Int. j. inf. tecnol. (June 2021) 13(3):1039–1046
https://doi.org/10.1007/s41870-021-00619-9

ORIGINAL RESEARCH

A forecasting method with efficient selection of variables in multivariate data sets

Pinki Sagar[1] · **Prinima Gupta**[1] · **Indu Kashyap**[1]

Received: 24 February 2020 / Accepted: 3 February 2021 / Published online: 28 February 2021

Abstract Regression is a kind of data analysis technique in which the relationship between the independent variable(x) and dependent variable(y) is modeled and for polynomial regression it is up to the nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted by E (y|x). In this paper polynomial regression analysis has been improved through efficient selection of variables that is coefficient of determination. Coefficient of determination is a square of the correlation between new predicted y values and actual y values and its values are in the range from 0 to 1. The main purpose of regression analysis is to discover the relationship among the independent and dependent variables or in other words it is an explanation of variation in one variable with another variable. In this paper, the main focus is on Multivariate data sets that have many attributes and it is not necessary that all variables are required for data analysis purposes. Using coefficient of determination (COD) irrelevant attributes get eliminated during analysis. The main objective of research is to reduce the cost of data maintenance, reduce the execution time and improve the prediction accuracy rate. COD helps in selecting suitable independent variables. It is a notch that is used in statistical analysis that assesses how well a model explains and forecasts upcoming outcomes. This method also helps in eliminating the irrelevant variables which are not required for the prediction model by this maintenance cost and size of data sets can be reduced.

✉ Pinki Sagar
pinkisagar9@gmail.com

[1] Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India

Keywords Polynomial regression · Coefficient of determination (COD) · Independent variable · And multivariate data sets

1 Introduction

Regression is an approach used for data analysis and helps in taking decisions. It is a data mining approach to predict the continuous values or range of numeric values. It is a prediction technique where the regression equation involves two variables: Unknown variable (predictor variable) and Response variable (values to predict). Manoj Kumar Gupta and Pravin Chandra [11] presented a systematic and detailed survey of different tasks and techniques of data mining. In addition, authors presented different real-life data mining applications. Authors explained the Data mining task realization and data mining techniques. Copeland, Karen [1] introduced non parametric methods which can relax assumptions on the outline of a regression method and can help to search for data which must be applicable for suitable regression function and for data sets as well. The use of these non-parametric functions with parametric techniques can yield immensely powerful data analysis tools. Eva Ostertagová [4] determined on the polynomial regression sculpt, if the relationship of two variables is curvilinear then polynomial regression is useful for prediction and characterizes the relationship between strains and drilling depth. Least square method is used to guesstimate the parameters of the model. After fitting and evaluating the model some frequent indicators are used to weigh up the truth of the regression model. Fawcett et al. [2] describes an automatic approach for fraud detection on the basis of transaction records, and the introduced system

will learn the features and generate confidence alarms for the users.

Samar Wazir, Sufyan Beg, Tanvir Ahmad [12] Proposed earlier Master Apriori algorithm which is used to measure estimated frequent Items for a combination of certain and uncertain databases with the help of UApriori for the uncertain database based on Apriori for Certain and Planned support. Researcher expanded the previous work for the uncertain database by using UApriori based on poisson and normal distribution. There is only one-time communication between sites where data is transmitted in the proposed algorithms, which decreases the overhead of communication. By using normal and synthetic databases, the scalability and efficiency of proposed algorithms were then tested. The performances were then calculated by comparing the time taken and each algorithm generated a number of frequent products.

Rimal and Almøy, [9] introduced a novel approach that can be evaluated based on various aspects of data. However, it is very limited for multiple response variables. A novel approach is used for real data and simulated data sets. Authors compared their approach with well-established prediction methods. This approach is specially designed by varying properties such as multi col-linearity, the correlation between multiple responses and position of relevant principal components of predictors. Tahani S. Gendy [6] discusses thermal formation of stabilized limited jet dispersal flames in the presence of various geometries of trick body burners which has been scientifically modeled. Two stabilizer disc burners the radial mean temperature is measured to develop and stabilize flames at multiple normalized axial distances. Stangierski, Weiss and Kaczmarek [10] compared the quality of multiple linear regression (MLR) and artificial neural network (ANN) to predict the whole quality of spreadable Gouda cheese during storage at 8 °C, 20 °C and 30 °C. The models were based on ANNs with high values of coefficients determination and lower RMSE values proved to be more accurate.

A polynomial mathematical model has been measured to learn this occurrence to find the finest connection on behalf of the new data. Least Squares regression study has been applied to guess the coefficients of the polynomial and inspect its satisfactions. In the study, it has been identified that for predictions in large data sets may cause of high cost for maintaining the data sets and it requires lot of execution time to work on large data sets. Proposed method will reduce the maintenance cost of large sets and with efficient selection of variable, which are required for prediction, it reduce the execution time and improve the prediction rates with low errors. Ramjeet Singh Yadav [13] found that the root mean square error of the sixth degree polynomial is much smaller in these six models compared to other quadratic, third degree, fourth degree, fifth degree, and exponential polynomials. Therefore, the sixth-degree polynomial regression model for COVID-2019 data analysis in India is a very good model for predicting the next 6 days. In this analysis, authors found that in the next 7 days, the sixth-degree polynomial regression models would enable Indian physicians and the government to prepare their plans. This model can be optimized for forecasting over long-term periods based on additional regression analysis studies. Apurbalal Senapati, Amitava Nag, Arunendu Mondal and Soumen Maji [14] found from the latest COVID-19 data review that the pattern of infection number per day follows linearly and then increases the exponentially. This property has been used in our prediction and the linear regression in the piece is the most suitable model for adopting this property. The experimental results indicated the superiority of the proposed scheme and that was a new approach to the COVID-19 prediction to the best of our knowledge.

Felix Schönbrodt [7] discusses the response surface analysis into psychological science and eliminates numerous problems of surrounding and introduces the concept of fit patterns, which provides the hypothetical base intended for difficult fit hypotheses with incommensurable scales. New-fangled statistical models, namely the shifted (and rotated) squared difference models and their extensions with rising ridges, extend the statistical toolbox and facilitate researchers to experiment fit hypotheses devoid of having to rely on impractical assumptions. These models have an advanced statistical authority to notice genuine fit patterns and provide easily interpret-able parameters. New hypotheses can be tested using these parameters which could be difficult or impossible to test with traditional methods. Lastly, new open-source software provides easy to use functions which hopefully make polynomial regression methodology easier to get to researchers from a wide range of scientific fields. In data mining analysis techniques various types of data sets are available like stream data, temporal data, continuous data, discrete data, spatial data etc. Few data sets consist of one independent variable and few consist of two or more independent variables. Generally, data sets are considered in three categories:

Uni-variate data sets consist of one variable.
Bi-variate data sets consist of two variables.
Multivariate data sets consist of more than two variables.

Uni-variate data is the simplest type of data set, only one variable in the data set is considered. This data set deals with information or data sets that contain a single entity. It does not focus on causes. The representation of pattern will be initiated in this kind of data and can find the assumptions using measures of central tendency like mean, mode and median. Bi-variate data sets include two dissimilar

variable quantities. The fields of bi-variate data set are quite less with result and analytic thinking. Bi-variate data sets are used to get the relationship between the two variables. The outcome relationship is involved in Bi-variate data sets, depth psychology, causes, comparison and account. Multivariate data sets are much similar to Bi-variate, but they contain more than one independent variable. Analysis in multivariate data sets is dependent on the results which are to be achieved through various algorithms and tools.

Multivariate data set consists of more than two independent variables. Generally, this data set is used for explanatory purposes. In this data set, analysis is done on the basis of two or more than one independent variables. On the basis of objectives of data analysis, various regression methods can be applied. Regression analysis, path analysis, factor analysis and multivariate analysis of variance are some of the techniques for data analysis. In Table 1 example of multivariate data set of energy consumption is shown: Humidity is recorded in every 10 min; humidity is estimated on the basis of various parameters: temperature (inside of building), windspeed, pressure (Press_m_hg), temperature outside (t_out), humidity outside (RH_out), humidity (humidity inside of building).

2 Regression components and data analysis

For prediction and analysis various types of regression techniques are used, Linear Regression (for numeric data sets), Logistic Regression (for binary data sets), Polynomial Regression, Stepwise Regression, Ridge Regression, Lasso Regression and Elastic Net Regression. Selection of prediction technique is based on the type of data sets, for example, if a data set involves logical data like 0 and 1, logistic regression is applied (Fig. 1).

All regression techniques have different levels of accuracy in predictions. These methods are regularly determined by three parameters: number of independent variables, type of dependent variables and shape of the regression line.

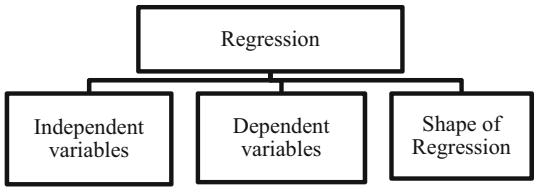


Fig. 1 Components of regression equation

2.1 Data analysis

It is the statistical standard of observations in statistics. Data analysis is a study of more than one statistical resultant variable at a time. It is a process of inspecting, cleansing, converting and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making. The aim of data analysis is to withdraw needful information from data and take the assessment based upon the data analysis. Analysis study of Multivariate data sets says that one variable is treating it as a dependent variable and others as independent variables. In data analysis process include various steps as shown in Fig. 2.

2.2 Data requirements specification

Identify necessary inputs for analysis and its types, like it should be continuous, logical or discrete.

2.3 Data collection

Process of gathering the required data or variables for targeted variables. Data collection should be accurate.

Table 1 Energy consumption multivariate data sets

Date and Time	Humidity	Temperature	T_out	Press_m_hg	RH_out	Windspeed
1/11/2016 17:00	50.91074	17.16741	6.60	733.5	92	7
1/11/2016 17:10	50.82722	17.14963	6.48	733.6	92	6.666667
1/11/2016 17:20	50.62889	17.1037	6.37	733.7	92	6.333333
1/11/2016 17:30	50.57481	17.06704	6.25	733.8	92	6
1/11/2016 17:40	50.73296	17.07074	6.13	733.9	92	5.666667
1/11/2016 17:50	50.79185	17.04852	6.02	734	92	5.333333
1/11/2016 18:00	50.78815	17.04074	5.90	734.1	92	5
1/11/2016 18:10	50.80296	17.01852	5.92	734.1667	91.83333	5.166667
1/11/2016 18:20	50.90194	17.01852	5.93	734.2333	91.66667	5.333333
1/11/2016 18:30	51.05074	17.03963	5.95	734.3	91.5	5.5
1/11/2016 18:40	51.22861	17.06676	5.97	734.3667	91.33333	5.666667

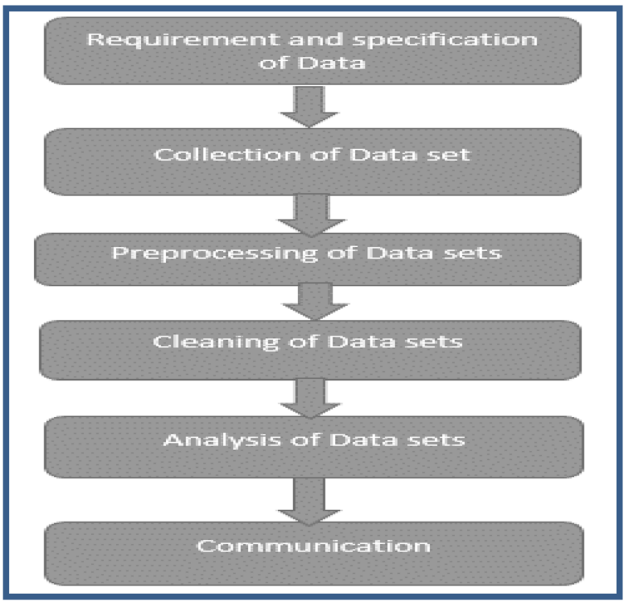


Fig. 2 Steps for data analysis

2.4 Data pre-processing

It is a method for structuring the data according to the analysis method.

2.5 Data cleaning

It is a method for detecting and avoiding the errors in data sets. Eradicate the replica values, irrelevant values, or incomplete data from data sets.

2.6 Data analysis

It is a technique to recognize, interpret, and to find conclusions that are based on the requirements for analysis.

2.7 Communication

Is the result of the data analysis, reported in a format as essential by the handlers to support their decisions and further action.

Benefits of Data analysis are following:

On the basis of changing scenarios in the market or organization, production can be increased or decreased.
Analysis of variances (ANOVA), is an analysis to help in decision making.
Data modeling is applied to reduce a large number of variables to a smaller number of variables.
To confirm a range or index by representing its constituent items load on the same factor. Drop proposed scale items which cross-load on more than one factor.

3 Forecasting method and its execution

Forecasting is an approach to do prediction on the basis of historical data, current data sets and on the basis of recent trends. In the earlier studies, it has been found that various algorithms are used to predict one dimensional and two-dimensional stream data. Various methods are used to improve the prediction accuracy rate and reduce the errors during the prediction. Sagar et al. [5] 11 introduced a prediction algorithm using regression for time series data sets to improve the performance of algorithms. Prediction method is a structure of multi regression equation up to the ordinal degree, its relationship edged by the experimental variable x and also the variable y. Polynomial regression fits a nonlinear relationship between the value of x and corresponding conditional mean of y. It is mentioned by E(y|x), and it has been accustomed to depict nonlinear phenomena that performs the calculations. Polynomial regression fits a nonlinear equation for estimation. Ostertagová [3] used a polynomial state control model developed using the sign of the relationship between the complexity and the depth of the flow. The model parameters are estimated using the least squares method. After fitting the model was evaluated using some of the most commonly indicators used to assess the accuracy of the regression model. Data was analyzed using the MATLAB computer program. Polynomial regression is taken into account as a special case of multiple linear regression.

Step 1 Find the total numbers of regression model 2^n, n is the number of independent variables. Find the ANOVAs for each regression equation shown in Table 2.

y1 = Temperature, y2 = T_out, y3 = Pressure, y4 = R_h, y5 = WindSpeed, x = Humidity

Step 2 Find the coefficient of determination and mean square error for each of the regression equations which are defined in Table 2. For example, a set of independent variables are four y1, y2, y3, y4 and coefficients are b0, b1, b2, b3, b4. All possible sets of independent variables are considered in regression equations. In Model 2, the regression equation includes one independent variable and two coefficients, and in every equation all independent variables are used with different combinations. In Model 3, all regression equations include three coefficients and two independent variables; in this model all possible sets of coefficients and independent variables are taken. Find the ANOVA for each regression equation, using ANOVA model coefficient of determination and MSE are calculated.

Table 2 The 2^n possible regression equations

Model 1	Model 2	Model 3	Model 4	Model 5
x = b0 + e	x = b0 + b1y1	x = b0 + b1y1 + b2 × 2	x = b0 + b1y1 + b2y2 + b3y3	x = b0 + b1y1 + b2y2 + b3y3 + b4y4
	x = b0 + b2y2	x = b0 + b1y1 + b3y3	x = b0 + b1 × 1 + b2y2 + b4y4	
	x = b0 + b3y3	x = b0 + b1y1 + b4y4	x = b0 + b1y1 + b3y3 + b4v4	
	x = b0 + b4y4	x = b0 + b2y2 + b3y3	x = b0 + b2y2 + b3y3 + b4y4	
		x = b0 + b2y2 + b4y4		
		x = b0 + b3y3 + b4y4		

$$\text{Coefficient ofdetermination} : ((\sum(\text{SumSq}(y1 + y2 + y3) + \text{residual})/(\sum(\text{Sum Sq}(y1 + y2 + y3) * 100))) \quad (1)$$

MSE = value of "Mean sq" corresponding to residuals in ANOVA model.

ANOVA models for regression equations shown in Table 1 are shown in Table 3a–d and in each ANOVA response is Humidity (x, dependent variable).

In each model the regression equation is selected with the highest coefficient of determination and minimum mean square error (MSE). The values of equations which have been selected (highlighted) are regression Eq. 3 from Model 2, regression Eq. 1 from Model 3, regression Eq. 1 from Model 4 and regression Eq. 1 from Model 5. By comparing all selected values of r2p and MSE in Table 4, value of Model 4 is finally selected because although r2p value 95.91 of Model 5 is higher than r2p value 95.87 of Model 4, but the MSE value 0.07 of Model 4 is lowest among all. All independent variables in selected regression Models (with high coefficients and lowest MSE) are more relevant for prediction and these variables are selected for structuring of improved prediction methods.

Step 3 Table 5 is having all selected values from Table 4. In Table 4 regression equation of Model 4 is chosen for identifying relevant variables for regression analysis. Regression Eq. 1 in Model 4 has 3 independent variables y1, y2, y3, which are most important and appropriate for the prediction model. X is a dependent variable that is converted in one column matrix as in Eq. (2). Independent variables are converted to matrices as in Eq. (3). Calculate values of inverses of x and y matrix. Matrix y' is the transpose of matrix y. Regression coefficients matrix b can be calculated as follows:

$$\mathrm{x} = \begin{bmatrix} x1 \\ x2 \\ x3 \\ . \\ . \\ xn \end{bmatrix} \quad \mathrm{b} = \begin{bmatrix} b0 \\ b1 \\ . \\ . \\ . \\ bk \end{bmatrix} \quad (2)$$

$$\mathrm{Y} = \begin{bmatrix} 1 & y1,1 & y2,2 & y1k1 \\ 1 & y2,1 & xy,2 & y2,k \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 1 & yn,1 & xn,2 & xn,k \end{bmatrix} \quad (3)$$

$$\mathrm{Y'X} = \mathrm{Y'Yb} \quad (4)$$

$$(\mathrm{Y'Y})^{-1}\mathrm{Y'Yb} = (\mathrm{Y'Y})^{-1}\mathrm{Y'Y} \quad (5)$$

$$\mathrm{b} = (\mathrm{Y'Y})^{-1}\mathrm{y'x} \quad (6)$$

$$\mathrm{X} = \mathrm{b}_0 + \mathrm{b}_1\mathrm{y}1 + \mathrm{b}_2\mathrm{y}2^2 + \mathrm{b}_3\mathrm{y}3^3 + \mathrm{b}_4\mathrm{y}4^4 + \cdots \mathrm{b}_N\mathrm{y}\mathrm{N}^N \quad (7)$$

Mean absolute error(MAE)

$$MAE = \frac{1}{n}\sum_{i=0}^{n} y - predicted\ y \quad (8)$$

4 Experimental results

Data set is collected from the UCI repository (https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction). Sample data set is presented in Table 1, humidity is dependent variable rest of variables are independent variables. Humidity (inside of a building) is to be predicted on the basis of wind speed, temperature (inside of building), temperature outside of building (t_out), pressure of wind etc. Experiments are implemented in r studio 3.3.2. In experiment general polynomial regression and improved polynomial regression are implemented. The improved method is based on selection of variables using coefficient of determination and mean square errors. In Fig. 3 blue plots represent the errors in forecasting using existing

Table 3 ANOVA for regression equations of (a) Model 2, (b) Model 3, (c) Model 4 and (d) Model 5

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
(a) Model 2					
Regression equation-1 of Model 2: x = b0 + b1y1					
y1	1	25.7695	25.7695	104.89	6.179e−09***
Residuals	18	4.4224	0.2457		
Regression equation-2 of Model 2: x = b0 + b2y2					
y2	1	3.6481	3.6481	2.4739	0.1332
Residuals	18	26.5439	1.4747		
Regression equation-3 of Model 2: x = b0 + b3y3					
y3	1	26.6034	26.6034	133.44	9.293e−10***
Residuals	18	3.5886	0.1994		
Regression equation-4 of Model 2: x = b0 + b4y4					
y4	1	1.1281	1.128	0.6986	0.4142
Residuals	18	29.0639	1.6147		
(b) Model 3					
Regression equation-1 of Model 3: x = b0 + b1y1 + b2y2					
y1	1	25.7695	25.7695	341.29	1.090e−12***
y2	1	3.1388	3.1388	41.57	6.017e−06***
Residuals	17	1.2836	0.0755		
Regression equation-2 of Model 3: x = b0 + b1y1 + b3y3					
y1	1	25.7695	25.7695	159.637	4.555e−10***
y3	1	1.6782	1.6782	10.396	0.00498**
Residuals	17	2.7442	0.1614		
Regression equation-3 of Model 3: x = b0 + b1y1 + b4y4					
y1	1	25.7695	25.7695	240.5	1.823e-11***
y4	1	2.6009	2.6009	24.273	0.0001278***
Residuals	17	1.8216	0.1072		
Regression equation-4 of Model 3: x = b0 + b2y2 + b3y3					
y2	1	3.6481	3.6481	17.32	0.0006534***
y3	1	22.9632	22.9632	109.02	8.19e−09***
Residuals	17	3.5807	0.2106		
Regression equation-5 of Model 3: x = b0 + b2y2 + b4y4					
y2	1	3.6481	3.6481	8.9029	0.008338**
y4	1	19.5779	19.57799	47.7783	2.514e−06***
Residuals	17	6.966	0.4098		
Regression equation-6 of Model 3: x = b0 + b2y3 + b4y4					
y3	1	26.6034	26.6034	128.5916	2.379e−09***
y4	1	0.0716	0.0716	0.3459	0.5642
Residuals	17	3.517	0.2069		
(c) Model 4					
Regression equation-1 of Model 4: x = b0 + b1y1 + b2y2 + b3y3					
y1	1	25.7695	25.7695	330.7466	4.122e−12***
y2	1	3.1388	3.1388	40.2861	9.690e−06***
y3	1	0.037	0.037	0.4748	0.5006
Residuals 16	16	1.2466	0.0779		
Regression equation-2 of Model 4: x = b0 + b1y1 + b2y2 + b4y4					
y1	1	25.7695	25.7695	327.7094	4.423e−12***
y2	1	3.1388	3.1388	39.9162	1.023e−05***
y4	1	0.0254	0.0254	0.3235	

Table 3 continued

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
Residuals	16	1.2582	0.0786		
Regression equation-3 of Model 4: x = b0 + b1y1 + b3y3 + b4y4					
y1	1	25.7695	25.7695	226.3686	7.301e−11***
y3	1	1.6782	1.6782	14.7418	0.001447**
y4	1	0.9228	0.9228	8.1064	0.011649*
Residuals	16	1.8214	0.1138		
Regression equation-4 of Model 4: x = b0 + b2y2 + b3y3 + b4y4					
× 2	1	3.6481	3.6481	22.3025	0.00023***
× 3	1	22.9632	22.9632	140.3845	2.473e−09***
× 4	1	0.9635	0.9635	5.8903	0.02740*
Residuals	16	2.6172	0.1636		
(d) Model 5					
Regression equation of Model 5: x = b0 + b1y1 + b2y2 + b3y3 + b4y4					
y1	1	25.7695	25.7695	313.4913	1.830e−11***
y2	1	3.1388	3.1388	38.1843	1.763e−05***
y3	1	0.037	0.037	0.4501	0.5125
y4	1	0.0136	0.0136	0.1653	0.6901
Residuals	15	1.233	0.0822		

Table 4 MSE and r2p (coefficient of determination)

Model 2		Model 3		Model 4		Model 5	
r2p (%)	MSE square	r2p (%)	MSE square	r2p (%)	MSE square	r2p (%)	MSE square
85.35	0.24	95.74	0.07	95.87	0.07	95.91	0.08
12.08	1.47	90.91	0.16	95.83	0.07		
88.11	0.19	93.96	0.11	93.96	0.11		
3.7	1.61	88.14	0.21	91.33	0.16		
		76.92	0.4				
		88.35	0.2				

Table 5 Selection of highest coefficient of determination with lowest mean square error

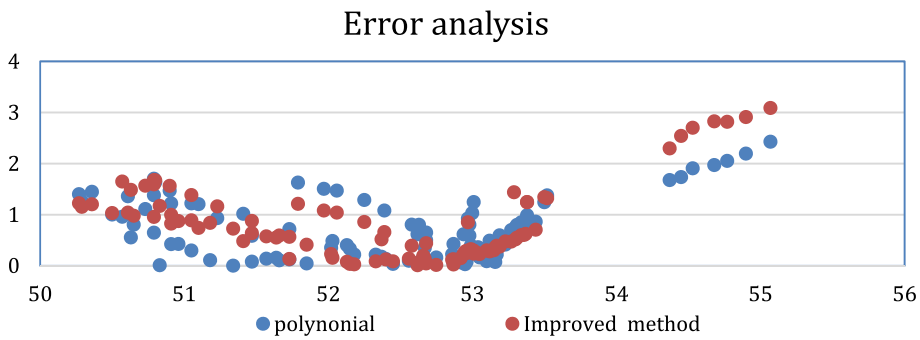
Model	r2p (%)	MSE
1	0	0.21
2	88.11	0.19
3	95.74	0.07
4	95.87	0.07
5	95.91	0.08

polynomial regression models. In this paper residuals, MAE, and average errors are analyzed.Red plots in Fig. 3 represent the errors in forecasting using enhanced polynomial regression models. By using the coefficient of determination an effective independent variable can be selected and these variables will be included in the prediction model and can reduce the errors during the prediction.

In Fig. 4 average errors in improved method are less in comparison of existing polynomial method. In Table 4 consider the coefficient of determination (must be highest) and mean square errors (must be lowest) of all prediction equations in each model. In Table 4, Model 4, regression Eq. 1 is selected with high coefficient with low mean square error.

It is corresponding to Table 2, Model 4 first regression equation "x = b0 + b1y1 + b2y2 + b3y3" presented as a forecasting model for prediction with absolute selection of variables. It means y1, y2, y3 variables are efficient models for prediction. Using appropriate selection of variables for prediction models can improve the accuracy rate of prediction (Fig. 5).

Fig. 3 Error analysis graph



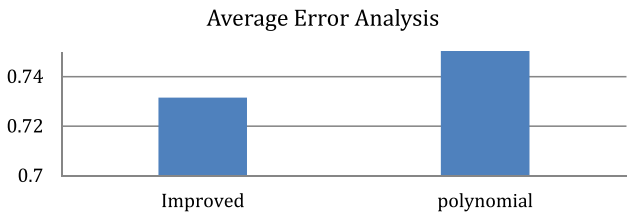


Fig. 4 Analysis of average error rate

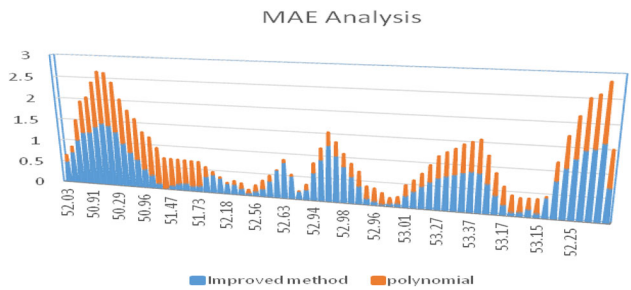


Fig. 5 Analysis of MAE

5 Conclusion and future scope

In multivariate data sets if there are many independent variables and it is not necessary that all independent variables will be included in the prediction model. Some variables have less weightage, in results using coefficient of determination appropriate independent variables are chosen to formulate an efficient prediction model with reduced error rates. Exempt the variables which are not required for forecasting. In a data set, reprocessing of variables consumes less time. Elimination of irrelevant variables from huge data sets will reduce the cost of data maintenance. In future, this improved algorithm can be applied in the area of agriculture for different zone of the states in country to estimate the production of crops on the basis water supply, temperature, uses of pesticides, humidity etc. prediction algorithms also can be used in the area production of industries, market analysis, weather forecasting, finding of disease on the basis of symptoms.

References

1. Copeland KAF (1997) Local polynomial modeling and its applications. J Qual Technol. 29:234. https://doi.org/10.1080/00224065.1997
2. Fawcett T, Provost F (1997) Adaptive fraud detection. Data Min Knowl Disc 1:291–316. https://doi.org/10.1023/A:1009700419189
3. Ostertagová E (2012) Modelling using polynomial regression. Proc Eng 48:500–506. https://doi.org/10.1016/j.proeng.2012.09.545
4. Ostertagová E (2013) Applied statistic (in Slovak), methodology and application of oneway ANOVA. Am J Mech Eng 1(7):256–261. https://doi.org/10.12691/ajme-1-7-21
5. Sagar P (2015) Regression based data mining techniques for frequent data stream (one dimensional and two dimensional stream data). Int J Comput Sci Eng 3(9):140–143. https://www.ijcseonline.org/pub_paper/27-IJCSE-012584.pdf **(E-ISSN: 2347-2693)**
6. Tahani S. GendyTaher M. El-ShiekhAmal S.Zakhary.,2015.A polynomial regression model for stabilized turbulent confined jet diffusion flames using bluff body burners Egyptian Journal of Petroleum Volume 24, Issue 4, December 2015, Pages 445–453,https://doi.org/https://doi.org/10.1016/j.ejpe.2015.06.001.
7. Schönbrodt, F. D. (2016, November 25). Testing fit patterns with polynomial regression models. https://doi.org/https://doi.org/10.31219/osf.io/ndggf.
8. Sagar P, Gupta P, Kashyap I (2018) Prediction technique for time series data sets using regression models. In: International conference on advanced informatics for computing research. https://doi.org/10.1007/978-981-13-3140-4_43
9. Rimal R, Almøy TS (2019) Comparison of multi-response prediction methods. Chemometr Intell Lab Syst 190:10–21. https://doi.org/10.1016/j.chemolab.2019.05.004
10. Stangierski J, Weiss D, Kaczmarek A (2019) Multiple regression models and Artificial Neural Network (ANN) as prediction tools of changes in overall quality during the storage of spreadable processed Gouda cheese. Eur Food Res Technol 245:2539–2547
11. Gupta MK, Chandra P (2020) A comprehensive survey of data mining. Int J Inf Technol 12:1243–1257. https://doi.org/10.1007/s41870-020-00427-7
12. Wazir S, Beg MMS, Ahmad T (2020) Comprehensive mining of frequent itemsets for a combination of certain and uncertain databases. Int J Inf Technol 12:1205–1216. https://doi.org/10.1007/s41870-019-00310-0
13. Yadav RS (2020) Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. Int J Inf Technol 12:1321–1330. https://doi.org/10.1007/s41870-020-00484-y
14. Senapati A, Nag A, Mondal A et al (2020) A novel framework for COVID-19 case prediction through piecewise regression in India. Int J Inf Technol. https://doi.org/10.1007/s41870-020-00552-3