**Advances in Artificial Intelligence Research (AAIR)**
**ISSN 2757-7422**
**www.dergipark.com/aair/**

**Vol. 2 (No. 2), pp. 38-44, 2022**
**doi: 10.54569/aair.1032607**
**Published online: Sept 15, 2022**

# An Approach Towards the Least-Squares Method for Simple Linear Regression

Hasan Halit Tali [1], Ceren Çelti [2,*]

[1] Matematik Bölümü, Fen Edebiyat Fakültesi, Haliç Üniversitesi, İstanbul, Türkiye;
[2] Matematik Bölümü, Lisansüstü Eğitim Enstitüsü, Haliç Üniversitesi, İstanbul, Türkiye;

## Abstract

This study approaches the least-squares method for simple linear regression model. The least-squares line does not comply with the data when there are outliers that have deceptive effects on the results in the dataset. The study aims to develop a method for obtaining a line that complies more with the data when there are outliers in the dataset.

***Keywords:*** *Applied mathematics, machine learning, simple linear regression, least-squares method, outliers.*

## 1. Introduction

Simple Linear Regression is a linear regression model that consists of one independent variable and one dependent variable. This model describes the linear relationship between the dependent and independent variables. In other words, the purpose of the model is to find a linear function between the dependent and independent variable. There are different regression methods for determining this function, and the least squares method, which aims to find a linear function that is as compatible as possible with the data, will be used in this study. For the Simple Linear Regression equation:

$$y = \beta_0 + \beta_1 x + e \tag{1}$$

the Least Squares Method is used and by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ values that minimize the equation:

$$q = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{2}$$

and the following simple linear regression model is obtained:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3}$$

[1,2]. However, datasets may contain observations that may have misleading effects on the results. Such observations are called outliers. Outliers can be found in one or more datasets. Outliers are observations that are incorrectly recorded or belong to another group. Therefore, they are not in accordance with the model. Thus, when there are outliers in the data set, the line $\hat{y}$ is incompatible with the data. Therefore, in case of outliers in the data set, the line $\hat{y}$ is incompatible with the data.

In Figure 1, a least squares line is drawn for the points $(1,1.9)$, $(2,1.8)$, $(3,1.6)$, $(4,1.5)$ and $(5,1.3)$**.** The least squares line is congruent with the points since the points are almost on a straight line.

*Corresponding author
*E-mail address:* hasantali@halic.edu.tr

Received: Dec 7, 2021; Received in revised form: Feb 19, 2022; Accepted: 01 Jul 2022.

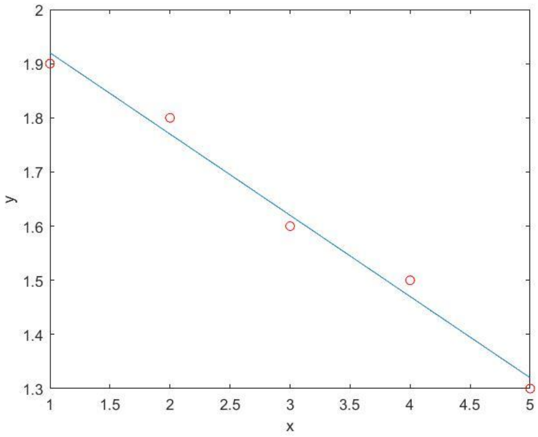

***Figure 1.*** *The least squares line for (x,y) points [3-6]*

However, in Figure 2, it is seen that the Least Squares Line obtained as a result of entering the point $(4{,}1.5)$ as $(4{,}5)$ due to the transfer error is incompatible with the points.

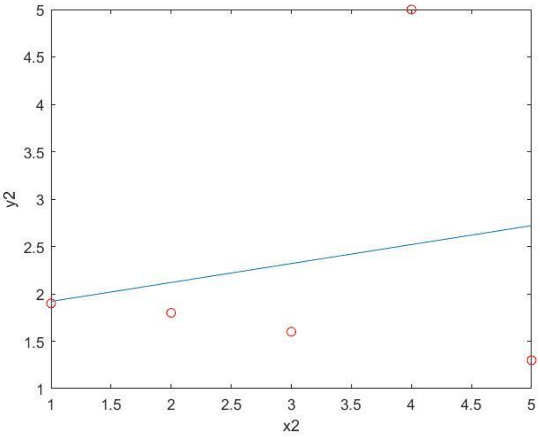

***Figure 2.*** *The least squares line for (x2, y2) points [3-6]*

Outliers may be present in one or more of the datasets and have a large impact on the least squares line, as shown in Figure 2. This situation poses a serious danger to least squares analysis and has attracted attention in the literature. There are basically two ways to eliminate this problem. The first of these is to perform a least square analysis of the remaining observations as a result of detecting outliers and deleting or correcting these values. Many methods are used to detect outliers. Most of these methods rely on the interpretation of $e_i$ residues. Least squares by definition select points with small residuals, but the outlier need not always have a large residual. Sometimes it has small residual and is included in the estimator by least squares. So least squares estimates fail. Another method used to detect outliers is to delete a different point from the dataset each time. The extent to which they affect the regression coefficients is examined by deleting individual data points. This method can be generalized to multiple outlier detection to highlight the simultaneous effect of several outliers by calculating for each case. At first glance it seems like a logical method, but it is not clear which subset of the data should be deleted. Some points may be effective when combined, but not individually. Calculation may not be possible due to the large number of subsets to consider. Another method used for the detection of outliers is the hat matrix. Linear model for $p$ −independent variables and $(n \times 1)$ dependent variable vector $y = (y_1, \dots, y_n)^T$ Eqn. It is expressed as 4 [3].

$$y = X\beta + e \tag{4}$$

$X$ is an $n \times p$ matrix,

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{5}$$

$\beta$ is the unknown vector and $e = (e_1, \dots, e_n)^T$ is the error vector. The matrix $H$, called the hat matrix, is Eqn. It is defined as 6.

$$H = X(X'X)^{-1}X' \tag{6}$$

This matrix pairs the observed values vector to the predicted values vector. The difference between the observed values and the predicted values gives the residuals. Using these residuals, outliers are detected. Many authors such as Hoaglin and Welsch (1978), Henderson and Vellemann (1981), Cook and Weisberg (1982), Hocking (1983), Paul (1983), Stevens (1984) have identified potential hotspots by looking at $H$.

Another approach to solving the problem of outliers for least squares analysis is Robust regression, which tries to design predictors that are not affected much by outliers. The Detection and Robust regression methods have the same goal, but the path they follow for outliers is a little bit different. In detection methods, outliers are detected first. Afterwards, the least squares method is applied to the clean data remaining as a result of deleting or editing these values. In robust regression, first, a correct line is found for most of the data. Points with large residuals on this line are determined as outliers. The next step is to think about the resulting model. The original dataset can be returned to, or the causes of outliers can be investigated using expert knowledge on the subject. Thus, it can be determined whether the deviations are a model error that can be repaired by adding terms or performing some transformations. There are many Robust estimators (Rousseeuw and Leroy, 1987). Edgeworth (1887) noticed that because of squaring the residuals, the least squares method becomes vulnerable to outliers. To deal with this, he proposed a method of minimizing the sum of the absolute values of the residuals rather than the sum of the squares of the residuals. This first $L$ −estimation method, which is more robust than least squares, is Eqn. It is the smallest absolute value regression defined as 7.

$$\widehat{\theta_{LAV}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} |r_i(\theta)| \tag{7}$$

This estimator protects against outliers on the $y$ −axis, but is useless at bad leverage points. This method, which has an efficiency of 64%, is called $L_1$ or Median Regression. Huber (1964) Median Regression, Eqn. Considering functions other than absolute value in 7, he generalized to a larger class of estimators called $M$ −Estimators. $M$ −Estimators protect Robustness against outliers in the $y$ −axis while increasing productivity. With ρ(·) being a symmetrical and less lossy function than the square function,

$$\widehat{\theta_M} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left\{\frac{r_i(\theta)}{\sigma}\right\} \tag{8}$$

Eqn. 8 would be an $M$ −estimator [7].

The $R$ − Estimators studied by Hodges and Lehman (1963) emerged from the inferences made from the rank tests. $R$ − Estimators are based on ordering residual values. $r_i$ residuals, $a_n(i)$ score function and $R_i$ rank of residuals as

$$\min_{\hat{\theta}} \sum_{i=1}^{n} a_n(R_i) r_i \tag{9}$$

Eqn. These expressions, defined by 9, are called $R$ −Estimators [8]. In Siegel's Estimator, another Robust method, a parameter vector $\left((x_{i1}, y_{i1}), \dots, (x_{ip}, y_{ip})\right)$ is calculated for any $p$ observation. This vector's j. coordinate is $\theta_j(i_1, \dots, i_p)$. Siegel's estimator, called the Repeated Median,

$$\widehat{\theta_J} = \operatorname*{med}_{i_1}(\dots(\operatorname*{med}_{i_{p-1}}(\operatorname*{med}_{i_p} \theta_j(i_1, \dots, i_p)))) \tag{10}$$

and defined as Eqn.10. The Least Squares method is known as the least squares sum [3]. As a result of this name, many people have tried to make this estimator robust by changing the squaring process without touching the sum operation. On the contrary, Rousseeuw developed a new method based on Hampel's idea. This method in Eqn. 11. is known as Least Median Squares [7].

$$\min_{\hat{\theta}} \operatorname*{med}_{i} {r_i}^2 \tag{11}$$

The Least Median Squares method is considered to be a very robust method for fitting regression models to the data. Although the breakpoint in this estimator reaches 50%, the estimator has important shortcomings that limit its use. The maximum efficiency of the estimator is 37% [8].

In this study, this study differs from existing methods due to the availability of sections from both approaches, easy programming of the developed method, and calculation speed. At the same time, it has been tried to ensure that the developed method is less affected by outliers than the least squares method. Thus, it is aimed to develop a method that can obtain a more consistent accuracy with the data.

## II. MATERIALS AND METHODS

The Least Squares estimates of the regression coefficients for a $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ dataset, the values minimizing the Eqn. 2. are:

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \tag{12}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \tag{13}$$

values [2]. By finding these values, The Simple Linear Regression model is obtained as Eqn. 3.

In this study, first of all, $\hat{y}$ least squares line is obtained for the dataset $\{(x_i, y_i); i = 1, 2, \ldots, n\}$. Afterwards, the perpendicular distance of each $(x_i, y_i)$ data point to the line $\hat{y} - \hat{\beta}_0 - \hat{\beta}_1 x = 0$ is calculated with:

$$d_i = \frac{|y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|}{\sqrt{{\hat{\beta}_0}^2 + 1}} \tag{14}$$

in Eqn. 14. According to the calculated $d_i$ distances,

$$A_1 = \left\{(x_i, y_i) \colon d_i \leq d_j \; for \;\; i \leq j \; and \; i, j = 1, 2, \ldots, n\right\} \tag{15}$$

has been obtained. In fact, the purpose of creating the $A_1$ set is to work with a certain number of data points that are closest to the $\hat{y}$ least squares line. Accordingly, the set below is obtained with the first $r$ element in the set $A_1$, where $r$ is the $\frac{3n}{8}$ real number rounded to an integer.

$$B = \{(x_i, y_i) \in A_1 \colon i = 1, 2, \ldots, r\} \tag{16}$$

has been obtained. By obtaining set $B$, the $\hat{y}$ least squares line is updated for the elements in set $B$. Then, by recalculating the perpendicular distances of the points in the data set to the $\hat{y}$ line, the dataset

$$A_2 = \left\{(x_i, y_i) \colon d_i \leq d_j \text{ for } i \leq j \; and \; i, j = 1, 2, \ldots, n\right\} \tag{17}$$

has been obtained. Then, $v = d_r + s$ value was determined, with $s$ being the standard deviation of the $d_i$ distances obtained for $i = 1,2, \ldots, n$ in the $A_2$ cluster. This value has been chosen in such a way that it can accept points that are at most one standard deviation away from the $d_r$ distance from the points in the B set to the $\hat{y}$ line.

Afterwards, the points with a distance greater than this value from the set $A_2$ to the $\hat{y}$ line were subtracted and the dataset

$$C = \{(x_i, y_i):\ d_i \leq v\} \tag{18}$$

has been obtained. Lastly, the $\hat{y}$ least squares line created for the data points in set $C$ has been determined as the estimation line, and the results were obtained by using this line.

## 3. Findings and Discussion

This method has been applied to various datasets. One of these is the dataset seen in Table 1, called Pilot-Plant and developed by Daniel and Wood (1971) and it consists of data that gives acid contents determined by titration and organic acid contents determined by core and weight. However, there are no outliers in this dataset. For this reason, let's consider that the x value of the 6th observation is recorded as 370 instead of 37 [3].

***Table 1.*** *Pilot-Plant dataset*

| **Observation** ($i$) | **Extraction** ($x_i$) | **Titration** ($y_i$) | **Observation** ($i$) | **Extraction** ($x_i$) | **Titration** ($y_i$) |
|---|---|---|---|---|---|
| 1 | 123 | 76 | 11 | 138 | 82 |
| 2 | 109 | 70 | 12 | 105 | 68 |
| 3 | 62 | 55 | 13 | 159 | 88 |
| 4 | 104 | 71 | 14 | 75 | 58 |
| 5 | 57 | 55 | 15 | 88 | 64 |
| 6 | 37 | 48 | 16 | 164 | 88 |
| 7 | 44 | 50 | 17 | 169 | 89 |
| 8 | 100 | 66 | 18 | 167 | 88 |
| 9 | 16 | 41 | 19 | 149 | 84 |
| 10 | 28 | 43 | 20 | 167 | 88 |

The least squares line for these distorted data was obtained as $\hat{y} = 0.081x + 58.939$, which is presented in Figure 1 with a dashed line [3-6]. On the other hand, if $r = round\left(\frac{20.3}{8}\right) = 8$ is selected in the new method and for $d_r$ value of the data whose $x$ value is 57 from the set $B$ is calculated with a standard deviation as $v = d_r + s$, $\hat{y} = 1.176x - 47.493$ line is obtained, which is presented in Figure 3 with a straight line [4-6]. It is observed in Figure 3 that the line obtained with the new method for this dataset is compatible with the data points.

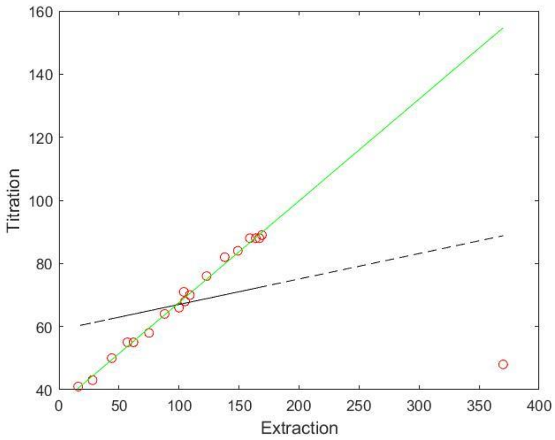

***Figure 3.*** *The least squares line (dashed line) and the new method (straight line) for the distorted Pilot-Plant dataset*

Another dataset is the dataset in Table 2 and this dataset consists of the number of international phone calls made from Belgium by years. Due to the difference in the registration system, the data from 1964 to 1969 contain excessive pollution [3].

**Table 2.** *Dataset of international calls from Belgium by year*

| **Year** $(x_i)$ | ***Number of calls** $(y_i)$ | **Year** $(x_i)$ | *** Number of calls** $(y_i)$ |
|---|---|---|---|
| 50 | 0.44 | 62 | 1.61 |
| 51 | 0.44 | 63 | 2.12 |
| 52 | 047 | 64 | 11.90 |
| 53 | 0.59 | 65 | 12.40 |
| 54 | 0.66 | 66 | 14.20 |
| 55 | 0.73 | 67 | 15.90 |
| 56 | 0.81 | 68 | 18.20 |
| 57 | 0.88 | 69 | 21.20 |
| 58 | 1.06 | 70 | 4.30 |
| 59 | 1.20 | 71 | 2.40 |
| 60 | 1.35 | 72 | 2.70 |
| 61 | 1.49 | 73 | 2.90 |

**One million*

The least squares line for these data was obtained as $\hat{y} = 0.504x - 26.01$, which is presented in Figure 4 with a dashed line [3-6]. On the other hand, if $r = \text{round}\ (\frac{24.3}{8}) = 9$ is selected in the new method and for $d_r$ value of the data whose $x$ value is 59 from the set $B$ is calculated with a standard deviation as $v = d_r + s$, $\hat{y} = 0.13x - 6.35$ line is obtained, which is presented in Figure 4 with a straight line [4-6]. It is observed in Figure 4 that the line obtained with the new method for this dataset is compatible with the data points.

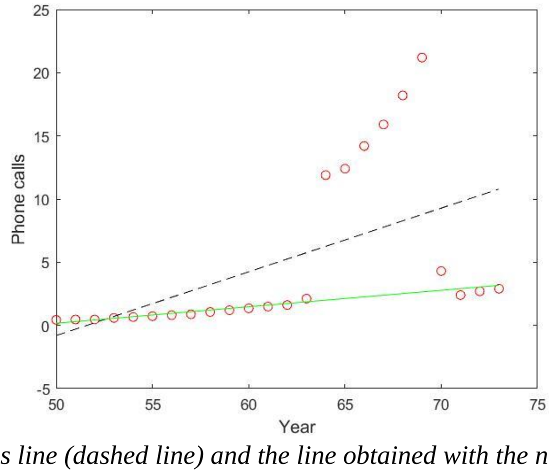

**Figure 4.** *The least squares line (dashed line) and the line obtained with the new method (straight line) for the dataset consisting of international calls made from Belgium by years*

In Figure 5, the line obtained for the $r = 15$ value is shown. However, it can be observed that this line is less compliant with the points than the line obtained for the $r = 9$ value.

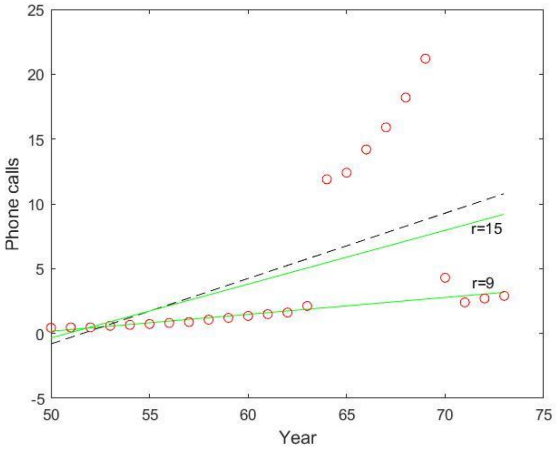

**Figure 5.** *The least squares line (dashed line) for the dataset consisting of the number of international calls made from Belgium by years and the lines (straight lines) obtained by the new method for the values of* $r = 9, r = 15$ *[4-6]*

## 4. Conclusion

In conclusion, a method has been developed in this study, which is expected to be more compatible with the data compared to the least squares line when there are outliers in the data set. This method is advantageous due to its easy programming and computational speed and differs from existing methods. Additionally, the method developed in the study was applied to 2 different data sets, and obtained lines are found to be more compatible with the data compared to the least squares line.

### Declaration of interest

The authors declare that there is no conflict of interest.

### Acknowledgements

It was presented at the ICAIAME 2021 conference and published as a summary.

### References

[1]. Miller I, Miller M. Mathematical Statistics, Prenttice-Hall, Inc, 1999 (Çev. Ümit Şenesen John E. Freund'dan Matematiksel İstatistik, Literatür Yayıncılık. 2007).
[2]. Arslan İ. Python ile Veri Bilimi, Pusula Yayıncılık, Türkiye, 2020.
[3]. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. John Wiley & Sons, 1987.
[4]. Attaway S. Matlab A Practical Introduction to Programming and Problem Solving. 5th ed. Cambridge, USA, Butterwoth-Heinmann, 2019.
[5]. Kubat C. Matlab Yapay Zeka ve Mühendislik Uygulamaları. 5. Baskı, İstanbul, Türkiye, Abaküs Kitap Yayın, 2021.
[6]. Güneş A, Yıldız K. Matlab Matematik ve Grafik Programlama Dili. İstanbul, Türkiye, Türkmen Kitabevi, 1997.
[7]. Verardi V, Croux C. "Robust regression in Stata". The Stata Journal, 9(3), 439-453, 2009.
[8]. Andersen R. Modern methods for robust regression (No. 152). Sage, 2008.