Analysis of COVID-19 vaccination

Sajja, Varsha   Namchaiya, Sudarat   Zhou, Mengmo

Morton, Evan   Becker, Jack

3/15/2021

Abstract

In the era of global pandemic, tracking COVID-19 vaccination rates is crucial to understand the scale of protection against the virus. This report deals with real-time data on covid-vaccination progress and analyzes how vaccinations have been distributed to the global population using statistical regression analysis techniques. The data from Pfizer and BioNTech pharmaceutical companies is being used for the study. Construction and validation of various regression models on dataset with respect to independent variables is provided in each member's point of view. A computer-based tool RStudio is used for calculations in this statistical project. The final model is selected by testing with Stepwise Regression methods. A linear regression line and equation for the model are generated to help observe and predict new deaths or increase in rate of total deaths due to virus over various locations in the world. The model also shows which variables play the most important roles in the prediction.

Introduction

As the global COVID-19 pandemic rages on, vaccines have been developed at a rapid pace. In the United States, multiple companies that have produced vaccines which have been approved for use such as Pfizer and BioNTech, Moderna, and Johnson and Johnson. Abroad, even more vaccinations have been available with companies like AstraZeneca joining the mix. While this rapid development is an unprecedented and encouraging step towards the end of a global tragedy, there are still many questions regarding vaccinations and their accessibility. While the virus is constrained to one particular country's boarder, not all countries are vaccinating their populations at the same rate. These countries all have different economies, health systems, and policies in place that have affected the outcomes of the pandemic in their nations. This poses the question of whether there are factors that can predict the number of vaccinated people in a population.

In order to explore our initial question, we analyzed data provided by Kaggle and Our World in Data, a free organization that tracks population level health data. The target dataset included entries on total vaccinations, country GDP, human development index, COVID-19 deaths, and COVID-19 cases among other variables of interest. During our data analysis, we explored how different variables related to total vaccination numbers and worked towards creating a model that was able to accurately predict vaccination numbers.

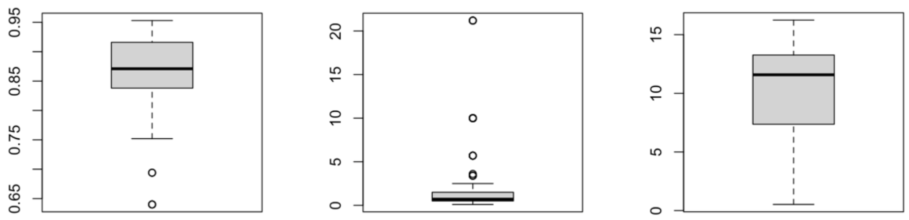
Data Preparation

Overall, the data from Kaggle (COVID-19 World Vaccination Progress Data) consists of 62,432 Rows and 59 columns which are numeric attributes except continent and location. This dataset was collected and published in 2020. It is a complete COVID-19 dataset including all historical data on the pandemic up to the date of publication.

We have selected total vaccinations, which is the number of COVID-19 vaccination doses administered for the given country, as our dependent variable that has been collected from 193 countries around the world. In terms of the parameters of study in this project, we focus on independent variables as they relate to total vaccination numbers. We have chosen 20 independent variables from 59 variables. Table 1, found on the last page, shows the name of each variable and the description of them.

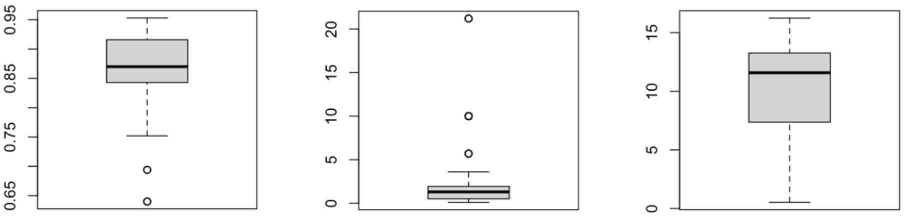
Regarding data cleanliness, there are a lot of missing values in most of our dependent variable. The missing values is more than percent of the observations. Therefore, we decided to cut all of the missing values off and then check if there is any missing value in the remaining observations.

We found that there are three variables with missing values which are human_dev_index, extreme_poverty and aged_70_older. We checked the distribution of each of those variables to consider which technique we could use for replacing them.



(a) human_dev_index (b) extreme_poverty (c) aged_70_older

Figure 1: Before cleaning



(a) human_dev_index (b) extreme_poverty (c) aged_70_older

Figure 2: After cleaning

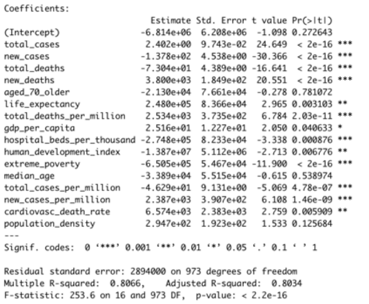
We can see in Figures 1a and 1b that the boxplots for human_dev_index and extreme_poverty show skewness in their distributions with some outliers, but in Figure 1c, the boxplot for aged_70_older shows skewness in its distribution with no outliers. To perform the cleaning method, we decided to replace the missing values of those variables by using mean of each group. The distribution of the cleaned variables looks better, shown in Figures 2a, 2b, and 2c. We decided to use this cleaned data for our further analysis to build the best model for predicting total vaccination.

Data Summary

After data preparation, the 62,432 observation were cleaned into 990 observations. Even though we cut a huge portion of data, we still have an acceptable chunk of data for creating a model. We have 990 observations with 21 variables. We are focusing on the numerical variables which are the 19 variables in Table 1 that are not location or continent. Thus, the relationship between each numerical variable is shown in the below picture. Then, we are building and improving the model by applying appropriate analysis tools, such as, validation, stepwise regression, data transformation, multicollinearity and residuals analysis to identify the useful subsets of the predictors. Then, we will choose the best regression model for predicting total vaccination.

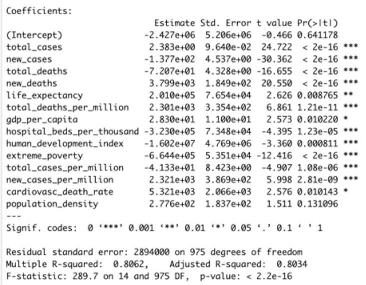
Model Building

As this dataset contains many numerical variables, stepwise regression appears to be an appropriate analysis tool to identify a useful subset of the predictors. We start using all of 18 numerical variables except date and handwashing facilities at first then applied forward and backward selection to see what variables should be kept in our regression model. After using stepwise regression, the results of both methods are similar. F-tests are significant and adjusted R-squares are 80.34% with MSE 8.233154e+12 which is pretty high error.



```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -6.814e+06  6.208e+06  -1.098 0.272643    
total_cases                 2.402e+00  9.743e-02  24.649  < 2e-16 ***
new_cases                  -1.378e+02  4.538e+00 -30.366  < 2e-16 ***
total_deaths               -7.304e+01  4.389e+00 -16.641  < 2e-16 ***
new_deaths                  3.800e+03  1.849e+02  20.551  < 2e-16 ***
aged_70_older              -2.130e+04  7.661e+04  -0.278 0.781072    
life_expectancy             2.480e+05  8.366e+04   2.965 0.003103 ** 
total_deaths_per_million    2.534e+03  3.735e+02   6.784 2.03e-11 ***
gdp_per_capita              2.516e+01  1.227e+01   2.050 0.040633 *  
hospital_beds_per_thousand -2.748e+05  8.233e+04  -3.338 0.000876 ***
human_development_index    -1.387e+07  5.112e+06  -2.713 0.006776 ** 
extreme_poverty            -6.505e+05  5.467e+04 -11.900  < 2e-16 ***
median_age                 -3.389e+04  5.515e+04  -0.615 0.538974    
total_cases_per_million    -4.629e+01  9.131e+00  -5.069 4.78e-07 ***
new_cases_per_million       2.387e+03  3.907e+02   6.108 1.46e-09 ***
cardiovasc_death_rate       6.574e+03  2.383e+03   2.759 0.005909 ** 
population_density          2.947e+02  1.923e+02   1.533 0.125684    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2894000 on 973 degrees of freedom
Multiple R-squared:  0.8066,	Adjusted R-squared:  0.8034 
F-statistic: 253.6 on 16 and 973 DF,  p-value: < 2.2e-16
```

(a) Model after applying forward elimination

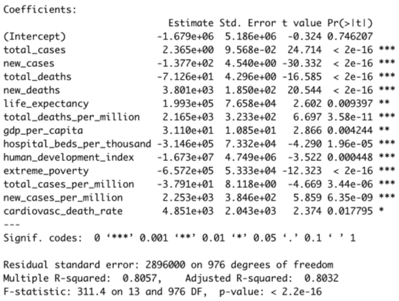


```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -2.427e+06  5.206e+06  -0.466 0.641178    
total_cases                 2.383e+00  9.640e-02  24.722  < 2e-16 ***
new_cases                  -1.377e+02  4.537e+00 -30.362  < 2e-16 ***
total_deaths               -7.207e+01  4.328e+00 -16.655  < 2e-16 ***
new_deaths                  3.799e+03  1.849e+02  20.550  < 2e-16 ***
life_expectancy             2.010e+05  7.654e+04   2.626 0.008765 ** 
total_deaths_per_million    2.301e+03  3.354e+02   6.861 1.21e-11 ***
gdp_per_capita              2.830e+01  1.100e+01   2.573 0.010220 *  
hospital_beds_per_thousand -3.230e+05  7.348e+04  -4.395 1.23e-05 ***
human_development_index    -1.602e+07  4.769e+06  -3.360 0.000811 ***
extreme_poverty            -6.644e+05  5.351e+04 -12.416  < 2e-16 ***
total_cases_per_million    -4.133e+01  8.423e+00  -4.907 1.08e-06 ***
new_cases_per_million       2.321e+03  3.869e+02   5.998 2.81e-09 ***
cardiovasc_death_rate       5.321e+03  2.066e+03   2.576 0.010143 *  
population_density          2.776e+02  1.837e+02   1.511 0.131096    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2894000 on 975 degrees of freedom
Multiple R-squared:  0.8062,	Adjusted R-squared:  0.8034 
F-statistic: 289.7 on 14 and 975 DF,  p-value: < 2.2e-16
```

(b) Model after applying backward elimination

Figure 3

We decided to cut some variables which are aged 70 older, median age, and population density that are not significant to the model. As a result, F-tests are significant and adjusted R-squares are 80.32% which are pretty high value. The screenshot from R studio is shown in Figure 4.

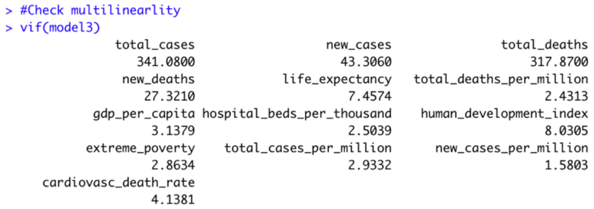


```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -1.679e+06  5.186e+06  -0.324 0.746207    
total_cases                 2.365e+00  9.568e-02  24.714  < 2e-16 ***
new_cases                  -1.377e+02  4.540e+00 -30.332  < 2e-16 ***
total_deaths               -7.126e+01  4.296e+00 -16.585  < 2e-16 ***
new_deaths                  3.801e+03  1.850e+02  20.544  < 2e-16 ***
life_expectancy             1.993e+05  7.658e+04   2.602 0.009397 ** 
total_deaths_per_million    2.165e+03  3.233e+02   6.697 3.58e-11 ***
gdp_per_capita              3.110e+01  1.085e+01   2.866 0.004244 ** 
hospital_beds_per_thousand -3.146e+05  7.332e+04  -4.290 1.96e-05 ***
human_development_index    -1.673e+07  4.749e+06  -3.522 0.000448 ***
extreme_poverty            -6.572e+05  5.333e+04 -12.323  < 2e-16 ***
total_cases_per_million    -3.791e+01  8.118e+00  -4.669 3.44e-06 ***
new_cases_per_million       2.253e+03  3.846e+02   5.859 6.35e-09 ***
cardiovasc_death_rate       4.851e+03  2.043e+03   2.374 0.017795 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2896000 on 976 degrees of freedom
Multiple R-squared:  0.8057,	Adjusted R-squared:  0.8032 
F-statistic: 311.4 on 13 and 976 DF,  p-value: < 2.2e-16
```

Figure 4

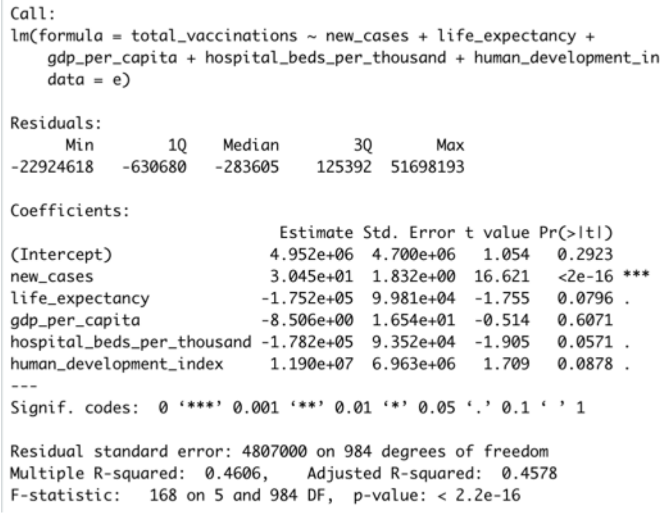
To make sure this model can predict the dependent variable precisely, I try to detect multicollinearity by using variance inflation factor (VIF). In order to make a predictive model, I need to protect against multicollinearity in my model.

From the results shown in Figure 5, there are four variables which have a VIF that is greater than 10. Thus, we decided to cut total cases, total deaths, and new deaths. Moreover, we cut total cases per million and new case per million and total deaths per million out because of redundancy, then run the model again. Then, variables that are highly not significant to the model are removed as well. Adjust R-square has been lowered from 80.32% to 45.78%. However, multicollinearity can lead misleading results.



```
> #Check multilinearlity
> vif(model3)
               total_cases                  new_cases               total_deaths
                  341.0800                    43.3060                   317.8700
                new_deaths            life_expectancy   total_deaths_per_million
                   27.3210                     7.4574                     2.4313
            gdp_per_capita hospital_beds_per_thousand    human_development_index
                    3.1379                     2.5039                     8.0305
           extreme_poverty    total_cases_per_million      new_cases_per_million
                    2.8634                     2.9332                     1.5803
     cardiovasc_death_rate
                    4.1381
```

Figure 5



```
Call:
lm(formula = total_vaccinations ~ new_cases + life_expectancy + 
    gdp_per_capita + hospital_beds_per_thousand + human_development_index, 
    data = e)

Residuals:
      Min        1Q    Median        3Q       Max 
-22924618   -630680   -283605    125392  51698193 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 4.952e+06  4.700e+06   1.054   0.2923    
new_cases                   3.045e+01  1.832e+00  16.621   <2e-16 ***
life_expectancy            -1.752e+05  9.981e+04  -1.755   0.0796 .  
gdp_per_capita             -8.506e+00  1.654e+01  -0.514   0.6071    
hospital_beds_per_thousand -1.782e+05  9.352e+04  -1.905   0.0571 .  
human_development_index     1.190e+07  6.963e+06   1.709   0.0878 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4807000 on 984 degrees of freedom
Multiple R-squared:  0.4606,	Adjusted R-squared:  0.4578 
F-statistic:   168 on 5 and 984 DF,  p-value: < 2.2e-16
```

Figure 6

After performing forward and backward elimination and applied multicolinearity check, we have five parameters of interest which are new cases, life expectancy, gdp per capital, hospital bed per thousand and human development index. We will use those variables as the independent variables to create the model for predicting total vaccination next. Figure 7 shows the scatter plot summary of each variable.

After checking the distribution of all variables, we found that total cases has a strong right skew, so we applied a log transformation on the variables. As a result, the distribution of new cases looks quite normal. This can be seen in Figures 8a and 8b

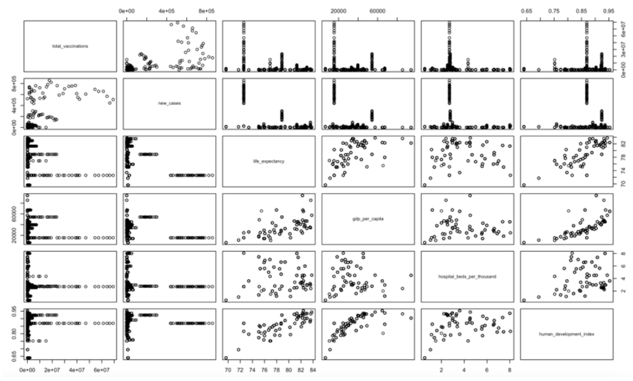
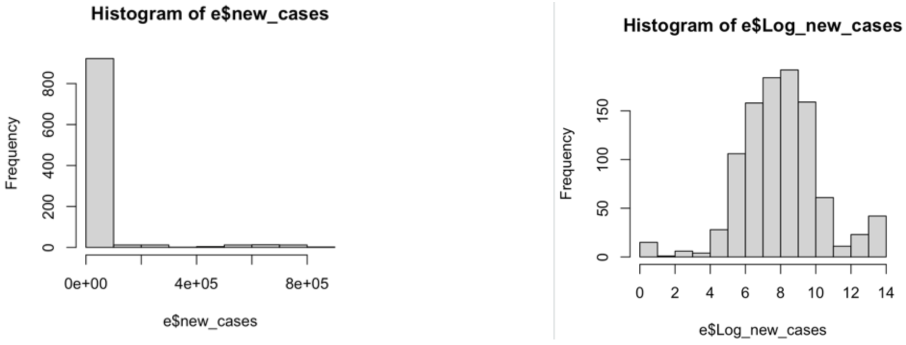


Figure 7: The scatter plot summary of each variable.



(a) Histogram of new cases

(b) Histogram of Log_new cases

Figure 8

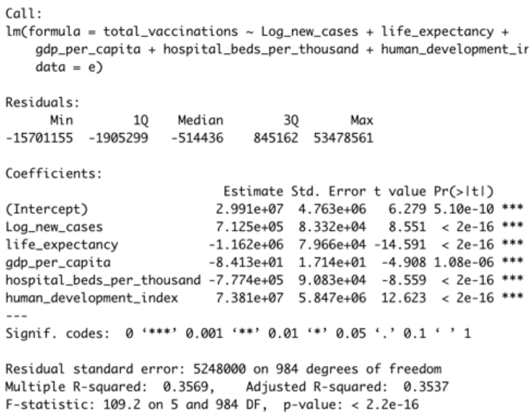
We created the first model of total vaccination that investigates the effect of making changes in new cases, life expectancy, gdp per capital, hospital bed per thousand, and human development. Figure 9 is the summary of this model.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{1}$$

where x_1 = new cases, x_2 = life expectancy, x_3 = gdp per capital, x_4 = hospital bed per thousand, and x_5 = human development.

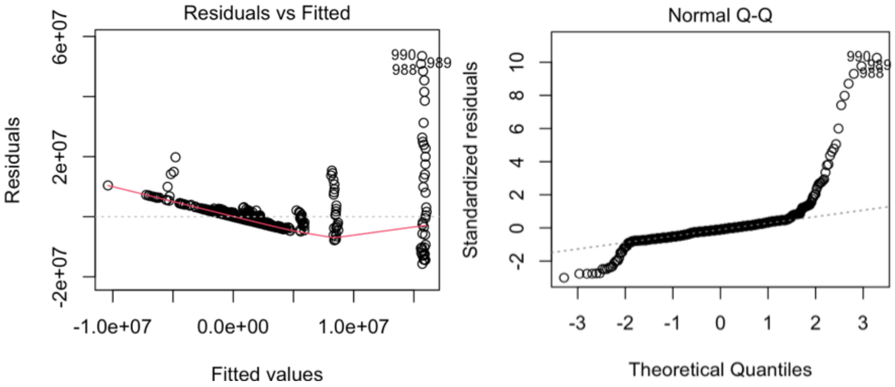
From this result, we can see that all of the independent variables are significant with an adjusted r-square of 0.3537, That means 35.37% of the variability of total vaccination can be explained by these independent variables. Even though the adjusted R-squared is quite low, the F-tests and p-value are significant.

Residuals plot vs fitted shows the cone shape and the normal QQ plot shows the right-skewed distribution with some outliers. These residual plots appear to be quite unhealthy residual plots resulting in an unidentified pattern. Therefore, we performed logarithm transformation to dependent variable to see if this



```
Call:
lm(formula = total_vaccinations ~ Log_new_cases + life_expectancy +
    gdp_per_capita + hospital_beds_per_thousand + human_development_ir
    data = e)

Residuals:
      Min        1Q    Median        3Q       Max
-15701155  -1905299   -514436    845162  53478561

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.991e+07  4.763e+06   6.279 5.10e-10 ***
Log_new_cases              7.125e+05  8.332e+04   8.551  < 2e-16 ***
life_expectancy           -1.162e+06  7.966e+04 -14.591  < 2e-16 ***
gdp_per_capita            -8.413e+01  1.714e+01  -4.908 1.08e-06 ***
hospital_beds_per_thousand -7.774e+05  9.083e+04  -8.559  < 2e-16 ***
human_development_index    7.381e+07  5.847e+06  12.623  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5248000 on 984 degrees of freedom
Multiple R-squared:  0.3569,    Adjusted R-squared:  0.3537
F-statistic: 109.2 on 5 and 984 DF,  p-value: < 2.2e-16
```

Figure 9: Summary of first order model.



(a) Residuals Vs Fitted

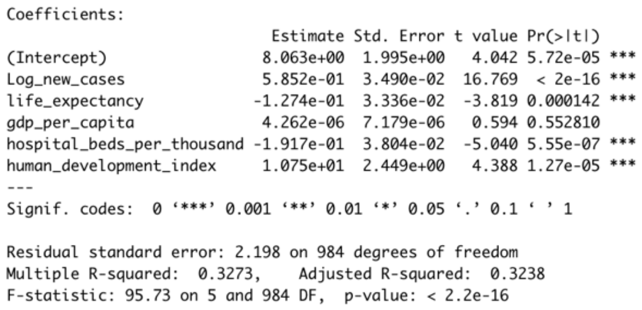
(b) Normal Q-Q plot

Figure 10

model performs better or not.

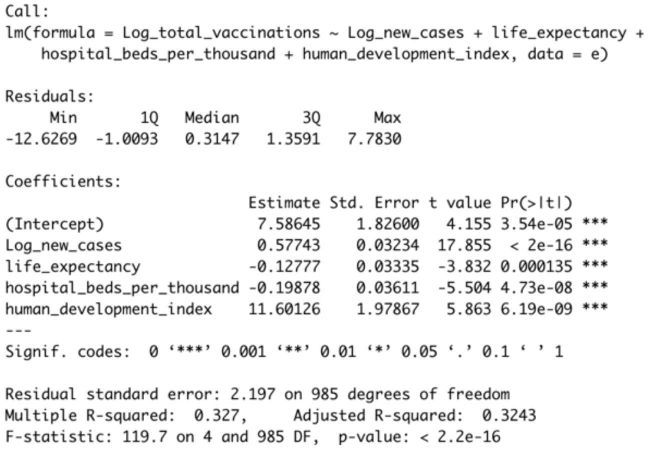
From the above result, we can see that gdp per capital is not significant then we decided to cut this variable out of our model and we can see that the adjusted r-square increased from 32.38% to 32.43% and all variables are significant. Moreover, the F-tests and p-value are significant.

With reference to the plot of residuals against fitted value, the plot such that the residuals seem to be accommodated in an inward opening funnel, then such pattern indicates that the variance of errors is not constant, but it is a decreasing function of y. In addition, there are a few potential outliers, so there may be a few issues with this model as well. Looking at the normal QQ plot, it displays that the residual distribution has a minimal left skewness compared to



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                8.063e+00  1.995e+00   4.042 5.72e-05 ***
Log_new_cases              5.852e-01  3.490e-02  16.769  < 2e-16 ***
life_expectancy           -1.274e-01  3.336e-02  -3.819 0.000142 ***
gdp_per_capita             4.262e-06  7.179e-06   0.594 0.552810    
hospital_beds_per_thousand -1.917e-01  3.804e-02  -5.040 5.55e-07 ***
human_development_index    1.075e+01  2.449e+00   4.388 1.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.198 on 984 degrees of freedom
Multiple R-squared:  0.3273,	Adjusted R-squared:  0.3238 
F-statistic: 95.73 on 5 and 984 DF,  p-value: < 2.2e-16
```

Figure 11: Summary of the model after applying transformation on dependent variable



```
Call:
lm(formula = Log_total_vaccinations ~ Log_new_cases + life_expectancy + 
    hospital_beds_per_thousand + human_development_index, data = e)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.6269  -1.0093   0.3147   1.3591   7.7830 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 7.58645    1.82600   4.155 3.54e-05 ***
Log_new_cases               0.57743    0.03234  17.855  < 2e-16 ***
life_expectancy            -0.12777    0.03335  -3.832 0.000135 ***
hospital_beds_per_thousand -0.19878    0.03611  -5.504 4.73e-08 ***
human_development_index    11.60126    1.97867   5.863 6.19e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.197 on 985 degrees of freedom
Multiple R-squared:  0.327,	Adjusted R-squared:  0.3243 
F-statistic: 119.7 on 4 and 985 DF,  p-value: < 2.2e-16
```

Figure 12: Summary of the model after cutting gdp per capital

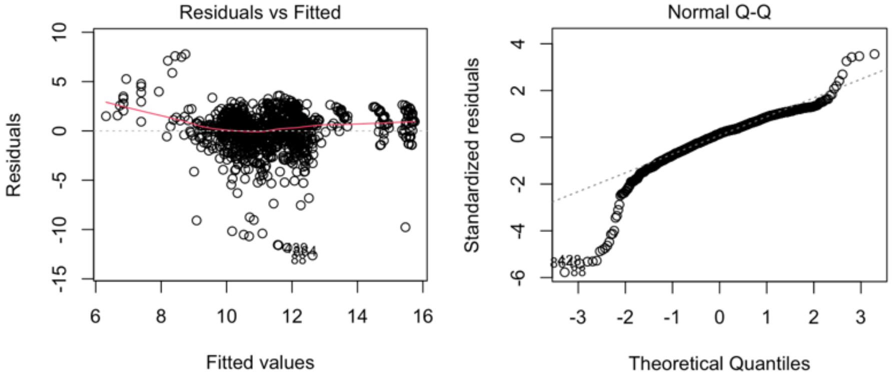
the standard normal distribution. Our data generally has uniform variance at the middle of our range and is somewhat heteroscedastic in the middle of the range too.

As the low value of adjusted R-squared, this might lead to under fitting. Thus, validation of the model is one technique to ensure this underfitting would not occur. As a result of model validation shown in Figure 14, a correlation between prediction and actual, the result came out is 51% which is a little bit higher than the results of mean absolute error. From the plot we can see some correlation even though there are some outliers occur.

The first order model

$$\mathrm{LOG}(y) = 7.586 + 0.577x_1 - 0.127x_2 - 1.988x_3 + 11.601x_4 \tag{2}$$

where y = total vaccinations x_1 = new cases, x_2 = life expectancy, x_3 = hospital bed per thousand, and x_4 = human development, exhibits a better



(a) Residuals Vs Fitted

(b) Normal Q-Q plot

Figure 13

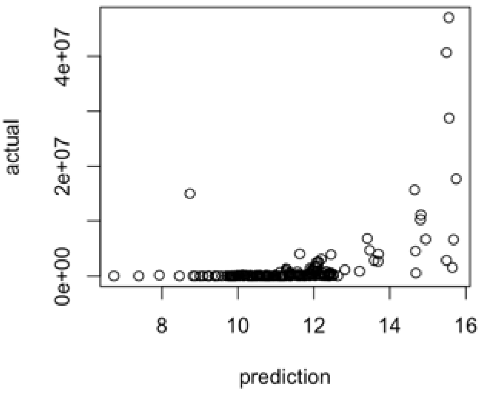
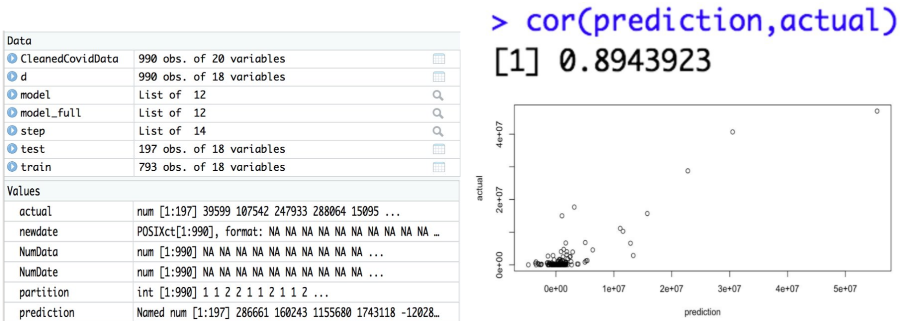


Figure 14: Scatterplot of actual value and prediction value from the model

performance than the first one that we made, but the adjusted R-squared is not very high. This might be because of our data is not large enough. However, including other potential variables would provide more efficiency in predicting the total vaccinations.

In the analysis, we apply other techniques to improve the model by using interaction terms. Furthermore, we could refine our model by using n-fold validation to improve the model predicting a number of total vaccinations as the dependent variable.

We then created a new data frame with the same number of observations as the full cleaned data set but two variables less. Then we partition the data set, creating a partition, using the rows from data frame d as we can see has 990 observations and 18 variables. After we create a partition, we put 80% in training and 20% in test. As we can see from the screenshot, after running the function, we get two sets from data frame d. then run the model using the training data and using prediction function on test data. Now after running a correlation between prediction and actual, the result came out is 89%. From the plot we do see that there are some correlation here even though couple of outliers occur, but we can mostly see the correlation.



```
Data
CleanedCovidData   990 obs. of 20 variables
d                  990 obs. of 18 variables
model              List of  12
model_full         List of  12
step               List of  14
test               197 obs. of 18 variables
train              793 obs. of 18 variables
Values
  actual           num [1:197] 39599 107542 247933 288064 15095 ...
  newdate          POSIXct[1:990], format: NA NA NA NA NA NA NA NA NA NA …
  NumData          num [1:990] NA NA NA NA NA NA NA NA NA NA ...
  NumDate          num [1:990] NA NA NA NA NA NA NA NA NA NA ...
  partition        int [1:990] 1 1 2 2 1 1 2 1 1 2 ...
  prediction       Named num [1:197] 286661 160243 1155680 1743118 -12028…
```

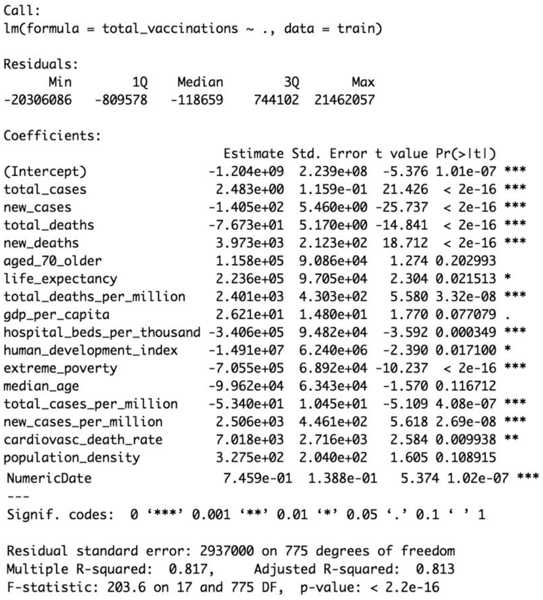
```
> cor(prediction,actual)
[1] 0.8943923
```

(a)

(b)

Figure 15

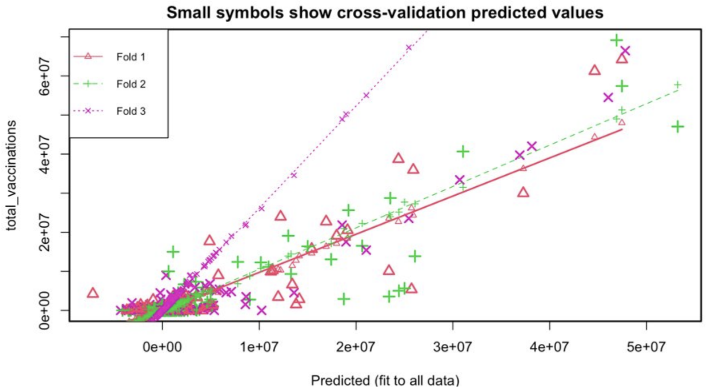
Figure 16 is a model summary on our training data. When we count all the variables and looking at the percentage of R squared is 81.7

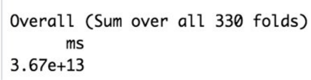


```
Call:
lm(formula = total_vaccinations ~ ., data = train)

Residuals:
      Min        1Q    Median        3Q       Max 
-20306086   -809578   -118659    744102  21462057 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -1.204e+09  2.239e+08  -5.376 1.01e-07 ***
total_cases               2.483e+00  1.159e-01  21.426  < 2e-16 ***
new_cases                -1.405e+02  5.460e+00 -25.737  < 2e-16 ***
total_deaths             -7.673e+01  5.170e+00 -14.841  < 2e-16 ***
new_deaths                3.973e+03  2.123e+02  18.712  < 2e-16 ***
aged_70_older             1.158e+05  9.086e+04   1.274 0.202993    
life_expectancy           2.236e+05  9.705e+04   2.304 0.021513 *  
total_deaths_per_million  2.401e+03  4.303e+02   5.580 3.32e-08 ***
gdp_per_capita            2.621e+01  1.480e+01   1.770 0.077079 .  
hospital_beds_per_thousand -3.406e+05  9.482e+04  -3.592 0.000349 ***
human_development_index  -1.491e+07  6.240e+06  -2.390 0.017100 *  
extreme_poverty          -7.055e+05  6.892e+04 -10.237  < 2e-16 ***
median_age               -9.962e+04  6.343e+04  -1.570 0.116712    
total_cases_per_million  -5.340e+01  1.045e+01  -5.109 4.08e-07 ***
new_cases_per_million     2.506e+03  4.461e+02   5.618 2.69e-08 ***
cardiovasc_death_rate     7.018e+03  2.716e+03   2.584 0.009938 ** 
population_density        3.275e+02  2.040e+02   1.605 0.108915    
NumericDate               7.459e-01  1.388e-01   5.374 1.02e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2937000 on 775 degrees of freedom
Multiple R-squared:  0.817,	Adjusted R-squared:  0.813 
F-statistic: 203.6 on 17 and 775 DF,  p-value: < 2.2e-16
```

Figure 16

Next, following is a cross validation can be used in validating the model. As we know that cross validation the function help measures of predictive accuracy for multiple linear regression. We store everything in varible OUT. CV.LM function, set m equal 3 which is 3-fold cross validation. After running the function, we can get the mean square for all 3 folds. From the plot we can evaluate our model is also good so far. There are still some outliers needs to be eliminated.



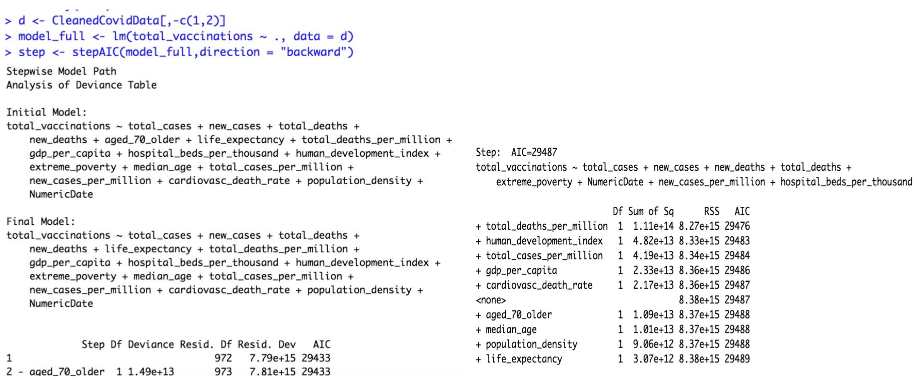


```
Overall (Sum over all 330 folds)
      ms
3.67e+13
```

Figure 17

we can dig further into that using stepwise regression, first by running the full model, using all the variables completely and eliminate one that least important in terms of adding predicting power to the model which is as we can see aged_70older. Based on AIC value. We eliminated couple of variables and from that. After couple testing, if we add adding more variables will decrease the value of r square.

Finally, we eliminate couple of variables, from the first plot we can see there might be couples Residual Multiplicative errors. Especially the red line which supposed to close to the straight line. But from QQ plot we can see all the way from -2 to 2 its very close to the line. Approximately more than 95 of the data so most of the residuals are following the normal distribution.



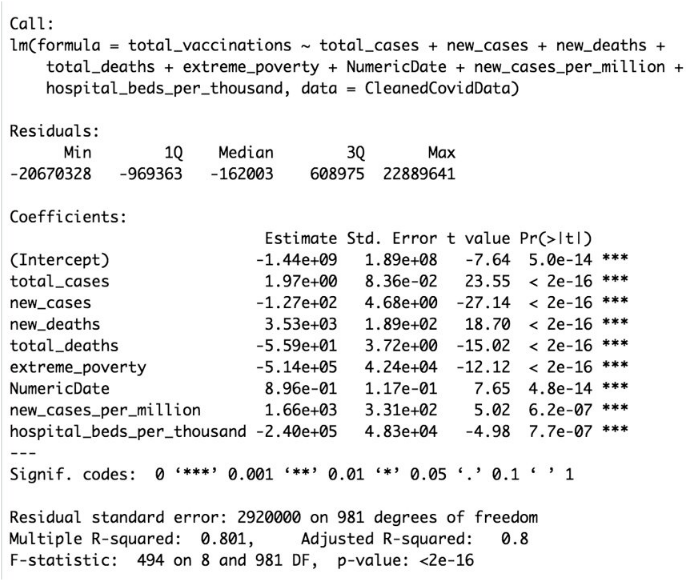
```
> d <- CleanedCovidData[,-c(1,2)]
> model_full <- lm(total_vaccinations ~ ., data = d)
> step <- stepAIC(model_full,direction = "backward")
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
total_vaccinations ~ total_cases + new_cases + total_deaths +
    new_deaths + aged_70_older + life_expectancy + total_deaths_per_million +
    gdp_per_capita + hospital_beds_per_thousand + human_development_index +
    extreme_poverty + median_age + total_cases_per_million +
    new_cases_per_million + cardiovasc_death_rate + population_density +
    NumericDate

Final Model:
total_vaccinations ~ total_cases + new_cases + total_deaths +
    new_deaths + life_expectancy + total_deaths_per_million +
    gdp_per_capita + hospital_beds_per_thousand + human_development_index +
    extreme_poverty + median_age + total_cases_per_million +
    new_cases_per_million + cardiovasc_death_rate + population_density +
    NumericDate


            Step Df Deviance Resid. Df Resid. Dev   AIC
1                                  972   7.79e+15 29433
2 - aged_70_older  1 1.49e+13      973   7.81e+15 29433
```

(a)

```
Step:  AIC=29487
total_vaccinations ~ total_cases + new_cases + new_deaths + total_deaths +
    extreme_poverty + NumericDate + new_cases_per_million + hospital_beds_per_thousand

                           Df Sum of Sq      RSS   AIC
+ total_deaths_per_million  1  1.11e+14 8.27e+15 29476
+ human_development_index   1  4.82e+13 8.33e+15 29483
+ total_cases_per_million   1  4.19e+13 8.34e+15 29484
+ gdp_per_capita            1  2.33e+13 8.36e+15 29486
+ cardiovasc_death_rate     1  2.17e+13 8.36e+15 29487
<none>                                  8.38e+15 29487
+ aged_70_older             1  1.09e+13 8.37e+15 29488
+ median_age                1  1.01e+13 8.37e+15 29488
+ population_density        1  9.06e+12 8.37e+15 29488
+ life_expectancy           1  3.07e+12 8.38e+15 29489
```

(b)

Figure 18



```
Call:
lm(formula = total_vaccinations ~ total_cases + new_cases + new_deaths +
    total_deaths + extreme_poverty + NumericDate + new_cases_per_million +
    hospital_beds_per_thousand, data = CleanedCovidData)

Residuals:
      Min        1Q    Median        3Q       Max
-20670328   -969363   -162003    608975  22889641

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.44e+09   1.89e+08   -7.64  5.0e-14 ***
total_cases                1.97e+00   8.36e-02   23.55  < 2e-16 ***
new_cases                 -1.27e+02   4.68e+00  -27.14  < 2e-16 ***
new_deaths                 3.53e+03   1.89e+02   18.70  < 2e-16 ***
total_deaths              -5.59e+01   3.72e+00  -15.02  < 2e-16 ***
extreme_poverty           -5.14e+05   4.24e+04  -12.12  < 2e-16 ***
NumericDate                8.96e-01   1.17e-01    7.65  4.8e-14 ***
new_cases_per_million      1.66e+03   3.31e+02    5.02  6.2e-07 ***
hospital_beds_per_thousand -2.40e+05   4.83e+04   -4.98  7.7e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2920000 on 981 degrees of freedom
Multiple R-squared:  0.801,     Adjusted R-squared:   0.8
F-statistic:  494 on 8 and 981 DF,  p-value: <2e-16
```

Figure 19

One of the potential models that we explored was how the human development index, extreme poverty rate, and the date the data was recorded all interact and affect the total amount of vaccine doses given. First, we explored the first order model, where we hypothesized that as the numeric date, human development index, and extreme poverty rate changes per unit, the number of total vaccinations given will change as well. After fitting the model, we were

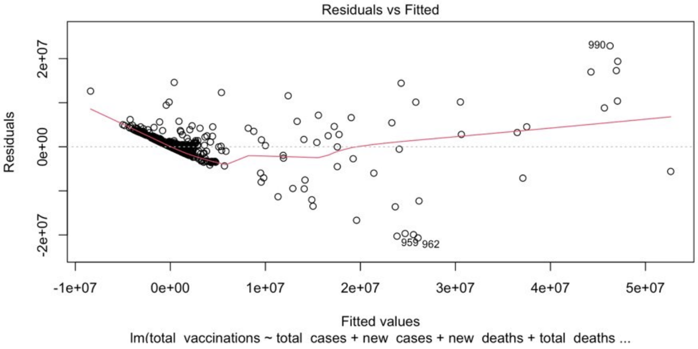


Figure 20

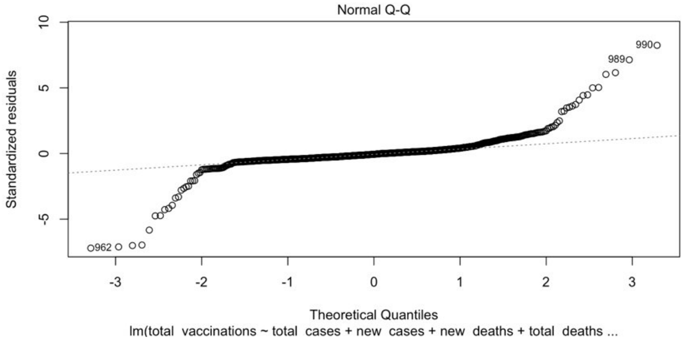
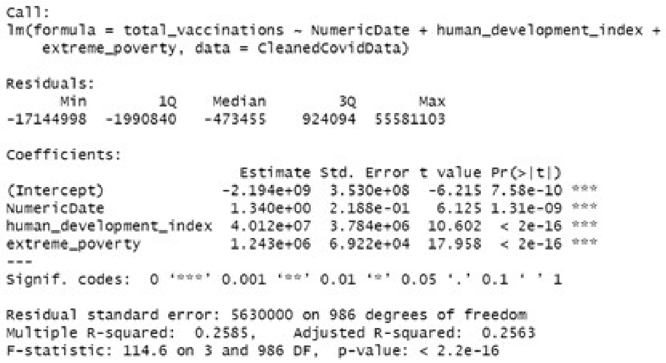


Figure 21

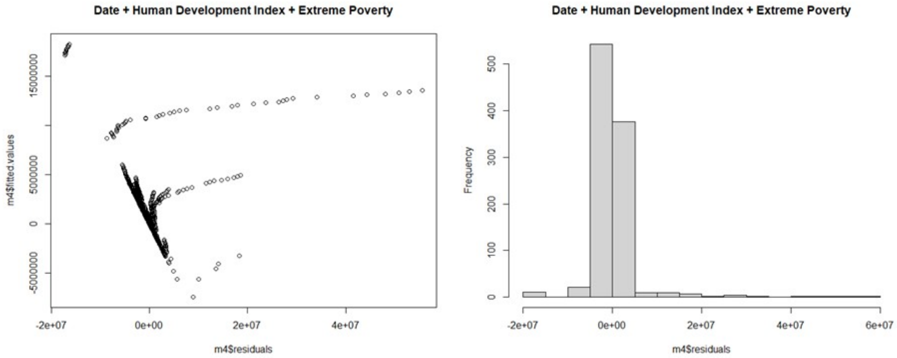
able to conclude that we have sufficient evidence to support the claim that the date, human development index, and extreme poverty rate all influence the total vaccination numbers. We can make this claim because the p-value is statistically significant when testing at a significance level of 0.05. However, this does not mean that this is a perfect model. First, as seen in the figure, the model only has an adjusted R-squared value of 0.2563. While this explains for a small amount of the variation, the model could still be much better. Additionally, when looking at the residual plot provided for this model, we can see that there is clearly a trend in the residual versus fitted values. As we look at the residual histogram, we can see that while the residuals are somewhat normally distributed, we can see improvement here as well.

This leads to the complete second order model where we fit the same variables of date, extreme poverty rates, and human development index to include more interactions between the variables. As seen in the output from the R software, we can continue to support our claim that these three variables influence the total vaccination numbers. While we have a similar p-value as the previous



```
Call:
lm(formula = total_vaccinations ~ NumericDate + human_development_index +
    extreme_poverty, data = CleanedCovidData)

Residuals:
     Min       1Q   Median       3Q      Max
-17144998 -1990840  -473455   924094 55581103

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.194e+09  3.530e+08  -6.215 7.58e-10 ***
NumericDate              1.340e+00  2.188e-01   6.125 1.31e-09 ***
human_development_index  4.012e+07  3.784e+06  10.602  < 2e-16 ***
extreme_poverty          1.243e+06  6.922e+04  17.958  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5630000 on 986 degrees of freedom
Multiple R-squared:  0.2585,    Adjusted R-squared:  0.2563
F-statistic: 114.6 on 3 and 986 DF,  p-value: < 2.2e-16
```

Figure 22

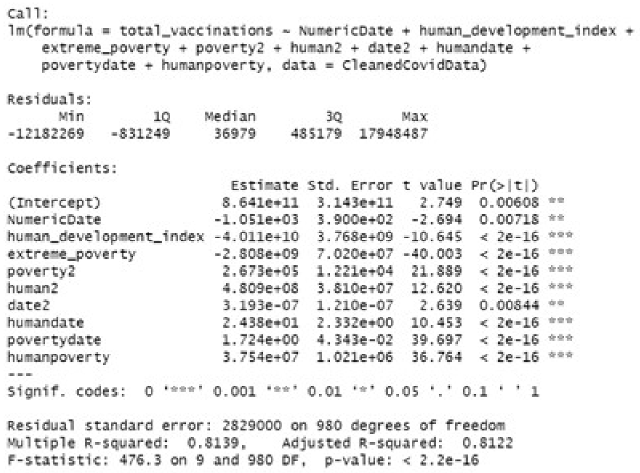


(a)

(b)

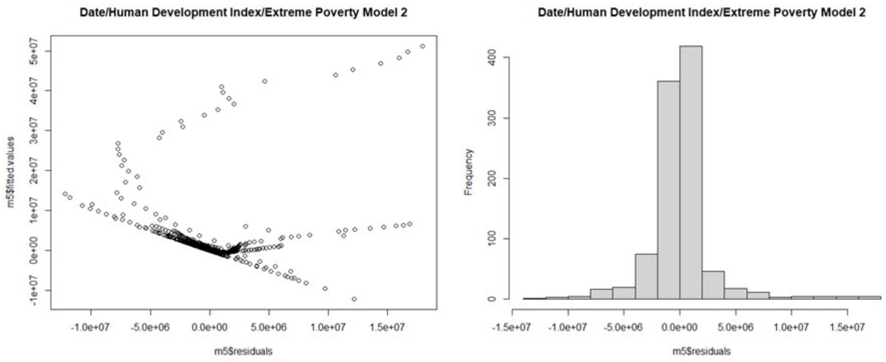
Figure 23

model, our adjusted R-squared is much better at 0.8122, or 81%. While the first order model can explain for 25% of the variation in the results, our new model can explain for over 81%, which is a huge improvement. Additionally, if we look at the plot of residual values versus fitted values for this model, we can see that the trend dissipated a little bit and improve. Unfortunately, we still observed some trends in the plot, telling us that we need to explore more models for a better fit. Still, our histogram of residual values has improved significantly, with a much more normal distribution of values.



```
Call:
lm(formula = total_vaccinations ~ NumericDate + human_development_index +
    extreme_poverty + poverty2 + human2 + date2 + humandate +
    povertydate + humanpoverty, data = CleanedCovidData)

Residuals:
     Min        1Q    Median        3Q       Max
-12182269   -831249     36979    485179  17948487

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              8.641e+11  3.143e+11   2.749  0.00608 **
NumericDate             -1.051e+03  3.900e+02  -2.694  0.00718 **
human_development_index -4.011e+10  3.768e+09 -10.645  < 2e-16 ***
extreme_poverty         -2.808e+09  7.020e+07 -40.003  < 2e-16 ***
poverty2                 2.673e+05  1.221e+04  21.889  < 2e-16 ***
human2                   4.809e+08  3.810e+07  12.620  < 2e-16 ***
date2                    3.193e-07  1.210e-07   2.639  0.00844 **
humandate                2.438e+01  2.332e+00  10.453  < 2e-16 ***
povertydate              1.724e+00  4.343e-02  39.697  < 2e-16 ***
humanpoverty             3.754e+07  1.021e+06  36.764  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2829000 on 980 degrees of freedom
Multiple R-squared:  0.8139,    Adjusted R-squared:  0.8122
F-statistic: 476.3 on 9 and 980 DF,  p-value: < 2.2e-16
```

Figure 24



(a)

(b)

Figure 25

Now, we further consider parameters cleaned data set for what we are calling here, Model Vaccination. These parameters are location, total_cases, new_cases, new_deaths, total_vaccinations, total_deaths_per_million, total_cases_per_million, and NumericDate. The plots in Figure 26 gives the interpretation of variables used for the model. These variables help us to predict the variation of total_deaths, total_cases with regard to the total_vaccinations over locations in the world.

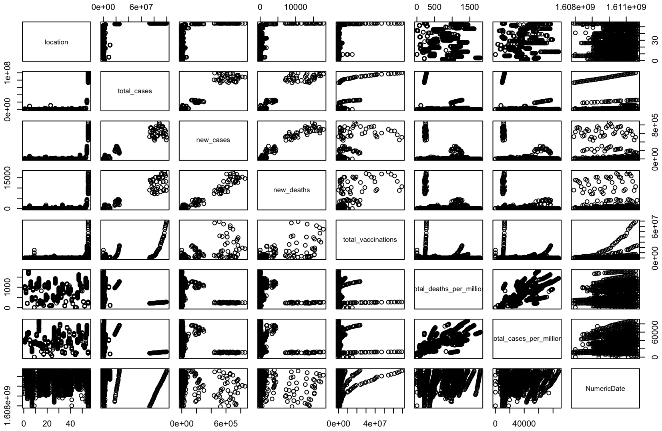
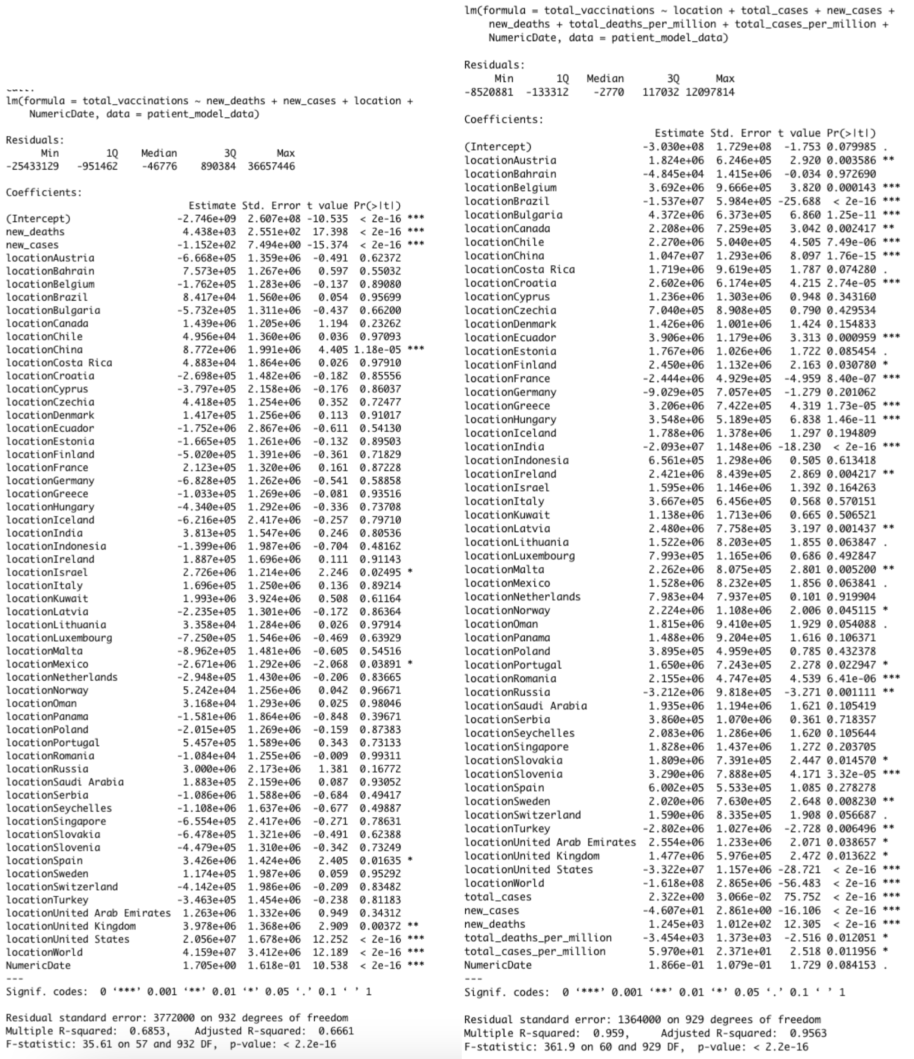


Figure 26: Plot demonstrating the parameters used in the model

Here, the response variable is total_vaccinations and the explanatory variables are location, total_cases, new_cases, new_deaths, total_deaths_per_million, total_cases_per_million, NumericDate. This is a Complete Second Order model. The model we call COVID Aggression describes the new deaths and new cases since the vaccination has been introduced around the world. The model we call Vaccination describes the total deaths and total cases around the world.

For analyzing the cleaned data, complete second order regression model was used as shown in Figure 27a. where the adjusted R2 obtained is 0.6661 which means 66.61% of variation in the total_vaccinations is accounted by the model. Also, a complete second order regression model was used as shown in Figure 27b. where the adjusted R2 obtained is 0.9563 which means 95.63% of variation in the total_vaccinations is accounted by the model.



```
lm(formula = total_vaccinations ~ new_deaths + new_cases + location +
    NumericDate, data = patient_model_data)

Residuals:
      Min        1Q    Median        3Q       Max 
-25433129   -951462    -46776    890384  36657446 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -2.746e+09  2.607e+08 -10.535  < 2e-16 ***
new_deaths                   4.438e+03  2.551e+02  17.398  < 2e-16 ***
new_cases                   -1.152e+02  7.494e+00 -15.374  < 2e-16 ***
locationAustria             -6.668e+05  1.359e+06  -0.491  0.62372    
locationBahrain              7.573e+05  1.267e+06   0.597  0.55032    
locationBelgium             -1.762e+05  1.283e+06  -0.137  0.89080    
locationBrazil               8.417e+04  1.560e+06   0.054  0.95699    
locationBulgaria            -5.732e+05  1.311e+06  -0.437  0.66200    
locationCanada               1.439e+06  1.205e+06   1.194  0.23262    
locationChile                4.956e+04  1.360e+06   0.036  0.97093    
locationChina                8.772e+06  1.991e+06   4.405 1.18e-05 ***
locationCosta Rica           4.883e+04  1.864e+06   0.026  0.97910    
locationCroatia             -2.698e+05  1.482e+06  -0.182  0.85556    
locationCyprus              -3.797e+05  2.158e+06  -0.176  0.86037    
locationCzechia              4.418e+05  1.254e+06   0.352  0.72477    
locationDenmark              1.417e+05  1.256e+06   0.113  0.91017    
locationEcuador             -1.752e+06  2.867e+06  -0.611  0.54130    
locationEstonia             -1.665e+05  1.261e+06  -0.132  0.89503    
locationFinland             -5.020e+05  1.391e+06  -0.361  0.71829    
locationFrance               2.123e+05  1.320e+06   0.161  0.87228    
locationGermany             -6.828e+05  1.262e+06  -0.541  0.58858    
locationGreece              -1.033e+05  1.269e+06  -0.081  0.93516    
locationHungary             -4.340e+05  1.292e+06  -0.336  0.73708    
locationIceland             -6.216e+05  2.417e+06  -0.257  0.79710    
locationIndia                3.813e+05  1.547e+06   0.246  0.80536    
locationIndonesia           -1.399e+06  1.987e+06  -0.704  0.48162    
locationIreland              1.887e+05  1.696e+06   0.111  0.91143    
locationIsrael               2.726e+06  1.214e+06   2.246  0.02495 *  
locationItaly                1.696e+05  1.250e+06   0.136  0.89214    
locationKuwait               1.993e+06  3.924e+06   0.508  0.61164    
locationLatvia              -2.235e+05  1.301e+06  -0.172  0.86364    
locationLithuania            3.358e+04  1.284e+06   0.026  0.97914    
locationLuxembourg          -7.250e+05  1.546e+06  -0.469  0.63929    
locationMalta               -8.962e+05  1.481e+06  -0.605  0.54516    
locationMexico              -2.671e+06  1.292e+06  -2.068  0.03891 *  
locationNetherlands         -2.948e+05  1.430e+06  -0.206  0.83665    
locationNorway               5.242e+04  1.256e+06   0.042  0.96671    
locationOman                 3.168e+04  1.293e+06   0.025  0.98046    
locationPanama              -1.581e+06  1.864e+06  -0.848  0.39671    
locationPoland              -2.015e+05  1.269e+06  -0.159  0.87383    
locationPortugal             5.457e+05  1.589e+06   0.343  0.73133    
locationRomania             -1.084e+04  1.255e+06  -0.009  0.99311    
locationRussia               3.000e+06  2.173e+06   1.381  0.16772    
locationSaudi Arabia         1.883e+05  2.159e+06   0.087  0.93052    
locationSerbia              -1.086e+06  1.588e+06  -0.684  0.49417    
locationSeychelles          -1.108e+06  1.637e+06  -0.677  0.49887    
locationSingapore           -6.554e+05  2.417e+06  -0.271  0.78631    
locationSlovakia            -6.478e+05  1.321e+06  -0.491  0.62388    
locationSlovenia            -4.479e+05  1.310e+06  -0.342  0.73249    
locationSpain                3.426e+06  1.424e+06   2.405  0.01635 *  
locationSweden               1.174e+05  1.987e+06   0.059  0.95292    
locationSwitzerland         -4.142e+05  1.986e+06  -0.209  0.83482    
locationTurkey              -3.463e+05  1.454e+06  -0.238  0.81183    
locationUnited Arab Emirates 1.263e+06  1.332e+06   0.949  0.34312    
locationUnited Kingdom       3.978e+06  1.368e+06   2.909  0.00372 ** 
locationUnited States        2.056e+07  1.678e+06  12.252  < 2e-16 ***
locationWorld                4.159e+07  3.412e+06  12.189  < 2e-16 ***
NumericDate                  1.705e+00  1.618e-01  10.538  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3772000 on 932 degrees of freedom
Multiple R-squared:  0.6853,	Adjusted R-squared:  0.6661 
F-statistic: 35.61 on 57 and 932 DF,  p-value: < 2.2e-16
```

(a) Model COVID aggression

```
lm(formula = total_vaccinations ~ location + total_cases + new_cases +
    new_deaths + total_deaths_per_million + total_cases_per_million +
    NumericDate, data = patient_model_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-8520881  -133312    -2770   117032 12097814 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -3.030e+08  1.729e+08  -1.753 0.079985 .  
locationAustria              1.824e+06  6.246e+05   2.920 0.003586 ** 
locationBahrain             -4.845e+04  1.415e+06  -0.034 0.972690    
locationBelgium              3.692e+06  9.666e+05   3.820 0.000143 ***
locationBrazil              -1.537e+07  5.984e+05 -25.688  < 2e-16 ***
locationBulgaria             4.372e+06  6.373e+05   6.860 1.25e-11 ***
locationCanada               2.208e+06  7.259e+05   3.042 0.002417 ** 
locationChile                2.270e+06  5.040e+05   4.505 7.49e-06 ***
locationChina                1.047e+07  1.293e+06   8.097 1.76e-15 ***
locationCosta Rica           1.719e+06  9.619e+05   1.787 0.074280 .  
locationCroatia              2.602e+06  6.174e+05   4.215 2.74e-05 ***
locationCyprus               1.236e+06  1.303e+06   0.948 0.343160    
locationCzechia              7.040e+05  8.908e+05   0.790 0.429534    
locationDenmark              1.426e+06  1.001e+06   1.424 0.154833    
locationEcuador              3.906e+06  1.179e+06   3.313 0.000959 ***
locationEstonia              1.767e+06  1.026e+06   1.722 0.085454 .  
locationFinland              2.450e+06  1.132e+06   2.163 0.030780 *  
locationFrance              -2.444e+06  4.929e+05  -4.959 8.40e-07 ***
locationGermany             -9.029e+05  7.057e+05  -1.279 0.201062    
locationGreece               3.206e+06  7.422e+05   4.319 1.73e-05 ***
locationHungary              3.548e+06  5.189e+05   6.838 1.46e-11 ***
locationIceland              1.788e+06  1.378e+06   1.297 0.194809    
locationIndia               -2.093e+07  1.148e+06 -18.230  < 2e-16 ***
locationIndonesia            6.561e+05  1.298e+06   0.505 0.613418    
locationIreland              2.421e+06  8.439e+05   2.869 0.004217 ** 
locationIsrael               1.595e+06  1.146e+06   1.392 0.164263    
locationItaly                3.667e+05  6.456e+05   0.568 0.570151    
locationKuwait               1.138e+06  1.713e+06   0.665 0.506521    
locationLatvia               2.480e+06  7.758e+05   3.197 0.001437 ** 
locationLithuania            1.522e+06  8.203e+05   1.855 0.063847 .  
locationLuxembourg           7.993e+05  1.165e+06   0.686 0.492847    
locationMalta                2.262e+06  8.075e+05   2.801 0.005200 ** 
locationMexico               1.528e+06  8.232e+05   1.856 0.063841 .  
locationNetherlands          7.983e+04  7.937e+05   0.101 0.919904    
locationNorway               2.224e+06  1.108e+06   2.006 0.045115 *  
locationOman                 1.815e+06  9.410e+05   1.929 0.054088 .  
locationPanama               1.488e+06  9.204e+05   1.616 0.106371    
locationPoland               3.895e+05  4.959e+05   0.785 0.432378    
locationPortugal             1.650e+06  7.243e+05   2.278 0.022947 *  
locationRomania              2.155e+06  4.747e+05   4.539 6.41e-06 ***
locationRussia              -3.212e+06  9.818e+05  -3.271 0.001111 ** 
locationSaudi Arabia         1.935e+06  1.194e+06   1.621 0.105419    
locationSerbia               3.860e+05  1.070e+06   0.361 0.718357    
locationSeychelles           2.083e+06  1.286e+06   1.620 0.105644    
locationSingapore            1.828e+06  1.437e+06   1.272 0.203705    
locationSlovakia             1.809e+06  7.391e+05   2.447 0.014570 *  
locationSlovenia             3.290e+06  7.888e+05   4.171 3.32e-05 ***
locationSpain                6.002e+05  5.533e+05   1.085 0.278278    
locationSweden               2.020e+06  7.630e+05   2.648 0.008230 ** 
locationSwitzerland          1.590e+06  8.335e+05   1.908 0.056687 .  
locationTurkey              -2.802e+06  1.027e+06  -2.728 0.006496 ** 
locationUnited Arab Emirates 2.554e+06  1.233e+06   2.071 0.038657 *  
locationUnited Kingdom       1.477e+06  5.976e+05   2.472 0.013622 *  
locationUnited States       -3.322e+07  1.157e+06 -28.721  < 2e-16 ***
locationWorld               -1.618e+08  2.865e+06 -56.483  < 2e-16 ***
total_cases                  2.322e+00  3.066e-02  75.752  < 2e-16 ***
new_cases                   -4.607e+01  2.861e+00 -16.106  < 2e-16 ***
new_deaths                   1.245e+03  1.012e+02  12.305  < 2e-16 ***
total_deaths_per_million    -3.454e+03  1.373e+03  -2.516 0.012051 *  
total_cases_per_million      5.970e+01  2.371e+01   2.518 0.011956 *  
NumericDate                  1.866e-01  1.079e-01   1.729 0.084153 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1364000 on 929 degrees of freedom
Multiple R-squared:  0.959,	Adjusted R-squared:  0.9563 
F-statistic: 361.9 on 60 and 929 DF,  p-value: < 2.2e-16
```

(b) Model vaccination

Figure 27

The data is randomly split into training and testing sets with the ratio 7:2

for 10 times and all compared algorithms are run on the same splits to take the average performance for evaluation. Model vaccination fits well with correlation coefficient of 0.98 which is significant to data.

QQ plot in Figure 28 shows that the two quantiles, standardized residuals and theoretical quantiles, do not provide accurate information regarding the administered COVID-19 vaccination doses. Even though the adjusted R-squared results in high variability of total_vaccinations; from QQ plot, we see that the regression line is not aligned with the values resulting in these residuals. This model interprets that the distribution of the administered COVID-19 doses over few locations around the would is not consistent with the counts of new_cases and new_deaths. Thus, using the parameter location is not best fit for the model.

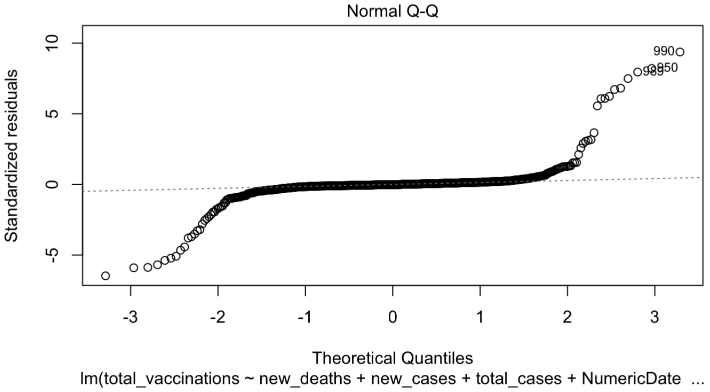
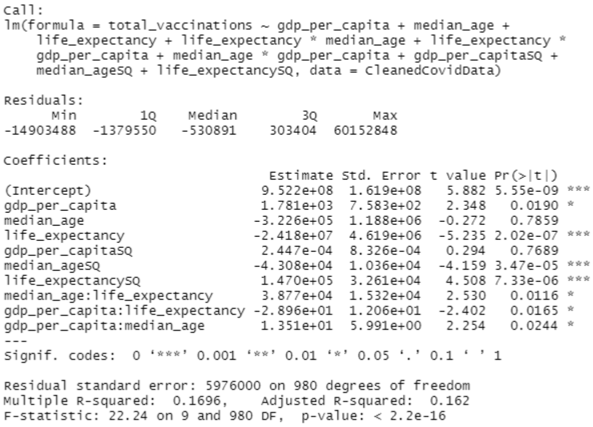


Figure 28: QQ plot for model vaccination

Another model that we built was for analyzing the utility of gdp_per_capita, median_age, and life_expectancy for predicting total_vaccinations. First, we look at the complete second order model using these three independent variables. We see in Figure 29 that the adjusted R-squared for this model is 0.162. The p-value for the overall utility is very small, but in an effort to achieve greater predictive utility, we look to the individual t-tests for each term.

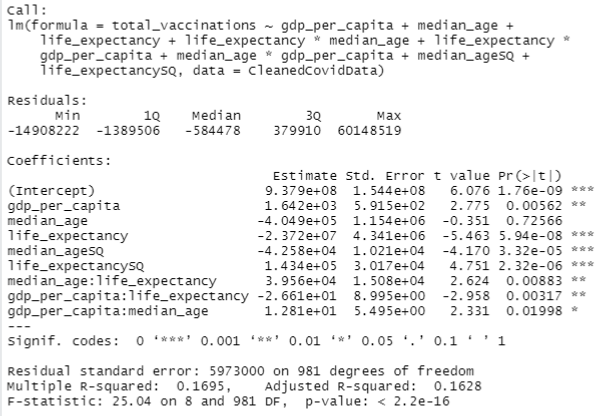
We see that the term associated with the square of the GDP per capita of the country fails the t-test. The median age term fails as well, but it's square term and interaction terms do no fail so we will retain each of those terms in the following model. In Figure 30, we see a slight improvement of the adjusted R-squared value and that all the terms pass their t-tests except the term for median age, but we have to keep it in the model when we include its higher order terms.

Removing all terms that involve median age does not improve our model and we omit those attempts here for simplicity. Instead, from this model, we see that the independent variables considered here are not good candidates for



```
Call:
lm(formula = total_vaccinations ~ gdp_per_capita + median_age +
    life_expectancy + life_expectancy * median_age + life_expectancy *
    gdp_per_capita + median_age * gdp_per_capita + gdp_per_capitaSQ +
    median_ageSQ + life_expectancySQ, data = CleanedCovidData)

Residuals:
     Min       1Q   Median       3Q      Max
-14903488  -1379550  -530891   303404  60152848

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    9.522e+08  1.619e+08   5.882 5.55e-09 ***
gdp_per_capita                 1.781e+03  7.583e+02   2.348   0.0190 *
median_age                    -3.226e+05  1.188e+06  -0.272   0.7859
life_expectancy               -2.418e+07  4.619e+06  -5.235 2.02e-07 ***
gdp_per_capitaSQ               2.447e-04  8.326e-04   0.294   0.7689
median_ageSQ                  -4.308e+04  1.036e+04  -4.159 3.47e-05 ***
life_expectancySQ              1.470e+05  3.261e+04   4.508 7.33e-06 ***
median_age:life_expectancy     3.877e+04  1.532e+04   2.530   0.0116 *
gdp_per_capita:life_expectancy -2.896e+01  1.206e+01  -2.402   0.0165 *
gdp_per_capita:median_age      1.351e+01  5.991e+00   2.254   0.0244 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5976000 on 980 degrees of freedom
Multiple R-squared:  0.1696,    Adjusted R-squared:  0.162
F-statistic: 22.24 on 9 and 980 DF,  p-value: < 2.2e-16
```

Figure 29: Complete second order model for analyzing the utility of gdp-per_capita, median_age, and life_expectancy for predicting total_vaccinations.



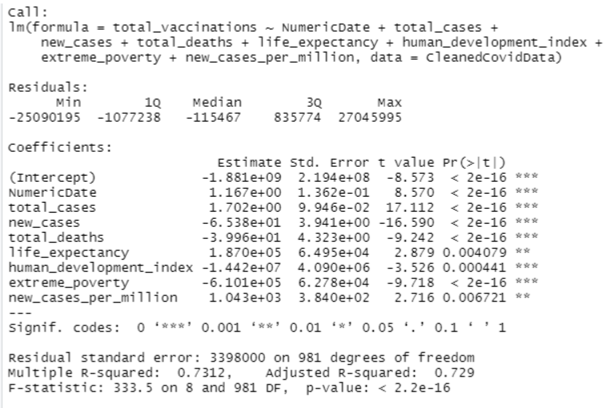
```
Call:
lm(formula = total_vaccinations ~ gdp_per_capita + median_age +
    life_expectancy + life_expectancy * median_age + life_expectancy *
    gdp_per_capita + median_age * gdp_per_capita + median_ageSQ +
    life_expectancySQ, data = CleanedCovidData)

Residuals:
     Min       1Q   Median       3Q      Max
-14908222  -1389506  -584478   379910  60148519

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    9.379e+08  1.544e+08   6.076 1.76e-09 ***
gdp_per_capita                 1.642e+03  5.915e+02   2.775  0.00562 **
median_age                    -4.049e+05  1.154e+06  -0.351  0.72566
life_expectancy               -2.372e+07  4.341e+06  -5.463 5.94e-08 ***
median_ageSQ                  -4.258e+04  1.021e+04  -4.170 3.32e-05 ***
life_expectancySQ              1.434e+05  3.017e+04   4.751 2.32e-06 ***
median_age:life_expectancy     3.956e+04  1.508e+04   2.624  0.00883 **
gdp_per_capita:life_expectancy -2.661e+01  8.995e+00  -2.958  0.00317 **
gdp_per_capita:median_age      1.281e+01  5.495e+00   2.331  0.01998 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5973000 on 981 degrees of freedom
Multiple R-squared:  0.1695,    Adjusted R-squared:  0.1628
F-statistic: 25.04 on 8 and 981 DF,  p-value: < 2.2e-16
```

Figure 30: This second order model only removes the squared term for gpd_per_capita.

a final model.

Analysis

After gaining further insights into the independent variables within this data, we were able to narrow our interest. In Figure 31 shows a first order model using predictor variables that we have determined throughout our previous models to have the best chance at creating a useful model when considered together. We can see that the F-statistic has a good p-value, each term passes it's t-test, and the adjusted R-squared value is 0.729.



```
Call:
lm(formula = total_vaccinations ~ NumericDate + total_cases +
    new_cases + total_deaths + life_expectancy + human_development_index +
    extreme_poverty + new_cases_per_million, data = CleanedCovidData)

Residuals:
      Min        1Q    Median        3Q       Max
-25090195  -1077238   -115467    835774  27045995

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.881e+09  2.194e+08  -8.573  < 2e-16 ***
NumericDate              1.167e+00  1.362e-01   8.570  < 2e-16 ***
total_cases              1.702e+00  9.946e-02  17.112  < 2e-16 ***
new_cases               -6.538e+01  3.941e+00 -16.590  < 2e-16 ***
total_deaths            -3.996e+01  4.323e+00  -9.242  < 2e-16 ***
life_expectancy          1.870e+05  6.495e+04   2.879 0.004079 **
human_development_index -1.442e+07  4.090e+06  -3.526 0.000441 ***
extreme_poverty         -6.101e+05  6.278e+04  -9.718  < 2e-16 ***
new_cases_per_million    1.043e+03  3.840e+02   2.716 0.006721 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3398000 on 981 degrees of freedom
Multiple R-squared:  0.7312,    Adjusted R-squared:  0.729
F-statistic: 333.5 on 8 and 981 DF,  p-value: < 2.2e-16
```

Figure 31: First order model using our top candidate predictor variables.

Now that we know which independent variables we are interested in using in our model, we look at correlation plots of these variables to determine which interaction terms and square terms to include. As we see in Figure 32, there are potential interactions between independent variables. In Figure 33, we show a model that includes our most useful independent variables for predicting as well as interaction terms between them that appeared useful.

We can see that this model does pass the F-test for overall utility as well as explaining 93% of the variability in total vaccinations according to the adjusted R-squared value. However, many of the β estimates do not pass their t-tests. Using stepwise regression to remove terms from the model, we find that we can retain an adjusted R-squared value of nearly 0.92 with only two independent variables and their interaction term.

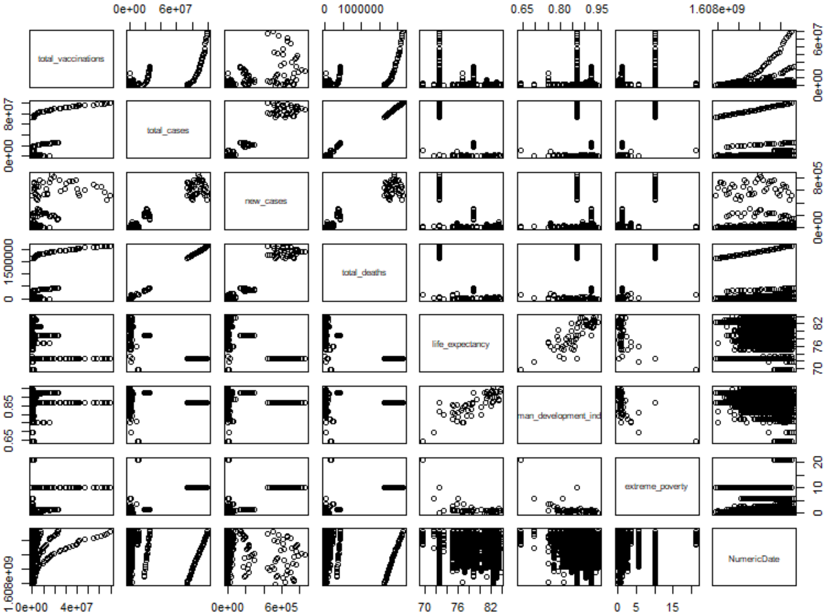
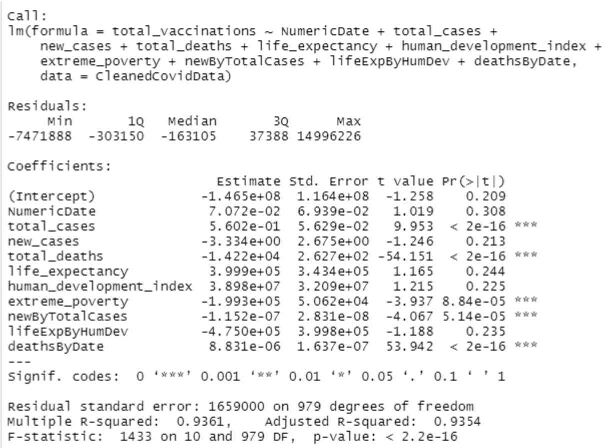


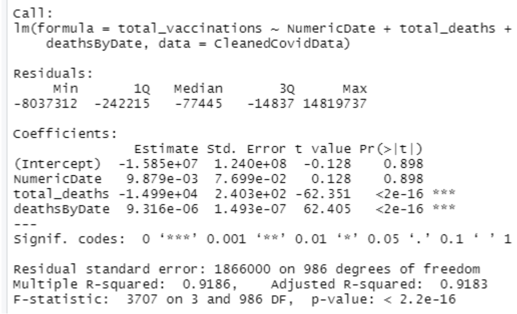
Figure 32: Correlation plots for top candidate variables.



```
Call:
lm(formula = total_vaccinations ~ NumericDate + total_cases +
    new_cases + total_deaths + life_expectancy + human_development_index +
    extreme_poverty + newByTotalCases + lifeExpByHumDev + deathsByDate,
    data = CleanedCovidData)

Residuals:
     Min       1Q   Median       3Q      Max
-7471888  -303150  -163105    37388 14996226

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.465e+08  1.164e+08  -1.258    0.209
NumericDate              7.072e-02  6.939e-02   1.019    0.308
total_cases              5.602e-01  5.629e-02   9.953  < 2e-16 ***
new_cases               -3.334e+00  2.675e+00  -1.246    0.213
total_deaths            -1.422e+04  2.627e+02 -54.151  < 2e-16 ***
life_expectancy          3.999e+05  3.434e+05   1.165    0.244
human_development_index  3.898e+07  3.209e+07   1.215    0.225
extreme_poverty         -1.993e+05  5.062e+04  -3.937 8.84e-05 ***
newByTotalCases         -1.152e-07  2.831e-08  -4.067 5.14e-05 ***
lifeExpByHumDev         -4.750e+05  3.998e+05  -1.188    0.235
deathsByDate             8.831e-06  1.637e-07  53.942  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1659000 on 979 degrees of freedom
Multiple R-squared:  0.9361,    Adjusted R-squared:  0.9354
F-statistic:  1433 on 10 and 979 DF,  p-value: < 2.2e-16
```

Figure 33: Our candidate variable terms as well as interaction terms.

In our simplest model that still explains almost 92% of the variability in total vaccinations (Figure 34), we only consider total deaths, numeric date, and their interaction term. Figure 35 shows the residual plot for this model and we see that we are not satisfied and that a clear pattern is hard to detect. Since we are dealing with very large numbers, we perform a log transformation on our independent variables. After looking at models with log transformations applied to each and both of our variables, we find that we can get our residuals to be more well behaved by simply applying the log transformation on the total deaths. Applying the log transformation to both independent variables provides the same results so for simplicity, we only consider the transformation on total deaths.



```
Call:
lm(formula = total_vaccinations ~ NumericDate + total_deaths + 
    deathsByDate, data = CleanedCovidData)

Residuals:
     Min       1Q   Median       3Q      Max 
-8037312  -242215   -77445   -14837 14819737 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.585e+07  1.240e+08  -0.128    0.898    
NumericDate   9.879e-03  7.699e-02   0.128    0.898    
total_deaths -1.499e+04  2.403e+02 -62.351   <2e-16 ***
deathsByDate  9.316e-06  1.493e-07  62.405   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1866000 on 986 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.9183 
F-statistic:  3707 on 3 and 986 DF,  p-value: < 2.2e-16
```

Figure 34: This simple model passes its tests and still has a high adjusted R-squared.

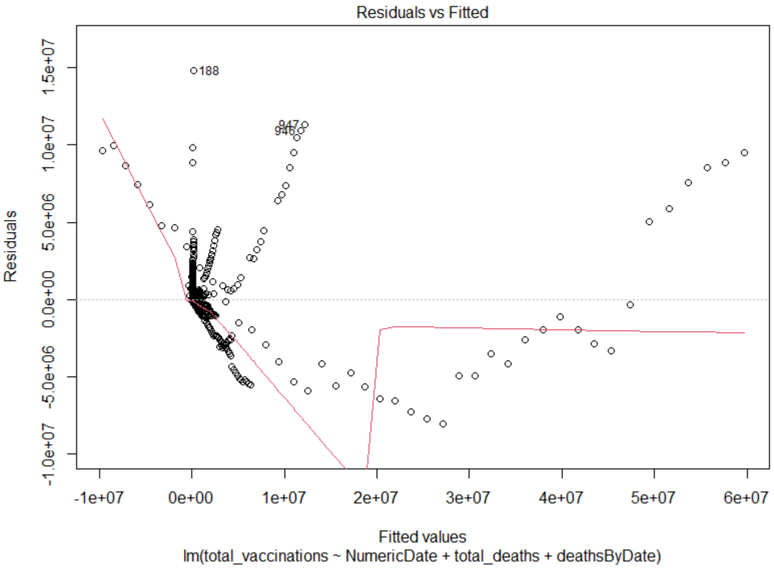


Figure 35: Residual plot of our simplest useful model.

The residual plot that we obtain by using the log transformed total deaths in our model shows us a better result as we see in Figure 36. This version of the residual plot emphasizes a pattern that was present in Figure 35, but was less obvious because of the greater spread in residuals.

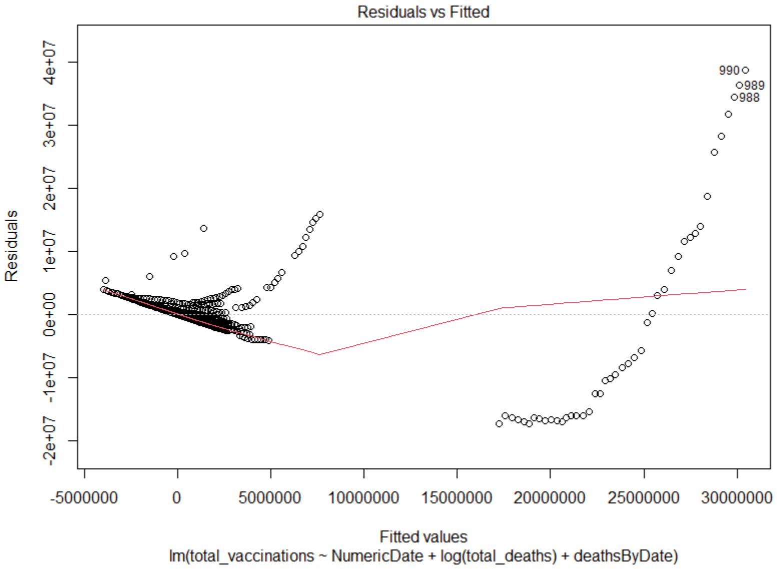
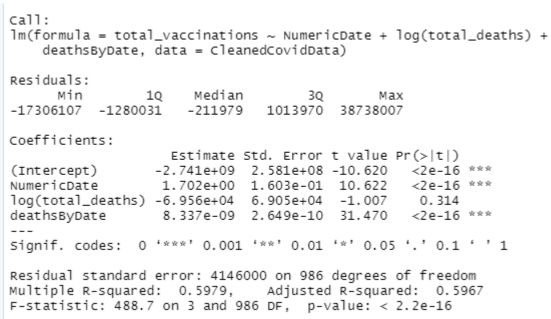


Figure 36: Residual plot of our simplest useful model after applying a log transformation to total deaths.

The pattern we recognize can be considered as three parts; the majority of the data, a small arm just beyond the majority, and a larger arm well beyond the majority. These arms are an indication that our two variables and their interaction are not all that is needed to explain variability in total vaccinations. The log transformation is helpful for getting a better understanding of our residuals, but as we see in Figure 37 by looking at the adjusted R-squared, this model is not nearly as useful as without the log transform.



```
Call:
lm(formula = total_vaccinations ~ NumericDate + log(total_deaths) +
    deathsByDate, data = CleanedCovidData)

Residuals:
     Min        1Q    Median        3Q       Max
-17306107  -1280031   -211979   1013970  38738007

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.741e+09  2.581e+08 -10.620   <2e-16 ***
NumericDate       1.702e+00  1.603e-01  10.622   <2e-16 ***
log(total_deaths) -6.956e+04  6.905e+04  -1.007    0.314
deathsByDate      8.337e-09  2.649e-10  31.470   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4146000 on 986 degrees of freedom
Multiple R-squared:  0.5979,    Adjusted R-squared:  0.5967
F-statistic: 488.7 on 3 and 986 DF,  p-value: < 2.2e-16
```

Figure 37: This model has the log transformed total deaths appears to be less useful for predicting total vaccinations.

Conclusion

During our exploration of COVID-19 vaccination numbers, we were able to identify variables that have relationships with the number of vaccines administered. Independent variables such as time and date, the number of deaths from COVID-19, and extreme poverty rate all had significant relationships to the total number of vaccines given. During our process of analyzing this vaccination data, we were able to develop a model that performed well and explained for a large amount of variability among the data. Interestingly, the final model that did the best was one of the simpler models that we tested. Through our testing, we have enough evidence to support the claim that we can predict the number of vaccinations using the independent variables of the number of deaths from COVID-19 and the date of interest.

It is also important to make note of the limitations of our study. The primary limitation that we faced is that we were working with a limited sample size due to the novel nature of world vaccination data. The first COVID-19 vaccination being given emergency-use authorization in the United States in December of 2020. While some countries were able to provide vaccinations sooner, the data was still very limited at the start because the data was very recent. If one was to conduct the same investigation a few years from now, where more information is available, they might be able to get more robust results.

To conclude, this analysis explored the relationship that vaccination numbers have with other potential explanatory variables. Our results suggest that overtime, more vaccinations are being administered at a significant rate. Additionally, the results infer that the number of deaths from the corona virus has a relationship with the number of vaccines given. These results hint at the idea that nations that are hardest hit by the virus might be the places that will eventually give out the most vaccinations over time. Still, we need more data to definitively make that claim. In future studies, exploring how well individual countries are doing in the vaccination race and comparing them would be an interesting way to test this claim.

Table 1: Independent variables and their descriptions

continent	Continent of the geographical location
location	Geographical location
date	Date of observation
total_cases	Total confirmed cases of COVID-19
new_cases	New confirmed cases of COVID-19
total_deaths	Total deaths attributed to COVID-19
new_deaths	New deaths attributed to COVID-19
Aged_70_older	Share of the population that is 70 years and older in 2015
life_expectancy	Life expectancy at birth in 2019
total_deaths_per_million	Total deaths attributed to COVID-19 per 1,000,000 people
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
hospital_beds_per_thousand	Hospital beds per 1,000 people, most recent year available since 2010
human_development_index	Summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
median_age	Median age of the population, UN projection for 2020
total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people
new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people
cardiovasc_death_rate	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)
handwashing_facilities	Share of the population with basic handwashing facilities on premises, most recent year available
population_density	Number of people divided by land area, measured in square kilometers, most recent year available