Neutrosophic Kernel Regression and Its Application

Ayman Orabi[1], Dalia Ziedan[2]

[1]Department of Management Information Systems Higher Institute for Specific Studies, Egypt
[2] Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

Abstract: The main purpose of utilizing neutrosophic statistics is to analyze data when there is uncertainty in the observations. The neutrosophic kernel regression is proposed in this paper for studying the relationship between the response variable and the study variable when the data is in neutrosophic numbers. The relationship between neutrosophic dietary fat level and neutrosophic death rate from prostate cancer has been studied using neutrosophic kernel regression.

Keywords: neutrosophic statistics; neutrosophic kernel regression**;** prostate cancer

(1) Introduction

The parametric technique is commonly used to illustrate the relationship between the independent variables and the dependent variable. Nevertheless, in some scenarios, the parametric model is inappropriate, and the resulting estimators do not outperform pure estimators in terms of efficiency. Kuo (1988) proposed a natural alternative for the distribution function that uses a nonparametric approach and does not put any constraints on the relationship between the independent variables and the dependent variable. Some noteworthy studies on this topic are Chambers et al. (1993), Drofman (1993), and Drofman and Hall (1993).

In sometimes, we cannot compute or supply exact values for statistical characteristics in real life, thus we must approximate them. As a result, neutrosophic statistics is an extension of classical statistics in which one works with set values rather than specific values. Smarandache (2014) is the first one introduced Neutrosophic Statistics as a generalization of classical statistics applied when the data under consideration is in neutrosophic numbers.

Many researchers are interested with neutrosophic statistics because it is more realistic in many problems. For example, Aslam and Albassam (2019) studied the statistical correlation between dietary fat level and prostate cancer risk. Alhabib and Salama (2020) presented a linear model for neutrosophic time series and used Student's distribution to determine the significance of its coefficient. Salem et al (2021) derived the mathematical properties of the neutrosophic lognormal distribution. Aslam and Saleem (2023) introduced the F-test of testing linearity under neutrosophic statistics. Alomair and Shahzad (2023) introduced the neutrosophic Hartley-Ross-type ratio estimators to estimate the population mean of neutrosophic data when the data is infected by the outliers.

Some of the more popular nonparametric regression methods are those based on kernel functions, spline functions and wavelets. Each of these methods has their own strengths and weaknesses but kernel functions have the advantage of mathematical simplicity. So, in this study, the neutrosophic kernel regression will be studied and applied to prostate cancer data. As far as we know, the neutrosophic structure of the kernel regression has never been addressed in previous studies.

(2) Preliminaries

Suppose that $T_N = [T_L, T_U]$ is a neutrosophic random variable, and the function $K(T_N)$ is called a neutrosophic kernel function, that has he following properties:

(i) $K(T_N) \geq 0$ and continous for all T_N

i.e. $K(T_L) \geq 0$ and $K(T_U) \geq 0$

(ii) $\int_{-\infty}^{\infty} K(T_N) dT_N = 1$,

i.e. $\int_{-\infty}^{\infty} K(T_L) dT_L = 1$ and $\int_{-\infty}^{\infty} K(T_U) dT_U = 1$

(iii) $K(-T_N) = K(T_N)$ for all T_N (K is a symmetric function about the origin)

i.e. $K(-T_L) = K(T_L)$ and $K(-T_U) = K(T_U)$.

The Gaussian kernel function is an example of the neutrosophic kernel function, and it has the form

$$K(T_N) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} T_N^2\right). \quad (1)$$

All proofs in this study are based on the algebraic framework of the neutrosophic numbers. [Smarandache (2014)].

Assuming two bounded real intervals $X_N = [X_L, X_U], and\ Y_N = [Y_L, Y_U]$ and Δ is the fundamental arithmetic operations, then

$$X_N \Delta Y_N = \left[\min(X_L \Delta Y_L, X_U \Delta Y_U), \max(X_L \Delta Y_L, X_U \Delta Y_U)\right].$$

Also, if we have D is a calculus operator (differentiation or integration) and f is any function, then

$$Df(X_N) = \left[\min\left(Df(X_L), Df(X_U)\right), \max\left(Df(X_L), Df(X_U)\right)\right].$$

(3) Neutrosophic Kernel Density Estimation

Theorem 1. Let $(x_{N1}, x_{N2}, ..., x_{Nn})$ be a neutrosophic sample selected from a univariate neutrosophic distribution with an unknown density f at any given interval x_N. Then, the neutrosophic kernel density estimation is

$$\hat{f}_h(x_N) = \frac{1}{n\,h_N}\sum_{i=1}^{n} K\left(\frac{x_N - x_{Ni}}{h_N}\right). \quad (2)$$

where $h_N > 0$ is a smoothing neutrosophic parameter called the neutrosophic bandwidth.

Proof. Using the properties of the neutrosophic kernel function, then we have

$$\frac{1}{n\,h_N}\sum_{i=1}^{n} K\left(\frac{x_N - x_{Ni}}{h_N}\right) \geq 0$$

Hence, the neutrosophic kernel density estimation more than or equal zero

i.e. $\hat{f}_h(x_N) \geq 0$ (3)

Also,

$$\int_{-\infty}^{\infty} \hat{f}_h(x_N)\,dx_N = \frac{1}{n\,h_N}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{x_N - x_{Ni}}{h_N}\right)dx_N$$

Let, $T_N = \dfrac{x_N - x_{Ni}}{h_N}$

Then, $J = \left|\dfrac{dx_N}{dT_N}\right| = h_N$

Therefore, $$\int_{-\infty}^{\infty} \hat{f}_h(x_N)\,dx_N = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(T_N)\,dT_N = 1 \quad (4)$$

From (3) and (4), theorem (1) is proved.

Theorem 2. Suppose that $(x_{N1}, y_{N1}), (x_{N2}, y_{N2}), ..., (x_{Nn}, y_{Nn})$ be a neutrosophic sample selected from a neutrosophic joint probability distribution with an unknown density $f(x_N, y_N)$ at any given pair of intervals (x_N, y_N). Then, the neutrosophic joint kernel density estimation of $f(x_N, y_N)$ is

$$\hat{f}_h(x_N, y_N) = \frac{1}{n\,h_N\,g_N}\sum_{i=1}^{n} K\left(\frac{x_N - x_{Ni}}{h_N}\right) K\left(\frac{y_N - y_{Ni}}{g_N}\right). \quad (5)$$

Proof. From the properties of the neutrosophic kernel function $K(T_N) \geq 0$ and $\int_{-\infty}^{\infty} K(T_N) dT_N = 1$, It is easy to prove that

$$K\left(\frac{x_N - x_{Ni}}{h_N}\right) \geq 0 \text{ and } K\left(\frac{y_N - y_{Ni}}{g_N}\right) \geq 0$$

Therefore, the neutrosophic joint kernel density estimation more than or equal zero

i.e. $\hat{f}(x_N, y_N) \geq 0$

(6)

Also,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{f}(x_N, y_N) dx_N dy_N = \frac{1}{n h_N g_N}\left[\sum_{i=1}^{n}\left(\int_{-\infty}^{\infty} K\left(\frac{x_N - x_{Ni}}{h_N}\right) dx_N\right)\left(\int_{-\infty}^{\infty} K\left(\frac{y_N - y_{Ni}}{g_N}\right) dy_N\right)\right]$$

(7)

Substituting $T_N = \frac{x_N - x_{Ni}}{h_N}$ and $S_N = \frac{y_N - y_{Ni}}{g_N}$ in (7) yielded

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{f}(x_N, y_N) dx_N dy_N = \frac{1}{n}\left[\sum_{i=1}^{n}\left(\int_{-\infty}^{\infty} K(T_N) dT_N\right)\left(\int_{-\infty}^{\infty} K(S_N) dS_N\right)\right] = 1$$

(8)

From (6) and (8), theorem (2) is proved.

(4) Neutrosophic Kernel Regression

Suppose that there is a neutrosophic set of data consists of *n* pairs of observations, $(x_{N1}, y_{N1}), (x_{N2}, y_{N2}), ..., (x_{Nn}, y_{Nn})$, and assume that there is a relationship between the independent variable X_N and the dependent variable Y_N. The relationship between X_N and Y_N can be represented by

$$Y_{Ni} = m(X_{Ni}) + \varepsilon_{Ni}, \qquad i = 1, ..., n.$$

(9)

The objective of neutrosophic kernel nonparametric regression is to estimate $m(x_N)$ using local averaging; the average will be formed in such a way that it is defined only in small neighborhoods around x_{Nj}.

Theorem 3. The estimation of $m(x_N)$ using neutrosophic kernel regression is

$$\hat{m}\left(x_{Nj}\right)=\sum_{i=1}^{n} w_{Nij} y_{Ni}$$

(10)

Where w_{Nij} is defined by the following:

$$w_{Nij}=\frac{K\left(\frac{x_{Nj}-x_{Ni}}{h_N}\right)}{\sum_{i=1}^{n} K\left(\frac{x_{Nj}-x_{Ni}}{h_N}\right)}$$

Proof:

$$\hat{m}\left(x_{Nj}\right)=E\left(y_N / x_{Nj}\right)=\int_{-\infty}^{\infty} \frac{y_N \hat{f}\left(x_{Nj}, y_N\right)}{\hat{f}\left(x_{Nj}\right)} dy_N$$

Using Theorem (1) and (2), then we have

$$\hat{m}\left(x_{Nj}\right)=\frac{\frac{1}{g_N}\sum_{i=1}^{n}\left[K\left(\frac{x_{Nj}-x_{Ni}}{h_N}\right)\left(\int_{-\infty}^{\infty} y_N K\left(\frac{y_N-y_{Ni}}{g_N}\right) dy_N\right)\right]}{\sum_{i=1}^{n} K\left(\frac{x_{Nj}-x_{Ni}}{h_N}\right)}$$

Let $S_N=\frac{y_N-y_{Ni}}{g_N}$, then the Jackobian is $J=\left|\frac{dy_N}{dS_N}\right|=g_N$ *and* $y_N=g_N S_N+y_{Ni}$,

and since $\int_{-\infty}^{\infty} K\left(S_N\right) dS_N=1$ *and* $\int_{-\infty}^{\infty} S_N K\left(S_N\right) dS_N=0$

Therefore,

$$\hat{m}\left(x_{Nj}\right)=\sum_{i=1}^{n}\frac{K\left(\frac{x_{Nj}-x_{Ni}}{h_N}\right) y_{Ni}}{\sum_{i=1}^{n} K\left(\frac{x_{Nj}-x_{Ni}}{h_N}\right)}=\sum_{i=1}^{n} w_{Nij} y_{Ni} \ .$$

(5) Real Application

Prostate cancer is one of the most frequent malignant illnesses in men and the second greatest cause of death from cancer (Van Booven et al, 2021). Recent studies have shown that an increased level of fat in the diet causes an increased incidence of prostate cancer. Dietary fat may be measured by measuring the fat in each ingredient in our daily food or by calculating the per percentage of calories in our daily food using glucose machines through blood. Measuring the diet using these approaches does not result in an exact number but can be recorded in an interval range. So, the relation between dietary level and prostate cancer death rate cannot be studied using

the traditional regression model when the variables are given as intervals. When the dietary fat level and prostate cancer death rate are in the interval range in an uncertain environment, neutrosophic statistics can be used for studying the relationship between the dietary fat level and prostate cancer death rate. In this study, we study the relationship between the predictor variable (the neutrosophic dietary fat level) and the response variable (the prostate cancer neutrosophic death rate) was studied using neutrosophic kernel regression. The dataset from Aslam and Albassam (2019) represented in table (1) shows the neutrosophic dietary fat level and prostate cancer neutrosophic death rate for 30 countries.

Table (1)

the neutrosophic dietary fat level and prostate cancer neutrosophic death rate for 30 countries.

Country No.	Diet Fat	D-Rate	Country No.	Diet Fat	D-Rate
1	[38,38]	[0.9,1.1]	16	[97,97]	[10.1,10.3]
2	[29,31]	[1.3,1.3]	17	[73,75]	[11.4,11.4]
3	[42,42]	[1.6,1.6]	18	[112,112]	[11.1,11.1]
4	[57,57]	[4.5,4.5]	19	[100,100]	[13.1,13.3]
5	[96,98]	[4.8,4.1]	20	[134,134]	[12.9,13.1]
6	[47,49]	[5.4,5.6]	21	[142,142]	[13.4,13.4]
7	[67,67]	[5.5,5.5]	22	[119,119]	[13.9,14.2]
8	[72,74]	[5.6,5.6]	23	[137,137]	[14.4,14.4]
9	[93,93]	[6.4,6.6]	24	[152,152]	[14.4,14.6]
10	[58,58]	[7.8,7.8]	25	[129,129]	[15.1,15.3]
11	[95,95]	[8.4,8.6]	26	[156,156]	[15.9,15.9]
12	[67,69]	[8.8,8.8]	27	[147,147]	[16.3,16.4]
13	[62,62]	[9,9]	28	[133,133]	[16.8,16.9]
14	[96,96]	[9.1,9.1]	29	[132,132]	[18.4,18.4]
15	[86,87]	[9.4,9.4]	30	[143,144]	[12.4,12.6]

In table (2), the estimation of response variable $\hat{Y}_{Nj} = \hat{m}\left(x_{Nj}\right)$ and the absolute neutrosophic error $\left|e_{Nj}\right| = \left|Y_{Nj} - \hat{Y}_{Nj}\right|$ were computed . Here the neutrosophic bandwidth $h_N = \left[h_L, h_U\right]$ is calculated as following

$$h_N = 0.9 \min\left(\hat{\sigma}_N, \frac{IQR_N}{1.35}\right). n^{-\frac{1}{5}}$$

(11)

Where $\hat{\sigma}_N$ and IQR_N are the estimated value of standard deviation and interquartile range for predictor variable, respectively and n is sample size. (see Silverman 1986)

From the above dataset, we found that

$\hat{\sigma}_N = [37.522, 37.785]$, $IQR_N = [67.75, 67.75]$, and $h_N = [17.118, 17.224]$

Table (2)

The $\hat{Y}_{Nj}$ and $|e_{Nj}|$ for prostate cancer neutrosophic dataset.

Country No.	$\hat{Y}_{Nj}$	$\|e_{Nj}\|$	Country No.	$\hat{Y}_{Nj}$	$\|e_{Nj}\|$
1	[4.007, 4.063]	[2.963,3.107]	16	[9.431,9.481]	[0.669,0.819]
2	[3.122,3.274]	[1.822,1.974]	17	[7.691,7.853]	[3.547,3.709]
3	[4.471, 4.537]	[2.871,2.937]	18	[11.644,11.658]	[0.544,0.558]
4	[6.232,6.3]	[1.732,1.8]	19	[9.769,9.808]	[3.331,3.492]
5	[9.329,9.584]	[4.529,5.484]	20	[14.419,14.514]	[1.414,1.519]
6	[5.146,5.312]	[0.254,0.288]	21	[14.678,14.777]	[1.278,1.377]
7	[7.213,7.239]	[1.713,1.739]	22	[12.866,12.902]	[1.034,1.298]
8	[7.627,7.76]	[2.027,2.16]	23	[14.546,14.644]	[0.146,0.244]
9	[9.054,9.117]	[2.517,2.654]	24	[14.791,14.883]	[0.283,0.391]
10	[6.339,6.405]	[1.395,1.461]	25	[14.093,14.175]	[1.007,1.125]
11	[9.233,9.29]	[0.69,0.833]	26	[14.816,14.904]	[0.996,1.084]
12	[7.239,7.384]	[1.416.1.561]	27	[14.749,14.846]	[1.551,1.554]
13	[6.749,6.801]	[2.199,2.251]	28	[14.367,14.46]	[2.433,2.44]
14	[9.329,9.383]	[0.229,0.283]	29	[14.308,14.399]	[4.001,4.092]
15	[8.533,8.671]	[0.729,0.867]	30	[14.695,14.81]	[2.21,2.295]

Also, the neutrosophic mean absolute error was computed as following $MAE_N = \frac{\sum |e_N|}{n} = [1.718, 1.847]$.

References

1. Alhabib, R., Salama, A. A. (2020). The neutrosophic time series-study its models (linear logarithmic) and test the coefficients significance of its linear model. Neutrosophic Sets and Systems. 33: 105-115. DOI:10.5281/zenodo.3782858
2. Alomair, A.M., Shahzad, U. (2023). Neutrosophic Mean Estimation of Sensitive and Non-Sensitive Variables with Robust Hartley–Ross-Type Estimators. *Axioms*, *12*(6): 578-592. https://doi.org/10.3390/axioms12060578
3. Aslam, M., Albassam, M. (2019). Application of neutrosophic logic to evaluate correlation between prostate cancer mortality and dietary fat assumption. *Symmetry*. *11*(3): 330-336. https://doi.org/10.3390/sym11030330
4. Aslam, M., Saleem, M. (2023). Neutrosophic test of linearity with application. AIMS Mathematics. 8(4): 7981-7989. doi: 10.3934/math.2023402

5. Chambers, R. L., Drofman, A. H., Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. J Am Stat Assoc. 88: 268–277. https://doi.org/10.2307/2290722
6. Dorfman, A. H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. Australian Journal of Statistics. 35(1): 29–41. https://doi.org/10.1111/j.1467842X.1993.tb01310.x
7. Dorfman, A. H., Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. The Annals of Statistics. 16(3): 1452–1475. https://doi.org/10.1214/aos/1176349267
8. Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data, In Proceedings of the section on survey research methods. Amer. Statist. Assoc., Alexandria, VA, 280-285.
9. Salem, S., Khan, Z., Ayed, H., Brahmia, A, Amin, A. (2021). The neutrosophic lognormal model in lifetime data analysis: properties .and applications. Journal of Function Spaces. 202:1–9. https://doi.org/10.1155/2021/6337759
10. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability. Vol. 26, Chapman and Hall, London. https://doi.org/10.1201/9781315140919
11. Smarandache, F. (2014). Introduction to Neutrosophic Statistics; Sitech & Education Publishing: Craiova, Romania; Columbus, OH, USA.
12. Van Booven, DJ., Kuchakulla, M., Pai, R., Frech, FS., Ramasahayam, R., Reddy, P., Parmar M. Et al. (2021). Systematic Review of Artificial Intelligence in Prostate Cancer. Research and Report in Urology, 13: 31-39. https://doi.org/10.1016/j.ifacol.2022.06.012