Diet's Effect on Covid-19 Confirmed Cases

Team Coreanos

Spring 2021, Stat 425

May, 11, 2021

Minwoo Sung: modeling

Yeonchul Choi: data handling & cleaning

Seunggyun Shin: Final report & analysis

Introduction

This project focuses on the analysis of effects of food diet on Covid-19 death rate. In this data set, there is different types of food, world population obesity, undernourished rate, Covid-19 statistics. This projects will inspect how a healthy eating style or specific diet pattern can help to defeat Covid-19 virus. The food data is provided in kcal showing the energy intake. Percentage of confirmed Rate of Covid-19 virus will be predicted using these features to find if the specific diet pattern can decrease or increase the confirmed rate. This food data is obtained from Food and Agriculture Organization of the United Nations FAO website. Data for population count for each country came from Population Reference Bureau PRB website. Covid-19 related data are obtained from Johns Hopkins Center for Systems Science and ENgineering CSSE website.

- *Alcoholic Beverages*: Alcohol, Non-Food; Beer; Beverages, Alcoholic; Beverages, Fermented; Wine
- *Animal Fats*: Butter, Ghee; Cream; Fats, Animals, Raw; Fish, Body Oil; Fish, Liver Oil
- *Animal Products*: Aquatic Animals, Others; Aquatic Plants; Bovine Meat; Butter, Ghee; Cephalopods; Cream; Crustaceans; Demersal Fish; Eggs; Fats, Animals, Raw; Fish, Body Oil; Fish, Liver Oil; Freshwater Fish; Marine Fish, Other; Meat, Aquatic Mammals; Meat, Other; Milk - Excluding Butter; Molluscs, Other; Mutton & Goat Meat; Offals, Edible; Pelagic Fish; Pigmeat; Poultry Meat
- *Aquatic Products, Other*: Aquatic Animals, Others; Aquatic Plants; Meat, Aquatic Mammals
- *Cereals - Excluding Beer*: Barley and products; Cereals, Other; Maize and products; Millet and products; Oats; Rice (Milled Equivalent); Rye and products; Sorghum and products; Wheat and products
- *Eggs*: Eggs
- *Fish, Seafood*: Cephalopods; Crustaceans; Demersal Fish; Freshwater Fish; Marine Fish, Other; Molluscs, Other; Pelagic Fish
- *Fruits* - Excluding Wine: Apples and products; Bananas; Citrus, Other; Dates; Fruits, Other; Grapefruit and products; Grapes and products (excl wine); Lemons, Limes and products; Oranges, Mandarines; Pineapples and products; Plantains
- *Meat*: Bovine Meat; Meat, Other; Mutton & Goat Meat; Pigmeat; Poultry Meat
- *Milk*: Excluding Butter Milk - Excluding Butter
- *Miscellaneous*: Infant food; Miscellaneous
- *edibl*: Offals, Edible
- *Oilcrops*: Coconuts - Incl Copra; Cottonseed; Groundnuts (Shelled Eq); Oilcrops, Other; Olives (including preserved); Palm kernels; Rape and - Mustardseed; Sesame seed; Soyabeans; Sunflower seed
- *Pulses*: Beans; Peas; Pulses, Other and products
- *Spices*: Cloves; Pepper; Pimento; Spices, Other
- *Starchy Roots*: Cassava and products; Potatoes and products; Roots, Other; Sweet potatoes; Yams
- *Stimulants*: Cocoa Beans and products; Coffee and products; Tea (including mate)
- *Sugar & Sweeteners*: Honey; Sugar (Raw Equivalent); Sugar non-centrifugal; Sweeteners, Other
- *Sugar Crops*: Sugar beet; Sugar cane
- *Treenuts*: Nuts and products
- *Vegetable Oils*: Coconut Oil; Cottonseed Oil; Groundnut Oil; Maize Germ Oil; Oilcrops Oil, Other; Olive Oil; Palm Oil; Palmkernel Oil; Rape and Mustard Oil; Ricebran Oil; Sesameseed Oil; Soyabean Oil; Sunflowerseed Oil
- *Vegetables*: Onions; Tomatoes and products; Vegetables, Other
- *Vegetal Products*: Alcohol, Non-Food; Apples and products; Bananas; Barley and products; Beans; Beer; Beverages, Alcoholic; Beverages, Fermented; Cassava and products; Cereals, Other; Citrus, Other; Cloves; Cocoa Beans and products; Coconut Oil; Coconuts - Incl Copra; Coffee and products; Cottonseed; Cottonseed Oil; Dates; Fruits, Other; Grapefruit and products; Grapes and products (excl wine); Groundnut Oil; Groundnuts (Shelled Eq); Honey; Infant food; Lemons, Limes and products; Maize and products; Maize Germ Oil; Millet and products; Miscellaneous; Nuts and products; Oats; Oilcrops Oil, Other; Oilcrops, Other; Olive Oil; Olives (including preserved); Onions; Oranges, Mandarines; Palm kernels; Palm Oil; Palmkernel Oil; Peas; Pepper; Pimento; Pineapples and products; Plantains; Potatoes and products; Pulses, Other and products; Rape and Mustard Oil; Rape and Mustardseed; Rice (Milled Equivalent); Ricebran Oil; Roots, Other; Rye and products; Sesame

seed; Sesameseed Oil; Sorghum and products; Soyabean Oil; Soyabeans; Spices, Other; Sugar (Raw Equivalent); Sugar beet; Sugar cane; Sugar non-centrifugal; Sunflower seed; Sunflowerseed Oil; Sweet potatoes; Sweeteners, Other; Tea (including mate); Tomatoes and products; Vegetables, Other; Wheat and products; Wine; Yams The report will compare four different models of Linear Regression Model, General Linear Regression, Non-parametric Regression etc. In order for the modeling, the dataset will be split into test and train set. All the analysis process will be done in R. There will be total 4 sections in this report including this introduction section. Section 2 will be Exploratory Data Analysis, Section 3 will be Methodology, and Sectoin 4 will be Discussion and Conclusions.

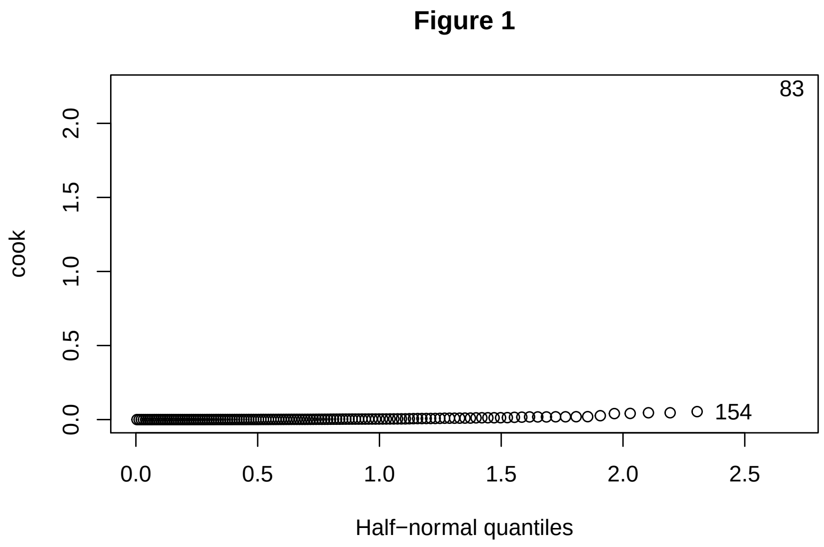
Exploratory Data Analysis

NA removal In order to use 'Death', 'Confirmed' columns for our predict value ,'Recovered','Active','Unit' which are irrelevant columns were dropped. Since this data might contain some missing values that can cause errors in our analysis, removal of missing value was done. In order to impute these missing values from each dataset, the columns with missing values were observed first, and then the observed missing values were imputed with their medians.

Each features were observed through R code and there was a feature that is not numeric. In order to make our model to fit, the non-numeric variables should be modified into a factor and consider it as a categorical variable. Undernourished feature contains character variable which is not numeric compared to the rest of the data.

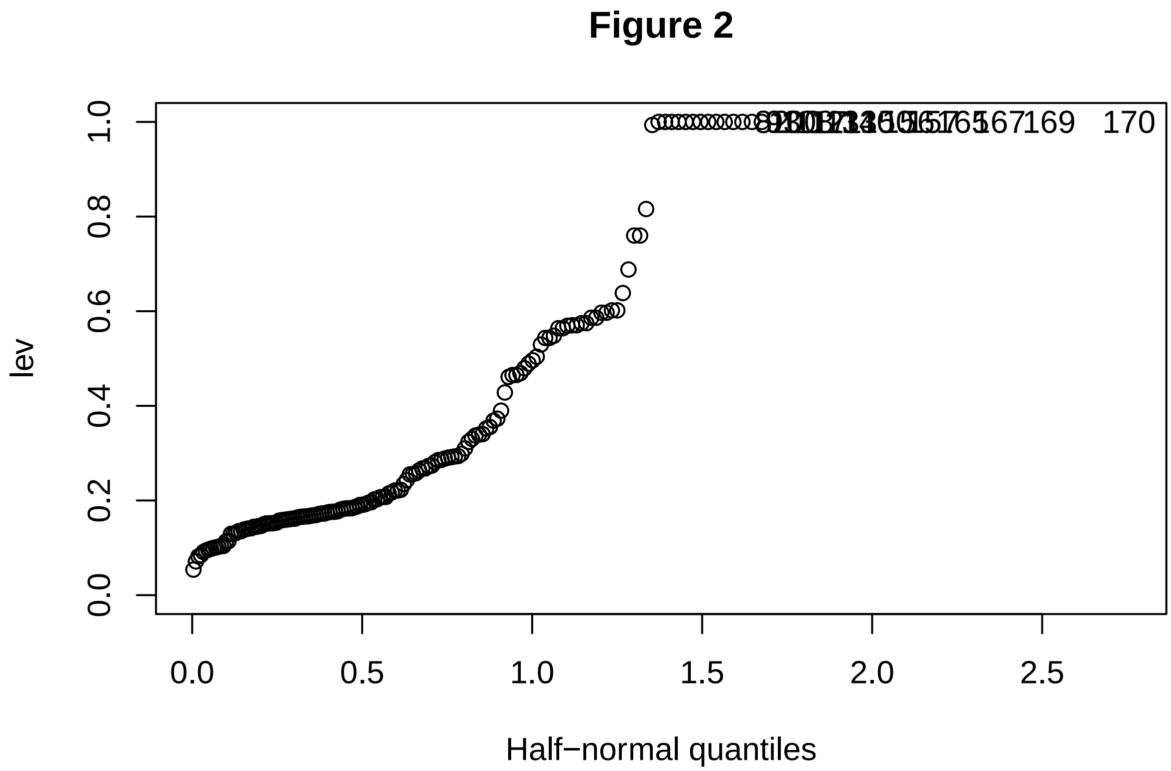
Undernourished had '<2.5' data, which caused it to be not-numeric but a certain rage. Therefore, since we can't find the exact value in that range, the data were split into 5 different portions of ranges, (0,2.5),(2.5,15),(15,30),(30,45),(45,60), and considered as a categorical variable.

Outlier Removal In order to check if there is an outlier or other influential points that can harm our model, several tests were done.



Cook distance was checked first in order to find unusual points. As we see in the **Figure 1** (Cook distance plot), there is the point 83 is either outliers or high-leverage points or both.

And then studentized residuals were checked to do the outlier tests to check if the previous abnormal points are outliers. The Outlier tests by studentized residuals show that those points are not outliers since all the values are lower than 3.711729, which is the critical value we observed using R.



Finally Leverage test was done to see if the points we found are high leverage points and if we should get rid of them. As we see on the **Figure 2**(the Leverage tests), those points are high-leverage points and the graph suggests that the data with index 83 is likely the Bad high leverage point so it is removed.

Plotting Visualizing the features were done to check if we can find some relationship between the input features and output feature.

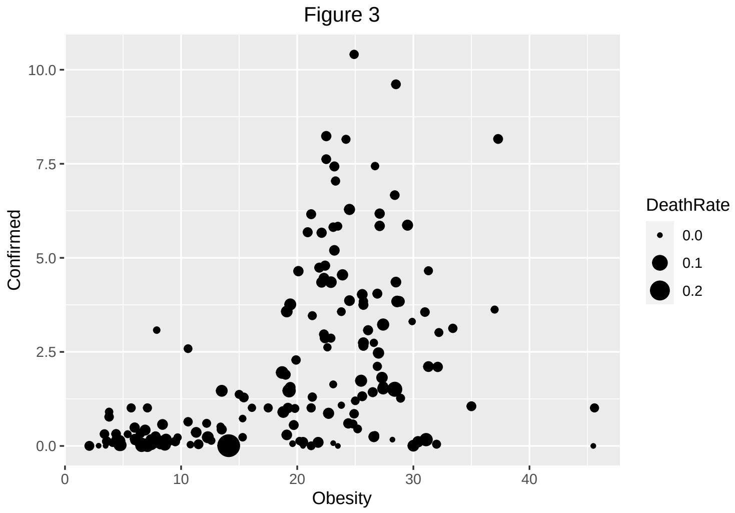


Figure 3 shows as Obesity rate goes up the Probability of Confirmed goes up.

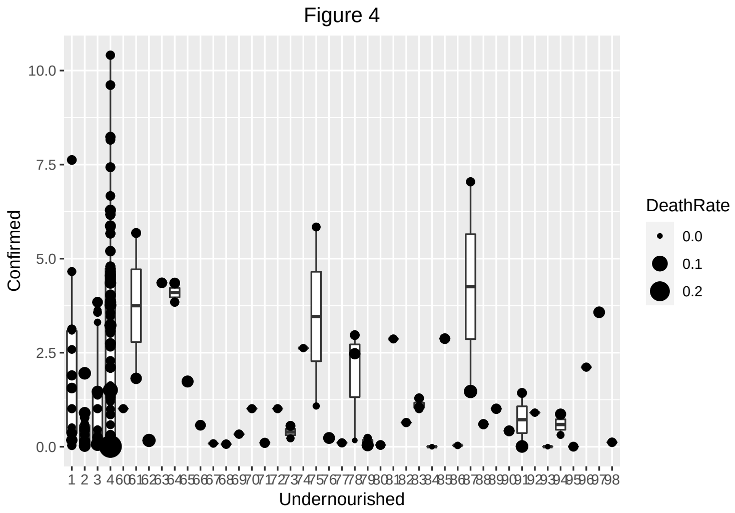


Figure 4 is the box plots show that Undernourished factor is negatively related with Confirmed probability.

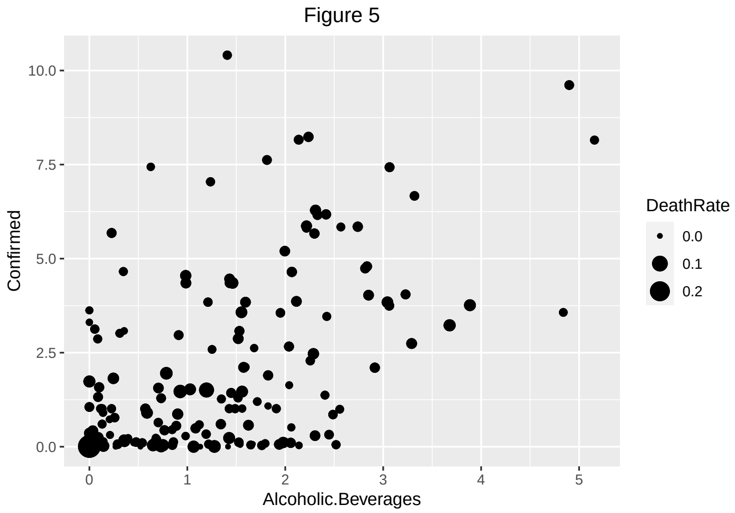


Figure 5 shows that Alcoholic Beverages factor is positively related with Confirmed probability.

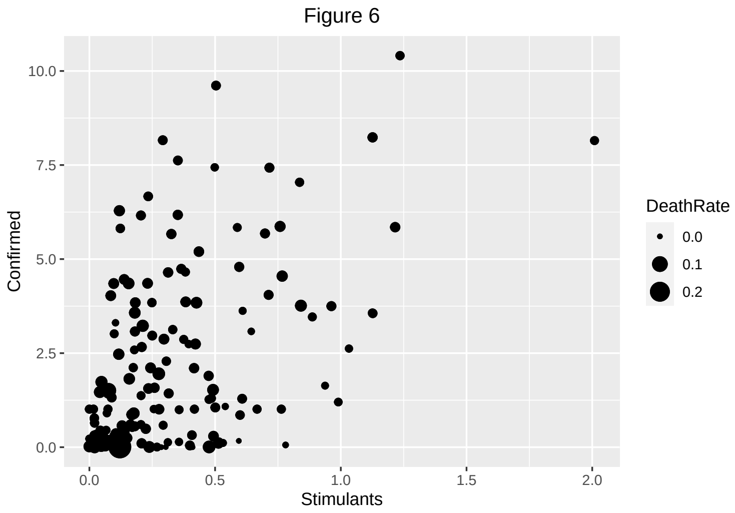


Figure 6 shows that Stimulants factor is positively related with Confirmed probability.

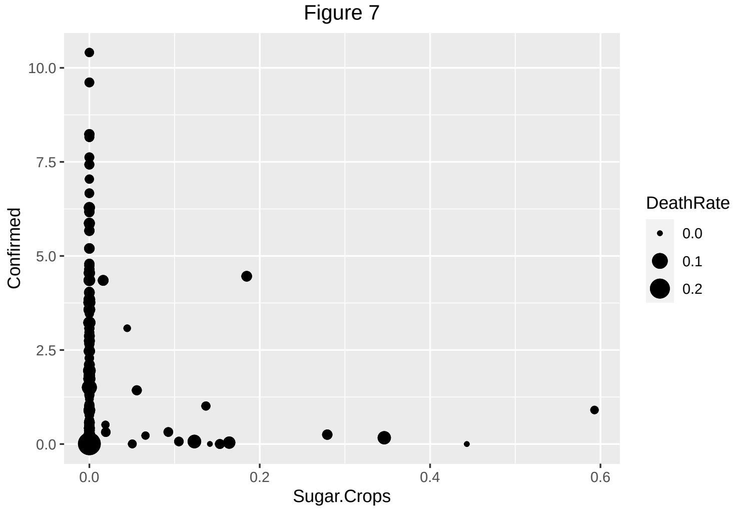
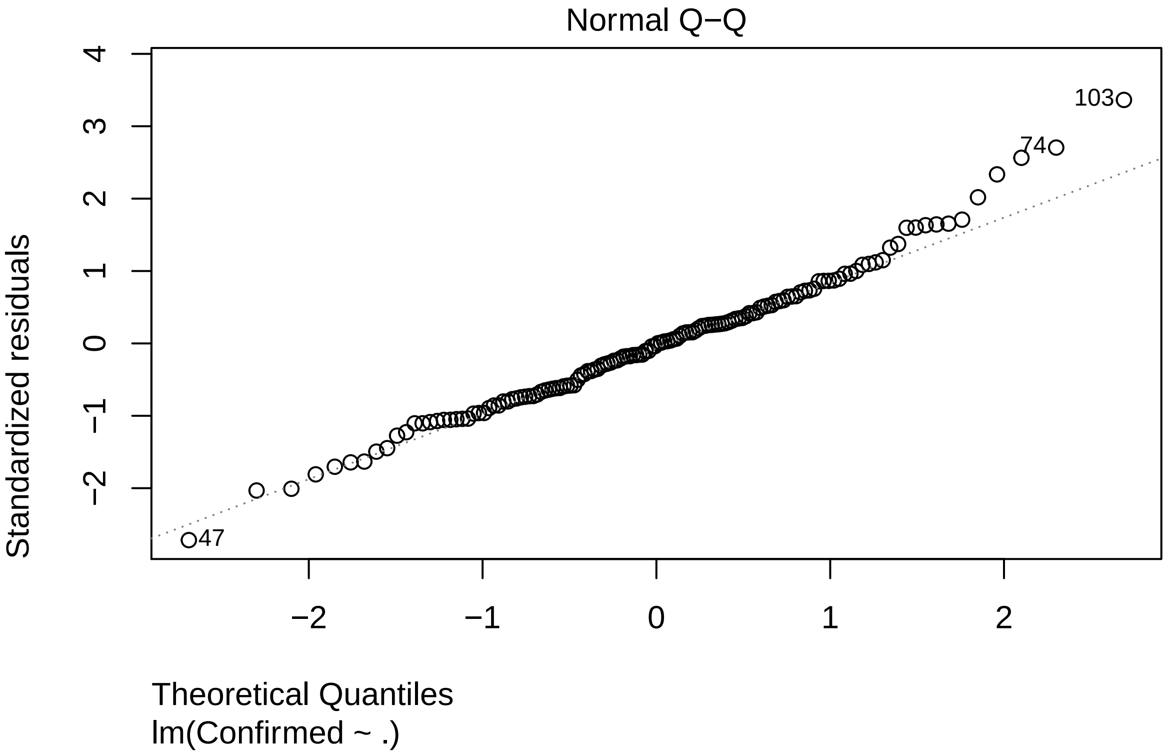


Figure 7 shows that Sugar Crops factor does not have significant relationship with Confirmed probability.

Correlated Features Correlated Features were looked up in order to reduce our features for efficiency. The high correlated value usually have similar impact on predicting so reducing it will reduce the chance of overfitting and increase efficiency.

Highly correlated features were looked up using cutoff of absolute value of correlation of .75. The result was that we should remove column Animal Products and Vegetal Products due to their high correlation to other predictors.

Figure 8

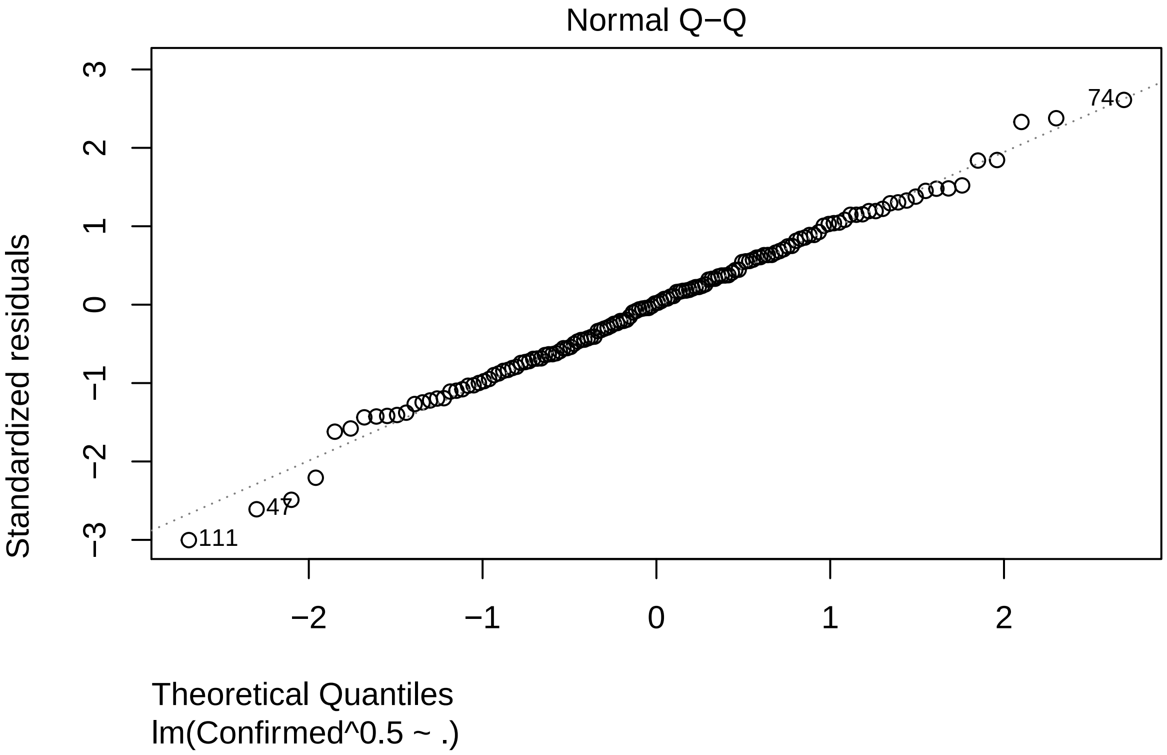


Since p-value is 0.001818, we reject the Null of Shapiro-Wilk normaility test, the normality error assumption doesn't hold yet.

Since p-value is 0.7626, Durbin-Watson test suggests that there is no auto correlation since we don't reject the Null.

So we tried transformation of response variable confirmed. We first tried log transformation, but it still did not pass the test. Then, we tried square root transformation and below is the result.

Figure 9



Since p-value is 0.6009, Shapiro-Wilk normality test suggest that the normality assumption is met. Since p-value is 0.7626, we fail to reject and there is no autocorrelation.

Modeling

First, we tried the liner regression with all feature variables using cross-validation with 10 folds.

Table 1: lm result

lm	metric
Adjusted R-squared	0.5858
F-statistic	9.8010
p-value	0.0000

Average RMSE of linear regression using all feature variables from the cross validation is 1.735572. Since we have 29 features, we decided to reduce the feature by using AIC and BIC.

By using AIC and BIC, we found the new models with reduced features.

AIC: Confirmedˆ0.5 ~ Milk…Excluding.Butter + Obesity + Alcoholic.Beverages + Oilcrops + Treenuts + Offals + Eggs + Miscellaneous + Stimulants,

BIC: Confirmedˆ0.5 ~ Milk…Excluding.Butter + Obesity + Alcoholic.Beverages + Oilcrops + Treenuts + Offals + Eggs

Table 2: aic result

lm	metric
Adjusted R-squared	0.5995
F-statistic	28.9400
p-value	0.0000

We used cross validation with our new models with reduced features which are using AIC and BIC.

Average RMSE of linear regression using reduced features by AIC from the cross validation is 1.653176.

Table 3: bic result

lm	metric
Adjusted R-squared	0.5769
F-statistic	33.7300
p-value	0.0000

Average RMSE of linear regression using reduced features by BIC from the cross validation is 1.666308.

We tried non-parametric regression which is Lasso regression to see whether fits better. We used the same cross validation in Lasso regression.

Average RMSE of Lasso regression from the cross validation is 1.720571.

Discussion of result and Conclusion

Result Discussion

For the prediction of COVID-19 effect, we have compared four models: Linear regression model with all the predictors, AIC model, BIC model and Lasso Regression model using RMSE as a comparison parameter. Depending on our testing result(table 4), we found out that AIC model has the lowest RMSE,

Table 4: rmse result

model	rmse
lm	1.735572
AIC	1.653176
BIC	1.666308
Lasso	1.720571

Therefore, the AIC selected featured linear regression is the best in the prediction of COVID-19 effect.

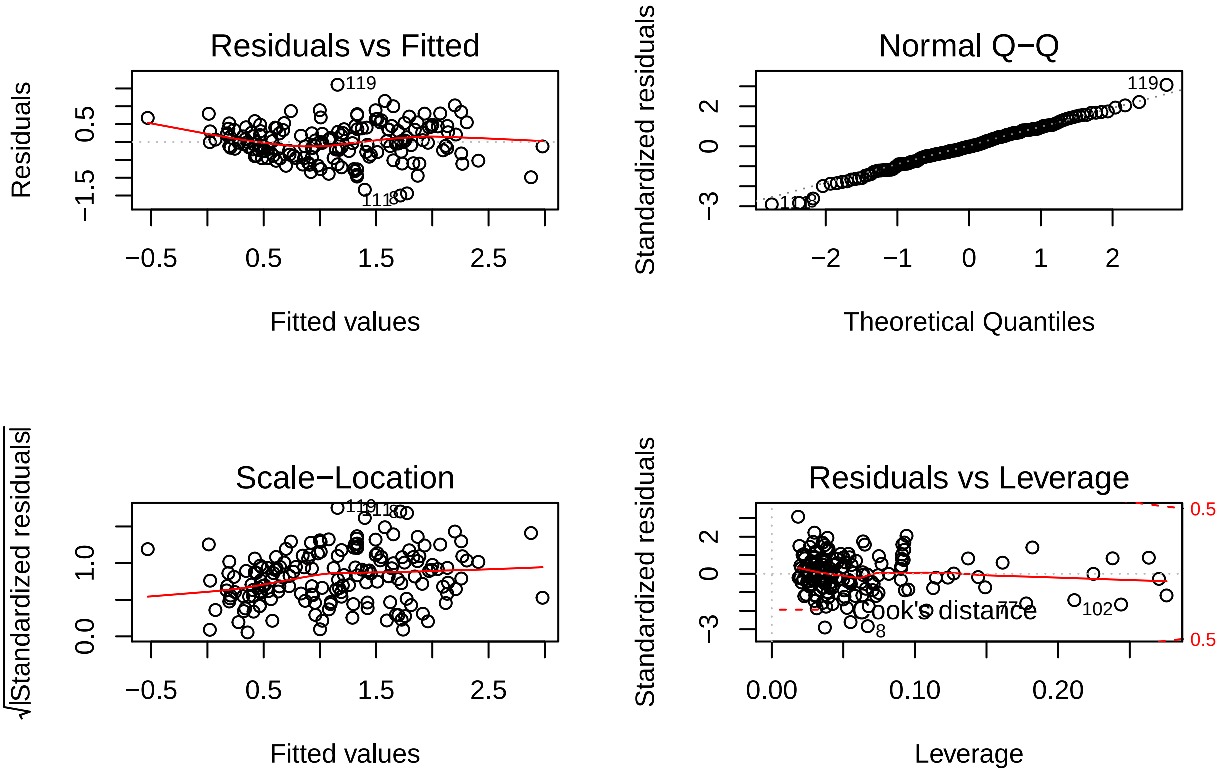
- AIC model

$$\sqrt{Confirmed} =$$

$$Milk...Excluding.Butter + Obesity + Alcoholic.Beverages + Oilcrops + Treenuts + Offals + Eggs + Miscellaneous + Stimulants$$

This means that features: Milk…Excluding.Butter, Obesity, Alcoholic.Beverages, Oilcrops, Treenuts, Offals, Eggs, Miscellaneous, Stimulants have significant effects on the COVID-19 confirmed rate, and we can use the variables above to predict COVID-19 confirmed rate.

Plots & Tests



P-value of Durbin-Watson Test : 0.6461

P-value of Shapiro-Wilk normality test: 0.868

As we cannot observe violation of assumptions from plots and tests, We conclude our final model as an AIC model.

Conclusion

In addition to selecting our final model to predict the response “Confirmed rate of Covid - 19”, through summary of the model, we can find p-values of each predictor, which means how significant each predictor is.

Table 5: P-values of Features

Predictors	P.value
Milk(butter X)	0.00636
Obesity	9.34e-06
Alcohol	0.00147
Oilcrops	0.00857
Treenuts	0.03399
Offals	0.01209
Eggs	0.02316
Miscellaneous	0.00413
Stimulants	0.01327

According to Table 5, p-values of each predictor, there are some factors with higher relationship with the response compare to others. Milk(butter X), Obesity, Alcoholic Beverage, Oilcrops and Miscellaneous are those, and among them, Obesity has the especially significant relation with the response.

By focusing on the factors above rather than just conducting vague data collection, it might be possible to save cost and time in research about "What makes people caught COVID-19 easier than others".

Here, we can conclude that appropriately regulating diet related to factors above such as milk, alcohol, oilcrop, treenuts, offals, eggs and stimulants in order to eventually prevent obesity might be helpful in decreasing the possibility to get infected by the virus in current situation of pandemic.