IASJ
المجلات الاكاديمية العراقية

Tikrit Journal of Pure Science
ISSN: 1813 – 1662 (Print) --- E-ISSN: 2415 – 1726 (Online)

Journal Homepage: http://tjps.tu.edu.iq/index.php/j

Comparison the Robust Estimators Nonparametric of Nonparametric Regressions

Ali Fadhil Abduljabbar , Afrah Mohammed Kadhim
Middle Technical University Kut Technical Institute
https://doi.org/10.25130/tjps.v28i1.1271

ARTICLE INFO.

Article history:
-Received: 23 / 8 / 2022
-Accepted: 26 / 10 / 2022
-Available online: 20 / 2 / 2023

Keywords: Mean Absolute Deviation (MAD), Mean Squares of Integral Error (MSIE), Real function k (h), Weight Series (w_i)

Corresponding Author:
Name: Ali Fadhil Abduljabbar
E-mail:
ali.fadhil@mtu.edu,iq
Afrahkadm88@gmail.com
Tel:

ABSTRACT

In order to get rid of or reduce the abnormal values of some phenomena that may be the reason for not obtaining the desired results. This makes us to get conclusions far from reality for the phenomenon we are studying. That the traditional nonparametric estimators are very sensitive to anomalous values, which prompted us to use the fortified estimators because they are not much affected by the anomalous values, as well as the nonparametric regression because it does not depend on the previous determinants or assumptions, but it depends directly and fundamentally on the data.

1. Introduction

Researchers face several problems, including that the data for some phenomena may contain anomalous values, which can lead to inaccurate results, so these inaccurate results will lead to conclusions that are far from the reality of the phenomenon under study, so it is necessary to conduct research on these are anomalies. In addition to ensuring their proportion to the total sample size that represents the phenomenon under study, the traditional nonparametric estimators are very sensitive to outliers, which prompts researchers to use immune estimators because they are not affected by the presence of outliers, which does not take a predetermined form of a function, but its estimates based on the data, and thus nonparametric regression has been addressed because it does not depend on previous determinants or assumptions, but rather depends directly and fundamentally on the data.

2. Traditional Nonparametric Capabilities:

2.1. Libyan functions [1-6]:-

It is the simplest form of nonparametric regression to find a data pattern without the need for a parameter model through a series of weights, and it has several names, including (a window function, a weight function, a basic function, and a shape function), and it is characterized by being a real function symbolized by the symbol, symmetric, continuous, and its integration equal to one and the derivative. The second is known and limited $k(h)$, meaning that

$$\int k(h)dh = 1 \quad , \int hK(h)dh = 0 \quad , \int h^2K(h)dz = k_2 < \infty \quad (1)$$

Among the most used Libyan functions are:
Gaussian kernel is known by the following formula:-

$$k(x) = \frac{1}{\sqrt{2x}} exp(-x^2/2) \quad (2)$$

And the Epanechnikov kernel is known by the following formula:

$$k(x) = \begin{cases} \frac{3}{4}(1-x^2) & if\ |x| \le 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

There are two series of Libyan functions: the Libyan functions of least variance, which reduce the variance, and the ideal Libyan functions, which reduce the mean of integrated error squares (MISE) and was derived in 1969 by the world*Epanch ni*Kov. that functionKernel can be written in the following form:

$$f(x) = E(y\backslash x) \quad = \int \frac{yf(x,y)}{f(x)}\ dy, \quad (4)$$

Since:

Y: is the (dependent) response variable x,: explanatory variable.

Also, the weight chain w_i Kernel estimates are written as follows:

$$w_i(x) = k_h(x - x_i)/\hat{f}_h(x) \quad (5)$$

Since:

- *h: the bandwidth is greater than zero.*
- $\hat{f}_h(x)$*: the estimated density function.*
- k_h*Kernel function.*

$$w_i(x_i) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right)} \quad (6)$$

2.2. The Core Estimator of Nadaria Watson [7- 12]

Nadaraya - Watson Kernel Estimator

It is the oldest and most widespread and widely used nonparametric capabilities, which were suggested by the two researchers Nadaraya - Watson Kernel in 1964, and it is derived based on the series of weights and is used in the static and random design. It is characterized by its continuous and positive function, which has an integral equal to one.

The estimate can be derived $m(x)$ as follows:-

$$m(x) = \int \frac{yf(x,y)}{f(x)}\ dy = \int \frac{y\frac{1}{n}\sum_{i=1}^{n} y_i K_{h1}(x-X_i)K_{h2}(y-Y_i)}{\frac{1}{n}\sum_{i=1}^{n} K_{h1}(x-x_i)}\ dy \quad (7)$$

$$f(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_{h1}(x-X_i)\int yK_{h2}(y-Y_i)dy}{\frac{1}{n}\sum_{i=1}^{n} K_{h1}(x-X_i)}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} K_{h1}(x-X_i)Y_i}{\frac{1}{n}\sum_{i=1}^{n} K_{h1}(x-X_i)}$$

$$\frac{\sum_{i=1}^{n} K_{h1}(x-X_i)Y_i}{\sum_{i=1}^{n} K_{h1}(x-X_i)}$$

$$\hat{m}_{NW}(x_i) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right)}, h > 0 \quad (8)$$

Since:-

$K(\frac{x-x_i}{h})$ The Libyan function is continuous and specific.

h: The boot parameter has a value greater than zero. It controls the bootstrap quantity of the output estimator.

Based on the weights series method:

$\hat{m}_{NW}(x_i) = \sum_{i=1}^{n} w_i(x)Y_i$ Becausew_iWeight function $\sum_{i=1}^{n} w_h(x - X) = 1$,

And to find the estimator of the function at the point x in its domain, we specify the bandwidth, which is the boot parameter h. It controls the width of the parameter around x. To ensure smoothing, the views close to x are given more weight than the farthest views, which is determined by the kernel functions in the form of weights.

2.3. K-Nearest Neighbor Estimator [13-16] K-Nearest Neighbor Estimator

This estimator depends on finding, the distance between each site in the area and adjacent siteswhich is represented by a pointIt is used in the random modelThat is, the values that are used in calculating the mean are those values corresponding to the values of X observed to the point X in the Euclidean distance at which we want to estimate m..

The series of weights is defined in the method w_{ni} K-N-N as it comes

$$W_{ni}(x) = \begin{cases} \frac{1}{k}, & if \quad i = 1,2,\dots,n \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$\hat{m}_{K-N-N}(x_i) = \frac{1}{k}\frac{\sum_{i=1}^{n} y_i k\left(\frac{x_i-x_j}{k_n}\right)}{\sum_{j=1}^{n} k\left(\frac{x_i-x_j}{k_n}\right)}, \quad i,j = 1,2,\dots,n \quad (10)$$

Since (k_n)K represents a constrained and nonnegative kernel function, the Euclidean distance between x and k from the nearest neighbor of x.

The introductory parameter is calculatedk through:

$$k_n = d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad , i,j = 1,2,\dots\dots,n \quad (11)$$

over here, K = Kn when ∞⟶kn and∞⟶n. mean ingelse, if applied function kernel on this estimator as in the Nadaraya-Watson model, k represents the smoothing parameter corresponding to the value h in the Nadaraya-Watson estimator.

3. Fortified capabilities]: Robust Estimators [17, 18, 19]

Several definitions of immunity have been provided Robust ness, Box first From mention it because method statistic it's called impregnableIf the statistical inference is not significantly affected by the violation of any of its prerequisites.

In general, immunity is to ignore outliers or reduce their impact on the data, and a immune estimate is an estimate that has little effect on outliers and has an efficiency similar to least squares estimators in the event of outliers or outliers.

Thus, the fortified capabilities can be explained as follows:-

When one of the regression assumptions is defective or there are outliers or random errors distributed in a distribution other than a normal distribution consistent with the method used, the estimator retains the expected properties of the estimator in the estimator and when the data violate the analysis is used in the case of one condition, the effect of the immunogenicity estimator is small for a broad distribution.

3.1. Appraisal methods For Tress

There are many methods of estimating the nonparametric regression through the regression equation represented by the explanatory variables (x)

and the response variable (y) to estimate the nonparametric regression function in the equation:xy

$y_i = f(x_i) + \varepsilon_i, i = 1,2,3 \ldots \ldots, n$ (12)

$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x-x_i}{h})$ (13)

3.2. The Core Estimator of the Hippocampal Nadaria Watson [20]

Robust Nadaraya - Watson Kernel Estimator

The hippocampal Nadaria-Watson core estimator is one of the most common estimators in the robust estimation methods, and when using the robust weight function.ψ The formula will be as follows:-

$\hat{m}_{NW}(x_i) = \psi n^{-1}\sum_{i=1}^{n} w_i(x)Y_i$ (14)

$W_{ni}(x) = \frac{\sum_{j=1}^{n} k(\frac{x_i-x_j}{h})}{n^{-1}\sum_{j=1}^{n} k(\frac{x_i-x_j}{h})}, \quad j = 1,2,\ldots.,n$ (15)

Since:-

$K(\frac{x-x_i}{h})$ The Libyan function is continuous and specific.

h is a boot parameter and its value is greater than zero.

w_i Endodontic weight function.

The robust weight function is used **(Huber)[21]** It takes the following form:

$$\psi = \begin{cases} 1 & if\ |e_i| < c \\ \frac{c}{|e_i|} & otherwise \end{cases} \quad (16)$$

Because C takes the default value 1.345

3.3. Robust K-Nearest Neighbor Estimator [22,23]

We use the nearest neighbor estimator K-N-N in Strong estimation methods .For the nonparametric regression by substituting the robust weight function (ψ) (as in the following formula:-

$$\hat{m}_{K-N-N}(x_i) = \frac{1}{k}\frac{\sum_{i=1}^{n} \psi y_i k(\frac{x_i-x_j}{kn})}{\sum_{j=1}^{n} k(\frac{x_i-x_j}{kn})} \quad (17)$$

$$W_{ni}(x) = \frac{k(\frac{x_i-x_j}{kn})}{\sum_{j=1}^{n} k(\frac{x_i-x_j}{kn})}, \quad i,j = 1,2,\ldots.,n$$

And the Probably The use of the robust weight function (Huber) As in the following formula:-

$$\psi = \begin{cases} 1 & if\ |e_i| \le k \\ \frac{k}{|e_i|} & if\ |e_i| > k \end{cases} \quad (18)$$

becausek takes the default value= 1.345,= standard deviation and it is calculated from the following formula:-σ σ

$\hat{\sigma} = \frac{MAR}{0.6745}$ (19)

MAD represents average the rest the absolute.

3.4. Method M Fortified: [24- 18]

This method is considered one of the mostaIt is safe because of the high efficiency of obtaining abilities using the method of least squares, where the idea is based on reducing somefunctionserror instead of reducing the sum of squares.This method has also received a lot of attention by researchers because a more flexible and T Provide the possibility of direct generalization to multiple regression It was suggested by the researcher Huber (1973) , and the idea of this method is to find the lowest value ofa and b and the hippocampal estimator is determined by a function test weight ψ.

The linear regression model is described by the following relationship:

$yi = \grave{x}i\beta + \varepsilon i$ (20)

Estimator can get M by decreasing or minimizingThe following amount:-:

$Min \sum_{i=1}^{n} \rho(\frac{e_i}{s}) = Min \sum_{i=1}^{n} \rho(\frac{y_i - \grave{x}_i\beta}{s})$ (21)

becauses is the measurement estimate and is found from the following equation

$s = \frac{median|e_i - median(e_i)|}{0.6745} = \frac{MAD}{0.6745}$ (22)

MAD is the mean absolute deviation.

(ρ): The objective function has many properties, includin

$\rho(e) \ge 0$

$\rho(0) = 0$

$\rho(e) = \rho(-e)$

$\rho(e_i) \ge \rho(e_i)\ for\ |e_i| \ge |e_i|$

And the assuming that $\psi = \grave{\rho}$ differentiable around, called the effect curve ρ

$\sum_{i=1}^{n} \psi(\frac{y_i - \grave{x}_i\beta}{s})\, x_i = 0$

The weight function must be tested ψ to determine the hippocampal estimator.

$\sum_{i=1}^{n} \rho(y_i - a - b(X_i - x))k\left(\frac{X_i - x}{h_n}\right)\delta_i$ (23)

Or positional estimation of equations:-

$\sum_{i=1}^{n} \psi(y_i - a - b(X_i - x))k\left(\frac{X_i - x}{h_n}\right)\delta_i = 0$ (24)

$\sum_{i=1}^{n} \psi(y_i - a - b(X_i - x))\left(\frac{X_i - x}{h_n}\right)k\left(\frac{X_i - x}{h_n}\right)\delta_i = 0$ (25)

Since:

$\rho(.)$ a convex and symmetric function, $\psi(.)$derived ρ which is the fortified weight function, $K(.)$ the Libyan kernel function, $h_{n,}$ a series of positive numbers.

Since the immunity of the estimator depends on the function of weightsψwhichcan defineHaIt is a set of functions on the basis of which the weights accompanying the observations are determined,It is possible to use the robust weight function (Huber) as in the following formula:-

$$\psi = \begin{cases} 1 & if\ d_i \le c \\ 0 & if\ d_i > c \end{cases} \quad (26)$$

Since it c is the cutoff constant, it is the tabular value of the chi-square distribution with a degree of freedom ρ and a significant level α.

There are several functions on which the method depends Mimmune to it

3.4.1. Function Bisquare [29]

It takes the following form

$$\psi = \begin{cases} (1 - (\frac{e_i}{c})^2)^2 & if\ |e_i| < c \\ 0 & otherwise \end{cases} \quad (27)$$

On the assumption thatc takes the value 4.685

3.4.2. Function Talkie [30]

It takes the following form

$$\psi = \begin{cases} x(1 - (\frac{x}{c})^2) & if\ |x| < c \\ 0 & otherwise \end{cases} \quad (28)$$

and c takes The following values are 4.685 and 6.0

3.4.3. Function Hampel [31]

$$\psi = \begin{cases} 1 & if\ |e| \le a \\ (\frac{a}{|e_i|}) & if a < |e_i| \le b \\ \frac{a(c-|e_i|)}{|e_i|(c-b} & if\ b < |e_i| \le c \\ 0 & otherwise \end{cases} \quad (29)$$

So that a,b,c constants

3.4.4. Function Andrews[32]

$$\psi = \begin{cases} (\frac{sin\frac{x}{c}}{\frac{x}{c}}) & if\ |x| \le \pi c \\ 0 & otherwise \end{cases} \quad (30)$$

3.4.5. Trimmed Least Squares Method (33) (34) (35) (11) (12) (13)

Least Trimmed Squares (LTS)

It was proposed for the first time by Rousseeuw in 1984, where the sum of the squares of errors is minimized after they are arranged in ascending order. It is calculated from the following equation:-

$$\hat{\varrho}_{lts} = \sum_{i=1}^{h}(r^2)i{:}n \quad (31)$$

because $h.(r^2)1{:}n \le (r^2)2{:}n \le (r^2)i{:}n \le \ldots.. \le (r^2)i{:}n$ Represents the square and ascending remainders, and h represents the observations that are adopted after eliminating outliers whose value is equal to

$h = \frac{n}{2} + \frac{p+1}{2}$, and represent, n and p are the given sample size and number of independent variables in the model, respectively.

This method is characterized by having a high breakdown point equal to 0.5, which can be found from the following relationship:

$BP = \frac{n-h}{n}$ but if the breaking point exceeds the aforementioned limit, then Na. The main disadvantage of can not distinguish between the good part and the abnormal part of the data The hippocampal LTS is the large number of operations required to sort the square values into a target function.

References

[1] Dilmaç, H., & Şişman, Y. New approaches for outlier detection: The least trimmed squares adjustment. International Journal of Engineering and Geosciences, 8(1), 26-31.

[2] M. Wand and B. Ripley, "KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)," R Packag. version, vol. 2, p. 19–22, 2006.

[3] CM Hurvich, JS Simonoff, and C. Tsai, "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion," JR Stat. Soc. Ser. B (Statistical Methodol., vol. 60, no. 2, pp. 271–293, 1998.

[4] Blanchett. J and Wadsworth. J. (2012). 5. "Applied Statistics", Institute of Mathematics, Analysis, and Applications EPF Lausanne An MSc Course for Applied Mathematicians, working paper.

[5] Racine, JS (2008). Nonparametric econometrics: A primer. Foundations and Trends® in Econometrics, 3(1), 1-88.

[6] Midi, H., Rana, MS, & Imon, AHMR (2009). The performance of robust weighted least squares in the presence of outliers and heteroscedastic errors. WSEAS Transactions on Mathematics, 8(7), 351-361.

[7] Hardle W. (1994), "Applied Nonparametric Regression", Gambridg MA: Cambirg University Press.

[8] Aydin D. (2007), "A Comparison of the Nonparametric Regression Models using Smoothing Spline and Kernel Regression", World Academy of Science, Engineering and Technology.

[9] Demir, S., & Toktamiş, Ö. (2010). On the adaptive Nadaraya-Watson kernel regression estimators. Hacettepe Journal of Mathematics and Statistics, 39(3), 429-437.

[10] Turlach, BA (1993, January). Bandwidth selection in kernel density estimation: A review. In CORE and Institut de Statistique. 11. Wand, MP, & Jones, MC (1994). Kernel smoothing. Chapman and Hall/CRC.

[11] KOUASSI, BC, Hili, O., & KATCHEKPELE, E. (2021). Nadaraya-Watson estimation of a nonparametric autoregressive model. Malaya Journal of Matematik, 9(4), 251-258.

[12] Cheng C.-B.& Lee ES, (1999), "Nonparametric fuzzy regressionk-NN and kernel smoothing techniques" Computers and Mathematics with Applications, 38, 239–251

[13] Wanas, MS, & AL-Sharoat, MH (2020). Studying some approaches to estimate the smoothing parameter for the nonparametric regression model. Periodicals of Engineering and Natural Sciences, 8(2), 673-683.

[14] Michael G. Schimek (2000), "Smoothing and Regression (Approaches, Computation, and Application)", University of Graz and University of Vienna, Karl-Franzens, Austria

[15] Gibran Abdel-Amir, Korkis Shahid, Muhammad Salem (2018), Estimating the parameters of multiple linear regression using robust methods (a comparative study), Al-Qadisiyah Journal of Computer Science and Mathematics, Volume (8), Issue (1), p. 12-21.

[16] Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. The Annals of Mathematical Statistics, 27(3), 832-837.

[17] R. Marona, R. Martin, VJ Yohai. 2006. Robust Statistics Theory and Methods. John Wiley & Sons Ltd., England.

[18] Härdle, W. (1984). Robust regression function estimation. Journal of Multivariate Analysis, 14(2), 169-180. [

[19] Huber, PJ, (1964). "Robust Estimation of location parameter." Ann .math .stat.35-73-101.

[20] Fox, J., & Weisberg, S. (2012). Robust regression. An R and S-Plus companion to applied regression, 91.
[21] Deribati, MM, Younso, A., & Al-shakh, D. Converges in of nearest neighbor regression function estimate for strong mixing processes 1.
[22] Chen C., (2002), “Robust regression and outlier detection with the Robustreg procedure” In Proceedings of the Twenty Seventh Annual SAS Users Group International Conference; SAS Institute: Cary, NC.
[23] Y. Susanti, H. Pratiwi, and T. Liana, Application of Mestimation to Predict Paddy Production in Indonesia, presented at IndoMS International Conference on Mathematics and Its Applications (IICMA), Yogyakarta, 2009.
[24] Yuliana and Y. Susanti, Estimasi M dan sifat-sifatnya pada Regresi Linear Robust, Jurnal Math-Info, 1, No. 11 (2008), 8-16.
[25] Y. Susanti and H. Pratiwi, Robust Regression Model for Predicting the Soybean Production in Indonesia, Canadian Journal on Scientific and Industrial Research, 2, No. 9 (2011), 318-328.
[26] Rasheed, BA, Adnan, R., Saffari, SE, & dano Pati, K. (2014). Robust weighted least squares estimation of regression parameter in the presence of outliers and heteroscedastic errors. Jurnal Teknologi, 71(1).
[27] HAWGLEN, DC, MOSTELLER, F & TUKEY, JW (1983) Understanding Robust and Exploratory Data Analysis. Wiley, New York.
[28] France, Jiri"Robust regression: Robust Estimation of Regression Coefficient in Linear Regression Model when orthogonality condition.
[29] Andrews, DF, Bickel, PJ, Hampel, FR, Huper, PJ, Rogers, WH, and Tukey, JW (1972). “Robust estimates of location: Survey and advances.” Princeton University Press , Princeton ,NJ
[30] M. Dr.. Ibtisam Karim Abdullah, & Hussein Karim Nashour. (2019). Estimates of least squares and least squares trimmed for linear regression with skewed normal errors. Journal of Administration and Economics, 1(121), 377-388
[31] Zuo, Y. (2022). Asymptotic normality of the least sum of squares of trimmed residuals estimator. arXiv preprint arXiv:2204.00700.
[32] Zuo, Y. (2022). Least sum of squares of trimmed residuals regression. arXiv preprint arXiv: 2202.10329.
[33] Zuo, Y. (2022). New algorithms for computing the least trimmed squares estimator. arXiv preprint arXiv: 2203.10387.
[34] Denis H. and Yan L. (2005), "Cross-Validation in Nonparametric Regression with Outliers", The Annals of Statistics, Vol. 33, No. 5, pp.2291-2310.
[35] Luo, S., & Zhang, CY (2016). Nonparametric $$ M $$ M-type regression estimation under missing response data. Statistical Papers, 57(3), 641-664.

مقارنة المقدرات الحصينه واللامعلميه للانحدار اللامعلمي

علي فاضل عبد الجبار ، أفراح محمد كاظم
المعهد التقني كوت ، الجامعة التقنية الوسطى

الملخص

من أجل التخلص أو التقليل من القيم الشاذة لبعض الظواهر التي قد تكون سببًا في عدم الحصول على النتائج المرجوة. هذا يجعلنا نستخلص استنتاجات بعيدة عن الواقع للظاهرة التي ندرسها. أن المقدرات اللامعلمية التقليدية حساسة للغاية للقيم الشاذة ، مما دفعنا إلى استخدام المقدرات المحصنة لأنها لا تتأثر كثيرًا بالقيم الشاذة ، وكذلك الانحدار اللامعلمي لأنه لا يعتمد على المحددات أو الافتراضات السابقة ، ولكن يعتمد بشكل مباشر وأساسي على البيانات.