COVID-19 Analysis

Haoyu Shen

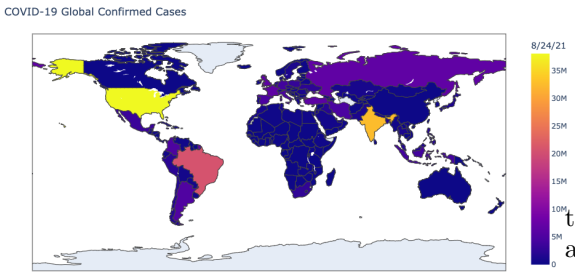
September 8, 2021

1 Introduction

COVID-19 is currently the most dangerous and contagious disease worldwide. It can spread from an infected person's mouth or nose in small liquid particles when they cough or breath. For the past year, more than 4 million people around the world have died because of it. In the next several parts, we will focus deeply on COVID-19 statistics in the US, Canada, and China. For instance, confirmed cases in each state/province on a particular date, the trend of COVID-19 for the past few months, and prediction of it use three different models: SVM, linear regression, and ARIMA models.

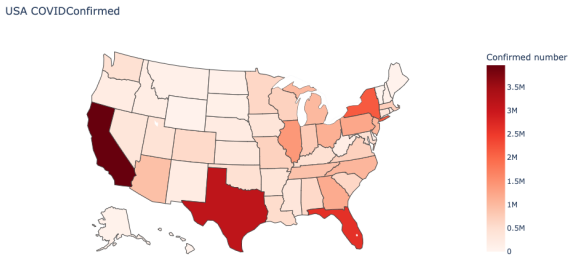
For those who are not familiar with these models, the SVM model, known as the support vector machine model, is a predictive analysis data classification algorithm that is usually used in the field of machine learning. Linear regression model predicts the value of a target variable based on our given predictor variable. ARIMA model is a class of statistical models for analyzing and forecasting time series data. We will apply the first two models to our data to figure out the future forecasting of COVID-19 in the US, Canada, and China. And at last, we will apply the ARIMA model.

2 COVID-19 spread - worldwide and in US

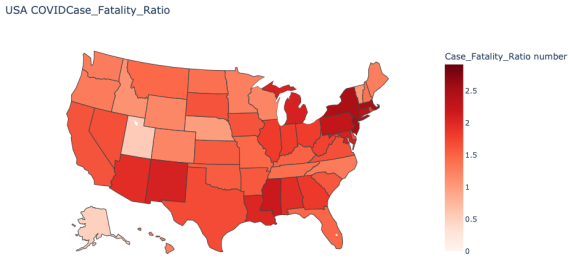


From the above graph of daily confirmed cases worldwide, we can see that US is the country with most confirmed cases as of Aug 24, 2021. As the amount of confirmed COVID-19 cases are shown as the scale on the right, we can also say that COVID-19 is spreading severely in both Brazil and India. In the next section, we will mainly focus on:

1. US, since it currently has the most confirmed cases over the world
2. Canada, as the author is currently located here, and
3. China, as it has done a fabulous job on controlling the spread of COVID-19 since the outbreak.



As the darker the red is, the severer the issue might be in that state, we can conclude that California is the state with most confirmed COVID-19 cases currently in US. In the next section, we will take a look at if the fatality rate of California also be the most in US.

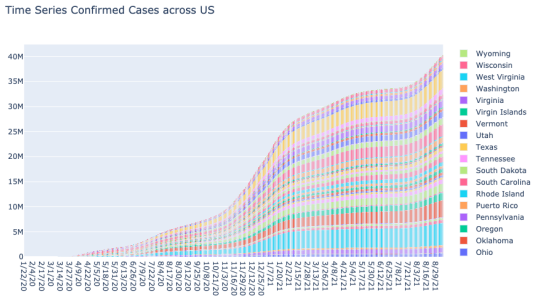


From the fatality rate map, we can conclude that states with the highest fatality ratio are around New Jersey, Miami and New York. As all these three states stack together, we might conclude that here is the most severe COVID spreading area in US. Also, as New York is one of the most developed states in US and it still has a relatively high fatality ratio, we can say

that we still haven't found an efficient way to protect people from getting COVID and/or cure the severe symptoms.

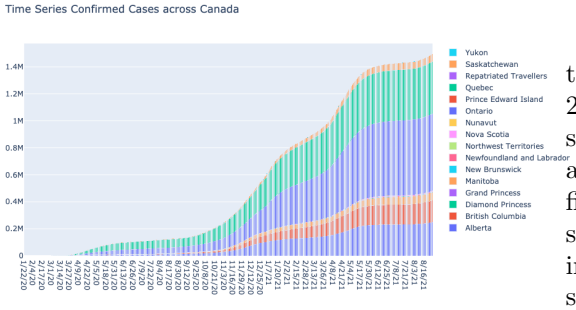
3 COVID-19 trends in US, Canada and China

3.1 US



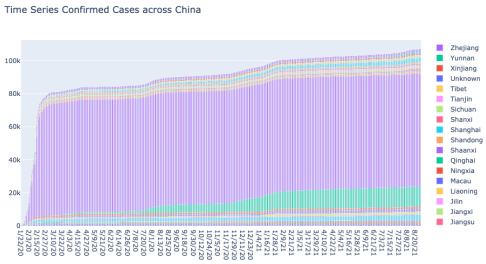
From bar graph above, we are able to see that the state with most confirmed cases since Jan 22, 2020 is California of around 4.4M. Also, we are able to see an obvious increasing trend in some states like California and Florida which means there may keep getting increasing COVID cases if there is no external factors to intervene.

3.2 Canada



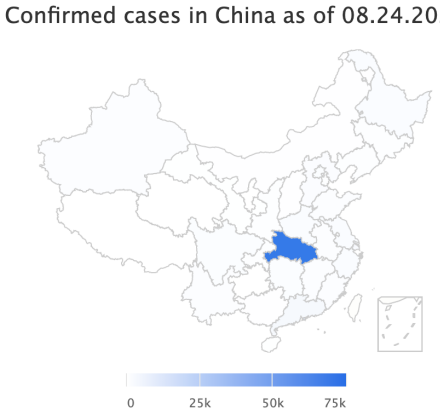
From the bar graph, we can conclude that Ontario is currently the province with most confirmed COVID cases and the other two top provinces are Quebec and Alberta. Although confirmed cases not increase rapidly in Canada, it also never stop as we see that confirmed cases keep increasing day after day. From the trends of data from these provinces, we may predict that confirmed cases will keep increasing in the future as we can't see any declining trend from the graph. To consolidate our opinion, we will make a future prediction based on the SVM and linear regression model in the next section.

3.3 China

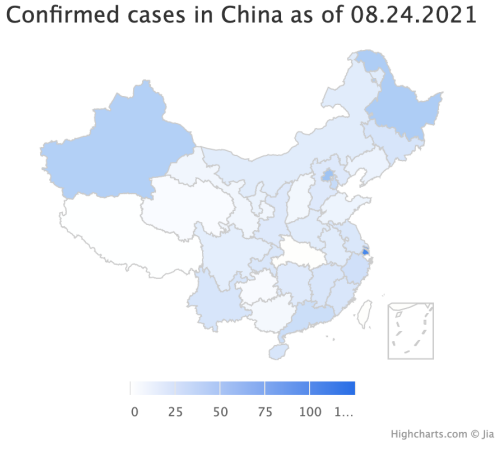


We can see there was a burst increasing in the amount of confirmed cases from Feb 13, 2020. Hubei was the province with the most severe situation as it had a confirmed amount of around 64k by Feb 23, 2020. Luckily, the confirmed cases in almost all provinces remain at a steady state, which means confirmed cases don't increase rapidly, except for Hong Kong, as we still can see increasing confirmed cases from July 27, 2020. From the trends of the provinces, we may assume that COVID has been under control in China, but we still make further more solid prediction based on SVM and linear regression models in the next section.

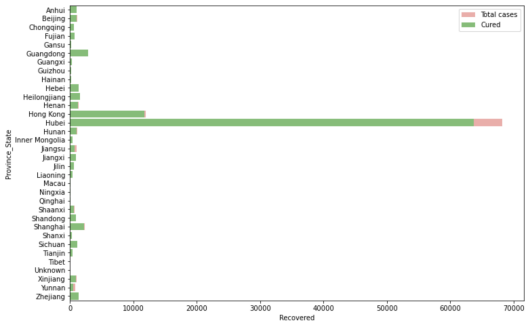
4 Confirmed, density and cured proportion of China



From the country map of China, we are able to get the idea of confirmed COVID cases as of Aug 24, 2021 intuitively. The result from the map consistent with our conclusion from bar graph of China that Hubei got the most confirmed cases which was around 60k. Actually we plotted confirmed cases of every province in China, but we can rarely show them on the map as those values are too small when comparing to 60k.



Then We take a look at the density graph of confirmed COVID cases in China as of Aug 24, 2021. The scale of this map is 1:1 million to avoid data sets to be huge. Also, I removed data of Hubei and Hongkong as their value are much greater than other provinces and may affect the outcome of our graph. From graph exclude Hubei and Hongkong, we can see that the province/city with the highest density of confirmed cases was Shanghai with around 101 in 1 million.

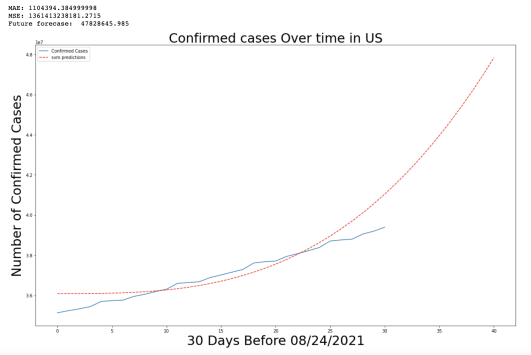


The bar graph indicates the ratio of total confirmed cases and cured cases in China as of Aug 24, 2021. The green part indicates the cure portion whereas the red part shows the total confirmed cases. We can say that almost every confirmed COVID cases in China from every province was cured, except for Hubei, that there still was a small portion of confirmed cases that was not cured. We are glad to see graph like this as this shows China has almost successfully controlled the spreading of COVID issue and remain the death rate being low.

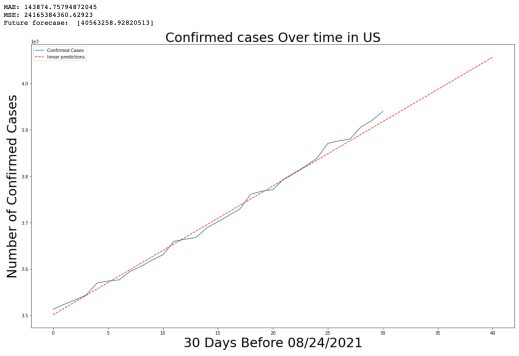
5 Prediction of COVID-19 trends

5.1 Prediction over 30 days

5.1.1 US

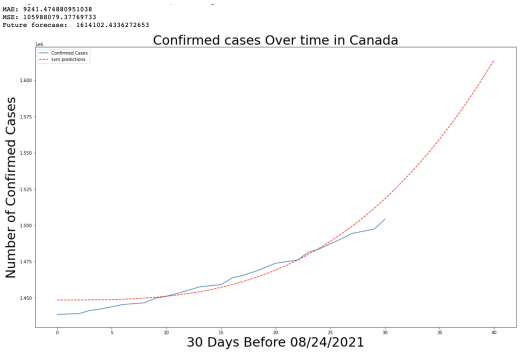


Firstly, we use the SVM model to make our prediction of the US 10 days after Aug 24, 2021. Intuitively, we can see that the SVM model might not be appropriate for the US as it tends to go up rapidly ignore the US's current trend. From the prediction of the SVM model, it shows there will be confirmed cases of 47,184,899 which is an increase of around 10M cases in 10 days. This result is not reliable as the US got an increased amount of 3M for the past 20 days. Therefore, we will take a look at the linear regression model of the US.

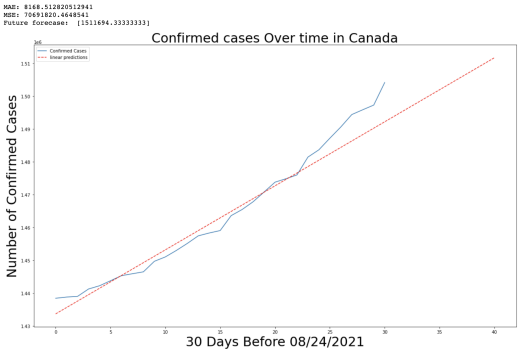


The linear regression model seems much more appropriate as most data lies on the linear regression line. The future prediction shows there will be an increased amount of 1M in confirmed cases over the 10-day period. Comparing to its past data with an increased amount of 3M over 20 days, the linear regression model is more valid.

5.1.2 Canada



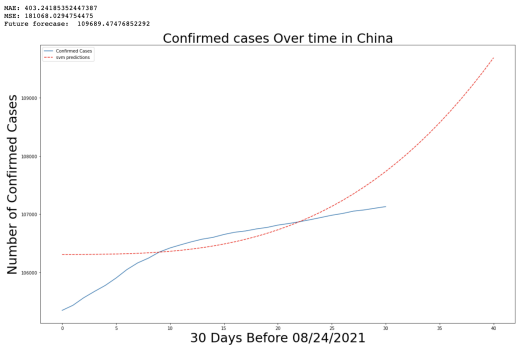
From SVM model of Canada, we can see most data lies on the curve where SVM model predicts a greater trend of increasing number in confirmed cases. It predicts that Canada may has an increase of 120k by 10 days after Aug 24, 2021. Although most data fits on the curve compare to data of China, the increasing amount still seems unreasonable as the confirmed cases only increased around 40k from July 25, 2021 to Aug 24, 2021. Therefore, we will try linear regression model on Canada.



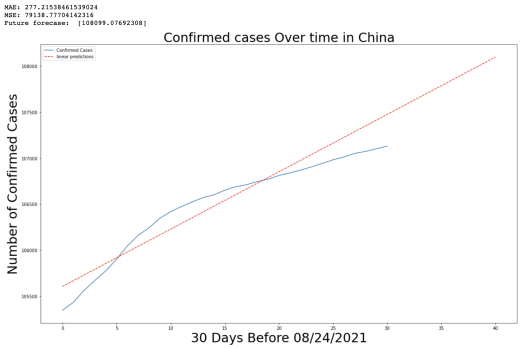
We can see that linear regression prediction model fits more appropriate compare to the SVM model. It predicts that Canada may has an increasing amount of around 10k 10 days after Aug 24, 2021. From the value we retrieve from the graph, we can also say that this prediction model is more reliable as confirmed cases

don't increase rapidly like the SVM model. The only problem is we see that the trend of confirmed cases in Canada tends to go up at a faster rate where our linear regression model only predicts it linearly, so it can't introduce any outer interference which may predicts lower than the actual value.

5.1.3 China



We are going to take a look at the prediciton of confirmed COVID cases 10 days after Aug 24, 2021 in China. First we use SVM model. SVM is a predictive analysis data-classification algorithm that assigns new data elements to one of labeled categories. From our model where red dashed line indicates the svm model, we can see prediction is not really accurate as trend of confirmed cases in China tends to decline whereas our predicted model tends to keep going up. One possible reason to cause this is our model doesn't involve external intervention like area lockdown, so we will take a look at linear regression prediction.

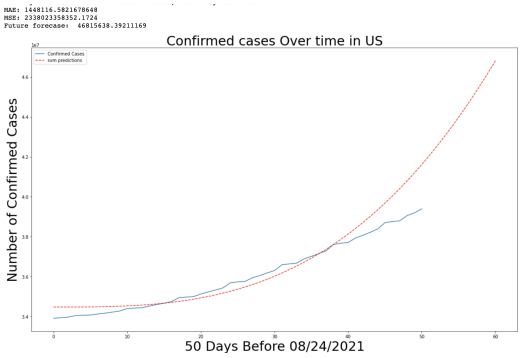


Linear regression is one of the most commonly used predictive modelling techniques based on dependent (target) and independent variable (predictor). We are able to see that the linear regression model fits more appropriate than our svm model. Our data of confirmed cases lies on the predicted regression line which provides a better accuracy then svm model. From our linear model, China may has around 2k confirmed cases increase 10 days after Aug 24, 2021.

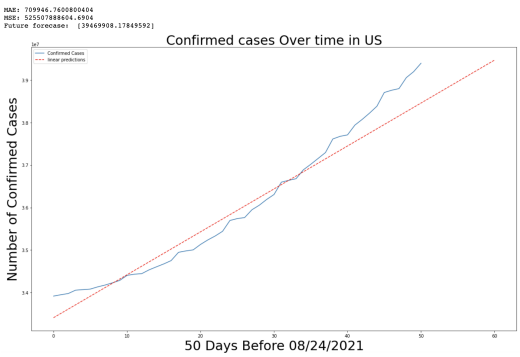
5.2 Prediction over 50 days

Although we have plotted prediction over 30 days, we want to make sure that our data is accurate based on separate models. Therefore, we want to check if the mean squared error which is the average of the squares of the errors get smaller when our input size gets bigger.

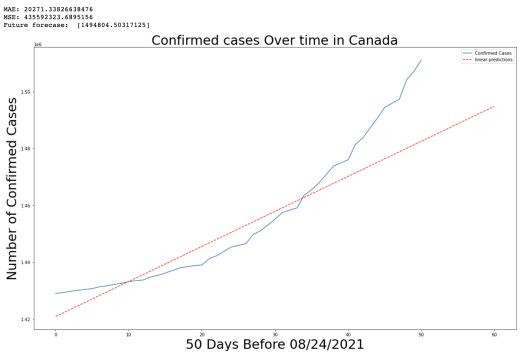
5.2.1 US



From our plot of US prediction over 50 days, we can see that both the SVM model and linear regression model have an increase in MSE which means both methods are not appropriate for data of US.

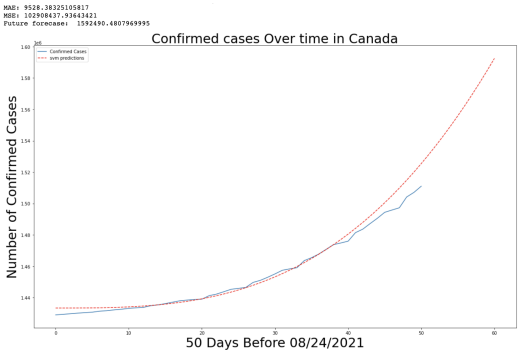


One possible reason to cause this might be confirmed cases in the US is too large for our model that many errors may involve. So, we will take a look at China and Canada to see whether these two methods work for them.



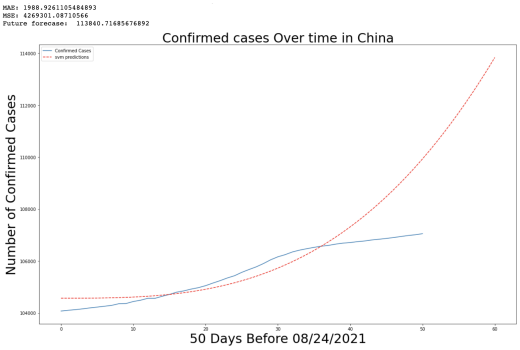
To ensure SVM model is better for Canada, we check MSE of linear regression model of Canada. MSE has a value smaller value over 30 days than over 50 days which shows an increasing trend with increase of input size. Therefore, we can conclude that SVM model is much better for Canada than linear regression.

5.2.2 Canada

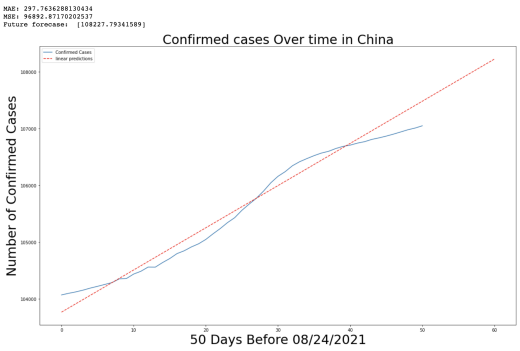


After US, we check our data on Canada. SVM model on Canada has a MSE value over 30 days that is greater than value over 50 days. This time we retrieve different test results from China that MSE value decreases as the size of input increase.

5.2.3 China



By comparing both SVM model and linear regression model, we got a value of MSE that is much greater for data over 50 days with svm model than data over 30 days. Its quite obvious that MSE value increases as we increase our data input which means SVM model is not appropriate for China. Then we check linear regression model, we also get MSE for 50 days is much smaller than MSE for 30 days. We are glad to see that MSE decreased a lot which means linear regression model is better than SVM model for China.



As we have proved linear regression is better for China and SVM model is better for Canada, we can see what prediction value do these two models have for two countries 10 days after Aug 24, 2021. for China, linear regression model predicts that it may has a total confirmed cases of 108274 and for Canada, SVM model predicts that it may has a total of 1,605,595 by 10 days after.

6 Prediction use ARIMA model

6.0.1 US

Best model: ARIMA(0,1,0)(0,0,0)[0] intercept
Total fit time: 0.683 seconds

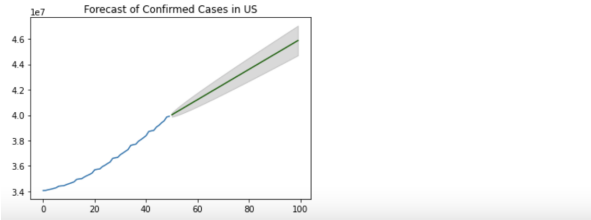
SARIMAX Results

Dep. Variable:	y	No. Observations:	50
Model:	SARIMAX(0, 1, 0)	Log Likelihood	-625.515
Date:	Sun, 05 Sep 2021	AIC	1255.030
Time:	15:57:42	BIC	1258.813
Sample:	0	HQIC	1256.465
	- 50		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	1.192e+05	1.57e+04	7.606	0.000	8.84e+04	1.5e+05
sigma2	7.171e+09	1.94e+09	3.693	0.000	3.37e+09	1.1e+10

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	6.19
Prob(Q):	0.94	Prob(JB):	0.05
Heteroskedasticity (H):	2.30	Skew:	0.87
Prob(H) (two-sided):	0.11	Kurtosis:	2.87

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).



As both SVM and linear regression model don't fit data of US well, we will try to make prediction of confirmed COVID cases of the same data set with ARIMA model. ARIMA model is a class of statistical models for analyzing and forecasting time series data. The shaded area indicates an approximate range of predicted value.It predicts that US may has an increased amount of 1M in 10 days and 5M in 50 days.

6.0.2 Canada

Best model: ARIMA(2,1,2)(0,0,0)[0] intercept
Total fit time: 1.389 seconds

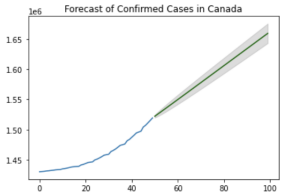
SARIMAX Results

Dep. Variable:	y	No. Observations:	50
Model:	SARIMAX(2, 1, 2)	Log Likelihood	-419.156
Date:	Sat, 04 Sep 2021	AIC	850.312
Time:	12:36:33	BIC	861.663
Sample:	0	HQIC	854.618
	- 50		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	514.9908	5110.169	0.101	0.920	-9500.757	1.05e+04
ar.L1	-0.0817	19.691	-0.004	0.997	-38.674	38.511
ar.L2	0.8988	17.871	0.050	0.960	-34.129	35.926
ma.L1	0.0770	19.690	0.004	0.997	-38.515	38.669
ma.L2	-0.9034	17.962	-0.050	0.960	-36.109	34.302
sigma2	1.486e+06	21.667	6.86e+04	0.000	1.49e+06	1.49e+06

Ljung-Box (L1) (Q):	0.69	Jarque-Bera (JB):	10.42
Prob(Q):	0.41	Prob(JB):	0.01
Heteroskedasticity (H):	3.46	Skew:	0.95
Prob(H) (two-sided):	0.02	Kurtosis:	4.22

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step)
[2] Covariance matrix is singular or near-singular, with condition number 3.02e+20.

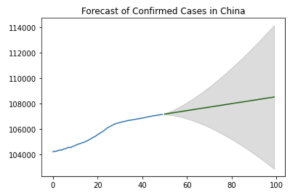


SVM model predicts that there will be an increased amount of 90,000 in 10 days where our ARIMA model predicts an amount of 50,000 in 10 days and 170,000 in 50 days. We can't really say which model is better for prediction as they are both future forecasting, they provide a general idea of how COVID may spreads around the world in the future.

6.0.3 China

```
Best model:  ARIMA(1,2,0)(0,0,0)[0]
Total fit time: 0.407 seconds
                               SARIMAX Results
==========================================================================================
Dep. Variable:                           y   No. Observations:                   50
Model:               SARIMAX(1, 2, 0)   Log Likelihood                -216.915
Date:                Sat, 04 Sep 2021   AIC                            437.830
Time:                        12:36:31   BIC                            441.572
Sample:                             0   HQIC                           439.244
                                  - 50
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5409      0.075     -7.257      0.000      -0.687      -0.395
sigma2       449.9128     79.928      5.629      0.000     293.257     606.569
===================================================================================
Ljung-Box (L1) (Q):                   1.05   Jarque-Bera (JB):                 1.01
Prob(Q):                              0.31   Prob(JB):                         0.60
Heteroskedasticity (H):               0.07   Skew:                            -0.33
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.29
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```



From linear regression model of China, which is the better model compare to SVM model, indicates that there will be an increased amount of 1277 in confirmed cases after 10 days. From our ARIMA model, it shows there will be an increased amount of 3500 in 10 days. As p-values both significant that both smaller than 0.05, ARIMA model is more reliable and accurate.

7 Discussion and recommendation

From all three models, we can see that the ARIMA model might be the best in future forecasting as it indicates the upper and lower bounds of the predicted value in the shaded area that gives an intuitive idea. Also, it is more precise than the SVM and linear regression model, as both those two models are pre-constructed. These models can provide a general idea of how COVID might be in the future but are not as accurate as of the ARIMA model.

No matter which model we use, we can see that the confirmed COVID cases in the US and Canada keep increasing. Due to policy difference, the US and Canada can't be like China, as they apply a strict quarantine policy, but overall they are doing a much better job than the past year.

Fortunately, citizens start to get vaccines of COVID that highly reduce the risk of death. Although they still might get sick because of COVID, it is much better than last year. Our prediction of the COVID trend suggests that the government can intervene, otherwise the confirmed cases will keep growing as we predicted even they are vaccinated.

Also, from official online resources, it shows 68.6% of population fully vaccinated in Canada. We can see a decline trend from May, the beginning month of the first dose of COVID vaccine, to the end of July, the beginning month of the second dose, with a confirmed cases dropped from 7k to 700, which indicates vaccine works well.