of this study are consistent with the result of previous studies (Calzada & Scariano, 2003; Carr, 2012; Ding et al., 2013; Elfessi & Hoar, 2001; Glaister, 2005; Kane & Mroch, 2010; Keleş & Altun, 2016).

Ding et al. (2013) investigated and compared the effect of measurement errors with the least squares and orthogonal regression for prediction of material property. They showed that orthogonal regression estimator has better performance than least squares estimator in the prediction of material property. Similarly, Ortiz et al. (2010) reviewed and discussed ordinary least squares method and orthogonal least squares regression method both theoretical and applicative. They pointed out that orthogonal least squares fitting is a very good method than ordinary least squares because it deals with errors in both variables. Calzada and Scariano (2003) studied on contrasting the ordinary least squares and total least squares regression method using two real data. They stated that total least squares regression is better than ordinary least squares and that total least squares regression presents a much better fit for the data. However, Calzada and Scariano (2003) pointed out that when both variables contain error, the total least squares method is the appropriate method to use. Keleş and Altun (2016) compared classical linear regression and orthogonal regression techniques with
respect to the sum of squared perpendicular distances. They pointed out that it is more accurate to use orthogonal regression method when there are measurement errors in both variables. In Elfessi and Hoar’s (2001) research, in which they studied to compare the performance of six different methods of estimating (as Ordinary Least Squares, Least Absolute Deviations, Least Absolute Perpendicular Deviations, Total Least Squares, Squares of Horizontal and Vertical Deviation, and Absolute Horizontal and Vertical Distances) the value of the slope according to mean square error on the results of the simulations, it is found the estimates from total least squares are better than the ordinary least squares estimates for most of the simulation results.

Similarly, Kane and Mroch (2010) examined the relationship between the two variables across different groups in ordinary least squares and orthogonal regression. They used two synthetic and two real data sets. The results revealed that ordinary least squares regression is appropriate if the goal is to use a known independent variable to predict an unknown dependent variable but the orthogonal regression method is more appropriate than the ordinary least squares regression method. Likewise, Carr (2012) examined and compared three different linear regression methods (Classical Linear Regression, Orthogonal Regression, and Reduced Major Axis) using geyser eruption data according to the mean squared error. He pointed out that classical linear regression gives the smallest mean squared error which implies it is the best method to predict the time for the next eruption. However, he suggested that when the goal is to compare the relationship between time and eruptions, then major axis and reduced major axis methods are preferred to the classical linear regression.

Calzada and Scariano (2003) and Glaister (2005), and Kane and Mroch (2010) emphasized that there is no clear-cut answer as to whether the classical least squares or orthogonal regression method should be preferred in a given application. However, Carr (2012), and Kane and Mroch (2010) stated that the goal of the analysis played a major role choice of method. They stated orthogonal regression method the most appropriate estimates when the goal is to examine the relationship between two variables but CLS regression is the best method to apply when the goal is to predict dependent variable from independent variable. Deming (1948), Glaister (2005), Golub and Loan (1980), Kane and Mroch (2010), Saraçlı (2011), Saraçlı et al. (2009b), and Stöckl et al. (1998) stated that there may be measurement errors in both variables in practice. In fact, the slope and intercept coefficient estimators for the CLS regression method are derived under the assumption that the only source of the measurement error is the dependent variable (Calzada & Scariano 2003; Glaister, 2005; Kane & Mroch, 2010; Leng et al., 2007; Scariano & Barnet, 2003). On the other hand, the slope and intercept coefficient estimators for the orthogonal regression method are derived under the assumption that the source of measurement error is both dependent and independent variable (Calzada & Scariano 2003; Carr, 2012; Glaister, 2005; Kane & Mroch, 2010; Scariano & Barnet, 2003). In addition, Calzada and Scariano (2003), Glaister (2005), Kane and Mroch (2010), Saraçlı et al. (2009a) emphasized that there is always no clear distinction between independent and dependent variable. In this case, orthogonal regression provides more accurate estimations about linear relationship between two variables than CLS estimator method (Calzada & Scariano, 2003; Carr, 2012; Kane & Mroch, 2010; Keleş & Altun, 2016; Saraçlı et al., 2009b; Warton et al., 2006). Likewise, the findings of this study indicate that orthogonal regression method the most appropriate estimate for the relationship between two variables.

The scope of social sciences is humans. It enables reaching accurate results by making appropriate analyzes to data obtained from humans (Büyüköztürk et al., 2012). When there are errors in both variables and there is no clear distinction between independent and dependent variable, the estimates in CLS regression are no longer accurate. In that case, using the orthogonal regression method which takes into account of measurement errors in both variables will give more accurate results. In this respect, orthogonal regression will make important contributions to researchers working in social sciences.

In this study has been investigated the effect of measurement errors on the estimation of simple linear regression and compared the performance of the CLS regression method and orthogonal regression according to the standard deviation of the residuals. Based on the findings and results of this study, it is recommended to be preferred the orthogonal regression method to estimate the linear relationship between two variables. Further research can be done on the effect of measurement errors on the estimation of multiple linear regression.
REFERENCES


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.


