Journal of Experimental Biology and Agricultural Sciences

<http://www.jebas.org>

ISSN No. 2320 – 8694

PREDICTING COVID-19 INFECTIONS PREVALENCE USING LINEAR REGRESSION TOOL

Syed Tahir Ali Shah¹ , Abeer Iftikhar² , Muhammad Imran Khan^{3,*} , Majad Mansoor⁴,
Adeel Feroz Mirza⁴, Muhammad Bilal^{5,*}

¹Iqra University Islamabad Sector H-9/1, Islamabad 45210, Pakistan

²National Institute of Heart Disease, Rawalpindi, Pakistan

³Hefei National Lab for Physical Sciences at the Microscale and the Centers for Biomedical Engineering, University of Science and Technology of China, Hefei 230027, Anhui, P. R. China

⁴Department of Automation, University of Science and Technology of China, Hefei, Anhui, 230027, China

⁵School of Life Science and Food Engineering, Huaiyin Institute of Technology, Huaian 223003, China

Received – September 03, 2020; Revision – September 14, 2020; Accepted – October 02, 2020

Available Online – October 08, 2020

DOI: [http://dx.doi.org/10.18006/2020.8\(Spl-1-SARS-CoV-2\).S01.S08](http://dx.doi.org/10.18006/2020.8(Spl-1-SARS-CoV-2).S01.S08)

KEYWORDS

COVID-19

Prediction

Linear Regression

ABSTRACT

Given the socio-economic impact of coronavirus disease 2019 (COVID-19), it is essential to gauge the spread of this disease. Pakistan is one of the countries, which initially did not suffer from this disease. To observe prediction of COVID-19 cases in Pakistan, the study would utilize a linear regression model. By this model, we can predict the number of infected cases in Pakistan in an efficient way. Linear regression and correlation are two parameters used in the estimation of the linear relationship between various parameters. Correlation tells about the direction and strength of an intervariable linear relationship without discrimination between dependent and independent variables (daily COVID-19 infections are the independent variable, and prediction value is the dependent variable). While linear regression explains the estimations that can predict the values based on given information (number of infected and number of prediction) and consider dependent and independent variables. A scatter plot is deemed to be a useful tool in the determination of relationship strength between relative variables. A correlation coefficient is the measure of association with numerical configuration between two comparable variables that can stand between [-1, 1]. By using this linear regression model, we can predict the number of cases in Pakistan.

* Corresponding author

E-mail: imran_almani@yahoo.com (MI Khan); bilaluf@hotmail.com (M. Bilal)

Peer review under responsibility of Journal of Experimental Biology and Agricultural Sciences.

Production and Hosting by Horizon Publisher India [HPI]
(<http://www.horizonpublisherindia.in/>).
All rights reserved.

All the articles published by [Journal of Experimental Biology and Agricultural Sciences](#) are licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#) Based on a work at www.jebas.org.

1 Introduction

The novel coronavirus is responsible for the pandemic that started in December 2019 from Wuhan, which is the capital city of Hubei province in China and has rapidly prevailed over the world (Chimmula & Zhang, 2020). The World Health Organization (WHO) declared coronavirus as a global pandemic on 11 March 2020. Many countries around the world are trying to restrict the further spreading of this virus (Tomar & Gupta, 2020). COVID-19 is a third Coronavirus outbreak globally and earned the 3rd highest death rate in the history of pandemics (Bilal et al., 2020). The world's economy has stood on the verge of destruction; even the world's largest economies could not control the virus at this time. So far, there has been no treatment of the disease. The COVID-19 case in Pakistan was reported in Karachi, Sindh province, on 26th February 2020 and was confirmed by the country's Ministry of Health (Waris et al., 2020).

Multiple linear regression (Rath et al., 2020), linear regression (Ghosal et al., 2020), advance integrated moving average (Singh et al., 2020), autoregressive integrated moving average (ARIMA) (Ilie et al., 2020), data-based forecasting (Anastassopoulou et al., 2020), and short-term and long-term prediction models are successfully utilized for gauging the impact of COVID-19 spread in different countries of the world (Ahmar & del Val, 2020; Seyed Mohammad Ayyoubzadeh et al. 2020). The common purpose is to identify and plan the peak time, scale of spread, and allocate resources for prevention, control, and cure of COVID-19.

Some prediction tools may help handle the crisis in the best administrative way. As recently, we have used the SIR model to predict the expected number of infected cases and deaths (Ali Shah et al., 2020). Here, we propose a linear regression method for the prediction of the number of COVID-19 cases during the upcoming days. Based on previous data, Shah et al. (2020) have used the Regression model to predict COVID-19 in Pakistan by using the said model. Here, we will predict the number of infected and diagnosed cases and deaths in September 2020. A different Beta and Gamma values are used to predict the COVID-19.

1.1. Related work

Waris et al. (2020) have compared Pakistan with other countries having a strong economy. They analyzed the number of available hospitals for COVID-19, inadequate health facilities in Pakistan, and also discussed quarantine facilities in the country. The authors have discussed the complete scenario of Pakistan's hospital capability and economic condition. Yang et al. (2020) used artificial intelligence by using the trained SARS 2003 data to predict current COVID-19. The authors have also used the susceptible, exposed, infected, and recovered (SEIR) model to derive the COVID-19 curve and to find peaks and magnitude of

the spread of COVID-19. The study has predicted that the peak of COVID-19 in China will be in late February and a gradual decline by the end of April (Yang et al., 2020).

A data-driven long-short term memory (LSTM) has been used to predict COVID-19 in India for the next month (May-2020). The number of positive cases and recovery rates per day is estimated, utilizing LSTM methods. The study evaluated the increase and deceased of cases and predicts that the spread of the virus could be reduced if preventative measures, like lockdown and social distance, be exercised (Tomar & Gupta, 2020). Chimmula & Zhang (2020) used LSTM by employing available data and to build a time series, sequential model, and to predict and find peak COVID-19 in Canada. The study also predicted that the current pandemic in Canada would end within three months (first week of August 2020), and a small number of infection clusters may appear until the end of 2020. The study also has compared the transmission rate in Canada with other countries, like Italy and America (Chimmula & Zhang, 2020).

Ayyoubzadeh et al. (2020) collected data from Google Trend and used LSTM and linear regression to estimate the number of positive cases of COVID-19 in Iran. The authors have claimed that the prediction could be very helpful to policymakers and the health care department to allocate health resources accordingly. Bandyopadhyay & Dutta (2020) have utilized deep learning and neural networks long short-term memory and gated recurrent unit (GRU) to train the model of neural network (NN). The combination of n of LSTM and GRU provides better prediction of COVID-19. Supplementary rechecking of frontline medical workers could be a promising strategy. It is crucial to improve the quality of results by implementing the accurately designed and optimized detection process (Bandyopadhyay & Dutta, 2020).

2 Materials and Methods

All the global data is taken from the WHO daily situation report (WHO 2020), and Pakistani data is obtained from the official page of the government of Pakistan for COVID-19 (Government of Pakistan 2020). The data are downloaded in the CVS file and Jupiter notebook and analyzed with the Python 3.5 software uploaded. To obtain information related to infections and death cases by country, many researchers often use linear regression to estimate the precise average numerical values Y for a known value of X by the application of a *regression line*. If the y -intercept and slope of the regression line given, the value of X can be driven. Through these x values, values of Y can be estimated by calculation, which can be considered as estimated values.

Linear regression analyses were used to establish a linear regression model that could successfully facilitate the average numerical values prediction of COVID-19 future in Pakistan. The cumulative figure of COVID-19 cases in Pakistan as the dependent variable and plotting the regional distribution of the number of cases.

3 Results and Discussion

Figure 1 explains the machine learning classification. Machine learning has three main types: reinforcement learning, unsupervised learning, and supervised learning. The proposed regression analysis in current manuscript falls into the category of supervised learning, which is an approach of creating artificial intelligence, where the program is given labelled input data and the expected output results. Supervised learning allows you to collect data or produce a data output from the previous experience, and the linear regression model predicts a single output value using training data. Linear Regression is not just a machine learning algorithm; it plays a considerable role in statistics. In supervised learning, each input is associated with a target label. The task of the model is to understand the pattern and find the best fit line that covers each input target pair.

Figure1 Classification of machine learning

Through the model, the value of y is calculated by the calculations through linear combinations of input values of variables x . This hypothesis is written as the following equation.

$$h_{\theta}(x) = \theta_0\alpha_0 + \theta_1\alpha_1 + \theta_2\alpha_2 + \dots + \theta_n\alpha_n \quad (1)$$

Where h is calculated value obtained from the dependent variable for the known value of the independent variable (x), θ_0 is the intercept, the predicted value of h when the α_1 is 0.

θ here store the coefficients/weights of the input features x and is of the same dimensionality as x . Note that to add support for a constant term in our model, we prefix the vector x with 1. With single input and output variable, the method is called simple linear regression

Figure 2 showed the number of deaths of top countries; the number of dead is very high in the USA, Brazil, and India as compared to other countries. Figure 3 shows cases and prediction of COVID-19 for 14 days in Pakistan by using Linear Regression; regression means output is a continuous variable. This Figure showed that COVID-19 cases decreasing after hit peak level next two weeks' cases of COVID-19 reducing in Pakistan.

3.1 Confirmed Cases

Table 1 shows the number of COVID-19 confirmed cases in the top 30 countries. Confirmed cases in Pakistan are gradually decreasing day by day, but cases still increased in the USA, Brazil, and India in the top 30 countries confirmed cases, Pakistan is in 17th number

Table 2 exhibits the number of death due to COVID-19 in the top 30 countries. United State, Brazil, and Mexico are the top three countries where the number of death due to COVID-19 is very high. In Pakistan, more than six thousand deaths due to COVID-19 are reported.

Table 3 illustrates the state of daily confirmed cases in Pakistan. From this table, it can be seen that in the beginning period, there were no cases reported in Pakistan. In Pakistan, the first case was registered on 26 February 2020, and then gradually, it increased, and then the peak of Covid-19 was observed in mid of July month. Whereas, from August 2020, a gradually decreasing trend in cases was recorded.

3.3 Coefficients of regression polynomials

Regression coefficients are estimated values of undetermined parameters, and they express the relationship of predictor variable with a response. For linear regression, values that multiply the predictor are known as coefficients. For example, in equation $y = 5X + 3$, the coefficient is +5 predictor is X while 3 is constant. There is another form of linear regression known as polynomial regression, in which the relationship between dependent and independent variables is reproduced as the specific degree polynomial like n th degree. The relationship adjusted by polynomial regression between dependents variable y and independent variables x is adjusted in a nonlinear relationship.

[0.00000000e+00 -2.96412582e+00 -5.40109773e+01
2.70260730e+00 -5.11969748e-02 4.53852559e-04 -1.85262628e-06 2.81785306e-09]

Results of current study indicate variance percentages of the dependent variable that the independent variables explain collectively. R^2 measures the strength of the relationship between the models and the dependent variable on a convenient 0 – 100% scale. The value of $R^2 = 0.9987181307458601$ in this case.

4 Conclusion and future work

In this research analysis, the linear regression model is commonly used to has an estimation of the expected number of patients and predict the expected peaks of diseases. Similarly, we have used the linear regression model for the prediction of the number of COVID-19 cases in Pakistan. If the same situation prevails, according to the

Figure 2 Number of COVID-19 Death

Table 1 Top 30 Countries/Regions until 04/09/20 Deaths

S.No.	Country	Confirmed Cases	S. No.	Country	Confirmed Cases
1	US	6200518	16	Saudi Arabia	319141
2	Brazil	4091801	17	Pakistan	298025
3	India	4023179,	18	Turkey	276555
4	Russia	1011987	19	Italy	274644
5	Peru	670145	20	Iraq	252075
6	Colombia	641574	21	Germany	250283
7	South Africa	635078	22	Philippines	232072
8	Mexico	623090	23	Indonesia	187537
9	Spain	498989	24	Ukraine	134069
10	Argentina	461882	25	Canada	133141
11	Chile	418469	26	Israel	126419
12	Iran	382772	27	Qatar	119637
13	France	347267	28	Bolivia	119580
14	United Kingdom	344686	29	Ecuador	117175
15	Bangladesh	321615	30	Kazakhstan	106225

Figure 3 COVID-19 Cases and Prediction

Table 2 Top 30 Countries/Regions till 04/09/20

S. No.	Country	Deaths	S.No.	Country	Deaths
1	US	187755	16	Argentina	9623
2	Brazil	125502	17	Germany	9327
3	India	69561	18	Canada	9190
4	Mexico	66851	19	Indonesia	7832
5	United Kingdom	41626	20	Iraq	7359
6	Italy	35518	21	Ecuador	6674
7	France	30730	22	Turkey	6564
8	Spain	29418	23	Pakistan	6340
9	Peru	29405	24	Netherlands	6270
10	Iran	22044	25	Sweden	5835
11	Colombia	20618	26	Egypt	5495
12	Russia	17598	27	Bolivia	5343
13	South Africa	14678	28	China	4728
14	Chile	11494	29	Bangladesh	4412
15	Belgium	9901	30	Saudi Arabia	4015

Table 3 The data of confirmed cases in Pakistan from 22nd January 202 till 04th September 2020

Date	Cases	Date	Cases	Date	Cases	Date	Cases
22-Jan	0	21-Mar	776	19-May	45898	17-Jul	261917
23-Jan	0	22-Mar	875	20-May	48091	18-Jul	263496
24-Jan	0	23-Mar	972	21-May	50694	19-Jul	265083
25-Jan	0	24-Mar	1063	22-May	52437	20-Jul	266096
26-Jan	0	25-Mar	1201	23-May	54601	21-Jul	267428
27-Jan	0	26-Mar	1373	24-May	56349	22-Jul	269191
28-Jan	0	27-Mar	1495	25-May	57705	23-Jul	270400
29-Jan	0	28-Mar	1597	26-May	59151	24-Jul	271887
30-Jan	0	29-Mar	1717	27-May	61227	25-Jul	273113
31-Jan	0	30-Mar	1938	28-May	64028	26-Jul	273113
01-Feb	0	31-Mar	2118	29-May	66457	27-Jul	274289
02-Feb	0	01-Apr	2421	30-May	69496	28-Jul	275225
03-Feb	0	02-Apr	2686	31-May	72460	29-Jul	276288
04-Feb	0	03-Apr	2818	01-Jun	76398	30-Jul	277402
05-Feb	0	04-Apr	3157	02-Jun	80463	31-Jul	278305
06-Feb	0	05-Apr	3766	03-Jun	85264	01-Aug	278305
07-Feb	0	06-Apr	4035	04-Jun	89249	02-Aug	279699
08-Feb	0	07-Apr	4263	05-Jun	93983	03-Aug	280461
09-Feb	0	08-Apr	4489	06-Jun	98943	04-Aug	280461
10-Feb	0	09-Apr	4695	07-Jun	103671	05-Aug	281136
11-Feb	0	10-Apr	5011	08-Jun	108317	06-Aug	281863
12-Feb	0	11-Apr	5230	09-Jun	113702	07-Aug	282645
13-Feb	0	12-Apr	5496	10-Jun	119536	08-Aug	283487
14-Feb	0	13-Apr	5837	11-Jun	125933	09-Aug	284121
15-Feb	0	14-Apr	6383	12-Jun	125933	10-Aug	284660
16-Feb	0	15-Apr	6919	13-Jun	132405	11-Aug	285191
17-Feb	0	16-Apr	7025	14-Jun	144478	12-Aug	285921
18-Feb	0	17-Apr	7638	15-Jun	148921	13-Aug	286674
19-Feb	0	18-Apr	8348	16-Jun	154760	14-Aug	287300
20-Feb	0	19-Apr	8418	17-Jun	160118	15-Aug	288047
21-Feb	0	20-Apr	9565	18-Jun	165062	16-Aug	289215
22-Feb	0	21-Apr	10076	19-Jun	171666	17-Aug	289515
23-Feb	0	22-Apr	11155	20-Jun	176617	18-Aug	289832
24-Feb	0	23-Apr	11940	21-Jun	181088	19-Aug	290445
25-Feb	0	24-Apr	12723	22-Jun	185034	20-Aug	290958
26-Feb	2	25-Apr	13328	23-Jun	188926	21-Aug	291588
27-Feb	2	26-Apr	13915	24-Jun	192970	22-Aug	292174
28-Feb	4	27-Apr	14612	25-Jun	195745	23-Aug	293261
29-Feb	4	28-Apr	15525	26-Jun	198883	24-Aug	293461

Table 3 The data of confirmed cases in Pakistan from 22nd January 2020 till 04th September 2020

Date	Cases	Date	Cases	Date	Cases	Date	Cases
01-Mar	4	29-Apr	16817	27-Jun	202955	25-Aug	293711
02-Mar	5	30-Apr	18114	28-Jun	206512	26-Aug	294193
03-Mar	5	01-May	19103	29-Jun	209337	27-Aug	294638
04-Mar	5	02-May	20084	30-Jun	213470	28-Aug	295053
05-Mar	6	03-May	20941	01-Jul	217809	29-Aug	295372
06-Mar	6	04-May	22049	02-Jul	221896	30-Aug	295636
07-Mar	6	05-May	24073	03-Jul	221896	31-Aug	295849
08-Mar	6	06-May	24644	04-Jul	225283	01-Sep	296149
09-Mar	16	07-May	26435	05-Jul	231818	02-Sep	297014
10-Mar	19	08-May	28736	06-Jul	234509	03-Sep	297512
11-Mar	20	09-May	30334	07-Jul	237489	04-Sep	298025
12-Mar	28	10-May	32081	08-Jul	240848		
13-Mar	31	11-May	34336	09-Jul	243599		
14-Mar	53	12-May	35298	10-Jul	246351		
15-Mar	136	13-May	35788	11-Jul	248872		
16-Mar	236	14-May	38799	12-Jul	251625		
17-Mar	299	15-May	38799	13-Jul	253604		
18-Mar	454	16-May	40151	14-Jul	255769		
19-Mar	501	17-May	42125	15-Jul	257914		
20-Mar	730	18-May	43966	16-Jul	257914		

predictions of research, COVID-19 will cross its peak level, and in the next fourteen days, cases will be decreased gradually. The study seconds the trends shown by the data. The addition of new cases has stagnated between 29000-30000 with high testing capacity showing a plateau with increased recovery rates (>99%). It is shown that the total number of active patients will plunge in the last weeks of Sep-2020. The death toll remains very low (32/million) and well within single digits in confirmed cases during the late Sep-2020.

In the future, the researchers and policymakers should use these machine learning models for the prediction of the number of cases to evaluate the usefulness and success of various policies implemented by the governments and organizations from time to time during pandemics, epidemics and endemics. Researchers from multiple countries could compare the policies and their effectiveness to control the infection prevalence of infections and their effects on local and global communities.

Conflict of Interest

Authors would hereby like to declare that there is no conflict of interests that could possibly arise.

References

- Ahmar AS, Del Val EB (2020) SutteARIMA: Short-term forecasting method, a case: Covid-19 and stock market in Spain. *Science of The Total Environment* 138883.
- Ali Shah ST, Mansoor M, Mirza AF, Dilshad M, Khan MI, Farwa R, Khan MA, Bilal M, Iqbal HMN (2020) Predicting COVID-19 Spread in Pakistan using the SIR Model. *Journal of Pure and Applied Microbiology* 14(2): 1423-1430. 10.22207/JPAM.14.2.40
- Anastassopoulou C, Russo L, Tsakris A, Siettos C (2020) Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS one* 15(3): e0230405.
- Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, Kalhori SRN (2020) Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study. *JMIR Public Health and Surveillance* 6(2): e18828.
- Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, Kalhori SRN (2020) Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR Public Health Surveill* 6(2): e18828. 10.2196/18828

- Bandyopadhyay SK, Dutta S (2020) Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release. MedRxiv 10.1101/2020.03.25.20043505 10.1101/2020.03.25.20043505
- Bilal M, Khan MI, Nazir MS, Ahmed I, Iqbal HMN (2020) Coronaviruses and COVID-19 – Complications and Lessons Learned for the Future. *Journal of Pure and Applied Microbiology* 14(suppl 1): 725-731. 10.22207/JPAM.14.SPL1.09
- Chimmula VKR, Zhang L (2020) Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals* 135(109864). 10.1016/j.chaos.2020.109864
- Ghosal S, Sengupta S, Majumder M, Sinha B (2020) Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*. DOI: DOI: 10.1016/j.dsx.2020.03.017.
- Government Of Pakistan (2020) CORONAVIRUS DASHBOARD OF PAKISTAN. Journal. 2020(Issue), <http://covid.gov.pk>
- Ilie OD, Cojocariu RO, Ciobica A, Timofte SI, Mavroudis I, Doroftei B (2020) Forecasting the spreading of COVID-19 across nine countries from Europe, Asia, and the American continents using the arima models. *Microorganisms* 8(8): 1158.
- Rath S, Tripathy A, Tripathy AR (2020) Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14(5): 1467–1474. doi: 10.1016/j.dsx.2020.07.045.
- Singh RK, Rani M, Bhagavathula AS, Sah R, Rodriguez-Morales AJ, Kalita H, Nanda C, Sharma S, Sharma YD, Rabaan AA (2020) Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model. *JMIR public health and surveillance*. 6(2): e19115.
- Tomar A, Gupta N (2020) Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of the Total Environment* 728(138762). 10.1016/j.scitotenv.2020.138762
- Waris A, Atta UK, Ali M, Asmat A, Baset A (2020) COVID-19 outbreak: current scenario of Pakistan. *New Microbes New Infect.* 35(100681). 10.1016/j.nmni.2020.100681
- World Health Organization (2020) Coronavirus disease (COVID-19) pandemic. Journal., <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>
- Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, Liu P, Cao X, Gao Z, Mai Z, Liang J, Liu X, Li S, Li Y, Ye F, Guan W, Yang Y, Li F, Luo S, Xie Y, Liu B, Wang Z, Zhang S, Wang Y, Zhong N, He J (2020) Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease* 12(3): 165-174. 10.21037/jtd.2020.02.64.