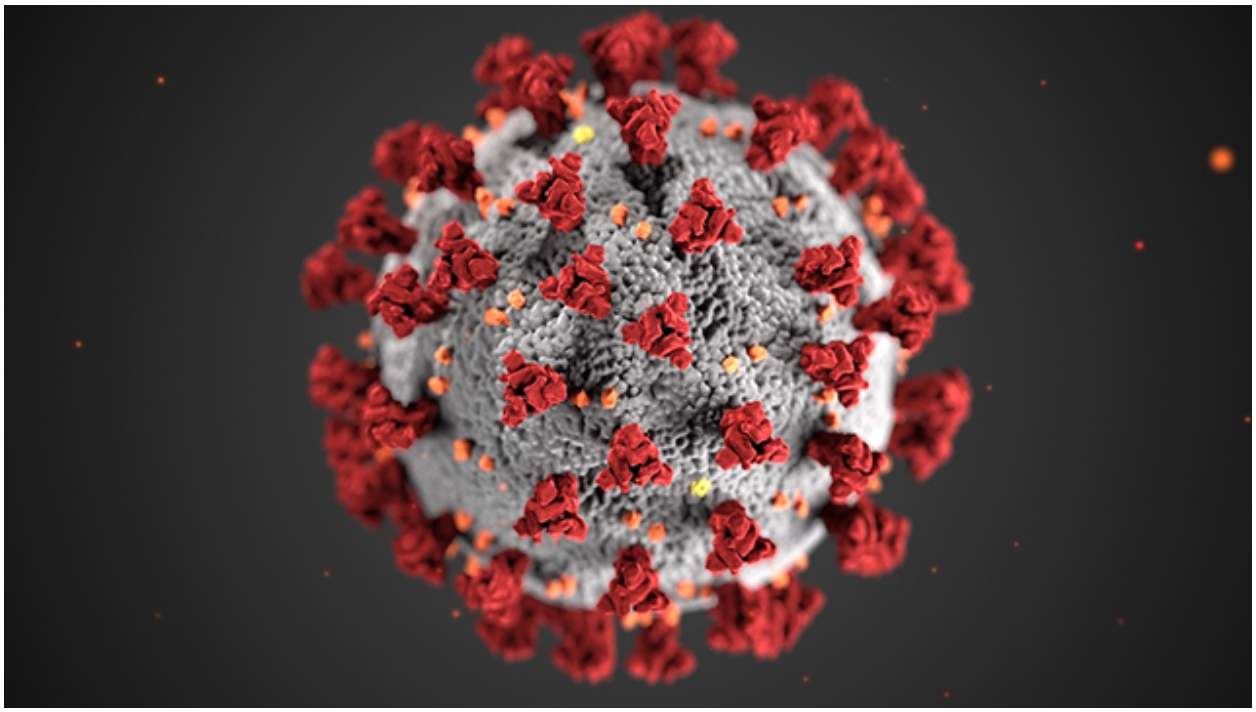
Corona Virus Linear Regression Project Report



Project Summary

Ryan McDonald
Dr Panangadan
Intro to Big Data/Data Science
23 April 2021

Data Sourcing

To begin, we must first gather our data that is going to be used for the linear regression modeling. The confirmed cases and the confirmed death counts are gathered from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University that is hosted on GitHub. The hospital beds data was information pulled from the World Health Organization(WHO).

Data Wrangling

Once the information had been gathered, it needed to be transformed into a format that is “tidy” and usable for linear regression analysis. To transform the confirmed cases and deaths from the Johns Hopkins University, the data needed to pivoted into a longer format. The data originally starts as one row per country with a column for each date. Once these datasets have been transformed the date columns are condensed into one single column, where each row has the date,country, and number of cases/deaths respectively. Next, the Hospital Bed data needed to contain only the number of beds from the most recent year. By removing data from older years, we removed unnecessary and redundant information. To tidy up the demographic data, the columns that were originally separated by gender were condensed into a single column that was the sum of the previous two columns. Once these four data sets were tidied, the data was merged into one single table using a combination of merges, and left joins in order to sort our data by country.

Linear Regression

The key and most important aspect of our model summaries listed below, is the Adjusted R-Squared Value. This value shows a correlation between the linear model that is created and the data that is provided. The higher our R-Squared value the more correlation there is between the data and our model. The Adjusted R-Squared value also takes into account the increased complexity that comes from adding in more variables to our model.

The title of each model presents the variables that are used for our linear regression analysis.

Model 1: Confirmed Cases vs Deaths

The first attempt at linear regression was the most simple and obvious case that compared the number of deaths with the confirmed number of cases. Our linear regression analysis, provides the following data about our model.

```
Residual standard error: 8827 on 85184 degrees of freedom
Multiple R-squared:  0.8895,    Adjusted R-squared:  0.8895
F-statistic: 6.855e+05 on 1 and 85184 DF,  p-value: < 2.2e-16
```

With a value of 0.8895 we see a very high correlation between our model and the data.

Model 2: All predictors vs Deaths

The second attempt at linear regression was perhaps the next most straightforward idea. In this attempt we use every predictor variable in order to create a model for our number of deaths.

```
Residual standard error: 7827 on 67785 degrees of freedom
  (17394 observations deleted due to missingness)
Multiple R-squared:  0.8411,    Adjusted R-squared:  0.841
F-statistic: 5.978e+04 on 6 and 67785 DF,  p-value: < 2.2e-16
```

While there is still a pretty large correlation, using all variables compared to only the cases provides a worse model given the complexity of the model. The other thing to note is that due to some of the information being missing for certain countries we are excluding ~17,000 observations. This could also explain the lower R-Squared value.

In order to create a better model using multiple predictor variables any rows containing missing data has been removed, so that each future model is created using all of the same observations.

Model 3: Revisiting Confirmed Cases vs Deaths

```
Residual standard error: 8885 on 63330 degrees of freedom
Multiple R-squared:  0.7984,    Adjusted R-squared:  0.7984
F-statistic: 2.508e+05 on 1 and 63330 DF,  p-value: < 2.2e-16
```

Our new model for the confirmed cases vs deaths shows a much lower adjusted R-Squared value than the 1st model, however this model gives us a baseline through which we can test all of our predictor variables.

Model 4: Revisiting all Predictors vs Deaths

```
Residual standard error: 7861 on 63325 degrees of freedom
Multiple R-squared:  0.8422,    Adjusted R-squared:  0.8422
F-statistic: 5.633e+04 on 6 and 63325 DF,  p-value: < 2.2e-16
```

With our new data set we can see that when creating a linear regression model using all of our data, we are able to come up with an equation that has a higher correlation than simply using only the number of confirmed cases.

Model 5: Confirmed Cases + Population Above 80 vs Deaths

This model shows a higher Adjusted R-Squared value than that of both the confirmed cases and the model using all predictors. This shows that when only utilizing these two predictor variables we are able to more accurately predict the number of deaths.

```
Residual standard error: 8884 on 63329 degrees of freedom
Multiple R-squared:  0.7984,    Adjusted R-squared:  0.7984
F-statistic: 1.254e+05 on 2 and 63329 DF,  p-value: < 2.2e-16
```

Model 6: Confirmed Cases + Total Population + Pop Above 80 vs Deaths

```
Residual standard error: 8533 on 63328 degrees of freedom
Multiple R-squared:  0.8141,    Adjusted R-squared:  0.814
F-statistic: 9.242e+04 on 3 and 63328 DF,  p-value: < 2.2e-16
```

This model shows an even higher R-Squared value than the previous models, showing that the size of the population as well as the population above 80 makes a better model for prediction.

Model 7: Total Population vs Deaths

```
Residual standard error: 18950 on 63330 degrees of freedom
Multiple R-squared:  0.08293,   Adjusted R-squared:  0.08292
F-statistic:  5727 on 1 and 63330 DF,  p-value: < 2.2e-16
```

By using a simpler model than some of the previous models we are able to come up with a model that can predict the number of deaths based solely on the population of the country.

Model 8: Confirmed Cases + Total Population + Pop Above 80 vs Deaths

```
Residual standard error: 8562 on 63327 degrees of freedom
Multiple R-squared:  0.8128,    Adjusted R-squared:  0.8128
F-statistic: 6.875e+04 on 4 and 63327 DF,  p-value: < 2.2e-16
```

When creating a model using the three predictor variables above, we lose a little bit of correlation compared to our previous models but gain correlation compared to the Model 5. This shows that the total population has a greater correlation than the pop above 80 to the number of deaths.

Model 9: Confirmed Cases + Hospital Beds(per 10k) + Total Population + Pop above 80 vs deaths

```
Residual standard error: 8421 on 63327 degrees of freedom
Multiple R-squared:  0.8189,    Adjusted R-squared:  0.8189
F-statistic: 7.16e+04 on 4 and 63327 DF,  p-value: < 2.2e-16
```

Model 10: Confirmed Cases + Hospital Beds(per 10k) + Total Population + Pop above 80 + Mortality Rate vs deaths

```
Residual standard error: 8404 on 63326 degrees of freedom
Multiple R-squared:  0.8197,    Adjusted R-squared:  0.8197
F-statistic: 5.757e+04 on 5 and 63326 DF,  p-value: < 2.2e-16
```

Model 11: Confirmed Cases + Total Population vs Deaths

```
Residual standard error: 8810 on 63329 degrees of freedom
Multiple R-squared:  0.8018,    Adjusted R-squared:  0.8018
F-statistic: 1.281e+05 on 2 and 63329 DF,  p-value: < 2.2e-16
```

Conclusions

Model Name	Adjusted R-Squared Value	Rank (1 being highest)
Confirmed Cases vs Deaths	.8895	1
All predictors vs Deaths	.841	3
Revisiting Confirmed Cases vs Deaths	.7984	9*
Revisiting all Predictors vs Deaths	.8422	2
Confirmed Cases + Population Above 80 vs Deaths	.7984	9*
Confirmed Cases + Total Population + Population Above 80 vs Deaths	.814	7
Total Population vs Deaths	.08292	11
Confirmed Cases + Total Population + Pop Above 80 vs Deaths	.8128	6
Confirmed Cases + Hospital Beds(per 10k) + Total Population + Pop above 80 vs deaths	.8189	5

Confirmed Cases + Hospital Beds(per 10k) + Total Population + Pop above 80 + Mortality Rate vs Deaths	.8197	4
Confirmed Cases + Total Population vs Deaths	.8018	8

Based on the Adjusted R-Squared values from the table above, when we are solely looking at the number of deaths with the number of confirmed cases we have the highest correlation between our model and our data. The downside to this, is that we are not able to accurately judge the effectiveness of other variables because of missing data. When we use only data that is complete, it shows that the use of all predictor variables provide the highest correlation to our data. When looking to use less than the total number of independent variables, it would appear that the combination of: Confirmed Cases + Hospital Beds(per 10k) + Total Population + Pop above 80 + Mortality Rate provides us with the highest correlation.

Based on this information, when only the number of confirmed cases is known we are able to generate the most accurate model. When we attempt to integrate multiple sets of data from various sources, we get a model that is not as accurate as the original model utilizing however, that is due to the fact that information must be excluded in our linear regression analysis due to missing information.

Variable Importance

While the models and variable choices above were done using trial and error. I sought to take another approach by utilizing the caret library in R. With this library we could take our model that contains all of our predictor variables and run an analysis to determine which variables have the most impact on our prediction.

```
                  overall
covid_data[, 4] 541.54093
covid_data[, 5]  28.60281
covid_data[, 6]  17.14121
covid_data[, 7]  94.54875
covid_data[, 8] 128.50147
covid_data[, 9]  10.10005
```

Looking at the information above we can see that column 4,7, and 8 in our covid_data dataset has the most impact on our prediction. Column 4 is the number of confirmed cases, as expected, column 7 is the urban population and column 8 is the total population. By using this information provided, a new model was created in order to test the accuracy of this.

Model 12: Confirmed Cases, Total Pop, and Urban Pop Vs Deaths

```
Residual standard error: 7980 on 63470 degrees of freedom
Multiple R-squared:  0.8386,    Adjusted R-squared:  0.8386
F-statistic: 1.1e+05 on 3 and 63470 DF,  p-value: < 2.2e-16
```

As seen above when we analyze the model created using these three variables we are able to have a r-squared value that comes very close to the r-squared value of the model using all the predictor variables. This R-Squared value outperforms all of the models created above through trial and error, which show that these three predictor variables do, in fact, have the most impact on the number of deaths. While the other variables do a play a part in determining the number of deaths, they are not as impactful on the prediction.