Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 1, No. 3, 2022

Forecast of COVID-19 Epidemic Trend Based on Multiple Linear Regression Model

Yixi Zou

Shenzhen New Channel School, Shenzhen, Guangdong China.

Abstract: At present, COVID-19 is prevalent all over the world, and countries are facing severe epidemic prevention and control problems. Most countries in the world have taken corresponding measures, and how to predict the dynamics and trends of the epidemic quickly and accurately plays a key role in the global joint efforts to fight the epidemic. Therefore, according to the spread characteristics of COVID-19 epidemic, this paper first analyzes the multiple linear regression model, and then predicts the epidemic trend of COVID-19 based on the multiple linear regression model. The experimental results showed that when COVID-19 broke out, the multiple linear regression model had a high accuracy, and its forecast results were in good agreement with the cumulative confirmed cases. This shows that this algorithm is more effective than other existing algorithms. Using multiple linear regression model to predict the epidemic trend of COVID-19 has certain theoretical and practical significance. Through the realization of this model, we can grasp the development and changes of epidemic situation in various regions in time in the management of epidemic prevention and control, prepare for epidemic prevention in advance through trend forecast, control epidemic situation in time and improve epidemic prevention effect.

Keywords: Multiple regression model; COVID-19; Epidemic trend; Forecast.

1. Introduction

In 2019, a sudden epidemic swept the world. At first, many unexplained cases were found in Wuhan, Hubei Province in China, which was named Covid-19 by WHO [1]. The virus has many ways of transmission and strong infectivity. In 2020, COVID-19 outbreak broke out in Europe and the United States. Because these countries have adopted a negative policy to fight the epidemic-natural immunization, the virus spreads very fast, leading to a large-scale outbreak of the epidemic in these countries [2]. At present, COVID-19 is spreading all over the world. Since March 2020, the COVID-19 epidemic situation in China has tended to calm down, while the epidemic situation in foreign countries has become increasingly serious [3]. Therefore, to find out the development characteristics and laws of the global epidemic situation, and accurately detect the transmission trend of COVID-19 confirmed cases and suspected cases; Give forecasts on the occurrence time, stopping time and final diagnosis scale of epidemic inflection point; Formulate feasible prevention and control measures to reduce casualties and economic losses caused by the sudden outbreak of COVID-19; This is essential to effectively control the global epidemic trend [4].

With the increasing number of international Covid-19 infections, the pressure of epidemic prevention in China is increasing, and with the coming of autumn and winter, sporadic or clustered cases also appear in some areas of China [5]. At present, the work of searching for close contacts mainly depends on big data screening and registration screening of various units [6]. With the outbreak of China in 2019 to a large-scale global outbreak, many scholars have made forecasts and analyses. The basic mathematical model of infectious diseases is mainly to study the dynamics theory, transmission mode, spatial range and transmission speed of infectious diseases, so as to effectively prevent and control infectious diseases [7]. At present, COVID-19 is spreading all over the world. Many achievements have been made in the research and analysis of COVID-19 epidemic spread trend through kinetic algorithm [8]. However, the latent patients of COVID-19 have strong transmission ability, and the epidemic dynamics algorithm ignores the transmission risk caused by the latent patients. Moreover, the current detection methods have not considered the containment effect of prevention and control measures and centralized treatment on epidemic situation [9-10]. Multiple regression model is a commonly used statistical tool, which uses a set of independent variables to explain one or more dependent variables. Multiple regression analysis allows us to evaluate the influence of one factor on the explained variables while keeping other factors unchanged. Based on this, this paper predicts the epidemic trend of COVID-19 based on multiple linear regression model.

2. COVID-19 trend forecast model based on multiple linear regression model

2.1. Model building

The incubation time of COVID-19 pathogen is evenly distributed with the infection time. The longer the incubation period, the stronger the infectivity. That is, the traditional algorithm thinks that people in the incubation period have no infectivity, but COVID-19 has certain infectivity in the incubation period [11]. Moreover, the algorithm is established without external intervention, and can't deal with the changes caused by management and control. Therefore, this paper aims at improving this problem. In order to establish the related multiple linear regression model, the choice of explanatory variables is extremely critical. In communication, we usually use the relevant meaning of communication to select the variables that need to be explained. If the use requirements of quantitative communication are met, we should continue to test the significance of the variables in the model, so that the end of the goal stops at the purpose of explaining the variables [12-13]. In multiple regression analysis, the significance test of regression coefficient is to

test whether the linear relationship between each independent variable and dependent variable in the model is significant. The significance test is carried out by calculating the T-test value of each regression coefficient. In the traditional Logistic function curve, P value can measure the speed of curve change; For COVID-19, the p value in this function curve indicates the speed at which the epidemic reaches its peak. If the P value is large, the epidemic will soon reach its peak, indicating that a country has taken strong and effective measures during the epidemic. For example, the hospital is admitted quickly, centralized isolation, etc. On the contrary, it takes a long time for the epidemic to reach its peak. Therefore, the value of P can measure the efficiency of a country's measures against the epidemic, the overall ability of the society to face the epidemic, and the attitude of the masses to the epidemic.

In COVID-19 trend forecast model based on multiple linear regression model, there may be multiple variables related to a certain variable y. Studying the quantitative relationship between variable y and other variables is called multiple linear regression analysis. The mathematical model is:

$$Y_\alpha = \beta_0 + \beta_{1X\alpha 1} + \Lambda\beta_{pX\alpha p} + \beta_{X\alpha} + \varepsilon_\alpha \quad (1)$$

$$E(\varepsilon_\alpha) = 0, D(\varepsilon_\alpha) = \sigma^2 \quad (2)$$

$$\alpha = 1,2,\Lambda,n \quad (3)$$

ε_α are independent of each other, and the least square method is used to estimate parameter β, and the multiple linear regression equation can be obtained:

$$Y = b_0 + b_{1X1} + b_{2X2} + \Lambda + b_{pXp} \quad (4)$$

The test methods of multiple regression model include: determination coefficient test (R test); Regression coefficient significance test (T test); Regression significance test (F test).

In the multiple linear regression model, it comprehensively considers the factors that affect the spread of the epidemic, and selects the influencing factors according to the actual situation of the region as explanatory variables in the forecast model; According to the regression analysis, the factors of autocorrelation and multicollinearity are eliminated. Then, the correlation coefficients between the influencing factors and the transmission quantity which conform to the model are obtained. Then, the theory of communication is used to establish the forecast model, and the numerical values of each variable in the forecast model are determined by statistical software. Finally, the data are sorted and substituted into the model to get the forecast results. The significance test of regression equation is to test whether there is a significant linear correlation between all independent variables as a whole and dependent variable. The significance test of multivariate regression equation is similar to that of univariate regression equation, so it will not be repeated here. The failure of the significance test of the regression equation may be due to the omission of important influencing factors in the selection of independent variables, or the nonlinear relationship between independent variables and dependent variables, so the forecast model should be re-established.

2.2. Experimental analysis

COVID-19 trend forecast model includes all four kinds of people. All four kinds of participants in this model are considered, among which the infected person keeps infecting the susceptible person; The susceptible person becomes The Infiltrator, The Infiltrator can still infect the susceptible person, and The Infiltrator has a certain probability of self-healing; The infected person can recover after receiving treatment, and can't infect the disease again after recovery. As time goes by, everyone in this model will eventually become a convalescent. In order to verify the effectiveness of COVID-19 trend forecast model based on multiple linear regression model, this section carries out experiments. Compared with the detection results of ARIMA algorithm and original SEIR algorithm, the simulation platform is Python, TensorFlow framework and Jupyternotebook. The given initial condition is that the conversion rate is 0.2.

To ensure the rationality of the model, the following assumptions are made: all data sources are true and reliable, and there are no false reports. First of all, some of the crawled data did not conform to the actual situation of the overall epidemic forecast, so the data were cleaned and corrected. Detailed cleaning rules are as follows: (1) The user selects a specific date, and the system feeds back the overall epidemic information data and data analysis of that date. (2) The user selects the map type, and the system feeds back the selected epidemic situation map information. (3) The user selects a specific city, and the system feeds back the epidemic information of the place and predicts the epidemic trend. In this paper, the reason for adding E to the model is that the contact infected person is not infected and the contact infected person is sick. However, due to the national isolation measures, the influence of isolation is considered, which reflects the importance of the national policy. Secondly, the contact infected person is sick, and the infection rate is probable. In order to verify the fitting and detection effect of the algorithm, this paper divides the data into two parts: training set and testing set. The data of training set is used to train the algorithm, and the data of testing set is used to verify the accuracy of the algorithm and judge the detection effect of the algorithm. Figure 1 shows the training results of the algorithm. Figure 2 shows the forecast accuracy results of the algorithm.

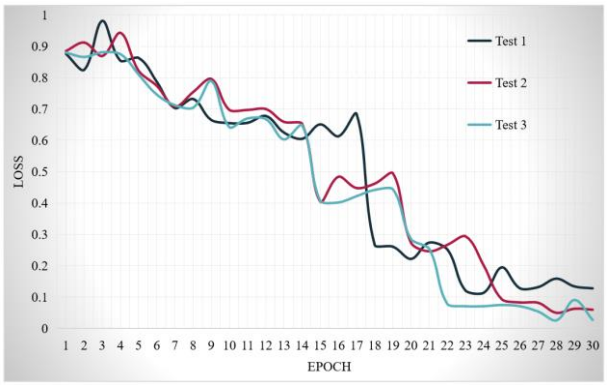

Figure 1. Training results of the algorithm

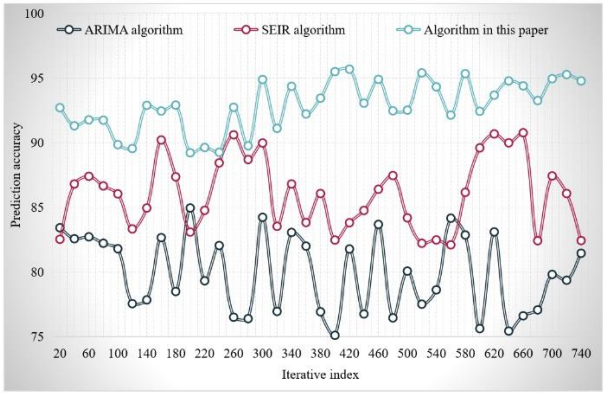

Figure 2. Forecast accuracy results of the algorithm

The accuracy and effectiveness of the multiple regression model are obviously better than that of the single regression model. Therefore, the multiple linear regression model constructed in this paper has a good effect in COVID-19 trend forecast. This forecast model has good forecast and analysis ability in the early stage of epidemic outbreak, and its accuracy rate is high. By fitting the cumulative confirmed cases and the cumulative number of deaths, a good fitting result can be obtained. Table 1 shows the comparison of experimental results of different algorithms.

Table 1. Comparison of experimental results of different algorithms

Index	ARIMA	SEIR	Paper algorithm
Optimum fitness value	0.9325	0.9235	0.9427
Training error	11.025	13.142	8.956
Forecast error	11.017	13.056	8.963
Node connection number	8	10	14

It can be seen from Table 1 that the trend of COVID-19 can be predicted well by using the forecast model after training. The accuracy of the model is high, and the predicted results are in good agreement with the cumulative confirmed cases.

3. Conclusions

At present, COVID-19 is spreading all over the world. Understand the development characteristics and laws of the global epidemic situation, and accurately detect the transmission trend of COVID-19 confirmed cases and suspected cases; Give forecasts on the occurrence time, stopping time and final diagnosis scale of epidemic inflection point; Formulate feasible prevention and control measures to reduce casualties and economic losses caused by the sudden outbreak of COVID-19; This is essential to effectively control the development trend of the global epidemic. According to the spread characteristics of COVID-19 epidemic, this paper first analyzes the multiple linear regression model, and then forecasts the epidemic trend of COVID-19 based on the multiple linear regression model. The experimental results showed that when COVID-19 broke out, the multiple linear regression model had a high accuracy, and its forecast results were in good agreement with the cumulative confirmed cases. The trained forecast model can predict the trend of COVID-19 well. This shows that this algorithm is more effective than other existing algorithms. In this paper, the multiple linear regression model was introduced into the epidemic trend forecast of COVID-19, which solved the problem of epidemic trend forecast of COVID-19. Through the realization of this model, we can grasp the development and changes of epidemic situation in various regions in time in the management of epidemic prevention and control, prepare for epidemic prevention in advance through trend forecast, control epidemic situation in time and improve epidemic prevention effect. However, this forecast method also has some shortcomings. Firstly, there may be some problems such as heteroscedasticity, autocorrelation and multicollinearity. Secondly, it is not enough to directly predict the accuracy by using the empirical regression formula generated by regression analysis. It can only be used for macro reference in decision-making. The follow-up will continue to be discussed in depth.

References

[1] Cui Hengjian, Hu Tao. Nonlinear regression method for forecasting epidemic situation in novel coronavirus [J]. China Science, 2021(051-008).

[2] Lin Tingkui, Wu Jiayuan, Liu Huafeng, et al. Forecast and analysis of the epidemic situation in novel coronavirus in western Guangdong and other prefecture-level cities-a study based on Holt's two-parameter exponential smoothing model [J]. Journal of Practical Cardiopulmonary Vascular Diseases, 2020, 28(2):5.

[3] Ding Zhongxing, Song Wenyu, Fang Xinyu, et al. Forecasting the epidemic trend of novel coronavirus in Wuhan, Hubei Province based on SEIAQR kinetic model [J]. China Health Statistics, 2020, 37(3):5.

[4] Zhang Jinfang, Niu Xiaohong, Ping Weiwei, et al. Analysis of the current situation of residents' quality of life and its influencing factors under the pressure of novel coronavirus epidemic [J]. China Folk Therapy, 2022, 30(3):4.

[5] Song Ge, Li Xiaoshan, Wang Kewei. Application of ARIMA and SVM combined model in novel coronavirus forecast [J]. chinese journal of nosocomiology, 2022, 32(1):5.

[6] Gong Wuqing, Peng Houxue, Chen Jianguo, et al. Analysis of the spatial distribution trend and related influencing factors of novel coronavirus prevalence in China [J]. Harbin Medicine, 2021, 41(1):4.

[7] Hong Bin, Chen Jinxiu, Wang Liansheng, Yu Rongshan. Analysis and forecast of novel coronavirus communication trend based on SEIR-LSTM mixed model [J]. Journal of Xiamen University: Natural Science Edition, 2020, 59(6):7.

[8] Li Zhongqi, Tao Bilin, Zhan Mengyao, et al. Comparative study on the effect of time series model applied to the forecast of epidemic situation in novel coronavirus [J]. Chinese Journal of Epidemiology, 2021, 42(3):6.

[9] Kang Guanlong, Liu Bingxiang. novel coronavirus Forecast Analysis Based on SIR Model [J]. China-Arab Science and Technology Forum (Chinese and English), 2020, 000(006):P.151-153.

[10] Zhong Deyan, Chen Lihua, Wu Ronghuo. novel coronavirus (COVID-19) epidemic forecast-based on residual autoregressive model [J]. Neijiang Science and Technology, 2021, 42(5):2.

[11] Dai Jiya, Guo Runing, Liu Guoheng. Analysis of the epidemic trend in novel coronavirus based on Joinpoint regression model [J]. Journal of Tropical Medicine, 2020, 20(10):5.

[12] Liu Zhongdian, Li Yanning. Forecast of epidemic situation in novel coronavirus, Guangxi based on ARIMA model [J]. Journal of Guangxi Medical University, 2021, 38(12):2367-2371.

[13] Cai Jie, Jia Haoyuan, Wang Ke. Based on SEIR model, the development trend of novel coronavirus epidemic in Wuhan was predicted [J]. Shandong Medicine, 2020(6):1-4.