Senior Honor Thesis Project

Preliminary Results

Rui Li

11/30/2020

1. INTRODUCTION

In 2020, a novel coronavirus named COVID-19 ravages the world. The first case of COVID-19 was reported on December 27th, 2019, in Wuhan, China, and was recognized as a pandemic in March 2020 by the World Health Organization (Qu, Cao, and Chen 2020). Due to COVID-19's devastating effect on human health and 1.47 million death till November 30th, it is essential to understand the spreading pattern of COVID-19. This project will predict the new COVID-19 case number of 212 countries by computing the new case number difference and adding to the current case number. We will conduct preliminary analysis by conducting linear regression models and then employ time series analysis to build specific models for 212 countries. We hope our results could provide insights to public health specialists and help them better manage the pandemic.

Data Sources

The data of the daily number of new reported cases of COVID-19 by country worldwide is obtained from the European Centre for Disease Prevention and Control (ECDC 2020). The dataset consists of daily data of 212 counties from January 1st, 2020, to the newest date. It summarizes daily COVID-19's case and death number of each country and the population data in 2019, from the World Bank. *Figure 1.1-1.4* visualize the COVID-19 dataset. Generally, the Americas and tropical regions have more severe epidemics, while Asian countries have better control over the disease.

Figure 1.1 The Logarithm of World Covid-19 Case Number

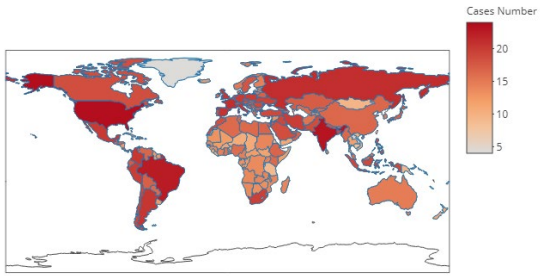


Figure 2.2 The Logarithm of World Covid-19 Death Number

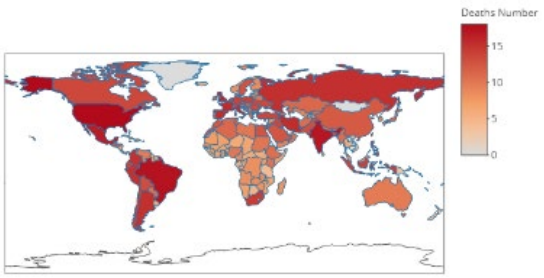


Figure 1.3 The Ratio of World Covid-19 Death to Case Number

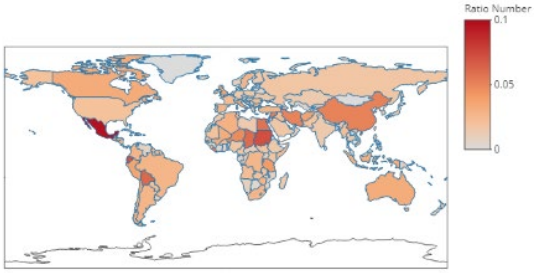
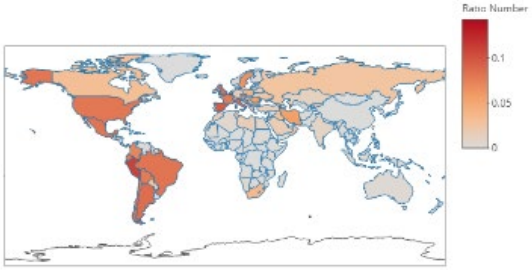


Figure 1.4 The Ratio of World Covid-19 Death/Pop*100



We also utilize the 'Land Area' dataset (WB 2018) from the World Bank, which describes the land area (square miles) of 201 countries in 2018. *Figure 1.5* and *Figure 1.6* presents the world population and the world population density in 2019.

Figure 1.5 The Logarithm of World Population in 2019

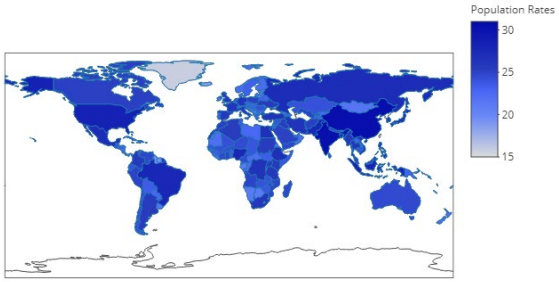
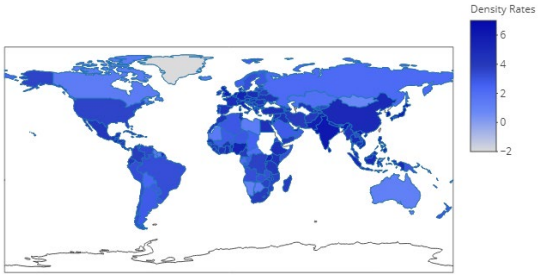


Figure 1.6 The Logarithm of World Population Density in 2019



2. METHODS

The primary method of predicting the new coronavirus case number in this project is to compute the new case number difference between two consecutive days (Dickey and Pantula 2002) by fitting linear regression models over the case number difference and death number difference of the past 14 consecutive days, as well as the population and the population density of each country in 2019. This paper tests four different models and compare them to find the best performance model.

Modeling

Let $C_t =$ be a column vector with the new coronavirus case number of each country at time t reversely, where $t = i, (i = 0,1,2 \cdots 15)$, representing i day past the current day $t = 0$. Then the response of the models is $C_0 - C_1$, and $C_i - C_{i+1}$ is the case number difference on day $i, (i = 1,2, \cdots, 15)$. The expression of the first model is

$$C_0 - C_1 = \alpha_0 + \sum_{i=1}^{14} \alpha_i \, (C_i - C_{i+1}), \tag{2.1}$$

where α_i are the estimators of i_{th} day's case number difference.

Let D_t be a column vector with new coronavirus death number of each country at time t reversely, where $t = i, (i = 0,1,2 \cdots 15)$, representing i day past the current day $t = 0$. For the second model, we want to consider the death number difference of the past 14 consecutive days as predictors, among which $D_j - D_{j+1}$ is the death number difference on day $j, (j = 1,2 \cdots 15)$. The expression of the second model is

$$C_0 - C_1 = \beta_0 + \sum_{i=1}^{14} \beta_i \, (C_i - C_{i+1}) + \sum_{j=1}^{14} \beta_{j+14} \, (D_j - D_{j+1}), \tag{2.2}$$

where $\beta_0 =$ intercept, $\beta_i, (i = 1,2, \cdots, 14) =$ the estimators of i_{th} day's case number difference, and $\beta_{j+14}, (j = 1,2, \cdots, 14) =$ the estimators of j_{th} day death number difference.

The third and fourth models respectively add the population and the population density of each country in 2019 to predictors. Let P be a column vector of every country's population in 2019, and $P_{density}$ be a column vector of every country's population density in 2019. Then the expression of the third and fourth models are

$$C_0 - C_1 = \gamma_0 + \sum_{i=1}^{14} \gamma_i \left(C_i - C_{i+1}\right) + \sum_{j=1}^{14} \gamma_{j+14} \left(D_j - D_{j+1}\right) + \gamma_{15} P, \tag{2.3}$$

where γ_0 = intercept, $\gamma_i, (i = 1,2,\cdots,14)$ = the estimators of i_{th} day case number difference, $\gamma_{j+14}, (j = 1,2,\cdots,14)$ = the estimators of j_{th} day death number difference, and γ_{15} = estimator of population, and

$$C_0 - C_1 = \kappa_0 + \sum_{i=1}^{14} \kappa_i \left(C_i - C_{i+1}\right) + \sum_{j=1}^{14} \kappa_{j+14} \left(D_j - D_{j+1}\right) + \kappa_{15} P_{density}, \tag{2.4}$$

where κ_0 = intercept, $\kappa_i, (i = 1,2,\cdots,14)$ = the estimators of i_{th} day case number difference, $\kappa_{j+14}, (j = 1,2,\cdots,14)$ = the estimators of j_{th} day death number difference, and κ_{15} = estimator of population density.

Adjust Models

Redefine $C_{diff}(i) = C_i - C_{i+1}, (i = 0,1,2,\cdots,14)$, and $D_{diff}(j) = D_j - D_{j+1}, (j = 0,1,2,\cdots,14)$, where $C_{diff}(i)$ = the case number difference of the past i_{th} day, and $D_{diff}(j)$ = the death number difference of the past j_{th} day, $i, j = 0,1,2,\cdots,14$. Therefore, the equations (2.1), (2.2), (2.3), and (2.4) can be rewritten as

- **Model 1:** $C_{diff}(0) = \alpha_0 + \sum_{i=1}^{14} \alpha_i \, C_{diff}(i)$
- **Model 2:** $C_{diff}(0) = \beta_0 + \sum_{i=1}^{14} \beta_i \, C_{diff}(i) + \sum_{j=1}^{14} \beta_{j+14} \, D_{diff}(j)$
- **Model 3:** $C_{diff}(0) = \gamma_0 + \sum_{i=1}^{14} \gamma_i \, C_{diff}(i) + \sum_{j=1}^{14} \gamma_{j+14} \, D_{diff}(j) + \gamma_{15} P$
- **Model 4:** $C_{diff}(0) = \kappa_0 + \sum_{i=1}^{14} \kappa_i \, C_{diff}(i) + \sum_{j=1}^{14} \kappa_{j+14} \, D_{diff}(j) + \kappa_{15} P_{density}$

Regression Analysis

After setting up four models, linear regression analysis is conducted and employ *stepwise selection* to find compelling variables. The final step is to use *anova()* function in R to compare models and get the one providing the best parsimonious fit of the data. We also evaluate the final result with a diagnostic test.

3. RESULTS

The results of linear regression analysis are presenting in *Table 3.1* and *Table 3.4*. From the summary of *Model 1* in *Table 3.1*, we can conclude that the case number differences of the past 14 days are statistically significant to the new coronavirus case number difference. The result of *Model 2* from *Table 3.1* shows the death number differences of the past 14 days are statistically significant to the new coronavirus case number difference, excluding the 3_{th} and 4_{th} day. When comparing the performance of *Model 1* and *Model 2*, the ANOVA test results from *Table 3.2* indicating that the death number differences have significant effects on the model ($p < 2.2e\text{-}16$) and should not be eliminated. Therefore, *Model 2* has a better performance than *Model 1*.

The analytic results of *Model 3* from *Table 3.1* prove that each country's population in 2019 has a significant effect on the new coronavirus case number difference ($p < 0.01$). Besides, the ANOVA test results from *Table 3.3* further corroborate that *Model 3* is a better choice than *Model 2* ($P = 9.75e\text{-}05$), and the variable P is essential to the response.

Table 3.1: Linear Regression Results of Models 1 & 2 & 3

	Dependent variable:		
		$C_{diff}(0)$	
	(Model 1)	(Model 2)	(Model 3)
$C_{diff}(7)$	0.486***	0.462***	0.461***
	(0.005)	(0.005)	(0.005)
$C_{diff}(14)$	0.114***	0.110***	0.110***
	(0.003)	(0.004)	(0.004)
$C_{diff}(1)$	-0.544***	-0.551***	-0.552***
	(0.004)	(0.004)	(0.004)
$C_{diff}(8)$	0.359***	0.337***	0.336***
	(0.005)	(0.005)	(0.005)
$C_{diff}(2)$	-0.435***	-0.444***	-0.445***
	(0.005)	(0.005)	(0.005)
$C_{diff}(9)$	0.222***	0.211***	0.210***
	(0.004)	(0.004)	(0.004)
$C_{diff}(6)$	0.073***	0.055***	0.054***
	(0.005)	(0.005)	(0.005)
$C_{diff}(3)$	-0.276***	-0.282***	-0.283***
	(0.005)	(0.005)	(0.005)
$C_{diff}(10)$	0.111***	0.107***	0.107***
	(0.004)	(0.004)	(0.004)
$C_{diff}(12)$	-0.110***	-0.109***	-0.109***
	(0.003)	(0.004)	(0.004)
$C_{diff}(13)$	-0.111***	-0.114***	-0.114***
	(0.004)	(0.004)	(0.004)
$C_{diff}(4)$	-0.140***	-0.145***	-0.146***
	(0.004)	(0.004)	(0.004)
$C_{diff}(5)$	-0.071***	-0.081***	-0.082***
	(0.005)	(0.005)	(0.005)
$D_{diff}(8)$		2.100***	2.098***
		(0.110)	(0.110)
$D_{diff}(7)$		1.908***	1.907***

		(0.101)	(0.101)
$D_{diff}(6)$		1.313***	1.312***
		(0.090)	(0.090)
$D_{diff}(5)$		0.672***	0.671***
		(0.070)	(0.070)
$D_{diff}(9)$		1.565***	1.563***
		(0.114)	(0.114)
$D_{diff}(10)$		1.401***	1.399***
		(0.112)	(0.112)
$D_{diff}(11)$		1.389***	1.386***
		(0.107)	(0.107)
$D_{diff}(12)$		1.048***	1.047***
		(0.100)	(0.100)
$D_{diff}(13)$		0.900***	0.899***
		(0.089)	(0.089)
$D_{diff}(14)$		0.569***	0.569***
		(0.070)	(0.070)
P			0.00000***
			(0.00000)
$D_{diff}(2)$		0.282***	0.282***
		(0.063)	(0.063)
$D_{diff}(1)$		0.182***	0.181***
		(0.063)	(0.063)
Constant	14.180***	13.018***	9.417**
	(3.778)	(3.761)	(3.892)
Observations	52,962	52,962	52,962
R^2	0.585	0.589	0.589
Adjusted R^2	0.584	0.588	0.589
Residual Std. Error	867.294 (df = 52948)	863.136 (df = 52936)	863.040 (df = 52935)
F Statistic	5,731.463*** (df = 13; 52948)	3,030.079*** (df = 25; 52936)	2,914.683*** (df = 26; 52935)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.2: ANOVA Test Results of Comparing Models 1 & 2

Model Name	Res.DF	RSS	Df	Sum of Sq	F	Pr(>F)
1	52547	39.17				
2	52534	38.78	13	388052910	40.438	<2.2e-16 ***

Table 3.3: ANOVA Test Results of Comparing Models 2 & 3

Model Name	Res.DF	RSS	Df	Sum of Sq	F	Pr(>F)
2	52534	38.78				
3	52533	38.77	1	11207507	15.187	9.75e-05 ***

However, the results of *Model 4* from *Table 3.4* reveal that the population density in 2019 does not have too much impact on the new coronavirus case number difference, which is also evaluated by the ANOVA test from *Table 3.5* (P = 9.75e-05). Therefore, *Model 3* is the best model among the four.

Table 3.4: Linear Regression Results of Models 3 & 4

	Dependent variable:	
	$C_{diff}(0)$	
	(Model 3)	(Model 4)
$C_{diff}(7)$	0.461***	0.462***
	(0.005)	(0.005)
$C_{diff}(14)$	0.110***	0.110***
	(0.004)	(0.004)
$C_{diff}(1)$	-0.552***	-0.551***
	(0.004)	(0.004)
$C_{diff}(8)$	0.336***	0.337***
	(0.005)	(0.005)
$C_{diff}(2)$	-0.445***	-0.444***
	(0.005)	(0.005)
$C_{diff}(9)$	0.210***	0.211***
	(0.004)	(0.004)
$C_{diff}(6)$	0.054***	0.055***
	(0.005)	(0.005)
$C_{diff}(3)$	-0.283***	-0.282***
	(0.005)	(0.005)
$C_{diff}(10)$	0.107***	0.107***
	(0.004)	(0.004)

$C_{diff}(12)$	-0.109***	-0.109***
	(0.004)	(0.004)
$C_{diff}(13)$	-0.114***	-0.114***
	(0.004)	(0.004)
$C_{diff}(4)$	-0.146***	-0.145***
	(0.004)	(0.004)
$C_{diff}(5)$	-0.082***	-0.081***
	(0.005)	(0.005)
$C_{diff}(8)$	2.098***	2.100***
	(0.110)	(0.110)
$C_{diff}(7)$	1.907***	1.908***
	(0.101)	(0.101)
$C_{diff}(6)$	1.312***	1.313***
	(0.090)	(0.090)
$C_{diff}(5)$	0.671***	0.672***
	(0.070)	(0.070)
$C_{diff}(9)$	1.563***	1.565***
	(0.114)	(0.114)
$C_{diff}(10)$	1.399***	1.401***
	(0.112)	(0.112)
$C_{diff}(11)$	1.386***	1.389***
	(0.107)	(0.107)
$C_{diff}(12)$	1.047***	1.048***
	(0.100)	(0.100)
$C_{diff}(13)$	0.899***	0.900***
	(0.089)	(0.089)
$C_{diff}(14)$	0.569***	0.569***
	(0.070)	(0.070)
P	0.00000***	
	(0.00000)	
$C_{diff}(2)$	0.282***	0.282***
	(0.063)	(0.063)
$C_{diff}(1)$	0.181***	0.182***

	(0.063)	(0.063)
Constant	9.417**	13.018***
	(3.892)	(3.761)
Observations	52,962	52,962
R^2	0.589	0.589
Adjusted R^2	0.589	0.588
Residual Std. Error	863.040 (df = 52935)	863.136 (df = 52936)
F Statistic	2,914.683*** (df = 26; 52935)	3,030.079*** (df = 25; 52936)

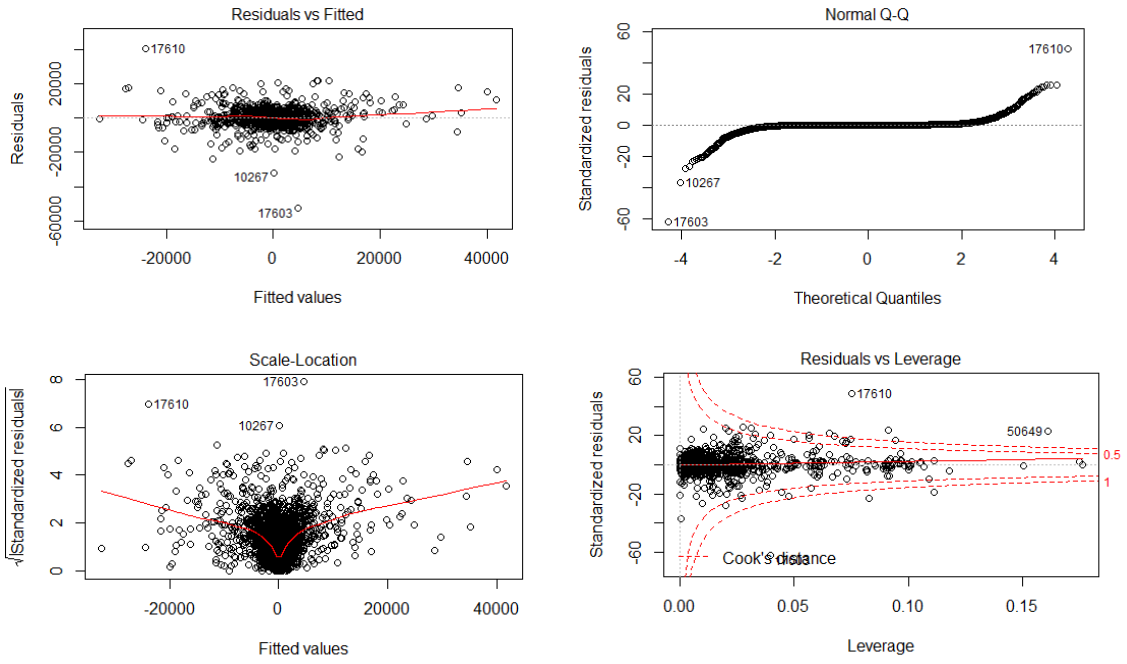
Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.5: ANOVA Test Results of Comparing Models 3 & 4

Model Name	Res.DF	RSS	Df	Sum of	F	Pr(>F)
3	52533	38.77				
4	52534	38.78	-1	-11207507	15.187	9.75e-05 ***

Unfortunately, when conducting the diagnostic test on *Model 3*, the results are not satisfactory. According to *Figure 3.1*, the variance of the data is not normally distributed, as the *Normal Q-Q* plot does not form a linear pattern. Additionally, there is a clear v-shape in the *Scale-Location* plot, showing that the residuals do not distribute randomly. Nevertheless, *Model 3* proves the strong influence that the past 14 days' case number difference, death number difference, and the population in 2019 have on the new coronavirus case difference. But for the next step of the research, it is necessary to update the design of *Model 3* to fit the data better.

Figure 3.1 Diagnostic Test Results



4. CONCLUSIONS & DISCUSSIONS

After conducting linear regression and ANOVA test, we conclude that *Model 3* is the best model among the four, which can be written as

$$
\begin{aligned}
&C_{diff}(0) \\
&= 10.232 + -0.560C_{diff}(1) - 0.430C_{diff}(2) - 0.258C_{diff}(3) - 0.146C_{diff}(4) \\
&- 0.082C_{diff}(5) + 0.054C_{diff}(6) + 0.463C_{diff}(7) + 0.330C_{diff}(8) + 0.198C_{diff}(9) \\
&+ 0.071C_{diff}(10) + 0.010C_{diff}(11) - 0.120C_{diff}(12) - 0.122C_{diff}(13) + 0.109C_{diff}(14) \\
&+ 0.176D_{diff}(1) + 0.333D_{diff}(2) + 0.638D_{diff}(5) + 1.272D_{diff}(6) + 1.948D_{diff}(7) \\
&+ 2.187D_{diff}(8) + 1.532D_{diff}(9) + 1.281D_{diff}(10) + 1.225D_{diff}(11) + 1.008D_{diff}(12) \\
&+ 0.863D_{diff}(13) + 0.573D_{diff}(14) + 1.032e - 07P
\end{aligned}
$$

While the results of *Model 3* confirm the assumption that the past 14 days' case number and death number difference, as well as the population in 2019, have strong relationships with the new coronavirus case difference, the diagnostic test and the estimators of the model reveal the deficiency of *Model 3*. Notably, the coefficient of the population is minimal ($P_{Coef} = 1.032e - 07$), and the population density shows no relationship with the responder, which is abnormal because the countries with higher population density are more susceptible to COVID-19.

One possible explanation of the deficiency is that *Model 3* is built to fit all countries with various behaviors, and a unified model considerably dampens the difference between each country and, therefore, decreases the efficiency of *Model 3*. We also need to consider other factors that could have significant impacts on the new COVID-19 case difference, such as GDP, hospital bed number, and ICU number of each country. Therefore, for the next step of the research, we should set up distinct models for each country and add other corresponding factors. We can then carefully examine the country with COVID-19 severe condition and check if the situation can be improved when increasing the coefficient of variables related to medical resources.

References

Dickey, David A, and Sastry G Pantula. 2002. "Determining the Order of Differencing in Autoregressive Processes." *Journal of Business & Economic Statistics* 20:1: 18 –24. https://doi.org/10.1198/073500102753410363.

ECDC. 2020. "The Daily Number of New Reported Cases of Covid-19 by Country Worldwide, Updated on November 30th." *European Centre for Disease Prevention and Control*. "https://opendata.ecdc.europa.eu/covid19/casedistribution/csv.

Qu, Jie-Ming, Bin Cao, and Rong-Chang Chen. 2020. *COVID-19: The Essentials of Prevention and Treatment*. Edited by Jie-Ming Qu, Bin Cao, and Rong-Chang Chen. Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-824003-8.09994-0.

WB. 2018. "Land Area." *The World Bank*. https://data.worldbank.org/indicator/AG.LND.TOTL.K2.