Data Analysis in Covid-19 Based on Regression and Seasonal ARIMA Model

Zhen Zhou (Graduate Student)
Netid: zhenz5

Abstract

In this project, I construct two models to discover the hidden information from Covid-19 data set. One is linear regression model to fit daily deaths by three other coefficients, another is SARIMA model to fit daily new cases. Based on the characteristics of Covid-19, I used a polynomial regression model with a time lag which will have flexible delay in different time series datasets, the RSE of the time lag regression model is 110.2 and all coefficients are significant. For seasonal dataset, after comparison, I use SARIMA$(1,1,2)(0,1,1)_7$ model to fit the dataset and all coefficients are significant, then, I calculate the predicted values for the next 5 days, and the errors are within the acceptable range.

1. Introduction

The year 2020 is a special year, in which coronavirus pandemic over the world. Covid-19 is a great challenge for governments and health care systems around the world. Community infections, lack of medical resources, vaccine development, vaccine policy, etc., for statistics, we can use these data to assist in finding solutions to this pandemic.

In this project, I focus on the data comes from Covid-19, including daily confirmed cases, the number of deaths per day, the number of vaccinations per day, etc.

Data such as daily deaths, daily confirmed, etc. These are all time series data, so we can use the models in the time series to build appropriate models to uncover the hidden information in the data.

In this study, firstly, I use regression analysis to find a model to fit the number of deaths per day. Secondly, I suggest an optimal seasonal ARIMA model to predict future confirmed cases.

All data sets were obtained from the article “A global database of COVID-19 vaccinations” [1]

2. Regression analysis of daily deaths from Covid-19

2.1 Preliminary analysis

Figure 1 shows the four data used for regression analysis, which are smoothed daily deaths, smoothed daily vaccinations, smoothed daily ICU patients, and smoothed daily tests. I use data other than the number of daily deaths to regress the number of daily deaths.

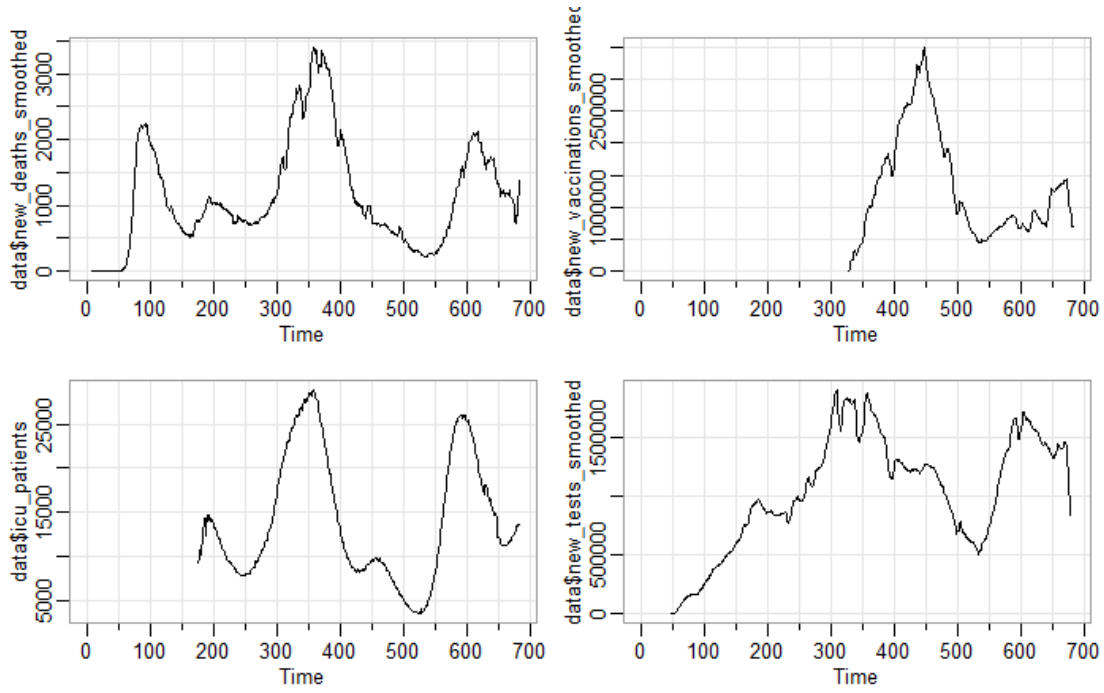

Figure 1 Four datasets

It is obviously that the first graph, which is the number of daily deaths, has a trend, it reaches peak at 350 and decreases until 550, then has an upper trend. And data definitely is not stationary.

Figure 2 shows the ACF of four datasets.

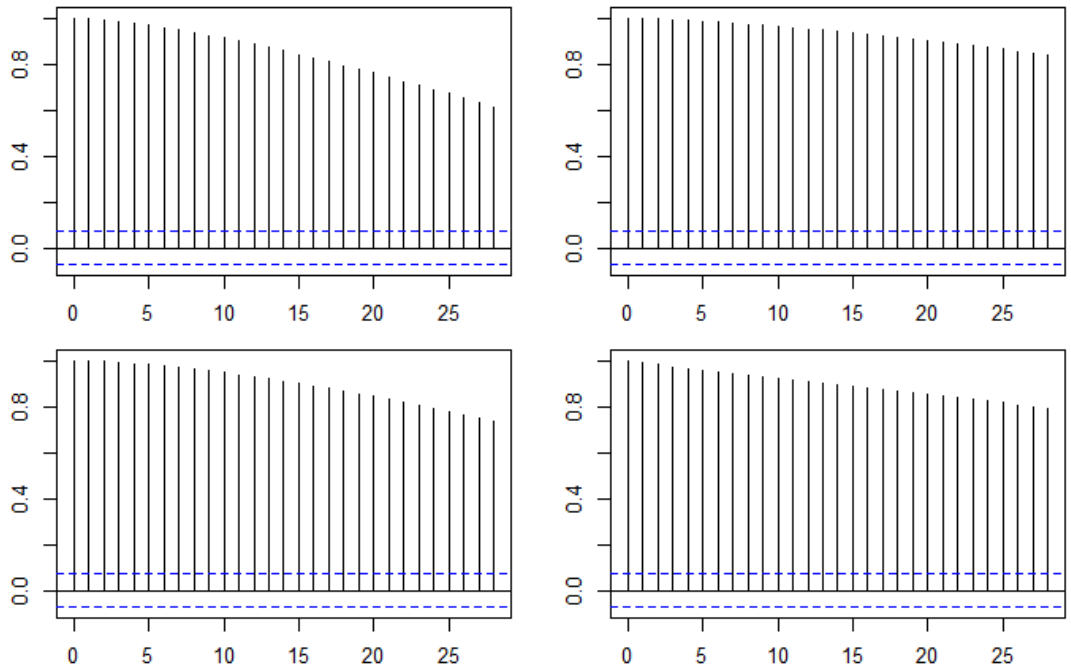

Figure 2 ACF of four datasets

2.2 Regression model selection

The covid-19 has a 14-day delay [2], meaning that the average test is positive about 14 days after infection. Then we can take this delay into account when building the regression model. I compare delay data with non-delay data in this section, also, I construct a flexible delay model to fit the data.

Firstly, non-delay model, it's a normal second order linear regression model, the equation can be written as below:

$$y_{death}(t) \sim t + t^2 + \mathrm{x_{vac}}(\mathrm{t}) + \mathrm{x_{icu}}(\mathrm{t}) + \mathrm{x_{test}}(t) + x_{vac}^2(\mathrm{t}) + x_{icu}^2(\mathrm{t}) + x_{test}^2(t)$$

Delay model can be written:

$$y_{death}(t) \sim t + t^2 + x_{\mathrm{vac}}(\mathrm{t-d}) + x_{\mathrm{icu}}(\mathrm{t-d}) + x_{\mathrm{test}}(t-d) + x_{vac}^2(\mathrm{t-d}) + x_{icu}^2(\mathrm{t-d}) + x_{test}^2(t-d)$$

Flexible delay model can be written as:

$$y_{death}(t) \sim t + t^2 + x_{\mathrm{vac}}(\mathrm{t-d_1}) + x_{\mathrm{icu}}(\mathrm{t}-d_2) + x_{\mathrm{test}}(t-d_3) + x_{vac}^2(\mathrm{t}-d_1) + x_{icu}^2(\mathrm{t}-d_2) + x_{test}^2(t-d_3)$$

In flexible delay model, I set $d_1 = 24,\ d_2 = 14,\ d_3 = 24.$

In these three models, t is the time, $y_{death}(t)$ is smoothed daily deaths, $x_{\mathrm{vac}}(\mathrm{t})$ is smoothed daily vaccinations, $x_{\mathrm{icu}}(\mathrm{t})$ is smoothed daily ICU patients and $x_{\mathrm{test}}(t)$ is smoothed daily tests, *d* is delay day, in this model, this value is 14.

Construct these three models and estimated coefficients and p value can be calculated

as below.

table 1 estimated coefficients and p value of two models

	Non-delay model		Delay model		Flexible delay model	
	estimated coefficients	P value	estimated coefficients	P value	estimated coefficients	P value
Intercept	-1.494e+03	0	-7.238e+02	0	1.088e+03	0
t	2.319e+03	0	2.569e+03	0	-8.047e+00	0
t^2	-2.499e+01	0	-1.239e+01	0	1.937e-02	0
x_{vac}	1.422e-03	0	7.845e-04	0	-1.562e-02	0
x_{icu}	2.874e-01	0.239	1.421e-01	0	-1.472e-04	0.0165
x_{test}	-2.408e-02	0	-8.287e-03	0.0115	1.324e-02	0
x_{vac}^2	-3.195e-10	0.603	-1.838e-10	0	0	0
x_{icu}^2	-5.182e-06	0.654	-1.782e-06	0.0088	7.838e-07	0
x_{test}^2	9.670e-08	0.561	4.882e-08	0	-1.710e-08	0

From table 1 we can conclude that p values of delay and flexible delay model are all below 0.05 which indicates that all coefficients are significant. By checking RSE, R square and p value of the whole models, we can get table 2 as below.

table 2 RSE, R square and p value of the whole models

	Non-delay model	Delay model	Flexible delay model
RSE	247.9	192.8	110.2
R square	0.7472	0.9027	0.9816
P value	0	0	0
AIC	4999.913	4443.962	3915.143

From table 2, all models are significant, RSE of non-delay model is 247.9, bigger than 192.8, flexible delay model has the least RSE, which is 110.2. The R square value of non-delay model is 0.7472, it means that 74.72% variation can be explained by non-delay model, flexible delay model has a R square of 0.9816, which is bigger than non-delay model and delay model. AIC of flexible delay model is 3915.143, which is smaller than what non-delay and delay model is. We can conclude that flexible delay model is better than non-delay and delay model.

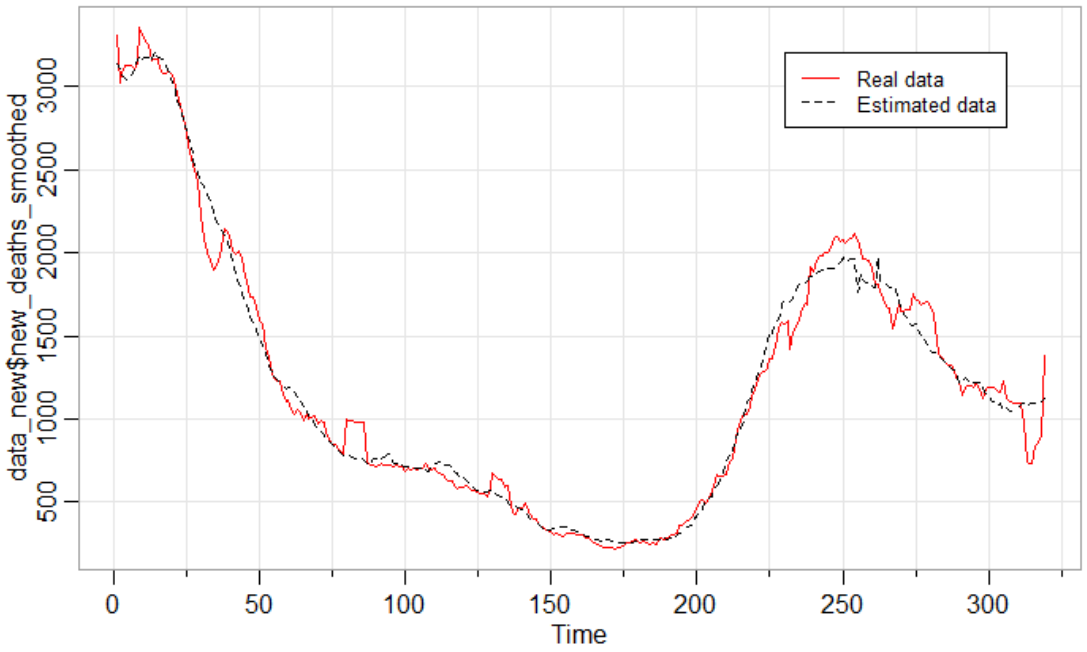

Figure 3 Estimated data and real data

Figure 3 shows the estimated curve and real curve of daily deaths from flexible delay model, the regression looks great.

2.3 Residual Analysis

Residuals are differences between the one-step-predicted output from the model and the measured output from the validation data set [3]. In this section, residuals are checked to discover further information.

Figure 4 shows the residuals plot, acf of residuals and histogram of residuals.

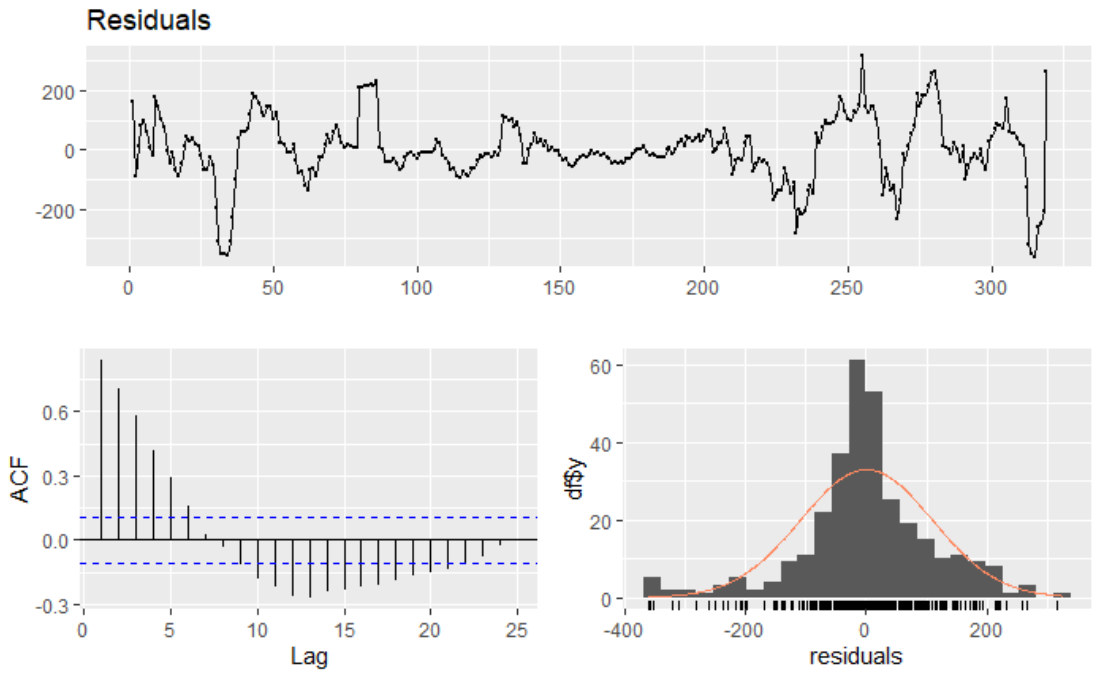

figure 4 Residual analysis for flexible delay model

By conducting Ljung-Box test, p value is less than 0.05, we reject the hypothesis. So, the residuals are not a white noise.

Objective errors in the time series of daily deaths are unavoidable due to several corrections in the number of deaths, and this is why the residuals do not look like white noise.

3. Covid-19 seasonal time series analysis

3.1 Data overview

We check the daily new cases of covid-19 in the United States, to avoid errors in data pre-processing, I select data from April 5, 2020 to December 2, 2021, which means I ignore data with a value of 0.

Figure 1 shows the daily new confirmed COVID-19 cases and logarithmization per million people in the United States. Looking at the first graph in Figure 1, I have used logarithmization for the data in order to reduce the effect of the magnitude.

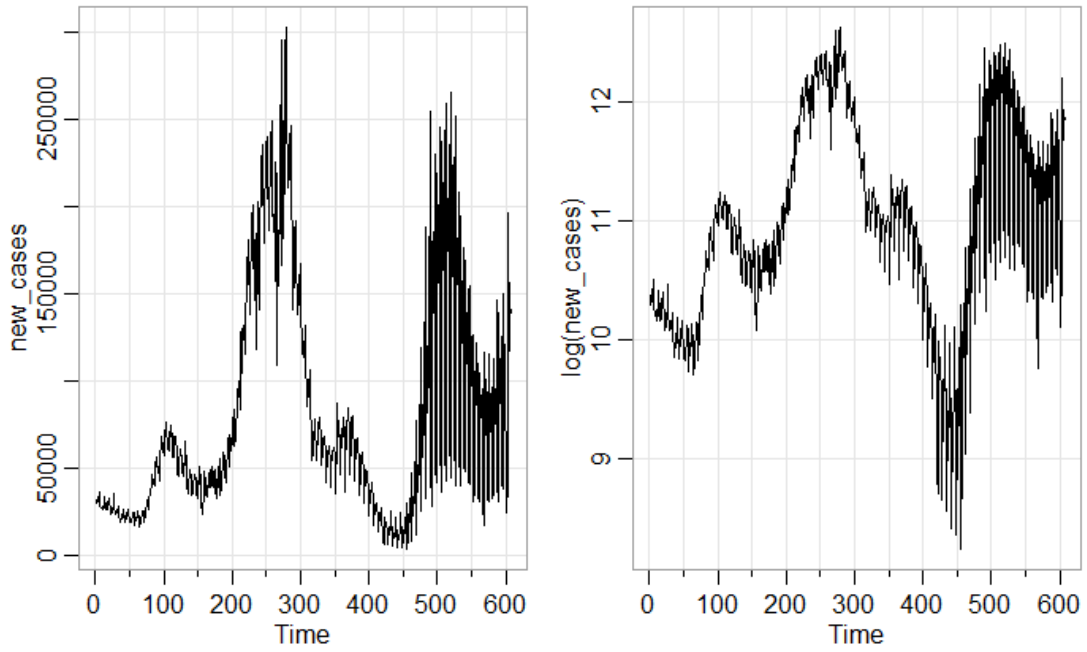

Figure 5 Daily new confirmed COVID-19 cases and logarithmization per million people in USA

3.2 Model Selection

I construct the model using logarithmic data, firstly I plot the acf and pacf of the data. In figure 2, the first graph shows the acf of the data after logarithmization, in this graph, the value of acf decreases slowly, which indicates an order 1 difference. The second graph shows the pacf of the data after logarithmization. From the acf and pacf plot which comes from logarithmic data, the value, also the lines has an interval pattern that approximately satisfies the equivariance series, which means it is a seasonal time series data with an interval of 7 days.

So, I use an order 7 after an order 1 difference. The third and the fourth graph is the acf and pacf of logarithmic data after order 1 and 7 difference.

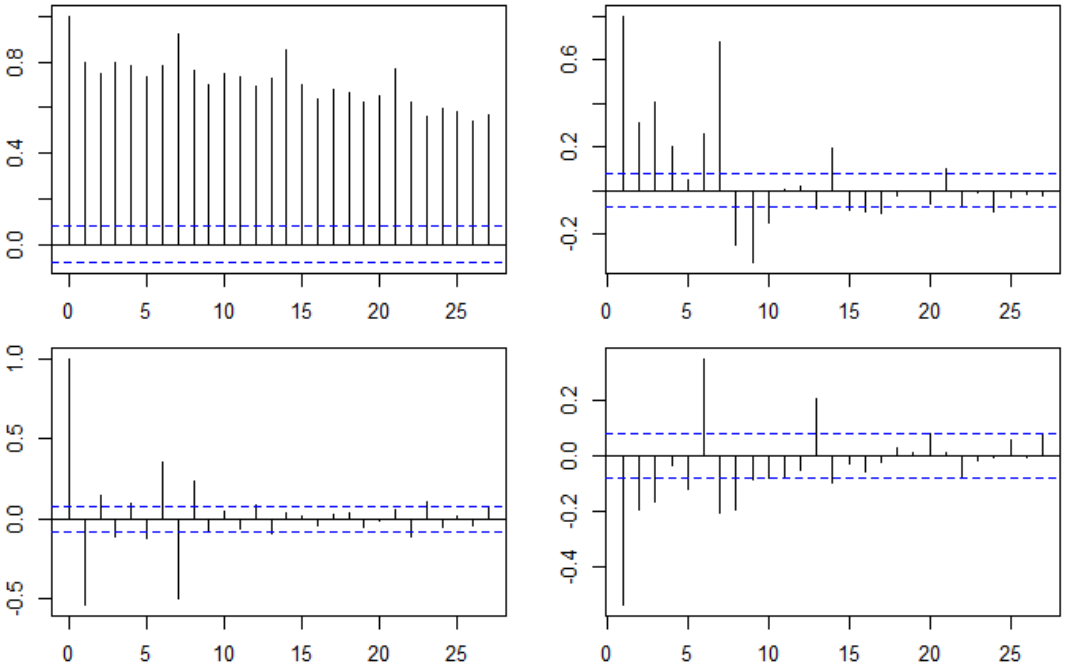

Figure 6 Acf and pacf of raw data and data after order 1 and 7 difference

I first check seasonal model, pacf tails off, acf cuts off, when we check non-seasonal model, pacf tails off and acf tails off, it suggests ARMA model.

The acf and pacf suggest an ARIMA model in non-seasonal model, and suggest a MA model in seasonal model.

In order to find the difference in seasonal model, the logarithmized and order 1 difference data is drawn as below.

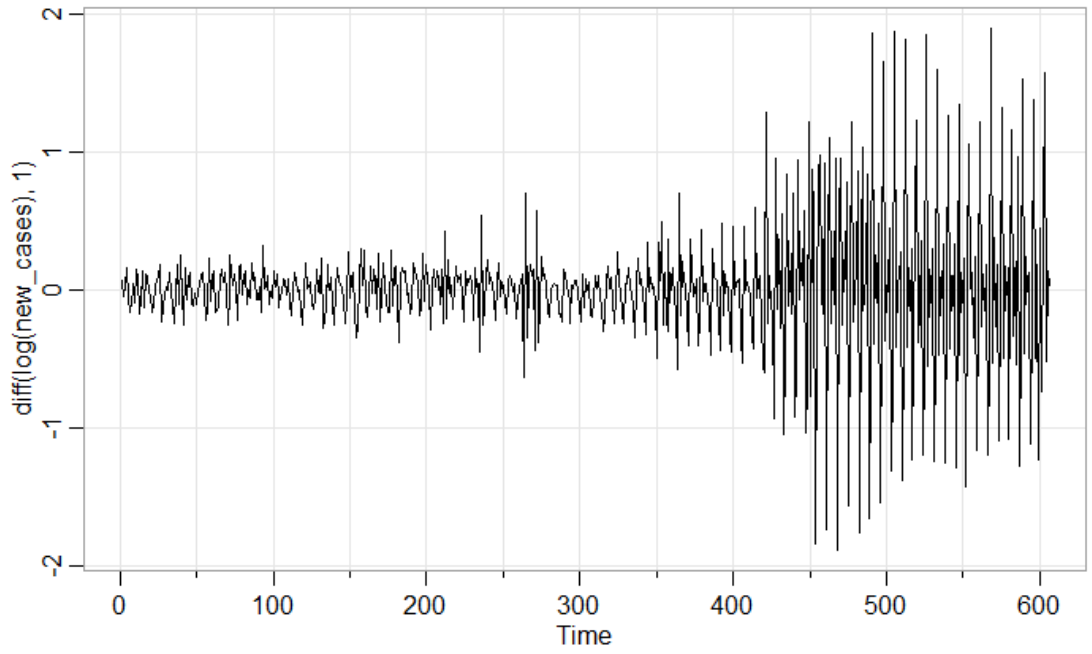

Figure 7 Logarithmic data after order 1 difference

Figure 7 shows the data after logarithmization and an order 1 difference. After difference, it looks like there's upper trend. So, in seasonal model, it should include an order 1 difference.

As a result, several SARIMA can be included, to find an optimal SARIMA model, I put the data into each model for calculation and compare between models. The AIC and BIC of each model shows below.

table 3 AIC and BIC of different models

Model	AIC	BIC
SARIMA(1,1,1)(0,1,1)$_7$	-0.4206759	-0.3916163
SARIMA(1,1,1)(0,1,2)$_7$	-0.4215356	-0.3852112
SARIMA(1,1,1)(1,1,1)$_7$	-0.4218981	-0.3855736
SARIMA(1,1,2)(1,1,1)$_7$	-0.423974	-0.3803846
SARIMA(1,1,2)(0,1,1)$_7$	-0.424726	-0.3884015
SARIMA(1,1,2)(0,1,2)$_7$	-0.4237542	-0.3801648
SARIMA(2,1,1)(0,1,1)$_7$	-0.4192789	-0.3829544

From table 1, I finally choose SARIMA(1,1,2)(0,1,1)$_7$ to construct the model, the next step is to check the reasonableness of the models.

Figure 8 is the validation of the SARIMA(1,1,2)(0,1,1)$_7$ model

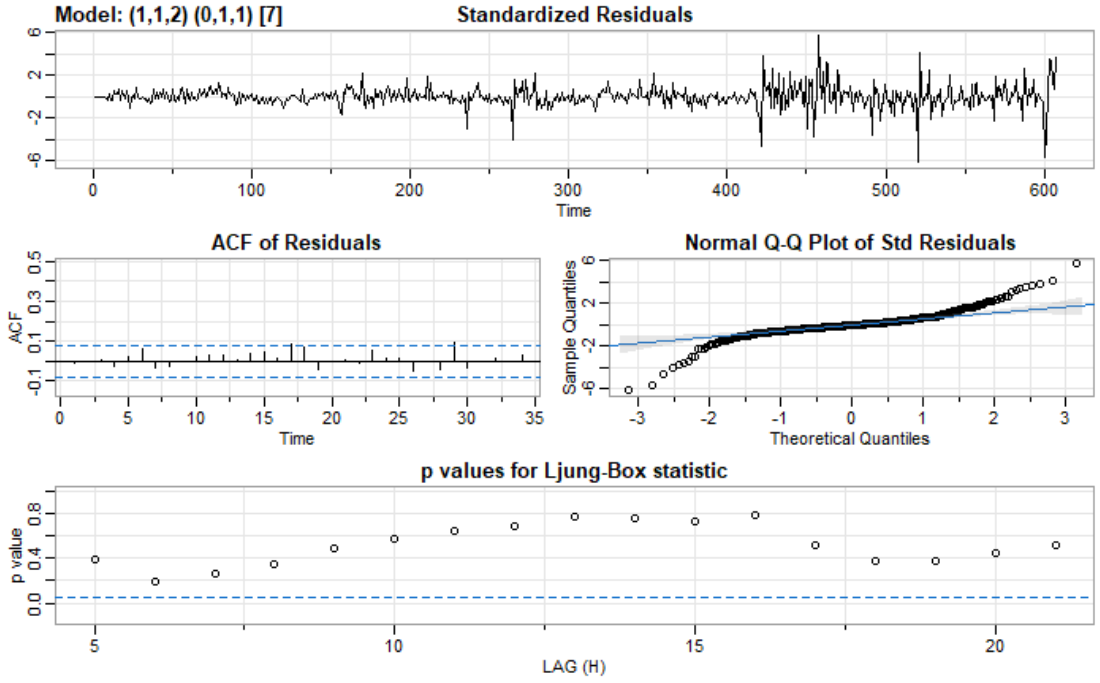

Figure 8 Validation of the models

In Standardized Residuals plot, the residuals tend to look like white noise, but at the quadrature, the amplitude of the residuals increases. In ACF of Residuals plot, all acf are inside the blue line, this plot suggest that we have white noise. In normal QQ plot, all points follow the straight line, it's normal distribution. In the plot of p values for Ljung-Box statistics, all points are above the blue line, which is bigger than 0.05.

Also, by checking t-table, the estimate value and p-value of each parameter shows in table 2.

table 4 T-table of the model

Parameters	Estimate value	P-value
Ar1	-0.8681	0
Ma1	0.1894	0.0493
Ma2	-0.4976	0
Sma1	-0.5964	0

All p-values are below 0.05, all parameters are significant. So, we accept SARIMA(1,1,2)$(0,1,1)_7$.

3.3 Forecast

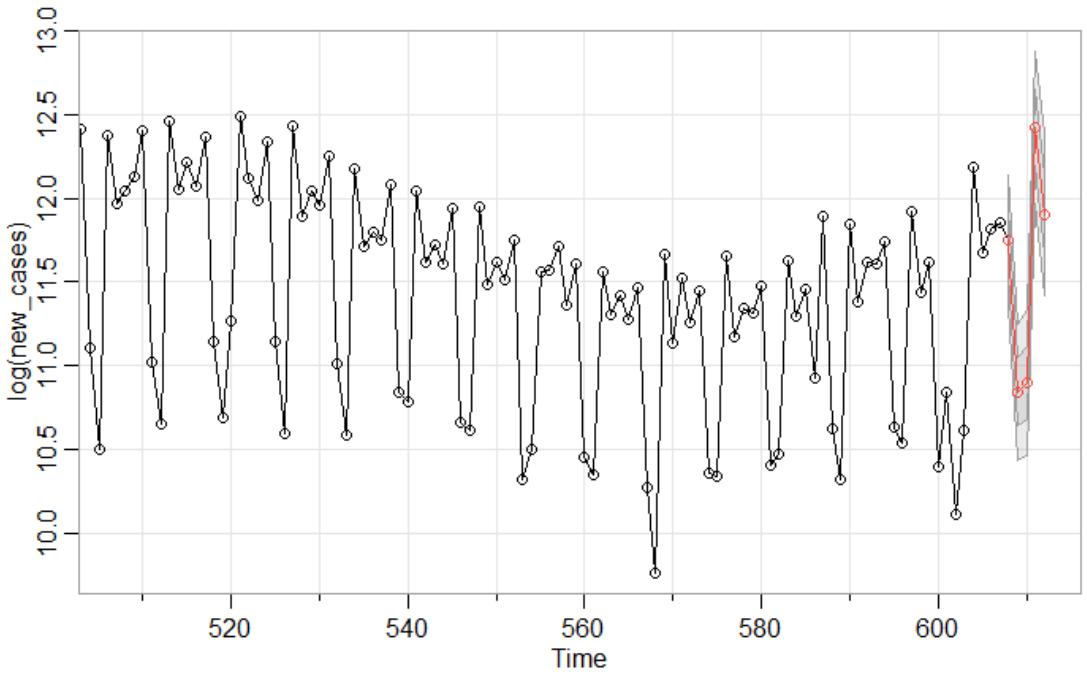

Figure 9 Future 5 days prediction

Figure 9 is the plot for the future 5 days by using SARIMA(1,1,2)(0,1,1)7, it is clear that the prediction value also has seasonal signal.

table 5 Prediction value and standard error of the model

	1	2	3	4	5
predict	126798.64	51221.02	54233.08	249472.01	147557.31
SE	0.1930159	0.2027375	0.2178350	0.2271427	0.2400525

Figure 5 and Table 3 shows the prediction values of the future 5 days covid-19 newly confirmed data and corresponding standard errors.

4. Results

In this project, I explore the regression and SARIMA models in covid-19 datasets.

In the regression model section, I used a quadratic polynomial regression model with flexible delay. After calculation, all parameters in the model are significant. After covariance analysis, the hypothesis that the covariance is white noise is rejected, and this problem may be caused by the correction of the number of deaths declared by the CDC.

In SARIMA model, by using logarithmization, I remove the trend, and then find a cycle to get seasonal information. The parameters of the SARIMA model can be found by acf and pacf of the dataset after difference. All coefficients are significant. And the prediction of future five days is conducted and drawn, standard error are within the acceptable range.

5. Discussion

This project has a number of issues that require follow-up research. In the regression model, I tried many different models, such as a once linear, up to three times polynomial model, with and without delay, but was never able to get the residuals detected as white noise. I think this is due to the objective error when counting the data. In a follow up study, we can look at the timing of the CDC correction for the number of deaths to see the residuals. Furthermore, we can use this to check the outliers of the dataset, which will help us justify the authenticity of the data.

References

[1] Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021)

[2] Covid, C. D. C., et al. "Characteristics of health care personnel with COVID-19—United States, February 12–April 9, 2020." Morbidity and Mortality Weekly Report 69.15 (2020): 477.

[3] https://www.mathworks.com/help/ident/ug/what-is-residual-analysis.html