Parameter Estimation in Multiple Linear Regression: A Neutrosophic Perspective with the Simple Averaging Method (SAM)

Kesavulu Poola[1,*], V. Pavankumari[2], J. Anil Kumar[3], Akkyam Vani[4], Asif Alisha S.[5], A. Srinivasulu[6]

[1]Associate professor, Center for Management Studies, Jain University, Bengaluru, India
[2]Assistant professor, G. Narayanamma Institute of technology and science, Hyderabad, India
[3]Associate professor, Sri Venkateswara College of Engineering (Autonomous), Tirupati, India
[4]Research Scholar, Department of Statistics, Sri Venkateswara University, Tirupati, India
[5]Associate Professor, Department of mathematics, School of Liberal Arts and Sciences, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, India
[6]Guest faculty, Department of Statistics, Vikrama Simhapuri University, Nellore. India
Emails: kesav.poola@gmail.com, pavani.stat@gmail.com, aniklkumar.jk@svcolleges.edu.in, akkyamvani1@gmail.com, asif.alisha@gmail.com, drsrinivasulu81@gmail.com.

Abstract

Regression modeling is a significant statistical tool aimed at quantifying and understanding the nature of relations between the predictor and response variables. The routine parameter estimation procedures, like OLS and ML, are based heavily on the assumption of normality in data, which will not be the case for most real-world data scenarios. The paper presents a Neutrosophic approach for the estimation of parameters in multiple linear regression models, making use of the Neutrosophic principles to treat uncertainties, indeterminacies, and inconsistencies in actual data, a proposed method is called the Simple Averaging Method, or SAM. This is a robust alternative to traditional methods and provides reliable results even if the assumptions of normality are not held. SAM performance is tested using real-time crime data in the USA and demonstrates its capabilities to deal with complex datasets. The comparative analysis between the OLS model and the same model is done via RMSE and MAD metrics. The results show that SAM significantly outperforms OLS with an RMSE of 34.37598 in contrast to 58.05248 for OLS. Graphical analysis further confirms SAM's performance over and above OLS. Critical issues of regression modeling with incorporation of neutrosophic logic cover their critical challenges, especially when standard assumptions are violated.

Keywords: Simple Averaging Method; Ordinary Least Squares method (OLS); Maximum likelihood estimation (MLE); RMSE; MAE

1. Introduction

Regression analysis is the statistical modeling backbone that, among other things, aims to estimate the relationship between explaining variables and one dependent variable-the so-called linear regression [1]. Among the methods of estimating parameters, the most used are the Ordinary Least Squares (OLS) method and Maximum Likelihood (ML) estimation. A critical technique in regression analysis, OLS is based on a few assumptions related to classical assumptions, which include linearity in parameters and error terms with a mean of zero and constant variance (σ^2) [2]. Regression modeling has traditionally analyzed average relationships between the variables so that predictions can be made based on independent variables. However, the introduction of multiple regression results in problems like errors with constant variance, no autocorrelation, multicollinearity, and heteroscedasticity. Violations of these assumptions can significantly degrade the accuracy of the model, often ending in overfitting or under fitting. One more challenge

DOI: https://doi.org/10.54216/IJNS.260216

in data is outliers, which are data points that lie far away from the regression line and do not follow the general trend. These outliers can distort the accuracy of the regression model, making it harder to draw reliable conclusions or predictions. Outliers can reduce correlation coefficients and skew regression relationships or inflate correlations, making conclusions incorrect. Robust regression techniques are thus needed in an effort to mitigate such impacts effectively [3].

Linear and multiple regression techniques are widely applied across different areas of research, especially for variable relationship analysis and forecasting. However, overfitting and under fitting are critical problems as overfitting provides spurious regression coefficients, p-values, and R^2 values, and under fitting provides inflexible and unreliable models. These problems require proper development of the model; therefore cross-validation and data-driven parameter estimation method can be helpful in that context. Accurate parameter estimation holds a very important place in the regression analysis. Here, it calls for assumptions like linearity, and avoiding multicollinearity and autocorrelation. Metrics such as RMSE are quite useful in assessing model accuracy for appropriate Gaussian error distribution. However, it is typically violated in real world data [4]. To address these limitations, this research introduces the Simple Averaging Method (SAM), a novel and robust approach to parameter estimation. SAM overcomes the constraints of traditional methods, delivering enhanced accuracy and reliability even under challenging data conditions. This method represents a significant advancement in regression modeling, providing a practical solution to address the complexities and inconsistencies inherent in real-world datasets.

2. Related Work

Regression parameter estimation had flexibility in terms of achieving precision, robustness, and managing data complexity. Traditional Ordinary Least Squares (OLS) and Maximum likelihood methods (ML) and its improved median-based version, address violations of regression assumptions [1]. These methods eliminate the reliance on error normality and independence assumptions, providing robust parameter estimates. Ridge regression and Lasso, which implement penalty terms for multicollinearity, and Elastic net tries to balance feature selection with shrinkage. Robust estimation includes methods such as M-estimators and Least Absolute Deviations, trying to reduce the effect of outliers. Heuristic methods, including Particle Swarm Optimization, and hybrid approaches improve the parameter tuning in the complex datasets. Besides, non-assumption-based methods [5]. The literature on parameter estimation in regression models highlights diverse methodologies tailored to specific challenges. For example, the Maximum Likelihood (ML) approach remains a dominant method for estimating parameters in distributions like Weibull, often enhanced by optimization techniques like Particle Swarm Optimization (PSO). Recent adaptations incorporate confidence intervals to narrow search spaces, improving PSO performance in parameter estimation. Studies validate these approaches using Monte Carlo simulations and real-world data, confirming their effectiveness in increasing accuracy and computational efficiency. Using metrics like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) demonstrate the efficiency of methods. Particularly when data assumptions are violated, offering a practical alternative for simplified linear regression models [4]. Additionally, innovative models like Intuitionistic Fuzzy Logistic Regression extend parameter estimation to ambiguous datasets, capturing vagueness and hesitation simultaneously. By refining fuzzy models with revised frameworks, these approaches show superior performance in fitting real-world datasets, such as birth weight data. Metrics like membership functions validate the model's effectiveness compared to traditional fuzzy regression [6]. Collectively, these advancements underscore the evolving nature of regression methodologies, addressing complexities in modern data analysis.

Novel techniques for estimating parameters in linear regression are constantly being developed to face new challenges in data analysis. Bayesian methods provide probabilistic approaches toward parameter estimation, with prior knowledge and credible intervals [7]. In Quantile Regression, one estimates the parameters based on the conditional quantiles, giving a robust variant to the methods based upon the mean, especially for skewed or heteroscedastic data. More and more, it exploits the techniques of Gradient Boosting and Neural Networks for parameter estimation, especially in non-linear and high-dimensional datasets [6]. Bootstrap and resampling improve the reliability of parameter estimation by providing robust confidence intervals. Therefore, these collectively improve the adaptability, precision, and interpretability of linear regression models.

Advances in recent times are targeted toward enhanced robustness and efficiency of model fitting in the estimation of parameters of linear regression. Among these techniques are Ridge and Lasso regressions, which are now very popular for dealing with multicollinearity and overfitting. In Ridge regression, L2 regularization is utilized. Lasso uses L1 regularization, allowing for the automatic selection of relevant variables. Elastic Net is the combination of both Lasso and Ridge. In high-dimensional data with correlated predictors, it balances the strengths of both approaches [8]. These procedures greatly improve the model because coefficients are shrunk, and variance in predictions decreases. Apart

DOI: https://doi.org/10.54216/IJNS.260216

from regularization, robust regression has emerged as a technique to be used in addressing outliers and non-normal errors in a given data set. M-estimators, such as Huber and Tukey's biweight, are widely used as they minimize a weighted loss function that reduces the impact of outliers on parameter estimates (Huber, 1964). Quantile regression has also emerged as an alternative to Ordinary Least Squares (OLS), focusing on estimating conditional quantiles rather than the conditional mean [9]. This methodology gives a better account of the data distribution, especially for skewed datasets or where the error variance is not constant (Koenker and Bassett, 1978). Such methods provide greater robustness in real-world applications where data often deviates from assumptions of normality and homoscedasticity. Machine learning approaches to parameter estimation are also gaining momentum, especially when dealing with large datasets or complex relationships those traditional regression techniques may not capture. For example, GBM and Random Forest have already proven to be highly competitive for regression problems with non-linear dependencies [10]. These tree-based techniques not only perform excellent parameter estimation but also provide a natural way of handling interactions between variables automatically. Although traditional linear models cannot be applied, neural networks, especially deep learning models, have recently been studied more intensively to formulate effective regression tools. These learn complex patterns within the data-adaptively; they therefore had better predict more accurately in high-dimensional space [11]. Bayesian methods also provide a probabilistic alternative framework for the estimation of parameters that includes prior beliefs in the regression model. Bayesian regression models allow one to derive and compute a full posterior distribution of the parameters, allowing one to quantify uncertainty and improve generalization [12]This method is particularly helpful in cases of small sample sizes or when there is knowledge already, where more accurate estimations are possible by incorporating uncertainty in terms of priors [13]. The applications of Bayesian regression are further facilitated by the use of MCMC algorithms in practice because these are used to easily compute the posterior distribution (Gilks et al., 1996). These developments allow greater flexibility and accuracy in parameter estimation, especially in a complex and uncertain environment [14].

3. Methods and Materials

The OLS method and the ML method have been the most common approaches to parameter estimation in regression analysis. Nevertheless, the latter relies on the fundamental assumption of normality in the generating distribution, which does not necessarily hold in various applications. To overcome the disadvantages, Cliff and Billy proposed the Simple Averaging Method, as a robust alternative for parameter estimation of linear regression models [15] . The advantage of SAM is that it is not based on the assumptions of normality; hence, it is more beneficial in case of non-normal distribution of the data set. This article extends the use of SAM from linear regression to multiple linear regression models. Developed as an extension of foundational work by Cliff and Billy in 2017, these mathematical formulations allow for the estimation of parameters associated with multiple regression models using SAM [16]. These calculations are designed uniquely for regression parameter estimation without the need to adhere to the assumptions of normality, thereby adding flexibility and versatility to regression analyses. In order to test the effectiveness of SAM, the method was applied to real-world datasets and parameter estimation was compared with OLS[17] .The analysis resulted in the regression parameters estimates by SAM to be as effective as using OLS, thus displaying its reliability and robustness. In addition, standard model accuracy metrics, like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), were used to measure the performance of SAM and OLS. Results showed that SAM can offer a comparable, if not better, accuracy level, further justifying its feasibility as an alternate method for parameter estimation in both linear and multiple linear regression models, even when assumptions made about a traditional regression model are violated [18] .

A. General linear regression model

In the recent era, prediction has become a standard part of analysis in all fields. General linear model (GLM) regularly preferred predictions using traditional linear and multiple regressions. Generalized linear model (GLM) plays a vital role in the area of forecasting [19].

i. General linear regression parameter estimation

Modeling between one dependent and independent variable is called simple linear regression model.

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

Let there are 'n' paired variables $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$ then the model can prescribe as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (2)$$

DOI: https://doi.org/10.54216/IJNS.260216

Estimation of the parameters β_0, β_1 by OLS method satisfies the minimum sum of squares of errors (SSE)

$$\therefore S(\beta_0, \beta_1) = \sum_{i=1}^{n} \in_i = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2$$

To obtain β_0, β_1, the partial differentiation w.r.to β_0, β_1 is

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \text{ And } \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

The OLS method gives the estimates $\hat{\beta}_0$ of β_0 and $\hat{\beta}_1$ of β_1 as

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = \frac{S_{XY}}{S_{XX}}$$

ii. Multiple Linear Regression Parameter estimation

General linear regression having multiple having multiple predictor $\left(X_k\right)$ which can influence the responsive variable $\left(Y\right)$ the model for general linear can be specified in the following way [20],

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i \qquad (3)$$

Here $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}$, $X = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n\times k}$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$

Y is regressand, X_1, X_2..... X_k are 'k' regressors, $\beta_1, \beta_2 ..... \beta_p$ are 'p' parameters, ε_i is the residual term and n is total annotations on each variable

Here $\beta_1, \beta_2 ..... \beta_p$ are the coefficients of regression with respect to the response variable $\left(i = 1, 2, 3, ..., p\right)$ and $X_{1i,} X_{2i,} \ldots\ldots$are the predicted variables. Although the generalized linear regression model by nature alleviates some forms of multicollinearity, other assumptions like homoscedasticity (equal errors' variance) or the absence of autocorrelation (independence of residuals) are equally important [21]. Such violations result in biased parameter estimates and non-reliable inferences; therefore, robust alternative estimation methods are needed under such conditions.

The OLS method minimizes the residual sum of secure (RSS) w.r.to the parameter vector β and hence obtained the 'k' normal equations.

Let $\hat{\beta}$ OLS estimate of β then $\hat{Y} = X\hat{\beta}$

$\therefore$ Vector residual $\varepsilon = Y - \hat{Y}$

Then, Residual Sum of Squares (RSS) $\varepsilon\varepsilon' = \left(Y - \hat{Y}\right)\left(Y - \hat{Y}\right)'$

$$\varepsilon\varepsilon' = \left(Y - X\hat{\beta}\right)\left(Y - X\hat{\beta}\right)'$$

DOI: https://doi.org/10.54216/IJNS.260216

According to draper & smith (1992), to minimize the Sum of Squares of residuals (RSS) with parametric vector β

$$\frac{\partial \varepsilon\varepsilon'}{\partial \hat{\beta}} = \frac{\partial}{\partial \hat{\beta}}\left(YY' - 2YX\hat{\beta}' + \hat{\beta}'X'X\hat{\beta}\right) = 0 \quad \text{where } \hat{\beta} = \frac{X'Y}{X'X}$$

$$\hat{\beta} = \left(X'X\right)^{-1}\left(X'Y\right) \tag{4}$$

Similarly, OLS estimation of $\beta_1, \beta_2, ...\beta_k$ are $\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_k$, Then

$$\hat{\beta}_0 = \bar{Y} - \left(\hat{\beta}_1\bar{X}_1 + \hat{\beta}_2\bar{X}_2 + ... + \hat{\beta}_k\bar{X}_K\right) \tag{5}$$

B. Parameter Estimation – Simple Averaging Method (SAM) Method

i. Simple linear regression method- SAM

The linear regression model for the variables $\left(X_i, Y_i\right), i = 1, 2, ..., n$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \tag{6}$$

The equation (6) can be stated as

$$\hat{\beta}_1 = \frac{1}{n}\sum_{i=2}^{n}\frac{Y_i - Y_{i-1}}{X_{1i} - X_{1i-1}}$$

$$\therefore \hat{\beta}_1 = \frac{1}{n-1}\left\{\frac{Y_2 - Y_1}{X_{12} - X_{11}} + \frac{Y_3 - Y_2}{X_{13} - X_{12}} + ... + \frac{Y_n - Y_{n-1}}{X_{1n} - X_{1n-1}}\right\}$$

$$\therefore \hat{\beta}_1 = \frac{1}{n-1}\beta_{1i}$$

$$= \frac{1}{n^*}\beta_{1i} \quad \text{where } n^* = n-1$$

$$\beta_{1i} = \left\{\frac{Y_2 - Y_1}{X_{12} - X_{11}} + \frac{Y_3 - Y_2}{X_{13} - X_{12}} + ... + \frac{Y_n - Y_{n-1}}{X_{1n} - X_{1n-1}}\right\}$$

$$\therefore \hat{\beta}_1 = \frac{1}{n^*}\beta_{1i} = \bar{\beta}_1 \tag{7}$$

The estimator of $'\beta_0'$ by using SAM method is almost similar to OLS method. Then the computation procedure of $'\beta_0'$ is

Since $\therefore \hat{\beta}_1 = \bar{\beta}_1 = \frac{1}{n^*}\beta_{1i}$

$$\hat{\beta}_0 = \frac{1}{n-1}\sum_{i=2}^{n}\left(Y_i - \bar{\beta}_1.X_{1i}\right)$$

$$\hat{\beta}_0 = \frac{1}{n-1}\left\{\left(Y_2 - \bar{\beta}_1.X_{12}\right) + \left(Y_3 - \bar{\beta}_1.X_{13}\right) + ... + \left(Y_n - \bar{\beta}_1.X_{1n}\right)\right\}$$

DOI: https://doi.org/10.54216/IJNS.260216

$$\hat{\beta}_0 = \frac{1}{n^*}\sum_{i=2}^{n}\left(Y_i - \bar{\beta}_1.X_{1i}\right) \quad (8)$$

And finally, the error term (ε_i) of β_0 & β_1 defined as

$$\varepsilon_i = \left(\hat{\beta}_0, \hat{\beta}_1\right)$$

$$\hat{\varepsilon}_i = \sum_{i=2}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i}\right)$$

$$= Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i}\right)$$

By using simple averaging method, the error term 'ε_i' can be determined as,

$$\hat{\varepsilon}_i = \frac{1}{n-1}\sum_{i=2}^{n}\left\{Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i}\right)\right\}$$

$$= \frac{1}{n-1}\left\{\left[Y_1 - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{11}\right)\right] + \left[Y_2 - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{12}\right)\right] + \ldots + \left[Y_n - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1n}\right)\right]\right\}$$

$$\hat{\varepsilon}_i = \frac{1}{n^*}\sum_{i=2}^{n}\left\{Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i}\right)\right\} \quad (9)$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots.. + \beta_k X_{ki} + \varepsilon_i \quad (10)$$

According to the equation (5), $\hat{\beta}_1$ can be written as

$$\hat{\beta}_1 = \frac{1}{n-1}\sum_{i=2}^{n}\frac{Y_i - Y_{i-1}}{X_{1i} - X_{1i-1}}$$

$$= \frac{1}{n-1}\sum_{i=2}^{n}\left\{\frac{Y_2 - Y_1}{X_{12} - X_{11}} + \frac{Y_3 - Y_2}{X_{13} - X_{12}} + \ldots. + \frac{Y_n - Y_{n-1}}{X_{1n} - X_{1n-1}}\right\}$$

Similarly,

$$\hat{\beta}_2 = \frac{1}{n-1}\sum_{i=2}^{n}\left\{\frac{Y_2 - Y_1}{X_{22} - X_{21}} + \frac{Y_3 - Y_2}{X_{23} - X_{22}} + \ldots. + \frac{Y_n - Y_{n-1}}{X_{2n} - X_{2n-1}}\right\}$$

$$\hat{\beta}_0 = \frac{1}{n-1}\left\{Y_i - \left[\bar{\beta}_1 X_{1i} + \bar{\beta}_2 X_{2i} + \ldots.. + \bar{\beta}_k X_{ki}\right]\right\}$$

$$= \frac{1}{n-1} Y_i \psi$$

Then the intercept $'\beta_0'$ can be defined as

$$\hat{\beta}_0 = \frac{1}{n^*}\psi \quad (11)$$

Were $\psi = \left\{Y_i - \left[\bar{\beta}_1 X_{1i} + \bar{\beta}_2 X_{2i} + \ldots.. + \bar{\beta}_k X_{ki}\right]\right\}$

Finally, the error term can explain by

DOI: https://doi.org/10.54216/IJNS.260216

$$\hat{\varepsilon} = \frac{1}{n-1}\left\{Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}\right]\right\}$$

$$= \frac{1}{\mathrm{n}^*}\gamma \qquad (12)$$

Were $\gamma = \frac{1}{n-1}\hat{\varepsilon} = \frac{1}{n-1}\left\{Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}\right]\right\}$

4. Results and Discussion

The researcher used real time crime data (Source: *Life in America's Small Cities*, By G.S. Thomas) of USA. Researcher predicted the overall reported crime rate per 1 million residents in small cities of USA. According to the researcher perception lot of factors, which are, influences the crime rate:

Y = total overall reported crime rate per 1 million residents

X_1 = reported Average violent crime rate per 100,000 residents

X_2 = annual police funding in $/resident

X_3 = % of people 25 years+ with 4 yrs. of high school

X_4 = % of 16- to 19-year-olds not in high school and not high school graduates

X_5 = % of 18- to 24-year-olds in college

X_6 = % of people 25 years+ with at least 4 years of college

Table 1: Crime data: *Life in America's Small Cities*, of USA

Y	X1	X2	X3	X4	X5	X6
507	27	10	16	11	9	20
494	14.3	14	22	12	23	18
643	8.79	6	20	10	10	16
341	6	4	11	11	1	19
773	15	11	22	30	9	24
603	21	6	12	8	10	15
484	13	10	9	12	9	14
546	19	4	11	13	11	11
424	8	6	16	7	16	12
548	11.5	3	17	9	18	15
506	9.87	5	18	13	11	23
819	19.4	2	22	4	31	36
541	12.5	14	16	9	11	12
491	8.7	2	17	11	7	16
514	11.8	3	19	12	13	11
371	4.69	4	14	10	21	14
457	7.98	3	16	12	9	10
437	4.98	6	11	18	19	27
500	16.7	3	12	7	11	10
432	11.1	1	10	12	10	15

DOI: https://doi.org/10.54216/IJNS.260216

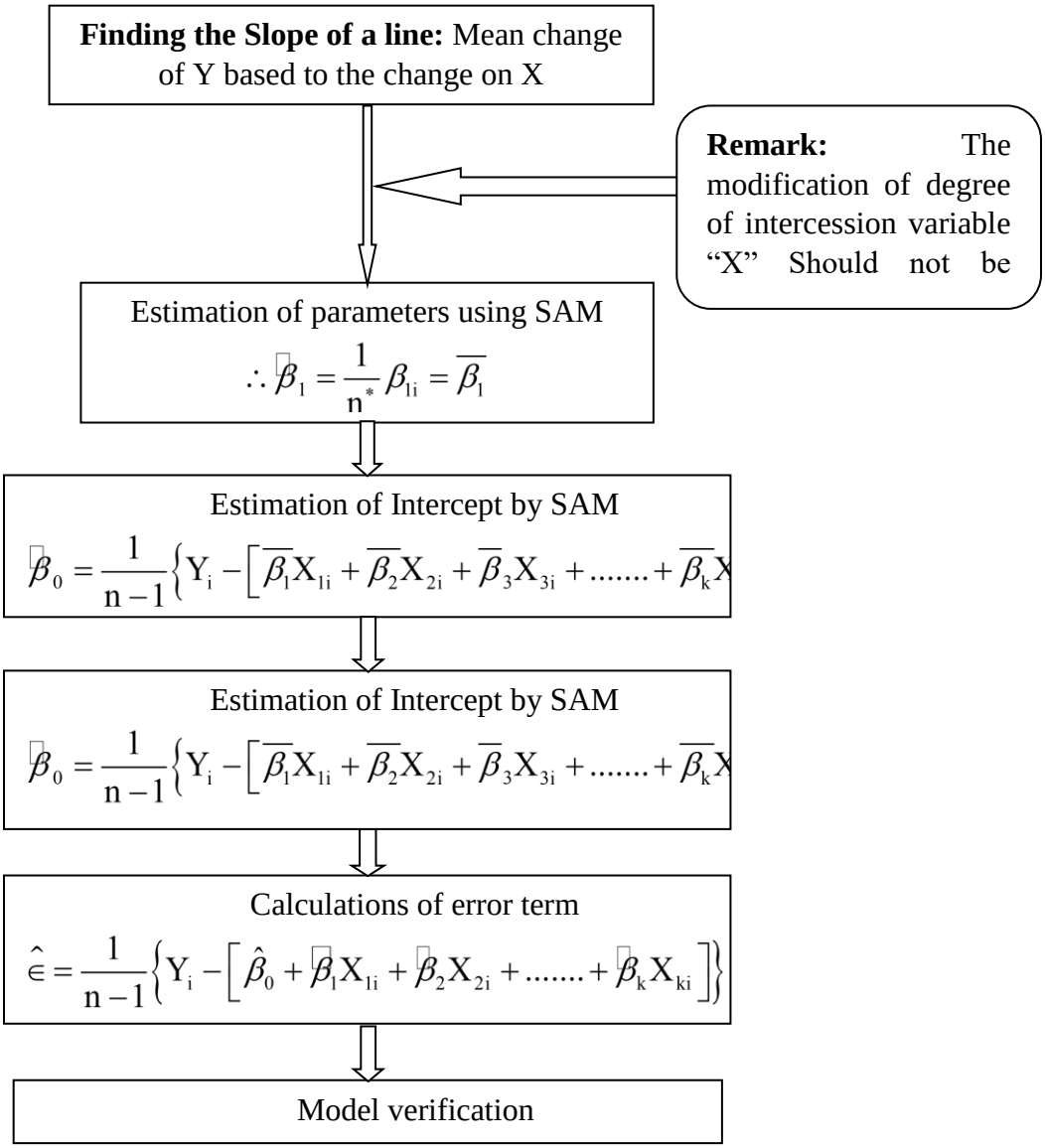

Figure 1. Parameter estimation algorithm for proposed SAM method**.**

A. Parameter estimation by OLS method

The Multiple linear regressions model, to predict the overall reported crime rate per 1 million residents can be defined as the following model[23]

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i}$$

Table 2: explain the parameter values estimated using ordinary least square method (OLS) that can be solved by using Table 1

Parameter	β_0	β_1	β_2	β_3	β_4	β_5	β_6
Estimated value	95.271	10.078	-4.850	13.103	4.315	0.851	3.755

DOI: https://doi.org/10.54216/IJNS.260216

Then fitted model for OLS method is:

$$\mathbf{Y_{OLS}} = 95.271 + 10.078X_1 - 4.850X_2 + 13.103X_3 + 4.315X_4 + 0.851X_5 + 3.755X_6$$

The regress and (Predicted) values of Table 1 and the error according to the estimated value when compare with the original data shown in the Table 3.

Table 3: Predicted and residuals of original data

Item	Original data (Y)	Predicted data $\mathbf{Y_{OLS}}$	Residuals
1	507	658.7454	-151.745
2	494	598.6963	-104.696
3	643	528.5567	114.4433
4	341	400.1358	-59.1358
5	773	708.3882	64.61175
6	603	534.3983	68.60175
7	484	407.7219	76.27807
8	546	518.2469	27.7531
9	424	445.3253	-21.3253
10	548	529.4466	18.55339
11	506	558.1698	-52.1698
12	819	748.5826	70.41742
13	541	455.9448	85.0552
14	491	509.504	-18.504
15	514	552.4441	-38.4441
16	371	420.176	-49.176
17	457	467.7811	-10.7811
18	437	455.7285	-18.7285
19	500	483.3743	16.62574
20	432	449.6334	-17.6334

A. Parameter estimation by SAM method

The General linear regressions model to predict the overall reported crime rate per 1 million residents can be defined as the following model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i}$$

Table 4 explain the parameter values estimated using Simple averaging Method (SAM) that can be solved by using Table.

DOI: https://doi.org/10.54216/IJNS.260216

Table 4: Regression Parameters estimation using SAM Method

$\beta_1 = \frac{yi - yi-1}{x1i - x1i-1}$	$\beta_2 = \frac{yi-yi-1}{x2i-x2i-1}$	$\beta_3 = \frac{yi - yi-1}{x3i - x3i-1}$	$\beta_4 = \frac{yi-yi-1}{x4i-x4i-1}$	$\beta_5 = \frac{yi - yi-1}{x5i - x5i-}$	$\beta_6 = \frac{yi - yi-}{x6i - x6i}$	$\overline{\beta_1}X_1$	$\overline{\beta_2}X_2$	$\overline{\beta_3}X_3$	$\overline{\beta_4}X_4$	$\overline{\beta_5}X_5$	$\overline{\beta_6}X_6$	$\beta_0 = \frac{1}{n}\psi$	**Predicted values**
1.024	-3.250	-2.167	-13.00	-0.929	6.500	277.0897	73.37498	121.1789	58.82531	89.34402	72.7877	-25.2371	695.6793
-27.042	1.625	6.500	6.500	1.000	6.500	170.3229	31.44642	110.1626	49.02109	38.84523	64.70018	123.7629	464.4985
108.244	6.500	1.444	-13.000	1.444	-4.333	116.2614	20.96428	60.58944	53.9232	3.884523	76.83146	-178.237	332.4543
48.107	-1.857	-1.182	-0.684	-1.625	-2.600	290.266	57.65177	121.1789	147.0633	34.9607	97.05026	253.7629	748.1709
-28.239	2.600	1.300	0.591	-13.000	1.444	406.9149	31.44642	66.09757	39.21687	38.84523	60.65641	83.76291	643.1774
14.875	-3.250	4.333	-3.250	13.000	13.000	251.8997	52.4107	49.57318	58.82531	34.9607	56.61265	-35.2371	504.2822
10.333	2.167	-6.500	-13.000	-6.500	4.333	368.1611	20.96428	60.58944	63.72742	42.72975	44.48137	26.76291	600.6534
11.091	-6.500	-2.600	2.167	-2.600	-13.000	155.0152	31.44642	88.1301	34.31477	62.15236	48.52513	-95.2371	419.584
35.838	4.333	-13.00	-6.500	-6.500	-4.333	222.0593	15.72321	93.63823	44.11898	69.92141	60.65641	28.76291	506.1175
26.415	-6.500	-13.00	-3.250	1.857	-1.625	191.25	26.20535	99.14636	63.72742	42.72975	93.0065	-13.2371	516.0654
32.706	4.333	-3.250	1.444	-0.650	-1.000	376.6869	10.48214	121.1789	19.60844	120.4202	145.5754	299.7629	793.952
39.885	-1.083	2.167	-2.600	0.650	0.542	241.6299	73.37498	88.1301	44.11898	42.72975	48.52513	21.76291	538.5089
13.263	1.083	-13.00	-6.500	3.250	-3.250	168.579	10.48214	93.63823	53.9232	27.19166	64.70018	-28.2371	418.5144
7.492	-13.000	-6.500	-13.000	-2.167	2.600	228.0661	15.72321	104.6545	58.82531	50.4988	44.48137	-5.23709	502.2493
20.198	-13.000	2.600	6.500	-1.625	-4.333	90.87766	20.96428	77.11384	49.02109	81.57498	56.61265	-148.237	376.1645
26.140	13.000	-6.500	-6.500	1.083	3.250	154.6277	15.72321	88.1301	58.82531	34.9607	40.43761	-62.2371	392.7046
6.667	-4.333	2.600	-2.167	-1.300	-0.765	96.49696	31.44642	60.58944	88.23797	73.80593	109.1815	-82.2371	459.7583
5.375	4.333	-13.00	1.182	1.625	0.765	323.5942	15.72321	66.09757	34.31477	42.72975	40.43761	-19.2371	522.8971
12.078	6.500	6.500	-2.600	13.000	-2.600	214.5023	5.24107	55.08131	58.82531	38.84523	60.65641	-87.2371	433.1516
β_1 =19.377	β_2 =5.241	β_3 =5.508	β_4 =4.902	β_5 =3.885	β_6 =4.044	228.6474	29.5155	85.52098	56.76127	51.11214	67.67979	3.078697	

DOI: https://doi.org/10.54216/IJNS.260216

Parameter estimation by SAM method for the above data using equation becomes, $\beta_0 = 3.078697$, $\beta_1 = 19.3777, \beta_2 = 5.241, \beta_3 = 5.508, \beta_4 = 4.902, \beta_5 = 3.885, \beta_6 = 4.0444$

Then, the fitted model for SAM method is

$\mathbf{Y_{SAM}} = 3.078697 + 19.377X_1 + 5.241X_2 + 5.508X_3 + 4.902X_4 + 3.885X_5 + 4.0444X_6$ The regress and (Predicted) values of Table 1 and the error according to the estimated value when compared with the original data shown in the Table 5.

Table 5: Predicted values of and the error according to the estimated value when compare with the original data

Item	Original data (Y)	Predicted data Y_{SAM}	Residuals
1	507		
2	494	464.498	29.50152
3	643	695.679	-52.6793
4	341	332.454	8.545697
5	773	748.171	24.82915
6	603	643.177	-40.1774
7	484	504.282	-20.2822
8	546	600.653	-54.6534
9	424	419.584	4.416026
10	548	506.118	41.88249
11	506	516.065	-10.0654
12	819	793.952	25.04801
13	541	538.509	2.491122
14	491	418.514	72.48557
15	514	502.249	11.75071
16	371	376.164	-5.1645
17	457	392.705	64.29541
18	437	459.758	-22.7583
19	500	522.897	-22.8971
20	432	433.152	-1.15161

B. Comparison of predictions

In order to understand the effectiveness Simple Averaging Method (SAM) parameter estimation, we have compared the results of both SAM and OLS methods. The below table **6**, explains that the predicted values of OLS and SAM methods are almost same.

Table 6: Predicted values of OLS and SAM

Item	Original data (Y)	Predicted Values Y_{SAM}	Predicted Values Y_{OLS}
1	507		659
2	494	464	599
3	643	696	529
4	341	332	400
5	773	748	708
6	603	643	534
7	484	504	408
8	546	601	518

DOI: https://doi.org/10.54216/IJNS.260216

9	424	420	445
10	548	506	529
11	506	516	558
12	819	794	749
13	541	539	456
14	491	419	510
15	514	502	552
16	371	376	420
17	457	393	468
18	437	460	456
19	500	523	483
20	432	433	450

The results offer strong evidence that SAM is an effective alternative to the OLS method. However, the researcher further seeks to carry out more analysis by using error metrics for the methods to establish the best model. This approach would ensure a complete comparison and verification of the performance of the SAM model against OLS.

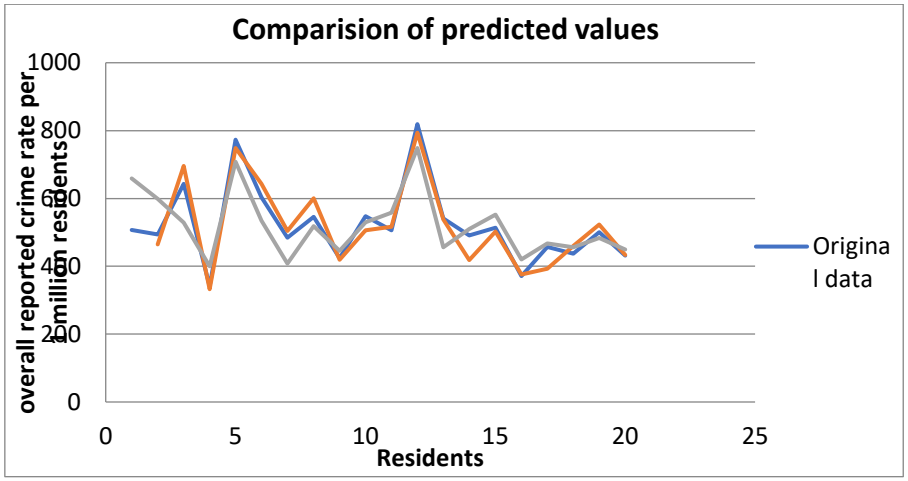

Figure 2. crime rate per 1 million residents in small cities of USA, for n=20

5. Model Evaluation

Over the years, researchers have been dependent on various error metrics in trying to evaluate model efficiency and accuracy. Traditional statistical methods apply metrics such as Root Mean Square Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Deviation (MAD) [24]to find the best fit of models applied across various fields, business, economics, and climate studies among others [25]. This study incorporates RMSE, MSE, and MAD in the process of accurately assessing the performance of the introduced models [26].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \qquad MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \qquad MAD = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

Table 7: Error metrics of OLS and SAM

Metrics	**SAM**	**OLS**
MAD	27.1092	49.10179
MSE	1181.708	3370.091
RMSE	34.37598	58.05248

DOI: https://doi.org/10.54216/IJNS.260216

From the above, SAM method perming well towards regression pater estimation because it got minimum RMSE, MSE and MAD when compare with OLS method [27].

6. Conclusion

This paper introduces the Simple Averaging Method, SAM, as a new approach to estimating parameters in multiple linear regression, based on Neutrosophic. It addresses certain uncertainties and inconsistencies in the data, providing a very robust alternative to the traditional OLS method, which is limited by strict assumptions such as normality and homoscedasticity. The superior accuracy of SAM for irregular datasets can easily be deduced from the empirical results supported by the RMSE, MSE, and MAD metrics. This innovative method highlights its potential in the improvement of regression modeling under tough conditions of the data. Further research could extend SAM for nonlinear multivariate data and datasets with missing values, significantly broadening its applicability in statistical analysis

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] A. M. Variyath and A. Brobbey, "Variable selection in multivariate multiple regression," *PLoS One*, vol. 15, no. 7, Jul. 2020, doi: 10.1371/journal.pone.0236067.

[2] P. Balestra, "On the efficiency of ordinary least-squares in regression models," *J. Am. Stat. Assoc.*, vol. 65, no. 331, pp. 1330–1337, 1970, doi: 10.1080/01621459.1970.10481168.

[3] B. Gafarov, "Generalized Automatic Least Squares: Efficiency Gains from Misspecified Heteroscedasticity Models," *arXiv preprint arXiv: 2304.07331*, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2304.07331

[4] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/peerj-cs.623.

[5] Ö. Türkşen, "A novel perspective for parameter estimation of seemingly unrelated nonlinear regression," *J. Appl. Stat.*, vol. 48, no. 13–15, pp. 2326–2347, 2021, doi: 10.1080/02664763.2021.1877638.

[6] M. R. Abonazel and I. M. Taha, "Beta ridge regression estimators: simulation and application," *Commun. Stat. Simul. Comput.*, vol. 52, no. 9, pp. 4280–4292, 2023, doi: 10.1080/03610918.2021.1960373.

[7] S. D. Permai and H. Tanty, "Linear regression model using Bayesian approach for energy performance of residential building," in *Procedia Comput. Sci.*, Elsevier B.V., 2018, pp. 671–677, doi: 10.1016/j.procs.2018.08.219.

[8] Y. M. Al-Hassan, "Performance of a new ridge regression estimator," *J. Assoc. Arab Univ. Basic Appl. Sci.*, vol. 9, no. 1, pp. 23–26, 2010, doi: 10.1016/j.jaubas.2010.12.006.

[9] M. S. Khan, A. Ali, M. Suhail, E. S. Alotaibi, and N. E. Alsubaie, "On the estimation of ridge penalty in linear regression: simulation and application," *Kuwait J. Sci.*, vol. 51, no. 4, Oct. 2024, doi: 10.1016/j.kjs.2024.100273.

[10] Y. J. Lei, B. Y. Wang, and Y. T. Yang, "Optimizing the loss function for bounding box regression through scale smoothing," *Ain Shams Eng. J.*, Nov. 2024, doi: 10.1016/j.asej.2024.103046.

[11] A. Yalçınkaya, İ. G. Balay, and B. Şenoğlu, "A new approach using the genetic algorithm for parameter estimation in multiple linear regression with long-tailed symmetric distributed error terms: an application to the COVID-19 data," *Chemom. Intell. Lab. Syst.*, vol. 216, Sep. 2021, doi: 10.1016/j.chemolab.2021.104372.

[12] S. Acitas, P. Kasap, B. Senoglu, and O. Arslan, "One-step M-estimators: Jones and Faddy's skewed t-distribution," *J. Appl. Stat.*, vol. 40, no. 7, pp. 1545–1560, 2013, doi: 10.1080/02664763.2013.788620.

[13] K. Poola, J. Anil Kumar, V. Pavankumari, P. Hemalatha, and N. M. Bhupathi, "Testing of multivariate nonlinear regression hypothesis using nonlinear least square (NLS) estimation," *Int. J. Stat. Appl. Math.*, vol. 5, no. 6, pp. 147–150, 2020.

[14] A. J. Telmoudi, M. Soltani, L. Chaouech, and A. Chaari, "Parameter estimation of nonlinear systems using a robust possibilistic c-regression model algorithm," *Proc. Inst. Mech. Eng. I J. Syst. Control Eng.*, vol. 234, no. 1, pp. 134–143, Jan. 2020, doi: 10.1177/0959651818756246.

DOI: https://doi.org/10.54216/IJNS.260216

[15] A. Prabowo, A. Sugandha, A. Tripena, M. Mamat, Sukono, and R. Budiono, "A new method to estimate parameters in the simple regression linear equation," *Math. Stat.*, vol. 8, no. 2, pp. 75–81, 2020, doi: 10.13189/ms.2020.080201.

[16] A. D. Al-Nasser and A. Radaideh, "Estimation of simple linear regression model using L-ranked set sampling," 2008.

[17] H.-H. Huang and Q. He, "Nonlinear Regression Analysis," Feb. 2024, doi: 10.1016/B978-0-12-818630-5.10068-5.

[18] F. Prihatmono, M. Y. Darsyah, and A. Karim, "Residual bootstrap resampling method for multiple linear regression model parameter estimation," *J. Litbang Edusaintech*, vol. 1, no. 1, pp. 35–43, Dec. 2020, doi: 10.51402/jle.v1i1.8.

[19] P. H. Sekhar, H. Sekhar, K. Poola, and B. Naidu, "Combined multiple forecasting model using regression," *Int. J. Stat. Appl. Math.*, vol. 5, no. 6, pp. 147–150, 2020.

[20] M. R. Abonazel, Z. Y. Algamal, F. A. Awwad, and I. M. Taha, "A new two-parameter estimator for beta regression model: Method, simulation, and application," *Front. Appl. Math. Stat.*, vol. 7, Jan. 2022, doi: 10.3389/fams.2021.780322.

[21] P. Kesavulu, M. Bhupathi Naidu, and P. Balasiddamuni, "Impact Factor: 5.2 IJAR," *Int. J. Appl. Res.*, vol. 2, no. 12, pp. 506–509, Nov. 2016.

[22] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," Jul. 2022, doi: 10.5194/gmd-15-5481-2022.

[23] J. Singthongchai, N. Thongmual, and N. Nitisuk, "An improved simple averaging approach for estimating parameters in simple linear regression model," *Math. Stat.*, vol. 9, no. 6, pp. 939–946, Nov. 2021, doi: 10.13189/ms.2021.090610.

[24] S. Yasin, S. Kamal, and M. Suhail, "Performance of some new ridge parameters in two-parameter ridge regression model," *Iran J. Sci. Technol. Trans. A Sci.*, vol. 45, no. 1, pp. 327–341, Feb. 2021, doi: 10.1007/s40995-020-01019-7.

[25] R. M. Dudley, "The speed of mean Glivenko-Cantelli convergence," *Ann. Math. Stat.*, vol. 40, no. 1, pp. 40–50, 1969, doi: 10.1214/aoms/1177698212.

[26] M. J. Wainwright, "High-dimensional statistics: A non-asymptotic viewpoint," *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, 2019, doi: 10.1017/9781108627771.

[27] G. C. Calafiore and L. El Ghaoui, "Optimization Models," *Encyclopedia of Systems and Control*, pp. 1–8, Springer, 2021, doi: 10.1007/978-1-4471-5102-9_237-2.

DOI: https://doi.org/10.54216/IJNS.260216