Plithogenic Logic and Computation

Journal Homepage: sciencesforce.com/plc

SCIENCES FORCE
International Academic Publisher

Plithogenic Log. Comp. Vol. 1 (2024) 120–126

Paper Type: Original Article

Integrating Neutrosophic Numbers in Regression Analysis for Enhancing Predictive Modelling through Uncertainty Representation

Maikel Leyva Vázquez [1],* **and Lorenzo Cevallos Torres [1]**

[1] Universidad de Guayaquil, Guayas, Ecuador. Emails: maikel.leyvav@ug.edu.ec, lorenzo.cevallost@ug.edu.ec.

Received: 06 Jan 2024 **Revised:** 08 Apr 2024 **Accepted:** 09 May 2024 **Published:** 12 May 2024

Abstract

This article explores the integration of neutrosophic statistics into regression analysis to bolster predictive modeling. Neutrosophic statistics, an extension of interval statistics, offer a robust framework for handling uncertainty and indeterminacy in datasets. By augmenting traditional interval predictions with measures of indeterminacy, neutrosophic numbers provide a nuanced representation of uncertainty, empowering decision-makers with a comprehensive view of potential outcomes. The methodology emphasizes the iterative refinement of predictive models to adapt to evolving data dynamics and escalating uncertainties. Future research directions include investigating the impact of different neutrosophic aggregation techniques on model performance, exploring synergies with other machine learning paradigms, and extending the applicability of neutrosophic statistics to diverse domains beyond regression analysis. By addressing these avenues, this study aims to advance the frontier of predictive modeling and facilitate more informed decision-making in complex and uncertain environments.

Keywords: Neutrosophic Statistics, Regression Analysis, Predictive Modeling, Uncertainty, Indeterminacy.

1 | Introduction

Neutrosophic statistics, an extension of interval statistics, provides a robust framework for handling various indeterminacies in statistical analysis, which makes it a powerful tool for predicting intervals in machine learning [1]. This form of statistics is particularly useful when data, inferential procedures, or probability distributions are not precisely defined but are instead characterized by uncertainty or imprecision. As machine learning increasingly permeates various domains, the ability to predict intervals rather than single-point estimates becomes crucial, especially in decision-making contexts where understanding the reliability and uncertainty of predictions is paramount [2].

In regression analysis, representing predictions as prediction intervals provides a more comprehensive view of the uncertainty associated with the predictions. A prediction interval offers a range within which we expect the true value of the dependent variable to fall with a certain probability, typically 95% or 99%. This is particularly useful because it accounts for the variability in the data that might not be captured by the prediction alone. To calculate a prediction interval, one must consider both the uncertainty in the estimate of the regression model and the inherent variability of the data. The interval is constructed around the predicted value and is usually symmetric, extending a certain amount above and below the predicted value [3].

Corresponding Author: maikel.leyvav@ug.edu.ec

https://doi.org/10.61356/j.plc.2024.1258

Utilizing prediction intervals in regression analysis is beneficial because they offer a realistic spectrum of possible outcomes, which aids in the decision-making process. This acknowledges that a single predicted value is not absolute but rather a likely scenario within a range of potential outcomes. To further refine this model, neutrosophic statistics can be applied, which excel at managing the ambiguity and indeterminacy of data. By converting the interval into a neutrosophic number [4], the traditional interval is enhanced to include an indeterminacy component. This addition captures the uncertainty and imprecision that are typically present in real-world data, offering a more nuanced understanding of the data's variability. This article explores the integration of neutrosophic statistics into regression analysis, demonstrating its efficacy in handling uncertainty and improving the reliability of predictive models.

2 | Preliminaries

Machine learning (ML) involves mathematical formulations to create models that can learn from data to make predictions or decisions without being explicitly programmed to perform those tasks. Interval prediction in machine learning refers to the technique of predicting a range of possible outcomes for a given input rather than a single-point estimate. By providing intervals, these methods offer not just predictions but also an insight into the reliability and uncertainty of the predictions, which is crucial for decision-making in uncertain environments [5].

For a dataset with independent variables $X = [x_1, x_2, \dots, x_n]$ and a dependent variable y, the objective of regression analysis is to model the relationship between X and y accurately. This relationship is mathematically expressed as [6]:

$$y \approx f(X;\ \theta) \tag{1}$$

where :

y is the dependent variable or the target that is to be predicted.

X represents the independent or explanatory variables that are used to predict

f is the regression function, which may vary in form depending on the type of regression model used (linear, polynomial, logistic, etc.).

θ are the parameters or coefficients of the model, adjusted during the training process to minimize a loss function, typically the Mean Squared Error (MSE) in regression [7].

In regression analysis, representing predictions as prediction intervals provides a more comprehensive view of the uncertainty associated with the predictions. A prediction interval offers a range within which we expect the true value of the dependent variable to fall with a certain probability, typically 95% or 99%. This is particularly useful because it accounts for the variability in the data that might not be captured by the prediction alone[8].

To calculate a prediction interval, one must consider both the uncertainty in the estimate of the regression model and the inherent variability of the data. The interval is constructed around the predicted value and is usually symmetric, extending a certain amount above and below the predicted value. This range is determined based on the standard error of the prediction and the residual standard deviation, which reflects the spread of the residuals or errors of the model [9].

For example, in a simple linear regression, the prediction interval for a new observation is given by [10]:

$$\hat{y}_0 \pm t_{\propto\ /2,n-2} \cdot SE \tag{2}$$

Where $\hat{y}_0$ is the predicted value of *y* from the t-distribution for a specified confidence level $\propto$ and $n-2$ degrees of freedom, and SE.

Utilizing prediction intervals in regression analysis is beneficial because they offer a realistic spectrum of possible outcomes, which aids in the decision-making process. This acknowledges that a single predicted value is not absolute but rather a likely scenario within a range of potential outcomes. This method of forecasting effectively incorporates the inherent uncertainties associated with future predictions, providing a more accurate depiction of what to expect. To further refine this model, neutrosophic statistics can be applied, which excel at managing the ambiguity and indeterminacy of data. By converting the interval into a neutrosophic number, the traditional interval is enhanced to include an indeterminacy component. This addition captures the uncertainty and imprecision that are typically present in real-world data, offering a more nuanced understanding of the data's variability. The neutrosophic treatment of the interval is as follows [11]:

$$\hat{y}_0 - t_{\propto/\ 2,n-2} \cdot SE + (\hat{y}_0 + t_{\propto\ /2,n-2} \cdot SE)I \qquad (3)$$

Here, I_N represents the indeterminacy factor associated with the prediction, where $I_N \in [I_l, I_u]$, This notation introduces the bounds of indeterminacy. I_l (lower indeterminacy) and I_u (upper indeterminacy), which defines the range of possible deviations due to uncertain elements affecting the forecast [12,13].

3 | Material and Methods

The Diabetes Dataset is a classic dataset used for regression analysis. Originating from the Diabetes Research Institute in Scania, Sweden, this dataset includes medical data from 442 patients. The target variable is a quantitative measure of diabetes progression one year after the baseline [14] as shown in Figure 1.

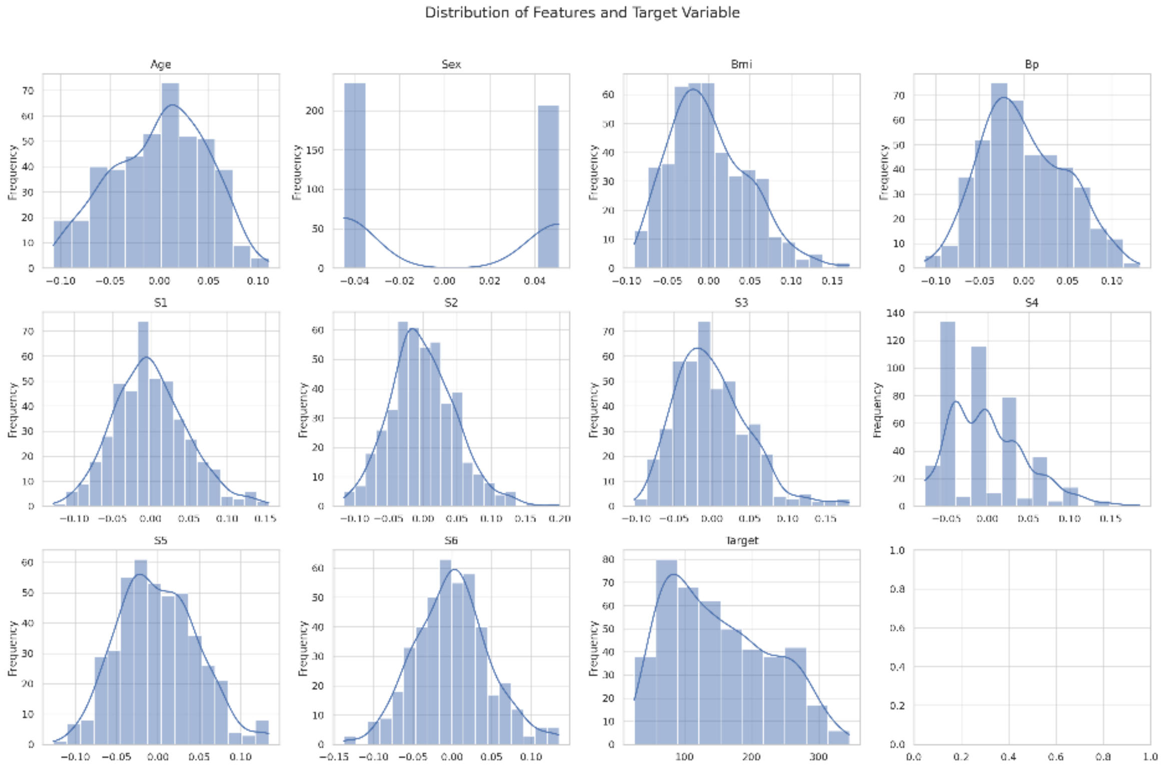

Figure 1. Visual analysis of variable distributions in diabetes research data.

The process of analyzing a dataset for regression and using neutrosophic numbers to represent uncertainty can be broken down into several key steps:

(i) Data Partitioning: The first step is to divide the data into training and testing sets. This split is crucial as it allows for the validation of the model on unseen data, ensuring the model's performance is not just a result of overfitting the training data. Typically, data scientists might use a 70-30 or 80-20 split where 70% or 80% of the data is used for training, and the remaining is for testing.

(ii) Training the Models: Each model is then trained on the training set. This involves adjusting the model parameters to best fit the data. Common regression models include Linear Regression [15], Ridge Regression [16], and Random Forest [17], among others. The training process involves finding the model parameters that minimize a loss function, essentially capturing the underlying pattern of the dataset.

(iii) Estimation of Prediction Intervals: After the models are trained, the next step is to estimate prediction intervals for new observations. This is where neutrosophic numbers come into play. Unlike traditional crisp intervals, neutrosophic intervals include measurements of truth, indeterminacy, and falsity, allowing for a more nuanced representation of uncertainty in predictions. Each model may require different methods to calculate these intervals, considering the model's specific characteristics and the data's nature.

(iv) Calculation and Analysis of Uncertainty Through Neutrosophic Numbers: The final step involves a detailed analysis of the uncertainty represented by the neutrosophic numbers. This includes evaluating how the indeterminate component of these numbers varies with different models and what it suggests about the data's complexity or variability. For example, a higher indeterminacy might indicate more significant external influences or inherent unpredictability in the dataset.

In this case, we employ neutrosophic means to combine interval predictions with other methods as part of a fusion approach in regression analysis. Neutrosophic means are particularly useful for integrating different predictive models because they allow for the incorporation of uncertainty, indeterminacy, and conflicting information which typically arise from diverse data sources or model outputs. This approach enhances the robustness and reliability of the predictive models by providing a more comprehensive framework that accounts for various aspects of uncertainty.

The neutrosophic mean is denoted as X_n, is calculated by considering the neutrosophic inclusion $I_N$that belongs to the interval $[I_l, I_u]$. This mean consists of two main elements: X_l, which is the mean of the lower part of the neutrosophic samples, and X_u, which is the mean of the upper part. The respective definitions are [18]:

$$X_l = \frac{\sum_{i=1}^{n_l} x_{il}1}{n_l} \tag{4}$$

$$X_u = \frac{\sum_{i=1}^{n_u} x_{iu}}{n_u} \tag{5}$$

where n_l and $n_u$represent the number of elements in the lower and upper parts of the neutrosophic samples, respectively. Therefore, the neutrosophic mean X_n, is expressed as the sum of X_l and X_u, adjusted by the interval of indetermination I_n:

$$X_N = X_l + X_u I_N;\ I_N \in [I_l, I_u] \tag{6}$$

$I_{l,}$=0, and I_u

$$I_u = \frac{X_u - X_l}{X_u} \tag{7}$$

4 | Results

The selected test case for prediction has the following standardized characteristics:

- Age (age): 0.045341
- Sex (sex): -0.044642 (indicates the female gender if we follow the standard encoding)
- Body Mass Index (bmi): -0.006206

- Mean blood pressure (bp): -0.015999
- S1 (blood - total cholesterol measurement): 0.125019
- S2 (blood - low-density lipoprotein measurement): 0.125198
- S3 (blood - high-density lipoprotein measurement): 0.019187
- S4 (blood - thyroid medication measurement): 0.034309
- S5 (blood - lamotrigine measurement): 0.032432
- S6 (blood-glucose level measurement): -0.00522

These values are standardized, meaning they are different from the average (expressed in number of standard deviations) in a reference population. For example, a value of 0.045 for age means that the subject's age is slightly above average.

Table 1. Prediction intervals and neutrosophic forms for Time Series models.

Model	Prediction Interval [Lower Bound, Upper Bound]	Neutrosophic form
Linear Regression	[39.11, 247.05]	39.11+247.05I; I∈[0,0.842]
Ridge Regression	[39.65, 252.67]	39.65+252.67I; I∈[0,0.843]
Random Forest	[67.00, 229.53]	67.00+229.53I; I∈[0,0.708]
Mean	[48.59,243.08]	N=48.59+243.08I;*I* ∈ [0,0,0.8]

In the presented Table 1, the prediction intervals and their corresponding neutrosophic forms for different regression models are compared. For Linear Regression, the prediction interval ranges from 39.11 to 247.05 and is represented neutrosophically as 39.11+247.05*I*, where I spans from 0 to 0.842. Ridge Regression shows a slightly tighter interval from 39.65 to 252.67, with its neutrosophic form given by 39.65+252.67, I in the range [0, 0.843]. The Random Forest model, however, presents a different interval, spanning from 67.00 to 229.53, with the neutrosophic representation as 67.00+229.53*I* I between 0 and 0.708. The average model provides a general overview with a prediction interval from 48.59 to 243.08 and its neutrosophic form expressed as *N*=48.59+243.08*I*, I ranges up to 0.8. These intervals and forms are crucial for understanding the variability and confidence in the predictions made by each model.

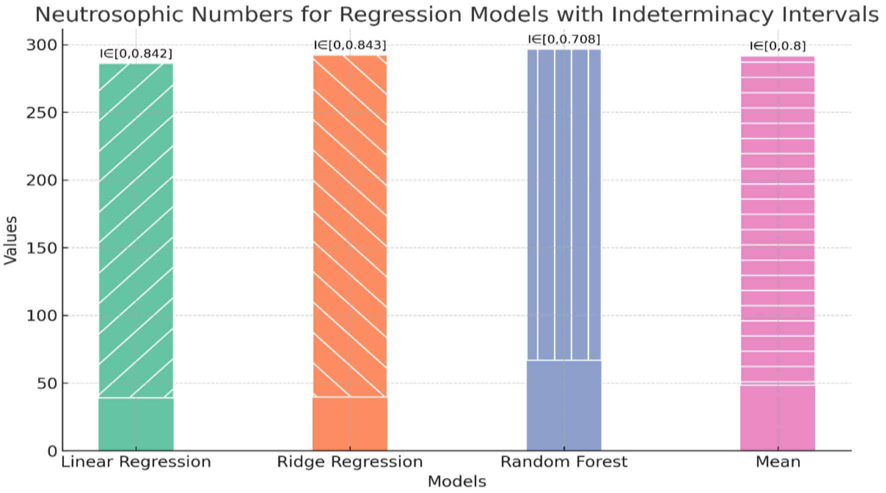

Figure 2. Neutrosophic number representation.

Figure 2 shows cases of the neutrosophic number representation of prediction intervals for various regression models, emphasizing the role of modeling indeterminacy to understand the dynamics within the analyzed data. The bars for each model—Linear Regression, Ridge Regression, Random Forest, and Mean—are marked with indeterminacy intervals, highlighting variations that may stem from complexities or uncertainties in the underlying data.

The indeterminate component of the neutrosophic numbers illustrates the growing complexity or variability of the data across different models. From a decision-making perspective, this analysis underscores the necessity for continuous monitoring and updates to predictive models to maintain alignment with the evolving data landscape. Such a practice ensures that decision-making processes remain robust despite the increasing uncertainty. This approach not only confirms the benefits of integrating neutrosophic statistics into regression analysis but also emphasizes the critical role these techniques play in enhancing our understanding of uncertainties within predictive modeling.

5 |Conclusion

In conclusion, the integration of neutrosophic statistics into regression analysis presents a promising avenue to enhance the reliability and robustness of predictive models. By augmenting traditional interval predictions with measures of indeterminacy, neutrosophic numbers offer a nuanced representation of the inherent uncertainty in datasets. This sophisticated methodology empowers decision-makers to make more judicious choices by acknowledging the variability inherent in predictions and considering a spectrum of potential outcomes. Furthermore, the incorporation of neutrosophic statistics underscores the imperative of iteratively refining predictive models to accommodate evolving data dynamics and escalating uncertainties.

Looking forward, potential research avenues could explore various facets to advance the utilization of neutrosophic statistics in predictive modeling. Firstly, investigating the impact of different neutrosophic aggregation techniques on model performance and uncertainty quantification could yield valuable insights for optimizing predictive accuracy. Additionally, exploring the synergistic potential of combining neutrosophic statistics with other machine learning paradigms, such as deep learning or ensemble methods, holds promise for enhancing predictive capabilities across heterogeneous datasets. Furthermore, extending the applicability of neutrosophic statistics beyond regression analysis to encompass diverse domains like classification or time series forecasting could broaden its utility and foster a more comprehensive understanding of uncertainty in machine learning frameworks. By addressing these avenues, future research endeavors stand poised to propel the frontier of predictive modeling and facilitate more informed decision-making in complex and uncertain settings.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Funding

This research has no funding source.

Author Contribution

All authors contributed equally to this work.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Smarandache, F. (2022). Neutrosophic statistics is an extension of interval statistics, while plithogenic statistics is the most general form of statistics (third version). Bulletin of Pure & Applied Sciences-Mathematics and Statistics, 41(2), 172-183.

[2] Xu, C., & Xie, Y. (2021, July). Conformal prediction interval for dynamic time-series. In International Conference on Machine Learning (pp. 11559-11569). PMLR.

[3] Dewolf, N., Baets, B. D., & Waegeman, W. (2023). Valid prediction intervals for regression problems. Artificial Intelligence Review, 56(1), 577-613.

[4] Mondal, K., Pramanik, S., Giri, B. C., & Smarandache, F. (2018). NN-Harmonic mean aggregation operators-based MCGDM strategy in a neutrosophic number environment. Axioms, 7(1), 12.

[5] Shrestha, D. L., & Solomatine, D. P. (2006). Machine learning approaches for estimation of prediction interval for the model output. Neural networks, 19(2), 225-235.

[6] Sun, Y., Wang, X., Zhang, C., & Zuo, M. (2023). Multiple Regression: Methodology And Applications. Highlights in Science, Engineering and Technology, 49, 542-548.

[7] Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development Discussions, 2022, 1-10.

[8] Carney, J. G., Cunningham, P., & Bhagwan, U. (1999, July). Confidence and prediction intervals for neural network ensembles. In IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339) (Vol. 2, pp. 1215-1218). IEEE.

[9] Carney, J. G., Cunningham, P., & Bhagwan, U. (1999, July). Confidence and prediction intervals for neural network ensembles. In IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339) (Vol. 2, pp. 1215-1218). IEEE.

[10] Wang, C. C., & Lee, W. C. (2019). A simple method to estimate prediction intervals and predictive distributions: summarizing meta-analyses beyond means and confidence intervals. Research Synthesis Methods, 10(2), 255-266.

[11] Smarandache, F., & Aslam, M. (Eds.). (2023). Cognitive Intelligence with Neutrosophic Statistics in Bioinformatics. Elsevier

[12] Aslam, M., & Arif, O. H. (2024). Simulating chi-square data through algorithms in the presence of uncertainty. AIMS Mathematics, 9(5), 12043-12056.

[13] Smarandache, F. (2022). Neutrosophic Statistics is an extension of Interval Statistics, while Plitogenic Statistics is the most general form of statistics (Fourth version)/La Estadistica Neutrosofica es una extension de la Estadistica de Intervalos, mientras que la Estadistica Plitogenica es la forma mas general de estadistica.(Cuarta version). Neutrosophic Computing and Machine Learning, 23, 21-39.

[14] Zhou, Y. (2023, March). Regression analysis for potential indicators of diabetes. In Second International Conference on Biological Engineering and Medical Science (ICBioMed 2022) (Vol. 12611, pp. 1323-1329). SPIE.

[15] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(2), 140-147.

[16] Hoerl, R. W. (2020). Ridge regression: a historical context. Technometrics, 62(4), 420-425.

[17] Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. The American Statistician, 63(4), 308-319.

[18] Castro Sánchez, F., Almeida Blacio, J. H., Flores Bracho, M. G., Andrade Santamaria, D. R., & Sánchez Casanova, R. (2021). Neutrosophic and Plithogenic Statistical Analysis in Educational Development. Neutrosophic Sets and Systems, 44(1), 26.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.