Project II: Predictive Analytics Project

A Cross-Sectional Data Analysis of COVID Variables

Paige Junker

University of San Diego

ECON 494: Introduction to Business Analytics

November 18, 2020

I. Executive Summary

For my project, I decided to analyze COVID-19 relevant data. This dataset was found on Kaggle and contains 14 variables pertinent to the current crisis. I chose to take a look at COVID-19 data because the current pandemic is prevalent in our lives. This dataset contains cross-sectional data on 187 countries and includes 14 distinct variables related to the pandemic. I know that this data is cross-sectional because the variables are recorded at a specific point in time or that the data is static. When looking at the different variables in this data set, I decided to focus on quantitative variables.

I made a few critical changes to my dataset from Project I to Project II. I decided to simplify the variables' names to allow for a more comfortable model building, and I removed the Deaths_100_Recovered variable. The Deaths_100_Recovered variable was missing many observations, which caused difficulty when examining the covariance and correlation matrices. Instead of replacing the missing observations with 'NA' I decided to exclude the variable entirely because I felt that the Deaths variable captured the same information. Additionally, I transformed my fixed effects variables to be binary, or that they return 1 if a country is in that region and a 0 if it is not. It allowed me to incorporate them into my regressions easily.

I separated the cleaned dataset into a training and testing partition. The training partition contains 70% of the data, including randomly selected observations, and the testing partition includes 30% of the data, the observations selected randomly. To include all of the data, the training partition was adjusted slightly and includes approximately 0.695% of the data or 130 observations, and the testing partition includes 57 observations.

II. Model Proposals

To determine the variables I wanted to examine closer, I created a correlation and covariance matrix. The correlation and covariance matrices only include quantitative variables. Initially, I was most interested in examining the number of New_Cases in each country. The highest correlation with New_Cases was Week_Change at 0.9595 and then Confirmed at 0.9142. I want to note that these high correlations may point to a great deal of multicollinearity existing within the New_Cases data.

Correlation Matrix:

	Confirmed	Deaths	Recovered	Active	New_Cases	New_Deaths	New_Recovered	Deaths_100_Cases	Recovered_100_Cases	Confirmed_Last_Week	Week_Change	Week_Percent_Increase
Confirmed	1.00000000	0.95426580	0.90690285	0.933326091	0.91424680	0.88699866	0.864612025	0.059023291	-0.10716221	0.999220306	0.96077095	0.012699992
Deaths	0.95426580	1.00000000	0.84948462	0.898606914	0.82694949	0.81651604	0.770650210	0.216612200	-0.16472569	0.959501618	0.87482446	-0.016662766
Recovered	0.90690285	0.84948462	1.00000000	0.695371907	0.83375200	0.83768919	0.937238104	0.040269076	-0.01607906	0.897750630	0.93064993	0.016797589
Active	0.93332609	0.89860691	0.69537191	1.000000000	0.85222965	0.80126506	0.679260108	0.056240717	-0.16552031	0.939452168	0.84856713	0.009162772
New_Cases	0.91424680	0.82694949	0.83375200	0.852229650	1.00000000	0.94972946	0.918372212	0.022822769	-0.11454479	0.901988498	0.95948463	0.053538158
New_Deaths	0.88699866	0.81651604	0.83768919	0.801265063	0.94972946	1.00000000	0.892784955	0.034290692	-0.11000503	0.878634734	0.90609803	0.049856916
New_Recovered	0.86461202	0.77065021	0.93723810	0.679260108	0.91837221	0.89278496	1.000000000	0.005593495	-0.06155048	0.845763363	0.95836386	0.060978894
Deaths_100_Cases	0.05902329	0.21661220	0.04026908	0.056240717	0.02282277	0.03429069	0.005593495	1.000000000	-0.21081789	0.065245187	0.01267794	-0.147237802
Recovered_100_Cases	-0.10716221	-0.16472569	-0.01607906	-0.165520311	-0.11454479	-0.11000503	-0.061550478	-0.210817885	1.00000000	-0.108273488	-0.09456506	-0.413758111
Confirmed_Last_Week	0.99922031	0.95950162	0.89775063	0.939452168	0.90198850	0.87863473	0.845763363	0.065245187	-0.10827349	1.000000000	0.94907195	0.007685586
Week_Change	0.96077095	0.87482446	0.93064993	0.848567135	0.95948463	0.90609803	0.958363860	0.012677938	-0.09456506	0.949071953	1.00000000	0.047356685
Week_Percent_Increase	0.01269999	-0.01666277	0.01679759	0.009162772	0.05353816	0.04985692	0.060978894	-0.147237802	-0.41375811	0.007685586	0.04735669	1.000000000

Covariance Matrix:

	Confirmed	Deaths	Recovered	Active	New_Cases	New_Deaths	New_Recovered	Deaths_100_Cases	Recovered_100_Cases	Confirmed_Last_Week	Week_Change	Week_Percent_Increase
Confirmed	2.035174e+11	6.982398e+09	8.927951e+10	1.072555e+11	2.800229e+09	5.567893e+07	1.930434e+09	96148.18773	-1194404.3970	1.790143e+11	2.450312e+10	125063.69866
Deaths	6.982398e+09	2.630686e+08	3.006632e+09	3.712698e+09	9.106318e+07	1.842750e+06	6.186212e+07	12686.28175	-66009.2533	6.180247e+09	8.021512e+08	-5899.41139
Recovered	8.927951e+10	3.006632e+09	4.761903e+10	3.865385e+10	1.235255e+09	2.543551e+07	1.012216e+09	31730.64345	-86688.1771	7.779857e+10	1.148094e+10	80013.69030
Active	1.072555e+11	3.712698e+09	3.865385e+10	6.488893e+10	1.473911e+09	2.840068e+07	8.563561e+08	51731.26253	-1041706.9666	9.503544e+10	1.222003e+10	50949.41975
New_Cases	2.800229e+09	9.106318e+07	1.235255e+09	1.473911e+09	4.609551e+07	8.972147e+05	3.085894e+07	559.51865	-19213.8160	2.431957e+09	3.682715e+08	7934.50406
New_Deaths	5.567893e+07	1.842750e+06	2.543551e+07	2.840068e+07	8.972147e+05	1.936129e+04	6.148186e+05	17.22901	-378.1714	4.855134e+07	7.127598e+06	151.43270
New_Recovered	1.930434e+09	6.186212e+07	1.012216e+09	8.563561e+08	3.085894e+07	6.148186e+05	2.449435e+07	99.96158	-7526.1588	1.662293e+09	2.681416e+08	6587.78751
Deaths_100_Cases	9.614819e+04	1.268628e+04	3.173064e+04	5.173126e+04	5.595186e+02	1.722901e+01	9.996158e+01	13.03869	-18.8076	9.356018e+04	2.588012e+03	-11.60549
Recovered_100_Cases	-1.194404e+06	-6.600925e+04	-8.668818e+04	-1.041707e+06	-1.921382e+04	-3.781714e+02	-7.526159e+03	-18.80760	610.4048	-1.062323e+06	-1.320810e+05	-223.14263
Confirmed_Last_Week	1.790143e+11	6.180247e+09	7.779857e+10	9.503544e+10	2.431957e+09	4.855134e+07	1.662293e+09	93560.17530	-1062323.3496	1.577071e+11	2.130716e+10	66623.83820
Week_Change	2.450312e+10	8.021512e+08	1.148094e+10	1.222003e+10	3.682715e+08	7.127598e+06	2.681416e+08	2588.01243	-132081.0474	2.130716e+10	3.195962e+09	58439.86047
Week_Percent_Increase	1.250637e+05	-5.899411e+03	8.001369e+04	5.094942e+04	7.934504e+03	1.514327e+02	6.587788e+03	-11.60549	-223.1426	6.662384e+04	5.843986e+04	476.49037

The following multiple linear regression model proposals focus on predicting the number of new cases given a number of different variables. I chose to focus on the quantitative variables, specifically Deaths, Week_Percent_Increase, Week_Change, Confirmed, Recovered, New_Deaths, Deaths_100_Cases, and Recovered_100_Cases, although I did choose to include dummy variables in some of the later models to account for country region.

Model 1

New_Cases = Deaths + Week_Percent_Increase + Week_Change + Confirmed + Recovered + New_Deaths + Deaths_100_Cases + Recovered_100_Cases + intercept

For my initial model, I wanted to focus on including the majority of my quantitative variables. Although Deaths and Deaths_100_Cases and Recovered and Recovered_100_Cases are very similar, I want to see if they have separate significance in predicting the number of New_Cases. Additionally, I predicted that the Week_Percent_Increase is significant in predicting the number of new COVID cases in a country. I hypothesize that an increase in Deaths, Week_Percent_Increase, Week_Change, Confirmed, New_Deaths, and Deaths_100_Cases will be associated with an increase in the number of New_Cases, OA HAEC, and Recovered and Recovered_100_Cases be associated with a decrease in the number of New_Cases, OA HAEC. I hypothesize that the intercept term will also be positive because it is impossible for the number of New_Cases to be negative.

Model 2

New_Cases = Deaths + Week_Change + Confirmed + Recovered + New_Deaths + Deaths_100_Cases

In order to determine which variables I wanted to utilize in Model 2, I ran a regression model using the framework from Model 1 and the entire dataset and found that these six variables were significant between the 0.001% level and the 0.01% level. Note that these significance levels will most likely change when using the Training dataset. For this model, I also wanted to remove the intercept term, forcing the model through the origin. The hope in doing this is that I would be able to decrease the out-of-sample and in-sample error. I believe that removing the intercept term makes sense for this model because the number of New_Cases can not be negative.

Model 3

New_Cases = Deaths + Week_Change + Confirmed + Recovered + New_Deaths + Deaths_100_Cases + Europe + Americas + Africa + Eastern_Mediterranean + `Western Pacific Dummy` + South_East_Asia

In this model, I wanted to focus on the significant quantitative variables from Model 2 (that showed up when using the entire dataset [although I know that these significances will most likely change when using the training dataset]) and include the fixed effects variables for regions. Note that I separated the countries into 6 regional fixed effects variables during the cleaning process. Additionally, this model still omits the intercept term in hopes that I can minimize the out-of-sample and in-sample error. I think this is justifiable because the number of new cases can not be negative. Note that the regional dummy variables are structured as binary variables or that they return a 1 if a country is in that region and a 0 if it is not.

Model 4

New_Cases = Deaths + Week_Change + Confirmed + Recovered + New_Deaths + Europe + South_East_Asia

In this model, I wanted to focus on the significant quantitative variables from Model 2 (that showed up when using the entire dataset [although I know that these significances will most likely change when using the training dataset]) and include the fixed effects variables for regions that similarly showed up as significant.

Lastly, I conducted this initial exploration using the entire dataset because I did not find any clear relationships in part one of the project. Although the quantitative variables that I chose to utilize in the models above came from reasoning that I discussed in Project I, I felt that I needed to conduct additional analysis to determine what variables to examine.

III. Diagnostic Results

The following models use the training dataset that I created during the data cleaning process to build regressions for the four models above:

Model 1

New_Cases =0.03872Deaths + 0.6451Week_Percent_Increase + 0.1360Week_Change - 0.003906Confirmed - 0.01366Recovered + 21.89New_Deaths + 22.07Deaths_100_Cases + 7.024Recovered_100_Cases - 352.2

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)           -3.522e+02  3.242e+02  -1.086  0.27948
Deaths                 3.872e-02  2.360e-02   1.641  0.10345
Week_Percent_Increase  6.451e-01  4.242e+00   0.152  0.87937
Week_Change            1.360e-01  7.612e-03  17.861  < 2e-16 ***
Confirmed             -3.906e-03  1.340e-03  -2.914  0.00425 **
Recovered             -1.366e-02  1.072e-03 -12.742  < 2e-16 ***
New_Deaths             2.189e+01  1.385e+00  15.802  < 2e-16 ***
Deaths_100_Cases       2.207e+01  2.749e+01   0.803  0.42382
Recovered_100_Cases    7.024e+00  3.946e+00   1.780  0.07758 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 914.3 on 121 degrees of freedom
Multiple R-squared:  0.983,     Adjusted R-squared:  0.9819
F-statistic: 874.1 on 8 and 121 DF,  p-value: < 2.2e-16
```

Model 1 forecasts New_Cases using only quantitative variables, specifically the Deaths, Week_Percent_Increase, Week_Change, Confirmed, Recovered, New_Deaths, Deaths_100_Cases and Recovered_100_Cases variables. I chose to omit the Active and Confirmed_Last_Week variables due to singularity pointed out by R. The results of the model showed that only four of these variables were significant. Week_Change, New_Deaths and Recovered were significant at the 0.001% level and Confirmed was significant at the 0.01% level. A one unit increase in the number of Deaths in a country would result in a 0.03872 increase in New Cases while a one unit increase in the number of Recovered individuals would result in a 0.01366 decrease. It makes sense that a country with a growing number of deaths would be experiencing a greater number of COVID new cases then a country with fewer deaths. I predicted that the more recovered individuals there were in a country, the lower the number of new cases would be because it would indicate that a large number of individuals had already contracted the virus. Model 1 confirmed this and shows that, the more recovered individuals there are in a country, the lower the number of new_cases would be. Note that the coefficients on Deaths, Week_Percent, Increase, Week_Change, New_Deaths, Deaths_100_Cases and Recovered_100_Cases are positive while the coefficients on Confirmed, Recovered and the intercept are negative. Note that I hypothesised that the Confirmed variable would have a positive relationship with New_Cases. This model shows that an increase in the number of Confirmed cases is associated with a decrease in the number of New_Cases. This result seems strange. I also want to point out that the residual error in this model is 914.3. The following models will attempt to lower this residual error as much as possible. Note that I continued to include Deaths_100_Cases and Deaths because eliminating them minimized the adjusted r-squared, or worsened the fit of the model.

Model 2

New_Cases = 0 + 0.020744Deaths + 0.131890Week_Change - 0.003131Confirmed - 0.012948Recovered + 21.922307New_Deaths + 38.333552Deaths_100_Cases

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
Deaths            0.020744   0.022153   0.936   0.3509
Week_Change       0.131890   0.007460  17.679   <2e-16 ***
Confirmed        -0.003131   0.001296  -2.415   0.0172 *
Recovered        -0.012948   0.001035 -12.514   <2e-16 ***
New_Deaths       21.922307   1.393399  15.733   <2e-16 ***
Deaths_100_Cases 38.333552  20.661275   1.855   0.0659 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 922.8 on 124 degrees of freedom
Multiple R-squared:  0.9831,    Adjusted R-squared:  0.9823
F-statistic:  1202 on 6 and 124 DF,  p-value: < 2.2e-16
```

The second model utilizes the Deaths, Week_Change, Confirmed, Recovered, New_Deaths and Deaths_100_Cases variables. For this model, I removed the intercept term, forcing the model through the origin. The hope in doing this is that I would be able to decrease the out-of-sample and in-sample error. Removing the intercept term increased the residual standard error from 914.3 in Model 1 to 922.8 in Model 2. I believe that removing the intercept term makes sense for this model, because the number of New_Cases can not be negative, but it is important to note that it did increase the in-sample error. By removing the insignificant variables and forcing the model through the origin, the coefficient on Confirmed decreased from being significant at the 0.01% level to being significant at the 0.05% level. Additionally, the coefficient on Deaths_100_Cases increased from having no significance to being significant at the 0.1% level. All of the other variable's significance and the signs of the variable coefficients remained the same from Model 1 to Model 2. Some important things to note is that a 1 unit increase in the number of New_Deaths is associated with a 21.92 unit increase in the number of new cases, OA HAEC. This makes sense logically, because the number of new deaths is most likely already included in the new cases count or if not would point to a country having a growing number of COVID cases. Additionally, a one unit increase in the number of Deaths_100_Cases is associated with a 38.333 unit increase in the number of New_Cases. Similarly to the previous example, it makes sense that a country with more Deaths would also have more New_Cases. A key statistic worth noting is the adjusted R^2. Moving from 0.9819 in Model 1 to 0.9823 in Model 2, the adjusted R^2 statistic can be interpreted as model 2 explaining 98.23% of the variance in the model. Note that the adjusted R^2 does take into account model complexity, so this increase in the explanation of the variance could be a result of a decrease in the model's complexity.

Model 3

New_Cases = 0 + 0.02552Deaths + 0.1283Week_Change - 0.002992Confirmed - 0.01301Recovered + 22.35New_Deaths + 13.24Deaths_100_Cases + 338.4Europe - 16.67Americas - 145.7Africa + 188.2Eastern_Mediterranean + 99.43`Western Pacific Dummy` + 1077South_East_Asia

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
Deaths                   2.552e-02  2.325e-02   1.098   0.2746
Week_Change              1.283e-01  7.763e-03  16.523   <2e-16 ***
Confirmed               -2.992e-03  1.334e-03  -2.244   0.0267 *
Recovered               -1.301e-02  1.025e-03 -12.691   <2e-16 ***
New_Deaths               2.235e+01  1.415e+00  15.789   <2e-16 ***
Deaths_100_Cases         1.324e+01  2.754e+01   0.481   0.6317
Europe                   3.384e+02  1.662e+02   2.036   0.0439 *
Americas                -1.667e+01  2.154e+02  -0.077   0.9385
Africa                  -1.457e+02  1.668e+02  -0.874   0.3841
Eastern_Mediterranean    1.882e+02  2.809e+02   0.670   0.5043
`Western Pacific Dummy`  9.943e+01  2.715e+02   0.366   0.7149
South_East_Asia          1.077e+03  3.917e+02   2.750   0.0069 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 894.4 on 118 degrees of freedom
Multiple R-squared:  0.9849,    Adjusted R-squared:  0.9833
F-statistic: 640.8 on 12 and 118 DF,  p-value: < 2.2e-16
```

The model above utilizes the quantitative variables from Model 2 and the regional fixed effect variables to predict New_Cases. Additionally, this model still omits the intercept term, because the number of New_Cases can not be negative, to minimize the out-of-sample and in-sample error. The Week_Change, Recovered and New_Deaths variables continue to be significant at the 0.001% level and the Confirmed variable is significant at the 0.05% level. Furthermore, the addition of the regional dummy variables decreases the Deaths_100_Cases variable's significance from the 0.1% level in Model 2 to no significance in Model 3. When examining the regional dummy variables, the South_East_Asia variable is significant at the 0.01% level and the Europe variable is statistically significant at the 0.05% level. The coefficients on the significant regional variables can be interpreted as follows: a country being located in Europe is associated with a 338.4 unit increase in New_Cases and a country being located in South_East_Asia is associated with a 1077 unit increase in New_Cases. Both of these variables have a direct positive relationship with New_Cases. Although the variables are not significant, it is interesting to note that a country being located in Africa and Americas is associated with a decrease in New_Cases. This result seems unreliable, given the disproportionate growth in the number of COVID cases in the United States. Some important things to note in this model is that the residual standard error decreases from 922.8 in Model 2 to 894.4 in Model 3, confirming that removing the intercept term decreased the error. Additionally, the Adjusted R^2 increases slightly from 0.9823 in Model 2 to 0.9833 in Model 3. The increase, although slight, is worth noting because of the increased complexity of Model 3.

Model 4

New_Cases = 0 + 0.03022Deaths + 0.1280Week_Change - 0.003153Confirmed - 0.01295Recovered + 22.40New_Deaths + 378.1Europe + 1104South_East_Asia

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
Deaths           3.022e-02  1.923e-02   1.572  0.11852
Week_Change      1.280e-01  7.367e-03  17.380  < 2e-16 ***
Confirmed       -3.153e-03  1.191e-03  -2.648  0.00915 **
Recovered       -1.295e-02  9.928e-04 -13.042  < 2e-16 ***
New_Deaths       2.240e+01  1.327e+00  16.886  < 2e-16 ***
Europe           3.781e+02  1.414e+02   2.674  0.00851 **
South_East_Asia  1.104e+03  3.844e+02   2.872  0.00480 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 883.3 on 123 degrees of freedom
Multiple R-squared:  0.9846,    Adjusted R-squared:  0.9838
F-statistic:  1126 on 7 and 123 DF,  p-value: < 2.2e-16
```

In this model, I examine the significant variables from Model 3. Although the Europe regional variable was only significant at the 0.05% level in Model 3, its significance increased in Model 4 to the 0.01% level. The significance of the coefficient for Confirmed also increased from the 0.05% level to the 0.01% level. The significance of the remaining variables remained consistent from Model 3 to Model 4. Additionally, I maintained forcing the model through the origin in order to decrease the in-sample and out-of-sample error. Note that the residual standard error decreased from 894.4 Model 3 to 883.3 in Model 4. Additionally, the Adjusted R^2 increased by 0.0005 from Model 3 to Model 4. Similarly, there is an increase in the Multiple R^2, which does not take into account model complexity, of 0.0003 from Model 3 to Model 4. Both of these point to Model 4 explaining more of the variation, although the change is so small that it can most likely be discounted as insignificant. Similar to Model 3, Europe and South_East_Asia have direct positive relationships with New_Cases. This can be interpreted as a country being in Europe is associated with a 378.1 unit increase in New_Cases and a country being in South_East_Asia being associated with a 1104 unit increase in New_Cases. There is an increase in both of these variable coefficients meaning that they more greatly influence the number of New_Cases.

I want to note that I tried removing the Deaths variable because it did not have a statistically significant relationship with New_Cases, but I found that this decreased the Adjusted R-squared and increased the residual standard error. Due to these reasons, I continued to include it in the models above.

Best Model

In order to determine which of the four models I examined was the best, I compared the models adjusted r-squared values. The adjusted r-squared statistic is the most important statistic to examine because it shows how much of the variance is explained by the model. Model 4 is the best model because it had the highest adjusted r-squared at 0.9838. This can be interpreted as 98.38% of the variance in the regression can be explained by the model. Additionally, I want to note that the variable coefficients that I previously discussed were all significant, with the exception of the Deaths variable which was included in order to increase the fit of the model.

IV. Predictions

Introduction

The following models use the testing partition of my data and examine the RMSE_OUT statistics. This allows us to show how well the data is able to be extrapolated to other datasets and makes sure that it is not just memorizing data. The primary reason why I created a training and testing partition is to ensure that the regression results are not just a representation of memorizing the data.

Model 1

In Model 1, the RMSE_IN is 882.0625 and the RMSE_OUT is 1540.188. This makes the generalization error 658.1255.

Model 2

In Model 2, the RMSE_IN is 901.2393 and the RMSE_OUT is 1454.233. The difference of these, or the generalization error, is 552.9937 for this model.

Model 3

In Model 3, the RMSE_IN is 852.1681 and the RMSE_OUT is 1444.892. This makes the generalization error 592.7239.

Model 4

In Model 4, the RMSE_IN is 859.2159 and the RMSE_OUT is 1461.93. The difference of these, or the generalization error, is 602.7141 for this model.

Best Model

In order to determine which of the four models I examined was the best, I compared the out-of-sample errors. Out-of-sample errors are the most important error metric to examine because they determine the model's ability to be extrapolated to other datasets, besides my cleaned dataset, and determine if the model is a decent approx. of the data generating process. Model 3 is the best model because it had the lowest out-of-sample error at 1444.892. In addition to having the lowest out-of-sample error, Model 3 has the lowest in-sample error at 852.1681.

V. Conclusion

The models above use cross-sectional data to analyze the number of new cases across 187 countries. The COVID dataset was found on Kaggle and contains 14 variables pertinent to the current crisis. Kaggle is a platform that allows users to post a variety of data sets and data collections. Initially, I chose to focus on the quantitative variables but later included fixed effects variables that accounted for a countries region in my later models.

The diagnostic results focused on four models that I thought best showed my exploration of the observations. Initially, I began by examining the relationships between all of the quantitative variables. After finding that not all variables were significant, I removed some of the variables and continued models, eventually adding in the dummy variables. My analysis found that only Europe and South_East_Asia had statistically significant relationships with New_Cases. After comparing the results of all four models, I determined that Model 4 is the best

model because it had the highest adjusted r-squared at 0.9838. This can be interpreted as 98.38% of the variance in the regression that can be explained by the model.

The following section, or the predictions section, focused on running all four models using the testing partition. The results of this showed that Model 3 was the best model because it had the lowest out-of-sample error at 1444.892.

Although the diagnostic results section determined that Model 4 was the best model, I believe that the predictions section's conclusion that Model 3 was the best model is correct. The out-of-sample error is the most important error metric to examine because it determines the model's ability to be extrapolated to other datasets and determines if the model is a decent approx. of the data generating process. I also want to point out that Model 3 had the second-highest adjusted r-squared statistic in the diagnostic results section, only falling 0.005 below Model 4. The results from Model 3 can be seen below.

Best Model: Model 3

New_Cases = 0 + Deaths + Week_Change + Confirmed + Recovered + New_Deaths + Deaths_100_Cases + Europe + Americas + Africa + Eastern_Mediterranean + `Western Pacific Dummy` + South_East_Asia

New_Cases = 0 + 0.02552Deaths + 0.1283Week_Change - 0.002992Confirmed - 0.01301Recovered + 22.35New_Deaths + 13.24Deaths_100_Cases + 338.4Europe - 16.67Americas - 145.7Africa + 188.2Eastern_Mediterranean + 99.43`Western Pacific Dummy` + 1077South_East_Asia

RMSE_IN 852.1681
RMSE_OUT 1444.892
Generalization error 592.7239

The Predictive Analytics Project gave insight into what variables increase and decrease the number of New_Cases in a county. The analysis found that many quantitative variables, such as confirmed cases, recovered cases, etc., all have statistically significant relationships with the number of new cases. Additionally, it found that countries within Europe and South_East_Asia had statistically significant relationships with the number of new cases. Although the results from this project did not prove to be groundbreaking, they allowed me to deepen my understanding of R and data analysis, and I believe that they could be useful in identifying relationships in further COVID research.