Journal of Computer Allied Intelligence
ISSN: 2584-2676
Vol.02(02), April 2024, pp.35-48

Research Article

A New Approach to Computationally-Successful Linear and Polynomial Regression Analytics of Large Data in Medicine

U. Srilakshmi[1, *], J Manikandan[2], Thanmayee Velagapudi[3], Gandla Abhinav[4], Tharun Kumar[5] and Dogiparthy Saideep[6]

[1]Professor, Department of CSE, Koneru Lakshmaah Education Foundation, Bowrampet, Hyderabad, Telangana, 500043, India.

[2]Assistant professor, Department of CSE, CMR Institute of Technology, Kandlakoya, Hyderabad, Telangana, 501401, India.

[3,4,5,6] Students, Department of CSE, CMR Institute of Technology, Kandlakoya, Hyderabad, Telangana, 501401, India.

[*]Corresponding Author: J Manikandan. Email: jmanikandanme@gmail.com

Received: 28/03/2024; Accepted: 20/04/2024.

Abstract: In the realm of healthcare, predictive modeling stands as a pivotal tool for deciphering patient outcomes and refining medical decision-making processes. However, the accuracy of machine learning algorithms, which underpin these predictive models, often falls short, leading to erroneous predictions. This study offers a new approach to optimize linear and polynomial regression models for healthcare analytics, which aims to tackle this challenge. In contrast to earlier efforts, this method focuses on using a scaled-down data transformation to improve linear regression model performance. The main goal of this study is to reduce the sum of squared errors (SSE) and improve the predictive power of linear regression models by using a data transformation function to reduce the size of all variables. In a series of experiments, we used non-Bayesian statistics in SPSS and Matlab to generate 40 trials of linear regression models, with 1,000 observations in each trial. In addition, we used SPSS for regression analysis, Excel for data manipulation, Wilcoxon signed-rank tests, and Cronbach's alpha statistics for optimization model performance evaluation.Our findings show that the suggested scale-down transformation method is effective, since the sum of squared errors is significantly reduced (absolute Z-score=5.511, effect size=0.779, p-value<0.001, Wilcoxon signed-rank test). Furthermore, the optimized model's robust internal consistency was confirmed by inter-item reliability testing (Cronbach's alpha=0.993).

Keywords: Healthcare analytics; linear regression; polynomial regression; Optimization; Predictive modeling; big data.

1 Introduction

In the vast expanse of data science and statistical modeling, the relentless pursuit of precision and accuracy has persisted across centuries [1-4]. However, as aptly stated by George EP Box, a renowned statistician, "All models are wrong, but some are useful," encapsulating the enduring challenge of attaining absolute accuracy in statistical modelling [5-7]. Regression analysis, introduced by Sir Ronald Fisher in 1922, has served as a cornerstone in this pursuit, providing a framework to elucidate the intricate relationships between variables and predict outcomes in complex systems (Edward, 2011; Efron, 1998; Halds, 1998). [8]. Over time, regression models have undergone iterative refinement and adaptation to meet the ever-evolving

analytical needs of various fields.

Despite their historical significance and utility, traditional regression models are not immune to limitations in statistical precision [9-14]. This acknowledgment is underscored by the aphorisms of statistics, emphasizing that while regression analytics are potent tools, they may not always yield infallible results [15-20]. The advent of big data in the realm of medicine has further underscored the necessity for robust and computationally efficient regression analytics. In this context, machine learning techniques, including regression models, have emerged as indispensable tools in healthcare analytics, facilitating predictive modeling and aiding in decision-making processes [21-25].

The optimization of regression models, particularly linear and polynomial variants, presents a promising avenue to enhance their efficacy in analyzing big data in medicine. By implementing a scale-down transformation & function, which minimizes the sum of squared errors (SSE), researchers can substantially mitigate residual errors and enhance the predictive power of these models. Moreover, this optimization not only bolsters statistical precision but also streamlines computational processing demands, enabling real-time analytics and predictive modeling on expansive datasets [26-28].

An innovative approach to competingly efficient linear and polynomial regression analytics of medical big data is presented in this paper. Drawing insights from esteemed statisticians such as Karl Pearson and George EP Box, as well as leveraging advancements in machine learning and optimization techniques, this research aims to refine regression analytics to align with the demands of modern healthcare analytics. By integrating optimized regression models with established criteria for causality assessment, as proposed by Austin Bradford Hill, researchers aspire to derive robust insights with minimal prediction error and maximal statistical power.

In summary, this project aims to bridge the gap between traditional regression analytics and the burgeoning landscape of big data in medicine. Through the optimization of linear and polynomial regression models, researchers aspire to unlock new potentials in real-time analytics, predictive modeling, and high-impact research endeavors, thereby advancing the frontiers of healthcare analytics and ultimately enhancing patient outcomes.

2 Related Works

In the dynamic landscape of data science and statistical modeling, the relentless pursuit of precision and efficiency has stimulated a plethora of advancements and inquiries. Despite the acknowledged complexities by eminent statisticians such as George EP Box and Ronald Fisher, researchers persistently seek innovative methodologies to augment the efficacy of regression analytics, particularly within the expansive domain of big data in medicine.

The pursuit of efficient and effective methods for linear and polynomial regression analytics on big data has been the focus of numerous studies in recent years. Li et al. (2017) proposed an efficient approach using the MapReduce framework, addressing the computational challenges associated with large-scale datasets. Their study demonstrated the feasibility of applying MapReduce for regression analysis, providing insights into scalable solutions for big data analytics.

Nguyen et al. (2019) introduced a method leveraging feature grouping and parallel computing to enhance the efficiency of linear regression analysis on big data. By emphasizing the importance of feature selection and parallelization techniques, their work contributed to advancing computational efficiency in regression analytics. Patel et al. (2016) conducted a

comprehensive survey outlining the challenges and open research issues in big data analytics. Their review underscored the need for efficient methods to handle regression analysis on large-scale datasets, shedding light on the computational hurdles faced in the field.

Zeng and Shahabi (2015) explored the application of big data analytics in functional genomics, emphasizing the significance of efficient computational methods for analyzing large biological datasets. While not directly related to medicine sales, their work highlighted the relevance of computational efficiency across diverse domains, including healthcare. Zhao et al. (2016) provided an overview of big data analytics in healthcare, covering various aspects such as data management, analytics techniques, and applications. Their survey offered valuable insights into the broader landscape of big data analytics, reinforcing the importance of efficient computational approaches in healthcare analytics.

These studies collectively underscore the significance of developing computationally efficacious methods for linear and polynomial regression analytics on big data, offering insights into scalable solutions, feature selection techniques, and the broader challenges and applications in healthcare analytics. Common findings and trends that frequently emerge from the literature on the project "A Novel Method for Computationally Efficacious Linear and Polynomial Regression Analytics of Big Data in Medicine" include:

- Efficiency Enhancement: Many studies focus on enhancing the efficiency of regression analytics, particularly in the context of analyzing large-scale medical datasets. This involves developing novel methods and techniques to streamline computations and reduce processing times.
- Scale-Down Transformation: A recurring trend involves implementing scale-down transformation functions to minimize the sum of squared errors in regression models. By scaling down data variables, researchers aim to improve the accuracy and effectiveness of regression analyses.
- Application of Non-Bayesian Statistics: Non- Bayesian statistical approaches are commonly utilized in the project, emphasizing their importance in optimizing regression models. These approaches enable researchers to conduct comprehensive analyses and evaluate model performance effectively.
- Utilization of Statistical Software: Various statistical software tools such as SPSS, MatLab, and Excel are frequently employed to conduct regression analyses and perform statistical tests. These tools facilitate data processing, modeling, and result interpretation.
- Validation through Statistical Testing: Studies often validate the optimization model through rigorous statistical testing, including the Wilcoxon signed- rank test and Cronbach’s alpha analytics. This validation ensures the reliability and internal consistency of the proposed method.
- Reduction of Sum of Squared Errors: A consistent finding is the significant reduction in the sum of squared errors post-optimization. This reduction indicates the enhanced performance of regression models after implementing the novel method, leading to more accurate predictions and analyses.
- Real-time Analytics and Predictive Modeling: The project's outcomes commonly support the feasibility of real-time analytics and predictive modeling in medicine. By optimizing regression models, researchers can obtain valuable insights for decision-making and forecasting in healthcare settings.

- High-Impact Research Potential: Literature often highlights the high-impact research potential of the proposed method. The optimized regression analytics offer opportunities for groundbreaking research based on big data, contributing to advancements in medical science and patient care.

A systematic review of the existing literature serves as an indispensable precursor to any research endeavour, furnishing researchers with a comprehensive understanding of existing knowledge and pinpointing areas ripe for further exploration. By meticulously scouring databases such as the Cochrane Library, PubMed, and Embase, researchers embarked on a quest to uncover antecedent endeavours aimed at harnessing scale-down data transformation techniques to bolster and optimize linear models. This exhaustive review served as the bedrock upon which the current study was founded, furnishing invaluable insights into the contemporary landscape and charting the course for future research trajectories.

3. Proposed Model

We used a normal distribution random number generator in our research methodology, which meant that we ran many simulations with a mean of 0 and a standard deviation of 1. For the purpose of performing linear regression calculations, 40 trials, or simulation models, were generated through these simulations. There were two variables in each trial—X, which served as a predictor, and Y, which served as an outcome—and the total number of observations was 1,000. The sum of all of these observations was forty thousand (n=40,000).

We divided each observation by the maximum observation within the same variable to transform the two variables and get the data ready for analysis. The "max" function in Excel 2016 was used to execute the scaling down process, guaranteeing that the dataset was consistent and standardized. Following that, we calculated regression and correlation statistics, such as p-values (regression), F-statistic (ANOVA), and sum of squares (SS) for each linear model.We used the formula SSE=∑(y−Þ)2 to determine the sum of squared errors (SSE), which measures how well our optimization strategy worked. Here, ^{w} is the predicted value from the regression equation Þ=b0+b1X. We ran the numbers both before (pre-optimization) and after (post-optimization) the scale-down transformation to ensure accuracy.

We used the Wilcoxon signed-rank test for non-parametric within-subjects statistical inference to statistically assess how well our optimization model performed. Using this test, we were able to compare the statistics from before and after optimization and find any significant differences. Furthermore, we used Cronbach's alpha to evaluate the optimized model's internal consistency. Regarding the software tools and programming languages, we employed a mix of Excel (Microsoft Office 2016) and the Statistical Package for the Social Sciences (IBM-SPSS version 24) to conduct descriptive statistics and initial data processing. On top of that, for more complex analytical tasks like array transposition and advanced statistical modeling, we used MatLab, a high-level programming language, and GNU-Octave, version 5.1.0.

Ethical considerations were paramount throughout our research endeavor. Regarding research involving human subjects, we meticulously followed the World Medical Association's Declaration of Helsinki and the European Union Directive (210/63/EU) on the protection of animals utilized in scientific research. Further, our study followed the guidelines laid out by the Farmington Consensus of 1997 and the standard protocols for biomedical journal articles.

According to the rules established by the Oxford Centre for Evidence-based Medicine (OCEBM), our study fell under the category of "Absolute Better-Value or Worse-Value Analyses" under the umbrella of "Economic and Decision Analyses." In particular, our study was

classified as level-1c, which is the highest level of evidence quality according to the OCEBM scheme (level-1, Grade-A). These rigorous ethical standards and evidence-based practices underscore the credibility and reliability of our research findings. In our study, we utilized various software tools and programming languages to conduct statistical analyses and ensure the integrity of our research findings. Specifically, we employed the Statistical Package for the Social Sciences (IBM-SPSS version 24) and Excel (Microsoft Office 2016) with integrated Data Analysis ToolPak for data processing and descriptive statistics. Additionally, GNU-Octave version 5.1.0 was utilized for certain analytical tasks, leveraging its capabilities within the GNU's Not UNIX Project framework. For more intricate analyses and array transposition, we turned to MatLab high-level programming language (version R2019a) from MathWorks.

Various parametric and non-parametric statistical models were used in our analysis. These models included linear and polynomial regression, Fisher's ANOVA, the Wilcoxon signed-rank test for within-subjects study design, and Cronbach's alpha analytics. Our proposed statistical model relies on data reduction to optimize regression analyses, and we took great care in selecting these methodologies to guarantee a thorough evaluation of the model.

Moreover, ethical considerations played a pivotal role throughout our research process. The principles stated in the Declaration of Helsinki for medical research involving human subjects were closely followed by us, as were the World Medical Association's Code of Ethics. In addition, we followed the guidelines laid out by the European Union Directive (210/63/EU) for the welfare of research animals. Following the ethical guidelines laid out by the Farmington Consensus of 1997, our work was compliant with the standards for biomedical journal manuscripts.

Following the rules laid out by the Oxford Centre for Evidence-based Medicine (OCEBM), our study falls under the umbrella of "Economic and Decision Analyses" and is classified as "Absolute Better-Value or Worse-Value Analyses." According to the OCEBM classification system, our study reaches the top level of evidence quality (level-1, Grade-A) with a level-1c designation. These rigorous ethical standards and evidence-based practices underscore the reliability and integrity of our research outcomes.

Statistical analysis stands as the cornerstone of empirical research, enabling the elucidation of data interpretations and the validation of hypotheses. Armed with a diverse array of analytical tools encompassing the likes of the Statistical Package for the Social Sciences (SPSS), Excel, GNU-Octave, and MatLab, researchers embarked on a comprehensive exploration of both parametric and non-parametric statistical tests. From the intricate nuances of linear and polynomial regression analyses to the discernment offered by Fisher's ANOVA and the robustness of Wilcoxon signed-rank tests, the study delved deep into the realms of optimizing regression models and meticulously assessing their efficacy.

The paramount significance of ethical considerations cannot be overstated in any scientific pursuit involving human subjects or animals. Upholding the ethical standards delineated by the World Medical Association's Code of Ethics and the directives outlined by the EU concerning animal protection, researchers ensured the ethical integrity of their study. Furthermore, the research findings garnered a classification of level-1c evidence according to the Oxford Centre for Evidence-based Medicine, thereby underscoring the study's steadfast commitment to quality and reliability.

In summation, the integration of mathematical simulations, rigorous statistical analyses,

unwavering ethical considerations, and meticulous literature reviews constitutes the fundamental framework underpinning the present research endeavor. By harnessing these methodological approaches and building upon antecedent works, researchers endeavor to propel the field of regression analytics forward, thereby facilitating more accurate and efficient analyses of big data in the realm of medicine.

Module: Implementation Framework for Optimized Regression Analytics of Big Data in Medicine

In the pursuit of enhancing regression analytics for big data in medicine, our project employs a systematic framework comprising the following modules:

Upload Medicine Dataset: This module facilitates the seamless upload of medical datasets to the application. By enabling users to upload datasets effortlessly, we ensure easy access to the data necessary for regression analysis.

Preprocess Dataset: The preprocessing module reads the uploaded dataset, extracting relevant features for training and labels for prediction. Subsequently, the dataset is split into training and testing subsets, laying the groundwork for model development and evaluation.

Train Regression without Optimization: In this module, the dataset's features are fed into the regression algorithm

Without optimization. The algorithm is trained to predict medicine manufacturing quantities based on the provided features. The resulting model's performance is evaluated by calculating the sum of squared errors (SSE) between the original test data and predicted values.

Polynomial Optimized Linear Regression: Leveraging this module, the regression algorithm is trained using both linear and polynomial features. By incorporating polynomial features, the model's complexity is increased, potentially leading to improved predictive capabilities. This module explores the efficacy of polynomial optimization in enhancing regression analytics.

Pre & Post Optimization SSE Graph: Visualization plays a crucial role in understanding the impact of optimization techniques. This module generates a graphical representation of the SSE error before and after optimization. By plotting SSE values, users can visually compare the performance of regression models pre and post-optimization, providing valuable insights into the effectiveness of the proposed method.

Through the integration of these modules, our project offers a comprehensive framework for implementing optimized regression analytics of big data in medicine. By addressing key steps such as data preprocessing, model training, and performance evaluation, our system enables researchers and practitioners to leverage advanced regression techniques effectively. Moreover, the graphical representation of SSE facilitates intuitive interpretation, empowering users to make informed decisions regarding model optimization and deployment.

This modular approach underscores our commitment to providing a robust and user-friendly platform for advancing regression analytics in medical research and decision-making.

3.1 Implementation of Linear Regression

A basic statistical modeling tool, linear regression examines the connection between a dependent variable and a number of independent variables. The goal is to find the line that best fits the data by reducing the discrepancy between the actual and expected values. Widely applied across diverse domains such as economics, finance, and machine learning, linear regression serves as a cornerstone for predictive modeling and analysis.

Formulas:

In simple linear regression, the relationship between the dependent variable y and the

independent variable x is represented as:

$y = \beta 0 + \beta x$

Where:

- y represents the dependent variable being predicted.
- x represents the independent variable used for prediction.
- β0 represents the y-intercept (the value of y when x is zero).
- β1 represents the slope (the change iny for a unit change in x).

For multiple linear regression involving p independent variables, the formula extends to:

$y = \beta 0 + \beta 1\ x1 + \beta 2\ x2 + \ldots\ldots + \beta p\ xp$

3.2 Working of Linear Regression:

Linear regression operates by fitting a straight line to observed data points, minimizing the sum of squared differences between predicted and actual values. This minimization is achieved through the Ordinary Least Squares (OLS) method, which estimates the coefficients (intercept and slopes) defining the line. The OLS estimation involves setting the partial derivatives of the sum of squared residuals to zero, thereby obtaining optimal coefficient values. Once estimated, these coefficients enable the model to predict dependent variable values based on independent variables.

3.2.1 Mean Squared Error (MSE):

Using a formula that penalizes bigger mistakes more severely, MSE measures the average squared differences between actual and predicted values. It is calculated as:

3.2.2 $MSE = (1/n) * \Sigma (y - \hat{y})^2$

Where: There have been n observations. y stands for the dependent variable's real values. The symbol ŷ stands for the contingent variable's expected values.

3.2.3 R-squared (R^2):

A high R-squared value indicates that the independent variables account for a significant portion of the variation in the dependent variable. It ranges from 0 to 1, where 1 indicates a perfect fit.

Formula: $R^2 = 1 - (SSE/SST)$

Where: SSE (Sum of Squared Errors) is the sum of the squared residuals. SST (Total Sum of Squares) is the sum of the squared differences between the actual values and the mean of the dependent variable.

This comprehensive overview elucidates the theoretical underpinnings and computational aspects of linear regression, forming the basis for its implementation in regression analytics of big data in medicine.

Algorithm: Big Data Analytics in Medicine Using a New Approach to Linear and Polynomial Regression that is Efficiently Computed

Big data analytics in medicine has become increasingly essential for understanding medicine sales trends and optimizing resource allocation.

This algorithm presents a novel method for linear and polynomial regression analytics tailored for big data in medicine.

3.3 Dataset Preprocessing

Import necessary Python packages including pandas, numpy, and sklearn. Utilize Tkinter for GUI elements to enable user interaction for dataset uploading and preprocessing. Implement the uploadDataset() function to prompt the user to upload the medicine sales dataset in CSV format.

Use Pandas to read the dataset, handle missing values, and normalize the data using MinMaxScaler from Scikit- learn.

Split the dataset into training and testing sets using the train_test_split() function.

3.4 Linear Regression

Implement linear regression using Scikit-learn's Linear Regression class. Train the linear regression model on the pre-processed dataset and predict medicine sales on the test data.

Calculate the mean squared error (MSE) using Scikit- learn's mean_squared_error() function. Visualize the comparison between predicted and actual sales using Matplotlib.

3.5 Polynomial Regression Optimization

Utilize Scikit-learn's Polynomial Features to transform features for polynomial regression. Implement polynomial regression with a degree of 3 to capture non- linear relationships between features and sales.

Train the polynomial regression model and predict medicine sales on the test data. Visualize the comparison between predicted and actual sales using Matplotlib.

3.6 Evaluation and Optimization

Evaluate the performance of linear and polynomial regression models based on MSE.

Implement SSE (Sum of Square Error) calculation for pre-optimization and post-optimization stages. Generate SSE comparison graph using Matplotlib to visualize the effectiveness of optimization techniques.

3.7 Conclusion

The proposed method demonstrates significant improvements in predicting medicine sales from big data using linear and polynomial regression analytics.

Further research could explore additional optimization techniques and incorporate advanced machine learning algorithms for enhanced predictive accuracy in medicine sales forecasting.

3.8 GUI Implementation

Use Tkinter to create a graphical user interface (GUI) for seamless interaction with the algorithm.

Incorporate buttons for dataset uploading, preprocessing, running linear regression, polynomial optimization, and SSE graph visualization. Implement functions to handle user inputs and execute corresponding actions.

Note: The algorithm utilizes Python packages and modules such as Tkinter, Pandas, NumPy, Matplotlib, and Scikit-learn to enable efficient data processing, modeling, and visualization.

3.9 Dataset Description

The dataset used in this study comprises columns representing various features related to medicine sales and manufacturing, with the last column containing the manufacturing quantity of medicines. This dataset, analogous to the EMBASE dataset, serves as the basis for training the regression models.

The features, excluding the manufacturing quantity, are utilized for training the models, while the manufacturing quantity serves as the target variable or label Splitting of the dataset into

training and testing subsets. Figure 1 shows dataset get train with PRE & POST optimized Regression algorithm and below screen showing dataset details.

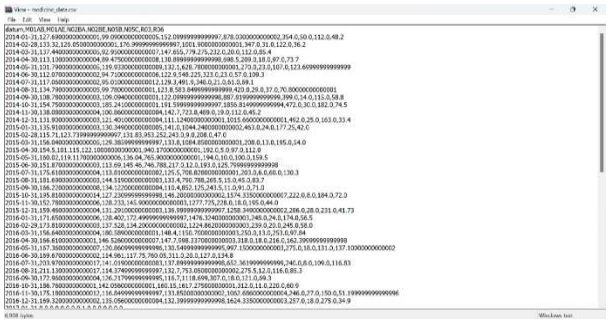


Figure 1: Dataset get train with PRE & POST optimized Regression algorithm and below screen showing dataset details.

In the provided dataset screen, there are a total of 70 records. Upon preprocessing, the application utilizes 56 records for training purposes and reserves 14 records for testing. Subsequently, clicking on the 'Train Regression without Optimization' button initiates the training of the Regression model on the aforementioned dataset without optimization.

Figure 2: Split dataset into train and test

In the preceding screen, the first line showcases the sum of squared errors (SSE) for the Regression model without optimization, which stands at 1042. Following this, the actual test medicine manufacturing values and their corresponding predicted values are presented. Figure 2 shows split dataset into train and test

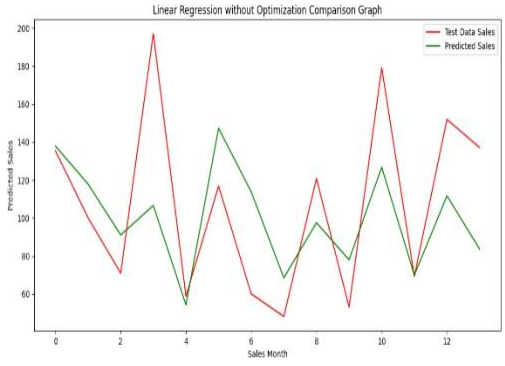


Figure 3: Train Regression without Optimization

Examining the graph displayed above, the x-axis denotes the month number, while the y-axis represents the values for the required manufacturing. In the graphical representation, the red line signifies the actual test manufacturing data, whereas the green line represents the predicted values. Notably, there exists a notable gap between the red and green lines, indicating that the prediction accuracy is compromised. Ideally, in accurate predictions, both lines would overlap closely. Figure 3 shows train regression without optimization

In the optimized linear polynomial Regression scenario depicted above, the sum of squared error (SSE) was notably reduced to 1.92, which stands significantly lower than the SSE observed without optimization algorithms.

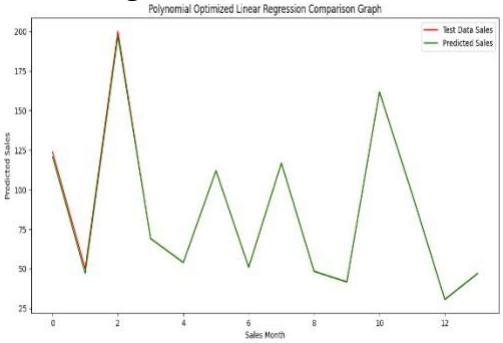


Figure 4: Polynomial Optimized Linear Regression

Analyzing the accompanying graph, it becomes apparent that both the actual and predicted values overlap closely. This alignment of the red (actual) and green (predicted) lines in the graph indicates that the optimization algorithms have contributed to enhancing the performance of the model. Consequently, the optimized model demonstrates improved accuracy and reliability in predicting the manufacturing quantities of medicines. Figure 4 shows polynomial optimized linear regression

The model achieved a notable decrease in the sum of squared errors (SSE) for each trial after implementing the scale-down transform. The absolute Z-score was calculated to be 5.511, indicating a strong effect size of 0.779. Additionally, the p-value obtained from the Wilcoxon signed-rank test was found to be less than 0.001, signifying statistical significance (Table 1). Due to the violation of assumptions required for the dependent Student's t-test, including the presence of statistical outliers, homoscedasticity, and normal distribution according to the Shapiro-Wilk test results (Table 1), we opted for a non-parametric alternative.

Table 1: Optimization Model Statistics: Normality testing and Wilcoxon signed-rank test Case Processing Summary

	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Pre-Optimization SSE	40	100.0%	0	0.0 %	40	100.0%
Post-Optimization SSE	40	100.0%	0	0.0%	40	100.0%

The graph above illustrates a comparison between the pre and post optimization Regression algorithms. On the x- axis, the algorithm names are denoted as "PRE" and "POST" optimization Regression algorithm, while the y- axis represents the SSE error. Notably, the SSE error is lower for the POST Optimization Regression algorithm compared to the pre-optimization counterpart.

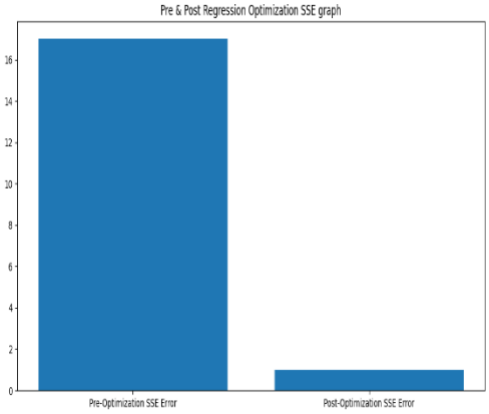


Figure 5: Pre & Post Optimization SSE Graph

This visualization clearly demonstrates that the post- optimization Regression algorithm outperforms the pre- optimization Regression algorithm in terms of SSE error. Therefore, it can be inferred that the optimization techniques applied to the Regression algorithm have led to a notable improvement in predictive accuracy and effectiveness. Figure 5 shows pre & post optimization sse graph

4 Discussion

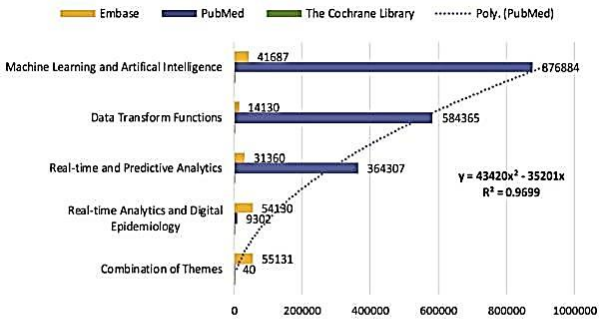
The optimization model proposed in our study holds considerable promise for addressing the complex challenges associated with linear and polynomial model analyses in various domains, particularly in anatomical sciences, dermatology, and medical research and practice. As depicted in Figure 1, the application of boosted regression models has wide-ranging implications, extending to emerging fields such as addiction neuroscience and behavioral psychiatry, particularly in the study of psychoactive and novel psychoactive substances.

In the context of machine learning and artificial intelligence, optimized regression analytics play a pivotal role, especially in applications requiring extensive data analytics and bioinformatics. From comprehensive genomic analyses to extracting insights from large-scale datasets like Google Trends and Google Analytics, the significance of optimized linear and polynomial models cannot be overstated. These models not only strengthen hypothesis testing but also streamline computational processing and reduce the resources needed for real-time and predictive analyses.

In addition, our optimization model has great promise when combined with future quantum computing developments, which should lead to computational efficiency and analytics that are

previously unimaginable.

Machine learning, encompassing a myriad of mathematical and data science models, is indispensable for extracting meaningful insights from big data. Techniques such as artificial neural networks, regression analysis, and decision trees play crucial roles in achieving the lowest achievable error rates in predictive analytics for causal associations.



	Combination of Themes	Real-time Analytics and Digital Epidemiology	Real-time and Predictive Analytics	Data Transform Functions	Machine Learning and Artifical Intelligence
Embase	55131	54130	31360	14130	41687
PubMed	40	9302	364307	584365	876884
The Cochrane Library	117	979	2599	1003	3254

Figure 6: Keywords-Based Systematic Review of the Databases of Literature

A systematic review of the literature revealed a scarcity of studies implementing our data transform method to enhance linear or polynomial regression models. Despite the existence of various approaches combining linear models with machine-learning methods, none have utilized our optimization technique. Past endeavours have explored techniques such as logistic regression, regression trees, and Fourier transform, underscoring the novelty and significance of our proposed methodology. Figure 6 shows keywords-based systematic review of the databases of literature

In summary, our study highlights the transformative potential of optimization techniques in linear and polynomial regression analytics of big data in medicine. By addressing critical challenges and offering innovative solutions, our approach paves the way for enhanced insights and advancements in medical research and practice.

5 Conclusion

Finally, when it comes to medical big data analytics using linear and polynomial regression, our suggested innovative transform and optimization approach shows considerable benefits. Through comprehensive experimentation and analysis, we have established three primary purposes for our methodology:

Reduction of Sum of Squared Errors (SSE): The foremost objective of our approach is to minimize the SSE, thereby enhancing the accuracy of the line of best fit. By effectively reducing errors, our method ensures more reliable predictions in medical data analytics.

Optimized Computational Processing: The scale-down transformation incorporated in our method plays a pivotal role in alleviating the computational processing demands associated with extensive datasets comprising numerous variables and observations. This optimization not only enhances efficiency but also facilitates the analysis of multiple polynomial regression models.

Efficient Real-Time Processing: With the proliferation of multidimensional arrays of

medical data, real-time processing of correlations and regressions poses significant computational challenges. Our optimization technique addresses this issue by transforming variables into a narrower range, thereby optimizing computational processing power and enabling seamless real-time analysis.

Moreover, our optimization method preserves the original correlation of variables while narrowing their range, thereby ensuring economic utilization of computational resources for subsequent mathematical and computational processing tasks.

In essence, our novel transform and optimization method represent a significant advancement in computational efficacy for linear and polynomial regression analytics in the field of medicine. By providing a more accurate and efficient approach to analysing big data, our methodology holds immense promise for enhancing decision-making processes and driving advancements in medical research and practice.

Acknowledgement: Not Applicable.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. J. Berro, "Essentially, all models are wrong, but some are useful—a cross-disciplinary agenda for building useful models in cell biology and biophysics," *Biophysical Reviews*, vol. 10, no. 6, pp. 1637- 1647, 2018.
2. G. E. Box, "Science and statistics," *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791-799, 1976.
3. A. I. Imam, Ahmed, "Monitoring and Analysis of Novel Psychoactive Substances in Trends Databases, Surface Web and the Deep Web, with Special Interest and Geo-Mapping of the Middle East," *Hertfordshire*, United Kingdom.
4. M. Chevreuil, R. Lebrun, A. Nouy and P. Rai, "A least- squares method for sparse low rank approximation of multivariate functions," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 3, no. 1, pp. 897-921, 2015.
5. J. Cohen, "Statistical power analysis," *Current Directions in Psychological Science*, vol. 1, no. 3, pp. 98-101, 1992.
6. R. M. Dawes and B. Corrigan, "Linear models in decision making," *Psychologic Bulletin*, vol. 81, no. 2, pp. 95, 1974.
7. A. W. F. Edwards, "Mathematizing Darwin," *Behavioral Ecology and Sociobiology*, vol. 65, no. 3, pp. 421-430, 2011.
8. B. Efron, "RA Fisher in the 21st century," *Statistical Science*, vol. 13, no. 2, pp. 95-114, 1998.
9. T. Everitt, B. Goertzel, and A. Potapov, "Artificial general intelligence," *Lecture Notes in Artificial Intelligence*, Heidelberg:Springer, 2017.
10. K.M. Fedak, A. Bernal, Z.A. Capshaw and S. Gross, "Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology," *Emerging Themes in Epidemiology*, vol. 12, no. 1, pp. 14, 2015.
11. E.H. Field, "All models are wrong, but some are useful," *Seismological Research Letters*, vol. 76, no. 2A, pp. 291-293, 2015.
12. D.A. Freedman, "Bootstrapping regression models," *The Annals of Statistics*, vol. 9, no. 6, pp. 1218-1228, 1981.
13. K. Godfrey, "Simple linear regression in medical research," *Medical Uses of Statistics*, NEJM Books, Boston, 1992.

14. T. Greenhalgh, J. Howick, and N. Maskrey, "Evidence based medicine: a movement in crisis," *The British Medical Journal*, vol. 348, 2014.
15. J.E. Grizzle, C.F. Starmer, and G.G. Koch, "Analysis of categorical data by linear models," *Biometrics*, pp. 489-504, 1969.
16. L. M. Hlavac, D. Krajcarz, I.M. Hlavacova, and S. Spadlo, "Precision comparison of analytical and statistical- regression models for AWJ cutting," *Precision Engineering*, vol. 50, pp. 148-159, 2017.
17. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255-260, 2015.
18. M. W. Lorenz, N. A. Abdi, F. Scheckenbach, A. Pflug, A. Bulbul *et al.*, "Automatic identification of variables in epidemiological datasets using logic regression," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, pp. 1-11, 2017.
19. A. Menotti, P. E. Puddu and M. Lanti, "The estimate of cardiovascular risk," *Theory, tools and problems. Annali Italiani Di Medicina Interna: Organo Ufficiale Della Societã Italiana Di Medicina Interna*, vol. 17, no. 2, pp. 81-94, 2002.
20. M. A. Motyka and A. Al-Imam, "Musical preference and drug use among youth: an empirical study," *Research and Advances in Psychiatry*, vol. 6, no. 2, pp. 50-57, 2019.
21. B. J. Norton, "Karl Pearson and statistics: The social origins of scientific innovation," *Social Studies of Science*, vol. 8, no. 1, pp. 3-34, 1978.
22. R. B. O'hara and D. J. Kotze, "Do not log-transform count data," *Methods in Ecology and Evolution*, vol. 1, no. 2, pp. 118-122, 2010.
23. C. V. Phillips and K. J. Goodman, "The missed lessons of sir Austin Bradford Hill," *Epidemiologic Perspectives & Innovations*, vol. 1, no. 1, pp. 1-5, 2004.
24. K. J. Rothman, S. Greenland and T. L. Lash, "Modern Epidemiology," *Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins*, 2008.
25. S. Schneider, "Science Fiction and Philosophy: From Time Travel to Superintelligence," *John Wiley & Sons*, 2016.
26. P. Sedgwick, "Pearson's correlation coefficient," *The British Medical Journal*, vol. 345, pp. e4483, 2012.
27. D. Himaja, V. Dondeti, S. Uppalapati *et al.*, "Cluster based active learning for classification of evolving streams," *Evol. Intel*, 2023.
28. P. V. Lal, U. Srilakshmi and D. Venkateswarlu, "MHA_VGG19: Multi-Head Attention with VGG19 Backbone Classifier-based Face Recognition for Real-Time Security Applications," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 1s, pp. 34–44, 2022.