BULLETIN OF ECONOMIC THEORY AND ANALYSIS

Journal homepage: http://www.betajournals.org

Use of Trimean in Theil-Sen Regression Analysis

Necati Alp ERİLLİ https://orcid.org/0000-0001-6948-0880

To cite this article: Erilli, A., N., (2021). Use of Trimean in Theil-Sen Regression Analysis. *Bulletin of Economic Theory and Analysis*, *6*(1), 15-26.

Received: 11 Nov 2020

Accepted: 19 Jan 2021

Published online: 30 Jun 2021

Bulletin of Economic Theory and Analysis

Volume VI, Issue 1, pp. 15-26, 2021
http://www.betajournals.org

Original Article / Araştırma Makalesi
Received / Alınma: 17.11.2020 Accepted / Kabul: 19.01.2021

Use of Trimean in Theil-Sen Regression Analysis

Necati Alp ERİLLİ[a]

[a]Assoc.. Prof. Dr., Sivas Cumhuriyet University, FEAS, Department of Econometrics, Sivas, TURKEY
https://orcid.org/0000-0001-6948-0880

ABSTRACT

Theil-Sen regression analysis is the most preferred method in non-parametric regression analysis. In the Theil-Sen method, calculations are made with the median parameter. In this study, it was proposed to calculate the trimean parameter instead of the median parameter. In this way, the effects of the outliers in the data on the model are fully reflected. In applications of one real-life and two simulation data, the results obtained with the use of trimean were more successful. It is recommended to use the trimean parameter instead of the median parameter in data structures with an excess of outliers.

Keywords
Theil-Sen Regression, Trimean, Non-Parametric Regression, MAPE

JEL Classification
C53, C14

CONTACT Necati Alp ERİLLİ aerilli@cumhuriyet.edu.tr Sivas Cumhuriyet University, FEAS, Department of Econometrics, Sivas, TURKEY

1. Introduction

Statistical estimation studies refer to the use of statistics based on historical data to predict what may happen in the future. The most used estimation method is regression analysis. Regression analysis is a statistical technique in which we use the observed data to correlate a variable called a dependent variable and one or more independent variables. The aim is to create a regression model or estimation equation that can be used to define, predict, and control the dependent variable based on independent variables (Gujarati, 2002). Many assumptions are required to obtain successful estimates by regression analysis. In applications, it is not easy to provide some of these assumptions. In cases where assumptions cannot be provided, it is recommended to use flexible but less powerful non-parametric methods.

Non-parametric regression analysis can also be defined as one of the alternative estimation methods. There are a small number of non-parametric regression analysis methods in the literature. The best known and used of these is the Theil-Sen method. This method was first proposed by Theil (1950) and the procedure is firstly known as Theil's Method. After Sen (1968) highlighted the relationship to Kendall's tau it is named as the Theil–Kendall or Theil-Sen method. Theil proposed estimating the slope of a regression line as the median of the slopes of all tines joining pairs of points with different *x* values (Theil, 1950). For a pair $\left(x_i, y_i\right)$ and $\left(x_j, y_j\right)$ the appropriate slope is $S_{ij} = \frac{\left(y_j - y_i\right)}{\left(x_j - x_i\right)}$. There will be $\frac{n\left(n+1\right)}{2}$ slopes for any data. The $\hat{\beta}_1$ statistic, which is the estimator of the parameter β_1 in simple regression analysis, is calculated as the median of the slope values: $\hat{\beta}_1 = Median\left(S_{ij}\right)$. Theil suggested for the estimation of the intercept as $\hat{\beta}_0 = Median(y_i) - \hat{\beta}_1 Median(x_i)$ (Theil, 1950; Sprent, 1989). In the Theil-Sen method, alternative methods for intercept parameter computation are also introduced, although the intercept parameter is calculated as given above. Some alternative calculations have been proposed in comparison to Theil's idea of finding the intercept parameter. Let us define $d_i = y_i - \hat{\beta}_1 x_i$ calculated for all observations where $\hat{\beta}_1$ is calculated with the Theil-Sen method. Hodges-Lehmann method for $\hat{\beta}_0$ is defined as the mean value of d_i $\left(\hat{\beta}_0 = Mean\left(d_i\right)\right)$ and the optimum method for $\hat{\beta}_0$ which is defined as the median value of d_i $\left(\hat{\beta}_0 = Median\left(d_i\right)\right)$(Hodges and Lehmann, 1963). The optimum approach does not require the assumption of symmetrically distributed d_i. It is better suited especially for data with outliers. On the other hand, the Hodges-Lehmann method may not be available for data with outliers (Lehmann and Dabrera, 1975; Erilli and Alakuş, 2016).

There are many papers studied with the Theil-Sen method in the literature (Akritas et al., 1995; Fernandes and Leblanc, 2005; Lavagnini et al., 2011; Hanxiang et al., 2008; Adichie, 1967;

Wang, 2005; Dang et al., 2008; Wilcox 1998). All of these have been studied on classical Theil-Sen estimates using the median parameter.

The study consists of five sections, including introduction and conclusion parts. In the second part, the trimean parameter is briefly introduced and expressed by the formula. In the third section, the proposed regression method using the trimean parameter and the significance test of the slope parameter are introduced. The strength of the proposed method in Chapter Four is compared on Theil regression method obtained with both median and trimean parameters. MAE and MAPE methods were used in comparisons and the results were evaluated. The study was completed with a conclusion section containing general assessments.

2. Trimean Parameter

A trimean is a number that represents the general tendency of a set of numbers or data set. Like the mean, median, and mode, it is a measure of central tendency. The trimean (TM) is a measure of a probability distribution's location defined as a weighted average of the distribution's median and its two quartiles:

$$TM = \frac{Q_1 + \left(2 \times Median\right) + Q_3}{4} \tag{1}$$

After Tukey has given this formula's name with a set of techniques in his book it is sometimes called Tukey’s Trimean (Tukey, 1977). It is considered ‘resistant’ or ‘robust’ because it is not very affected by outliers.

3. Trimean with Theil-Sen Regression

With this study, it is proposed to use trimean instead of the median parameter in the calculation of both the slope parameter and the intercept parameter in the Theil-Sen regression method. The slope parameter is calculated by using trimean instead of the median of slope values calculated from dependent and independent variable binaries in the Theil-Sen method. Similarly, the intercept parameter was also found by calculating the trimean of d_i values: $\hat{\beta}_0 = Trimean\left(d_i\right)$.

3.1. Test of Significance of Slope Parameter

To test $\beta_1 = 0$, we can use the test statistics given in Equation.3.1 and 3.2:

$$|t| = \frac{|U|}{SD(U)} \tag{2}$$

where

$$U = \sum\left[rank(y_i) - \frac{n+1}{2}\right]x_i \text{ and } SD(U) = \sqrt{\frac{n(n+1)}{12}\sum(x_i - \bar{x})^2} \tag{3}$$

The approximate p-value of the test is calculated to be Prob$\left[|Z| \geq |t|\right]$, where Z is a random variable having a standard normal distribution (Birkes and Dodge, 1993:119).

4. Application

In the application part Theil-Sen regression is performed with Trimean and Median parameters separately. The proposed method is tested in 1 real-time data and 2 simulation data sets where the outliers were added by 10% to 40% to the real-time data. MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) values were examined to test the validity of the results.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_i - \hat{y}_j\right| \tag{4}$$

$$MAPE = \left(\frac{1}{n}\sum\frac{|Actual - Forecast|}{|Actual|}\right)\times 100 \tag{5}$$

MAE is more robust to outliers since it does not make use of square and MAPE is asymmetric and reports higher errors if the forecast is more than the actual and lower errors when the forecast is less than the actual.

The first data set is Blood Pressure data and given in Table.1 which has 30 samples and taken from Spath (1992: 304). Sample data consist of age (independent variable) and systolic blood pressure (dependent variable) values for 30 individuals aged 17 to 69 years.

Table 1

Data Set.1

Variables	**Data**														
Y (Blood Pressure)	144	220	138	145	162	142	170	124	158	154	162	150	140	110	128
X (Age)	39	47	45	47	65	46	67	42	67	56	64	56	59	34	42
Y (Blood Pressure)	130	135	114	116	124	136	142	120	120	160	158	144	130	125	175
X (Age)	48	45	17	20	19	36	50	39	21	44	53	63	29	25	69

The scatterplot of the variables is also given in Figure.1.

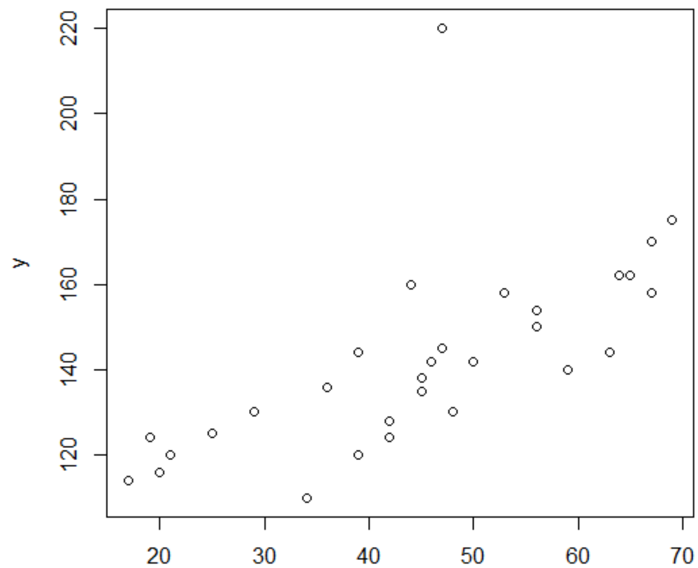

Figure 1. Scatterplot for Data.1

Parameter estimates were obtained for 9 different models using median and trimean. MAE and MAPE values were compared for 4 models where the slope parameter is estimated with median and 5 models where the slope parameter is estimated with trimean. In Table 2, the results of the estimates obtained for the original version of the data are given.

Table 2

Parameter Estimations and MAE-MAPE Results for Blood Pressure Data

	β_0 **Calculation**	**Original Data**			
		β_0	β_1	**MAE**	**MAPE**
β_1 calculation with median	Theil-Sen (Median)	64,12	1,36	50,57466667	305,2356789
	d_i (Mean)	31,652	1,36	56,7536	222,7109698
	d_i (Median)	65,8	1,36	50,53466667	309,7311069
	d_i (Trimean)	45,68375	1,36	53,94725	258,2766554
β_1 calculation with trimean	Theil-Sen (Median)	37,1528241	1,952685185	49,09053704	291,1776195
	Theil-Sen (Trimean)	13,0485243	1,952685185	52,21621489	228,9357882
	d_i (Mean)	4,90214198	1,952685185	53,84549136	208,2874961
	d_i (Median)	35,673287	1,952685185	49,02344444	287,1833844
	d_i (Trimean)	18,7998032	1,952685185	51,0659591	243,5133126

As for the results given in Table.2, we can clearly say that β_1 calculation with trimean has the best scores for both MAPE and MAE. It is found β_0 calculation with d_i median has minimum MAE and β_0 calculation with d_i mean has minimum MAPE.

Secondly, the above calculations were repeated by creating 10%, 20%, 30%, and 40% outliers for the original data. The aim is to investigate the effect of the proposed method on deviating values in the data. Figure.2 shows the scatterplot for the data with 10%, 20%, 30%, and 40% outliers.

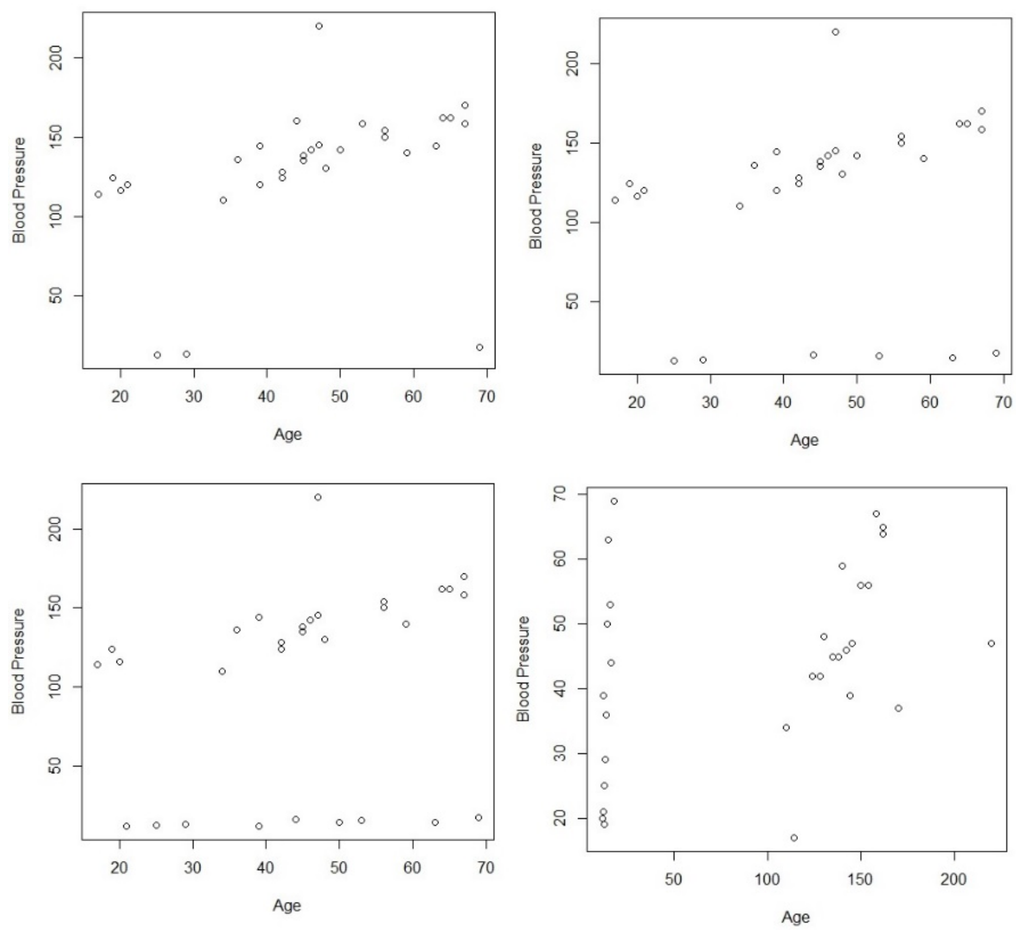

Figure 2. Scatterplot with 10%, 20%, 30%, and 40% outliers for Blood Pressure Data

In Tables 3,4,5 and 6, the results of the Blood Pressure data with outliers added are given. When we look at the results in Table.3, it is seen that the best model is the calculation with trimean according to MAPE and the calculation with median according to MAE.

Table 3

Parameter Estimations and MAE-MAPE Results for Blood Pressure Data with 10% outliers

	β_0 Calculation	**Blood Pressure Data with 10% Outlier**			
		β_0	β_1	**MAE**	**MAPE**
β_1 calculation with median	Theil-Sen (Median)	93,5	1	21,26666667	90,01196426
	d_i (Mean)	84,5	1	23,83333333	85,79972089
	d_i (Median)	95	1	21,23333333	90,98957131
	d_i (Trimean)	92,6875	1	21,34166667	89,5215978
β_1 calculation with trimean	Theil-Sen (Median)	86,1749226	1,160990712	21,61971104	89,56132911
	Theil-Sen (Trimean)	84,6669892	1,160990712	21,71344169	88,62290537
	d_i (Mean)	77,2339525	1,160990712	23,85265222	85,17595114
	d_i (Median)	86,1455108	1,160990712	21,61971104	89,54172461
	d_i (Trimean)	85,2474845	1,160990712	21,65357327	88,96768046

As for the results given in Table.4, it is clearly said that β_1 calculation with trimean has the best scores for both MAPE and MAE. It is found β_0 calculation with d_i median has minimum MAE and β_0 calculation with d_i mean has minimum MAPE just like results given in Table.2.

Table 4

Parameter Estimations and MAE-MAPE Results for Blood Pressure Data with 10% outliers

	β_0 Calculation	**Blood Pressure Data with 20% Outlier**			
		β_0	β_1	**MAE**	**MAPE**
β_1 calculation with median	Theil-Sen (Median)	89,6764706	0,941176471	33,60117647	167,1870134
	d_i (Mean)	73,294902	0,941176471	41,12815686	152,4432758
	d_i (Median)	95,2941176	0,941176471	32,70117647	173,5558044
	d_i (Trimean)	93,7463235	0,941176471	32,8604902	171,7378408
β_1 calculation with trimean	Theil-Sen (Median)	91,3715686	0,903921569	33,58503268	167,1555246
	Theil-Sen (Trimean)	89,5386029	0,903921569	33,98473856	165,1562573
	d_i (Mean)	74,9763399	0,903921569	41,09785621	152,3765839
	d_i (Median)	97,3803922	0,903921569	32,62928105	173,9750902
	d_i (Trimean)	95,7231618	0,903921569	32,80802288	172,0324633

Results in Table.5 shows that β_1 calculation with trimean has the best scores for both MAPE and MAE. It is found β_0 calculation with Theil-Sen Trimean has minimum MAPE and β_0 calculation with d_i median has minimum MAE.

Table 5

Parameter Estimations and MAE-MAPE Results for Blood Pressure Data with 30% outliers

	β_0 **Calculation**	**Blood Pressure Data with 30% Outlier**			
		β_0	β_1	**MAE**	**MAPE**
β_1 calculation with median	Theil-Sen (Median)	87	0,923076923	44,40461538	247,7264811
	d_i (Mean)	62,6517949	0,923076923	53,28235897	206,7326515
	d_i (Median)	94,8076923	0,923076923	43,82512821	262,6567125
	d_i (Trimean)	70,4350962	0,923076923	50,16903846	219,6004656
β_1 calculation with trimean	Theil-Sen (Median)	80,8444444	1,058363858	44,31442409	246,790076
	Theil-Sen (Trimean)	56,1626603	1,058363858	53,31658018	205,2601872
	d_i (Mean)	56,5458445	1,058363858	53,16330647	205,8936901
	d_i (Median)	87,2316239	1,058363858	43,71779406	258,9097462
	d_i (Trimean)	64,0612408	1,058363858	50,15714794	218,3185878

Table.6 results Show that β_1 calculation with trimean has the best scores for both MAPE and MAE. It is found β_0 calculation with d_i median has minimum MAE and β_0 calculation with d_i mean has minimum MAPE just like results given in Table.2 and Table.4.

Table 6

Parameter Estimations and MAE-MAPE Results for Blood Pressure Data with 40% outliers

	β_0 **Calculation**	Blood Pressure Data with 20% Outlier			
		β_0	β_1	MAE	MAPE
β_1 calculation with median	Theil-Sen (Median)	64,12	1,36	50,57466667	305,2356789
	d_i (Mean)	31,652	1,36	56,7536	222,7109698
	d_i (Median)	65,8	1,36	50,53466667	309,7311069
	d_i (Trimean)	45,68375	1,36	53,94725	258,2766554
β_1 calculation with trimean	Theil-Sen (Median)	37,1528241	1,952685185	49,09053704	291,1776195
	Theil-Sen (Trimean)	13,0485243	1,952685185	52,21621489	228,9357882
	d_i (Mean)	4,90214198	1,952685185	53,84549136	208,2874961
	d_i (Median)	35,673287	1,952685185	49,02344444	287,1833844
	d_i (Trimean)	18,7998032	1,952685185	51,0659591	243,5133126

Model significance was also calculated for the models which parameter estimates given above. Thus, β_1 is found significant at 0,05 and 0,01 level with values of $|t| = \frac{|U|}{SD(U)} = \frac{3213}{75,0646} = 4,4313$.

Thirdly, the proposed method was applied on two simulation data, one with 14 observations and one with 7 observations. In Figure.3 it is given scatterplot Simulation-1 and Simulation-2 Data.

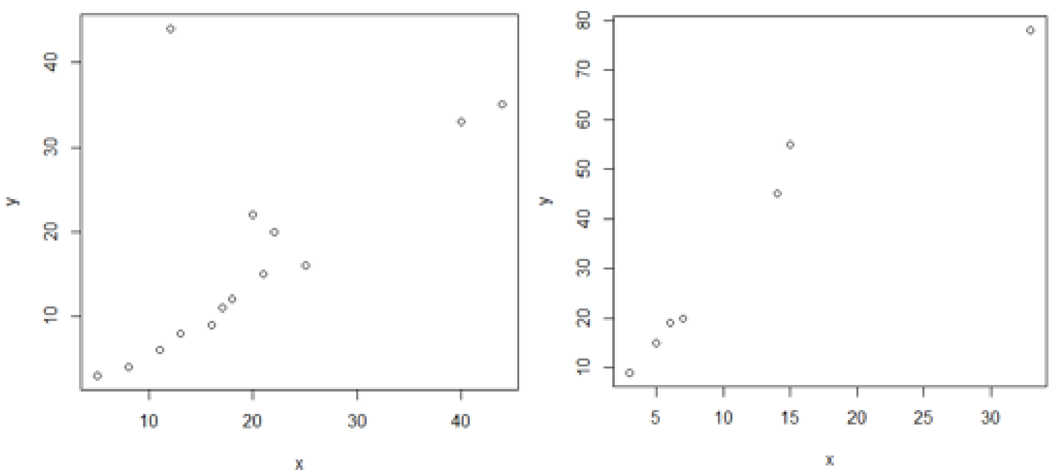

Figure 3. Scatterplot of Simulation-1 and Simulation-2 Data

Results of simulation data are given in Table.7 and Table.8 separately. For the first simulation data, the best MAE value was obtained by β_1 calculation with median, while the best MAPE value was obtained by β_1 calculation with trimean given in Table.7.

Table 7

Parameter Estimations and MAE-MAPE Results for Simulation Data-1

	β_0 Calculation	**Simulation Data-1**			
		β_0	**β_1**	**MAE**	**MAPE**
β_1 calculation with median	Theil-Sen (Median)	-1,3076923	0,846153846	4,527472527	23,18496979
	d_i (Mean)	-2	0,846153846	4,32967033	20,69790329
	d_i (Median)	-2,7692308	0,846153846	4,186813187	18,15427628
	d_i (Trimean)	-2,5192308	0,846153846	4,222527473	18,95034281
β_1 calculation with trimean	Theil-Sen (Median)	-0,8413462	0,819505495	4,556456044	24,30225605
	Theil-Sen (Trimean)	0,05103022	0,819505495	4,938903061	31,75746647
	d_i (Mean)	-1,467033	0,819505495	4,399489796	21,00146286
	d_i (Median)	-2,556044	0,819505495	4,293406593	17,86369402
	d_i (Trimean)	-2,2754808	0,819505495	4,293406593	18,48987884

In Table.8, it is seen that the calculation of β_1 with trimean results has minimum MAE and MAPE.

Table 8

Parameter Estimations and MAE-MAPE Results for Simulation Data-2

	β_0 Calculation	Simulation Data-2			
		β_0	β_1	MAE	MAPE
β_1 calculation with median	Theil-Sen (Median)	-2,75	3,25	3,107142857	7,100165916
	d_i (Mean)	-4,1071429	3,25	3,591836735	10,08581122
	d_i (Median)	-0,75	3,25	2,75	4,605029473
	d_i (Trimean)	-1	3,25	2,767857143	4,7383501
β_1 calculation with trimean	Theil-Sen (Median)	-2,0208333	3,145833333	2,973214286	6,607061311
	Theil-Sen (Trimean)	0,01041667	3,145833333	2,707589286	5,004651325
	d_i (Mean)	-2,8720238	3,145833333	3,277210884	8,479637096
	d_i (Median)	-0,4375	3,145833333	2,675595238	4,532538611
	d_i (Trimean)	-0,4270833	3,145833333	2,676339286	4,543517977

When we look at the model significances; β_1 is found significant at 0,05 and 0,01 level with values of $|t| = \frac{|U|}{SD(U)} = \frac{563}{166.988} = 3.3713$ in Simulation Data-1 and β_1 is found significant at 0,05 level with values of $|t| = \frac{|U|}{SD(U)} = \frac{118}{54.85} = 2.151$ in Simulation Data-2.

5. Conclusion

In this study, it was proposed to use the trimean parameter instead of the median parameter in Theil-Sen regression analysis. Thus, the contribution of the effect of the outliers to the model was tried to be investigated. In the proposed method, trimean was used separately for both the slope parameter and the intercept parameter. As a result of applications on one real-time data and two simulation data, model comparisons were made according to MAE and MAPE criteria. Besides, the efficiency of the method was tested by adding 10%, 20%, 30%, and 40% outliers to the real-time data. The results of the analysis showed that the calculations with trimean were more successful than those with the median. The best model estimation methods can be said to be β_1 calculation with trimean and β_0 calculation with di (Mean) and di (Median).

The most common method of non-parametric regression analysis is perhaps the Theil-Sen method. In this study, the estimation results obtained by the proposed trimean parameter instead of the median parameter were successful. Finally, the use of the Trimean mean in other non-parametric statistical methods is also proposed to be investigated.

References

Adichie, J. N. (1967). Estimates of regression parameters based on rank tests. *Annals of Mathematical Statistics,* 38, 894-904.

Akritas, M. G., Murphy, S. A., & LaValley, M. P. (1995). The Theil–Sen estimator with doubly censored data and applications to astronomy. *J. Amer. Statist. Assoc*., 90, 170–177.

Birkes, D. & Dodge, Y. (1993). *Alternative Methods of Regression*. John Wiley & Sons Inc., NY, USA.

Dang, X., Peng, H., Wang, X. & Zhang, H. (2008). *Theil-Sen Estimators in a Multiple Linear Regression Model*, Olemiss Edu.

Erilli, N. A. & Alakuş, K. (2016). Parameter Estimation In Theil-Sen regression analysis with Jackknife method. *Eurasian Econometrics, Statistics & Empirical Economics Journal*, 5, 28-41.

Fernandes, R. & Leblanc S. G. (2005). Parametric (modified least squares) and non-parametric (Theil–Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. *Remote Sensing of Environment*, 95, 303–316.

Gujarati, D. N. (2002). *Basic Econometrics*. McGraw Hill pub., NY, USA.

Hanxiang, P., Shaoli W. & Xueqin, W. (2008). Consistency and asymptotic distribution of the Theil–Sen estimator. *Journal of Statistical Planning and Inference*, 138, 1836–1850.

Hodges, J. L. & Lehmann, E. L., (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* 34, 598–611.

Lavagnini, I., Badocco, D., Pastore, P. & Magno, F. (2011). Theil–Sen nonparametric regression technique on univariate calibration, inverse regression and detection limits. *Talanta*, 87, p.180-188.

Lehmann, E. L., & Dabrera H. J. M. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. SF, USA: Holden-Day Inc.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association,* 63, 1379–1389.

Spath, H. (1992). *Mathematical Algorithms for Linear Regression*. London: Academic Press.

Sprent, P. (1989). *Applied Nonparametric Statistical Methods*. Chapman and Hall Pub., London, UK.

Theil, H. (1950). A-Rank invariant method of linear and polynomial regression analysis. *III. Nederl. Akad. Wetensch.Proc., Series A*, 53, 1397-1412.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, M A: Addison-Wesley, 46-47.

Wang, X. Q. (2005). Asymptotics of the Theil–Sen estimator in simple linear regression models with a random covariate. *Nonparametric Statist.* 17, 107–120.

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