A Multiple Linear Regression Approach to Covid-19

Naomi Shields

STA4211

Abstract

This project aims to address the number of deaths occurring due to the covid-19 outbreak. There are many ongoing studies and tools ongoing since it is an evolving problem. This project uses data provided by Johns Hopkins University with a multiple linear regression approach. The model created uses the number of confirmed cases, number of recovered cases, and country and independent variables and the number of deaths as the dependent variable. Ultimately the model and data hold under all the assumptions and the model fits the data well.

Introduction

Since December 10th 2019, when one of the earliest known cases of covid-19 was detected, the virus has become an issue for much of the world's population. The World Health Organization stated that the virus was a public health problem that warrants international concern on January 30th 2020. Many countries then proceeded to institute measures to limit the spread of the virus, including the United States where most states issued "stay at home" orders by late March.

Johns Hopkins University offers an up to date public dataset[1] of all recorded cases of covid-19. For every serial number, they keep track of the observation date, province/state, country/region, confirmed number of cases, number of deaths, and number of recovered cases. For the purpose of this project, only the country, confirmed number of cases, number of deaths, and number of recovered cases will be considered. Furthermore, this project will only look at Mainland China, South Korea, and the United States in regards to county.

The data being used throughput this project is from the timespan of January 22nd 2020 – March 22nd 2020. Within this timepsan, there were 3,533 serial numbers containing the figures for the three countries. Figure 1.1 (left) displays the number of deaths against the number of confirmed cases by country and Figure 1.1 (right) displays number of deaths against recovered cases by country. By March 22nd the median number of confirmed cases was 62 and the mean number of cases was 1,057. Also by this date, the median number of deaths was 0 and the mean was 33.76. Also, the median number of recovered cases was 1 and the mean was 488.8.

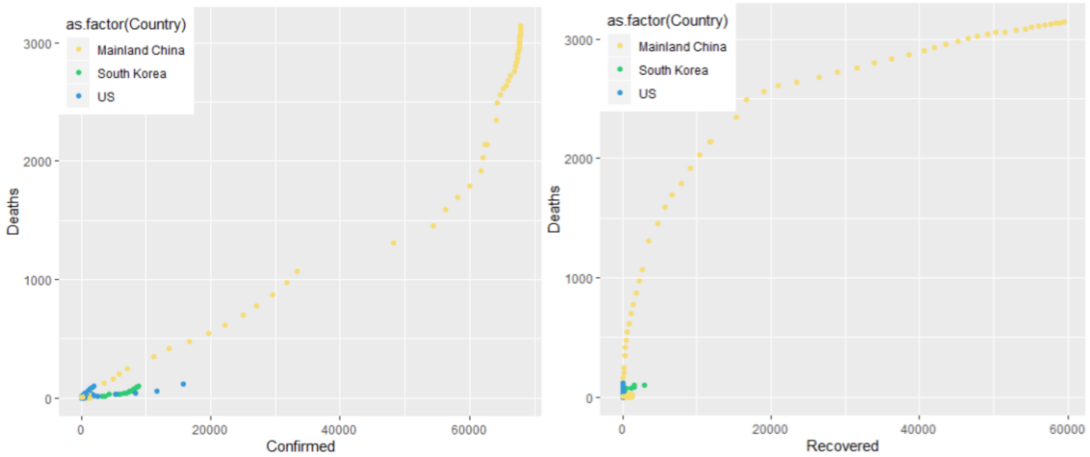

Figure 1.1

[1] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Description of Problem

As covid-19 has continued and spread throughout the world, it has presented many problems to our society. It has affected health care systems, governmental policies, economies, and people's health throughout the world. Although there are myriad problems that need to be addressed when considering covid-19, this project will be focusing on the latter of the issues mentioned, people's health.

Ultimately, this project sets out to explore what factors are useful predictors for the number of deaths occurring due to covid-19. Using the Johns Hopkins data, this project will be looking at the number of deaths based on country, number of confirmed cases, and number of recovered cases.

Literature Review

Since the outbreak has become an international concern, there have been many studies that model covid-19. Since there was not a lot of available data when the virus first began, there have been some studies that have tried to use statistical modeling to infer how many cases there were in the beginning of the outbreaks. Wu et al. [2] aimed to estimate the number of cases of the virus in Wuhan using data that provided the number of exported cases from Wuhan. They then forecasted the spread of the virus both nationally and globally. They eventually concluded that other major cities in China were experiencing outbreaks along with overseas major cities that had transportation links with China.

There have also been a number of predictive tools created to aid in providing estimates of the number of covid-19 cases. University of Pennsylvania built Covid-19 Hospital Impact Model for Epidemics (CHIME)[3] for hospitals to use throughout the course of the pandemic. The model uses the number of confirmed cases in the hospital's region and how many patients the hospital is currently treating to infer how many patients the hospital might receive.

Methodology and Assumptions Associated with Methodology

Within this project, a multiple linear regression approach will be used. Multiple linear regression models the linear relationship between a set of independent variables and a dependent variable. This project will use country, number of confirmed cases, and number of recovered cases as the independent variables. Both the number of confirmed cases and the number of recovered cases will be used as quantitative variables and country will be used as a qualitative variable. The dependent variable in this project will be the number of deaths.

There are several assumptions for the multiple linear regression approach. The first assumption is that there is a linear relationship between the dependent and independent variables. The second is that the residuals are normally distributed. The third is that there is no multicollinearity between the independent variables. The last assumption is that there is homoscedasticity, meaning that the variance of the errors is equally distributed for all the independent variables.

---

[2] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30260-9/fulltext
[3] https://penn-chime.phl.io/

Model Form

In this project, the first-order regression model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

Where:

X_{i1} = number of confirmed cases

X_{i2} = number of recovered cases

$X_{i3} = \begin{cases} 1 & \text{if South Korea is the country} \\ 0 & \text{otherwise} \end{cases}$

$X_{i4} = \begin{cases} 1 & \text{if US is the country} \\ 0 & \text{otherwise} \end{cases}$

Regression Coefficients:

Table 1.1 gives the regression coefficients and their respective standard errors and t-values when the data is put into the regression model.

Using the values from Table 1.1, the fitted equation is:

$$\hat{Y} = -12.02 + 0.02979X_1 + 0.02092X_2 - 53.05X_3 - 10.83X_4$$

Based on this equation, several interpretations can be made based on the β's. For fixed levels of recovered cases and country, we expect the number of confirmed cases to increase by 0.02979 for every one unit increase in the number of deaths. Additionally, for fixed levels of confirmed cases and country, we expect the number of recovered cases to increase by 0.02092 for every one unit increase in the number of deaths. Compared to Mainland China, we would expect South Korea to have 53.05 fewer deaths, on average, at the same level of confirmed cases and recovered cases. Also, compared to Mainland China, we would expect the US to have 10.83 fewer deaths, on average, at the same level of confirmed cases and recovered cases.

Table 1.1

	Estimate	***Std. Error***	***t-value***
Intercept	-12.02	.5239	-22.94
Number of Confirmed Case	.02979	.0001115	267.09
Number of Recovered Cases	.02092	.0001926	108.61
South Korea	-53.05	2.941	-18.04
US	10.83	.7605	14.24

Diagnostics

Checking Model Assumptions:

1. To check for linearity, the quantitative independent variables were plotted against the dependent variable. Figure 1.2 (left) shows the relationship between confirmed number of cases and number of deaths. The plot indicates a generally linear relationship between the variables. Figure 1.2 (right) shows the relationship the relationship between the number of recovered cases and the number of deaths. This also shows a linear relationship between the variables, so the data holds under this assumption.

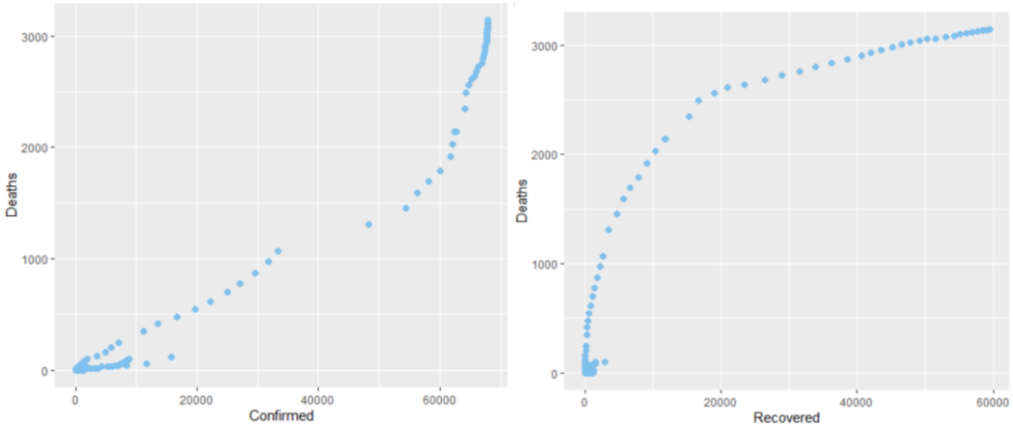

Figure 1.2

2. The normality of the errors was checked by the Q-Q Plot seen in Figure 1.3. Most of the points fall on the line, so this assumption holds.

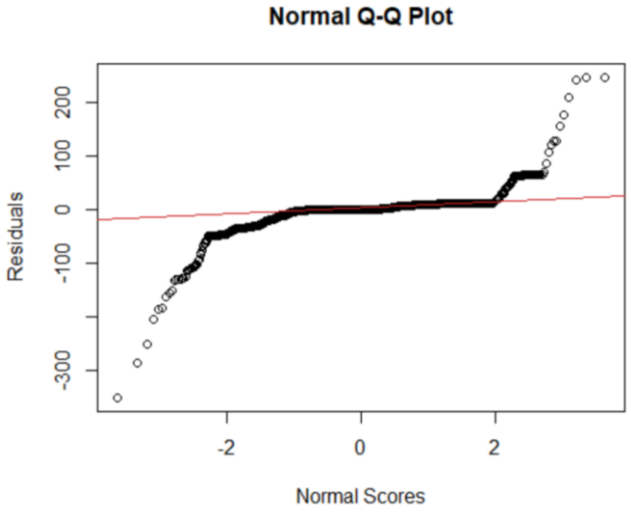

Figure 1.3

3. In order to ensure that there was no multicollinearity the Variance Inflation Factor(VIF) was checked. The number of confirmed cases had a VIF of 4.32, number of recovered

cases had a VIF of 4.30, and country had a VIF of 1.02. Since none of the VIFs were above 10, the assumption of no multicollinearity holds.

4. Finally, to check the assumption of homoscedasticity Figure 1.4 shows a scatterplot of the residuals of the model versus the predicted values. Since there is no clear pattern, this assumption holds

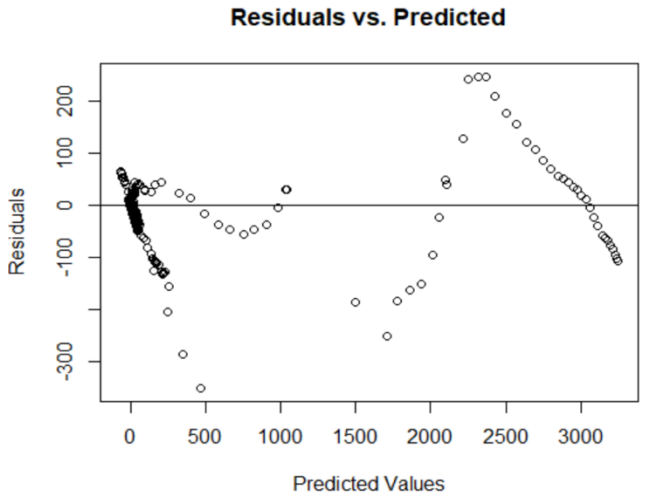

Figure 1.4

Checking R^2 and Adjusted R^2:

The coefficient of multiple determination (R^2) for the model is 0.9939. This implies that the model fits the data very well. However, R^2 increases with more predictors and may allow for an overfit model. The adjusted R^2 also is 0.9939 for the model, so this indicates that the model is not overfit.

ANOVA, F-test, & t-tests:

Table 1.2 gives the analysis of variance (ANOVA) table from the model. Based on the values in the table both an F-test and t-tests assuming $\alpha = 0.05$ can be completed.

Table 1.2

	df	*Sum of Squares*	*Mean Squares*
Confirmed	1	276317124	276317124
Recovered	1	5986984	5986984
Country	2	293119	146560
Residuals	3,528	1735095	492

The null hypothesis for the F- test is H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$ and the alternative hypothesis is H_a: at least one $\beta_j \neq 0$. The test statistic F* = MSR/MSE = 143596.1519 >>> F(0.95, 4, 3528) = 2.374. Hence, the p-value is less than 0.05. So, we reject H_0 and conclude that at least one $\beta_j \neq 0$. Based on this conclusion, the independent variables are significant.

In order to see if each of the individual independent variables are significant, a t-test can be conducted for each of the β's. Table 1.1 gives the values for the test statistic t* in every individual test. For each test, The null hypothesis is $H_0: \beta_j = 0$ and the alternative hypothesis is $H_a: \beta_j \neq 0$.

For beta1:

The test statistic $|t^*| = 267.09 >>> t(0.975, 3528) = 1.96$. Thus, the p-value is less than 0.05. Therefore, we reject H_0 and conclude that beta1 does not equal 0, making it a significant coefficient.

For beta2:

The test statistic $|t^*| = 108.61 >> 1.96$. So, we can conclude that this is also a significant coefficient

For beta3:

The test statistic $|t^*| = 18.04 > 1.96$. So, we can conclude that this is also a significant coefficient

For beta4:

The test statistic $|t^*| = 14.24 > 1.96$. So, we can conclude that this is also a significant coefficient. Therefore, all the coefficients in this model are significant

Conclusions and Possible Improvements

This project used the Johns Hopkins covid-19 data to perform multiple linear regression. The model created used the number or confirmed cases, number of recovered cases, and country as independent variables and the number of deaths as the dependent variable. All of the assumptions for the multiple linear regression approach held. Also, based on the R^2 and adjusted R^2 values the model fit the data well. Additionally, the F-test and t-tests showed that the predictors used in the model were significant.

An improvement that could be made would be to use a more complex model with additional predictors and interactions effects and see if it is still a good fit. Also, since it is an evolving problem the use of more recent data would be beneficial. Another possible improvement would be to manipulate the data set to make the dependent variable, death, binary and perform logistic regression.