Available online at www.alphanumericjournal.com

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

Volume 9, Issue 2, 2021

Received: September 21, 2021
Accepted: November 22, 2021
Published Online: December 31, 2021

AJ ID: 2021.09.02.STAT.03
DOI: 10.17093/alphanumeric.998384
Research Article

Contributions to Theil-Sen Regression Analysis Parameter Estimation with Weighted Median

Cem Öztaş

Department of Econometrics, Faculty of Economics and Administrative Sciences, Sivas Cumhuriyet University, Sivas, Turkey, cemoztas5800@gmail.com

Necati Alp Erilli, Ph. D.

Assoc. Prof., Department of Econometrics, Faculty of Economics and Administrative Sciences, Sivas Cumhuriyet University, Sivas, Turkey, aerilli@cumhuriyet.edu.tr

* Sivas Cumhuriyet Üniversitesi Merkez Kampüsü, İİBF, Ekonometri Bölümü 58140 Sivas /Türkiye

ABSTRACT

Regression analysis is one of the most commonly used estimation methods. In statistical studies, some assumptions must be fully met to make good estimations with regression analysis. Some of these assumptions are not always fulfilled in real life data. For such cases, alternative methods are used. One of them is Theil-sen method, which is one of the non-parametric regression analysis techniques. In this study, different analysis techniques were proposed by using the weighted median parameter instead of the median parameter used in the Theil-Sen regression method. With the proposed four different algorithms, new approaches to Theil-Sen regression analysis estimation have been introduced. It has been seen that the obtained results are successful compared to the classical Theil-Sen results.

Keywords: Theil-Sen Method, Weighted Median, Non-Parametric Regression, Mean Absolute Error

This study is derived from the thesis "Contributions to Theil-Sen Regression Analysis Parameter Estimation by Weighted Median" written by Cem Öztaş and under the supervision of N. Alp Erilli.

1. Introduction

Regression analysis is one of the statistical techniques that can be applied in almost any field today. In addition to being an easily understandable and applicable method, it finds wide usage and application areas with the help of many statistical package programs. In general, regression analysis can be defined as a method that studies the relationship between a dependent and one or more independent variables, form the effect of independent variables on dependent variables. Regression analysis is divided into three groups: parametric, non-parametric and semi-parametric according to the field of use. Many mathematical and statistical assumptions are required to apply parametric regression models. These assumptions are the most powerful methods in prediction studies when they are provided. However in real-life applications, these assumptions are not always fully met. In cases where these assumptions are not met, like the shape of the functional relationship between the dependent and independent variables is not known, non-parametric or semi-parametric regression models are used.

Problems of not being able to provide assumptions, especially in data with few observations, cause problems for the researcher and may cause the study to be blocked or stopped in a certain place. Non-parametric regression techniques, which ignore such situations and are more flexible, are relatively preferred especially in data with assumption problems and data with few observations. In this study, new approaches have been proposed in Theil-Sen regression analysis method, which is one of the non-parametric regression analysis techniques, by using the weighted median parameter instead of the median parameter. In these approaches, Theil-Sen regression analysis with weighted median was applied to all possible sub-datasets ($n>2$) that could be created from the data set, and dependent and independent variable parameter estimations were obtained with different algorithms using all the sets that will emerge. Thus, it has been tried to contribute to the actual parameter estimation of all parameter estimates (including outlier observations) to be obtained from the data set.

In the literature, there are studies that contribute to Theil-Sen regression analysis with different algorithms. Some of these can be summarized as follows: Fernandes and Leblanc (2005) made a regression estimation under measurement errors with Theil-Sen regression. The Theil-Sen approach is proposed as a potential alternative to OLS for linear regression in remote sensing applications. Lavagnini et al. (2011) made an inverse regression estimation with the help of Theil-Sen regression. Also in article it is reported the combined use of the nonparametric Theil-Sen regression technique and of the statistics of Lancaster-Quade concerning the linear regression parameters to solve typical analytical problems, like method comparison, calculation of the uncertainty in the inverse regression, determination of the detection limit. Zhou and Serfling (2008) performed multivariate spatial estimations with the Theil-Sen estimator. Writers suggested a new statistics based on spatial U-quantiles are presented for nonparametric estimation of multiple regression coefficients, extending the classical Theil-Sen nonparametric simple linear regression slope estimator, and for robust estimation of multivariate dispersion. Shen (2009), used Theil-Sen regression and asymptotic multiple linear regression estimation studies. It is also a simple and robust (point as well as interval) estimator of β based on Kendall's

rank correlation τ is studied. The point estimator is the median of the set of slopes joining pairs of points with ti ≠ ti , and is unbiased. Erilli and Alakuş (2016) suggested the use of Jackknife in Theil-Sen estimations, and Alakuş and Erilli (2014) suggested the use of adjusted mean for non-parametric regression calculations of data with equal independent values. In both studies, some alternative suggestions were made for the classical Theil-Sen regression. In Akritas et al. (1995) a second extension of the Theil-Sen estimator, based on a direct estimation of the median of pairwise slopes is given. In this article they proposed an estimator obtained by inverting a suitable version of Kendall's τ . Authors derived the asymptotic normality of this estimator and obtained a class of simple estimates of its asymptotic variance. Wilcox (1998) examined some comments on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. In these studies in the literature, the median parameter was used. In this study, the weighted median parameter has been used instead of the median parameter to contribute to the calculations of the Theil-Sen method.

2. Material and Method

Regression analysis is in our lives and shows itself in many areas. It is an analysis that shows researchers the strength, direction and degree of interaction between variables. It is an effective and successful analysis method in prediction and forecasting studies. The aim is to create a regression model or prediction equation that can be used to define, predict and control the dependent variable based on the independent variables (Bowerman et al., 2015).

Regression analysis is examined under different headings according to different situations: According to the number of dependent and independent variables (univariate - multivariate), according to the mathematical form (linear - non-linear), according to the case of independent shadow variables (analysis of variance - analysis of covariance models) and according to the condition of providing the assumptions (parametric - non-parametric). These headings are also divided into different subheadings. (Gujarati, 1999). The model or models that will be used in the studies are determined by experts in the field on the condition that they follow certain rules.

One of the most important assumptions of regression analysis is that the shape of the relationship between dependent and independent variables is known. Estimates made in cases where these assumptions are not met will lose their ability to be a good estimate. In order to make better predictions in such cases, regression methods are needed that help to stretch the linearity assumption in parametric regression. These methods are regression models known as semi-parametric and non-parametric regression methods (Erilli 2015).

Parametric regression is the expression of dependent and independent variables and the average relationship between these variables with a mathematical function. In order for the parametric regression analysis to be applied successfully, it is necessary to provide assumptions such as normal distribution, homoscedasticity and autocorrelation. These methods are the strongest regression methods if the assumptions are true.

If some assumptions that apply to parametric regression methods cannot be met, the methods used can be defined as non-parametric methods. They are effective methods for data with a very low sample size or a large number of outlier observations. In statistical studies, there are powerful (robust) parametric methods that deal with the effects of outlier observations in different ways. However, even these powerful methods may not produce suitable solutions because the parameters are corrupted due to outlier observations and the real structure of the data may not be reflected in the model. In this case, non-parametric regression provides preliminary information (Hardle, 1994).

Although the nonparametric regression method does not have restrictive assumptions when making predictions, it has some drawbacks. It is difficult to make an estimate when there are too many independent variables, and the resulting graphs have a complex structure. In addition, it is difficult to take discrete variables into account using the non-parametric method and interpret the individual effects of the dependent variable depending on the increase in the number of independent variables. The disadvantages of nonparametric methods can be eliminated using the semiparametric regression models (Horowitz, 1993).

The semi-parametric regression model uses both the parametric and non-parametric regression models together. For this reason, the semi-parametric regression model is not affected by the restrictive assumptions of parametric models dec but combines the attractive features of non-parametric methods (such as Cox, Kernel Regression). Semi-parametric regression models are used in cases where non-parametric regression methods cannot make good predictions, or when the distribution of errors is unknown, although the researcher wants to use parametric methods. Normality assumption is not required when estimating parameters with these models (Takezawa, 2006).

2.1. Theil-Sen Method

This method, suggested by Theil (1950), is one of the most commonly used slope finding methods by researchers. With the correction of Sen (1968), it is called Theil-Sen method in many sources.

In this method, the slopes of all pairs of the sample units are used to find the estimator of the β_1 parameter. The slopes of pairs are calculated by the formula $S_{ij} = \frac{y_j - y_i}{x_j - x_i}$. Accordingly, the estimator of the parameter β_1 is formulated as $\hat{\beta}_1 = Med\left\{S_{ij}\right\}$. The constant term of the model is found by substituting the $\left(x_i, y_i\right)$ and $\left(x_j, y_j\right)$ points of the i-th and j-th variables that give the $Med\left\{S_{ij}\right\}$ value after the slope of the regression line is found, that is, by solving either of the $\hat{\beta}_0 = y_i - \hat{\beta}_1 x_i$ or $\hat{\beta}_0 = y_j - \hat{\beta}_1 x_j$ equations.

In the Theil-Sen regression analysis, the test statistics given in Equation 1 and Equation 2 are used together to test hypothesis $H_0 : \hat{\beta}_1 = 0$ (Birkes and Dodge, 1993).

$$|t| = \frac{|U|}{SD(U)} \quad (1)$$

In the above equation, the values of U and $SD(U)$ are calculated as follows:

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

The approximate p-value of the test is calculated as $\left[|Z| \geq |t| \right]$, where *z* is a random variable with a standard normal distribution (Hussain and Sprent, 1983).

2.2. Weighted Median

The weighted median is a method of averaging calculated by adding weight from the measures of the central trend to the median parameter. The median of a list of numbers is obtained by sorting the numbers from small to large and selecting the number in the center of the ordered list. The weighted median of the numbers w_i and x_i with weights is calculated as follows:

First, numbers are sorted from small to large. By changing the indices, an interval is created so that it is $x_1 \leq x_2 \leq \mathrm{L} \leq x_{k-1} < 0.5$. $w_i = \frac{|x_i - x_j|}{\sum_{i,j=1}^{n} |x_i - x_j|}$ weights should not be negative and their sum should be 1. Then an index *k* is found which satisfies the following equation:

$$\begin{aligned} w_1 + w_2 + \mathrm{L} + w_{k-1} &< 0.5 \\ w_1 + w_2 + \mathrm{L} + w_{k-1} + w_k &> 0.5 \end{aligned} \quad (3)$$

In this case, x_k is the weighted median. Sometimes it happens that there is an index in the form of a $w_1 + w_2 + \ldots + w_{k-1} = 0.5$. In this case, $(x_{k-1} + x_k)/2$ is the weighted median (Birkes and Dodge, 1993).

2.3 Proposed Algorithm

In this study, weighted median parameter was used instead of median parameter in Theil-Sen regression analysis and calculation suggestions were made using all possible subsamples for parameter estimation. First, new sample groups $(n > 2)$ were created by drawing all possible sub-samples from the sample data used in the study. The number of all sub-samples used without repeating the one observation used will be as much as $N^* = {}_nC_3 + {}_nC_4 + {}_nC_5 + \mathrm{L} + {}_nC_{n-1} + {}_nC_n$. Then, Theil-Sen regression is applied to all calculated sub-sample groups (n=3, n=4, ...) with individually weighted median and N^* unit parameter estimates are obtained. The four different parameter estimates proposed in the study will be as follows:

(1) The mean and median values of the N^* unit parameter estimates are calculated separately for the new parameter estimate values.

(2) The mean and median values of the ${}_nC_3$, ${}_nC_4$, ${}_nC_5$, ... sample groups that make up the N^* parameter estimation are calculated separately (n=3, n=4, ...) and the new parameter estimation values are calculated.

(3) Certain proportions of N^* unit parameter estimates obtained in the first step (5%, 10%, etc.) mean and median values and new parameter estimation values are calculated from the new data set to be created by truncating. Truncating is performed by applying the cropping ratio to be applied to a data set sorted from small to large in equal amounts to the lowest and highest order (For example 5% cropping is observed from the bottom and up to 2.5% from the top for 2.5%).

(4) From the N^* unit parameter estimates obtained in the first step, certain proportions of samples (resampling) are taken (5%, 10%, etc.) with the mean and median values and the new parameter estimation values are calculated from the new data set created.

3. Numerical Example

In practice, for a 10-observation data set, the above-mentioned four different estimation models were obtained by using the Weighted Median parameter instead of the median parameter in the Theil-Sen regression method, and the obtained estimation results were compared with the results of the classical Theil-Sen method. 968 sub-datasets $\left({}_{10}C_3 + {}_{10}C_4 + \mathrm{L} + {}_{10}C_{10} = 968\right)$ were obtained from a 10-observation dataset. The parameter estimation values obtained using the four different estimation methods briefly introduced above were compared using the Mean Absolute Error (MAE) model selection criterion and the best model estimate was determined.

$$MAE = \sum_{i=1}^{n} \left| \frac{y_i - x_i}{n} \right| \tag{4}$$

In all calculations, MATLAB.2009b. and Microsoft Excel package programs were used. The confidence level was determined as 0.05 in the coefficient significance tests. The data set created from a dependent and an independent variable used in the study is as given in Table 1.

Y	X
15	26
16	30
20	32
21	35
26	29
30	36
29	37
32	34
35	39
33	8

Table 1. Sample Data Set

As shown in Table 1, this value was considered an outlier because the last observation value of variable *X* was too small than the other observation values, and it affected the arithmetic mean and other parameters more than it should have been.

Table 2 shows the estimated values of Theil-Sen model coefficients obtained using Weighted Median and Median parameters and the results of the model selection criteria.

	β_0	β_1	**MAE**
Theil-Sen Regression Result with Median	-6,66667	0,833333	7,633333
Theil-Sen Regression Result with W. Median	-12,068137	1,1727782	7,1722652

Table 2. Theil-Sen Model Coefficient Estimate Values

It can be seen that the estimation result made with the weighted median according to the results given in Table 2 gives a better MAE value. In Table 3, the mean and median values of the weighted median parameters calculated separately (n=3, n=4, ..., n=10) for each sub-sample group of 968 sample data and the new parameter estimation values are given.

	β_0	β_1	**MAE**
Mean of all sub-samples groups with 3 observations	-8,07481	1,021203	6,929684
Median of all sub-samples groups with 3 observations	-14,4091	1,272727	7,609091
Mean of all sub-samples groups with 4 observations	-12,6487	1,185962	7,169536
Median of all sub-samples groups with 4 observations	-13,1932	1,209615	7,264056
Mean of all sub-samples groups with 5 observations	-6,69171	1,012488	6,917484
Median of all sub-samples groups with 5 observations	-5,25	0,9	6,95
Mean of all sub-samples groups with 6 observations	-15,5147	1,282897	7,429977
Median of all sub-samples groups with 6 observations	-13,8	1,272727	7,852727
Mean of all sub-samples groups with 7 observations	-18,4515	1,381773	7,780469
Median of all sub-samples groups with 7 observations	-14,9	1,339161	8,502238
Mean of all sub-samples groups with 8 observations	-15,5864	1,292248	7,496055
Median of all sub-samples groups with 8 observations	-12,1	1,272727	8,532727
Mean of all sub-samples groups with 9 observations	-16,9287	1,346678	7,814056
Median of all sub-samples groups with 9 observations	-16,8636	1,386364	8,490909
All Sub-Sample Groups with 10 Observations	-27,5	1,666667	8,833333

Table 3. Results of all sub-samples groups with *n* observations

According to the results in Table 3, it can be seen that the mean and median values of the parameter estimation values obtained from all subgroups with 5 observations give better MAE values than other results. In Table 4, the new parameter estimation values obtained from the new dataset created by truncated the 968 parameter estimates obtained in the first step at certain rates are given. Here, the observation numbers and parameter estimation values of the datasets obtained with 5% to 50% truncated rates are given.

Truncated Ratio	**Sample Size**	β_0	β_1	**MAE**
5%	920	-12,686628	1,1942762	7,2334499
10%	872	-12,878747	1,20203	7,2617086
15%	823	-13,033762	1,2070486	7,2738654
20%	775	-12,9716	1,2062346	7,2792975
25%	726	-13,006874	1,2065626	7,2750637
30%	678	-12,901145	1,2042273	7,2761254
35%	630	-12,872639	1,2034609	7,2752361
40%	581	-12,937173	1,2033327	7,2612262
45%	533	-12,665831	1,2009033	7,2946021
50%	484	-12,579344	1,2018598	7,3201253

Table 4. Prediction values of data truncated at certain ratios

When the results given in Table 4 are examined, it is seen that 5% truncated gives the best result. Although it seems that the second best result is 40% truncated, it would not be wrong to say that the results come out a little worse as the truncated rate increases.

In Table 5, 50 and 100 random samples were taken from the 968 parameter estimates obtained in the first step at the rates of 5%, 10% and 20%, and new parameter estimate values obtained with the mean and median parameter were given.

	β_0	β_1	**MAE**
Mean of 50 Samples taken at 5%	-11,6775	1,16449	7,179108
Median of 50 Samples taken at 5%	-11,8818	1,2	7,443636
Mean of 50 Samples taken at 10%	-11,7595	1,163772	7,15654
Median of 50 Samples taken at 10%	-11,6636	1,2	7,514545
Mean of 50 Samples taken at 20%	-12,4635	1,183398	7,184527
Median of 50 Samples taken at 20%	-11,9	1,2	7,44
Mean of 100 Samples taken at 5%	-12,0799	1,173808	7,178768
Median of 100 Samples taken at 5%	-11,3068	1,2	7,657273
Mean of 100 Samples taken at 10%	-11,7512	1,165645	7,174304
Median of 100 Samples taken at 10%	-11,6416	1,186084	7,372002
Mean of 100 Samples taken at 20%	-12,0487	1,171167	7,162285
Median of 100 Samples taken at 20%	-11,8662	1,208567	7,574012

Table 5. 50 and 100 samples results for 5,10 and 20% rates

According to the results in Table 5, it can be seen that the best model estimate is obtained from the sub-samples taken at a rate of 10%. Looking at all the results, it seems that the best model estimate is the model obtained with the mean of the parameter estimates calculated with the weighted median of all sub-sample groups with 5 observations:

$$\hat{Y} = -6,69171 + 1,012488X$$

4. Conclusion

Since nonparametric methods do not require many assumptions, they are often preferred in the study of statistics and econometrics. Nonparametric regression analysis is also one of the preferred methods in forecasting studies. One of the most preferred among these methods is the the Theil-Sen method. In the Theil-Sen method, parameter estimation is performed with the median value of the slope values between all observation pairs. In calculations with the median parameter, the effect of outliers in the data can not be included too much in the model. In this study, different parameter calculations were proposed using the weighted median parameter instead of the median to address this problem.

The aim of the four different proposed calculation methods is to contribute to Theil-Sen regression analysis, to provide the researcher with alternative calculation methods and to provide various trial opportunities to obtain the best model estimation. According to the results obtained from this study, the results of some of the proposed methods for weighted median parameter in Theil-Sen regression analysis were found to be more successful than the results of classical Theil-Sen analysis estimation. The testing of different calculation methods is important for the development of such estimation methods. In order to achieve the best modeling result, working on new algorithms will pave the way for future studies.

References

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

Alakuş, K., and Erilli, N.A. (2014). Non-Parametric Regression Estimation for Data with Equal Value, European Scientific Journal (ESJ) ,2014, 4, 1857- 7431.

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

Bowerman, B.L., O'Connell, R.T., Murphree, E.S. and Orris, J.B. (2015). Essentials of Business Statistics, 5th edition. McGraw and Hill pub. USA.

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

Erilli, N.A. (2015). İstatistik-2. Seçkin Pub., Ankara.

Fernandes, R. and 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), 3, 303-316.

Gujarati, D. (1999). Temel Ekonometri. Translate: Ümit Şenesen, G. Günlük Şenesen. Literatür Pub., İstanbul.

Hardle, W. (1994). Applied Nonparametric Regression. Cambridge University, UK.

Horowitz, J.L. (1993). Semiparametric Estimation of a Work-Trip Mode Choice Model, Journal of Econometrics, 58, 49-70.

Hussain, S.S. and Sprent, P. (1983). Non-Parametric Regression. Journal of The Royal Statistical Society. Ser., A., 146, 182-191.

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

Sen, P.K. (1968). Estimates of The Regression Coefficient Based on Kendall's Tau. J. Amer. Statist. Ass., 63, 1379-1389.

Shen, G. (2009). Asymptotics of a Theil-Sen-type estimate in multiple linear regression

Statistics & Probability Letters, volume 79, Issue 8, pp. 1053-1064.

Takezawa, K. (2006). Introduction to Nonparametric Regression. Wiley-Interscience, Canada.

Theil, H. (1950). A Rank Invariant Method of Linear and Polynomial Regression Analysis. III. Nederl. Akad. Wetensch. Proc., Series A, 53, 1397-1412.

Zhou, W. and Serfling, R. (2008). Multivariate spatial U-quantiles: A Bahadur–Kiefer representation, a Theil- Sen estimator for multiple regression, and a robust dispersion estimator. Journal of Statistical Planning and Inference, 138:6, Pages 1660-1678.

Wilcox, R. (1998). A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. Biometrical J. 40, 261–268.