Impact of COVID 19 on diet

Stat425 Final Project

Hyunjoon Rhee

5/10/2021

Contents

Introduction

December 19th, there was a first confirmed case of COVID-19 from Wuhan, China. Since then highly infectious COVID-19 has transferred the virus all over the world in short time. It has been more than a year since the pandemic has started. People are living a different life, comparing the before and after the pandemic started. One of the major difference that the people are experiencing is the diet. The way people are eating has changed ever since. The restrictions that people were given prevented people from dining outside and started to consume food in home more. Additional to such phenomenon, there must have been more changes toward the dietary patterns of people.

In this project, I am going to analyze and compare the energy consumption level and the active cases of COVID-19, in order to see the relationship between those two.

The data is obtained from Kaggle(2021). The data shows the energy intake level from various products in 170 different countries. The data is composed of 32 columns of different variables and 170 rows. The meaning of each variables are explained in Kaggle, and is summarized below:

- Country(Categorical): Countries around the world
- Alcoholic Beverages(Numeric): Percentage of energy intake (kcal) from alcoholic beverages
- Animal Products(Numeric): Percentage of energy intake (kcal) from animal products
- Animal Fats(Numeric): Percentage of energy intake (kcal) from animal fats
- Aquatic Products, Other(Numeric): Percentage of energy intake (kcal) from aquatic product
- Cereals - Excluding Beer(Numeric): Percentage of energy intake (kcal) from cereal - excluding beer
- Eggs(Numeric): Percentage of energy intake (kcal) from eggs
- Fish, Seafood(Numeric): Percentage of energy intake (kcal) from fish, seafood
- Fruits - Excluding Wine(Numeric): Percentage of energy intake (kcal) from fruits - excluding wine
- Meat(Numeric): Percentage of energy intake (kcal) from meat
- Milk - Excluding Butter(Numeric): Percentage of energy intake (kcal) from milk - excluding butter
- Miscellaneous(Numeric): Percentage of energy intake (kcal) from miscellaneous
- Offals(Numeric): Percentage of energy intake (kcal) from offals
- Oilcrops(Numeric): Percentage of energy intake (kcal) from oilcrops

- Pulses(Numeric): Percentage of energy intake (kcal) from pulses
- Spices(Numeric): Percentage of energy intake (kcal) from spices
- Starchy Roots(Numeric): Percentage of energy intake (kcal) from starchy roots
- Stimulants(Numeric): Percentage of energy intake (kcal) from stimulants
- Sugar Crops(Numeric): Percentage of energy intake (kcal) from sugar crops
- Sugar & Sweeteners(Numeric): Percentage of energy intake (kcal) from sugar and sweeteners
- Treenuts(Numeric): Percentage of energy intake (kcal) from treenuts
- Vegetable Products(Numeric): Percentage of energy intake (kcal) from vegetal products
- Vegetable oils(Numeric): Percentage of energy intake (kcal) from vegetable oils
- Vegetables(Numeric): Percentage of energy intake (kcal) from vegetables
- Obesity(Numeric): Obesity rate (%)
- Undernourished(Numeric): Undernourished rate (%)
- Confirmed(Numeric): Percentage of confirmed COVID-19 cases
- Deaths(Numeric): Percentage of confirmed COVID-19 cases
- Recovered(Numeric): Percentage of COVID-19 recovered
- Active(Numeric): Percentage of COVID-19 active cases
- Population Count(Numeric): Population count
- Unit(Categorical): Unit for data (%)

As it is shown above, all the variables are numeric except for the two categorical variable(countries, unit). With such vast data, the analysis would be done by mainly comparing each numeric variables with that of active COVID19 cases in order to see the relationship. All the analysis would be done in R.

Exploratory Data Analysis

Data Cleansing The main purpose of the analysis is to see the overall relationship among the diet and number of COVID-19 cases. Therefore, rather than showing the every detail of the variables, there will be an comparison among 'Active' and other variables.

For data cleaning process, process was made to clear out rows that contains empty values and remove Active cases that has negative value in it.

The massive data contains columns that does not have many significant values. For example column 'Aquatic Products, Other' ranges from 0 to 0.4, and includes value other than 0 in less than 5 rows. Compared to 170 rows of data in the other columns, it is relatively unlikely to give any significant meaning to the data. Although proper testing is required, but with such significantly small number of data, it was unnecessary for me to test it out. Columns of 'Sugar Crops', 'Unit' is removed to similar reasons.

Also, since in this project, I am only going to compare the remaining variables with Active cases of COVID19, other columns of 'Confirmed', 'Deaths', 'Recovered' is unnecessary.

Moreover, column 'Undernourished' has a character value of <2.5 which indicates any value that is below 2.5. For convenience, it is going to be assumed as a value of 1. Since this Undernourished column is in class of character, it is necessary to change it to numeric value in order to find a better understanding of the rate.

Analysis of product intakes After the data cleaning process, further analysis is required in order to view the average amount of product that people intake.

In order to look at the overall percentage of products that people intake, bar graph is made. Bar graph in Figure 1 shows that the vegetal products takes most part of the energy that people intake. Therefore, further analysis with vegetal products must be made in order to further analyze. In a following order, top 4 products takes about 75% of the total energy consumption. The top 4 products are Vegetal Products, Cereals excluding beer, animal products, and vegetable oils.

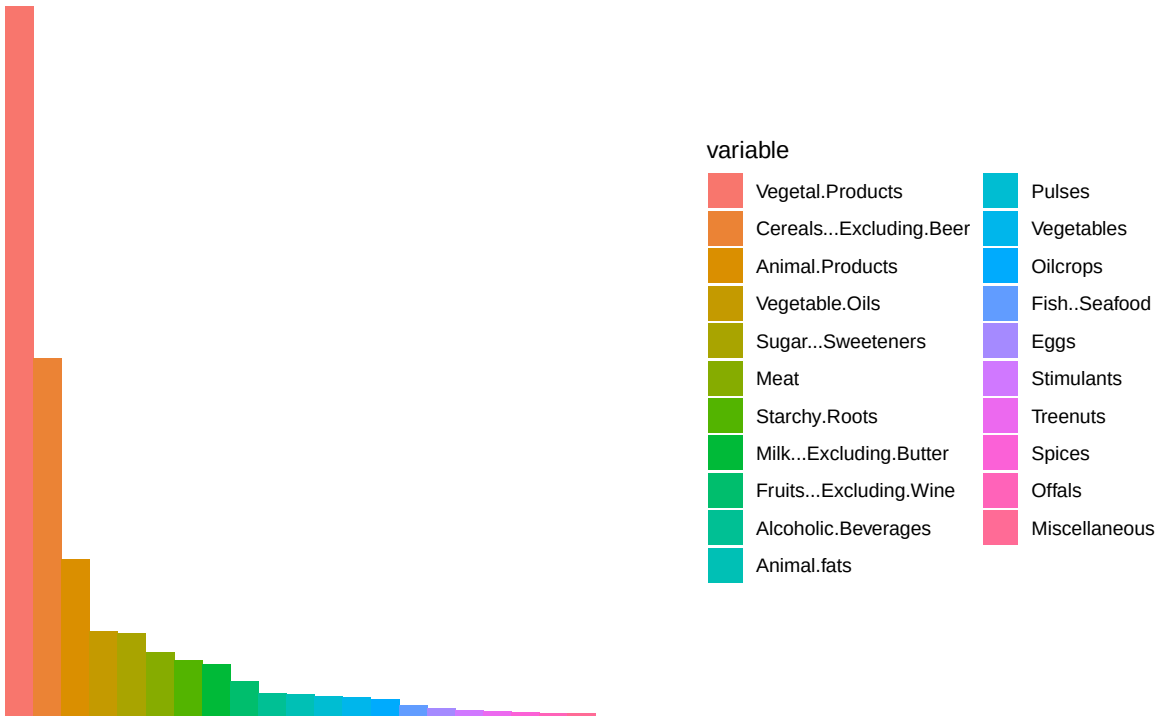

Figure 1: Figure1: Energy intake products

There are products that takes less than 1.2% of the total consumption rate, which includes, 'Pulses', 'Vegetables', 'Oilcrops','Fish..Seafood','Eggs','Stimulants','Treenuts','Spices','Offals','Miscellaneous'. Considering the scope of work that what the major dietary patterns there are related to Active covid cases, it is decided to remove these terms from the data.

Correlation Matrix The correlation plot further guides the method of analysis, which will be covered in later for detailed analysis. Figure 2 provides a correlation plot of the data. Value that is close to 1 means that there are strong linear relationship among the variables. Due to its complexity of the variables, the correlation coefficient that is greater or smaller than 0.7 and -0.7 would be good to take a look at.

List: (Animal Products, animal fats, 0.74), (animal products, meat, 0.85), (animal products, milk excluding butter, 0.81), (animal products, vegetal products, -1), (animal fats, vegetal product, -0.74), (meat, vegetal product, -0.85), (milk excluding butter, vegetal product, -0.81).

The above list shows the combination that has a correlation coefficient that is greater than 0.7. It is noticeable that the animal fats, meat, milk excluding butter are the variables that exist between the correlation with both animal products and vegetal products.

However, the correlation matrix does not show the relationship among individual variables with the rate of active Covid cases. Therefore, further analysis with the linear model is required.

4 mostly consumed products vs Active Covid-19 rate To discover more on the the prior relationship of active COVID-19 rate and top 4 products that are taken into consideration of energy intake, 4 graphs were used.

From Figure 3, it is hard to analyze if there are any linear relationship among Active COVID cases and the products. Just as it is shown in the correlation matrix, further analysis is required.

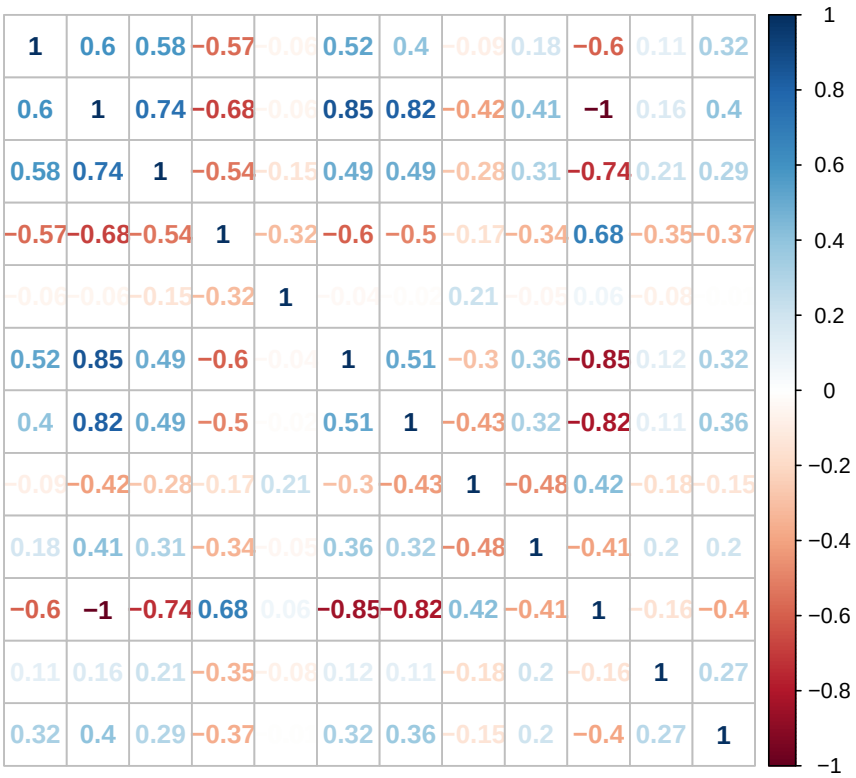

Figure 2: Figure2: Correlation Plot

Method

In order to find out the pattern of foods that people consume after COVID outraged, linear model is used.

3.1 Model selection In order to analyze the model, first the data is split into two parts. First is the training dataset, and the second is the test dataset. The ratio of training and testing is 9 to 1.

From the previous brief analysis of the data, it was hard to find any specific relationship between the active covid cases and other variables. Therefore, in order to find the variables that has significant relationship with the active cases, the step function is used.

After using the step function on both sides, the variables that are left with is Alcoholic Beverages, Milk excluding butter, and vegetable oils.

Check Model Assumption In order to see if the model that the step function has created is meeting the assumption criteria, the assumptions are checked. Constant variance, normality, and independence assumptions are checked.

Constant Variance It is hard to realize from the residual vs fitted graph in Figure 4, but the bptest has a p value of 0.0002, which rejects the null hypothesis that error variances are all equal when the $\alpha = 0.05$.

Check for normality Although it was hard to observe if the data is normal in the qqplot in Figure 4, the shapiro-wilk test had a p value that was small with a value that is close to 0. This rejects the null hypothesis that the variables are normally distributed in some population.

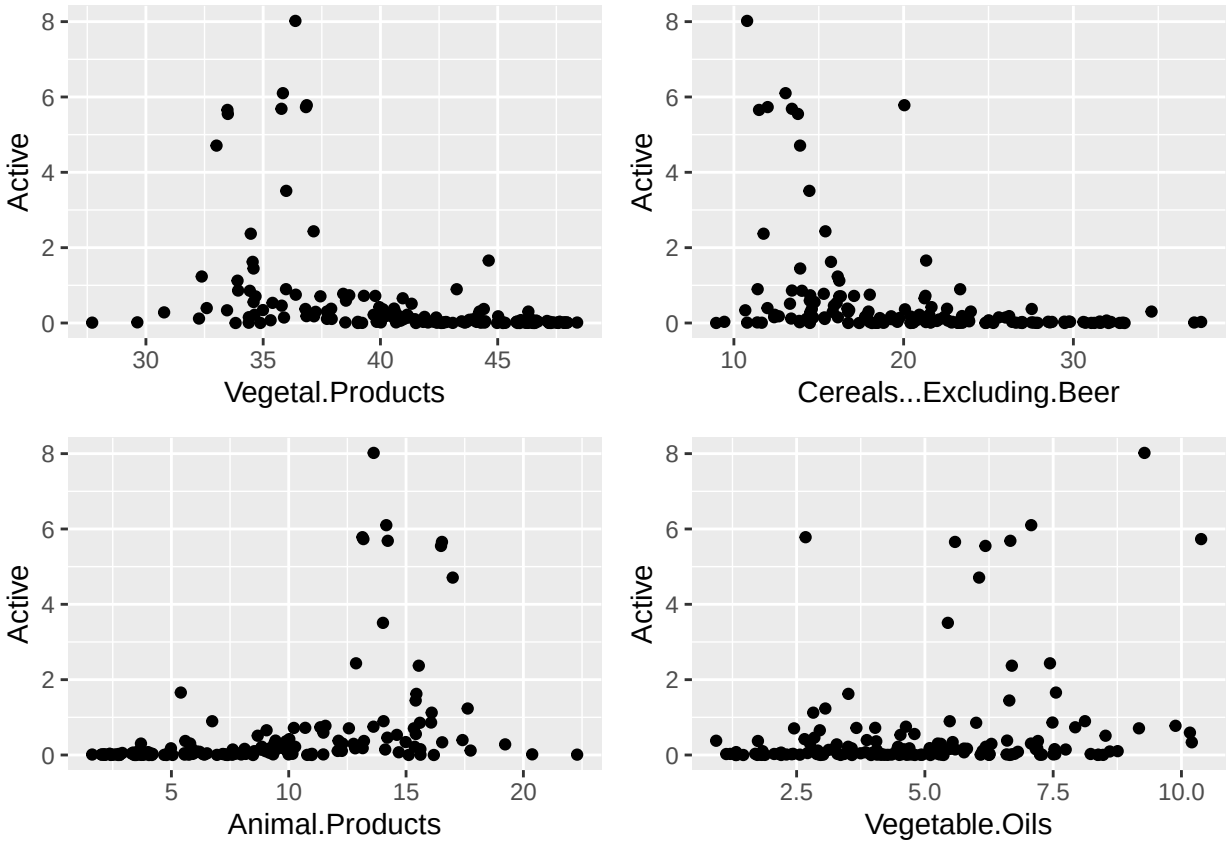

Figure 3: Figure3: Active vs Top 4 Energy intake products

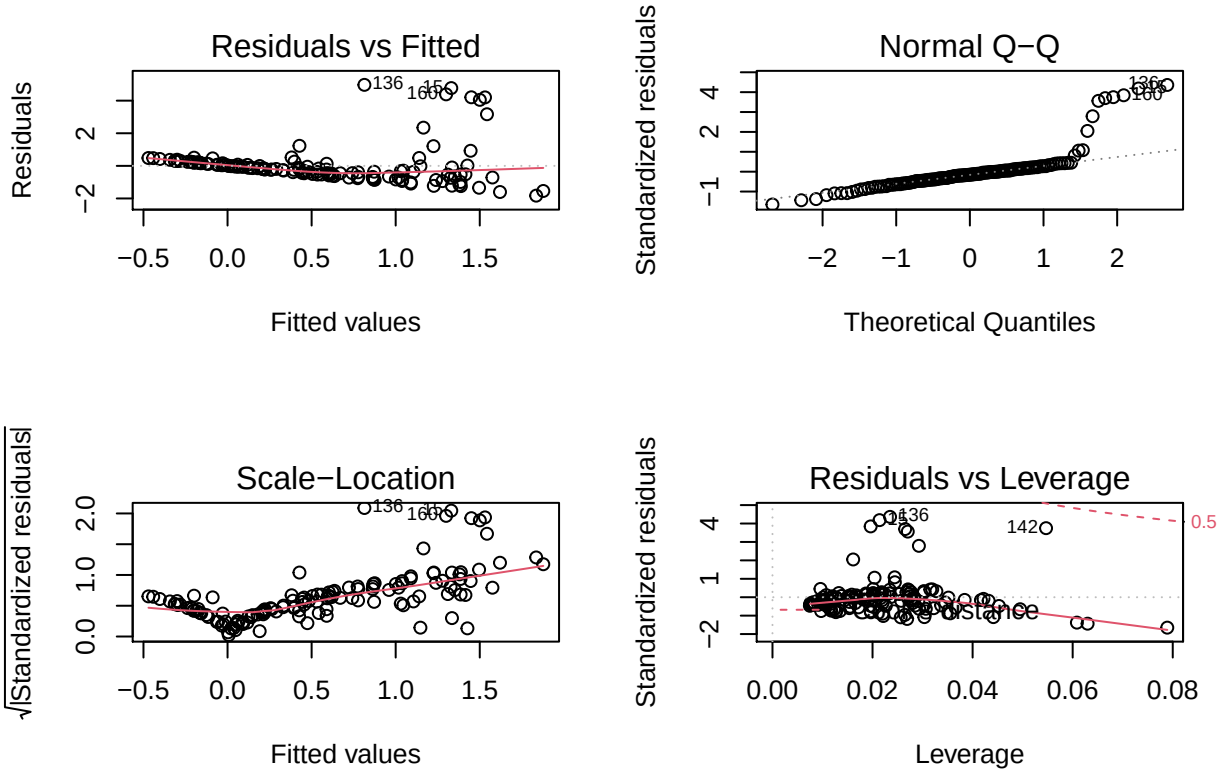

Figure 4: Figure4: Model Assumption Check

Check for independence The dw test shows that there is no autocorrelation since the p value is large with 0.6994, which means that it fails to reject the null hypothesis.

Just by looking at the normality assumption and the homogeneity of the data, it is easy to say that the transformation of a data is needed.

In order to transform the data, box-cox transformation is implemented. The box-cox graph is provided in Figure 5.

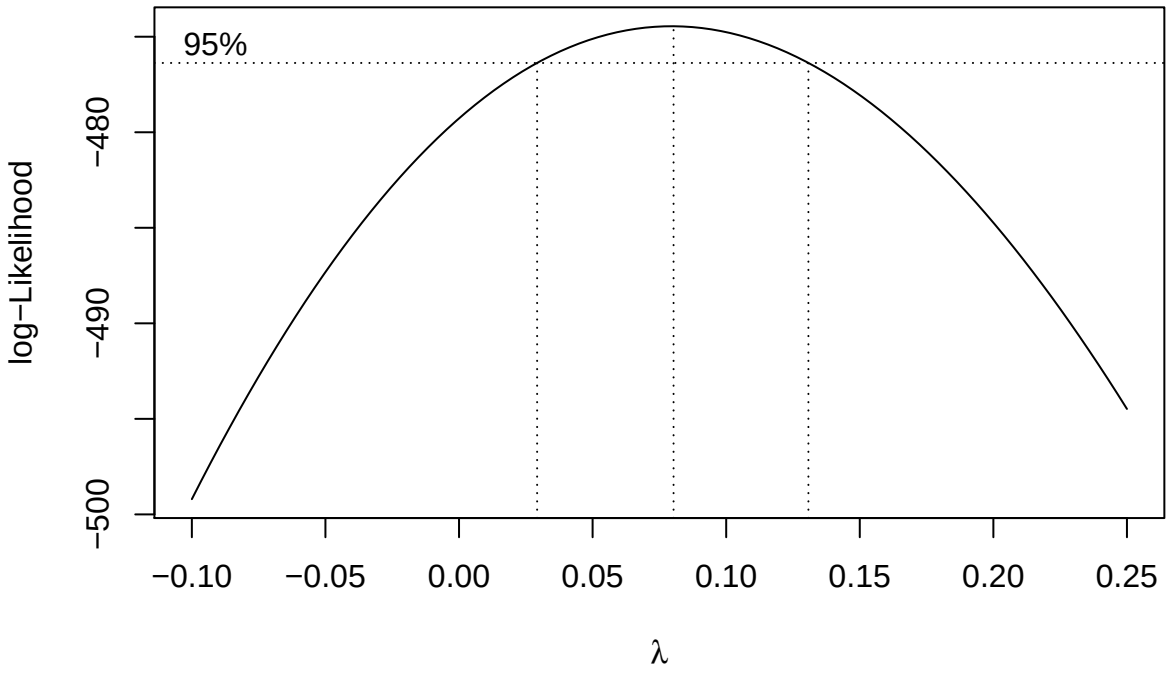

Figure 5: Figure 5: Box-Cox Graph

According to the box cox graph, the lambda is 0.08. This shows that the response variable should be recalculated in a form of $(y^{\lambda} - 1)/\lambda$.

After the recalculation, the model is rearranged and the assumptions are met with BP test Shapiro-Wilk, and Durbin-Watson test showing p value that is greater than $\alpha = 0.01$. BP test, Shapiro test, and DW test each had 0.4913, 0.3455, and 0.4293.

Variable	Value
BP test	0.4913631
Shapiro test	0.3455563
DW test	0.4292971

Check for high-leverage, outlier, influential points As the assumptions are met, effort to improve the model was made. Leverage, outlier, and cook's distance was calculated, and as a result, there were 3 leverage points, with no outlier, and 1 highly influential point.

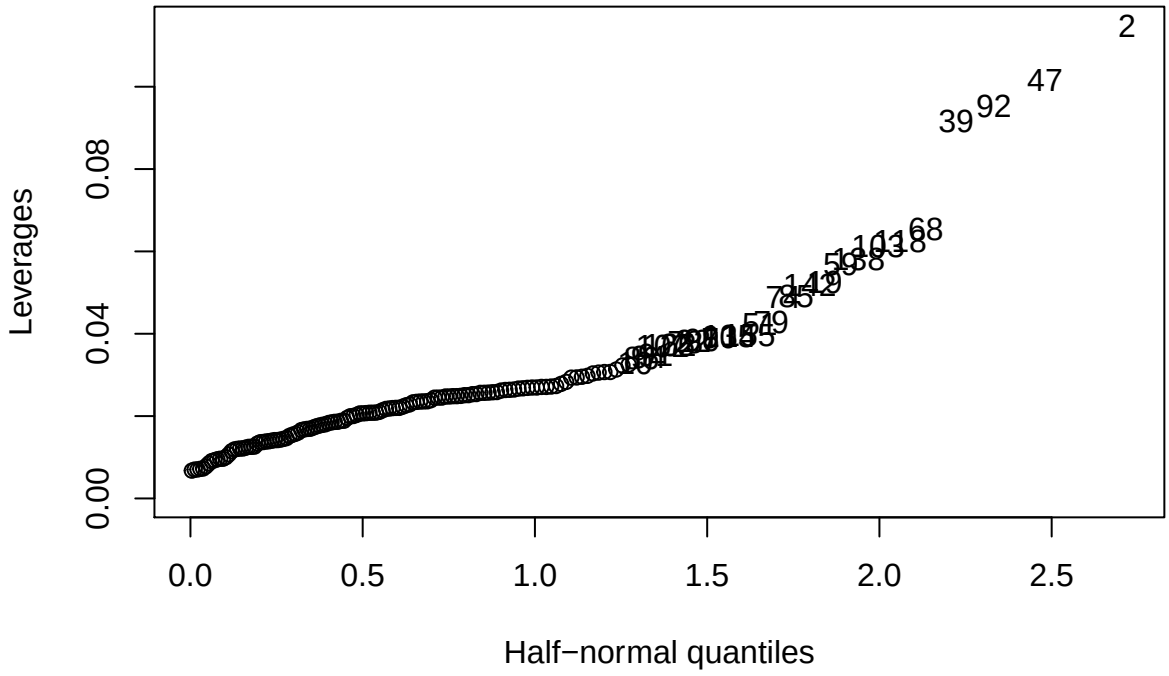

Figure 6: Figure 6: Leverage points

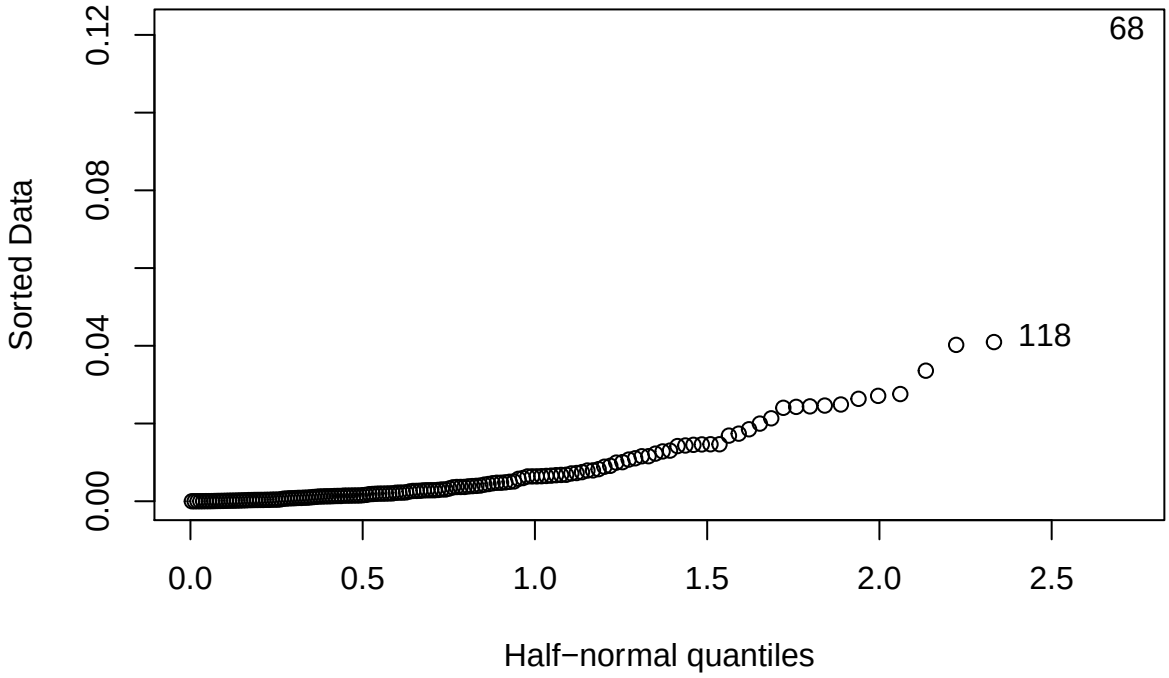

Figure 7: Figure 7: Influential points

Check for highly influential points By getting rid of the leverage points and the influential point, it much better fitted for the model.

After getting rid of the influential points, model was processed again due to change in data set. Therefore, new training dataset and testing dataset were created.

3.2 After the modification of the model, the training and testing process were applied. RMSE with training data and testing data was applied.

Variable	Value
Train RMSE	1.460722
Test RMSE	1.631980

After the calculation, it could be noticed that the test rmse is smaller than train rmse, which shows a good result for the diagnosing the model.

After the analysis, it could be summarized that alcoholic.beverages, milk, and vegetable oils form a linear relationship with Active covid cases. By looking at the coefficients of the model, all of the variables show positive relationship with the active covid cases.

3.3

Variable	Value
Train RMSE	1.170988
Test RMSE	1.479298

As another method to analyze the data, random forest model was used. With having 500 regression trees, each trees were built on random subset of the data, meaning that by every split, there was a random subset of variables. Comparing the RMSE value of training data set and testing data for this model, just like the analysis above, the test rmse has a reduced value. This shows that the testing data set show decrease in error.

Discussion and Conclusion

To summarize the process of analysis, first with a given data, the method of analysis was searched. Because I wanted to find the relationship between active covid cases and the foods, the linear model was applied to the data. By leaving the active covid cases as the response variable, the predictor variables were chosen with careful examination. Step function was used to find out the reliable predictor variables. As a result, Alcohol, Milk, and vegetable oils were used as predictor variables. However, during the process of checking the model assumption, it could be found that all the assumptions were violated. Therefore, box-cox transformation was needed in order to solve the issue. After applying a box-cox transformation, leverage point, outlier, and influential points were checked, in order to improve the model with greater accuracy. After modifying the model, training and testing process were applied to the data and the model. As a result, the training RMSE and testing RMSE were calculated. It could be found that the testing RMSE was smaller than the training RMSE, which shows that it has slightly better fitting model for the testing data.The random forest model showed a similar result, with smaller testing RMSE. Looking at the values of the coefficient for the model, it shows that alcohol, milk, and vegetable oils have positive linear relationship with active covid cases. It is hard to find the reason behind the positive linear relationship, however, with statistical analysis, it was able to find that places with large active covid cases have large number of alcohol, milk, and vegetable oils consumption.