Available online at www.alphanumericjournal.com

alphanumeric journal

The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 3, Issue 2, 2015

2015.03.02.STAT.05

SOME ROBUST ESTIMATION METHODS AND THEIR APPLICATIONS

Tolga ZAMAN *

Kamil ALAKUŞ †

Research Assistant, Department of Statistics, Faculty of Science and Arts, Ondokuz Mayıs University, Samsun

Assoc. Prof. Dr., Department of Statistics, Faculty of Science and Arts, Ondokuz Mayıs University, Samsun

Received: 16 November2015

Accepted: 26 December 2015

Abstract

This study examines robust regression methods which are used for the solution of problems caused by the situations in which the assumptions of LSM technique, which is commonly used for the prediction of linear regression models, cannot be used. Robust estimators are not influenced by small deviations and discrepancies. For this purpose, some robust regression techniques which are used in situations in which the assumptions cannot be made were introduced and parameter estimation algorithms of these techniques were analyzed. Regression models of the methods of Lad, Weighted –M regression, Theil regression and Least Median Squares, coefficients of determination and average absolute deviations were calculated and the results were discussed as to which of these methods gave better results.

Keywords: Robust Regression Methods, Least Square Errors Methods, Average Absolute Deviations, Coefficient of Determination

Jel Code: C40

BAZI ROBUST TAHMİN YÖNTEMLERİ VE UYGULAMALARI

Özet

Bu çalışmada doğrusal regresyon modellerinin tahmininde yaygın olarak kullanılan EKK tekniğinin varsayımlarının sağlanmamasından kaynaklanan problemlerin çözümü için kullanılan Robust regresyon yöntemleri incelenmiştir. Robust tahmin ediciler küçük sapmalardan, aykırılıklardan etkilenmezler. Bu amaçla, çalışmada varsayımların sağlanmadığı durumlarda kullanılan bazı robust regresyon teknikleri tanıtılmıştır ve bu tekniklere ait parametre tahmin algoritmaları incelenmiştir. Uygulamada Lad, Ağırlıklı –M regresyon, Theil regresyon ve En küçük Medyan Kareler yöntemlerine ait regresyon modeli, belirleme katsayıları ve ortalama mutlak sapmalar hesaplanmış olup, bu tahmin edicilerden hangisinin daha iyi sonuç verdiği tartışılmıştır.

Anahtar Kelimeler : Robust Regresyon Methodları, En Küçük Kareler Methodu, Ortalama Mutlak Sapma, Belirleme Katsayısı

Jel Kodu : C40

1. INTRODUCTION

Nowadays, with statistical analysis becoming more and more important, LSM method still continues to be one of the most used methods among regression parameters estimation techniques. However, when a data set has an outlier, using LSM method by excluding these outliers from the data or including them as they are may give wrong results. In that case, using regression methods which will decrease the effect of outliers will yield more reliable results. Studies on robust estimators started when

* tolga.zaman@omu.edu.tr (Corresponding author)

† kamilal@omu.edu.tr

the Least Absolute Deviation (LAD, L1) regression technique was put forward by Roger Joseph Boscovich in 1757. However, it was not used much since it was too long and complicated to calculate (Birkes D. and Dodge,Y. 1993). Later, with the developments in computer programming, studies on robust regression started again. Tukey in 1960 and Huber in 1964 studied regression and Huber who studied theoretically between the years 1972 and 1973 was followed by Hampel with his studies between 1973 and 1978 (Neter, J., Kutner, M.H., Nachtsheim, 1993). In a simple linear model, Theil (1950) proposed the median of pairwise slopes as an estimator of the slope parameter. Later, Sen (1968) extended this estimator to handle ties. The Theil-Sen estimator (TSE) is robust with a high breakdown point 29.3%, has a bounded influence function, and possesses a high asymptotic efficiency. Thus it is very competitive to other slope estimators (e.g., the least squares estimators), see (Sen, 1968, Dietz,1989and Wilcox,1998). The TSE has been acknowledged in several popular textbooks on nonparametric and robust statistics, e.g., (Sprent,1993), (Rousseeuw and Leroy 1986).

2. PARAMETER ESTIMATION

2.1. Estimation of regression parameters with the help of Least Absolute Deviations Method (Lad, L1)

LSM method is calculated in a way that $\hat{\beta}_0$ and $\hat{\beta}_1$estimators minimize the total of error squares (Genceli, 2001). Least Absolute Deviations Method is a method that minimizes the total of absolute errors and it is stated as follows:

$$min \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$$

There is no mathematical expression to calculate estimators with Least Absolute Deviations Method. Thus, an algorithm has been developed to calculate L1 estimators. The basis of the algorithm aims to find the best line among all the lines that pass from a given $(x_0, y_0$ line.

The following steps are followed in finding out the regression line for L2 technique (Yorulmaz, 2003):

1. Generally, the first of observation pairs is chosen.
2. By using the observation pair chosen, slope values for each observation pair and the corresponding $x_i - x_0$ values are obtained.
3. The absolute values of $x_i - x_0$ values which correspond to slope values ordered from the smallest to the biggest are found.
4. The cumulative sum of the $x_i - x_0$ values found is calculated.
5. Half of the cumulative sum found in the previous step equals the critical value.
6. To find the slope value which equals the critical value, the observation value in the third step is referred to. The first observation value higher than the critical value is the point looked for. The slope value of the corresponding value is checked. This value is the value found in the third step.
7. The original order of the point which gives this slope value is calculated. This point is the new starting point for the next step.
8. When two consequent same values are found as a result of such iterations, the process is stopped.

2.2. Estimation of Regression Parameters through Weighted M-Regression Technique.

In Huber M- Regression Technique, $\rho(z)$, which is the function of error terms, is minimized. Thus, when the $\rho(z)$ function is defined for error terms in the technique proposed by Huber (1973), the following is found;

$$\rho(\varepsilon) = \begin{cases} \varepsilon^2, -k \le \varepsilon \le k \\ 2k|\varepsilon| - k^2, \varepsilon < -k \vee k < \varepsilon \end{cases}$$

(Jabr, 2005). Here, $k = 1{,}5 * MSM$ and calculated as

$$MSM = \frac{Med\{|\varepsilon_i - med(\varepsilon_i)|\}}{0{,}6745}, i = 1{,}2, \dots, n$$

Here Med (.) shows the median value.

In Huber's M- Regression Technique, parameter estimations can also be calculated by using Huber weight function. The expression $\sum_{i=1}^{n} \varepsilon_i{}^2 \rightarrow$ is minimized by LSM. When w_i weights are also taken into consideration, the minimum function will be as $min \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$. Some important weights are given as summarized in Table 1.

Table 1. Some weight functions for the estimation of simple liner regression model.

Name of the method	Weight function
Huber M- Weighted Regression	$w_i = \begin{cases} 1, & \lvert r\rvert \leq 1,5 \\ \frac{1,5}{\lvert r\rvert}, & \lvert r\rvert > 1,5 \end{cases}$
Hampel Weighted Regression	$w_i = \begin{cases} 1, & 0 < \lvert r\rvert \leq 1,7 \\ \frac{1,7}{r} sgn(r), & 1,7 < \lvert r\rvert \leq 3,4 \\ \frac{1,7}{r}\left[\frac{8,5-\lvert r\rvert}{5,1}\right] sgn(r), & 3,4 < r \leq 8,5 \\ 0, & 8,5 < \lvert r\rvert \end{cases}$
Andrews Weighted Regression	$w_i = \begin{cases} \frac{\sin(\frac{r}{1,5})}{r}, & \lvert r\rvert \leq 1,5\pi \\ 0, & \lvert r\rvert > 1,5\,\pi \end{cases}$
Tukey Weighted Regression	$w_i = \begin{cases} (1-(\frac{r}{5}))^2, & \lvert r\rvert \leq 1,5 \\ 0, & \lvert r\rvert > 1,5 \end{cases}$

The r value in the functions given in Table 1 is calculated as $r = \frac{\varepsilon_i}{MSM}$. sgn(.) in the Hampel weighted

method is the sign function and it is expressed as $sgn(x) = \begin{cases} -1, x < 0 \\ 0, x = 0 \\ 1, x > 0 \end{cases}$. $sin(.)$ in Andrews Weighted Regression shows the sine value.

The following steps are followed in finding out the regression line for Weighted M-Regression techniques:

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ estimation values are found through LSM method.
2. Next, MSM and ε_i values are found by using these estimation values.
3. Weight values are calculated.
4. $\hat{\beta}_{00}$ and $\hat{\beta}_{10}$ estimation values are found through weighted LSM method.
5. The process is finished if the difference between estimations is < 0,001(Ergül, B., 2006).

2.3. Estimation of Regression Parameters through Theil-Sen Method.

Theil-Sen method is also expressed as Theill-Kendall or Theil method in literature. Brown-Mood method which is recommended for finding the slope is a fast, but not very reliable method. Thus, Theil method, which is especially recommended to find the slope coefficient, is more useful. In this method, the linear regression model is expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

Here, β_0 is the cut parameter, while β_1 is the slope parameter and these parameters are estimated. There are some assumptions to estimate these parameters of the simple linear regression. These assumptions are:

1. For each X_i value, a lower mass of Y's and εi's are mutually independent.
2. X_i 's are non-repetitive and they are in $X_1 < X_2 < \cdots < X_n$ line.
3. The data set consists of n observation pairs as $(X_1, Y_1), \ldots, (X_n, Y_n)$.

In line with these assumptions, all the possible $Sij = \frac{(Y_j - Y_i)}{(X_j - X_i)}$ slopes $(for i < j)$ are calculated to reach $\hat{\beta}_1$ estimation. $N = \binom{n}{2}$ Sij slopes are obtained. $\hat{\beta}_1$ estimation is calculated as the median of Sij values. That is, if $\hat{\beta}_1 = Median(Sij)$ and a constant term, $\hat{\beta}_0 = Median(Y) - \hat{\beta}_1 Median(X)$ (Kıroğlu, 2001). In addition, there are other methods to calculate $\hat{\beta}_0$ estimation. (Wilcox, 2013), (Granato, 2006) and (Erilli and Alakus, 2014) can be seen for these methods.

2.4. Estimation of Regression Parameters through Least Median of Squares Method.

Least Median of Squares regression is a robust method used to find out outliers. It was put forward by Rousseeuw and developed by Rousseeuw and Leroy. The method has the idea of minimizing median of error squares instead of sum of error squares. The function to be minimized is given as follows:

$$minmedian({\varepsilon_i}^2)$$

(Rousseeuw and Leroy, 1987).

This estimator is robust for outliers in the direction of both x and y. Breakdown point is 0.5 and it has the highest possible breakdown point (Rousseeuw and Leroy, 1987).

The following steps are followed in finding out the regression line for Least Median of Squares method:

1. $\hat{\beta}_0$ and $\hat{\beta}_1$estimation values are calculated for all point pairs.
2. For each calculated $\hat{\beta}_0$ and $\hat{\beta}_1$ value, error terms with n number of observation pairs are found and the median is found by squaring these error terms.
3. $\hat{\beta}_0$ and $\hat{\beta}_1$estimation values which correspond to the least median of squares value within the calculated median of squares are taken.
4. Weighted LSM technique is applied by using the weighted values in the fifth step. For the method, the weights are obtained with the following expression:
5. $w_i = \begin{cases} 1, \left|\frac{\varepsilon_i}{s_0}\right| \leq 2,5 \\ 0, \left|\frac{\varepsilon_i}{s_0}\right| > 2,5 \end{cases}$

and $s_0 = 1,4826 * \left[1 + \frac{5}{n-p}\right] * \sqrt{med(\varepsilon_i{}^2)}$

and the coefficient of determination is found as; $R^2 = 1 - \left(\frac{med|\varepsilon_i|}{mad(y_i)}\right)$.

Here, $mad(y_i) = med\{|y_i - medy_j|\}$ (Rousseeuw and Leroy, 1987).

3. REAL DATA EXAMPLE

In this practice, rainfall between the years 1970 and 1975 and annual sugar production yields are discussed. The response variable (Y) was taken as yield, while the independent variable was taken as rainfall (X) (Clarke and Cooke, 1992). Assumptions should be proved to be able to apply the LSM method. We can check the Q-Q graph of error terms in order to be able to check visually whether normal distribution assumption is proved.

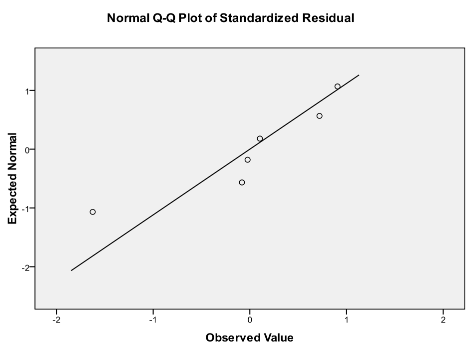

Figure.1 Q-Q graph of the error terms found in the practice

When Figure 1 is analyzed, it can obviously be seen that although Q-Q graph is one of the test methods for goodness of fit, results can be misleading in such small size samples. In samples of such sizes, both visual and other goodness of fit test can give misleading results. For example, although the data seems to have normal distribution, using robust methods rather than LSM method will give more reliable results.

Parameter estimation results for the simple linear regression model L1 technique given with Model (1) are as summarized in Table 2.

Table 2. Analysis results for L1 technique

Results of the first iteration						
y_i	x_i	m	Ordered m	$x_i - x_0$	$\lvert x_i - x_0 \rvert$	cluster $\lvert x_i - x_0 \rvert$
63	20	*	*	*	*	*
77	26	2,333333	-4,5	6	4	4
61	17	0,666667	0,166667	-3	6	10
73	**22**	**5**	**0,666667**	**2**	**3**	**13**
45	24	-4,5	2,333333	4	6	19
62	14	0,166667	5	-6	2	21
$(x_0, y_0) = (20{,}63)$					$\boldsymbol{Criticalvalue = 21/2 = 10.5}$	
Results of the first iteration						
63	20	0,166667	-1,7	6	10	10
77	26	1,25	-0,33333	12	3	13
61	17	-0,33333	0,166667	3	6	19
73	**22**	**1,375**	**1,25**	**8**	**12**	**31**
45	24	-1,7	1,375	10	8	39
62	14	*	*	*	*	*
$(x_0, y_0) = (14{,}62)$					$\boldsymbol{Criticalvalue = 39/2 = 19.5}$	

For the Lad Technique, iterations were continued until the same slope value was found. Finally, as a result of the 3rd and 4th iteration, the slopes were found as equal and the process stopped after 4 iteration. $\hat{\beta}_1 = \frac{(y_k - y_0)}{(x_k - x_x)} = 1{,}25$ and $\hat{\beta}_0 = y_0 - \hat{\beta}_1 x_0 = 44{,}5 \rightarrow \hat{Y}_i = 44{,}5 + 1{,}25X_i$. In the light of these results, coefficient of determination is found as $R^2 = \frac{\sum(\hat{Y}_i - \acute{Y})^2}{\sum(Y_i - \acute{Y})^2} = \frac{418{,}8125}{623{,}5} = 0{,}671712$. In other words, according to Lad technique, rainfall accounts for 67,1% of the variance of yield.

Table 3. Huber –M weighted regression results.

Results of the first iteration								
y_i	x_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$\varepsilon_i - med\varepsilon_i$	$\lvert \varepsilon_i - med\varepsilon_i \rvert$	r_i	$\lvert r_i \rvert$	w_i
63	20	63,28643	-0,28643	-0,781407035	0,78141	-0,03917	0,03917	1
77	26	65,84925	11,15075	10,65577889	10,65578	1,524927	1,524927	0,983654
61	17	62,00503	-1,00503	-1,5	1,5	-0,13744	0,13744	1
73	22	64,1407	8,859296	8,364321608	8,364322	1,211557	1,211557	1
45	24	64,99497	-19,995	-20,48994975	20,4899	-2,73442	2,73442	0,548562
62	14	60,72362	1,276382	0,781407035	0,781407	0,174552	0,174552	1
$\boldsymbol{med\varepsilon_i = 0.494975}$				$\boldsymbol{med\lvert \varepsilon_i - med\varepsilon_i \rvert = 4.932161}$			$\boldsymbol{MSM = 7.31232172}$	
Results of the final iteration								
63	20	65,6038779	-2,603877	-2,8827692	2,88276923	-0,4362377	0,43623777	1
77	26	71,4475702	5,5524297	5,27353845	5,27353845	0,93022011	0,93022011	1
61	17	62,6820317	-1,682031	-1,9609230	1,96092307	-0,2817973	0,28179730	1
73	22	67,5517753	5,4482246	5,16933332	5,16933332	0,91276222	0,91276222	1
45	24	69,4996727	-24,49967	-24,778564	24,7785646	-4,1045252	4,10452528	0,36545
62	14	59,7601856	2,2398144	1,960923078	1,96092308	0,37524480	0,375244803	1
$\boldsymbol{med\varepsilon_i = 0.27889133}$				$\boldsymbol{med\lvert \varepsilon_i - med\varepsilon_i \rvert = 4.026051282}$			$\boldsymbol{MSM = 5.968941855}$	

The weight values in Table 3 were found by using the Huber-M weighted technique in Table 1. Later, the best estimation value was found as a result of technique results and first and final iteration analysis results were summarized as in Table 4.

Table 4. Huber –M weighted regression results.

Variable	First iteration			Final Iteration		
	$\hat{\beta}_i$	Std. Error	$t_{calculation}$	$\hat{\beta}_i$	Std. Error	$t_{calculation}$
Constant	49.117	21.289	2.307	46.124	18.461	2.498
Rainfall	0.785	1.033	0.756	0.973	0.900	1.081
Correlation, r	0.355			0.476		
Coefficient of determination, R^2	0.126			0.226		

Thus, the regression equation estimated as a result of the ninth iteration according to Huber –M weight regression technique was calculated as $\hat{Y}_i = 46.124 + 0.973X_i$ and the amount of rainfall explains 22.6% of the yield according to Huber –M weight regression method.

Table 5. Hampel –M weight regression results.

y_i	x_i	$\hat{y}_i$	$\varepsilon_i = y_i - \hat{y}_i$	$\varepsilon_i - med\varepsilon_i$	$\lvert\varepsilon_i - med\varepsilon_i\rvert$	r_i	$\lvert r_i\rvert$	w_i
First iteration results								
63	20	63,28643	-0,28643216	-0,78140703	0,78141	-0,0391711	0,03917	1
77	26	65,84925	11,15075377	10,65577889	10,65578	1,52492658	1,524927	1
61	17	62,00503	-1,00502512	-1,5	1,5	-0,1374426	0,13744	1
73	22	64,1407	8,859296485	8,364321608	8,364322	1,21155726	1,211557	1
45	24	64,99497	-19,9949749	-20,4899497	20,4899	-2,7344222	2,73442	0,621703553
62	14	60,72362	1,276381912	0,781407035	0,781407	0,17455220	0,174552	1
$\boldsymbol{med\varepsilon_i = 0.4949748877}$				$\boldsymbol{med\lvert\varepsilon_i - med\varepsilon_i\rvert = 4.93216}$			$\boldsymbol{MSM = 7.31232172}$	
Final Iteration results								
63	20	65,67102	-2,671019	-2,9540595	2,9540595	-0,4492539	0,449253	1
77	26	71,60977	5,390235	5,1071945	5,1071945	0,90661447	0,906614	1
61	17	62,70165	-1,701646	-1,9846865	1,9846865	-0,2862095	0,286209	1
73	22	67,6506	5,349399	5,0663585	5,0663585	0,89974603	0,899746	1
45	24	69,63018	-24,630183	-24,9132235	24,9132235	-4,1426914	4,142691	0,350602071
62	14	59,73227	2,267727	1,9846865	1,9846865	0,38142198	0,381421	1
$\boldsymbol{med\varepsilon_i = 0,2830405}$				$\boldsymbol{med\lvert\varepsilon_i - med\varepsilon_i\rvert = 4,010209}$			$\boldsymbol{MSM = 5,945454411}$	

The weight values in Table 5 were calculated by using the weight function of Hampel –M weight regression technique in Table 1 and the results of the information obtained as a result of 16 iterations were summarized in Table 6.

Table 6. Hampel–M weight regression results.

Variable	First Iteration			Final Iteration		
	$\hat{\beta}_i$	Std. Error	$t_{calculation}$	$\hat{\beta}_i$	Std. Error	$t_{calculation}$
Fixed	50.034	22.185	2.255	45.876	18.191	2.522
Amount of rainfall	0.726	1.073	0.676	0.989	0.887	1.115
Correlation, r	0.320			0.487		
Coefficient of determination, R^2	0.103			0.237		

ISSN 2148-2225
http://www.alphanumericjournal.com/

Thus, the regression equation estimated as a result of the 16 iterations for Hampel –M weight regression technique is $\hat{Y}_i = 45{,}875 + 0.989X_i$ and according to this technique, the amount of rainfall as a result of the final iteration explains 23,7% of the variance of yield.

Table 7. Andrews weighted regression results.

y_i	x_i	$\hat{y}_i$	$\varepsilon_i = y_i - \hat{y}_i$	$\varepsilon_i - med\varepsilon_i$	$\lvert\varepsilon_i - med\varepsilon_i\rvert$	r_i	$\lvert r_i\rvert$	$sin\left(\frac{\lvert r_i\rvert}{1,5}\right)$	w_i
First Iteration results									
63	20	63,28643	-0,2864321	-0,781407	0,78141	-0,039171	0,03917	0,018651	0,47616
77	26	65,84925	11,150753	10,655778	10,65578	1,5249265	1,524927	0,664	0,43543
61	17	62,00503	-1,0050251	-1,5	1,5	-0,137442	0,13744	0,0654009	0,47585
73	22	64,1407	8,8592964	8,3643216	8,364322	1,2115572	1,211557	0,545455	0,450209
45	24	64,99497	-19,994974	-20,48995	20,4899	-2,734422	2,73442	0,964119	0,352586
62	14	60,72362	1,2763819	0,7814070	0,781407	0,1745522	0,174552	0,083024	0,47564
$\boldsymbol{med\varepsilon_i = 0,4949748}$				$\boldsymbol{med\lvert\varepsilon_i - med\varepsilon_i\rvert = 4,932161}$				$\boldsymbol{MSM = 7,312321}$	
Final Iteration results									
63	20	64,5258	-1,525882	-1,720983	1,720983	-0,240231	0,240231	0,1141467	0,4751525
77	26	68,8205	8,179474	7,984373	7,984373	1,2877600	1,287760	0,5755030	0,4469023
61	17	62,3785	-1,37856	-1,573661	1,573661	-0,217037	0,217037	0,1031674	0,4753431
73	22	65,9574	7,04257	6,847469	6,847469	1,1087681	1,108768	0,5037936	0,4543723
45	24	67,3889	-22,38897	-22,584079	22,58407	-3,524876	3,524876	0,9942042	0,2820536
62	14	60,2312	1,768762	1,573661	1,573661	0,2784703	0,278470	0,1322166	0,4747961
$\boldsymbol{med\varepsilon_i = 0,195101}$				$\boldsymbol{med\lvert\varepsilon_i - med\varepsilon_i\rvert = 4,284226}$				$\boldsymbol{MSM = 6,351706}$	

The results of the information obtained as a result of 12 iterations were summarized in Table 8.

Table 8. Andrews weighted regression results.

Variable	First Iteration			Final Iteration		
	$\hat{\beta}_i$	Std. Error	$t_{calculation}$	$\hat{\beta}_i$	Std. Error	$t_{calculation}$
Fixed	52.553	23.487	2.237	50.210	21.863	2.297
Amount of rainfall	0.568	1.137	0.499	0.716	1.062	0.673
Correlation, r	0.242			0.319		
Coefficient of determination, R^2	0.059			0.102		

The regression equation estimated as a result of the 12 iterations for Andrews weighted regression is $\hat{Y}_i = 50{,}210 + 0.715X_i$ and according to this technique, the amount of rainfall explains 10,2% of the variance of yield.

Table 9. Tukey weighted regression results.

y_i	x_i	$\hat{y}_i$	$\varepsilon_i = y_i - \hat{y}_i$	$\varepsilon_i - med\varepsilon_i$	$\lvert\varepsilon_i - med\varepsilon_i\rvert$	r_i	$\lvert r_i\rvert$	$1 - (\frac{\lvert r_i\rvert}{5})^2$	w_i
First Iteration results									
63	20	63,2864	-0,286432	-0,78140703	0,781407	-0,039171	0,039171	0,9999386	0,9998772
77	26	65,8492	11,15075	10,65577889	10,65577	1,5249266	1,524926	0,9069839	0,8226198
61	17	62,0050	-1,005025	-1,5	1,5	-0,137442	0,137442	0,9992443	0,9984893
73	22	64,1407	8,859296	8,364321608	8,364321	1,2115573	1,211557	0,9412851	0,8860177
45	24	64,9949	-19,99497	-20,4899497	20,48994	-2,734422	2,734422	0,7009173	0,4912851
62	14	60,7236	1,276381	0,781407035	0,781407	0,1745522	0,174552	0,9987812	0,9975640

$\boldsymbol{med\varepsilon_i = 0,494974}$				$\boldsymbol{med\lvert\varepsilon_i - med\varepsilon_i\rvert = 4,932160}$				$\boldsymbol{MSM = 7,312321}$	
Final Iteration results									
63	20	67,5179	-4,517926	-3,868626	3,868626	-0,877385	0,877385	0,9692077	0,9393637
77	26	76,0434	0,956578	1,605878	1,605878	0,1857683	0,185768	0,9986196	0,9972411
61	17	63,2551	-2,255178	-1,605878	1,605878	-0,437957	0,437957	0,9923277	0,9847143
73	22	70,3597	2,640242	3,289542	3,289542	0,5127375	0,512737	0,9894840	0,9790786
45	24	73,2015	-28,20159	-27,55229	27,55229	-5,476776	5,476776	-0,1998031	0
62	14	58,9924	3,00757	3,65687	3,65687	0,5840730	0,584073	0,9863543	0,9728948
$\boldsymbol{med\varepsilon_i = -0,6493}$				$\boldsymbol{med\lvert\varepsilon_i - med\varepsilon_i\rvert = 3,473206}$				$\boldsymbol{MSM = 5,149304}$	

Table 10. Tukey weighted regression results

Variable	First Iteration			Final Iteration		
	$\hat{\beta}_i$	Std. Error	$t_{calculation}$	$\hat{\beta}_i$	Std. Error	$t_{calculation}$
Fixed	49.802	20.609	2.416	39.099	8.163	4.789
Amount of rainfall	0.744	1.013	0.734	1.420	0.403	3.524
Correlation, r	0.345			0.898		
Coefficient of determination, R^2	0.119			0.806		

The weight values in Table 9 were calculated by using the weight function Tukey weighted regression technique in Table 1 and the results obtained as a result of the 7 iterations were summarized in Table 10. Thus, the regression model estimated as a result of the 7 iterations for Tukey weighted regression method is $\hat{Y}_i = 39.099 + 1,420X_i$ and according to this technique, the amount of rainfall explains 80,6% of the variance of yield.

Table 11. LMS regression results

y_i	x_i	$\hat{\beta}_1$	$\hat{\beta}_0$	First $\hat{\beta}_0$ and $\hat{\beta}_1$ Results			15th $\hat{\beta}_0$ and $\hat{\beta}_1$ Results		
				$\hat{y}_i$	ε_i	ε_i^2	$\hat{y}_i$	ε_i	ε_i^2
63	20	2,333	16,333	63	0	0	51,8	11,2	125,44
77	26	0,667	49,667	77	0	0	41,6	35,4	1253,16
61	17	5	-37	56	5	25	56,9	4,1	16,81
73	22	-4,5	153	67,667	5,333	28,44444	48,4	24,6	605,16
45	24	0,167	59,667	72,333	-27,333	747,1111	45	0	0
62	14	1,778	30,778	49	13	169	62	0	0
		1	51		$med\varepsilon_i{}^2$	26,72222		$med\varepsilon_i{}^2$	71,125
		16	-339						
		1,25	44,5						
		2,4	20,2						
		-2,286	99,857						
		-0,333	66,6667						
		-14	381						
		1,375	42,75						
		-1,7	85,8						

Table 12**.** Median results of error squares in LMS regression analysis.

$\hat{\beta}_0$ and $\hat{\beta}_1$ Values	$\hat{\beta}_1$	$\hat{\beta}_0$	$med\varepsilon_i^2$
1.	2,333333	16,33333	26,722222
2.	0,666667	49,66667	42,055556
3.	5	-37	212,5
4.	-4,5	153	300,625
5.	0,166667	59,66667	47,847222
6.	**1,777778**	**30,77778**	**10,395062**
7.	1	51	29
8.	16	-339	5162
9.	1,25	44,5	11,78125
10.	2,4	20,2	29,2
11.	-2,28571	99,85714	56,377551
12.	-0,33333	66,66667	97,888889
13.	-14	381	2522
14.	1,375	42,75	14,257813
15.	-1,7	85,8	71,125

By using the slope information of the line, it was calculated through $\hat{\beta}_1 = \frac{y_j - y_i}{x_j - x_i}$, $i = 0 < j$ and $\hat{\beta}_0 = y_0 - \hat{\beta}_1 x_0$ for all possible situations. $med\varepsilon_i{}^2$ value was calculated for all possible data pairs. In the next step, $\hat{\beta}_0$ and $\hat{\beta}_1$ estimation coefficients with $minmed\varepsilon_i{}^2$ value were calculated. In the light of this information, ($\binom{6}{2} = 15$) $\hat{\beta}_0$ and $\hat{\beta}_1$ were calculated for all possible situations in Table 11. Later, the median of the error squares of these regression parameters were found as in Table 12 and estimation values which had $minmed\varepsilon_i{}^2$ value were expressed as regression coefficients for LMS.

As a result, the regression line of LMS was obtained as $\hat{Y}_i = 30{,}778 + 1{,}778x_i$. Coefficient of determinacy was calculated as $R^2 = 1 - \left(\frac{med|\varepsilon_i|}{mad(y_i)}\right)^2 = 1 - \left(\frac{3{,}2222}{6}\right)^2 = 0{,}711$ and according to this method, the amount of rainfall explains 71,1% of the variance of yield.

Table 13. Weighted LSM technique for LMS method.

y_i	x_i	$\hat{y}_i$	ε_i	ε_i^2	$\frac{\varepsilon_i}{s_0}$	$\left\lvert\frac{\varepsilon_i}{s_0}\right\rvert$	w_i
63	20	66,33338	-3,33338	11,11142	-0,30993087	0,309930879	1
77	26	77,00006	-6E-05	3,6E-09	-5,57868E-0	5,57868E-0	1
61	17	61,00004	-4E-05	1,6E-09	-3,71912E-0	3,71912E-0	1
73	22	69,88894	3,11106	9,678694	0,289260018	0,289260018	1
45	24	73,4445	-28,4445	809,0896	-2,64471163	2,64471163	0
62	14	55,6667	6,3333	40,11069	0,588857326	0,588857326	1
			$med\varepsilon_i{}^2$	10,39506			
			$\sqrt{med\varepsilon_i{}^2}$	3,224137			
			$1+\frac{5}{n-p}$	2,25			
			s_0	10,75524			

Table 14. Weighted LSM technique for LMS method.

Variable	$\hat{\beta}_i$	Std. Error	$t_{calculation}$
Fixed	39.134	8.258	4.739
Amount of rainfall	1.417	0.408	3.473
Correlation, r	0.895		
Coefficient of determination, R^2	0.801		

Regression coefficients in weighted LSM technique for LMS method were calculated by using regression coefficients obtained by LMS technique and according to this method, the amount of rainfall explains 80,1% of the variance of yield.

When Table 11 is examined for Theil method, the median of all possible slopes were taken to reach $\hat{\beta}_1$estimation and it was calculated as 1. It is calculated as $\hat{\beta}_0 = Median(Y) - \hat{\beta}_1 Median(X) = 62.5 - 1 * 21 = 41.5$

4. CONCLUSION AND RECOMMENDATIONS

In this study, regression line, standard error, coefficients of determination and average absolute deviations were calculated and interpreted for regression models and parameter estimations of techniques used on real life data by using simple linear robust regression techniques. According to the results, the method which gave the best result in terms of the percentage of independent variable explaining the dependent variable was Tukey-weighted regression method. Although weighted least median of squares method was close to Tukey-weighted regression method, its R^2 was found to be a bit lower. The percentage of explanations obtained by non-weighted least median of squares method was calculated as $R^2 = 0{,}712$. However, when the methods analyzed were taken into consideration, it was seen that Tukey, least median of squares and Lad methods gave significantly better results than the other regression models analyzed. In the light of this information, it is seen that Tukey, least median of squares and Lad methods gave more reliable results than LSM method. In addition, when average absolute deviation values ($OMS = \frac{\sum_{i=1}^{n}|(Y_i - \hat{Y}_i|}{n}$) were taken into consideration for the methods, it can be said that the techniques which have high coefficients of determination have lower average absolute deviations.

The summary of the information about the methods used are as follows:

Table 15. Summary Information

Method	Estimation Equation	Average Absolute Deviation
LSE	$\hat{Y}_i = 54.744 + 0.427X_i$	7.095
LAD & L1	$\hat{Y}_i = 44.500 + 1.250X_i$	6.958
Huber M Weighted Reg.	$\hat{Y}_i = 54.124 + 0.974X_i$	7.004
Hampel-M Weighted Reg.	$\hat{Y}_i = 45.875 + 0.989X_i$	7.001
Andrews Weighted Reg.	$\hat{Y}_i = 50.210 + 0.716X_i$	7.047
Tukey Weighted Reg.	$\hat{Y}_i = 39.099 + 1.420X_i$	6.929
LMS	$\hat{Y}_i = 30{,}777 + 1{,}778X_i$	6.870
Weighted LMS	$\hat{Y}_i = 39{,}134 + 1{,}417X_i$	6.930
Theil Reg.	$\hat{Y}_i = 41{,}500 + 1.000X_i$	8.330

The results obtained and our interpretations are valid for the data set we used. No generalizations can be made. Robust regression methods for simple linear regression were analyzed in this study. Similarly, studies can be made on robust methods for multiple linear regression. In future studies, it can be recommended to be used together with the robust methods we discussed with jackknife method.

References

Birkes, D. and Dodge, Y. (1993), Alternative Methods of Regression. NY: Wiley.

Clarke, G. M. and Cooke, D., (1992). A Basic Course in Statistics. 3rd Edition. P. 354-356, Exercises on Chapter 20, Excercis No: 9.

Dietz, E. J. (1989), Teaching Regression in a Nonparametric Statistics Course. The American Statistician. 43, 35-40.

Ergül, B. (2006), Robust Regresyon ve Uygulamaları. Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı Yüksek Lisans Tezi, Eskişehir.

Granato, G. E. (2006), Kendall-Theil Robust Line (KTRLine-version 1.0) –A Visual Basic Program for Calculating and Graphing Robust Nonparametric Estimates of Linear-Regression Coefficients Between Two Continous Variables. Chapter 7, Section A, Statistical Anlaysis, Book 4, Hydrologic Analysis and Interpretation. U. S. Geological Survey Techniques and Methods 4-A7.

Genceli, M. (2001), Ekonomide İstatistik İlkeler, İstanbul, Filiz Kitabevi.

Huber, P. J. (1964), Robust Estimation of a Location Parameter. Ann. Math. Statist., 35, 73-101.

Jabr, R. (2005), Power System Huber-M Estimation with Equilaty and Inequality Constraints, Electric Power System Research, 74, 239-246.

Kıroğlu, G. (2001), Uygulamalı Parametrik Olmayan İstatistiksel Yöntemler. Mimar Sinan Üniversitesi Fen-Edebiyat Fakültesi, İstanbul.

Erilli, N. A. and K., Alakus. (2014), Non-Parametrıc Regression Estimation for Data With Equal Values. European Scientific Journal. February. Edition Vol. 10, No.4 ISSN:1857-7881 (Prınt) e-ISSN 1857-7434.

Neter, J., Kutner, M. H., Nachtheim, C. J., Wasserman, W. (1993), Applied Linear Statistical Methods, Wiley.

Rousseeuw, P. J. and Leroy, A. M. (1987), Robust Regression and Outlier Detection. New York: John Wiley& Sons, Inc.

Wilcox, R. R.. (2013), A Heteroscedastic Method for Comparing Regression Lines at Specified Design Points When Using a Robust Regression Estimator. Journal of Data Science 11,281-291.

Sen, P. K. (1968), Estimates of the regression coefficient based on Kendall's tau., J.Amer. Statist. Assoc., 63, 1379-1389.

Sprent, R. (1993), Applied nonparametric statistical methods. 2 nd Ed. CRC Press, NY.

Theil, H. (1950), A rank-invariant method of lineer and polynomial regression analysis, I. Proc. Kon. Ned. Akad. v . Wetensch. A53, 386-392.

Yorulmaz, Ö. (2003), Robust Regresyon ve Mathematica Uygulamaları. Marmara Üniversitesi, Yüksek Lisans Tezi, Ankara.

Wilcox, R. (1998), Simulations on the Theil-Sen regression estimator with right-censored data. Stat.& Prob. Letters 39, 43-47.