The Spreading of Coronavirus and the Impact on China

Introduction

Predicting the situation of COVID-19 is undoubtedly of great significance to ensure that the serious situation of preventing the spread of the epidemic from accelerating and to guide our orderly resumption of production and working. In this report, in order to do a good job in predicting the epidemic situation of pneumonia infected by new coronavirus, the following specific points are made:

First of all, make a macro grasp of the prediction and analysis of the data of coronavirus, understand the calculation system and evaluation method of the COVID-19 situation; then, conduct a macro overview of each variable given by the task, and find the usual variable processing methods and correlation.

For different Types of variables (categorical variables, numerical variables, grading variables, text variables) should adopt different methods to make the data meet the requirements of regression analysis; then, perform a preliminary statistical analysis of the dependent variable and each variable to observe whether its distribution has statistical significance, Whether each statistical parameter meets the requirements of linear regression;

This is followed by the main body of the linear regression analysis task, which is helpful to grasp whether the regression model meets the statistical significance, and to use the obtained model to guide the prediction of COVID-19 situation; finally summarize the conclusions and propose feasible improvements.

According to the analysis results of the model, among all the indicators given, the three principal components of 'cumulative factor', 'new factor', and 'existing factor' have a significant impact on new cases on the second day.

Data overviewing and parameters design

According to the observation of the data structure, the variables are mainly divided into three categories:

1. **'New' variables**

   'New' variables include the three indicators of 'new diagnosed on the same day', 'new close contact on the same day', and 'new diagnosed tomorrow'. There is no doubt that the two columns of new-day additions and second-day additions are almost completely linear, so we estimate that there may be some collinearity problems in the model, and it is the new additions that have the values Only

in this way can the value in the data frame increase almost cumulatively, which can effectively reflect the epidemic situation of the day.

2. **‘Cumulative’ variables** "Cumulative" variables "cumulative diagnoses", "cumulative deaths", "cumulative cures", and "cumulative close contacts" are four indicators. Although the cumulative value will inevitably increase, due to the existence of the 'new type' variable, the cumulative number can well reflect the trend of the entire epidemic growth and reflect the significant effect after the introduction of control measures. Although due to the long incubation period of the virus, many measures will not be effective until some time after implementation, but we can still see the obvious anti-epidemic effect through multiple days of records.
3. **‘Existing’ variables** The ‘existing’ indicators include the ‘existing critical illnesses’ and ‘existing medical observations’ indicators. Existing indicators are the best indicators to observe the current epidemic situation. Since February 3, the number of newly diagnosed cases outside Hubei Province has ushered in seven consecutive declines, from 890 cases on the same day to 381 cases on February 10. In addition, newly suspected cases outside Hubei Province also reached a peak of 2,211 on the 6th and continued to drop to 1,722 on the 10th. From this set of data, in addition to the steady control of the epidemic in Hubei Province, the epidemic prevention and control data in 30 provinces and regions in China has released a clear positive signal. However, the judgment of the inflection point is largely measured and estimated based on the existing values.

As far as specific operations are concerned, this article first deals with the original data table. First, fill in the vacancies. After checking the data from all aspects for verification, or giving a reasonable data estimate based on the original data distribution, we filled in all the new cases observed on the second day (that is, the dependent variable to be predicted), and the missing cumulative discharge The number of new suspected cases was added. At this point, we have obtained the complete design matrix and dependent variable sequence.

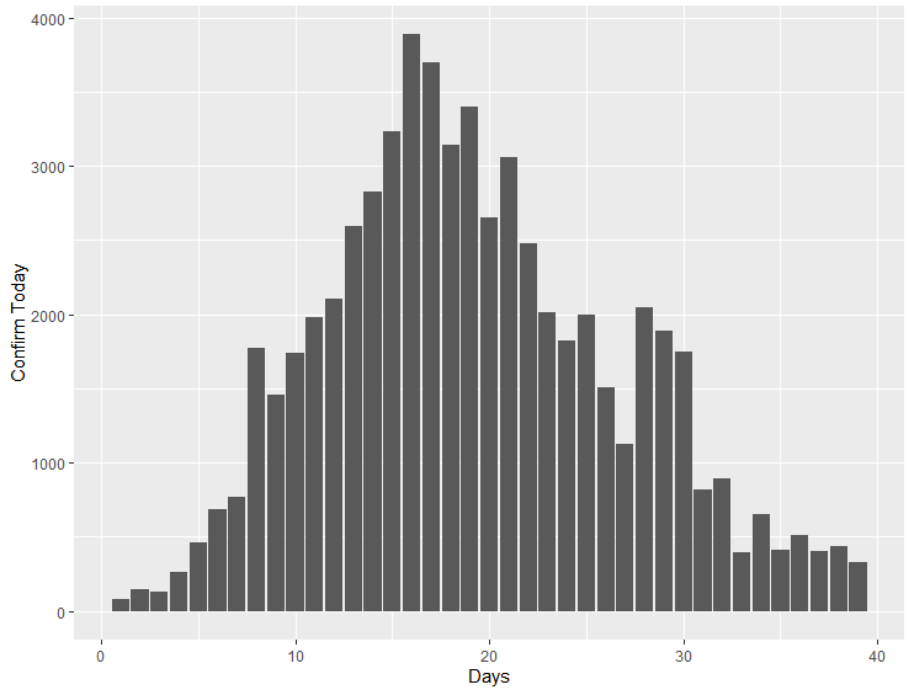
At the same time, we should realize that this article is based on the analysis of the national epidemic situation. It has not been specifically expanded to every province, city or region, nor has it analysed other countries and regions in the world.

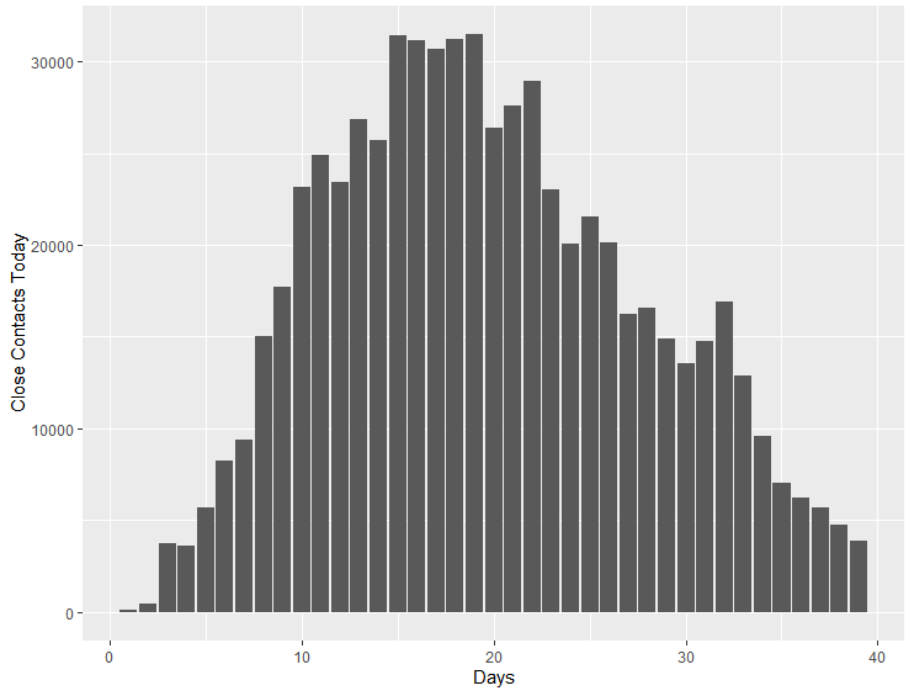
1. An epidemiological analysis of a wider area is not rigorous and precise. If you want to analyse the situation in various places, there are still many indicators that need to be included in the model.
2. The data is updated in real time and changes in real time. We should realize that for a more rigorous model, the data in different time periods of each day is also of great significance. This article simply performs regression prediction analysis.

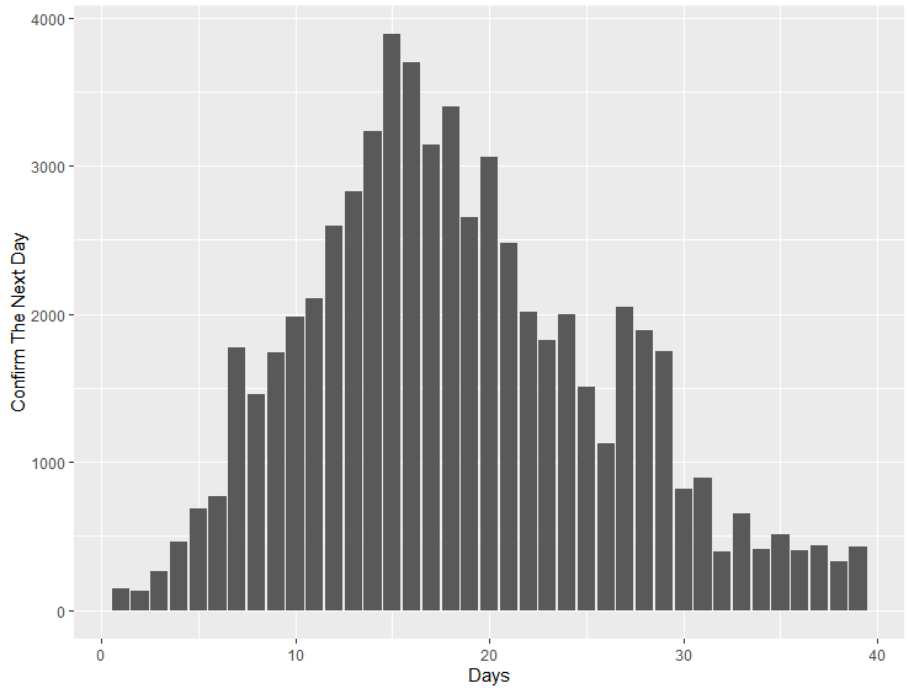
Descriptive statistical analysis

Because there are many classification indicators in the original data, based on the above indicator design, the descriptive statistical analysis in this paper directly analyses the data quality of the variable layer analysis, and initially determines the correlation between the dependent variable and each potential influencing factor for further research.

First, plot a bar graph for "new" variables:



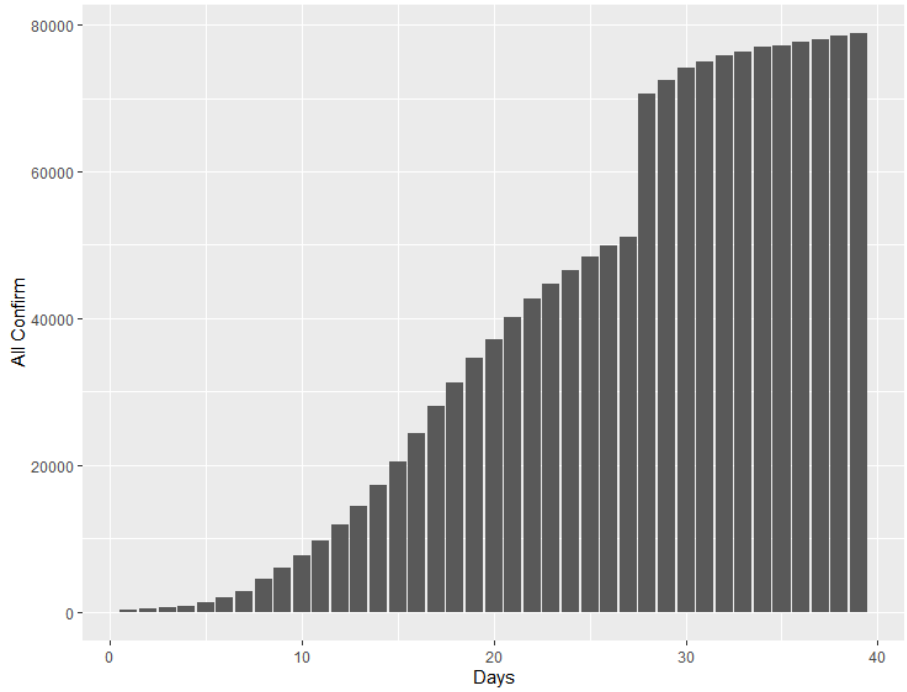




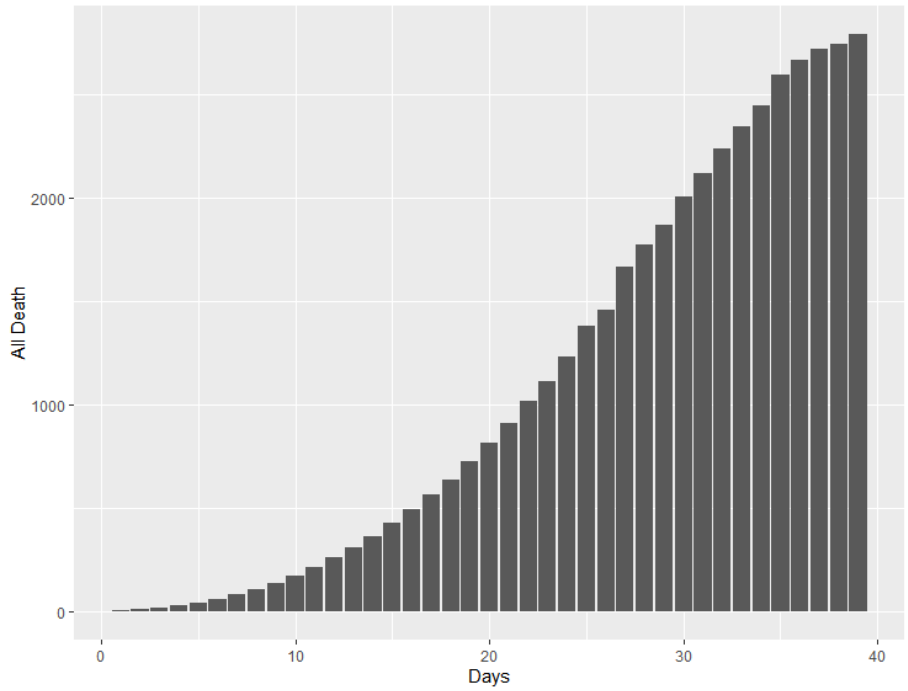
By observing the shape of the data distribution, it can be found that the number of newly diagnosed cases every day, the number of daily close contacts and the number of newly confirmed cases on the second day are stable and in line with the expected distribution. Among them, in mid-February, Wuhan announced the inclusion of clinical diagnostic criteria in the diagnosis of cases, which led to a large ups and downs in the data.

At the same time, we also saw that on February 3rd, the sixteenth observation, new confirmed cases and new close contacts began to fluctuate and fell to the current low level. This shows that the prevention and control of the epidemic situation has a very good effect, and the epidemic situation has been basically controlled.

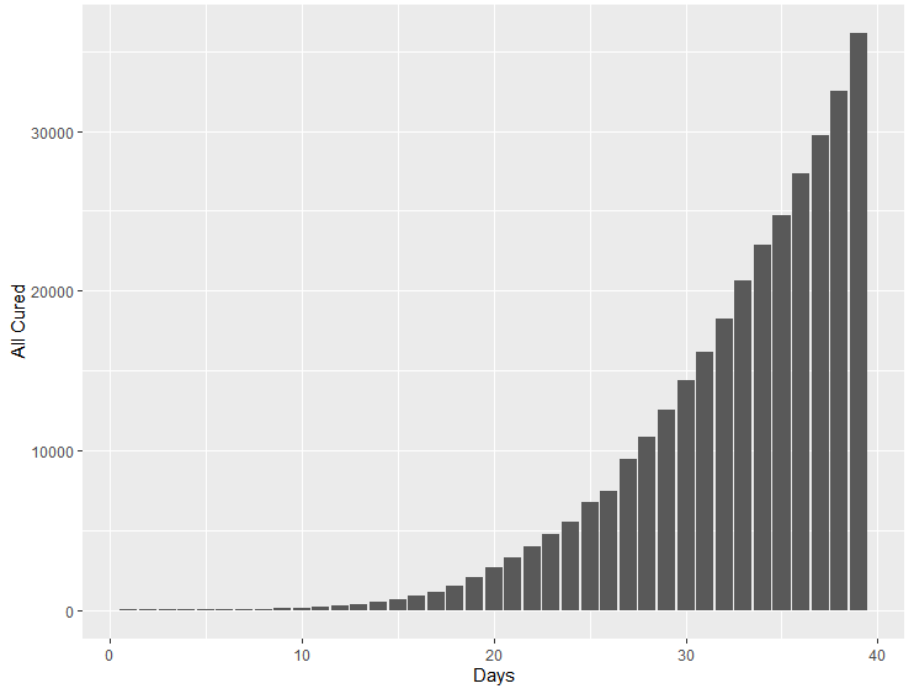
Then, draw a bar graph for the "cumulative" variable:



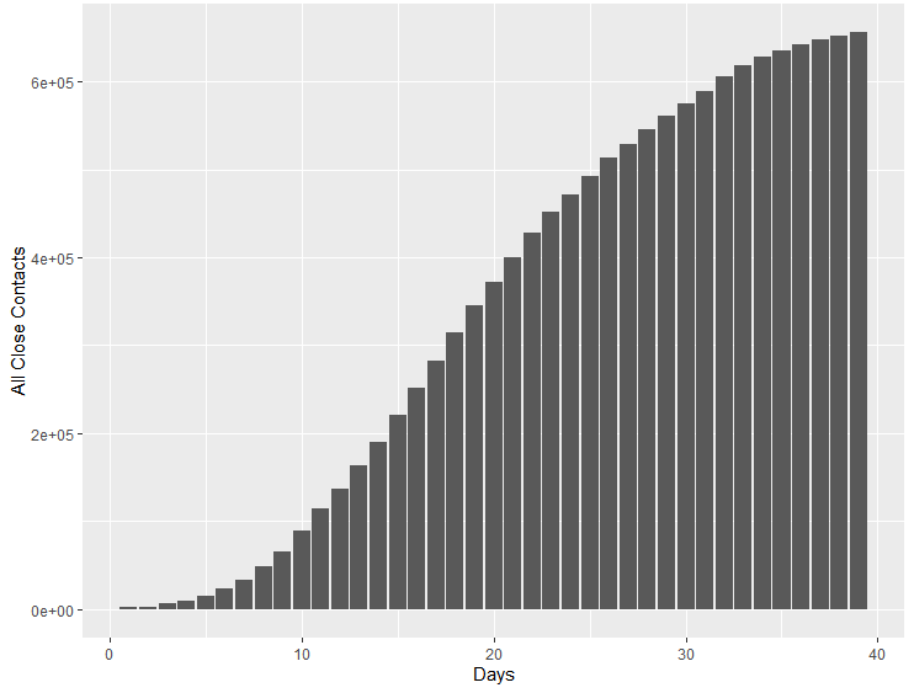
It can be found that the cumulative number of confirmed cases has accelerated from the beginning, and stabilized the growth trend in mid-March, and then Wuhan announced the inclusion of clinical diagnostic criteria in the diagnosis of cases. Contributions to prevention and treatment have effectively prevented the dilemma that suspected cases cannot be diagnosed for a long time. So from then on, the epidemic situation was clearly controlled and the growth trend gradually slowed down. Next, draw a bar graph of the cumulative number of deaths:



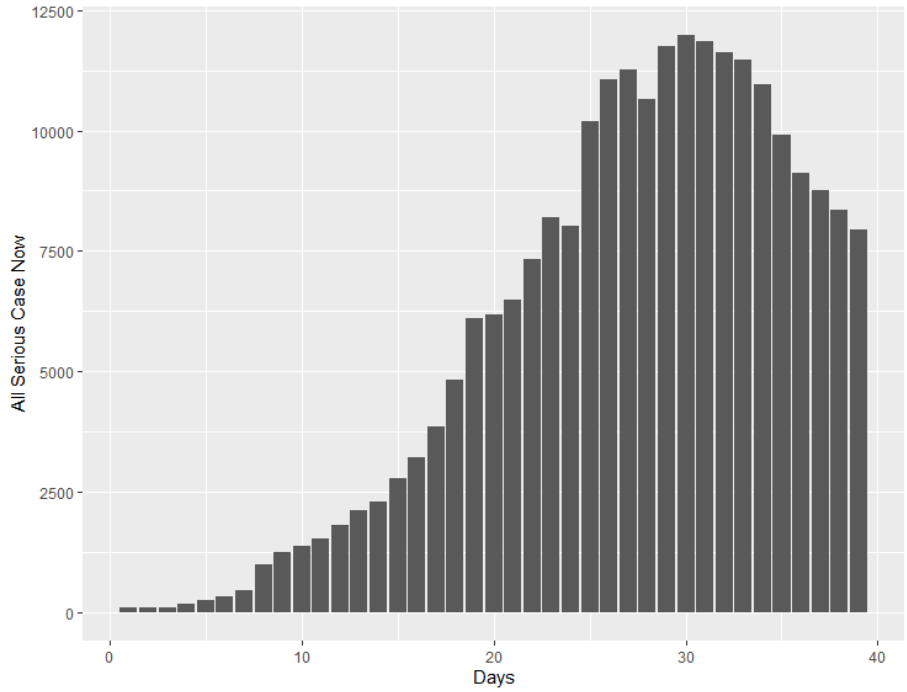
It can be found that from the accelerated growth at the beginning until the increase in the number of deaths on February 27 has slowed down, and has experienced very strict control and prevention measures, the current growth trend has slowed significantly, and the data is stable and credible, which reflects Control effect. Next, draw a bar graph of the cumulative number of cured cases:

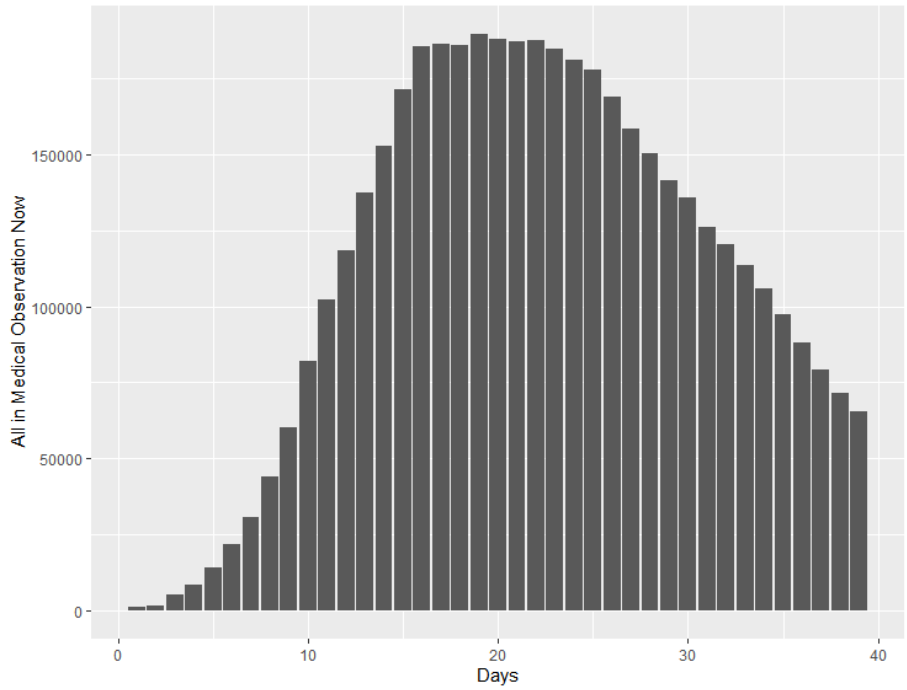


It can be found that the cumulative number of patients is increasing exponentially. With the timely availability of medical supplies and the implementation of prevention and control measures, it is suspected that the patients should be collected and more medical power is being invested in the treatment of patients It is believed that in the near future, due to the decline of cumulative cases, the growth trend of cumulative cure cases will also slow down as the number of cases decreases. Next, draw a bar graph of the cumulative number of close contacts tracked:



Similar to the increasing trend of the cumulative number of confirmed cases, the cumulative tracking of close contacts will increase with the decrease in the number of new cases per day. In summary, the value of the cumulative variable meets the analysis requirements, and the next step can be explored. Next, we draw a bar graph of 'existing' variables that reflect the current treatment pressure:





Based on the above two images, this article believes that since the beginning of February, due to the reduction of new close contacts, the pressure of staff responsible for monitoring the status of close contacts has begun to decrease. Work pressure also began to gradually ease. It is believed that according to this trend, the number of severe cases and close contacts will continue to decline until the end of the epidemic. Therefore, from the above descriptive statistical analysis, it can be seen that the data conforms to the actual situation, there are no missing values and specific values, and it meets the requirements of linear regression analysis. The above-mentioned preliminary inference will provide a great help in establishing a linear regression model.

Simple linear regression model

The premise of regression analysis is the pre-processing of logarithm based on the above data classification method. The author has established a total data frame in the R language program, including the various types of independent variables and dependent variables mentioned above.

First, establish a full regression model of the data. The relevant parameter estimation and test results of the full model are as follows:

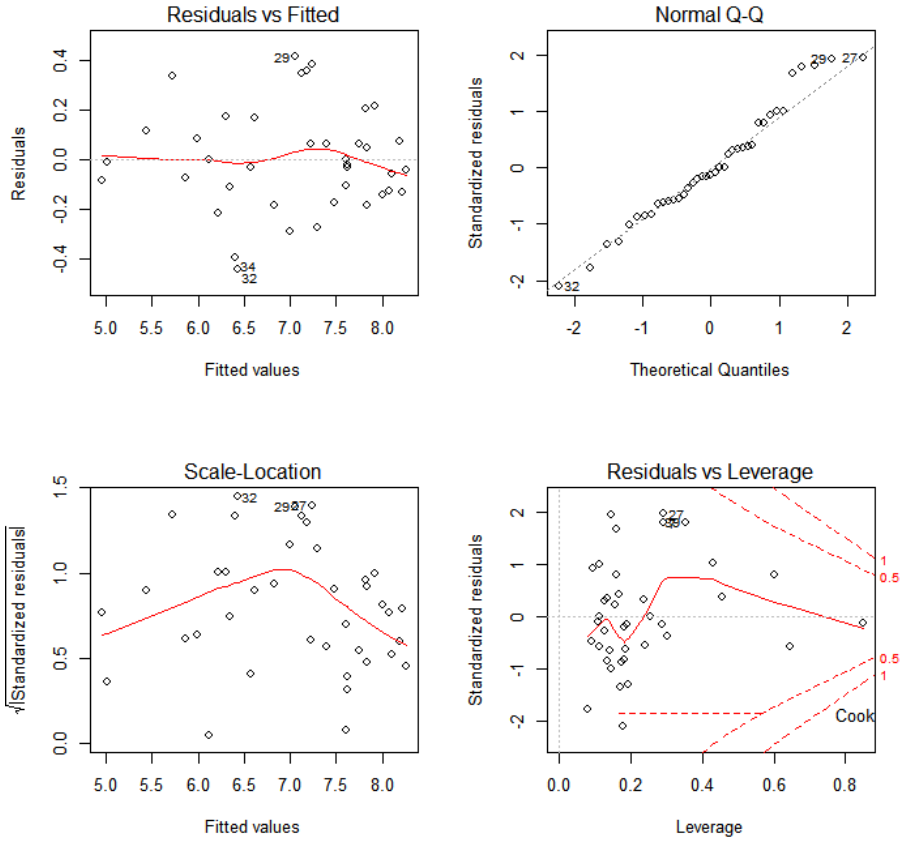
NAME	***COEF***	***STD.ERROR***	***P-VALUE***
Intercept	3.03120	3.56843	0.40236
confirm.new.	0.18212	0.19367	0.35453
confirm.all.	-0.07185	0.54124	0.89528
serious.now.	-0.22472	0.25083	0.37742
death.all.	0.49488	0.75953	0.51965
cured.all.	0.09006	0.16052	0.57892
close.all.	-1.71982	1.18608	0.15743
medical(observation)	2.49197	0.73960	0.00208

NAME	COEF	STD ERROR	P-VALUE
New.	462		

The overall F-test of the model is highly significant (P-value <2.2e-16), which indicates that at least one explanatory variable is related to the height of the new case index on the next day; the adjusted judgment coefficient of the overall model is Radj2 = 0.9381.

Examine and analyse the t test results for each explanatory variable. The t-test results in the table show that the relationship between close contacts who are undergoing medical observation and the newly added close contacts on the same day and the newly confirmed cases on the second day are very significant.

In the case that the model results are reasonable and credible, this article uses visual means to test the reliability of the regression model, the drawing is as follows:



The results obtained from the summary function are already known, and the model as a whole is significant. According to the image, the residuals follow a normal distribution and the number of outliers is small. The model fitting results are good.

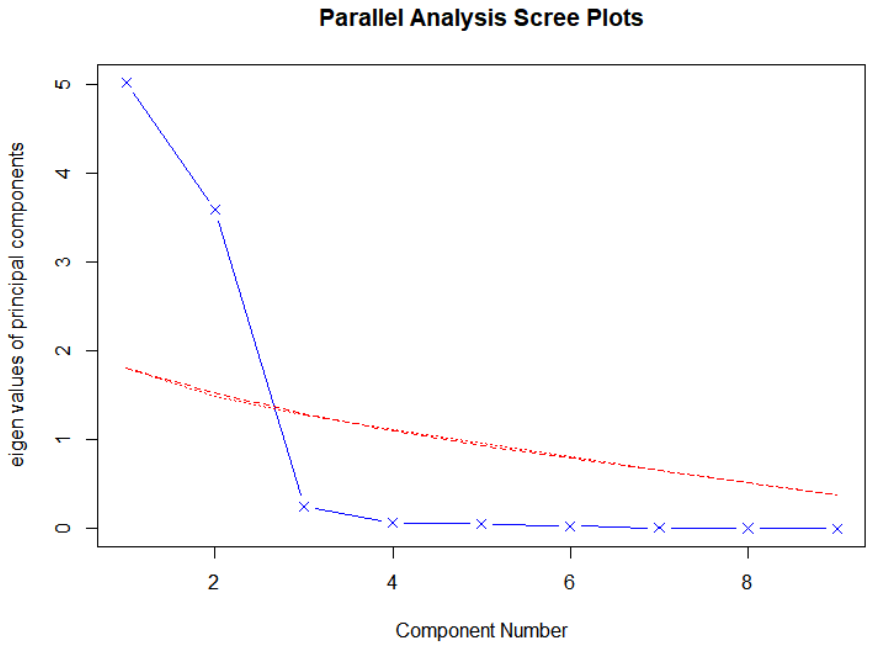
Using the test method provided in the R language package, the VIF value of each explanatory variable is calculated to be large, and there may be multiple collinearity problems.

There's serious multicollinearity problem. This is because for the time series data of the epidemic situation, from the definition of each variable, it can be known that each variable must affect each other and the correlation between the variables is high, and the problem of multicollinearity cannot be avoided.

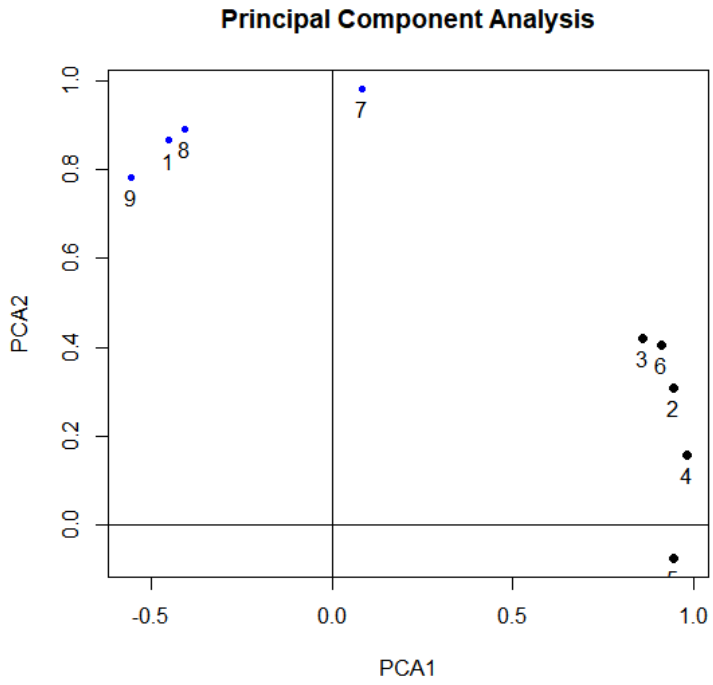
PCA Model (Principal component analysis)

Subsequently, the kappa coefficient in the regression model was calculated; when the condition number k <100, the degree of multicollinearity of the design matrix X is very small; when $100 \leq k \leq 1000$, the design matrix X has strong multicollinearity; At 1000, there is severe multicollinearity.

In this model, the kappa value is 2877.892, it can be determined that the model has strong collinearity, we will continue to carry out principal component analysis, and the analysis can get a better explanation of the principal components of the model. This article first draws the gravel map:



It can be determined that this model should retain the two principal components as the most reasonable. Then the load matrix of the two principal components is calculated, and the weight parameters of each variable are obtained. It can be seen that the eight variables are very different from the propensity of the two principal components:



So according to the tendency of the eight variables to the two principal components, we divide them into two categories:

1. The first principal component represents the "cumulative" variable, so we call it the "cumulative factor".
2. The second principal component represents the ‘additional’ variable, so we call it the ‘additional factor’.

Through these two principal components, 97% of the model changes can be explained by calculation, and the result is very satisfactory.

Next, we use the above two principal components in the actual regression, that is, the ‘cumulative factor’ and ‘new factor’ to regress the design matrix after rotation. The regression results are as follows:

NAME	***COEF***	***STD.ERROR***	***P-VALUE***
Intercept	559.457679	158.121179	0.00113
PCA1	-0.041811	0.003048	<<0.0001
PCA2	0.051607	0.003773	<<0.0001

Both factors are very significant, as expected, and the model fits well.

The comparison with SARS

After 17 years in 2003, the coronavirus brought disaster and sorrow to China and the world. Not only the Chinese economy, but this time the world economy has also been hit hard. The new coronavirus(COVID-19) has become another global pandemic after the Spanish flu in the 1910s.

Under such circumstances, we must not only defend against viruses, but also study viruses, understand viruses, integrate other pandemic diseases encountered by human beings, and estimate and judge the loss and impact of new coronaviruses(COVID-19), in case that the next epidemic will suffer severe losses again.

In this report, we will focus our attention on China, combined with the SARS epidemic in 2003, to assess the similarities and differences between the two coronaviruses on China's economy, culture and other fields, to deepen our understanding of the spreading of coronavirus and the impact on China.

The impact of SARS on the entire Chinese economy is relatively serious. The GDP in the second quarter fell from 9.9% in the first quarter to 6.7%, and it fell to the lowest point since 1992. Regarding the impact of SARS on exports, it should be said that SARS has not had a significant impact on exports in the first half of the year, because a large number of orders were determined last year and before last year.

Therefore, the impact of SARS on exports should be shown in the second half of the year. Because during the occurrence of SARS, it seriously affected the business transactions between China and international. This can also be seen in the monthly data on exports. In May, China's foreign exports increased by 37.3%, but by June it fell to 32.6%. The impact of SARS on exports will appear in the coming months, which means that China's current high-speed export growth will be affected.

However, in general, as China achieves a staged or decisive victory in the prevention and control of SARS, the impact of SARS on the entire economy, including exports, will become smaller and smaller. As far as June is concerned, the consumer price index in June rose by 0.3%; the total retail sales of goods in June was 8.3%; and the fixed asset investment in June was 35.3%.

While according to the relevant basic data and the method of national economic accounting, the main results of the preliminary accounting of China's GDP in the first quarter of 2020 (hereinafter referred to as GDP) are as follows, in which all the values are calculated using Quarterly GDP Accounting Instructions of China protocol.

	Absolute Value (100 million yuan)	Growth Rate over the Same Period Last Year (%)
	Q1	Q1
Gross Domestic Products	**206504**	**-6.8**
Primary Industry	**10186**	**-3.2**
Secondary Industry	**73638**	**-9.6**
Tertiary Industry	**122680**	**-5.2**
Farming, Forestry, Animal Husbandry, and Fishery	10708	-2.8
Industry	64642	-8.5
#Manufacturing	53852	-10.2
Construction	9378	-17.5
Wholesale and Retail Trades	18750	-17.8
Transport, Storage, and Post	7865	-14.0
Accommodation and Restaurants	2821	-35.3
Finance	21347	6.0
Real Estate	15268	-6.1
Information Transmission, Software and Information Technology Services	8928	13.2
Renting and Leasing Activities and Business Services	7138	-9.4
Others	39660	-1.8

The Y/Y Growth Rate on GDP

	Q1	Q2	Q3	Q4
2015	7.1	7.1	7.0	6.9
2016	6.9	6.8	6.8	6.9
2017	7.0	7.0	6.9	6.8
2018	6.9	6.9	6.7	6.5
2019	6.4	6.2	6.0	6.0
2020	-6.8			

The Q/Q Growth Rate on GDP

	Q1	Q2	Q3	Q4
2015	1.8	1.8	1.7	1.6
2016	1.5	1.8	1.7	1.6
2017	1.7	1.8	1.6	1.5
2018	1.7	1.7	1.5	1.5
2019	1.6	1.5	1.3	1.5
2020	-9.8			

As we can see from the numbers and tables, considering that China is still in a period of rapid economic growth in 2003, and the affected areas were limited to a few areas such as Beijing and Hong Kong, the recent outbreaks have been caused a huge negative impact on the Chinese economy, especially on various figures and reports.

The above-scale industrial enterprises is going down over the same month of the previous year, and the growth rate was down by over 11.3% over the previous quarters due to COVID-19.

At a press conference held by the Information Office of the State Council of China on April 23, the person in charge of the Ministry of Industry and Information Technology stated that as a series of counter-cyclical macro-control policies gradually emerged, the major industrial indicators improved significantly. In the first quarter, the added value of industrial enterprises above designated size fell by 8.4% year-on-year, of which 1.1% in March, which was 12.4 percentage points narrower than the previous two months.

Among the 41 industrial major industries, 16 industries achieved growth in March, and 15 of them grew from negative to positive. At the same time, he said that the industrial enterprises above the designated size have basically achieved resumption of production. As of April 21, the average operating rate and resumption rate have reached 99.1% and 95.1% respectively.

Summary

From the original regression model, when the number of new close contacts increases and the cumulative number of close contacts under medical observation remains at a high level, the new cases will also remain at a high level on the second day; at the same time, the new day When the number of confirmed diagnoses is kept at a relatively high level, the newly confirmed cases on the second day will also be relatively high.

From the regression model after the rotation of principal components, the three principal components of ‘cumulative factor’, ‘new factor’ both have a significant impact on new cases on the second day. As the values of both factors increase and decrease, the number of new cases on the second day will increase and decrease accordingly.

When we compared COVID-19 with the SARS that occurred in 2003, we found that in 2003, the impact of SARS on China was limited to certain regions and a relatively short period of time. And this year, China, whose economy is slowly advancing, has received nationwide impact. As the epidemic continues to spread abroad, this impact will have a long-term impact on China's economic development.