Stock Price Forecasting in ARIMA Model by Evolutionary Regressor Selection in Covid-19 Outbreak

Xue (Kristi) Gong[a]

a. Information and Computer Sciences, University of Hawaii at Manoa

Abstract In this research, the relationship between the statistics of Coronaviruses and stock price during the Covid-19 outbreak is investigated. I have forecast the stock index price (Dow Jones Index) during the period by using the ARIMA model. Moreover, to improve the performance of ARIMA model, I select the best regressors by using the evolutionary computation. The result shows that the model with the EC regressors improved the stock market forecasts. The AR-1, AR-2 and also the MA-1 terms can best explained the stock index return in this period.

Keywords Covid-19, Dow Jones Index, ARIMA Model, Evolutionary Algorithm, Stock Market

1 Introduction

The world is shocked by covid-19, which is a new virus belongs to Corona virus's family which is transmitter from animal to human and was recognised in China, in December 2019. This virus cause serious illness and death around the world, until now, it has spread in 210 Countries and Territories in the world have reported a total of 2,332,004 confirmed cases of the COVID-19 and a death toll of 160,767 deaths.

Undoubtedly, this global pandemic has changed the world, people's lifestyle and also the economics, it has an unprecedented impact on stock markets. From February 24 to March 24, 2020, there were 22 trading days and 18 market jumps—more than any other period in history with the same number of trading days. (Kellogg Insight, 2020)

The situations may be better in the future, however, before the vaccine is invented and used, we need to understand the relationship between the covid-19 and stock market prices, what and to what extent is the impact of the covid-19 to the stock market. Is the confirmed cases or death rate of covid-19 drive the markets? If so, if the recovered cases make the market less volatile. Are the "fears" of disease drive the stock market? not the virus itself, therefore maybe we can give some more rational information to the public and spread the correct knowledge about this disease.

Xue (Kristi) Gong. *Stock Price Forecasting in ARIMA Model by Evolutionary Regressor Selection in Covid-19 Outbreak*. . In ICS674 Final Project, pages 1–7.

In this paper, I use the evolutionary computation to select the best regressors in ARIMIA models, which also included exogenous regressors, such as the confirmed cases, death rate, and also recovered rate during the period, Therefore I can analyse which factors are important to the stock market.

The paper is organized as follows. Section 2 shows ARIMA approach and evolutionary computation to selected regressors Section 3 presents data I use in this study. Section 4 gives the results and the conclusions are written in section 5.

2 ARIMA Approach and Evolutionary Algorithm to Select Regressors

2.1 ARIMA Approach for Stock Price Prediction

Auto Regressive Integrated Moving Average(ARIMA) is a time series models that explains a given data by its own past values, that is its lags and lagged forecast errors. In the other hand, this model uses the past to predict the future.

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 z_{t-1} - ... - \theta_q z_{t-q}$$

where the p is the order of the AR term, Auto regressive term is the $y_{t-1}, \dots y_{t-p}$ in the formula above and z is the MA term, which is the moving average error term here. In this paper, I have added some more exogenous variables (ARIMA-X model), therefore, I rewrite the formula as

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 z_{t-1} - ... - \theta_q z_{t-q}$$

where x is the exogenous variable terms, such as confirmed cases (x_1), death cases (x_2), and recovered cases (x_3).

2.2 Evolutionary Computation for Variable Selection

Evolutionary Computation (EC) have been used before for optimization within time series forecasting in many different research. Lukoseviciute and Minvydas (2010) uses the Evolutionary algorithms for the selection of time lags for time series forecasting by fuzzy inference systems. Stoean *et.al.*, (2017) have done a research about stock price time series forecasting in ARIMA model, and select the most relative variables by evolutionary algorithm, the results show that by using the evolutionary algorithm, the forecasts accuracy is improved.

In this study, the EC used here does not change the ARIMA model, instead it selects the most important variables to explain the stock market price return. I measure the "important" variables here as Akaike Information Criterion (AIC) value, it is an estimator of out-of-sample prediction error which represent the quality of each model, the lower AIC value, the better the statistical models.

To my knowledge, there is one research (Stoean *et.al.*, 2017) use EC to select the relative variables in ARIMA model, their research is about the individual stock prices, but my research focus on predicting on the stock market index which could represent the business cycle in the extreme economic period.

The length of an individual gene corresponds to the number of regressors. When the value is 1, it indicates the attribute is taken into account, and 0 means the attribute is ignored. The fitness is measured by the AIC value of the ARIMA-X model. I have

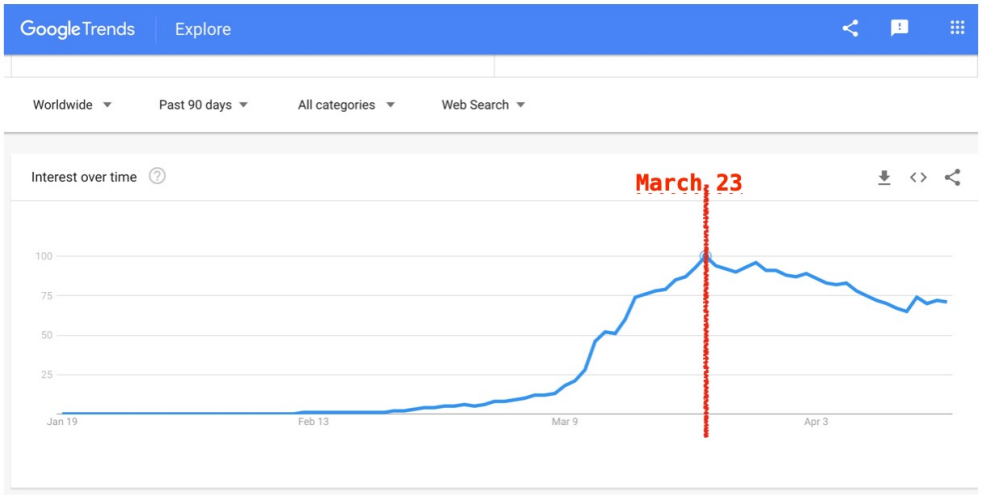

Figure 1 – "Covid-19" Google Trend

used the GA one-point recombination and bit-flip mutation variation to evolve to the next generation.

3 Data Description

I have collected the data from Jan 25, 2020 to April 17, 2020 (refer to Figure.2), the confirmed cases, recovered cases and deaths cases are shown in the left axis while the adjusted close price are shown in the right axis. Since the stock market will close at the weekend, to make the data parallel, I have deleted the weekend data in both series, the data description can be found as below Table. 1 and Table. 2. Besides, A measure of worldwide concern about the pandemic can be obtained from Google Trend, which gives the daily number of Google searches on keyword "COVID-19". The results are as in Figure 1, there is a peak date in March 23.

Table 1. Dow Jones Price (p_t)and Covid-19 Statistics

	Confirmed	Recovered	Deaths	Adjust close price
Total Observation	61	61	61	61
Mean	434811.705	108302.082	20793.934	25325.641
Stdev	636187.989	148433.343	35876.525	3425.653
Min	555	28	17	18591.929
Max	2240190	568343	134176	29551.419

Since the ARIMA model requires data to be stationary, to make the data series iid. I have preprocess the data by differencing it, and then divide by the previous term, I get the stationary data: $r_t = (p_t - p_{t-1})/p_{t-1}$

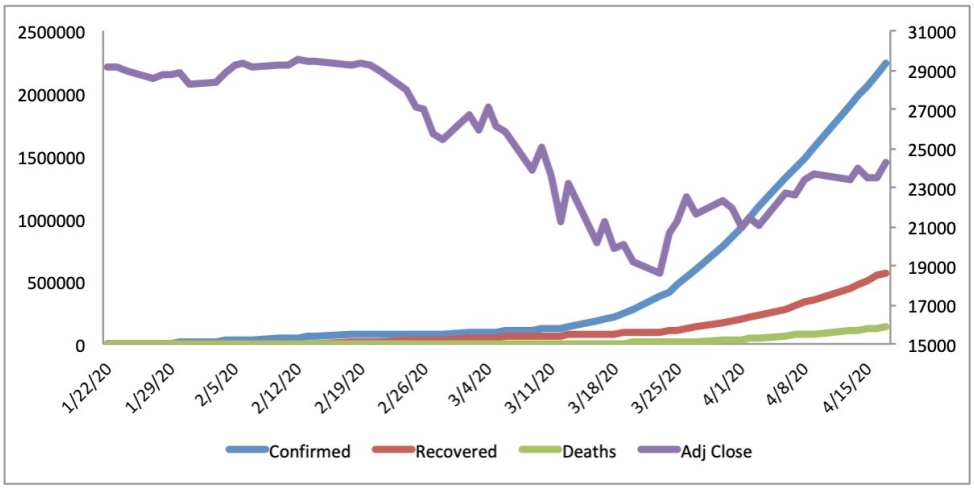

Figure 2 – The Dow Jones Price and Covid-19 Statistics

Table 2. Dow Jones Price Return (r_t) and Covid-19 Statistics

	Confirmed	Recovered	Deaths	Adjust close price
Total Observation	60	60	60	60
Mean	0.173	0.203	0.183	-0.002
Stdev	0.312	0.290	0.296	0.0416
Min	0.006	0.0197	0.004	-0.129
Max	2.110	1.806	2.153	0.113

And then I get the correlation between the exogenous variables and dependent variables in Table.3.

Table 3. The Pearson Correlation between the Covid-19 and Adjusted Close Price

Whole Period			
	Confirmed	Recovered	Deaths
Adjust close price	-0.461	-0.489	-0.381
Period before March 23			
	Confirmed	Recovered	Deaths
Adjust close price	-0.902	-0.944	-0.921
Period after March 24			
	Confirmed	Recovered	Deaths
Adjust close price	0.826	0.809	0.816

The Google Trend data for the number of searches on "COVID-19", Figure. 1, shows a peak at March 23. This gives us an independent basis to divide the time series data into two parts, March 23 and before, and March 24 and after. This division is not used in the evolutionary algorithm, from the data I can see that there is negative correlation between the stock market index and the covid-19 related statistics. However, recall the figure in google trend above, when I separate the data into two period at March 23rd. I have observed something unusual, before the March 23, there is strong negative correlation between the covid-19 and stock market price, but after that it seems that the stock market start to bounce back, therefore the positive trend shows.

4 ARIMA Model and Evolutionary Algorithm Results

The total data observation is 61, but since I use the lag terms, the total data is 55 observations. The first 35 observations are considered in the training set, while the rest represent the test part (that is 20 observations).

As mentioned in section above, beside the 3 exogenous indicators, there are 5 lag AR terms, and 5 lag MA terms obtained from the daily return. This conducts to 13 indicators in total, so the GA encoding will have 13 binary genes, each corresponding to taking or not (1 or 0) into account the regressor for building the model. A population size of 50 individuals is considered, the mutation probability is taken as 0.3 and the iterative process stops after 50 steps.

The training set is used for building the ARIMA model. The genes from the GA individual that are equal to 1 decide the representative predictors that are engaged in constructing the model. I have used the AIC to measure the fitness of the model The ARIMA-X models are running in Python notebook in Google Colab. As we know, the Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models.

Generally, the AIC is calculated as:

$$AIC = 2k - 2ln(L)$$

where k is the number of estimated parameters in the model. L is the maximized likelihood function for the estimated model.

To be more specific in our model, let u be the bit string of regressor usage. AIC is a function of u.

$$AIC(u) = 2k - 2ln(L(u))$$

The maximum likelihood for the model, $L(u)$ is calculated as follows:

$$L(u) = \sum (y_t - \hat{y}_t(u))^2/n$$

where $\hat{y_t} = u_1 * \beta_1 x_{1t} + u_2 * \beta_2 x_{2t} + u_3 * \beta_3 x_{3t} + u_4 * \phi_1 y_{t-1} + ... + u_8 * \phi_5 y_{t-5} - u_9 * \theta_1 z_{t-1} - ... - u_{13} * \theta_q 5_{t-5}$ in my case.

4.1 Evolutionary Computation Result

The details of the implementation of Evolutionary Algorithm can be found in the appendix, the best, worst and average fitness is in the Figure.3. Here I have shown one of the EC evlove, the best, worst and average fitnesses quickly dropped to the -2200. And the fitness gene is [0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0], the respective variables are AR-1, AR-2, and MA-1 term.

To figure out the performance of EC evolve, I have done a exhaustive global search, I tried out all of combinations of different regressors (totally 8191 possible combinations), and calcuate the AIC for each model (the detail of the implementation can be found in the Appendix), the results of selected regressors are the same as the most run in EV results, the corresponding AIC is -2248.426. However by using the exhastive search, I have used up 240.43 seconds in Google Colab by my labtop, while

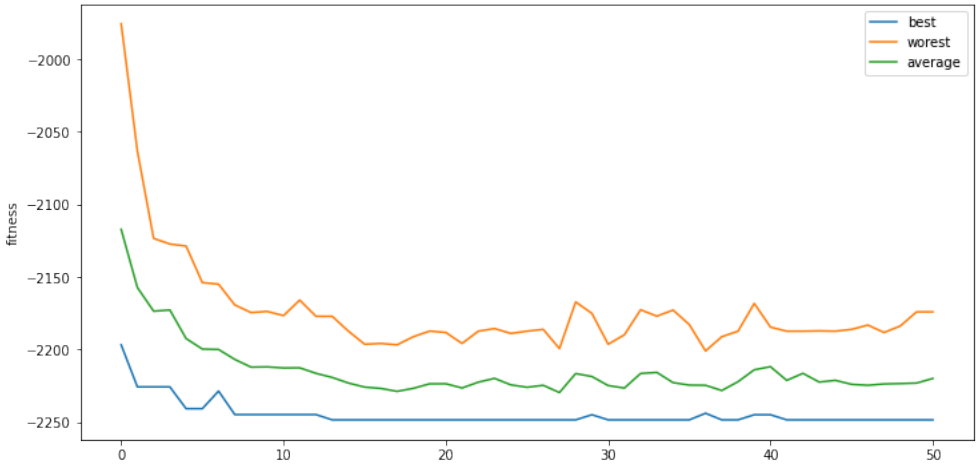

Figure 3 – The Best, Average and Worest Fitness along the Iterations

by using EC, the time for 50 iterations are just 135.57 seconds, which is more than 100 seconds less than exhaustive search time.

Also I have run EC algorithm for 50 different times with different random initial values, each time is 50 iterations, 44 out of 50 times (88% of total runs) the EC achieved the global best. I have summarized the results as Table 3:

Table 4. The Results of EC Evolve with Different Initial Values

Selected Regressor	Frequency
[4, 5, 8]	44
[4, 8, 10]	2
[8]	1
[5, 8]	1
[3,5,8]	1
[8,9,10]	1

4.2 ARIMA-X Model Result

From the evolutionary algorithm results, I use the selected regressors to do the forecasts in ARIMA model. Moreover, I have compared the results with the ARIMA(1,1) and ARIMA(1,1)-X models.

The ARIMA(1,1) can be expressed as:

$$y_t = \phi_1 y_{t-1} - \theta_1 z_{t-1}$$

And ARIMA(1,1)-X models is:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \phi_1 y_{t-1} - \theta_1 z_{t-1}$$

The details of the Models can be refered to the section 3. SSE obtained from the ARIMA model with the best GA solution is as below:

Table 5. The Comparison of predicted MSE of 3 Models

	Model with best GA regressors	ARMA(1,1)	ARIMA(1,1)-X
SSE	0.0795*	0.0903	0.0967

The results show that the ARIMA model with GA-selected regressors outperformed other two other popular models.

5 Conclusion

In this paper, I have collected stock market data and three different statistics from the Covid-19 break, that is, confirmed cases, recovered cases, and deaths. I have used the evolutionary algorithm to select the best regressors in ARIMA model based on its fitness, AIC value, that is the AR-1, AR-2 and MA-1 term and then use these regressors to forecast the stock market price return. It turns out the ARIMA model with EA regressors make the most accurate forecasts compared with the other two classic models.

6 Reference

1. Kellogg Insight, 2020, The Unprecedented Stock-Market Reaction to COVID-19: https://insight.kellogg.northwestern.edu/article/what-explains-the-unprecedented-stock-market-reaction-to-covid-19

2. Lukoseviciute, Kristina, and Minvydas Ragulskis. "Evolutionary algorithms for the selection of time lags for time series forecasting by fuzzy inference systems." Neurocomputing 73, no. 10-12 (2010): 2077-2088.

3. Stoean, Ruxandra, Catalin Stoean, and Adrian Sandita. "Evolutionary regressor selection in ARIMA model for stock price time series forecasting." In International Conference on Intelligent Decision Technologies, pp. 117-126. Springer, Cham, 2017.

Acknowledgments Thanks for Prof Lee's kind help through the whole course and whole semester!