Analyze Influencing Factors of COVID-19 with R Language and Machine Learning Models

Suizhi Ma[*] Yuxing Pei[†] Ming Wang[‡] Longling Geng[§] Jiaming Han[¶] and Yunfei Liu[‖]

Department of Computing, The Hong Kong Polytechnic University

Abstract

We have analyzed possible economical, political, and medical factors which may affect the spread of COVID-19 with R language and several machines learning regression models including multivariate linear regression and multivariate polynomial regression. Comparisons are made between countries and in countries to figure out the possible influencing factor's effect on daily new confirmed cases. According to the result of our training models, political and economic factors may be more important in some aspects than vaccines.

***K*eywords** COVID-19 · Machine Learning · R Language · Data Analysis

1 Introduction

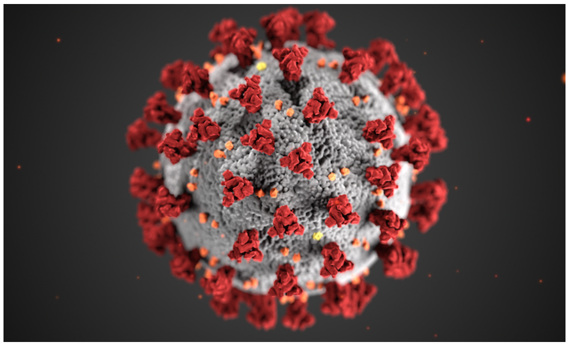


Figure 1: COVID-19

COVID-19 has received concern from the whole world, and people from the whole world are fighting against the epidemic together. To understand the spread of the virus, the human must figure out what factors affect the spreading speed. As global citizens, we feel that we should undertake the responsibility to make a contribution to the fight against the virus, and we are eager to implement what we have learned in this subject to achieve this target. In this project, we try to find what can affect the virus and the relation between spread speed and possible factors [1].

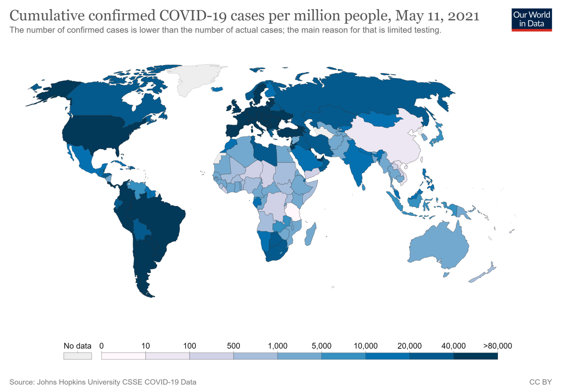

Cumulative confirmed COVID-19 cases per million people, May 11, 2021
The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.
Our World in Data
No data
0
10
100
500
1,000
5,000
10,000
20,000
40,000
>80,000
Source: Johns Hopkins University CSSE COVID-19 Data
CC BY

Figure 2: Total confirmed COVID-19 cases in the world

2 Motivation

The very first intuition we choose the topic "Data analysis on the influencing factors of development of Covid-19" is deep interest. There is a large number of people in Kong

[*] Algorithm Engineer

[†] Test Engineer

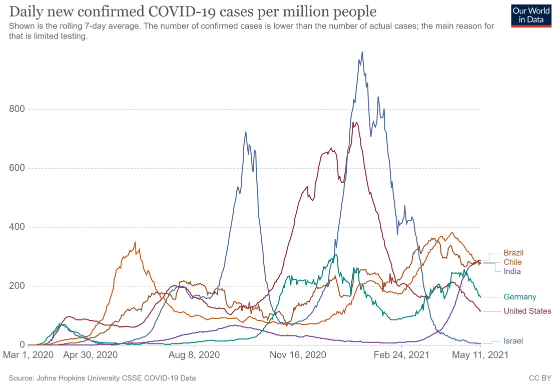
[‡] Algorithm Engineer

[§] Leader, Report Writter

[¶] Data Engineer

[‖] Systems Architect, Report Writter

Kong starting to take the vaccine [2]. Some people do not know much about the effectiveness of the vaccine. Some people are questioning the effectiveness. Therefore, as a group of students who are interested in the facts and truths behind the vaccine. We have a strong desire to find out the real impact of the vaccine [3].

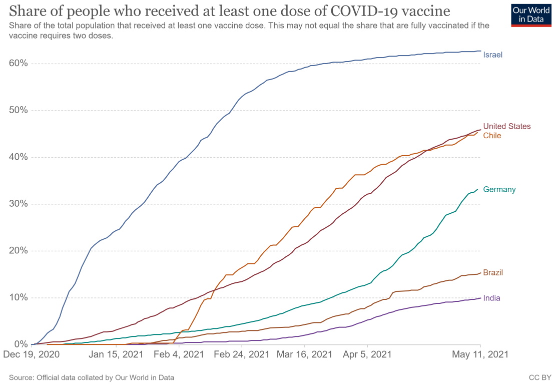

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Source: Johns Hopkins University CSSE COVID-19 Data

Figure 3: Confirmed COVID-19 cases fluctuation in some countries

However, before scientific research is conducted, only interest is not enough. The research topic should be proved valuable and worthy to discuss. After looking into various papers, we find out that theoretical value and reality research are two main issues that should be taken into consideration. On the theoretical side, a large number of related papers can be found and "Covid-19" and "related vaccine" are both the hot keywords in the academic library. On the reality side, this topic is proven to be useful because its result will benefit the government and the healthcare organization to take the measurement. Also, this topic is close to the hot and crucial social issue nowadays.


Share of people who received at least one dose of COVID-19 vaccine

Share of the total population that received at least one vaccine dose. This may not equal the share that are fully vaccinated if the vaccine requires two doses.

Source: Official data collated by Our World in Data

Figure 4: COVID-19 vaccination rate in some countries

Finally, although this topic is challenging, it allows us to combine what we learn in this course with creative thinking. We can apply the linear regression model, basic techniques of processing data in this research. We can also learn by ourselves and explore new models and algorithms in this research. Therefore, based on "interest, research value and learning exercise" these three aspects, this topic is decided.

3 Background

The first research is about the influencing factors of Covid-19 based on the 30 provinces in China.

It investigates multiple influencing factors, for example, the GDP of each province, the population density of each province, to find out which factors influence the spread of Covid-19. It's conducted during the period from 2020.1.26 to 2020.2.29 with its data from Baidu and the official China government website. The result shows that the scale of the disease in the first stage, the migration rate of Wuhan's population, and the multi-factor set of the urban population are seriously related to the expansion of the second stage of new cases.

However, there are still 3 limitations: The data is not up to date; the data is limited; the import case is not accurate, which are stated in their report, too. Therefore, inspired by its multiple factors research on Covid-19 and based on its limitation, our research steps forward to look at the data all over the world. This does not only solve the problem that data is limited and old because we use the most recent data collected by Hopkins University, which is proven to be authoritative as well. By adjusting our independent factors, the "inaccurate import data" problem is solved as well [4]. In summary, we make good use of its researching process and methodology but migrate this idea to a better data source. However, apart from its researching thinking, we also integrate what we have learned in class and by ourselves into this research.

4 Description

We have concluded that COVID-19 may be affected by a few factors, and we have classified these factors into groups including time, virus, vaccine, government, population, and economic groups. All these groups may have more or fewer effects on virus spread. Also, different models including linear regression are all used to test for the best fit [5].

In this project, we define a few independent variables which are possibly related to the spread of COVID-19. For example, the date may affect the result because the overall confirmed cases will vary by date. We list more than 10 possible factors which cover different areas to figure out whether the relationship exists between these factors and COVID-19[6, 7].

We have distributed our task into a few parts: Research target determination, data collection, data process, model construction, code implementation, test validation, result finalization, and report writing. We have distributed these stages of research to different group members so that we can increase efficiency and avoid time conflict.

5 Implementation

5.1 Code in R

R has a large variety of packages that offers numerous APIs for machine learning. As the saying goes, "Don't reinvent the wheels.". To write the implementation of the algorithms is not feasible and our programming level is far from "invent the wheels" [8, 9]. Meanwhile, even if we write the code to implement algorithms by ourselves, due to the shortage of time, we cannot improve it and it won't perform as well as the APIs offered by mature advanced packages. [10]To balance the workload and adjust our actual ability, we choose to use built-in functions to implement our ideas [11, 12, 13].

5.2 Models

R has a powerful build-in function "lm()", which we can use to construct a linear regression model. As mentioned above, we choose to use the new confirmed case rate, which is "7-day moving average of newly confirmed cases per million", as the dependent. Then, we select a variety of independents including the rate of people that get vaccinated (represent the vaccine), stringency index (an index which can show the government regulation fight against the virus, such as knockdown and mandatory COVID-19 test), population density (reflect the difficulty that COVID-19 can spread between people), GDP per capita (reflect the economic development of this country), handwashing facilities(hand wash can prevent the spread of the virus) and human development index (reflect the medical facilities and people's awareness of getting treatment) as an independent [14].

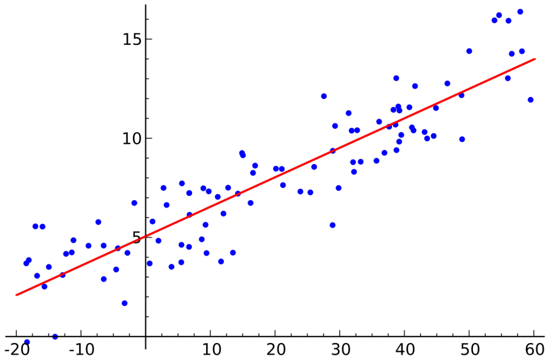


Figure 5: A sample of linear regression

We tried multivariate linear regression to fit the data at the beginning and based on the result of the first round of model training and testing, we select the independents that have a good fit with dependent (i.e. high p-value) to do the next round of training and testing. After that, we tried to conclude and display the results [15, 16].

Also, we tried to use multivariate polynomial regression to fit the data, and compare it with the multivariate linear regression model [17]. To implement the idea of a control variable, we use the same dataset and the same dependents and independents to guarantee the correctness of the result [18, 19]. Similar to the multivariate linear regression model, we also select independents with a high p-value for the next round of training and testing [20].

5.3 Test and Validation

As mentioned above, four different machine learning models are implemented and finally one model are expected to be found as the best fit. To test whether our result is correct, we use split the data into training dataset and testing dataset, the "p-value" and "R-value" are calculated to evaluate the models, and after that, the model with the smallest p-value and "R-value" which is the closest to 1 has the highest accuracy.

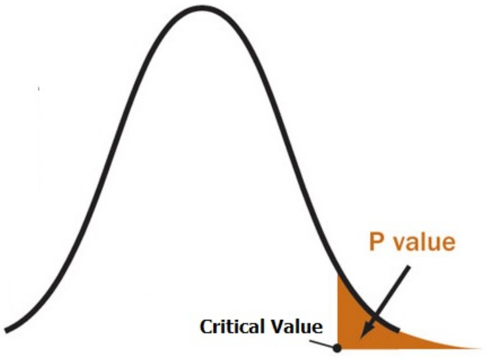


Figure 6: The visualization of p-value

6 Data

6.1 Data Collection

Data collection is an important beginning of this project, as the data will determine the goodness of fit of models and the correctness of results. Reliable sources will increase the credibility of the final result while non-official sources may lead to wrong answers. To take responsibility for humanity, we decided to only use data from official department announcements such as CDC. Data sources are attached in the appendix.

When collecting the data, we attempted a few times to get the final raw data. We tried to use a web crawler to collect the data at first, but we found that some websites do not support crawlers, or they do not provide convenient access. Then, we tried to download the statistical data manually, but we meet genuine difficulties when we tried to recognize the language different from Chinese and English. Finally, we found an organization *Our World In Data*, which provides a summary of all data offered on the government or official institution website. The organization *Our World In*

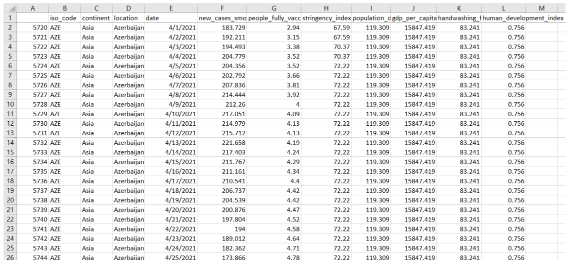
Data provided the sources of the data, after checking, we finally determined to use the data on its GitHub repository.

6.2 Data Selection

The data provided by *Our World In Data* is quite complete and it contains information that we do not need for analysis. Moreover, some countries with a small population are not very feasible for analysis, and part of the data in some countries is missed. To get rid of this problem, proper data selection is of vital importance, and it can also affect the result.

Because we are planning to analyze data horizontally and vertically, so we should filter the data and create a different dataset for machine learning. The dataset should be the subset of the raw table and we only keep key features that we want to analyze in the dataset. Meanwhile, we deleted countries without required features because they are not useful for the current experiment.

We initially classified the datasets into the horizontal group and vertical group. For the vertical group, we divided the data in different times, such as 2020 spring, summer, autumn, and winter, which can help us figure out the change of epidemic with time. Also, we can find how the virus is spread in specific countries and regions. For the horizontal group, we compare data from different countries. Since each country has its economic or medical situation and different policy against COVID-19. By analyzing different countries' data, we can find the relationship between COVID-19 and country development and figure out whether a knock-down policy is necessary [21, 1].



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		iso_code	continent	location	date	new_cases_smo	people_fully_vacc	stringency_index	population_d	gdp_per_capita	handwashing_f	human_development_index	
2	5720	AZE	Asia	Azerbaijan	4/1/2021	183.729	2.94	67.59	119.309	15847.419	83.241	0.756	
3	5721	AZE	Asia	Azerbaijan	4/2/2021	192.211	3.15	67.59	119.309	15847.419	83.241	0.756	
4	5722	AZE	Asia	Azerbaijan	4/3/2021	194.493	3.38	70.37	119.309	15847.419	83.241	0.756	
5	5723	AZE	Asia	Azerbaijan	4/4/2021	204.779	3.52	70.37	119.309	15847.419	83.241	0.756	
6	5724	AZE	Asia	Azerbaijan	4/5/2021	204.356	3.52	72.22	119.309	15847.419	83.241	0.756	
7	5725	AZE	Asia	Azerbaijan	4/6/2021	202.792	3.66	72.22	119.309	15847.419	83.241	0.756	
8	5726	AZE	Asia	Azerbaijan	4/7/2021	207.836	3.81	72.22	119.309	15847.419	83.241	0.756	
9	5727	AZE	Asia	Azerbaijan	4/8/2021	214.444	3.92	72.22	119.309	15847.419	83.241	0.756	
10	5728	AZE	Asia	Azerbaijan	4/9/2021	212.26	4	72.22	119.309	15847.419	83.241	0.756	
11	5729	AZE	Asia	Azerbaijan	4/10/2021	217.051	4.09	72.22	119.309	15847.419	83.241	0.756	
12	5730	AZE	Asia	Azerbaijan	4/11/2021	214.979	4.13	72.22	119.309	15847.419	83.241	0.756	
13	5731	AZE	Asia	Azerbaijan	4/12/2021	215.712	4.13	72.22	119.309	15847.419	83.241	0.756	
14	5732	AZE	Asia	Azerbaijan	4/13/2021	221.658	4.19	72.22	119.309	15847.419	83.241	0.756	
15	5733	AZE	Asia	Azerbaijan	4/14/2021	217.403	4.24	72.22	119.309	15847.419	83.241	0.756	
16	5734	AZE	Asia	Azerbaijan	4/15/2021	211.767	4.29	72.22	119.309	15847.419	83.241	0.756	
17	5735	AZE	Asia	Azerbaijan	4/16/2021	211.161	4.34	72.22	119.309	15847.419	83.241	0.756	
18	5736	AZE	Asia	Azerbaijan	4/17/2021	210.541	4.4	72.22	119.309	15847.419	83.241	0.756	
19	5737	AZE	Asia	Azerbaijan	4/18/2021	206.737	4.42	72.22	119.309	15847.419	83.241	0.756	
20	5738	AZE	Asia	Azerbaijan	4/19/2021	204.539	4.42	72.22	119.309	15847.419	83.241	0.756	
21	5739	AZE	Asia	Azerbaijan	4/20/2021	200.876	4.47	72.22	119.309	15847.419	83.241	0.756	
22	5740	AZE	Asia	Azerbaijan	4/21/2021	197.804	4.52	72.22	119.309	15847.419	83.241	0.756	
23	5741	AZE	Asia	Azerbaijan	4/22/2021	194	4.58	72.22	119.309	15847.419	83.241	0.756	
24	5742	AZE	Asia	Azerbaijan	4/23/2021	189.012	4.64	72.22	119.309	15847.419	83.241	0.756	
25	5743	AZE	Asia	Azerbaijan	4/24/2021	182.362	4.71	72.22	119.309	15847.419	83.241	0.756	
26	5744	AZE	Asia	Azerbaijan	4/25/2021	173.866	4.78	72.22	119.309	15847.419	83.241	0.756	

Figure 7: Selected data for training

6.3 Data Cleaning

Sometimes, the government didn't provide sufficient information on some holidays, but they provided it on workdays so that it is not wise to simply delete all the data of the country. Instead, we removed the rows which do not have enough data and set date as a feature so that machine learning models can understand that some rows are removed, and the models will make adjustments for these missed data.

Moreover, the raw data contains the repeated part, which may affect the result. For example, if we only want to analyze the data from different countries, the data of the continent will become useless. Therefore, we deleted these unnecessary data.

6.4 Data Standardize

This step is preprocessed before we use the data from Hopkins University. The feature list and the meaning of its value is listed as below.

6.4.1 new_cases_smoothed_per_million

- **Source:** COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
- **Description:** New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people

6.4.2 new_deaths_smoothed_per_million

- **Source:** COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
- **Description:** New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people

6.4.3 people_fully_vaccinated_per_hundred

- **Source:** National government reports
- **Description:** Total number of people who received all doses prescribed by the vaccination protocol per 100 people in the total population

6.4.4 total_vaccinations_per_hundred

- **Source:** National government reports
- **Description:** Total number of COVID-19 vaccination doses administered per 100 people in the total population

6.4.5 people_vaccinated_per_hundred

- **Source:** National government reports
- **Description:** Total number of people who received at least one vaccine dose per 100 people in the total population

6.4.6 stringency_index

- **Source:** Oxford COVID-19 Government Response Tracker, Blavatnik School of Government
- **Description:** Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)

6.4.7 population_density

- **Source:** World Bank World Development Indicators, sourced from Food and Agriculture Organization and World Bank estimates

- **Description:** Number of people divided by land area, measured in square kilometers, most recent year available

6.4.8 gdp_per_capita

- **Source:** World Bank World Development Indicators, source from World Bank, International Comparison Program database
- **Description:** Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available

6.4.9 handwashing_facilities

- **Source:** United Nations Statistics Division
- **Description:** Share of the population with basic handwashing facilities on premises, most recent year available

6.4.10 human_development_index

- **Source:** United Nations Development Programme (UNDP)
- **Description:** A composite index measuring average achievement in three basic dimensions of human development long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506

7 Result

7.1 Overview

We are trying to find out the relationships between the spread of Covid-19 and several influencing factors, especially the vaccination. So it is crucial to use proper data analysis methods and build some models to fit the conclusions. After serious consideration, we chose the linear regression model and exponential regression model. Now let me show you what we have found out [22].

The influencing factors are complicated and various. But we have chosen several key influencing factors, which are considered to have a great impact on the epidemic though some of them may not have significant influence in fact.

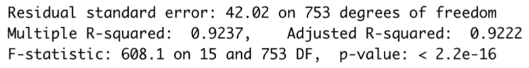
1. people_fully_vaccinated_per_hundred
2. total_vaccinations_per_hundred
3. people_vaccinated_per_hundred
4. stringency_index
5. population_density
6. gdp_per_capita
7. handwashing_facilities
8. human_development_index

As for the value to monitor the epidemic, we chose:

1. new_cases_smoothed_per_million
2. new_deaths_smoothed_per_million

7.2 Vaccination and other contributing factors

We have tried four different models, and finally, an exponential model is considered to be the most accurate and reasonable. We would like to explain it in detail.



```
Residual standard error: 42.02 on 753 degrees of freedom
Multiple R-squared:  0.9237,	Adjusted R-squared:  0.9222
F-statistic: 608.1 on 15 and 753 DF,  p-value: < 2.2e-16
```

Figure 8: R-square and p-value of models

As we can see, the R-squared value is close to 1 and p-value is extremely small, which have proved the accuracy of this model.

Here is the primary residual graph. The y-axis is $\frac{model\_value}{true\_value}$ and the x-axis is the time index. If most value is close to 1, the we can know the model is accurate enough. This graph is not clear enough, so we have another graph.

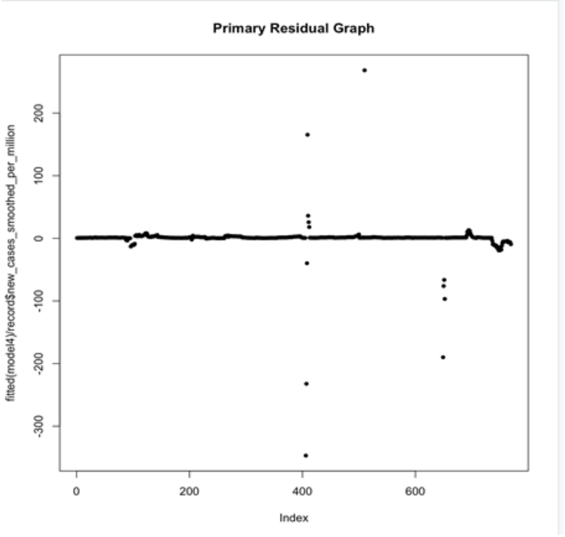


Figure 9: Accuracy Test, $\frac{model\_value}{true\_value}$

In the next graph, we remove some points that are extremely large or extremely small. And we minor 1 to every value. And we could see that most points are close to 0. So the model is relatively accurate. And it has some practical meanings for prediction.

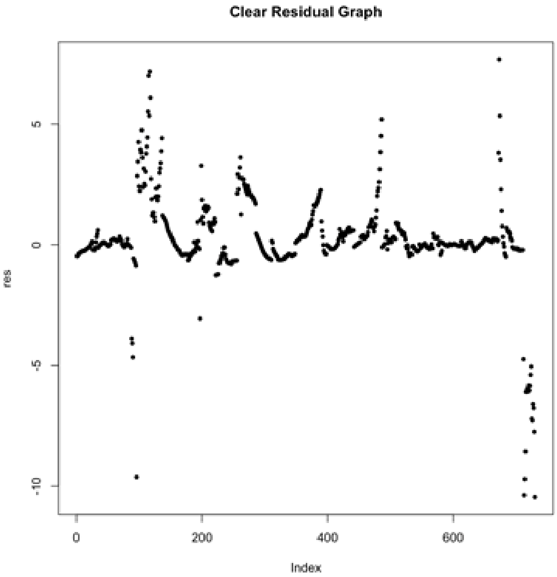
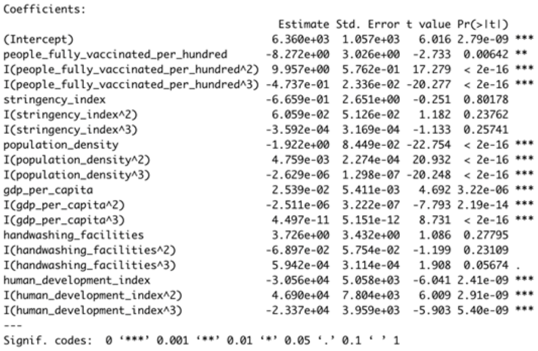


Figure 10: Accuracy Test, delete outliers, $\frac{model\_value}{true\_value}$

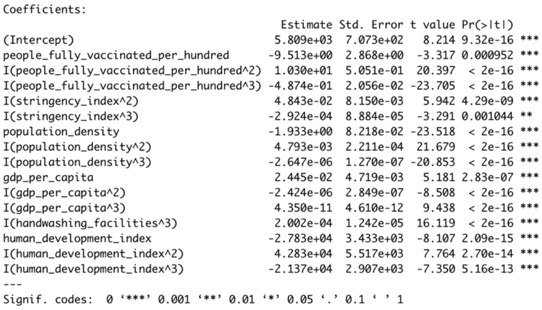
How the factors influence the epidemic? We can find answers in the models.



```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              6.360e+03  1.057e+03   6.016 2.79e-09 ***
people_fully_vaccinated_per_hundred     -8.272e+00  3.026e+00  -2.733  0.00642 ** 
I(people_fully_vaccinated_per_hundred^2)  9.957e+00  5.762e-01  17.279  < 2e-16 ***
I(people_fully_vaccinated_per_hundred^3) -4.737e-01  2.336e-02 -20.277  < 2e-16 ***
stringency_index                        -6.659e-01  2.651e+00  -0.251  0.80178    
I(stringency_index^2)                    6.059e-02  5.126e-02   1.182  0.23762    
I(stringency_index^3)                   -3.592e-04  3.169e-04  -1.133  0.25741    
population_density                      -1.922e+00  8.449e-02 -22.754  < 2e-16 ***
I(population_density^2)                  4.759e-03  2.274e-04  20.932  < 2e-16 ***
I(population_density^3)                 -2.629e-06  1.298e-07 -20.248  < 2e-16 ***
gdp_per_capita                           2.539e-02  5.411e-03   4.692 3.22e-06 ***
I(gdp_per_capita^2)                     -2.511e-06  3.222e-07  -7.793 2.19e-14 ***
I(gdp_per_capita^3)                      4.497e-11  5.151e-12   8.731  < 2e-16 ***
handwashing_facilities                   3.726e+00  3.432e+00   1.086  0.27795    
I(handwashing_facilities^2)             -6.897e-02  5.754e-02  -1.199  0.23109    
I(handwashing_facilities^3)              5.942e-04  3.114e-04   1.908  0.05674 .  
human_development_index                 -3.056e+04  5.058e+03  -6.041 2.41e-09 ***
I(human_development_index^2)             4.690e+04  7.804e+03   6.009 2.91e-09 ***
I(human_development_index^3)            -2.337e+04  3.959e+03  -5.903 5.40e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Coefficients of independents

This image shows the significance of the factors. As we can see, **handwashing_facilities** and **stringency_index** have low significance to the epidemic, while the others have high significance. Remove some of the factors and we have another model summary.



```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              5.809e+03  7.073e+02   8.214 9.32e-16 ***
people_fully_vaccinated_per_hundred     -9.513e+00  2.868e+00  -3.317 0.000952 ***
I(people_fully_vaccinated_per_hundred^2)  1.030e+01  5.051e-01  20.397  < 2e-16 ***
I(people_fully_vaccinated_per_hundred^3) -4.874e-01  2.056e-02 -23.705  < 2e-16 ***
I(stringency_index^2)                    4.843e-02  8.150e-03   5.942 4.29e-09 ***
I(stringency_index^3)                   -2.924e-04  8.884e-05  -3.291 0.001044 ** 
population_density                      -1.933e+00  8.218e-02 -23.518  < 2e-16 ***
I(population_density^2)                  4.793e-03  2.211e-04  21.679  < 2e-16 ***
I(population_density^3)                 -2.647e-06  1.270e-07 -20.853  < 2e-16 ***
gdp_per_capita                           2.445e-02  4.719e-03   5.181 2.83e-07 ***
I(gdp_per_capita^2)                     -2.424e-06  2.849e-07  -8.508  < 2e-16 ***
I(gdp_per_capita^3)                      4.350e-11  4.610e-12   9.438  < 2e-16 ***
I(handwashing_facilities^3)              2.002e-04  1.242e-05  16.119  < 2e-16 ***
human_development_index                 -2.783e+04  3.433e+03  -8.107 2.09e-15 ***
I(human_development_index^2)             4.283e+04  5.517e+03   7.764 2.70e-14 ***
I(human_development_index^3)            -2.137e+04  2.907e+03  -7.350 5.16e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12: Coefficients of independents, after the removal of low relevance factors

7.3 Summary of program results

After repeating experiments, we found that the spread of COVID-19 may related to with the following expressions:

$$f \propto \Big[-0.49 \times (C_1)^3 + 10.3 \times (C_1)^2 - 9.5 \times C_1 \Big]$$

$$f \propto \Big[-2.6 \times 10^6 \times (C_2)^3 + 0.0048 \times (C_2)^2 - 1.93 \times (C_2) \Big]$$

$$f \propto \Big[4.35 \times 10^{-11} \times (C_3)^3 - 2.424 \times 10^{-6} \times (C_3)^2 + 2.445 \times 10^{-2} \times C_3 \Big]$$

$$f \propto \Big[-2.137 \times 10^4 \times (C_4)^3 + 4.283 \times 10^4 \times (C_4)^2 - 2.783 \times 10^4 \times C_4 \Big]$$

In the above expressions, f refers to the newly confirmed cases everyday (which can represent the spread of COVID-19) , while the C_i refers to:

$$C_1 = people\_fully\_vaccinated\_per\_hundred$$
$$C_2 = population\_density$$
$$C_3 = gdp\_per\_capita$$
$$C_4 = human\_development\_index$$

8 Discussion

In this research, we have got some results. However, due to the regression model and other possible impact features, we cannot 100 percent assert that the result is correct. Considering that the data source is reliable, and the academic integrity of the group members can be guaranteed, we can prove that the result is as correct as possible. Here are some of the possible errors that we think may affect the answer.

- **Data accuracy** Some cases may be ignored by the government and they are not counted into the confirmed cases
- **Model Choice** The models we used may not the best model, and there may be other models which is better than current models
- **Change of constant values** We assume some independents as constant values, such as human development index and stringency index. However, they will change with the time.
- **Data selection** Because of the lack of data, only countries will complete data are selected, and this may lead to error for some developing countries.
- **Math error** Most of the countries have a low vaccination rate due to the large population, and it may cause math error and affect the result.

Considering the possible errors above, the result may have some differences with the reality, and we cannot fully convinced that the result is correct. More detailed and well-designed experiments required to make the final decision.

9 Conclusion

The development of epidemic is not influenced by one single factor, but affected by several factors. In common, more people are vaccinated or lower population density or higher gdp (more prosperous economy) or lower human_development_index (which may mean less educated) will contribute to the development of epidemic.

If the Hong Kong government intends to control Covid-19, I am afraid that the result will not be satisfying enough if the government only vaccines more citizens. Because according to our research, vaccination only has slight influence to the epidemic. But more importantly, as a high advanced city, control the population density (maybe control the social distance) and let more people understand the importance of it will contribute to the decrease of new cases.

10 Appendix

10.1 Useful Links

- **Project Repository**
  GitHub `https://github.com/comp1433/Magic-Vaccine`
- **Data Set**
  Our World In Data `https://ourworldindata.org/coronavirus`
- **Data Set Repository**
  GitHub `https://github.com/owid/covid-19-data/tree/master/public/data`
- **Reference Source**
  Google Scholar `https://scholar.google.com/`

References

[1] Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, and Joe Hasell. Coronavirus pandemic (covid-19). *Our world in data*, 2020.

[2] Chor-Cheung Frankie Tam, Kent-Shek Cheung, Simon Lam, Anthony Wong, Arthur Yung, Michael Sze, Yui-Ming Lam, Carmen Chan, Tat-Chi Tsang, Matthew Tsui, et al. Impact of coronavirus disease 2019 (covid-19) outbreak on st-segment–elevation myocardial infarction care in hong kong, china. *Circulation: Cardiovascular Quality and Outcomes*, 13(4):e006631, 2020.

[3] Kin On Kwok, Kin Kit Li, Ho Hin Chan, Yuan Yuan Yi, Arthur Tang, Wan In Wei, and Yeung Shan Wong. Community responses during the early phase of the covid-19 epidemic in hong kong: risk perception, information exposure and preventive measures. *MedRxiv*, 2020.

[4] Elizabeth J Williamson, Alex J Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E Morton, Helen J Curtis, Amir Mehrkar, David Evans, Peter Inglesby, et al. Opensafely: factors associated with covid-19 death in 17 million patients. *Nature*, 2020.

[5] Mohammed AA Al-Qaness, Ahmed A Ewees, Hong Fan, and Mohamed Abd El Aziz. Optimization method for forecasting confirmed cases of covid-19 in china. *Journal of Clinical Medicine*, 9(3):674, 2020.

[6] Anthony S Fauci, H Clifford Lane, and Robert R Redfield. Covid-19—navigating the uncharted, 2020.

[7] Rachel E Jordan, Peymane Adab, and KK32217618 Cheng. Covid-19: risk factors for severe disease and death, 2020.

[8] Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

[9] Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. strucchange. an r package for testing for structural change in linear regression models. 2001.

[10] Brad Boehmke and Brandon M Greenwell. *Hands-on machine learning with R*. CRC Press, 2019.

[11] Michael J Crawley. *The R book*. John Wiley & Sons, 2012.

[12] R Core Team et al. R: A language and environment for statistical computing. 2013.

[13] Brett Lantz. *Machine learning with R*. Packt publishing ltd, 2013.

[14] Smita Rath, Alakananda Tripathy, and Alok Ranjan Tripathy. 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, 2020.

[15] Roseline O Ogundokun, Adewale F Lukman, Golam BM Kibria, Joseph B Awotunde, and Benedita B Aladeitan. Predictive modelling of covid-19 confirmed cases in nigeria. *Infectious Disease Modelling*, 5:543–548, 2020.

[16] Julien Arino and Stéphanie Portet. A simple model for covid-19. *Infectious Disease Modelling*, 5:309–315, 2020.

[17] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. An interpretable mortality prediction model for covid-19 patients. *Nature machine intelligence*, 2(5):283–288, 2020.

[18] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, Marta Colaneri, et al. A sidarthe model of covid-19 epidemic in italy. *arXiv preprint arXiv:2003.09861*, 2020.

[19] Jesús Fernández-Villaverde and Charles I Jones. Estimating and simulating a sird model of covid-19 for many countries, states, and cities. Technical report, National Bureau of Economic Research, 2020.

[20] Ulrike Grömping et al. Relative importance for linear regression in r: the package relaimpo. *Journal of statistical software*, 17(1):1–27, 2006.

[21] Mark G Thompson, Jefferey L Burgess, Allison L Naleway, Harmony L Tyner, Sarang K Yoon, Jennifer Meece, Lauren EW Olsho, Alberto J Caban-Martinez, Ashley Fowlkes, Karen Lutrick, et al. Interim estimates of vaccine effectiveness of bnt162b2 and mrna-1273 covid-19 vaccines in preventing sars-cov-2 infection among health care personnel, first responders, and other essential and frontline workers—eight us locations, december 2020–march 2021. *Morbidity and Mortality Weekly Report*, 70(13):495, 2021.

[22] Paulino Pérez, Gustavo de Los Campos, José Crossa, and Daniel Gianola. Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in r. *The plant genome*, 3(2), 2010.