THEIL-SEN REGRESSION ESTIMATORS WITH TRIMEAN FAMILY

Necati Alp Erilli

Abstract

Theil-Sen regression can be defined as a robust method that is frequently used in non-parametric regression analysis. In this method, In this method, parameter estimations are made with the median parameter. Trimean is a measure of central tendency that can be used in data especially when there are agglomeration or outliers and it is a variation of mean based on quartiles. In this article it is proposed similar means with the help of other quantile values and they are defined as trimean family means: Tetramean, pentamean, hexamean, heptamean, octamean, nonamean and decamean. With these newly defined means, parameter estimations were made in Theil-Sen regression and the results were compared with the classical Theil-Sen estimation results. In the application, datasets derived from 7 different distributions and 3 different datasets studied in the literature were used. According to the results obtained, trimean family results in real life data and median parameter results in simulation data were more successful. According to the results of the analysis, if there are no significant outliers in the data, it can be said that the results of the trimean family parameters are effective.

Key words: Theil-Sen regression, trimean, nonparametric regression

JEL Code: C13, C14

Introduction

Regression analysis is one of the statistical techniques frequently used in forecasting studies. Regression analysis is a subject that examines the dependence of one explanatory variable on other explanatory variables with the aim of estimating the mean value of the former in terms of the constant values of the others (Gujarati, 1999). The aim here is to reveal and interpret the relationship between a dependent variable with one or more independent variables. This relationship does not have to be dependent on a function, nor may it have a cause-effect relationship. Parametric regression is the expression of the explained and explanatory variables and the mean relationship between these variables with a mathematical function. In order for the parametric regression analysis to be successfully applied, assumptions such as

normal distribution, equal variance, and autocorrelation must be provided. They are the most powerful regression methods in case the assumptions are realized (Erilli & Alakuş, 2016). These assumptions are often difficult to provide in real-life data. The methods used in cases where some assumptions valid for parametric regression methods cannot be met are generally called non-parametric methods. These methods emerge as effective methods when the sample size is very low or when there are outliers in the data (Hardle, 1994).

Nonparametric regression techniques rely more on data than parametric techniques to obtain information about the regression function. Therefore, it is suitable for inference problems. Non-parametric estimators are more appropriate to use when a suitable parametric form for the regression function cannot be obtained. Because when the parametric model is valid, the efficiency of non-parametric models will be less. In addition, non-parametric models can also be used to test the validity of the parametric model (Eubank, 1988). Nonparametric methods use the median parameter instead of the arithmetic mean, which is susceptible to outliers. In this study, it is suggested to use new means defined as trimean family instead of median parameter in Theil-Sen method, which is one of the most frequently used non-parametric regression analysis methods in the literature. The suggested means were applied to real life data with simulation data obtained from different distributions and the results were investigated.

1 Theil-Sen regression analysis

Theil-Sen regression can be defined as a robust method that is frequently used in non-parametric regression analysis. This method proposed in Theil (1950) study and was named Theil-Sen method after the corrections in Sen (1968). In this method, in order to find the $\hat{\beta}_1$ statistic, when the sample units are considered in pairs, the slopes of all cases $\left[S_{ij} = \frac{y_i - y_j}{x_i - x_j}, j > i \right]$ are calculated and the median of these slope values is defined as $\hat{\beta}_1$ $\left[\hat{\beta}_1 = Median(S_{ij}) \right]$. The constant parameter of the model is found with the $\hat{\beta}_0 = Median(Y_i) - \hat{\beta}_1 Median(X_i)$ equation.

Although there are many studies with Theil-Sen method in the literature, the median parameter was used in almost all of these studies. In recent years, studies using alternative measures of central tendency instead of median have been increasing. It has been observed

that successful results have been obtained in recent studies with weighted median, trimean, and weighted trimean parameters (Erilli, 2021; Erilli, 2022; Öztaş and Erilli, 2021).

1.1 Trimean and Trimean family

The trimean parameter is a measure of central tendency that can be used in data, especially when there are agglomerations or outliers. It is a mean variety based on Trimean quartiles, first introduced in Tukey (1977) and is calculated as given in Equation 1:

$$TriMean = \frac{Q_1 + 2xQ_2 + Q_3}{4} \qquad (1)$$

The quantiles are the values that divide a series ordered from smallest to largest into equal parts. Similar to the trimean parameter, similar means can be defined with the help of other quantile values. In this study, new mean definitions were made with the help of quantiles dividing a series into 6, 8, 10, 12, 14, 16 and 18 equal parts. Based on the trimean parameter, these obtained mean parameters were named Tetramean, pentamean, hexamean, heptamean, octamean, nonamean and decamean.

The formulas of these proposed means are given in Equation 2-8:

$$TetraMean = \frac{Q_1 + 2Q_2 + 3Q_3 + 2Q_4 + Q_5}{9} \qquad (2)$$

$$PentaMean = \frac{Q_1 + 2Q_2 + 3Q_3 + 4Q_4 + 3Q_5 + 2Q_6 + Q_7}{16} \qquad (3)$$

$$HexaMean = \frac{Q_1 + 2Q_2 + 3Q_3 + 4Q_4 + 5Q_5 + 4Q_6 + 3Q_7 + 2Q_8 + Q_9}{25} \qquad (4)$$

$$HeptaMean = \frac{Q_1 + 2Q_2 + 3Q_3 + 4Q_4 + 5Q_5 + 6Q_6 + 5Q_7 + 4Q_8 + 3Q_9 + 2Q_{10} + Q_{11}}{36} \qquad (5)$$

$$OctaMean = \frac{Q_1 + 2Q_2 + 3Q_3 + 4Q_4 + 5Q_5 + 6Q_6 + 7Q_7 + 6Q_8 + 5Q_9 + 4Q_{10} + 3Q_{11} + 2Q_{12} + Q_{13}}{49} \qquad (6)$$

$$NonaMean = \frac{Q_1 + 2Q_2 + 3Q_3 + 4Q_4 + 5Q_5 + 6Q_6 + 7Q_7 + 8Q_8 + 7Q_9 + 6Q_{10} + 5Q_{11} + 4Q_{12} + 3Q_{13} + 2Q_{14} + Q_{15}}{64} \qquad (7)$$

$$DecaMean = \frac{Q_1 + 2Q_2 + 3Q_3 + 4Q_4 + 5Q_5 + 6Q_6 + 7Q_7 + 8Q_8 + 9Q_9 + 8Q_{10} + 7Q_{11} + 6Q_{12} + 5Q_{13} + 4Q_{14} + 3Q_{15} + 2Q_{16} + Q_{17}}{81} \qquad (8)$$

1.2 Test of significance of slope parameter

To test $H_0 : \hat{\beta}_1 = 0$, we can use the test statistics given in Equation.9 and 10 (Birkes and Dodge, 1993).

$$|t| = \frac{|U|}{SD(U)} \qquad (9)$$

where

$$U = \sum \left[rank(y_i) - \frac{n+1}{2} \right] x_i \text{ and } SD(U) = \sqrt{\frac{n(n+1)}{12} \sum (x_i - \bar{x})^2} \qquad (10)$$

The approximate *p*-value of the test is calculated to be $\text{Prob}\left[|Z| \geq |t|\right]$, where Z is a random variable having a standard normal distribution.

2 Application

In application part, the means of the trimean family introduced above were used in the Theil-Sen regression analysis instead of the median parameter, and the obtained estimation results were compared with the classical Theil-Sen results. Analyzes were tested on 2 different data structures. In the first, simulation data produced from 7 different distributions were used, and in the second, some real-life data used in the literature were used. The MAPE criterion was used in model comparisons. The MAPE formula is as given in Equation 11.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|A_i - F_i|}{A_i} \qquad (11)$$

In Equation 11, A_i represents actual value, F_i represents forecast value of the model. The mean absolute percentage error is one of the most widely used measures of forecast accuracy, due to its advantages of scale-independency and interpretability (Kim and Kim, 2016). All analyzes were made in R package programme and Microsoft Excel with codes written by the author.

In the first part of the application, the data from the determined 7 distributions were produced separately for both dependent and independent variables and Theil-Sen estimates were obtained with the proposed means. STATSGRAPHIC.16 (Trial version) package program was used to generate data. Data sets obtained by Monte-Carlo simulation technique from normal, exponential, log-normal, erlang, beta, weibull and gamma distributions with different parameter ratios were created and the performances of the proposed means were evaluated. Each dataset has 50 observations and 10 different datasets were created from each distribution. In Table 1, the parameter information of the produced data, the classical and suggested averages, and the best model parameters versus the MAPE values obtained from Theil-Sen estimates are given. The best model parameters are the parameters in which the best

results are obtained from the results obtained from 10 different data sets (11th or 12th model predictions are looked at if there is an equal best result).

Tab. 1: Data analysis results generated from 7 distributions

Method	General (Including Median)	Only Trimean Family
Normal (Y:Mean 0, Deviation 1 ; X: Mean 0, Deviation 2)	Median	Hexa
Normal (Y:Mean 0, Deviation 1 ; X: Mean 1, Deviation 2)	Tetra	Tetra
Normal (Y:Mean 0, Deviation 1 ; X: Mean 2, Deviation 2)	Median	Penta
Exponential (Y:Scale:0,1 threshold:1,0 ; X:Scale:0,1 threshold:1,0)	Hepta	Hepta
Exponential (Y:Scale:0,1 threshold:1,0 ; X:Scale:0,2 threshold:2,0)	Median	Tetra
Exponential (Y:Scale:0,1 threshold:1,0 ; X:Scale:0,3 threshold:3,0)	Median	Tri
Log-Normal (Y:Mean 10, Deviation 1; X:Mean 10, Deviation 1)	Median	Tetra
Log-Normal (Y:Mean 10, Deviation 1; X:Mean 10, Deviation 5)	Median	Hepta
Log-Normal (Y:Mean 10, Deviation 1; X:Mean 10, Deviation 10)	Median	Hexa
Erlang (Y:Shape:1 Scale:1; X:Shape:1 Scale:1)	Median	Nona
Erlang (Y:Shape:2 Scale:2; X:Shape:2 Scale:3)	Nona	Nona
Erlang (Y:Shape:3 Scale:3; X:Shape:3 Scale:6)	Median	Tri
Beta (Y:Shape:1 Shape:1, X:Sahpe:1 shape:1)	Hexa	Hexa
Beta (Y:Shape:2 Shape:2, X:Sahpe:2 shape:4)	Hepta	Hepta
Beta (Y:Shape:3 Shape:3, X:Sahpe:3 shape:6)	Median	Tri
Weibull (Y:Shape:2 Scale:1; X:Shape:2 Scale:1)	Median	Deca
Weibull (Y:Shape:2 Scale:2; X:Shape:2 Scale:4)	Hexa	Hexa
Weibull (Y:Shape:3 Scale:3; X:Shape:3 Scale:6)	Median	Hepta
Gamma (Y:Shape:1 Scale:1; X:Shape:1 Scale:1)	Hexa	Hexa
Gamma (Y:Shape:2 Scale:2; X:Shape:2 Scale:4)	Median	Deca
Gamma (Y:Shape:3 Scale:3; X:Shape:3 Scale:6)	Nona	Nona

When the results in Table 1 are examined, it is seen that the median parameter is very successful in the estimation results made with all mean types, including the median parameter. The superiority of the median parameter is not seen in the estimation results only in the Beta and Gamma distributions. It is seen that 3 best results are obtained in the estimations made with Hexamean and 2 best results are obtained in the estimations made with Heptamean and Nonamean. In Theil-Sen regression analysis, we can say that using the median parameter in data structures where both variables have the same distribution is sufficient to obtain the best result. When the results obtained from the means other than the median parameter are examined, the most successful results are seen as Hexamean with 5 and Heptamean with 4. No data structure has found the best results with Octamean.

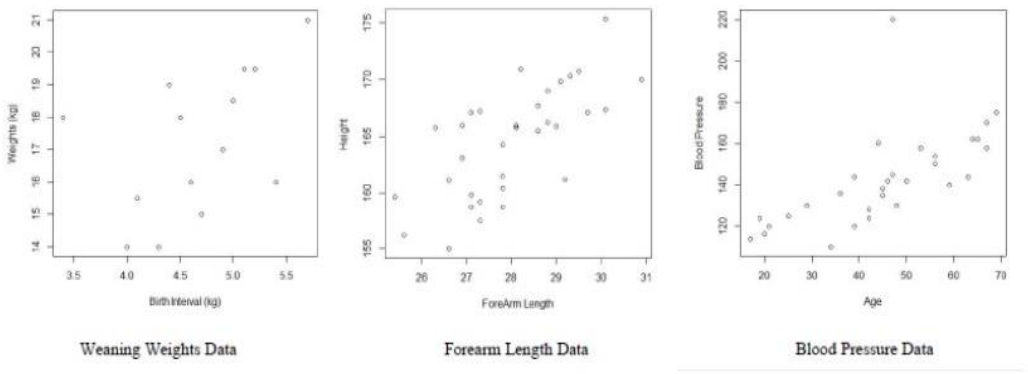
In the second part of the application, some real-life data used in the literature and the results obtained with the proposed means in different data structures that emerged as a result

of the changes in these data were investigated. These data are briefly introduced in Table 2 and the distribution chart of the data is given in Figure 1.

Tab. 2: Real-time data definitions used in application

Data	Simple Definiton	Sample Size	Source
Forearm Length	The heights (Y) in cm and forearm lengths (X) in cm of 33 black female applicants	33	Birkes and Dodge (1993)
Weaning Weights	Birth intervals in kg. (X) and weaning weights in kg. (Y) of ships	14	Topla (1999)
Blood Pressure	The systolic blood pressure (Y) and age (X)	30	Spaeth (1991)

Fig. 1: Data spread graph



Weaning Weights Data Forearm Length Data Blood Pressure Data

In Table 3, new data definitions and analysis results obtained with 3 different data structures and the changes made in them are given.

Tab. 3: Data analysis results generated from 3 real-life time data

Data	General (Including Median)	Only Trimean Family
ForeArm Length	Hexa	Hexa
ForeArm (Last 3 observations on X reduced by 40%)	Median	Tetra
ForeArm (Last observation on X reduced by 90%)	Median	Tetra
ForeArm (Last observation on Y reduced by 90%)	Nona	Nona
Blood pressure	Hepta	Hepta
Blood pressure (Last observation on X reduced by 90%)	Median	Tetra
Blood pressure (Last observation on Y reduced by 90%)	Median	Tetra
Weaning Weights	Tri	Tri
Weaning Weights (Last observation on X reduced by 90%)	Median	Hexa
Weaning Weights (Last observation on Y reduced by 90%)	Median	Nona

When the results in Table 3 are examined, it is seen that the best estimates of all 3 data in their original form are not the estimates made with the classical median. In addition, it is seen that the median parameter obtained the best estimates in the data with at least 1 outlier in the data. Although generalization cannot be made with these analyzes, it can be said that trimean family parameters can be used instead of the median parameter if they do not have extreme values in the data.

In addition, all of the models that gave the best predictive value in the model prediction results in Table 1 and Table 3 were statistically significant ($p<0.05$).

Conclusion

Non-parametric statistical methods use the median parameter instead of the arithmetic mean, which is affected by the slightest changes in the data. In this study, new means under the name of trimean family were introduced instead of the median parameter for non-parametric regression analysis and their use in Theil-Sen regression analysis was suggested. In the application, simulation data obtained from 7 different distributions and 3 real life data were used. In most of the simulation data, it was seen that the estimation results obtained with the median parameter were more successful. In addition, while the Trimean family results were more successful in real life data, it was determined that the median parameter gave successful results in the outliers formed in these data.

According to the simulation results obtained from 7 different distributions, it is seen that the suggested averages give the best model results in different data structures. After that, it was seen that TetraMean parameters in the data obtained from the normal distribution, Heptamean in the exponential and beta distributions, median in the log-normal and erlang distributions, and Nonamean parameters in the weibull and gamma distributions gave more successful results. It is also interesting that no best model results came out with the Octamean.

As a result, it is thought that using the proposed trimean family means in non-parametric regression methods will be a good alternative. It can give successful results, especially in cases where outliers are not excessive in the data. Although the assessments are made according to the estimation results obtained from this study, in future studies, stronger generalizations can be made by testing these means in different data structures as well as having low observations, right or left skewed structures or high rates of outliers.

References (Times New Roman, 14 pt., bold)

Birkes, D., and Dodge, Y. (1993). *Alternative methods of regression*. NY, USA: John Wiley & Sons Inc.

Erilli, N. A. and Alakuş, K. (2016). Parameter estimation in Theil-Sen regression analysis with Jackknife method. *Eurasian Econometrics, Statistics & Emprical Economics Journal,* 5; 28-41.

Erilli, N. A. (2021). Use of trimean in Theil-Sen regression analysis. *Bulletin of Economic Theory and Analysis*; 6(1); 15-26.

Erilli, N. A. (2022). Weighted Trimean as a Regressor in the Estimate of Theil-Sen Regression. *International Journal of Innovative Technology and Interdisciplinary Sciences.* 5, 2, 892-906

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric regression*. Marcel Dekker Inc., New York, USA.

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

Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32, 3, 669-679.

Öztaş, C. and Erilli, N. A. (2021). Contributions to Theil-Sen Regression Analysis Parameter Estimation with Weighted Median, *Alphanumeric Journal,* 2148-2225, 9, 2, 259-268.

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

Spaeth, H. (1991). *Mathematical Algorithms for Linear Regression*. London: Academic Press, UK.

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

Topal, M. (1999). *A Study on nonparametric regression methods. (*Unpublished Ph.D. Thesis*)*. Ataturk Univ., Graduate School of Natural and Applied Sciences, Erzurum, Turkey.

Tukey, J. W. (1977) *Exploratory Data Analysis.* Reading, Addison-Wesley, USA.

Contact

Necati Alp Erilli

Sivas Cumhuriyet University, Department of Econometrics

Sivas Cumhuriyet University, IIBF, Ekonomettri Bolumu, Kayseri Caddesi 57 58000

Sivas/Sivas, Turkey

e-Mail: aerilli@cumhuriyet.edu.tr