Educational Sciences: Theory & Practice - 13(4) • 2151-2158
©2013 Educational Consultancy and Research Center
www.edam.com.tr/estp
DOI: 10.12738/estp.2013.4.1867

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

Orkun COŞKUNTUNCEL[a]
Mersin University

Abstract

The purpose of this study is two-fold; the first aim being to show the effect of outliers on the widely used least squares regression estimator in social sciences. The second aim is to compare the classical method of least squares with the robust M-estimator using the "determination of coefficient" (R^2). For this purpose, analyzes were performed on three data sets. The first set of data is hypothetical, consisting of 15 students' general mathematic and linear algebra final scores. The second set of data was collected from 231 adolescents attending different high schools in Turkey. The data were collected using the Scale of Aggressiveness, Academic Self-efficacy Scale, Scale of Peer Pressure, and Trait Anxiety Inventory. The third set of data was collected from 1,385 high school students. This data were collected using the Maslach Burnout Inventory-Students Survey, Coping Styles of Stress Scale, Test Anxiety Inventory, Adolescence Self-Efficacy Scale, and Parental Attitude Scale. It was seen that, comparisons with small, medium and large volume samples, especially for the data sets including outlier/outliers, R^2 in M estimate is better alternatives than those having least squares. The findings are discussed in light of the recommendations presented in the literature.

Key Words

Coefficient of Determination, Least Squares, M-Estimator, Outlier, Regression Analysis, Robust Statistics.

In scientific research projects, finding a relationship between two or more variables and then expressing it in a mathematical equation is an important dimension needed in order to make future predictions. This mathematical relationship does not only refer to functional relationship, but also shows that one of the variables of a predetermined value provides estimation of the other.

The method that permits one to depict the relationship between variables in an equation is called "regression analysis," a method which has applications in almost every field (Arıcı, 1991). Regression analysis has an important role in scientific research projects because it allows a researcher to predict the future, which is one of the most important missions of science. In fact, regression analysis may be the most widely used statistical technique (Büyüköztürk, 2005; Büyüköztürk, Çokluk, & Köklü, 2011).

In general, the simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (1)$$

where the intercept β_0 and the slope β_1 are unknown constants and where e is a random error component.

a Orkun COŞKUNTUNCEL, Ph.D., is an assistant professor of Mathematics Education. His research interests include mathematics education and robust statistical methods. *Correspondence:* Mersin University, Faculty of Education, Department of Elementary Education, 33110 Yenişehir, Mersin, Turkey. Email: orkunct@gmail.com Phone: +90 324 341 2815/1733.

The errors are assumed to have a mean of 0 and a variance of σ^2. Additionally, it is usually assumed that the errors are uncorrelated. Customarily, x is called the independent variable, and y is called the dependent variable. The dependent variable, y, is a function of the independent variable, x (Büyüköztürk, 2005).

A regression model that involves more than one regressor variable is called a multiple regression model. The model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad (2)$$

is called a multiple linear regression model with k regressor. The parameters β_j, j = 1, ... , and k are called regression coefficients (Draper & Smith, 1998; Montgomery, Peck, & Vining, 2001).

It is more convenient to deal with multiple regression models if they are expressed in matrix notation. This allows for a very compact display of the model, data, and results. In matrix notation, the model given by Eq. (2) is:

$$y = Xb + \varepsilon \qquad (3)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad (4)$$

In general, y is an *n* x 1 vector of the observations, X is an *n* x *p* matrix of the levels of the regressor variables, b is an *p* x 1 vector of the regression coefficients, and ε is an *n* x 1 vector of random errors (Montgomery et al., 2001). The major assumptions made thus far in the present study of regression analysis are as follows:

i) The relationship between the response y and the regressors is linear, at least approximately.

ii) The errors term ε has a zero mean.

iii) The errors term ε has a constant variance σ^2.

iv) The errors are uncorrelated.

v) The errors are normally distributed.

Assumption (v) is required to test a hypothesis and to estimate intervals. Assumptions (iv) and (v) imply that the errors are independent random variables (Draper & Smith, 1998; Montgomery et al., 2001).

Least squares estimation is widely used to estimate unknown parameters in regression analysis. The least squares function is: min $\Sigma\varepsilon_i^2$ where e = y – Xβ. This function must be minimized with respect to β. Thus, the least squares estimator of β is:

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (5)$$

and the fitted regression model is:

$$\hat{y} = X\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \qquad (6)$$

Standardized residuals for least squares:

$$r_i = \frac{e_i}{\hat{\sigma}} \qquad (7)$$

where $e_i = y_i - \hat{y}_i$ and:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 \qquad (8)$$

Errors have a zero mean, a standard error of s, and are identically distributed. Therefore, $\hat{\sigma}^2$ is an unbiased estimator for σ^2 (Birkes & Dodge, 1993; Chatterjee, Hadi, & Price, 2000; Draper & Smith, 1998; Montgomery et al., 2001).

It is often assumed in the social sciences that data conform to a normal distribution. The least squares method is a suitable method and has good statistical properties when the data are normally distributed. However, in the case of deviations from normality, the least squares method is not an effective estimator. In this situation, robust estimators can be a suitable alternative method (Arslan & Billor, 2000).

Robust statistics refers to the stability theory of statistical procedures. It systematically investigates the effects of deviations from modeling assumptions on known procedures and, if necessary, develops new, better procedures. Common modeling assumptions are those of normality and of independence of random errors.

The implicit or explicit hope that under approximate (instead of exact) normality the least squares method would still be approximately optimal was thwarted by Tukey (1960). Soon after Tukey's (1960) inspiring paper, the foundations for four closely related robustness theories were laid by Huber (1964; 1965), Hample (1968), and Rousseeuw (1984).

As has been mentioned above, one of the main problems of regression analysis is outliers; that is, observations far from the bulk of the data. The main target of robust statistical methods is to develop a method that will combat outliers. For this, in the social sciences, the least squares method is preferred to robust methods because of easy computation (Aktaş, 2005; Aluçdibi & Ekici, 2012; Coşkuntuncel, 2005; Gündüz & Çelikkaleli, 2009; Güneş & Tulçal, 2002; İnandı, 2009; Rahman & Amri, 2011; Şahin & Anıl, 2012).

There are many procedures for robust regression estimation proposed in the literature. Among the most commonly used method is the robust M-estimator. M-estimation is known as the classical robust regression estimator (Arslan, 1992, 2004a, 2004b; Arslan & Billor, 1996; Arslan, Edlund, & Ekblom, 2001; Belsley, Kuh, & Welsch, 1980; Hample, Ronchetti, Rousseeuw, & Stahel, 1986; Huber, 1981; Rousseeuw, 1984; Rousseeuw & Leroy, 1987; Rousseeuw & Yohai, 1984; Rousseeuw & Zomeren, 1990).

The M-estimator for the unknown coefficient β given in Eq. (3) is:

$$\min \sum_{i=1}^{n} \rho(e_i) = \min \sum_{i=1}^{n} \rho(y_i - x_i'\beta) \qquad (9)$$

where ρ(e) is a function that satisfies the following conditions:

i) $\rho(e) \geq 0$ *ii)* $\rho(0) = 0$ *iii)* $\rho(e) = \rho(-e)$ *iv)* $|e_i| > |e_j|$, $i \neq j$ while $r(e_i) \geq \rho(e_j)$, $e_i = y_i - x_i' b$

The two most widely used ρ functions are the Huber and Tukey r functions. The Huber ρ function is:

$$\rho\ (e) = \begin{cases} e^2/2 & ,-k \leq e \leq k \\ k\,|\,e\,| - (k^2/2) & ,\ \text{d.y.} \end{cases}$$

where k = 1,345, and the Tukey ρ function is:

$$\rho\ (e) = \quad \rho(e) = \begin{cases} (c^2/6)\left(1 - \left[1 - (e/c)^2\right]^3\right), & |\,e\,| \leq c \\ c^2/6 & ,|\,e\,| > c \end{cases}$$

where c = 5 or 6 (Hample et al., 1986; Huber, 1981; Maronna, Martin, & Yohai, 2006; Rousseeuw & Leroy, 1987). In this study, the Tukey ρ function has been used. Since the Tukey ρ function is a differentiable function, the researcher has obtained the following estimating equation after setting the derivative of Eq. (9) with respect to β to 0:

$$\sum_{i=1}^{n} \frac{\rho'(e_i)}{e_i} . e_i x_i = 0\,, \qquad (10)$$

where $e_i \neq 0$. Further, if $w_i = \rho'(e_i)/e_i$ then the following weighted form of the estimator for β is obtained:

$$\hat{\beta}_M = \left(\sum_{i=1}^{n} w_i x_i x_i' \right)^{-1} \sum_{i=1}^{n} w_i x_i y_i = (X'WX)^{-1} X'Wy \qquad (11)$$

where W = diag(w_1, ..., w_n) is a diagonal matrix. Here, the weight function is a bounded function of the residuals so that the observations with large residuals will receive smaller weights and will hence exert less of an effect on the estimator (Birkes & Dodge, 1993; Coşkuntuncel, 2009).

In order to assess the quality of the fit in multiple linear regression, the coefficient of determination, or R^2, is a very simple tool, yet the most used by practitioners. Indeed, it is reported in most statistical analyzes, and although it is not recommended as a final model selection tool, it does provide an indication of the suitability of the chosen explanatory variables in predicting the response. R^2 is often understood to be the proportion of variation explained by the regressor, x. For the least squares coefficient of determination may be computed by the ANOVA table given in Table 1 (Montgomery et al., 2001).

The quantity for the coefficient of determination is:

$$R^2 = \frac{KT_R}{KT_T} = 1 - \frac{KT_H}{KT_T} \qquad (12)$$

Since $0 \leq KT_H \leq KT_T$, it follows that $0 \leq R^2 \leq 1$. The values of R^2 that are close to 1 imply that most of the variability in y is explained by the regression model.

In the classical setting, it is well known that the least-squares fit and the coefficient of determination may be arbitrary and/or misleading in the presence of a single outlier. In many applied settings, both the assumption of normality of the errors and the absence of outliers are difficult to establish. In these cases, Renaud and Feser (2010) have suggested that robust coefficient of determination (R_w^2). R_w^2 is:

Table 1.
Analysis of Variance (ANOVA) Table

Source of Variation	Sum of Squares (KT)	Degrees of Freedom	Mean Square (KO)	F
Regression	$KT_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	K	$KO_R = KT_R / k$	KO_R / KO_H
Error	$KT_H = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	n – k – 1	$KO_H = KT_H / (n - k - 1)$	
Total	$KT_T = \sum_{i=1}^{n} (y_i - \overline{y})^2$	n – 1		

$$R_w^2 = \left(\frac{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)(\hat{y}_i - \bar{\hat{y}}_w)}{\sqrt{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)^2 \sum_{i=1}^{n} w_i (\hat{y}_i - \bar{\hat{y}}_w)^2}} \right)^2 \qquad (13)$$

where $\bar{\hat{y}}_w = (1/\Sigma w_i)\Sigma w_i y_i$, $\bar{\hat{y}}_w = (1/\Sigma w_i)\Sigma w_i \bar{\hat{y}}_i$, w_i and $\bar{\hat{y}}_i$ are weights and the fitted value of the robust M-estimator. The total sum of squares form of (R_w^2) is:

$$\tilde{R}_w^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)^2} \qquad (14)$$

and (R_w^2) = ($\tilde{R}_w^2$) (Renaud & Feser, 2010).

Method

In this study, both simple and multiple linear regression methods were performed on various data sets. The classical method of least squares and the Robust M-regression estimator are compared with respect to the coefficient of determination.

The Research Data

In this study, three sets of data have been studied. The first set of data is hypothetical and contains 15 randomly selected student' final grades in Linear algebra and General Mathematics from Mersin University's Education Faculty. For the data collection simple linear regression was used. Whether the success of general mathematics predicted the success of linear algebra was investigated.

The second and third sets of data consist of real data. These data were analyzed using the least squares method and published as two papers before this study. The first of them has been carried out by Gündüz and Çelikkaleli (2009). This study aimed to analyze the prediction of male and female students' levels of aggressiveness in terms of the following variables: belief of academic efficacy, peer pressure, and anxiety. The research group consisted of 231 high school students (129 females; 102 males) aged between 14 and 19. In the research project, the Aggressiveness Scale, Academic Self-Efficacy Scale, Peer Pressure Scale, and Anxiety Inventory were used as measurement devices.

The second set of data was derived from Çapulcuoğlu (2012). This study is a descriptive study aiming to examine the burnout level of students according to gender, grade level, school type, and perceived level of academic achievement; as well as to investigate the relationship between student burnout with factors such as coping with stress, test anxiety, academic self-efficacy, and parental attitudes. The study group consisted of 1,385 high school students in various distics of the city of Mersin in Turkey during the 2010-2011 academic year. The Maslach Burnout Inventory-Student Survey (MBI-SS) was used to measure students' burnout levels; the Coping with Stress Styles Scale was used to measure the styles of used to cope with stress; the Test Anxiety Inventory was administered to gauge test anxiety level; the Adolescence Self-Efficacy Scale was applied to evaluate self-efficacy; the Parental Attitude Scale was used to measure parental attitudes; and the Personal Information Sheet was used to gather demographic data in the study.

Data Analysis

The data were analyzed using the R v2.15.1 program (Chambers, Eddy, Hardle, Sheather, & Tierney, 2002; Delgaard, 2008; R Core Team, 2013; Wilcox, 2005). The codes needed for (R_w^2) are given below:

```
ekk<-lsfit(x,y,intercept=T)#least squares
estimation

M_tah<-rreg(x, y, int=T, iter=100)#Robust
M estimator

w<-M_tah$w

ysapka<-M_tah$fitted.values

ywcizgi<-(1/sum(w))*sum(w*y)

ywsapkacizgi<-(1/sum(w))*sum(w*ysapka)

RKAREw<-((sum(w*(y-ywcizgi)*(ysapka-
ywsapkacizgi)))/(sqrt(sum(w*(y-
ywcizgi)^2)*sum(w*(ysapka-
ywsapkacizgi)^2))))^2

RKAREwtilda<-1-(sum(w*(y-ysapka)^2)/
sum(w*(y-ywcizgi)^2))
```

Results

The Results of Hypothetical Data

Table 2 shows the simple linear regression results for least squares and the M-estimator of both.

As shown in Table 2, observation 11 holds zero weight in reference to the M-estimator and the least squares estimator is more different than the

Table 2.
Least Squares and M-Estimator for Simple Linear Regression

#	GMat. (x)	LCeb. (y)	$\hat{y}_{EKK}$=32,807 + 0,235x	$\hat{y}_{M}$=12,224 + 0,572x	Weights obtained by M-estimator (w)
1	80	50	51,59	58,04	0,96
2	67	50	48,54	50,59	1,00
3	57	56	46,19	44,87	0,92
4	35	30	41,02	32,27	0,99
5	80	60	51,59	58,04	1,00
6	60	30	46,89	46,58	0,82
7	80	70	51,59	58,04	0,90
8	82	51	52,06	59,18	0,96
9	70	50	49,24	52,31	1,00
10	72	54	49,71	53,46	1,00
11	**90**	**10**	**53,94**	**63,76**	**0,00**
12	85	73	52,76	60,90	0,89
13	60	56	46,89	46,58	0,94
14	84	52	52,53	60,33	0,96
15	71	52	49,47	52,88	1,00

M-estimator. This means that there is a conflict between the M-estimator and this single outlier. Because the classical R^2 has a value of 0,044, whereas (R^2_w) equals 0,466, it is observed that the effect of the outlier/outliers and the resistance of the M- estimate can be easily seen by excluding outlier/outliers from data. Table 3 shows the results of estimates with and without observation 11 while also depicting the effects of single outliers

Aggressiveness Level Data

For this data (Gündüz & Çelikkaleli, 2009), regression analysis was conducted separately for 102 male and 129 female students. The results are shown in Table 4 for the 102 male students.

If the weights obtained by the M-estimators are analyzed, then observations 52, 65, and 101 each hold near zero weight, whereas the others hold an approximate weight of 1. This means that these 3 observations are outliers. The robust coefficient of determination is higher than the classical one, as expected. The same analysis performed on the male students was also performed on the 129 female students, the results of which are shown in Table 5. Here, observations 9, 20, 69, 78, 79, 92, 95, and 128 hold weights near zero and the others of about 1. Again, the robust coefficient of determination is higher than the classical one, as is to be expected.

Table 4.
Comparison of Estimators for 102 Male Students

Independent variables	LS Estimation	Std. Err.	M-estimation	Std. Err.
SABİT	125,151	12,53	113,213	11,85
SK	0,084	0,23	0,165	0,22
AB	0,243	0,07	0,255	0,06
AYİ	-0,667	0,26	-0,429	0,24
R^2	0,235		0,280	

Table 5.
Comparison of Estimators for129 Female Students

Independent variables	LS Estimation	Std. Err.	M estimation	Std. Err.
SABİT	99,058	13,62	103,188	12,34
SK	0,801	0,22	0,756	0,20
AB	0,234	0,07	0,174	0,06
AYİ	-0,960	0,23	-0,948	0,20
R^2	0,297		0,363	

Table 6 shows the results of estimates with and without observations for both male and female students. It should be noted that the least squares

Table 3.
Least Squares and M-estimate results with and without observation 11.

	Least Squares for 15 observation		M estimator for 15 observation		**Least Squares without observation 11**		**M-estimator without observation 11**	
Term	Estimate	Std.Er.	Estimate	Std.Er.	Estimate	Std.Er.	Estimate	Std.Er.
Constant	32,807	22,01	12,224	14,05	11,257	13,03	11,984	13,65
GMat	0,235	0,30	0,572	0,19	0,586	0,18	0,576	0,19
R^2	**0,044**		**0,466**		**0,461**		**0,468**	

method without outliers fits approximately the same as M-estimates method based on full data.

Normal Q-Q plots may provide a suitable approach for a researcher to detect outliers and to gauge goodness of fit. Realizing that Q-Q plots and other graphical techniques are highly subjective, more formal tests were required in order to identity a plausible distribution for the data as well as to identify outliers, inliers, and other data anomalies (Tiku & Akkaya, 2004).

For the sub-dimension of exhaustion, of the 1385, 85 weights are close to zero, and the rest 1300 observations are close to 1. The M-estimate provides a better coefficient of determination value than the least squares method, despite outliers. Table 8 shows the results for the sub-dimension of desensitization.

Table 6.
Least Squares and M-Estimate Results without Outliers

	Results for 99 male students				Results for 121 female students			
Independent variables	LS Estimate	Std. Err.	M Estimate	Std. Err.	LS Estimate	Std. Err.	M Estimate	Std. Err.
CONSTANT	114,250	11,24	115,305	11,84	110,743	10,74	110,126	11,45
SK	0,142	0,20	0,154	0,21	0,610	0,17	0,613	0,19
AB	0,269	0,06	0,255	0,06	0,156	0,05	0,158	0,06
AYI	-0,481	0,24	-0,501	0,25	-0,953	0,18	-0,972	0,19
R^2	0,301		0,316		0,342		0,372	

Burnout Level Data

For this data, regression analysis was conducted separately for MBI-SS' sub-dimensions of exhaustion, desensitization, and self-efficacy. The results are shown in Table 7 for the sub-dimension of exhaustion.

Table 7.
Comparison of Estimators for the sub-Dimension of Exhaustion

Independent Variables	LS Estimate	Std. Err.	M-Estimate	Std. Err.
CONSTANT	14,930	1,22	14,375	1,18
SELF-CONFIDENT APPROACH	0,008	0,03	0,028	0,03
HELPLESS APPROACH	0,063	0,03	0,050	0,03
SUBMISSIVE APPROACH	0,137	0,04	0,144	0,03
OPTIMISTIC APPROACH	-0,277	0,04	-0,279	0,04
SEEKING OF SOCIAL SUPPORT	-0,082	0,04	-0,066	0,04
ACADEMIC COMPETENCY	-0,213	0,02	-0,238	0,02
SOCIAL COMPETENCY	0,047	0,02	0,039	0,02
EMOTIONAL COMPETENCY	0,063	0,02	0,046	0,02
TEST ANXIETY	0,043	0,01	0,039	0,01
DEMOCRATIC ATTITUDE	-0,010	0,01	0,007	0,01
PROTECTIVE HEADREQUESTOR ATTITUDE	0,029	0,01	0,033	0,01
AUTHORITARIAN ATTITUDE	-0,017	0,02	-0,015	0,02
R^2	0,239		0,324	

Table 8.
Comparison of Estimators for the sub-Dimension of Desensitization

Independent Variables	LS Estimate	Std. Err.	M-Estimate	Std. Err.
CONSTANT	9,437	1,02	9,990	0,95
SELF-CONFIDENT APPROACH	0,010	0,03	0,024	0,03
HELPLESS APPROACH	0,018	0,03	0,015	0,02
SUBMISSIVE APPROACH	0,184	0,03	0,163	0,03
OPTIMISTIC APPROACH	-0,127	0,03	-0,138	0,03
SEEKING OF SOCIAL SUPPORT	-0,080	0,04	-0,061	0,03
ACADEMIC COMPETENCY	-0,135	0,02	-0,149	0,02
SOCIAL COMPETENCY	0,018	0,02	-0,004	0,02
EMOTIONAL COMPETENCY	0,046	0,02	0,030	0,02
TEST ANXIETY	0,039	0,01	0,039	0,01
DEMOCRATIC ATTITUDE	-0,016	0,01	-0,009	0,01
PROTECTIVE HEADREQUESTOR ATTITUDE	-0,008	0,01	-0,010	0,01
AUTHORITARIAN ATTITUDE	0,028	0,02	0,034	0,02
R^2	0,203		0,302	

Here, 72 of the observations hold weights of near zero, and again, the robust coefficient of determination is better than the classical one. Table 9 shows the results for the sub-dimension of self-efficacy.

Table 9.
Comparison of Estimators for the sub-Dimension of Self-efficacy

Independent Variables	LS Estimate	Std. Err.	M-Estimate	Std. Err.
CONSTANT	4,139	0,89	3,743	0,91
SELF-CONFIDENT APPROACH	0,080	0,02	0,083	0,03
HELPLESS APPROACH	-0,041	0,02	-0,047	0,02
SUBMISSIVE APPROACH	0,055	0,03	0,067	0,03
OPTIMISTIC APPROACH	0,111	0,03	0,125	0,03
SEEKING OF SOCIAL SUPPORT	-0,205	0,03	-0,223	0,03
ACADEMIC COMPETENCY	0,171	0,02	0,178	0,02
SOCIAL COMPETENCY	0,096	0,02	0,111	0,02
EMOTIONAL COMPETENCY	0,003	0,02	-0,001	0,02
TEST ANXIETY	-0,014	0,01	-0,014	0,01
DEMOCRATIC ATTITUDE	0,024	0,01	0,023	0,01
PROTECTIVE HEADREQUESTOR ATTITUDE	-0,005	0,01	-0,007	0,01
AUTHORITARIAN ATTITUDE	0,022	0,01	0,022	0,01
R^2	0,288		0,344	

In contrast to the previous sub-dimensions, here, only eight observations have very low weight. The coefficient of determination for the least square method was affected too much by theese outliers.

Discussion

Regression analysis is an important statistical tool routinely applied in science. Out of the many possible regression techniques, the least squares method has been generally adopted due to tradition and the ease of computation it provides (Maronna et al., 2006; Rousseeuw & Leroy, 1987). However, the techniques used and assumptions are of equal importance (Huber, 1981).

To assess the quality of the fit in a multiple linear regression, the coefficient of determination, or R^2, despite being a simple tool, is the most used by practitioners. Indeed, it is reported in most statistical analyzes, and although it is not recommended as a final model selection tool, it does provide an indication of the suitability of the chosen explanatory variables in predicting the response. In the classical setting, it is well known that the least-squares fit and coefficient of determination can be arbitrary and/or misleading in the presence of a single outlier. In many applied settings, both the assumption of normality of the errors and the absence of outliers are difficult to establish. In these cases, robust procedures for estimation and inference in linear regression are available thereby providing a suitable alternative (Renaud & Feser, 2010).

In this paper, it two important points have been illustrated by means of both hypothetical and real data. These points are the robust coefficient of determination and identifying outliers. The robust coefficient of determination has shown that outliers unduly affect the least squares estimator and that the M-estimator may be a suitable alternative to the least squares method when data contain an outlier or outliers. Because, as shown in the data analysis mentioned, the effects of outliers minimized with the M-estimate and fitted model has a larger R^2 value. This means that the proportion of variation explained by the independent variables is better with fitted model by M-estimator. Moreover, with the M-estimator, it was shown that the M-weights provides researchers to identify outliers and to be able make data analyses without needing to remove the outliers from data sets.

On the other hand, data may contain outliers in x and/or y directions in which the M-estimator may not be robust in regards to outliers in x direction. In such a situation, GM-estimations, which are robust to the outliers in both x and y directions, may be the more appropriate estimator for this kind of data (Arslan & Billor, 1996; Coşkuntuncel, 2010).

References/Kaynakça

Aktaş, C. (2005). Türkiye'nin turizm gelirlerini etkileyen değişkenler için en uygun regresyon denkleminin belirlenmesi. *Doğuş Üniversitesi Dergisi, 6*(2), 163-174.

Aluçdibi, F. ve Ekici, G. (2012). Ortaöğretim öğrencilerinin biyoloji dersi motivasyon düzeylerine biyoloji öğretmenlerinin sınıf yönetimi profillerinin etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi dergisi, 43*, 25-36.

Arıcı, H. (1991). *İstatistik: Yöntem ve uygulamalar.* Ankara: Meteksan.

Arslan, O. (1992). *Multivariate robust analysis based on the t-distribution and the EM algorithm* (Doctorial thesis, University of Leeds, Department of Statistics, Leeds, U.K.).

Arslan, O. (2004a). Convergence behavior of an iterative reweighting algorithm to compute multivariate M-estimates for location and scatter. *Journal of Statistical Planning and Inference, 118*, 115-128.

Arslan, O. (2004b). Family of multivarate generalized t distribution. *Journal of Multivariate Analysis, 89,* 329-337.

Arslan, O., & Billor, N. (1996). Robust ridge regression estimation based on the GM-estimators. *Journal of Mathematics & Computer Sciences (Math. Ser.), 9*(1), 1-9.

Arslan, O., & Billor, N. (2000). Robust liu estimator for regression based on an M-estimators. *Journal of Applied Statistics, 27*(1), 39-47.

Arslan, O., Edlund, H., & Ekblom, H. (2001). Algorithms to compute CM- and S-estimates for regression. *Metrika, 55*, 37-51.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: Wiley.

Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression.* New York: John Wiley & Sons.

Büyüköztürk, Ş. (2005). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem A.

Büyüköztürk, Ş., Çokluk, Ö. ve Köklü, N. (2011). *Sosyal bilimler için istatistik.* Ankara: Pegem A.

Chambers, J., Eddy, W., Hardle, W., Sheather, S., & Tierney, L. (2002). *Basics of S-plus.* New York: Springer-Verlag.

Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression analysis by example.* New York: John Wiley & Sons.

Coşkuntuncel, O. (2005). *Karma denemelerde ve modellerde robust istatistiksel analizler* (Doktora tezi, Çukurova Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Anabilim Dalı, Adana). https://tez.yok.gov.tr adresinden edinilmiştir.

Coşkuntuncel, O. (2009). Eğitimle ilgili sapan değer içeren veri kümelerinde en küçük kareler ve robust-M tahmin edicilerinin karşılaştırılması. *Mersin Ün. Eğitim Fakültesi Dergisi, 5*(2), 251-262.

Coşkuntuncel, O. (2010). Sosyal bilimlerde yanlı regresyon tahmin edicilerin kullanılması. *Eğitim ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(2), 100-108.

Çapulcuoğlu, U. (2012). *Öğrenci tükenmişliğini yordamada stresle başaçıkma, sınav kaygısı, akademik yetkinlik ve anne-baba tutumları değişkenlerinin incelenmesi,* (Yüksek lisans tezi, Mersin Üniversitesi, Eğitim Bilimleri Enstitüsü, Mersin). https://tez.yok.gov.tr adresinden edinilmiştir.

Delgaard, P. (2008). *Introductory statistics with R.* New York: Springer.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3th ed.). New York: John Wiley and Sons.

Gündüz, B. ve Çelikkaleli, Ö. (2009). Ergen saldırganlığında akademik yetkinlik inancı, akran baskısı ve sürekli kaygının rolü. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 10*(2), 19-38.

Güneş, M. ve Tulçal, R. (2002). Faiz oranlarını etkileyen faktörlerin regresyon analizi ile tespiti üzerine bir uygulama. *Trakya Üniversitesi Bilimsel Araştırmalar Dergisi, 2*(1), 49-56.

Hample, F. R. (1968). *Contributions to the theory of robust estimation.* (Doctoral Thesis, Department of Statistics, University of California, Berkeley).

Hample, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influential functions.* New York: Wiley.

Huber, P. J. (1964). Robust estimation of location parameters. *The Annals of Mathematical Statistics, 35*, 73-101.

Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics, 36*, 1753-1758.

Huber, P. J. (1981). *Robust statistics.* New York: John Wiley & Sons.

İnandı, Y. (2009). The barriers to career advancement of female teachers in Turkey and their levels of burnout. *Social Behavior And Personality, 37*(8), 1143-1152.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics.* New York: John Wiley & Sons.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis* (3th ed.). New York: John Wiley and Sons.

R Core Team. (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rahman M. S., & Amri, A. A. (2011). Effect of outlier on coefficient of determination. *International Journal of Education Research, 6*(1), 9-20.

Renaud, O., & Feser, M. P. V. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference, 140,* 1852-1862.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*, 871-880.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection.* New York: Wiley.

Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time series,* J. Franke, W. Hardle & R. D. Martin (Eds.), *Lectures notes in statistics* (v. 26, pp. 256-272). New York: Springer.

Rousseeuw, P. J., & Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of American Statistical Association, 85*(411), 633-651.

Şahin, M. D. ve Anıl, D. (2012). 7. sınıf öğrencilerinin SBS 2010 fen ve teknoloji testi başarılarını etkileyen bazı faktörler [Özel Sayı]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 2*, 162-170.

Tiku, M. L., & Akkaya, A. D. (2004). *Robust estimation and hypothesis testing.* New Delhi: New Age International.

Tukey, J. (1960). A Survey of sampling from contaminated distributions. I. Olkin (Ed.), *Contributions to Probability and Statistics* (pp. 448-485). Standford, CA: Stanford University Press.

Wilcox, R. R. (2005). *Introduction to Robust estimation and hypothesis testing.* U.K.: Elsevier Academic Press.