e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:03/March-2023 Impact Factor- 7.868 www.irjmets.com

LINEAR REGRESSION COMPREHENSIVE IN MACHINE LEARNING: A SURVEY

Manisha Keer*1, Dr. Harsh Lohiya*2, Mr. Sudeesh Chouhan*3

*1Research Scholar, SSSUTMS, Sehore, Madhya Pradesh, India.

*2,3Assistant Professor, SSSUTMS, Sehore, Madhya Pradesh, India.

ABSTRACT

Perhaps one of the most common and comprehensive statistical and machine learning algorithms are linear regression. Linear regression is used to find a linear relationship between one or more predictors. The linear regression has two types: simple regression and multiple regression (MLR). This paper discusses various works by different researchers on linear regression and polynomial regression and compares their performance using the best approach to optimize prediction and precision. Almost all of the articles analyzed in this review is focused on datasets; in order to determine a model's efficiency, it must be correlated with the actual values obtained for the explanatory variables.

Keywords: Regression, Simple Linear Regression, Multiple Linear Regression, Polynomial Regression, Least Square Method.

I. INTRODUCTION

Machine learning [1-5] is commonly used in diverse fields to solve difficult problems that cannot be readily solved in based on computer approaches. The linear regression is one of the simplest and most common machine learning algorithms. It is a mathematical approach used to perform predictive analysis. Linear regression allows continuous/real or mathematical variables projections. Sir Francis Galton first suggested the concept of linear regression in 1894.Linear regression [6-8] is a mathematical test used for evaluating and quantifying the relationship between the considered variables. Univariate regression analyses (Chi-square, exact testing by Fisher and t Type equation here.testing and variance analysis (ANOVA) cannot be used to take into account the outcomes of the other covariates/founders in the analysis. Therefore, partial correlation and regression are tests which enable scientists in understanding the relationship between two variables to assess the impact of confusions [4, 9, 10].Linear regression [11] is commonly used in mathematical research methods, where it is possible to measure the predicted effects and model them against multiple input variables. It is a method of data evaluation and modeling that establishes linear relationships between variables that are dependent and independent. This method would thus model relationships between dependent variables and independent variables from the analysis and learning to the current training results. In this article, an inclusive summary of researchers' recent and most popular approaches in linear regression data processing, various statistics and machine learning over the last five years was performed. The particulars of each process are often summarized, such as used algorithms, databases, accuracy and performance[12]. This paper is organized as follows: Introduction is explained in section I. Theoretical background are presented in section II. Next, A Review and experimental Comparison are explained in section III. Discussion are explained in section IV. Section V describes conclusion.

II. THEORY

Regression

Regression [13] is a technique used for two theories. First, regression analyzes are usually used for forecasting and prediction, in which their application has major overlaps with the area of machine learning. Second, regression analysis can be used in some cases to determine causal relations between the independent and dependent variables. Importantly, regressions alone show only relations between a dependent variable and a fixed dataset collection of different variables.

Regression Models

According to the regression models, the independent variables predict the dependent variables [14]. Regression analysis estimates dependent 'y' variable value due to the range of independent variable values 'x'

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:03/March-2023 Impact Factor- 7.868 www.irjmets.com

[15]. We discuss linear regression and polynomial regression in this paper that better fits the predictive model[16]. Regression [17] may either be a simple linear regression or multiple regression.

• Simple Linear Regression

Simple Linear Regression is a case model with a single independent variable [18]. Simple Linear regression defines the dependence of the variable. y = β0 +β1x + . Simple regression distinguishes the influence of independent variables from the interaction of dependent variables[19].

• Multivariate linear regression (MLR)

MLR is a statistical technique to predict the result of an answer variable, using a number of explanatory variables. The object of (MLR) is to model the linear relationship between the independent variables x and dependent variable y that will be analyzed [20].The basic model for MLR is:

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$ The formula to determine the formula matrix is: [21] .

$\beta^{\text{ന}} = (X^T X)^{-1} X^T y$ Where

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

• Polynomial Regression

Polynomial regression [22, 23] is a type of regression analyse in the nth degree polynomial modeling of the relationship between independent and dependent variables. Polynomial regression is a special case of MLR in which the polynomial equation of data blends in with curvilinear interplay of the dependent and independent variables [24]. Model of polynomial [25, 26] is: $y = \beta_0 + \beta_1 x + \beta_2 x$

$_2 + \cdots + \beta_h x^h + \varepsilon$ Where h is named the polynomial degree [27, 28].

III. A REVIEW ON (LINEAR REGRESSION)

Xingang Wang. [37] Used MLR algorithm to calculate its weight, which eliminates redundancy between attributes, proposed a weighted naive Bayesian algorithm on the basis of the multiple regression (MLWNBC). Simultaneously, each attribute will also determine the impact size of each attribute on the basis of weight. MLWNBC makes WNBC more rational (weighted naive bias classification algorithm). The study results of which classification of 10 data sets in UCI database indicate that the algorithm has strong properties and is capable of enhancing accuracy, reducing consumption time. The data collection estimates all attributes, and certain properties have no influence on the results.

Zhihao PENG. [38] It is proposed to use a multivariate statistical method, i.e. factor analysis, to identify predictor variables by their relationships and importance, in order to approximate portfolio sensitivities to 4 chosen macroeconomic factors (Market Performance, Real GDP, Inflation, and Unemployment). (Market Performance, Real GDP, Inflation, and Unemployment). Introduces and applies a multi-factor model for portfolio management of stocks. First, the model is established, the portfolio will then be refined and multi-factors will eventually be used to estimate portfolio sensitivity. Results show that improved results can be obtained by choosing the less associated variables.

Hyun-il Lim. [11]a framework has been developed for the use of linear regression in the evaluation of software features defined applications using code vectors based on software instructions. Experiments have been conducted to test the suggested method, although experimental findings suggest that linear regression can be an efficient way to classify related software in software analysis. To conclude, a well-designed machine learning model can be easily used in software analysis. The use of machine learning in information analysis would also enhance comprehension of software functionality.

Qingxiang Feng. [39] Proposed enter-based weighted kernel linear regression (CWKLR) classifier is proposed for objects and face recognition. The middle of each class is used in CWKLR to provide information. CWKLR can then use the Tikhonov Matrix to achieve weighted classification projection coefficients. Experimental results show that, relative to many state-of-the-art approaches, the classifier proposed achieves improved efficiency,

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:03/March-2023 Impact Factor- 7.868 www.irjmets.com

analyzes and preliminary findings in three datasets indicate the efficacy and face detection of the suggested algorithms for artifacts.

Xuan Feng. [40] Centered on the 110KV high voltage switchgear contact temperature results. Using the Map Reduce model, the temperature regression model is developed by MLR models to analysis and process the monitoring point data. The effects of the estimation are evaluated by the F regression criterion. The results show that the longitudinal regression in the MLR could well be appropriate for the long-term tempering estimation for the high-voltage switchgear communication with a slight variance. The inference is that the longitudinal regression of MLR has a high precision in long-term temperature prediction.

Tadahiko Maeda. [41] Automatic design software for human-equivalent phantoms with linear and exponential regression analyzes was proposed to increase the production performance of human-equivalent phantoms for antenna calculation. The components of the human phantoms are water, silicone emulsion, glycerin, sodium chloride and agar. The software uses MLR and exponential regression analyses to create compositions that target the target fantasy. The article describes the findings of measurements for brain dreams and mind dreams developed with the software as examples of the new software. Fabricated phantoms show that it takes an additional 9% brain fantasy value and 13% stomach fantasy value to get closer to the real world. It is confirmed.

Zhaobin Zhang. [20] Proposed new approach for intra-coding video based on MLR. The proposed method uses a linear regression model to learn from end-to-end projections and the best intra-predictive block. The technology is developed in the realm of pixels rather than physical space. A separate model is qualified to optimize the model by using intra-prediction. The clean and succinct style but also delivers promising results. A suggestion is implemented into the HEVC reference program, outperforming a matched anchor. These findings offer valuable information for video coding in the future. The experimental findings indicated the reliability of the proposed system and provided important insights into how classical video coding algorithms could be further manipulated.

Ethan C. Jackson. [42] They contrast two contemporary methods in task-based functional magnetic resonance imaging (fMRI) for a MLR: linear Regression with ridge regularization and nonlinear Symbolic Regression by genetic programming. The data for this project reflect an experimental fMRI framework for visual stimulation. For 10 topics, linear and non-linear models were developed, with a further 4 refused for validation. Model consistency is measured by comparing R values (Pearson product-moment correlation) in different contexts, including single run self-compatibility, generalization of the subjects and generalization between subjects. The suitability for modeling overfit strategies is determined with a separate resting state scan. The findings show that neither approach is necessarily or statistically superior to the other.

R. Harimurti. [43] The article focuses on the processing of educational data to predict the psychomotor domain of students. In this case, the method of linear regression is used. Four regularizations were used during this point, namely: no regularization, ridge regression, lasso regression and elastic net regression. In comparison, utilizing as an appraisal tool two sampling methods: cross-validation sampling and random sampling as examples. The experimental result shows that an elastic net regression is the best regularization for cross validation and random sampling, as this regularization yields the lowest predictive error. For cross-validation, MSE, RMSE and MAE values are respectively 40.079, 6.330 and 5.183. In comparison, for random sampling, the MSE, RMSE and MAE values are 86.910, 8.428 and 6.511 respectively.

Yanming Yang, [44] I worked on a statistical model and used MLR. The MLR analyses interval projections. A MLR model has been developed which forecasts airplane material consumption. Based upon a study of the cases, fitness test, t-testing and residual tests, and a detailed and reliable regression model have been validated and evaluated. The model indicates the use of aero-material replacement parts is permanent and successful. The results provide a realistic and analytic estimate of aero material consumption.

Dejian Wei. [45] MLR methods are used to quantify the details in the simulated world on Chinese medicine bone setting manipulation. A linear regression is used to predict the content of abstract knowledge from the manipulation. We model the displacement and angle knowledge of bone manipulation. Both medications and physical exercise helps accomplish the bone settling and bone movement mechanism. A true and science forum

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:03/March-2023 Impact Factor- 7.868 www.irjmets.com

assists students in understanding bone setting manipulation and practicing bone setting manipulation. There are efforts to enhance the education level, treatment standard and heritage of bone setting in Chinese medicine. A linear regression analysis can be used to determine the strength of the relationship between a treatment and its effect on a dependent variable.

Sreehari. [46] A described article explains MLR rainfall prediction. It will help farmers determine crop yields. Floods or droughts can be evaluated together at the same time. The MLR technique was implemented in the Andhra Pradesh district of Nellore. Our analysis is designed to take advantage of the relevant rainfall findings as a basic linear regression. The researchers have applied the MLR method, estimate the values, and the MLR error rate much less simple linear regression. MLR are stronger than just linear regressions. Vapor pressure is a dependent variable, others are independent variables, and MLR is applied.

Gopalakrishnan T. [47] Worked to evaluate the sales of a big store and estimate future sales in order to maximize their revenues and make their brands much better and more competitive by generating customer loyalty. The technology used for revenue prediction is the Deep Learning Linear Regression Algorithm. The revenue figures are from 2011-2013 and the data are expected for 2014. In addition, real-time 2014 data are taken and real 2014 data are compared with the expected data to measure the predictability. They took data for 2014 and compared it to their estimated sales volume and found our projections 84% accurate, which is very similar indeed.

Luminto. [48] Worked on MLR model to predict the rice cultivation time and the result showed highest farmer's exchange rate. The weather data shall be obtained using National Statistical Authority's weather forecast and Farmer Exchange Rate data, and the obtained data will be used to construct a regression model using MLR to detect the weather FR association. The factors are "Average Temperature," "Average Moisture," "Rainfall" and "Radiation from Solar." Their effects are estimated, but their effects are mainly derived from the other factors. Prediction can be achieved by checking all the combinations of variables that cause a low RMSE value. This then is seen by the line diagram. Evaluations show a cumulative RMSE of 0.39 – 1.34 in a suggested study in two separate regions.

Dehua WANG. [49] The average measurement data processing technique was introduced by EXCEL to draw the color readings and content concentration dispersion diagram using a linear regression analysis system to calculate linear regression equation levels for color readings and material model. We use the least square approach to derive the regression equation and we use the cumulative square sum, residual square sum and regression sum and model error to evaluate model errors.

Timur Bakibayev, [50] Proposed an algorithm for processing spatial co-ordinates using polynomial regression to measure the movement's common behaviour. The key benefit of this algorithm is that a trajectory map is usable in every area. Used Python programming language for machine learning and viewing along with Science-Learning and Mattplotlib libraries. At the end they attempt to forecast the movement of all points on the map over several stages.

Franc¸ois Grondin, [51] Proposed a simple 2-D approach by creating and overlaying the acoustic image with the visual field of a camera with the auditory field of an array microphone. Polynomial regression can effectively resolve non-linear video distortion using a low-cost microphone array and off-stage camera and that SVD-PHAT, a newly suggested approach for real-time analysis of sound sources can be tailored for this role. Used polynomial regression to match an acoustic picture with a simple method of calibration that needs little calculation. The findings also suggest that SVD-PHAT is effective in producing the acoustic picture in real time with a reduction of 47 compared with SRP-PHAT.

Soon-Jong Kwon, [52] Propose a method for estimating remaining useful life (RULs) by the application of the IR voltage and capacitance correlation to the regression method of the polynomials. In addition, an accelerated degradation test and the ESI test were performed using LIBs (LINixCoyMn1-x-yO2 (NCM) with separate nickel material (Ni), life properties and alternating current (AC) impedance properties. In polynomial regression analysis, the association between internal resistance (IR) voltage and the power extracted from the accelerating degradation test was applied, which revealed that the NCM LIB was projected for the remaining useful life (RUL).

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:03/March-2023 Impact Factor- 7.868 www.irjmets.com

Ismail El kafazi, [53] Proposed two ways to predict renewable energy. Wind and solar power integration and system improvement and availability assure continued output and ensure supply of necessary amount of energy. The energy from renewable sources is tailored to customer needs. Historical statistics were used to analyze energy production over time. This experiment was to show the feasibility of the power output projection from 2016-2030. Output predictions are often inaccurate because of meteorological data. A linear model suggested a polynomial model and reliability estimates. The Maxent model was the maximum R-square and modified R-square, meaning that both models were moresuitable. For applications of the output, polynomial curve fitting models are suggested. The simulation showed how the market for electricity affects the market for power.

Ahmed Al-Imam, [54] Improve the accuracy of linear regression models by, 1) remove the square root of errors. Alternative (2) would reduce the need for statistical analysis of large-scope data simulations, a lengthy list of variables, and for polynomial regression tests, for each variable. 3) Efficiently analyzing a time-series analysis of multidimensional data would result in a more complex computational burden due to the computational capacity restrictions. Techniques include non-Bayesian statistics using SPSS and MatLab. Using Excel to produce 40 experiments using SPSS to run all the statistical tests, the signed-rank test, the test used to identify statistically optimal operational procedures. Results: A downward transition significantly decreased squared errors by 5,511 units.

Shen-Chuan Tai, [18] the approach suggested is based on image self-likeness and the basic linear regression used to establish a reconstruction model adaptively to enhance the visual qualities of updated images. The findings of the tests demonstrate that the proposed approach produces finer corners and less artifacts than previous approaches and is excellent for both visual consistency and objective parameters.

H. Roopa, [14]the main purpose of this paper is to develop a diabetes data mathematical model to achieve improved rating accuracy. The evidence is provided in this research work on characteristics extraction and mathematical modelling of Pima Indian Diabetes. Extracted characteristics of the diabetes data are projected to a new space via the key component analysis, and then modeled on these newly developed features using the linear regression approach. The accuracy reached with this approach is 82.1 percent for diabetes estimation that has improved according to other current classification systems.

Suvidha Jambekar, [21] Applying data mining techniques to forecast future crop production in relation to various observed parameters during the time, such as rice, wheat and maize (1950-2013). The parameters were precipitation, medium temperature, irrigated region, area, output and yield. The MLR and random forest regression are used in this analysis (Earth). Findings indicated that multivariate adaptive retrenchment and random forest retrenchment and multilinear retrenchment and MLR retrieval were better than random forest retrieval and multilinear retrenchment for maize data collection.

IV. DISCUSSION

Most of the papers included in this review are observational studies which used linear regression models through (2016 to 2020). Table 1. presents a summary of each study selected (reviewed) in this paper , the summary includes the dataset of each paper, technique/ method, pros and cons of the used method and accuracy of the results. As given in Table 1 there are three regression models (methods), two papers (or 18%) are used SLRM, 7 papers (or 63%) are used MLRM and 2 papers (or 18%) are used polynomial regression. About (40%) used simulation to generate data to be modeled and (60%) are used the historical data sets, Fewer than (20%) of the papers predicted of future values, examined collinearity. Statistical significance testing or confidence intervals, goodness of fit were reported in all papers. The best-achieved accuracy of the reviewed papers and from those who depended on the MLRM algorithms is research [44] because the fitting of the MLRM is very good and the prediction error is small. [11] Used SLM method, the average accuracy was 90% or almost 10% of the forecasts were inaccurate Based on a test run of the regression program, there were some drawbacks to the software's predictive results. We may implement advanced machine learning algorithms, such as help vector machine and deep neural network, to solve problems. [53] Two methods were proposed for estimating energy output from renewables, using the R-square figure for the curve fitting of the polynomial method is considerably higher than the one of the linear regression method. However, the obtained regression

equation will be used in the future to analyze and study the relationship and interactions between demand and energy production using machine learning in order to model electrical transport. The accuracy value of three reviewed papers were used MLRM method [14, 37, 49] is about (82%), fitting a model with low accuracy can cause by including not significant variables, in these situations to identify candidate variables the stepwise regression can be used, because Specification of the appropriate model depends on the proper variables. In [37] it is considered that the standard MLRM is applied and the accuracy of the data is improved. Early stage regression modelling using Stepwise and Best Subsets Regression can help in model specification. The authors in [14] Discrete features (derived from sampled diabetes data) were projected to a new space using PCA, after which they were analyzed using MLRM.

V. CONCLUSION

Regression modeling is a statistical method commonly used in research, particularly for observational studies. The proper choice of regression model, the choosing and presence of model variables are the key actions which should be established and properly controlled in order to achieve valid statistical results because the unavailability or misapplication of an appropriate regression modeling may cause to inaccuracies results. This review utilized (23) papers appeared in the last 5 years on three regression models: Simple Linear Regression Model is suitable to data contains a linear relationship between two variables, a MLR Model is a linear relation between two or more independent variables a Polynomial Regression Model would be used in case of variables having a polynomial relationship. The results of this review illustrate that almost all of the provided research papers estimated the models utilizing data sets, the accuracy of the models was measured and the predictive ability of the method is really important, to measure the performance of a regression method, a comparative study has been done between predicted and sample values.

VI. REFERENCES

[1] S. Shalev-Shwartz and S. Ben-David, Understanding machine learning: From theory to algorithms: Cambridge university press, 2014.

[2] K. P. Murphy, Machine learning: a probabilistic perspective: MIT press, 2012.

[3] P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, pp. 78-87, 2012.

[4] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Machine learning and Region Growing for Breast Cancer Segmentation," in 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, pp. 88-93.

[5] Bargarai, F., Abdulazeez, A., Tiryaki, V., & Zeebaree, D. (2020). Management of Wireless Communication Systems Using Artificial Intelligence-Based Software Defined Radio.

[6] B. Akgün and Ş. G. Öğüdücü, "Streaming linear regression on Spark MLlib and MOA," in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015, pp. 1244-1247.

[7] M. H. Dehghan, F. Hamidi, and M. Salajegheh, "Study of linear regression based on least squares and fuzzy least absolutes deviations and its application in geography," in 2015 4th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), 2015, pp. 1-6.

[8] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review," Machine Learning, vol. 62, 2020.

[9] Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020). Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. IEEE Access, 8, 203097-203116..

[10] Abdulazeez, A, M. A. Sulaiman, and D. Q. Zeebaree "Evaluating Data Mining Classification Methods Performance in Internet of Things Applications," Journal of Soft Computing and Data Mining, vol. 1, pp. 11-25, 2020.

[11] H.-I. Lim, "A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software," in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 942-943.

[12] M. R. Sarkar, M. G. Rabbani, A. R. Khan, and M. M. Hossain, "Electricity demand forecasting of Rajshahi City in Bangladesh using fuzzy linear regression model," in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2015, pp. 1-3.

[13] J. Wu, C. Liu, W. Cui, and Y. Zhang, "Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression," in 2019 IEEE International Conference on Power Data Science (ICPDS), 2019, pp. 139-142.

[14] H. Roopa and T. Asha, "A linear model based on principal component analysis for disease prediction," IEEE Access, vol. 7, pp. 105314-105318, 2019.

[15] G. A. Seber and A. J. Lee, Linear regression analysis vol. 329: John Wiley & Sons, 2012.

[16] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis vol. 821: John Wiley & Sons, 2012.

[17] S. Kavitha, S. Varuna, and R. Ramya, "A comparative analysis on linear regression and support vector regression," in 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-5.

[18] Abdulazeez, A., Salim, B., Zeebaree, D., & Doghramachi, D. (2020). Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol.