Multimedia Systems (2022) 28:113–120
https://doi.org/10.1007/s00530-021-00798-2

REGULAR PAPER

# Predicting COVID-19 statistics using machine learning regression model: Li-MuLi-Poly

Hari Singh[1] · Seema Bawa[2]

Received: 20 March 2021 / Accepted: 17 April 2021 / Published online: 6 May 2021

## Abstract

In this paper, linear regression (LR), multi-linear regression (MLR) and polynomial regression (PR) techniques are applied to propose a model Li-MuLi-Poly. The model predicts COVID-19 deaths happening in the United States of America. The experiment was carried out on machine learning model, minimum mean square error model, and maximum likelihood ratio model. The best-fitting model was selected according to the measures of mean square error, adjusted mean square error, mean square error, root mean square error (RMSE) and maximum likelihood ratio, and the statistical *t*-test was used to verify the results. Data sets are analyzed, cleaned up and debated before being applied to the proposed regression model. The correlation of the selected independent parameters was determined by the heat map and the Carl Pearson correlation matrix. It was found that the accuracy of the LR model best-fits the dataset when all the independent parameters are used in modeling, however, RMSE and mean absolute error (MAE) are high as compared to PR models. The PR models of a high degree are required to best-fit the dataset when not much independent parameter is considered in modeling. However, the PR models of low degree best-fits the dataset when independent parameters from all dimensions are considered in modeling.

**Keywords** Machine learning · Linear regression · Polynomial regression · *t*-Test · COVID-19 · Accuracy

## 1 Introduction

The terms endemic, epidemic, outbreak, and pandemic are very closely related. An endemic is a disease that has a constant presence in a particular location or region. For example, Ice is an endemic to Antarctica and Malaria is an endemic to Africa and in some parts of India also. However, an epidemic is a disease that is localized to a region but the number of new cases of the disease spreads very fast than expected. In an epidemic the problem becomes out of control, for example, the time when the COVID-19 was limited to Wuhan city of China only, it was an epidemic. Going one step further, the endemic becomes an outbreak when the rise in number of cases of the disease is more than anticipated. If at this point the outbreak is not controlled then it becomes an endemic. When the epidemic is more geographically spread, over multiple countries or continents, then it becomes a pandemic [1].

The dataset to be used for analysis should be viewed from different angles for pre-processing. Multi-view methods can well preserve the diverse characteristics of data [2–6]. Many researchers have analyzed the spread pattern of diseases and tried to predict the impact of diseases so as to develop some policies to combat it and prevent the destruction from it. A number of statistical models are developed towards it. In this paper, mostly machine learning-based linear and polynomial regression models have been surveyed and analyzed. A non-linear regression model for modeling and forecasting the malaria disease incidence with a high confidence level and high degree of efficiency is developed [7]. The authors used three types of data, long and small-time series, and spatial data on non-linear regression analysis, and tested the models on statistical ANOVA tests. A support vector regression mechanism is applied to predict the number of COVID-19 cases and found that non-linear models, having the highest degree of non-linearity on the basis of Gaussian Kernel

Communicated by Y. Zhang.

✉ Hari Singh
hsrawat2016@gmail.com

[1] Computer Science and Engineering Department, Jaypee University of Information Technology, Solan, Waknaghat, India

[2] Computer Science and Engineering Department, Thapar University, Patiala, Punjab, India

function are good but these suffer from over-fitting of data [8]. An exponential, polynomial, and auto-regressive integrated moving averages (ARIMA) regression mechanism is used for predicting the growth of COVID-19 cases in India. The authors traced the growth of COVID-19 cases and found that it follows a power regime i.e. from exponential to quadratic and then quadratic to linear. Models were fitted using *p* values, R-Square error values, and ANOVA test, and experimentation revealed that the ARIMA models are the best one [9]. An auto-regression technique is used to improve the predictive ability of linear and multi-linear regression model for predicting the death-rate in India. However, it was found that predicted death-rate did not pass the test of statistical significance [10]. In another research work, the authors proposed a susceptible-infectious-recovered-dead (SIDR) model for estimating the growth in the COVID-19 cases that uses parameters basic reproduction number, mortality, and recovery rates on linear regression with least square as the cost function. The accuracy of the model is checked on R-Square and RMSE [11].

Some other research works based on advanced techniques have also been studied. A deep learning and artificial intelligence framework is used for categorizing the illness [12]. A long short-term memory (LSTM) based model [13] and a deep learning approach to the LSTM network [14] is used for showing the trend in infection rate and death-rate from the COVID-19 pandemic. The impact of COVID-19 is predicted for confirmed, recovered, and death cases through a linear regression model for many countries [15]. A mathematical model accounts for the parameters having an impact on the spread of COVID-19 cases and a Fourier decomposition-based non-parametric model is presented to best-fit the available data [16]. A trust-region-reflective (TRR) algorithm-based model that uses a real-time optimization technique has been presented to fit the COVID-19 data and the uncertainty of the mechanism has been quantified with LHS-PRCC coefficient test [17]. The effect of population density and climatology factors are used in modeling the COVID-19 statistics [18, 19]. In another research, the authors consider the effect of the availability of health care facilities in controlling COVID-19 cases and developed a SEIR epidemic model [20].

The statistical machine learning regression models have also been applied in a number of domains ranging from business, climate control, education and academia, sports, etc. to name a few. In another proposed approach for analyzing the pattern and relationship among dependent and independent parameters, the authors used lagged polynomial fractional regression (LPFR) which is an extension of the polynomial fractional regression (PFR). The proposed approach was proven better on the basis of R-Square error and Adjusted R-Square error metrics [21]. A polynomial regression model is applied to predict the relationship between strains and drilling depth, and parameters of the model are estimated with least-square method [22]. The market value of football players is predicted through a multiple linear regression model on the basis of physical and past performance features [23] and multiple linear regression is applied on the academic evaluation of students [24]. In another research, the authors studied COVID-19 impact on the educational system globally [25].

The work presented in this paper comes out with machine learning-based linear and polynomial regression models that best-fits the COVID-19 pandemic statistics from Johns Hopkins dataset [26]. The dataset covers COVID-19 statistics from 58 states. The prediction of the number of deaths occurring is modeled on four independent parameters using linear and polynomial regression. The experiment was carried out on a machine learning model, minimum mean square error model, and maximum likelihood ratio model. The best-fitting model was selected according to the measures of mean square error, adjusted mean square error, mean square error, RMSE, and maximum likelihood ratio, and the statistical *t*-test was used to verify the results. Data sets are analyzed, cleaned up, and debated before being applied to the proposed regression model. The correlation of the selected independent parameters was determined by the heat map and the Carl Pearson correlation matrix. The magnitude of the correlation between data is presented as colors or variation of color intensity in two dimensions. The high values in the matrix, which are in a range of 0–1 reveals that the five fields are strongly correlated and can be considered as independent parameters in fitting models.

The rest of the paper is organized as follows. Section 2 describes methods for data pre-processing and proposed models. Section 3 presents the accuracy results for linear and polynomial regression models on the basis of four accuracy evaluation metrics. Section 4 presents a discussion on the results and validates models through statistical *t*-tests. Section 5 concludes the paper.

## 2 The proposed Li-MuLi-Poly model

After preprocessing of the dataset such as data collection, analysis, cleaning, wrangling, and independent parameters selection, we propose a machine learning-based linear and polynomial regression models according to the flow diagram presented in Fig. 1. The following four accuracy metrics are used to check the models: R-Square error, Adjusted R-Square error, root mean square error (RMSE), and mean absolute error (MAE).

The following four independent parameters are used from the dataset for LR and PR and the number of deaths 'Deaths' is predicted as the dependent parameter.

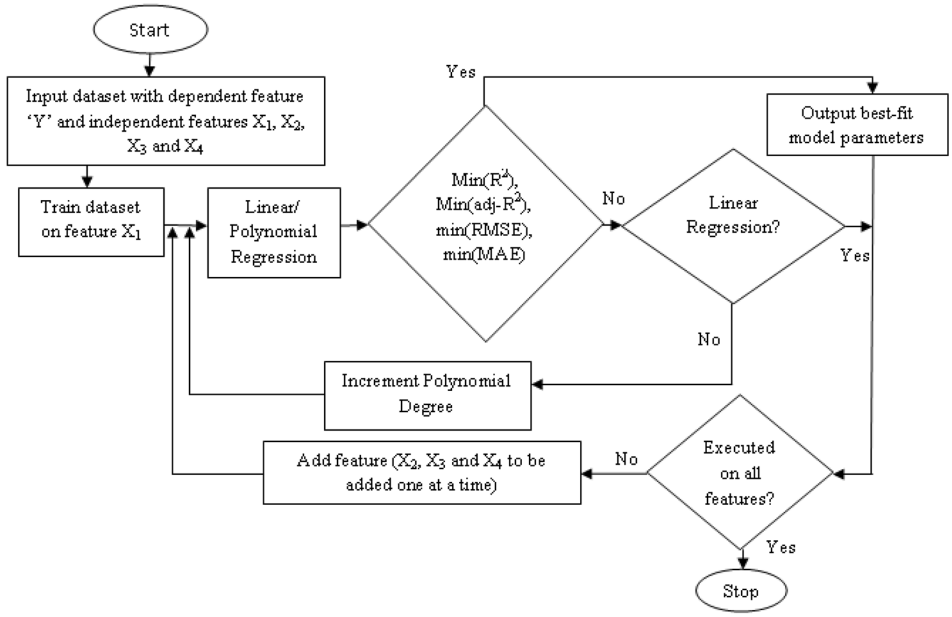
Start
Input dataset with dependent feature 'Y' and independent features X1, X2, X3 and X4
Train dataset on feature X1
Linear/ Polynomial Regression
Min(R²), Min(adj-R²), min(RMSE), min(MAE)
Yes
No
Linear Regression?
Output best-fit model parameters
Yes
No
Increment Polynomial Degree
Add feature (X2, X3 and X4 to be added one at a time)
No
Executed on all features?
Yes
Stop

**Fig. 1** Flow diagram for obtaining the proposed Li-MuLi-Poly model

$x1$ = 'Day', $x2$ = 'People_Tested', $x3$ = 'Active' and $x4$ = 'Confirmed'.

The model is trained on 80% input dataset using machine learning techniques on linear and polynomial regression models on a single parameter, two parameters, three parameters, and four parameters from the feature set '$x1$', '$x2$', '$x3$' and '$x4$'. At each stage, the models are evaluated with four evaluation metrics and, intercept and coefficients are obtained.

The input dataset is from a '$p$' dimensional real space and the output is also from a real space. The data comes from some joint distribution unknown a priori.

$$x \epsilon R^p,$$

$$y \epsilon R.$$

We try to learn a function $f(x)$ on a training sample dataset $(x_1, y_1)$, $(x_2, y_2)$, …, $(x_N, y_N)$ and validate on the test dataset.

$$f(x) : R^2 \to R,$$

$$\hat{y} = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

where '$x$' comprises of $(x_1, x_2,.., x_p)$; each corresponds to an attribute that describes the data.

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

It can also be written as:

$$f(x) = \sum_{j=0}^{p} \beta_j x_j,$$

where set $x_0 = 0$.

It is represented in vector form as:

$$f(x) = x^{\mathrm{T}}\beta,$$

$$f(x) = \begin{bmatrix} x_{01} & \cdots & x\ 0p \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Polynomial regression is used where the dataset to be fitted shows a curvilinear pattern in nature. This class helps in providing features to add a polynomial term to a simple linear regression model. Then an object of the class is created that helps in transforming matrix of features into a new matrix of features. This new matrix of features contains independent parameters like $x$, $x^2$ which represents additional polynomial terms. In other words, the transformation converts a parameter '$x$' into new matrix that contains additional independent parameters with power 2, 3, 4, etc.

For a degree = '$n$' polynomial with one independent parameter '$x$', the general form of the equation is as:

$$\hat{y} = \beta_{0+}\beta_1 x + \beta_2 x^2 + \ldots\ldots + \beta_n x^n$$

For a degree two polynomial with two independent parameters '$x_1$' and '$x_2$', the equation predicting the number of deaths is given as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_1}^2 + \beta_4 x_1 x_2 + \beta_5 {x_2}^2$$

where $\beta_0$ is the intercept and $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$ and $\beta_5$ are coefficients. This can be generalized for a polynomial of degree three and higher.

The following equations were obtained for LR and PR with one independent parameter $x_1$:

$$\hat{y} = 44099.58 + 998.08x_1,$$

$$\hat{y} = 32469.98 + 1495.94x_1 - 3.59{x_1}^2,$$

$$\hat{y} = 20166.54 + 2568.63x_1 - 22.745{x_1}^2 + 9.15e - 2{x_1}^3,$$

$$\hat{y} = 18295.42 + 2.84782e + 3x_1 - 31.84{x_1}^2 + 0.19{x_1}^3 - 3.61742e - 4{x_1}^4,$$

$$\hat{y} = 20261.29 + 2.4821e + 3x_1 - 9.928{x_1}^2 - 2.2848e - 1{x_1}^3 + 2.9977e - 3{x_1}^4 - 9.6088e - 6{x_1}^5,$$

$$\hat{y} = 19393.42 + 2.639869e + 3x_1 - 2.8680e + 1{x_1}^2 + 3.091e - 1{x_1}^3 - 4.12e - 3{x_1}^4 + 3.528e - 5{x_1}^5 - 1.08e - 7{x_1}^6,$$

$$\hat{y} = 18929.89 + 2.870903e + 3x_1 - 4.767232e + 1{x_1}^2 + 1.064{x_1}^3 - 1.89e - 2{x_1}^4 + 1.88e - 4{x_1}^5 - 9.026e - 7{x_1}^6 + 1.63e - 9{x_1}^7,$$

$$\begin{aligned}\hat{y} = 26718.82 + 2.462e + 1x_1 + 2.3669e + 2{x_1}^2 \\ - 1.161e + 1{x_1}^3 + 2.8158e - 1{x_1}^4 - 3.87e - 3{x_1}^5 \\ + 3.0469e - 5{x_1}^6 - 1.27362e - 7{x_1}^7 + 2.189e - 10{x_1}^8.\end{aligned}$$

The following equations were obtained for LR and PR with two independent parameter $x_1$ and $x_2$:

$$\hat{y} = 34728.076 + 1.7459768e + 3x_1 - 1.4014e - 3x_2,$$

$$\hat{y} = 22801.69 + 2.7482e + 3x_1 - 1.51e - 3x_2 - 2.14e + 1x_1x_2 + 2.311\mathrm{e} - 5{x_1}^2 + 1.0e - 11{x_2}^2,$$

$$\begin{aligned}\hat{y} = 36439.88 + 1.83e - 9x_1 + 1.25e - 9x_2 - 2.52\mathrm{e} \\ - 13{x_1}^2 - 5.4\mathrm{e} - 8x_1x_2 + 8.14\mathrm{e} - 10{x_2}^2 - 3.43\mathrm{e} \\ - 11{x_1}^3 - 5.37\mathrm{e} - 6{x_1}^2x_2 + 7.0\mathrm{e} - 12{x_2}^2x_1 - 6.28\mathrm{e} - 18{x_2}^3\end{aligned}$$

The coefficients of equations with the above-mentioned one and two parameters as well as for three and four parameters were generated with a machine learning library. However, the equations with three and four parameters become very complex for polynomials.

## 3 Results measuring accuracy of the model

A number of experimental runs were performed with linear regression (LR) and polynomial regression (PR) models with varying correlated independent parameters of interest. The observations are recorded in the form of evaluation metrics obtained for the run models as presented in Table 1. Here liner regression models are evaluated on a single parameter (days) LRP1, two parameters (days and people tested) MLRP2, three parameters (days, people tested and active cases) MLRP3, and four parameters (days, people tested, active cases, and confirmed cases) MLRP4. Similarly, the polynomial regression models are also evaluated on varying parameters and varying degree PRPxDy, where *x* represents the number of parameters and *y* represents the degree of polynomial.

The best models were chosen on the basis of the accuracy evaluation metrics: R-Square error, Adjusted R-Square error, root mean square error (RMSE), and mean absolute error (MAE). A good model has the characteristics of maximizing R-Square error and Adjusted R-Square error and minimizing RMSE and MAE. Table 1 presents the linear regression (LR) and polynomial regression (PR) models evaluated on a varying number of independent parameters and varying degrees for polynomial regression.

## 4 Discussions

The results obtained for linear and polynomial regression models on the four accuracy evaluation metrics are discussed in this section.

**Table 1** Linear and polynomial regression model evaluation

| Model used | R-Square error | Adjusted R-Square error | RMSE | MAE |
|---|---|---|---|---|
| LRP1 | 0.9692 | 0.966 | 6810.99 | 5912.39 |
| PRP1D0 | −0.0689 | −0.176 | 40156.72 | 32411.64 |
| PRP1D1 | 0.9692 | 0.966 | 6810.99 | 5912.39 |
| PRP1D2 | 0.9874 | 0.982 | 4343.29 | 3690.51 |
| PRP1D3 | 0.9997 | 0.982 | 729.14 | 559.03 |
| PRP1D4 | 0.9997 | 0.982 | 655.65 | 512.85 |
| PRP1D5 | 0.9998 | 0.982 | 671.23 | 523.83 |
| PRP1D6 | 0.9997 | 0.982 | 724.43 | 545.21 |
| PRP1D7 | 0.9997 | 0.982 | 778.27 | 603.93 |
| PRP1D8 | 0.9995 | 0.982 | 862.04 | 641.77 |
| PRP1D9 | 0.9829 | 0.982 | 5076.83 | 3985.43 |
| MLRP2 | 0.9895 | 0.9894 | 3971.39 | 3374.88 |
| PRP2D2 | 0.9995 | 0.9995 | 780.61 | 659.80 |
| PRP2D3 | 0.9567 | 0.9524 | 8074.50 | 6963.00 |
| MLRP3 | 0.9922 | 0.9909 | 3430.16 | 2991.10 |
| PRP3D2 | 0.9995 | 0.9994 | 819.49 | 653.94 |
| PRP3D3 | 0.9969 | 0.9965 | 2133.56 | 1767.00 |
| MLR P4 | 0.9983 | 0.9979 | 1597.53 | 1286.83 |
| PRP4D2 | 0.9996 | 0.9995 | 720.10 | 575.82 |
| PRP4D3 | 0.9994 | 0.9992 | 941.56 | 756.04 |

### 4.1 Linear and polynomial regression model evaluation with one independent parameter

It is observed from Table 1 that polynomial regression of degree zero is of no use as it is giving negative values for R-Square error and Adjusted R-Square error. The simple LR and PR of degree = 1 are same. The adjusted R-Square error metric can be ignored for LR and PR with one parameter as this metric actually measures the impact of including a greater number of parameters on the predicted output. It can be seen here that the R-Square error metric for PR is slightly better than the simple LR. Among PR models, the R-Square error improves on moving from degree = 1 to degree = 3, and it remains almost stable on further increasing the degree of PR model up to degree = 7 and then it decreases on further increasing the degree. The values obtained for RMSE, and MAE are large in most of the observations because of the pattern of data distribution. Though the values for these parameters are very large still they can be used to compare the models.

Similar to the trend observed with R-Square error metric; RMSE and MAE metrics also follow the same pattern and the minimum values are obtained for PR of degree = 3 and 4. So, PR models of degree 3 and 4 with a single independent parameter 'Days' best-fit the dataset on the basis of the four model evaluation metrics used here. Here LRPx means LR with '*x*' number of independent parameters and PRPxDy means PR with '*x*' number of independent parameters and '*y*' degree of polynomial. It is also shown graphically in Fig. 2, where the PRP1D3 (red-fitted curve) and PRP1D4 (orange fitted-curve) best-fits the original scattered dataset though RMSE and MAE values are slightly less for PRP1D4. But RMSE and MAE values are significantly low for PR models. However, between the two polynomials then lower degree polynomials are always preferred due to its low complexity over the higher degree polynomial. So, PRP1D3

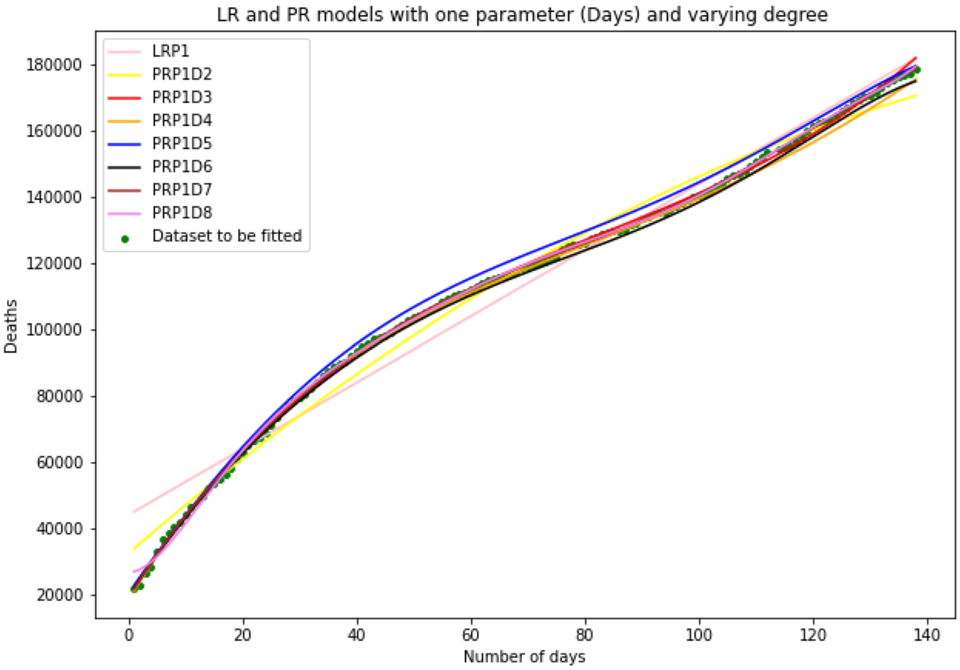

**Fig. 2** LR and PR models with one parameter (number of days) and varying degree, green curve represents the original scattered dataset, LRPx means LR with '*x*' number of independent parameters and PRPxDy means PR with '*x*' number of independent parameters and '*y*' degree of polynomial

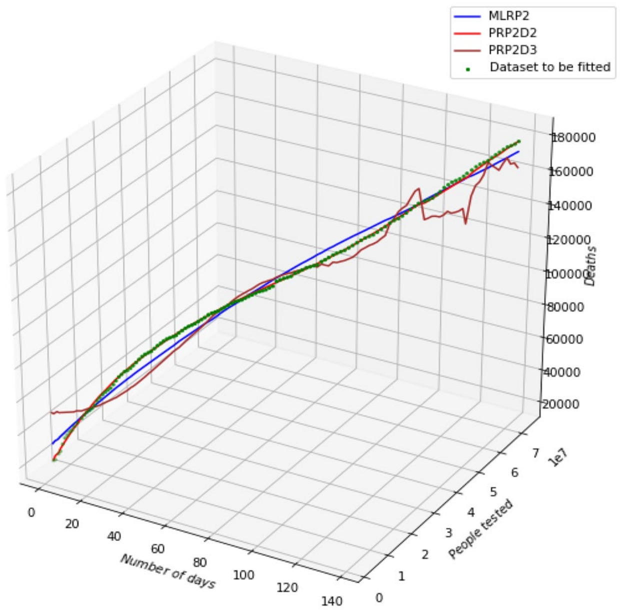

**Fig. 3** MLR and PR models with two independent parameters (days and people tested) and varying degree, green curve represents the original scattered dataset, MLRPx means MLR with '*x*' number of independent parameters and PRPxDy means PR with '*x*' number of independent parameters and '*y*' degree of polynomial

is selected as the best-fit model when a single independent parameter is considered.

## 4.2 Multi-linear and polynomial regression model evaluation with multiple independent parameters

It is evident from Table 1 that PR of degree two polynomial best-fits the dataset over the MLR and other high order polynomials because it is having maximum values for R-Square error and Adjusted R-Square error and low values for RMSE and MAE. It is also evident from the three-dimensional plot of two parameters against the predicted death-toll in Fig. 3. The green scattered plot represents the dataset to be fitted i.e., 'Days' and 'People_Tested' versus 'Deaths' occurring. Here MLRPx means MLR with '*x*' number of independent parameters and PRPxDy means PR with '*x*' number of independent parameters and '*y*' degree of polynomial. The red curve representing PRP2D2 best fits the dataset can be seen in Fig. 3.

It is also observed from Table 1 that the MLR with two parameters is better than LR with one parameter on all model evaluation metrics. The PR of degree = 2 is very close to approximating the dataset. The MLR with three parameters is even better than the MLR with two parameters and PR of degree = 2 with three parameters is better than the PR of degree = 2 with two parameters. It is also observed that the MLR with four parameters is better than all the MLR with less than four parameters as it predicts the output on the basis of all the possible dimensions. Similarly, PR of degree = 2 with four parameters is also superior to the PR of degree = 2 with three parameters. One more observation is recorded from here that PR of degree more than two are not good over the corresponding PR of degree = 2 with two, three, or four parameters. From all these observations it can be said that the dataset is closely fitted with MLR with four parameters, PR of degree = 3 with one parameter and PR of degree = 2 with three and four parameters.

**Table 2** Independent sample *t*-test or two sample *t*-test for the best metrics evaluated models

| Sr. | Model | *t*-test_stat | *p* value | Outcome |
|---|---|---|---|---|
| 1 | PRD3P1 | 0.0063 | 0.995 | Same distribution (fail to reject H0) |
| 2 | PRD2P3 | − 0.029 | 0.976 | Same distribution (fail to reject H0) |
| 3 | PRD2P4 | 0.008 | 0.993 | Same distribution (fail to reject H0) |
| 4 | MLRP4 | 0.066 | 0.947 | Same distribution (fail to reject H0) |

As the chosen three models are very close to each other, the equation with a lower degree or order is always preferred due to its low complexity. Moreover, the MLR model with four parameters encompasses the impact of all the four parameters in the equation. It was very difficult to present models on more than two parameters/dimensions, so it has not been presented here graphically.

## 4.3 Hypothesis testing

The *t*-value can be described in terms of variance as:

$$t\text{-value} = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

The *z*-test is more suitable for sample sizes more than 30. Sample size is related to 'degree of freedom (df)'. For *t*-test, "df = sample size − 1". Here, the independent-samples *t*-test is used to compare mean of the predicted death-toll values from the test dataset against the death-toll values predicted through the LR, MLR, and PR models.

The following two hypotheses are assumed for performing *t*-test.

$H_0$: The death-toll mean from the test dataset and the predicted death-toll mean are almost the same. There is no statistically significant difference between the samples.

$H_1$: The death-toll mean from the test dataset and the predicted death-toll mean are totally different and the equation fitting the death-toll does not closely represent the real scenario.

Here, we take one sample of size = 28 values, which is 20% of the test dataset, as the test dataset of 'Deaths' on the basis of the independent parameters from the actual dataset and another sample of the same size is taken from the predicted values of 'Deaths' for the same independent parameters. Using the SciPy library in Python, *t*-test was conducted on the PRD3P1, PRD2P3, PRD2P4 and MLRP4 regression models with a 95% level of confidence. These models claim to best-fit the dataset on the basis of the four-evaluation metrics discussed earlier. It means with '$\alpha$' = 0.05 level of significance. The results obtained are recorded in Table 2. The observations from the table suggest that for all the three models *p* value is very large than the '$\alpha$' value. Further the smaller *t*-values also claim a closely resembling test dataset and predicted dataset. It strongly validates that the predicted values obtained from these models qualifies the null hypothesis and the predicted death-toll data is quite similar to the actual death-toll values in the test dataset.

## 5 Conclusions and future scope

The experiments were carried out on machine learning linear, multiple-linear and polynomial regression models and the best-fitting models were selected according to the measures of R-Square error, Adjusted R-Square error, MSE, RMSE, and MAE metrics, and the results were validated using statistical *t*-tests. The dataset for the experimental run was taken from Johns Hopkins dataset [26]. Data sets are analyzed, cleaned up and debated before being applied to the proposed regression model. The correlation of the chosen independent parameters was ascertained through the heat-map and Karl Pearson's correlation matrix. It was found that the MLR model with four independent parameters model quite closely approximates polynomial regression of degree = 2 with three independent parameters model and the polynomial regression of degree = 3 with single independent parameter model. However, all three models clearly pass the *t*-test and validate the null hypothesis that the dataset of predicted values of a number of deaths is similar to the dataset taken for testing the regression models with a 95% level of confidence. In the future, we wish to use spatial data processing and parallel processing platforms to analyze COVID-19 datasets [27–29].

**Funding** No funding was received.

## References

1. What's the difference between a pandemic, an epidemic, endemic, and an outbreak? Intermountain Healthcare. https://intermountainhealthcare.org/blogs/topics/live-well/2020/04. Accessed 20 Jan 2021
2. Chenggang, Y., Biao, G., Yuxuan, W., Yue, G.: Deep multi-view enhancement hashing for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **43**(4), 1445–1451 (2021)
3. Chenggang, Y., Biyao, S., Hao, Z., Ruixin, N., Yongdong, Z., Feng, X.: 3D room layout estimation from a single RGB image. IEEE Trans. Multimed. **22**(11), 3014–3024 (2020)
4. Jin, X., Wah, B.W., Cheng, X., Wang, Y.: Significance and challenges of big data research. Big Data Res. **2**(2), 59–64 (2015). https://doi.org/10.1016/j.bdr.2015.01.006
5. Chenggang, Y., Zhisheng, L., Yongbing, Z., Yutao, L., Xiangyang, J., Yongdong, Z.: Depth image denoising using nuclear norm and learning graph model. ACM Trans. Multimed. Comput. Commun. Appl. **16**(4), 315–337 (2020)
6. Chenggang, Y., et al.: Task-adaptive attention for image captioning. IEEE Trans. Circuits Syst. Video Technol. **14**(8), 1–9 (2015)
7. Chatterjee, C., Sarkar, R.R.: Multi-step polynomial regression method to model and forecast malaria incidence. PLoS ONE (2009). https://doi.org/10.1371/journal.pone.0004726
8. Peng, Y., Nagata, M.H.: An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. Chaos Solitons Fractals (2020). https://doi.org/10.1016/j.chaos.2020.110055
9. Sharma, V.K., Nigam, U.: Modelling and forecasting of COVID-19 growth curve in India. Trans. Indian Nat. Acad. Eng. **5**, 697–710 (2020)
10. Ghosal, S., Sengupta, S., Majumder, M., Sinha, B.: Linear regression analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). Diabetes Metab. Syndr. Clin. Res. Rev. **14**(January), 311–315 (2020)
11. Anastassopoulou, C., Russo, L., Tsakris, A., Siettos, C.: Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS ONE **15**(3), 1–21 (2020). https://doi.org/10.1371/journal.pone.0230405
12. Saeed, S., Humayun, M.: Quantitative analysis of COVID-19 patients: a preliminary statistical result of deep learning artificial intelligence framework. In: Book Series: ICT Solutions for Improving Smart Communities in Asia, IGI Gobal, pp. 218–242 (2021)
13. Basu, S., Campbell, R.H.: Going by the numbers: learning and modeling COVID-19 disease dynamics. Chaos Solitons Fractals **138**, 110140 (2020). https://doi.org/10.1016/j.chaos.2020.110140
14. Chimmula, V.K.R., Zhang, L.: Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals (2020). https://doi.org/10.1016/j.chaos.2020.109864
15. Hoseinpour Dehkordi, A., Alizadeh, M., Derakhshan, P., Babazadeh, P., Jahandideh, A.: Understanding epidemic data and statistics: a case study of COVID-19. J. Med. Virol. **92**(7), 868–882 (2020). https://doi.org/10.1002/jmv.25885
16. Singhal, A., Singh, P., Lall, B., Joshi, S.D.: Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. Chaos Solitons Fractals **138**, 110023 (2020). https://doi.org/10.1016/j.chaos.2020.110023
17. Nabi, K.N.: Forecasting COVID-19 pandemic: a data-driven analysis. Chaos Solitons Fractals **139**, 110046 (2020). https://doi.org/10.1016/j.chaos.2020.110046
18. Behnood, A., Mohammadi Golafshani, E., Hosseini, S.M.: Determinants of the infection rate of the COVID-19 in the U.S. using ANFIS and virus optimization algorithm (VOA). Chaos Solitons

Fractals **139**, 110051 (2020). https://doi.org/10.1016/j.chaos.2020.110051

19. Malki, Z., Atlam, E.S., Hassanien, A.E., Dagnew, G., Elhosseini, M.A., Gad, I.: Association between weather data and COVID-19 pandemic predicting mortality rate: machine learning approaches. Chaos Solitons Fractals **138**, 110137 (2020). https://doi.org/10.1016/j.chaos.2020.110137
20. Çakan, S.: Dynamic analysis of a mathematical model with health care capacity for COVID-19 pandemic. Chaos Solitons Fractals (2020). https://doi.org/10.1016/j.chaos.2020.110033
21. Adesanya, K.K., Taiwo, A.I., Adedodun, A.F., Olatayo, T.O.: Modeling continuous non-linear data with lagged fractional polynomial regression. Asian J. Appl. Sci. **6**(5), 315–320 (2018). https://doi.org/10.24203/ajas.v6i5.5492
22. Ostertagová, E.: Modelling using polynomial regression. Proc. Eng. **48**, 500–506 (2012). https://doi.org/10.1016/j.proeng.2012.09.545
23. Kologlu, Y., Birinci, H., Kanalmaz, S.I., Özyılmaz, B.: A multiple linear regression approach for estimating the market value of football players in forward position. https://deepai.org/publication/a-multiple-linear-regression-approach-for-estimating-the-market-value-of-football-players-in-forward-position (2018)
24. Uyanık, G.K., Güler, N.: A study on multiple linear regression analysis. Proc. Soc. Behav. Sci. **106**, 234–240 (2013). https://doi.org/10.1016/j.sbspro.2013.12.027
25. Khalil, M.I., Humayun, M., Jhanjhi, N.Z.: COVID-19 impact on educational system globally. In: Emerging Technologies for Battling COVID-19 Applications and Innovations, vol. 324, pp. 257–269 (2021)
26. Hopskin, J.: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_daily_reports_us. https://github.com/CSSEGISandData. Accessed Feb 2021
27. Mittal, M., Singh, H., Paliwal, K., Goyal, L.M.: Efficient random data accessing in MapReduce. In: International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions), IEEE Explore, pp. 552–556 (2017)
28. Singh, H., Bawa, S.: Spatial data analysis with ArcGIS and MapReduce. In: Proceedings of International Conference on Conference Computing, Communication and Automation, IEEE Explore, pp. 45–49 (2016)
29. Singh, H., Bawa, S.: IGSIM: an integrated architecture for high performance spatial data analysis. Int. J. Comput. Sci. Inf. Secur. **14**(11), 302–309 (2016)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.