Journal of Physical Sciences, Vol. 30, 2025, 55-62
ISSN: 2350-0352 (print), www.vidyasagar.ac.in/publication/journal
Published on 31 December 2025
DOI: http://doi.org/10.62424/jps.2025.30.00.06

Neutrosophic Correlation and Regression with Applications

Provat Ghosh[1*], Jayanta Sarkar[2] and Purbasa Giri[3]

Department of Applied Mathematics, Vidyasagar University
Midnapore-721102, India
Email: [1]provatghosh1996@gmail.com; [2]jsarkarmath@gmail.com;
[3]giripurbasa@gmail.com. [*]Correspnding author.

ABSTRACT

Classical statistical techniques assume that observations are precise and unambiguous. However, real-world data are often affected by uncertainty, vagueness, and incomplete information. Neutrosophic statistics, based on neutrosophic numbers comprising deterministic and indeterminate components, provides a realistic mathematical framework for addressing such situations. In this paper, we develop rigorous formalisms for neutrosophic correlation and regression using theorem–proof structures. Explicit expressions for neutrosophic mean, variance, covariance, correlation coefficient, and regression models are derived. A comprehensive numerical case study of the relationship between working time and income is presented, with all intermediate computations and final results reported in detail. The proposed approach demonstrates clear advantages over classical regression by naturally incorporating uncertainty in both explanatory and response variables.

Keywords: Neutrosophic number, Neutrosophic variance, Neutrosophic covariance, Correlation coefficient, Neutrosophic regression

AMS Mathematics Subject Classification (2010): 62J05, 62H20, 03E72, 62P99

Abstract in Bengali

প্রচলিত পরিসংখ্যানগত পদ্ধতিগুলি পর্যবেক্ষণসমূহকে নির্ভুল ও দ্ব্যর্থহীন বলে ধরে নেয়। তবে বাস্তব জীবনের তথ্য প্রায়ই অনিশ্চয়তা, অস্পষ্টতা এবং অসম্পূর্ণ তথ্য দ্বারা প্রভাবিত হয়। নির্ধারিত (deterministic) ও অনির্ধারিত (indeterminate) উপাদান নিয়ে গঠিত নিউট্রোসফিক সংখ্যার উপর ভিত্তি করে নিউট্রোসফিক পরিসংখ্যান এই ধরনের পরিস্থিতি মোকাবিলার জন্য একটি বাস্তবসম্মত গাণিতিক কাঠামো প্রদান করে। এই প্রবন্ধে উপপাদ্য–প্রমাণ কাঠামো ব্যবহার করে নিউট্রোসফিক সহসম্বন্ধ (correlation) ও রিগ্রেশন বিশ্লেষণের কঠোর আনুষ্ঠানিক বিকাশ উপস্থাপন করা হয়েছে। নিউট্রোসফিক গড়, বিচ্যুতি, সহ-বিচ্যুতি, সহসম্বন্ধ সহগ এবং রিগ্রেশন মডেলের সুস্পষ্ট সূত্র নির্ণয় করা হয়েছে। কাজের সময় ও আয়ের মধ্যে সম্পর্কের উপর একটি বিস্তৃত সংখ্যাতাত্ত্বিক ক্ষেত্রে অধ্যয়ন উপস্থাপন করা হয়েছে, যেখানে সমস্ত মধ্যবর্তী গণনা ও চূড়ান্ত ফলাফল বিস্তারিতভাবে প্রদর্শিত হয়েছে। প্রস্তাবিত পদ্ধতিটি ব্যাখ্যাকারী ও প্রতিক্রিয়াশীল উভয় চলকে

স্বাভাবিকভাবে অনিশ্চয়তা অন্তর্ভুক্ত করার মাধ্যমে প্রচলিত রিগ্রেশনের তুলনায় সুস্পষ্ট সুবিধা প্রদর্শন করে।

1. Introduction

Statistical regression analysis has long been a central tool for modelling and quantifying relationships between variables in diverse fields such as economics, engineering, agriculture, and the social sciences. In the classical statistical framework**,** regression methods are based on crisp numerical observations and deterministic assumptions. The least squares principle, developed by Gauss, provides the mathematical foundation for classical regression analysis and has been widely adopted due to its simplicity and strong theoretical properties [1]. However, classical regression models implicitly assume that data are precise and free from ambiguity. In practice, this assumption is often violated, as real-world data are influenced by various sources of uncertainty.

In economic systems, fluctuations in markets and income levels introduce variability; in agriculture, environmental conditions such as rainfall, temperature, and soil quality affect crop yield; in engineering applications, measurement errors and material imperfections are unavoidable; and in social sciences, human behaviour is inherently imprecise. These factors lead to vagueness, incompleteness, and indeterminacy in data that cannot be adequately represented using real numbers alone. Consequently, the applicability of classical regression analysis becomes limited when uncertainty plays a dominant role.

To overcome these limitations, fuzzy regression analysis emerged as an extension of classical regression following the introduction of fuzzy set theory by Zadeh [2]. Tanaka et al. proposed the first fuzzy linear regression model [3], in which regression coefficients are represented as fuzzy numbers rather than crisp values. This approach allows uncertainty and vagueness to be incorporated directly into the regression structure. Later, Diamond developed fuzzy least-squares techniques [4] to improve parameter estimation in fuzzy environments. Fuzzy regression models have since been applied successfully in decision-making, forecasting, control systems, and risk analysis, where precise data are rarely available.

Despite their advantages, fuzzy regression models primarily address vagueness and do not explicitly capture indeterminacy arising from incomplete or conflicting information. To address this gap, neutrosophic statistics was introduced as a further generalization of classical and fuzzy statistics. Based on the neutrosophic set theory proposed by Smarandache [5], neutrosophic statistics incorporates an explicit indeterminate component alongside deterministic information. Neutrosophic numbers allow simultaneous representation of known values and associated uncertainty, providing a more realistic mathematical framework for complex data.

The present paper systematically develops neutrosophic correlation and regression and illustrates their applicability through a fully worked numerical example. The progression from classical to fuzzy and neutrosophic regression reflects the growing demand for robust statistical tools capable of handling uncertainty, vagueness, and indeterminacy in real-world data analysis.

2. Neutrosophic numbers

A neutrosophic number is defined as

$$N = a + bI, \tag{1}$$

where $a, b \in \mathbb{R}$ and $I \subset [0,1]$ denotes indeterminacy satisfying

$$I^2 = I, \qquad I \cdot 0 = 0, \qquad I \neq 1.$$

Here, a represents the deterministic (observed) part, and bI represents the indeterminate (uncertain) part. If $a = 0$, then N is called a pure neutrosophic number.

Real-life examples of neutrosophic numbers

Typical real-life quantities are often affected by uncertainty due to incomplete information, measurement errors, environmental factors, human judgment or many others. Such situations can be effectively modelled using neutrosophic numbers**,** which consist of a determinate part and an indeterminate component. For example, monthly income can be represented as (50,000 + 12,000I), where the fixed salary constitutes the determinate part and the indeterminate part reflects bonuses, overtime payments, or irregular earnings. Similarly, household expenditure**,** expressed as (30,000 + 8,000I), accounts for routine expenses and uncertain costs such as medical bills and emergency spending. In agriculture, crop yield varies due to weather conditions, soil quality, and pest attacks, and can be modelled as (40 + 15I) quintals per hectare. Likewise, the project completion time, given by (180 + 40I) days, captures the planned schedule along with possible delays caused by resource shortages or unforeseen circumstances. These representations provide a more realistic framework for decision-making under uncertainty.

3. Algebra of neutrosophic numbers

Let $N_1 = a_1 + b_1 I$ and $N_2 = a_2 + b_2 I$. Then

$$N_1 + N_2 = (a_1 + a_2) + (b_1 + b_2)I, \tag{2}$$

$$N_1 - N_2 = (a_1 - a_2) + (b_1 - b_2)I, \tag{3}$$

$$N_1 N_2 = a_1 a_2 + [(a_1 + b_1)(a_2 + b_2) - a_1 a_2]I. \tag{4}$$

(Division) If $N_2 = a_2 + b_2 I$ with $a_2 \neq 0$ and $a_2 + b_2 \neq 0$, then

$$\frac{N_1}{N_2} = \frac{a_1}{a_2} + \frac{b_1 a_2 - a_1 b_2}{a_2(a_2 + b_2)} I. \tag{5}$$

4. Neutrosophic statistical measures

Neutrosophic mean: For $N_i = a_i + b_i I,\ i = 1,2,\dots,n,$

$$\overline{N} = \overline{a} + \overline{b}I,$$

where $\overline{a} = \frac{1}{n}\sum a_i$ and $\overline{b} = \frac{1}{n}\sum b_i$.

Lemma 1. (Neutrosophic variance)

Let $N_i = x_i + y_i I$. Then

$$\mathrm{Var}(N) = \mathrm{Var}(x) + [\mathrm{Var}(y) + 2\mathrm{Cov}(x, y)]I.$$

Proof.

$$\begin{aligned} \mathrm{Var}(N) &= \frac{1}{n}\sum (N_i - \overline{N})^2 \\ &= \frac{1}{n}\sum [(x_i - \overline{x}) + (y_i - \overline{y})I]^2 \\ &= \mathrm{Var}(x) + \mathrm{Var}(y)I + 2\mathrm{Cov}(x, y)I. \end{aligned}$$

Lemma 2. (Neutrosophic covariance)

Let $u_i = x_i + y_i I$ and $v_i = p_i + q_i I$ be a bivariate sample of size *n*. Then

$$\mathrm{Cov}(u, v) = \mathrm{Cov}(x, p) + [\mathrm{Cov}(x, q) + \mathrm{Cov}(y, p) + \mathrm{Cov}(y, q)]I.$$

Proof. $\mathrm{Cov}(u,v) = \frac{1}{n}\sum(u-\bar{u})(v-\bar{v})$

$$= \frac{1}{n}\sum\{(x_i + y_i I) - (\bar{x} + \bar{y}I)\}\{(p_i + q_i I) - (\bar{p} + \bar{q}I)\}$$

$$= \frac{1}{n}\sum\{(x_i - \bar{x}) + (y_i - \bar{y})I\}\{(p_i - \bar{p}) + (q_i - \bar{q})I\}$$

$$= \frac{1}{n}\sum\{(x_i - \bar{x})(p_i - \bar{p}) + (x_i - \bar{x})(q_i - \bar{q})I + (y_i - \bar{y})(p_i - \bar{p})I + (y_i - \bar{y})(q_i - \bar{q})I\}$$

$$= \mathrm{Cov}(x,p) + \mathrm{Cov}(x,q) + \mathrm{Cov}(y,p) + \mathrm{Cov}(y,q).$$

Note that if there is no uncertainty in both u_i and v_i, then the covariance between *u* and *v* is equal to the covariance between *x* and *p*.

4.1. Neutrosophic correlation and regression

The neutrosophic correlation coefficient is defined as

$$r = \frac{\mathrm{Cov}(u,v)}{\sigma_u \sigma_v},$$

where $\sigma_u = \sqrt{\mathrm{Var}(u)}$ and $\sigma_v = \sqrt{\mathrm{Var}(v)}$.

A neutrosophic regression model of Y on T is given by

$$Y = \alpha + \beta T,$$

where α and β are neutrosophic coefficients. These two parameters can be determined based on the given sample values.

5. Numerical results and case study

To illustrate the proposed regression model within a neutrosophic framework, we consider a dataset comprising 10 families. The dataset contains these families' monthly income and the time spent working each month. In practical situations, a family's working time and income are not always fixed or precisely known. Apart from regular, fixed work, families may engage in additional, non-specific, irregular, or uncertain activities that contribute variably to both working time and income.

Accordingly, in the given dataset, fixed working time and fixed income are recorded along with additional indeterminate components representing other uncertain or non-specific work and the corresponding income. These indeterminate components may arise due to overtime work, temporary jobs, seasonal employment, or informal economic activities. Such uncertainty cannot be adequately modelled using classical regression techniques, which assume precise numerical observations.

By representing both working time and income in neutrosophic form, the dataset captures the deterministic and indeterminate aspects of the variables simultaneously. This makes the data particularly well-suited for demonstrating the effectiveness of the proposed neutrosophic regression model, which can incorporate indeterminacy directly into the regression structure. The numerical analysis based on this dataset highlights that neutrosophic regression provides a more realistic and flexible modelling approach than classical regression when dealing with uncertain socio-economic data.

The time represented as $T = a_T + b_T I$ and income represented as $Y = a_Y + b_Y I$, where aT, bT represent the fixed time spent in a month, and the uncertain amount of time spent for a month. Similarly, a_Y and b_Y represent the fixed monthly income and the extra income obtained by spending bT time.

Table 1: Neutrosophic data for working time and income

Obs.	$T = a_T + b_T I$ (hours/month)	$Y = a_Y + b_Y I$ (Rs. in K/month)
1	$160 + 10I$	$40 + 8I$
2	$170 + 12I$	$42 + 8.5I$
3	$180 + 15I$	$45 + 9I$
4	$190 + 15I$	$47 + 10I$
5	$200 + 18I$	$50 + 11I$
6	$210 + 20I$	$52.5 + 12I$
7	$220 + 22I$	$55 + 13I$
8	$230 + 25I$	$57.5 + 14I$
9	$240 + 25I$	$60 + 15I$
10	$250 + 28I$	$62.5 + 16I$

The statistics for this table of data are as follows.

$$\overline{a_T} = 205,\ \overline{b_T} = 19,\ \overline{a_Y} = 51.15,\ \overline{b_Y} = 11.65.$$

$$\overline{T} = 205 + 19I,\ \overline{Y} = 51.15 + 11.65I.$$

This data shows that the average time spent on the work is 205 hours (fixed), and 19 hours may be spent for extra income. The minimum fixed average income for this group of families is Rs. 51.15 thousand, and the uncertain income is Rs. 11.65 thousand.

The variances of the crisp data for A_T, B_T, A_Y and B_Y are determined as

$Var(A_T) = 825.00,$ $Var(B_T) = 32.60,$
$Var(A_Y) = 52.8525,$ $Var(B_Y) = 7.1025.$

And the covariances are

$\mathrm{Cov}(A_T, B_T) = 163.00,$ $\mathrm{Cov}(A_T, A_Y) = 208.75,$ $\mathrm{Cov}(A_T, B_Y) = 76.25,$
$\mathrm{Cov}(B_T, A_Y) = 41.30,$ $\mathrm{Cov}(B_T, B_Y) = 15.05,$ $\mathrm{Cov}(A_Y, B_Y) = 19.3025.$

The variance of neutrosophic sample $N_i = x_i + y_i I, i = 1, 2, \dots, n$ is

$$\mathrm{Var}(N) = \mathrm{Var}(x) + [\mathrm{Var}(y) + 2\mathrm{Cov}(x, y)]I.$$

Then

$$\mathrm{Var}(T) = 825 + (32.60 + 2 \times 163)I = 825 + 358.60\ I,$$
$$\mathrm{Var}(Y) = 52.8525 + (7.1025 + 2 \times 19.3025)I = 52.8525 + 45.7075\ I.$$

For $u = x + yI$ and $v = p + qI$,

$$\mathrm{Cov}(u, v) = \mathrm{Cov}(x, p) + [\mathrm{Cov}(x, q) + \mathrm{Cov}(y, p) + \mathrm{Cov}(y, q)]I.$$

Thus, the covariance between time spent T and income Y is

$$\mathrm{Cov}(T, Y) = \mathrm{Cov}(A_T, A_Y) + [\mathrm{Cov}(A_T, B_Y) + \mathrm{Cov}(B_T, A_Y) + \mathrm{Cov}(B_T, B_Y)]I$$
$$= 208.75 + [76.25 + 41.30 + 15.05]I = 208.75 + 132.6I.$$

5.1. Correlation coefficients

The correlation coefficient r is determined by the conventional formula

$$r = \frac{\mathrm{Cov}(u, v)}{\sigma_u \sigma_v}.$$

where $\sigma_u = \sqrt{\mathrm{Var}(u)}$ and $\sigma_v = \sqrt{\mathrm{Var}(v)}$.

Note that the computation of the square root of a neutrosophic number is not so easy. It provides four possible values, unlike crisp numbers, which provide only two. The value of $\sigma_u \sigma_v$ can be computed in two different ways. By separately finding the values of σ_u and σ_v, we can compute their product. It makes it more complicated. Alternatively, we can compute the variances of *u* and *v*, then multiply them and take the square root. This process reduces the confusion.

Let $p = \mathrm{Var}(T)\mathrm{Var}(Y)$ $= (825 + 358.60I)(52.8525 + 45.7075I)$
$= 43603.3125 + 73052.3035\mathrm{I}$

Square roots are:

$$z_1 = 208.814062 + 122.71\,I, \qquad z_2 = 208.814062 - 540.33\,I,$$
$$z_3 = -208.814062 + 540.33\,I, \qquad z_4 = -208.814062 - 122.71$$

The principal values are

$z_1 = 208.814062 + 122.71\,I$ and $z_3 = -208.814062 + 540.33\,I$.

The four correlation coefficients are,

$$r_1 = 0.9997 + 0.0300\,I, \qquad r_2 = 0.9997 - 2.0300\,I,$$
$$r_3 = -0.9997 + 2.0300\,I, \qquad r_4 = -0.9997 - 0.0300\,I.$$

Note that the deterministic parts of all correlation coefficients are the same in terms of magnitude. In this particular application, if the time spent on other work increases, the income will increase, and vice versa. So, the correlation coefficient must be positive. Thus, only two correlation coefficients are valid for this application. These are

$$r_1 = 0.9997 + 0.0300\,I, \qquad r_2 = 0.9997 - 2.0300\,I.$$

The deterministic correlation coefficient is 0.9997. This shows that time spent and income are highly correlated, and this is obvious.

6.2. Regression equations

Regression equations are very important in regression analysis and are used for prediction.

The regression equation for income on time spent is

$Y - \bar{Y} = b_{YT}(T - \bar{T})$, where

$$b_{YT} = r\frac{\sigma_Y}{\sigma_T} = \frac{\mathrm{Cov}(Y,T)}{\sigma_Y\sigma_T}\ \ \frac{\sigma_Y}{\sigma_T} = \frac{\mathrm{Cov}(Y,T)}{\sigma_T^2} = \frac{\mathrm{Cov}(Y,T)}{\mathrm{Var}(T)}.$$

This formula gives a single value for b_{YT}. But if we evaluate this expression for different values of r, we get two distinct values of b_{YT}.

Thus $b_{YT} = \frac{208.75+132.6I}{825+358.60I} = 0.2530 + 0.0354\,I$.

Thus, the regression equation of income on time spent is

$$Y - \bar{Y} = b_{YT}(T - \bar{T})$$

That is,

$$Y = b_{YT}T + (\bar{Y} - b_{YT}\bar{T})$$
$$= (0.2530 + 0.0354\,I)T - (0.715 + 1.0866\mathrm{I}).$$

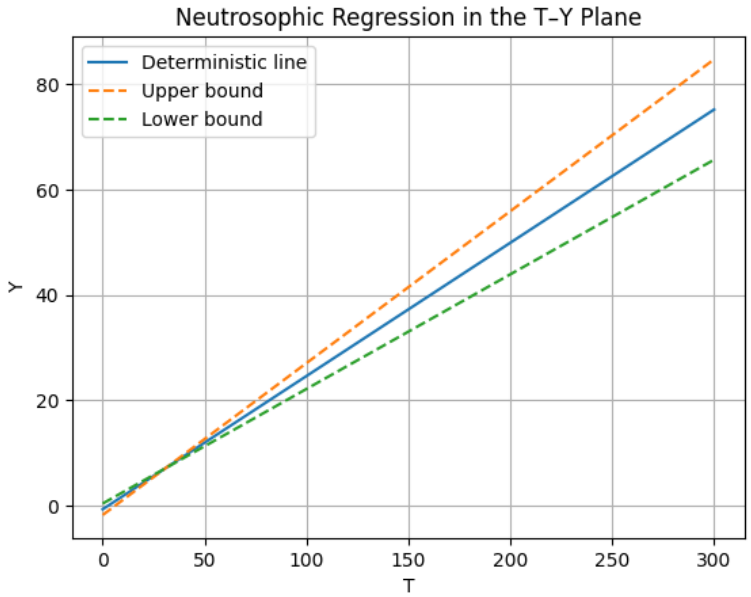

Figure 1: Neutrosophic regression line with indeterminacy band
Deterministic line Y = 0.2530T - 0.715
Upper bound: Y = (0.2530+0.0354)T - (0.715+1.0866),
Lower bound: Y = (0.2530-0.0354)T - (0.715-1.0866).

Prediction

For $T = 110 + 50I$,

$$Y = 27.115 + 17.2274\, I.$$

6. Conclusion

This paper demonstrates that neutrosophic correlation and regression provide a mathematically rigorous and realistic framework for statistical analysis under uncertainty. Unlike classical regression, the proposed approach captures uncertainty in both independent and dependent variables, yielding an indeterminacy band rather than a single prediction line.

REFERENCES

1. C.F. Gauss, Theoria combinationis observationum erroribus minimis obnoxiae, Göttingen (1823).
2. L.A. Zadeh, Fuzzy sets, Information and Control, 8 (1965) 338–353.
3. H. Tanaka, S. Uejima and K. Asai, Linear regression analysis with fuzzy model, IEEE Transactions on Systems, Man, and Cybernetics, 12 (1982) 903–907.
4. P. Diamond, Fuzzy least squares, Information Sciences, 46 (1988) 141–157.
5. F. Smarandache, Neutrosophic set—A generalization of the intuitionistic fuzzy set, International Journal of Pure and Applied Mathematics, 24(3) (2005) 287–297.
6. M. Pal, Recent Developments of Fuzzy Matrix Theory and Applications. Springer, 2024.
7. M. Pal, Neutrosophic matrix and neutrosophic fuzzy matrix. In: Recent Developments of Fuzzy Matrix Theory and Applications. Springer, 2024.