Impacts of Diet on COVID Cases of 170 countries

Weiyang Wang, Yulin Song, and Yuhan Zhou

University of Illinois at Urbana-Champaign

STAT 425 Final Project

Prof. Lelys Bravo de Guenni, PhD

May 10, 2021

Contribution

Weiyang Wang is in charge of the code of Random Forest Models and PCA. He is also responsible for checking the code and report as well as providing advice.

Yulin Song is responsible for writing introduction, the report and code parts for linear regression models and criteria based model selection.

Yuhan Zhou is in charge of writing the report of Random Forest Models, exploratory data analysis, and the final discussion.

Section 1: Introduction

The ultimate goal of this project is to find a relationship between calorie intake composition of 170 countries and their performance in containing the spread of the COVID-19 disease. Young adults born in years around the start of the millennium haven't experienced a global pandemic of such scale, spreading speed, and deadliness. Thus, we are particularly interested in discovering what is causing the different results of different countries around the globe combating the virus.

After careful discussion on choosing the aspect appropriate for association with COVID-19 cases, we decided to focus on the role of diet for the population of different countries. The 170 countries are ranked in alphabetical order. And, from the four different datasets that Professor Bravo has provided, we have chosen a dataset that is comprised of:

- people's consumption of 23 different food or food products and their percentage contribution to the total calorie intake for each country,
- the obesity and undernourished rate as a percentage of total population for each country,
- the percentage of "confirmed", "death", "recovered" and "active" COVID-19 cases as a percentage of total population for each country as of the Week of February 6, 2021 according to the original Kaggle data set (Ren, 2021) that the Professor obtained.

After discussion, we decided that "death" is the most appropriate response variable because the analysis on it would be the most straightforward. Therefore, we want to predict the effects of diet composition and health conditions (obesity rate and undernourished rate), a total of 25 predictors, on COVID-19 death cases, the response variable. We will be fitting several different models, calculating the performance score of each model with test data, and choosing the best model based on the performance score.

Section 2: Exploratory Data Analysis

As a first step, we want to check the summary statistics of predictors and the response.

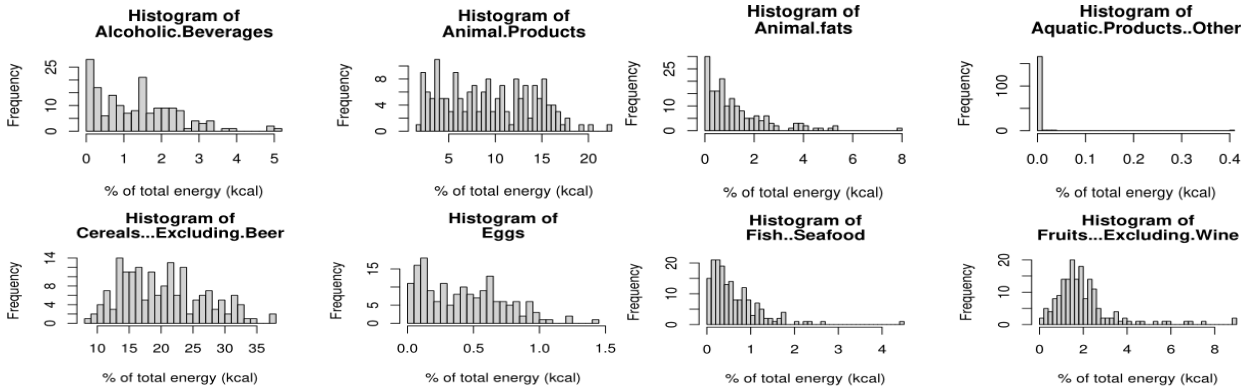
	Alcoholic Beverages	Animal Products	Animal Fats	Aquatic Products, Other	Cereals - Excl. Beer	Eggs	Fish, Seafood	Fruits - Excl. Wine	Meat	Milk - Excl.Butter
Minimum	0.0000	1.624	0.0000	0.000000	8.957	0.0188	0.0000	0.1471	0.298	0.0000
1st Quarter	0.3613	5.083	0.3428	0.000000	15.306	0.1410	0.2402	1.2245	2.081	0.3613
Median	1.2446	9.034	0.8755	0.000000	19.620	0.4037	0.4783	1.6948	3.687	1.2446
Mean	1.3252	9.295	1.2674	0.002786	20.365	0.4285	0.6315	2.0120	3.896	1.3252
3rd Quarter	2.0280	13.175	1.7632	0.000000	24.841	0.6330	0.8697	2.3707	5.278	2.0280
Maximum	5.1574	22.291	7.8007	0.400700	37.526	1.4461	4.4183	8.8540	10.567	5.1574

	Misc.	Offals	Oil Crops	Pulses	Spices	Starchy Roots	Stimulants	Sugar Crops
Minimum	0.00000	0.00000	0.0179	0.0000	0.00000	0.2938	0.00000	0.00000
1st Quarter	0.02470	0.08725	0.2993	0.2967	0.03635	1.1123	0.07765	0.00000
Median	0.08805	0.11825	0.6363	0.7084	0.08590	1.5449	0.20675	0.00000
Mean	0.15933	0.14122	1.1035	1.1089	0.18320	3.0839	0.30537	0.01788
3rd Quarter	0.19173	0.17663	1.1902	1.5472	0.22798	2.9245	0.42080	0.00000
Maximum	1.18220	0.80150	10.4822	7.5638	1.22020	19.6759	2.00900	0.59300

	Sugar & Sweeteners	Tree Nuts	Vegetal Products	Vegetable Oil	Vegetables		Obesity (3missing)	Undernourished (51 missing)		Deaths (%) (6 missing)
Minimum	0.6786	0.00000	27.71	0.9325	0.0957		2.10	2.50		0.000000
1st Quarter	3.4222	0.04662	36.83	3.1263	0.6026		8.50	5.70		0..002013
Median	4.6784	0.17400	40.97	4.6607	1.0031		21.20	9.90		0.011998
Mean	4.8212	0.26162	40.71	4.8724	1.0863		18.71	14.46		0.039370
3rd Quarter	6.3458	0.38958	44.94	6.4279	1.3670		25.70	18.95		0.069503
Maximum	9.5492	1.42100	48.39	10.3839	3.3524		45.6	59.60		0.185428

Figure 2.1 Summary Statistics of the predictors and the response variable

From the summary statistics, we doubt that some of the variables have many zero values. In addition, some have median values relatively different from the mean. We will confirm this by drawing histograms. These are the variables we might want to pay closer attention to.



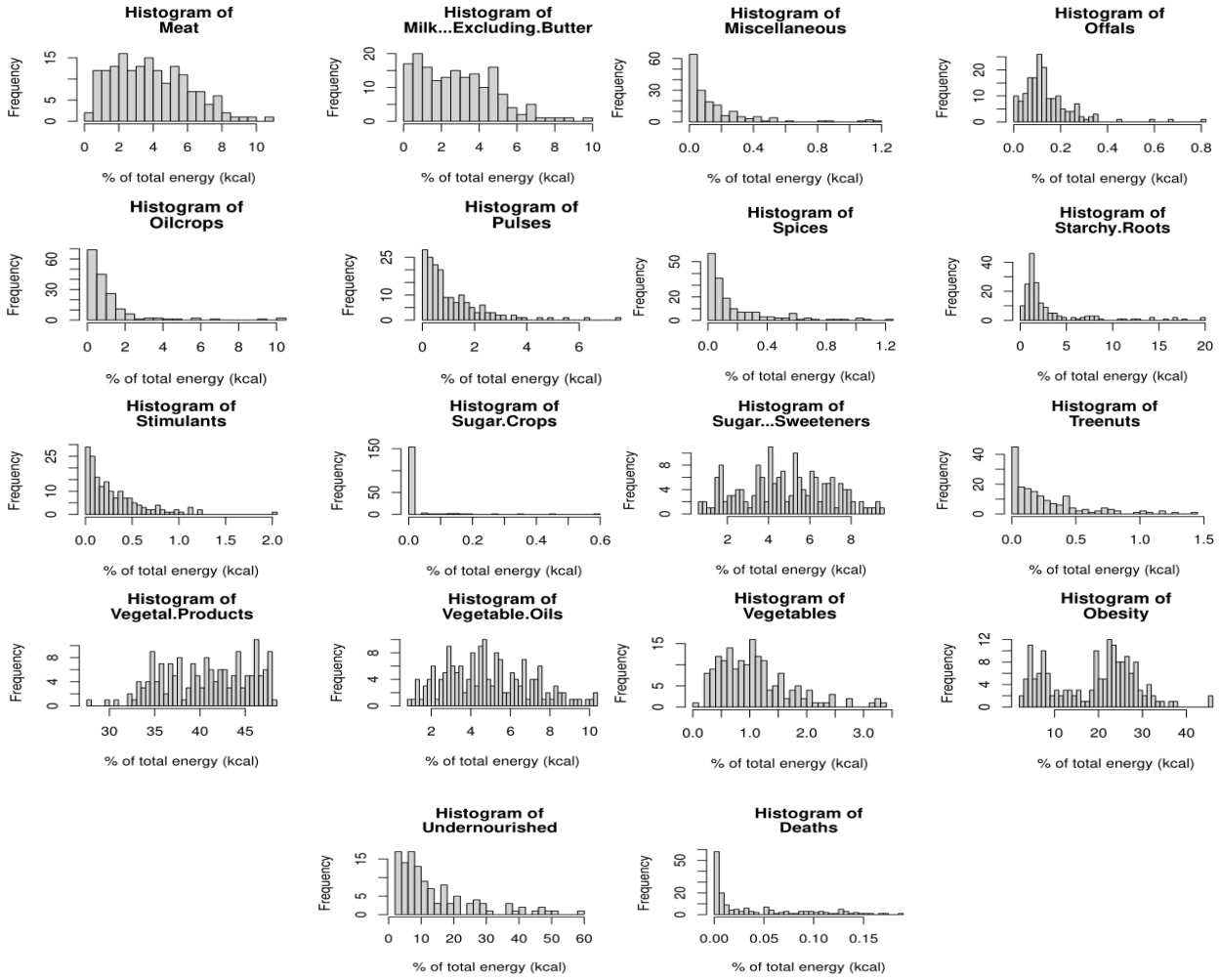


Figure 2.2 Histogram of the 25 predictors and the response variable

We can tell from the histograms that many of the variables do not follow an approximately normal distribution, and variables like "**Sugar Crops**" and "**Aquatic Products**" are mostly 0 percent of the country's total energy intake. We might need to consider transformations of the variables; and many of the extreme cases we might need to consider exclude them when constructing models. To do that, we also need to consider the pairwise correlation between each variable. We will use a correlation plot to present the correlations.

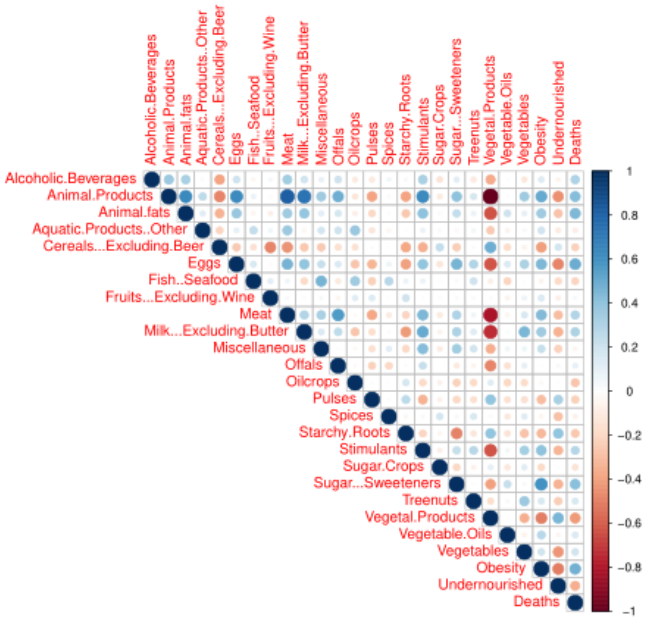


Figure 2.3: Plot of Correlation matrix

In the correlation plot, larger the positive correlation of two variables, larger the dot in each cell, and darker the blue color. Likewise, larger the negative correlation of two variables, larger the dots in each cell, and darker the red color. We can see that many variables have a large negative correlation with "**Vegetal Products**" and many variables have a large positive correlation with "**Animal Products**". We also need to pay closer attention to other pairs of variables having high absolute correlation values; we might need to consider adding an interaction term between the pairs or avoiding including both at the same time in our models. And other correlations are close to zero. Therefore, when we are trying to calculate the total variance of this dataset, we might consider using some dimensional reduction method.

Section 3: Methodology

Section 3.1: Linear Regression

To start, we will attempt fitting a linear regression with all the predictors. However, after inspecting the data frame for the 25 predictors and the response variable, we found out that out of 170 countries, 51 of them contain missing values for the predictor "Undernourished" which is way more than the amount of missing values of any other predictors. Since we believe it's inappropriate to drop so many entire rows (countries), we consider eliminating "Undernourished" first before fitting any models. Following that, only 7 rows contain any missing values and will be disregarded in conducting regression, which is considered to be acceptable. Before fitting models, we separate the 163 rows into two, a training data frame and a testing dataframe, each respectively containing 130 and 33 rows.

For the first model, we consider simply fitting a linear regression model that uses all the remaining 24 predictors. We will call it **"Full Model"**. Some summary results are presented below:

R^2	Adjusted R^2	F	p-value for F
0.5607	0.4603	5.584	2.4e-10

Figure 3.1.1: Summary of Full Model

From summary output of the regression, the 23 predictors (without Obesity) are all individually insignificant at level of 0.05 but significant at level of 0.1; while Obesity isn't even significant at level of 0.1. While Adjusted R^2 looks ok, it still can be improved. Next, we will do diagnostics for the Full Model.

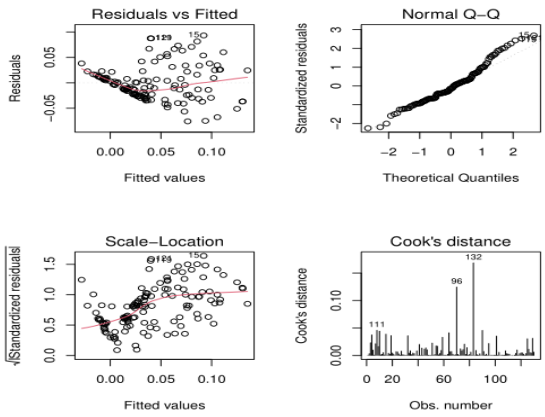


Figure 3.1.2:Diagnostic Plots of Full Model

The errors seem to be independent of each other, since there is no apparent pattern of residuals; they are approximately randomly scattered around the zero residual line. There is no strong vision evidence of a non-linear relationship from the residual plot. However, the constant variance assumption is violated: there is apparent less variance in residuals for low fitted values. And, the Normal Q-Q plot shows deviation from the normal line in the two tails, meaning a violation of the normality assumption. We will attempt to transform the response variable to solve the violation of assumptions.

We discovered that some of the Death values are 0, so we will add a small constant (0.00001) to Death and then take the natural log. We will call it the **"Log Full Model"**. Some summary results are presented below:

R^2	Adjusted R^2	F	p-value for F
0.6047	0.5143	6.692	2.134e-12

Figure 3.1.3:Summary of Log Full Model

We can see that both the R^2 and the Adjusted R^2 has been improved. However, now all the individual significance have been dropped; all predictors now have around 0.2 to 0.3 p-value. But we should check the diagnostics anyways.

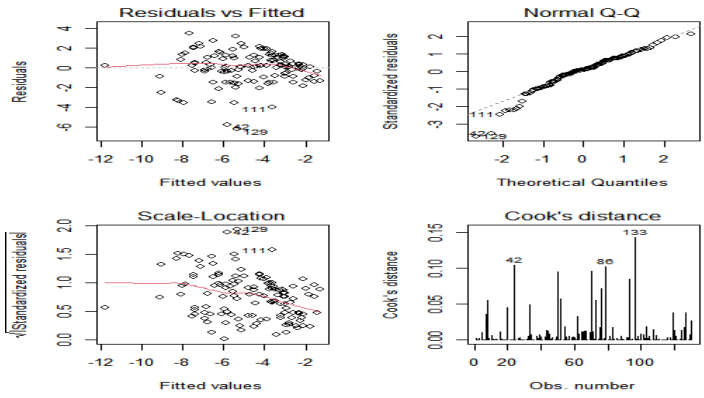


Figure 3.1.4:Diagnostic Plots of Log Full Model

Now it seems that the normality assumption and constant variance assumptions have been met. Although the p-values for each predictor are high, they are very close to each other so we should use this model to perform prediction on the testing data frame and calculate the test MSE (Mean Squared Error). Though, in order for comparison to be meaningful, we need to undo the transformation made in the creation of the Log Full Model. After transformation, we acquired a test MSE of 0.04172. Next, we will be performing model selections based on AIC and BIC.

Section 3.2: Criteria-Based Selection

We will do stepwise model selection in both directions together. We will examine AIC and BIC as our criterion for model selection, and call them "**AIC model**" and "**BIC model**". For AIC, the result is a model with 20 predictors. Some summary statistics are shown below:

R^2	Adjusted R^2	F	p-value for F
0.5869	0.5112	7.744	2.811e-13

Figure 3.2.1:Summary of AIC Model

We will perform stepwise model selection for BIC. Its result is a model with 15 predictors. As usual, some summary statistics are shown below:

R^2	Adjusted R^2	F	p-value for F
0.5237	0.4837	13.09	3.343e-15

Figure 3.2.2:Summary of BIC Model

The R^2 and Adjusted R^2 for the BIC model is slightly lower than all of the above models, but now all of the 15 predictors are individually significant at level of 0.05, while most of them are even significant at level of 0.01. While for the AIC model, most predictors have a p value of around 0.09 while the p value of "Obesity" is about 0.15. So we will proceed further with the BIC model.

We want to do model diagnostics for the BIC model first. We want to check the independence, linearity, constant variance, and normality assumption.

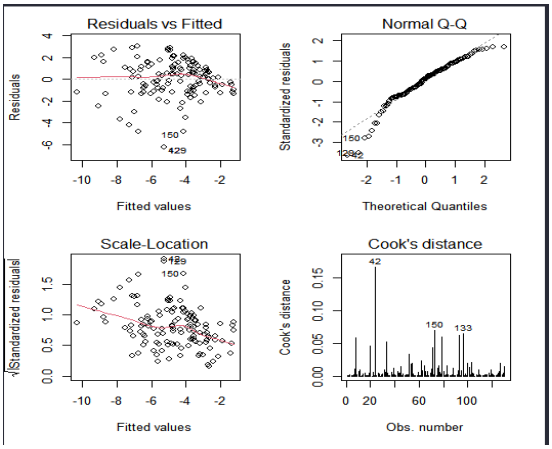


Figure 3.2.3:Diagnostic of BIC Model

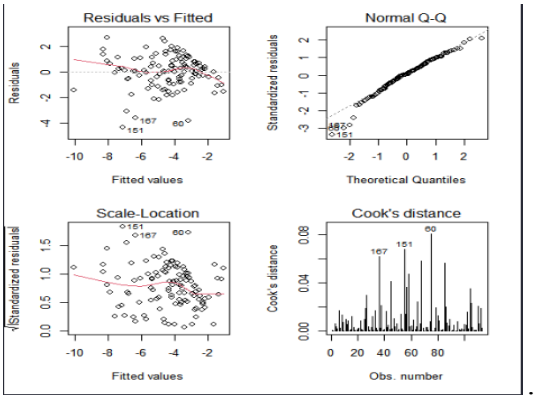
There might be outliers/highly influential points problem. We need to detect these kinds of points before proceeding.

We detected that row 42 and 129 are outliers, 10 rows are of high leverage, and 9 rows are of strong influences. The total amount of unique abnormal values is 18. We then refitted the model removing these and called it "**BIC Model 2**". The summary is presented below:

R^2	Adjusted R^2	F (24 and 105 df)	p-value for F
0.6628	0.6297	20.05	2.2e-16

Figure 3.2.4:Summary of BIC Model 2

We can see that the R^2 and the Adjusted R^2 improved a lot compared to previous models. Most of the variables are significant at level of 0.05. We check the usual assumptions.



	D-W Test for Autocorrelation	S-W Test for Normality	B-P Test for Heteroskedasticity
p-value	0.3696	0.08872	0.06324

Figure 3.2.5:Diagnostics of BIC Model 2

The residuals also seem to be randomly scattered around the zero mean line, which indicates linearity.BIC Model 2 looks like an appropriate model for us to predict on the testing data frame and calculate the test MSE. Since the response variable is in the same format as in the Log Full Model, we should do the same undo transformation. After transformation, we acquired a test MSE of 0.00635, which is smaller than the one calculated before.

Section 3.3: Non-Parametric Method: Random Forest Model

The non-parametric model that we choose to apply is the random forest regression model. Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. In this report, we use the library caret. And to make sure we get the same results every time, we set seed of 1.

First, we use 5-folds cross validation as the training control method to find the best model. Since random forest does not require normality or other assumptions, we can use deaths as response directly. In this model, we take 130 samples and 24 predictors into consideration. Now, we want to see what number of splits can result in the lowest test RMSE. The results of parameter tuning are shown in the table below:

mtry	2	3	5	6	8	9	11	13	14	16	17	19	20	22	24
RMSE	0.03699	0.03701	0.03627	0.03596	0.03640	0.03615	0.03610	0.03643	0.03649	0.03668	0.03671	0.03682	0.03681	0.03664	0.03684

Figure 3.3.1:Table of RMSE in Random Forest Model

In this table, we find that the final value used for the model should be mtry = 6 because it has the lowest RMSE 0.03596. Therefore, the best parameter mtry is 6.

Now, we want to check the variables important plot to find which predictors affect the deaths seriously. And the plot is shown below:

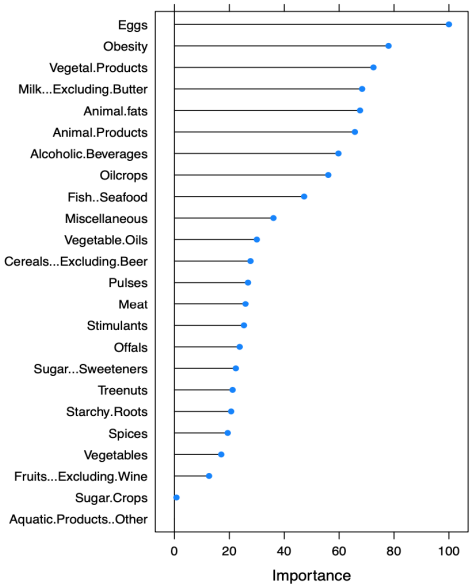


Figure 3.3.2:Diagnostic of Random Forest Model

Since the top ones in the variable important plot are the most important variables; from figure 3.3.2, the top six important variables related to COVID death rate are eggs, obesity, vegetal.products, milk-excluding butter, animal fats, and animal products.

According to the results above, we calculated the MSE of the random forest model, the test MSE is 0.00095, which is much lower than MSEs of those linear models and the random forest model mentioned before.

Section 3.4: Principle Component Analysis

We also took great use of the dimension reductions to help us determine what we should include in our model. The method we applied for is the Principle Component Analysis. In its essence, we want to reduce the components (parameters) to only a few such that much of the data's original variation is preserved. We applied the Principle Component Analysis to the data frame that had "Undernourished" removed and then all NA-containing rows are removed. The approximated cumulative proportion of variance explained when an additional component is added can be summarized in the following chart:

1 Comp.	2 Comps.	3 Comps.	4 Comps.	5 Comps.	6 Comps.
64.18%	79.69%	91.02%	93.62%	95.32%	96.48%

Figure 3.4.1 First Few Principle Components Cumulative Variance Preserved

Usually we want to choose the number of components that are closest to a "great change" of slope but no significant change of slope occurs after it. This can be visualized in a Scree's plot:

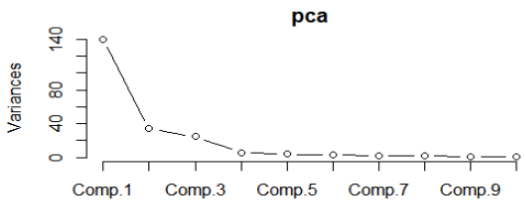


Figure 3.4.2 Scree's Plot for Principle Component Analysis

We might be tempted to choose 2 components, but it's only after the 4th component there isn't much change in slope anymore, so it would be reasonable to choose 4 components. In each

principle component, every predictor has a different weight in the component. In other words, in each component, some sort of linear combination of several of the important "contributors" are included. If a predictor has a higher weight (in absolute value), then it is important in forming that component. We used the first 4 Principle Components to build a regression model that is based on the original data frame, the response "Death", and this "linear combination". As before, we separated this newly created data frame into training and testing dataset, fitted the response onto the four weighted Principle Components, and obtained a linear model with following summary statistics: We will call this model the "**Principle Component Model**".

R^2	Adjusted R^2	F	p-value for F
0.3254	0.3038	15.07	4.404e-10

Figure 3.4.3 Summary of Principle Component Model

The R^2 and Adjusted R^2 are less than before as expected, but we should try predicting too. After calculating the test MSE, we got a value of 0.001167658, which is slightly higher than the one we got in the random forest model, but still is significantly lower than the first two.

Section 4: Discussion and Conclusion

We want to compare the four models chosen, namely, Log Full Model, BIC Model 2, Random Forest Model, and Principle Component Model by looking at their test MSE score. The different MSE are summarized into a chart below:

Types of Model	Test MSE
Log Full Model	0.04172
BIC Model 2	0.00635
Random Forest Model	0.00095
Principle Component Model	0.00117

Figure 4.1 Types of Model and their Test MSE

After selecting a model by a criteria (BIC), the test accuracy has been greatly increased instead of a linear model. Compared to other fitted models that can easily get Adjusted R squared statistics from, the BIC model 2 also has the highest one. This indicates after going through transformation of the response variable, stepwise model selection, and removing the abnormal observations (from the first BIC Model), not only many model assumptions have now been met, but also a higher proportion of variance can be explained by the model.

However, using non-parametric method and principle component method can further improve the accuracy of prediction. We can observe that the test MSE of random forest model and Principal component model vary slightly between each other but are both great improvements to the two models before. Extra advantages of the “most accurate” model, the random forest model, including presenting several most important predictors that have serious impact on the COVID death rate of various countries. Among the top six predictors ranked by seriousness, four of them came from animal products, indicating the correlation of the consumption of these food products to people’s death rate. In addition, Obesity rate is also

associated with COVID death rate. This may encourage the public to maintain a healthier meal composition and lifestyle.

So, based on testing MSE, we would choose the random forest model. However, we acknowledge that some of our limitation of our analysis include that if we could find the training and testing MSE together (to see if it's indeed the same pattern as we observed) or calculate the Adjusted R^2 for every model (to observe the variation in residuals with respect to variation between observed values and their mean), then we might get different results. In addition, we did not consider fitting interaction terms or polynomial terms to fit our model. They might improve our model fitting and model selection, as well. We also could try other dimension reduction methods so that when being applied to real life, better economic decisions can be made.

References

Ren, M. (2021, February 06). COVID-19 Healthy Diet Dataset, Version 66. Retrieved March 18, 2021 from https://www.kaggle.com/rtatman/r-vs-python-the-kitchen-gadget-test.