NON-PARAMETRIC REGRESSION ESTIMATION FOR DATA WITH EQUAL VALUES

N. Alp Erilli, PhD
Department of Econometrics, University of Cumhuriyet, Sivas, Turkey
Kamil Alakus, PhD
Department of Statistics, University of 19 Mayis, Samsun, Turkey

Abstract

Parametric regression analysis depends on some assumptions. One of the most important of assumption is that the type of relationship between dependent and independent variable or variables is known. Under such circumstances, in order to make better assumptions, regression methods which enable flexibility in the linearity assumption of the parametric regression are needed. These methods are nonparametric methods known as semi parametric regression methods. Estimation of parameters in a parametric regression which has independent variables of different values has been studied extensively in literature. Sometimes, one or more observation series of independent variable values can be equal while dependent variable values are different. This study offers a new method for the estimation of regression parameters under such data. Proposed method and other nonparametric methods such as Theil, Mood-Brown, Hodges-Lehmann methods and OLS method were compared with the sample data and the results were evaluated.

Keywords: Non-Parametric Regression, median, mean, theil

Introduction

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, and the other independent variables are held fixed.

Since the regression term was introduced by Sir Francis Galton, regression analysis is applied in many areas of academic and applied sciences such as social sciences, medical researches, economics, agriculture, biology, meteorology and chemistry. It is an easily comprehensible method and today, it has a wide area of usage and application with the help of statistical package software.

During the evaluation of an observed phenomenon, researching the underlying effects forms the basis of regression analysis. Regression analysis tries to find answers to questions such as “Is there a relationship between dependent and independent (explanatory) variables? If there is, what is the power of this relationship? What kind of a relationship is there between the variables? In the case of taking certain conditions under control, what is the effect of a specific variable or group of variables on the other variable or group of variables and how does it change?” During regression analysis, observation values and affected phenomena should be expressed through a mathematical presentation, that is, with the help of a function.

Regression analysis can be defined as the expression of a mean relationship between dependent and independent variables in the form of a mathematical function. This definition assumes a linear relationship between independent and dependent variables. A regression analysis in which only one independent variable is used is called univariate regression analysis whereas a regression analysis in which more than one independent variable is used is called multivariate regression analysis.

Regression analysis depends on some assumptions. The most important of these assumptions is that the type of relationship between dependent and independent variables is known. Estimates which are made when there are no assumptions cannot be good estimates. Under such circumstances, in order to make better assumptions, regression methods which enable flexibility in the linearity assumption of the parametric regression are needed. Therefore, these methods are regression models known as nonparametric and semi-parametric regression methods.

There are many regression analysis studies in literature. However, nonparametric regression analysis is not preferred much since it does not have new methods added and it doesn’t have a wide application area. Fernandes and Leblanc (2005) studied regression estimation under measurement errors with Theil regression; Lavagnini et al.(2011) studied reverse regression with the help of Theil regression; Zhou and Serfling (2008) studied multivariable spatial estimations with Theil estimator and Shen (2009) studied asymptotic multi linear regression estimation with Theil.

Parameter estimation in parametric regression in which the independent variable has different values has been widely studied in the literature. Sometimes, when one or more observation series from independent variables are equal, independent variable values can be different or same. For such data, nonparametric regression methods such as Theil, Hodges-Lehmann results in calculation errors since $\frac{y_2 - y_1}{x_2 - x_1}$ statistics is used for the estimation of $\hat{\beta}_1$. In this study, it is developed a new method for the estimation of nonparametric regression parameters under such data was developed. The offered method and other nonparametric methods such as Theil, Mood-Brown, Hodges-Lehmann and OLS methods were compared with the sample data and the results were discussed.

Regression Analysis

Regression analysis studies the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view of estimating or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter (Gujarati, 2003).

According to another definition, regression analysis is a statistical technique that uses the observed data in order to relate a variable called dependent (or responding) variable and one or more independent (or predictor) variables. The purpose is to form a regression model or estimating equation that can be used to define, estimate and control the dependent variable based on independent variables (Bowerman et al., 2010).

Depending on the use of regression analysis, it is divided into two categories; parametric regression and non-parametric regression.

Parametric – Nonparametric Regression

The most important difference between parametric and nonparametric regression methods depends on the trust in the information taken from the researcher and the data about the regression function. In parametric regression, the researcher chooses a possible family of curves from all the curves and needs very special quantitative data about the form of regression function (Eubank, 1988).

Nonparametric regression techniques depend on data more than parametric regression techniques in order to get information about the regression function. Thus, they are suitable for inference problems. Also, it is more suitable to use nonparametric estimators when there is no parametric form for the regression function, because when the parametric model is valid, nonparametric models will be less efficient. In addition, nonparametric models can be used to test the validity of parametric models (Eubank, 1988).

Parametric Regression Analysis

Parametric regression is the expression of dependent and independent variables and the average relationship between these variables is expressed through a mathematical function and the clear representation of parameter vectors in this function. For a successful application of parametric regression analysis, assumptions such as normal distribution, homoscedasticity and autocorrelation should be provided. Thus, they are the most powerful regression methods in the event of assumptions' becoming a reality.

Non-Parametric Regression

These are methods used when some assumptions valid for parametric regression methods are not provided. They are effective methods for data which have low sample size or contradictory sample. In statistical studies, there are robust parametric methods which can address the effects of outliers differently. However, since parameters are spoiled because of outliers, even these robust methods may not generate suitable solutions and the real form of the data may not be reflected in the model. Thus, nonparametric regression provides preliminary information (Härdle, 1994).

Although nonparametric regression does not have restrictive assumptions while making estimations, it has some disadvantages. When there are too many independent variables, it is difficult to make estimations and the graphics may become complicated. In addition, with nonparametric method, it is difficult to take discrete independent variables into consideration and to comment on the individual effects of dependent variables because of the increase in independent variables. The disadvantages of nonparametric methods can be resolved by using semi parametric regression models (Horowitz, 1993).

Semi Parametric Regression Analysis

Semi parametric regression model uses both parametric and nonparametric regression models. Thus, it not only affected by the restrictive assumptions of parametric models, but it also brings together the attractive features of nonparametric methods (such as Cox and Kernel Regression).

Semi parametric regression models can be defined as a combination of parametric and nonparametric regression models. They are used when nonparametric regression methods cannot make better estimations or when the researcher wants to use parametric methods but does not know the distribution of errors. Therefore, assumption of normality is not needed while making parameter assumptions with these models (Sprent and Smeeton, 2001).

While it is possible to work with two explanatory variables at most in order to get interpretable results in nonparametric model estimation, it is possible to analyze the dependent effect of *k* numbered explanatory variable in semi parametric method. In addition, since it does not make as many assumptions as the parametric model does, semi parametric models are advised to be used in practical studies (Horowitz, 1993).

Some Non-Parametric Regression Methods
Mood-Brown Method

Suppose that we have data, which consist of independent variable *X* and dependent variable *Y* which is measured in terms of least equally spaced level. Thus, the data will be comprised of $(x_1, y_1),(x_2, y_2),\cdots,(x_n, y_n)$ pairs. In addition, the suitable model for scatter diagram is the $Y = \beta_0 + \beta_1 X + u$ linear regression model. Mood-Brown method is used only for estimating the parameters of simple linear regression model (Mood and Brown, 1951). The algorithm of the method is given as follows.

i. Sample units are lined from smallest to the biggest in terms of the values they get for *X* variable (called the natural line of *X*).

ii. Scatter diagram is made for the obtained data (sequential pairs) from sample units in terms of *X* and *Y* variables.

iii. A straight line is drawn to the middle of *X*: on this straight line, half of the dots on the scatter diagram will stay on the left side of the straight line while the other half will remain on the right side. Let's call the small dots from the median of *X* variable as the first group and the big dots as the second group dots.

iv. Medians of sample units in the first group in terms of *X* and *Y* variables are found. Similarly, the same process is done for the sample units in the second group. Here;

x_1 : The median of sample units in the first group in terms of *X* variable.

y_1 : The median of sample units in the first group in terms of *Y* variable.

x_2 : The median of sample units in the second group in terms of *X* variable.

y_2 : The median of sample units in the second group in terms of *Y* variable.

v. Let's show the pairs in the fourth step with (x_1, y_1) and (x_2, y_2) $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ line is the line passing from (x_1, y_1) and (x_2, y_2) points. The slope of the straight line passing from these two points is found by; $\hat{\beta}_1 = \dfrac{y_2 - y_1}{x_2 - x_1}$.

The point where the straight line cuts the *Y* axis can be determined with the help of the equivalences $y_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$ or $y_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2$ (Hill, 1962; Hogg and Randles, 1975).

Theil Method

After Theil (1950) introduced this method, it became one of the most frequently used nonparametric regression methods. When the sample units are taken in pairs in order to find the $\hat{\beta}_1$ statistic which is the estimator of β_1 parameter, the slopes of all cases are calculated through $S_{ij} = \frac{y_j - y_i}{x_j - x_i}$ (j>i) equivalence. The number of all S_{ij} estimates calculated which depends on (j > i) requirement can be as much as $\frac{(n-1)n}{2}$. $\hat{\beta}_1$ statistics which is the estimator of β_1 parameter is the median of slope values. For these, $\hat{\beta}_1$ is equal to $Med(S_{ij})$(Hussain and Sprent, 1983). After the slope of regression straight line is found, (x_i, y_i)or (x_j, y_j) that belong to the *i*-th or *j*-th variables can give $Med(S_{ij})$value which are found by putting the dots in place $\hat{\beta}_0 = y_i - \hat{\beta}_1 x_i$ or $\hat{\beta}_0 = y_j - \hat{\beta}_1 x_j$

Methods Which Estimate the Slope (β_1) and Intercept (β_0) Parameters with Common Way

Here, two methods are introduced which were developed from Theil method. Let's, $Y = \beta_0 + \beta_1 X$ shows simple regression model and $\hat{\beta}_1$ shows Theil estimation. The following two different methods are methods introduced for β_0 estimation. Therefore, both two methods obtained from all data which depend on getting estimation values with $d_i = y_i - \hat{\beta}_1 x_i$

Optimum Estimation Method Based On Sign Test

Optimum estimation value are calculated from $d_i = y_i - \hat{\beta}_1 x_i$ and the median of these values is calculated as $\hat{\beta}_0$. In this method, d_i values are not supposed to be symmetrically distributed. However, this method gives more successful results especially in data which has extreme values.

Hodges-Lehmann Method

This method is the developed version of Theil-Kendall method. $d_i = y_i - \hat{\beta}_1 x_i$ variable is also defined to find $\hat{\beta}_1$ statistics. This approach necessitates the assumption that d_i values are distributed symmetrically around β_0. Here, $d_i = y_i - \hat{\beta}_1 x_i$ values are calculated for all data and

$\hat{\beta}_0$ value calculated with the arithmetical median of d_i values, which is $\hat{\beta}_0 = \frac{\sum d_i}{n}$. A big number of outliers decrease the robustness of the method (Lehmann, 2006).

Proposed Method

Let's take a data in which some of the x_i observations are equal and which consists of $(x_1, y_1),(x_2, y_2),\cdots,(x_n, y_n)$ consecutive pairs. Thus, when two observations are equal $(x_i = x_j)$ then $S_{ij} = \frac{y_i - y_j}{x_i - x_j}$ $(i \neq j)$ slope values will be unclear and it would not be possible to calculate S_{ij} slopes. In such circumstances, both the median and mean of y_i observations which have equal x_i will be found separately and each x_i value will be represented with one dependent variable value. Thus, it will be possible to get two different regression data sets and it will be easier to apply nonparametric regression techniques to the data.

In Theil method, $\hat{\beta}_1$ statistic is the median of slope values. In addition to slope parameter values and optimum and Hodges-Lehmann calculations, $\hat{\beta}_0$ was calculated with the help of introduced methods. The calculated slope parameters were placed in the $d_i = y_i - \hat{\beta}_1 x_i$ equivalence and d_i values were calculated for each (x_i, y_i) observation pairs and they were lined and placed in order from the smallest to the biggest. This calculation was found separately for the series found with both median and mean.

Application

Data.1 (Simulation Data)

Simulation data set consists of 22 surveys. Dependent variable (Y_i) and independent variable (X_i) scores are given in Table.1

Table.1. Simulation data

Y_i	30	40	45	50	50	55	60	70	60	75	80	90	85	80	90	92	95	98	96	100
X_i	4	5	5	5	6	8	9	9	10	10	10	10	10	11	11	11	12	13	14	15

Firstly, we must sort data from smallest to largest according to X_i scores. Then, we calculate mean and median scores for same X_i surveys. Organized and listed data, according to mean and median is given Table.2.

Table.2 Ranking simulation data according to Mean and Median

Median Ranking		Mean Ranking	
x_i	y_i	x_i	y_i
4	30	4	30
5	45	5	45
6	50	6	50
8	55	8	55
9	65	9	65
10	***80***	**10**	***78***
11	***90***	**11**	***87,3***
12	95	12	95
13	98	13	98
14	96	14	96
15	100	15	100

According to table above, data from the median and mean size is calculated and ranked from smallest to largest. Regression models parameters are calculated as described above which is shown in Table.3a and Table.3b.

Table.3a Regression Parameter Estimation Results

β_0	β_1	Used Method
-14,875	9,625	Mood-Brown
10,364	6,562	OLS

Table.3b Regression Parameter Estimation Results

Mean		Median		Used Method	
β_0	β_1	β_0	β_1	β_0	β_1
3,6	6,6	11,875	6,625	Theil	
10,5	6,6	10,3125	6,625	Optimum	Theil
10,01	6,6	9,775	6,625	Hodges-Lehman	

RMSE results calculated for regression models given in example.1 are shown in Table.4. RMSE for Mood-Brown estimated value and Theil estimator values are far away from RMSE for OLS. Other estimated results are quite closed to OLS results.

Table.4 RMSE Results for all regression models for study time data

RMSE value	Used Method	
153,9703	Mood-Brown	
55,63696	OLS	
96,738	Theil	Median
55,89	Optimum and Theil	
55,6499	Hodges-Lehman and Theil	
60,08281	Theil	Mean

55,96172	Optimum and Theil	
55,67281	Hodges-Lehman and Theil	

Data 2. (Real Data)

The relationship of Samsunspor soccer team away goals and scoring goals were investigated. The numbers of away and scoring goals, which they played between 1995-2013 seasons, are given in Table.5.

Table.5 Samsunspor Data

Season	Scoring Goals	Away Goals	Season	Scoring Goals	Away Goals
1994-95	54	60	**2004-05**	40	55
1995-96	45	45	**2005-06**	45	62
1996-97	49	52	**2006-07**	31	38
1997-98	42	42	**2007-08**	45	61
1998-99	38	53	**2008-09**	35	47
1999-00	51	43	**2009-10**	49	47
2000-01	55	52	**2010-11**	35	**14**
2001-02	32	43	**2011-12**	36	47
2002-03	42	59	**2012-13**	38	39
2003-04	46	47			

Organized and listed in order of data according to mean and median is given in Table.6.

Table.6 Ranking Samsunspor data according to Mean and Median

Median Ranking		Mean Ranking	
Scoring Goal	**Away Goal**	**Scoring Goal**	**Away Goal**
31	38	31	38
32	43	32	43
35	30,5	35	30,5
36	47	36	47
38	46	38	46
40	55	40	55
42	50,5	42	50,5
45	**61**	45	**56**
46	47	46	47
49	49,5	49	49,5
51	43	51	43
54	60	54	60
55	52	55	52

According to the table above, data from the median and mean size is calculated and ranked from smallest to largest. Regression models parameters are calculated as described above which is given in Table.7a and Table.7b.

Table.7a Regression Parameter Estimation Results

β_0	β_1	**Used Method**
18,0772	0,6923	Mood-Brown
17,378	0,713	OLS

Table.7b Regression Parameter Estimation Results

Mean		**Median**		**Used Method**	
β_0	β_1	β_0	β_1	β_0	β_1
19,9177	0,5833	19,6604	0,5916	Theil	
21,7932	0,5833	23,0116	0,5916	Optimum	Theil
23,4183	0,5833	22,5256	0,5916	Hodges-Lehman	

RMSE results calculated for regression models given in example.2 are shown in Table.8. RMSE for OLS is the minimum value but Mood-Brown, Optimum and Hodges-Lehman calculated from mean and median are quite close to OLS result.

Table.8 RMSE Results for all regression models for Samsunspor data

RMSE	**Used Method**	
89,65279	Mood-Brown	
89,60539	OLS	
98,52479	Theil	Median
90,55134	Optimum and Theil	
90,31518	Hodges-Lehman and Theil	
99,18288	Theil	Mean
91,59401	Optimum and Theil	
90,70716	Hodges-Lehman and Theil	

Data 3.

Here, we changed one of the observations from the data given in example.2. Season 2010-2011 datum changes as home goal:88 and away goal:35. Thereby, datum for 2010-2011 became as a outlier. Hence, we can test the power of the proposed methods within outlier data. Scatterplot of new data is given Figure.3.

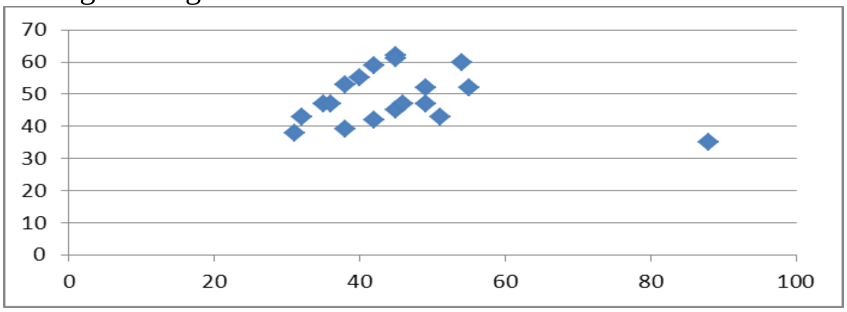

Figure.1 Scatterplot for Changed Samsunspor data

Regression models parameters are calculated as described before which is given in Table.9a and Table.9b.

Table.9a Regression Parameter Estimation Results

β_0	β_1	**Used Method**
25,891	0,5555	Mood-Brown
29,79	0,419	OLS

Table.9b Regression Parameter Estimation Results

Mean		**Median**		**Used Method**	
β_0	β_1	β_0	β_1	β_0	β_1
32,51	0,355	32,51	0,355	Theil	
32,475	0,355	32,475	0,355	Optimum	Theil
32,7023	0,355	32,7023	0,355	Hodges-Lehman	

RMSE results calculated for regression models given in example.3 are shown in Table.10. RMSE values for OLS is the maximum value with Mood-Brown statistics. Outlier of dependence variable impresses OLS and Mood-Brown as expected. Adjusted mean results are quite better esp. with Theil β_1 both mean and median ranking.

Table.10 RMSE Results for all regression models for Samsunspor data-2

RMSE	**Used Method**	
125,7441	Mood-Brown	
97,7355	OLS	
88,94584	Theil	Median
88,96053	Optimum and Theil	
88,90884	Hodges-Lehman and Theil	
88,94584	Theil	Mean
88,96053	Optimum	
88,90884	Hodges-Lehman and Theil	

Conclusion

Nonparametric regression analysis is a regression method which is preferred when classical regression analysis assumptions are not valid or when the sample number is very low. These methods which generally give more successful results in simple regression are more effective in cases when there are outliers.

In parameter estimation of parametric regression in which the independent variable has different values, if an observation series or more of independent variable values are equal, the dependent variable values can be different or same. For such data, nonparametric regression methods like Theil, Hodges-Lehmann cannot be calculated since $\frac{y_2 - y_1}{x_2 - x_1}$ statistic is used for the estimation of $\hat{\beta}_1$. This study proposes a new method for the estimation of nonparametric regression parameters under such data. The method proposed and other nonparametric methods such as Theil, Mood-Brown, Hodges-Lehmann methods and OLS method were compared with the sample data.

In the data set which the independent variable has outliers, the OLS estimators gave incorrect values as expected. The proposed method produced more successful results like other nonparametric regression methods. Moreover, the proposed methods' results are close to OLS results in the data set which were close to normal distribution and in the data set which the dependent variable has outliers. This shows that the proposed method can be among the alternative nonparametric regression family.

The method proposed present methods and other nonparametric regression methods such as Theil, Mood-Brown, Hodges-Lehmann methods and OLS method which were compared with the sample data. However, since the analysis were made without searching if the data had the linear regression assumptions for the OLS method or not, the analysis results were in favor of OLS. The validity of these assumptions should certainly be researched in real life. However, if we take this weak side into account, both the proposed method and parametric methods would give away quite valuable results.

References:

Bowerman B.L., O'Connell R.T., and Murphree E.S. Business Statistics in Practice. McGraw-Hill Irwin Pub., NY, USA., 2010.

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

Fernandes R., and Leblanc S.G. 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, Volume 95, Issue 3,Pages 303-316, 2005.

Gujarati D. Basic Econometrics. McGraw-Hill Education Pub.,NY, USA., 2003.

Härdle, W. Applied nonparametric regression. Cambridge University Press Cambridge, UK., 1994.

Hill, B.M. A test of linearity versus convexity of a median regression curve. Annals of mathematical statistics, 33, 1096-1123, 1962.

Hogg R.V., and Randles R.H. Adeptive distribution free regression methods and their applications. Technometrics, 17 (4), 399-407, 1975.

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

Hussain S.S., and Sprent P. Non-Parametrik Regression. Journal of the royal statistical society. Ser., A., 146, 182-191, 1983.

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

Lehmann, E. L. Nonparametrics: Statistical methods based on ranks. With the special assitance of H.J. D'Abrera. New-York: Springer, pp. 463, 2006.

Mood, A.M., and Brown, G.W. On Median Tests for Linear Hypotheses. Proceedings of the Second Berkeley Symposium On Mathematical Statistics and Probability, Berkeley and Los Angeles: The Universitey of California Pres., USA, 1951.

Shen G. Asymptotics of a Theil-type estimate in multiple linear regression Statistics & Probability Letters, Volume 79, Issue 8, 15 April 2009, Pages 1053-1064, 2009.

Sprent P., and Smeeton N.C. Applied nonparametric statistical methods. Chapman & Hall/CRC, USA, 2001.

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

Zhou W., and Serfling R. 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, V. 138, Issue 6, Pages 1660-1678, 2008.