STAT 425 Final Project Progress Report

Group name: YNX

Group members:
Yijun Zhao: Exploratory Data Analysis, Methodology
Nishant Balepur: Introduction, Outlier Detection
Xi Cheng: Interaction & linearity exploration

Section 1: Introduction

The main goal of this project is to accurately model the COVID death case percentage as a combination of various measures relating to food supply quantities. Through a variety of models, selection techniques, cross-validation methodologies, and visualizations, we hope to build an accurate model and gain some insight onto which variables are the most significant in predicting this COVID death case percentage quantity. We have obtained the necessary data from *Food_Supply_Quantity_kg_Data.csv*, which was provided by the STAT 425 course staff. The framework of this database has country as the primary key, followed by a variety of data corresponding to food quantity. The next couple of columns contain obesity and undernourished rates, while the last few columns contain possible predictor variables, such as confirmed, death, recovered, and active COVID-19 cases rates.

This data is quite interesting, which has motivated us to try and answer a few questions with our final model. First off, we want to see if it is even reasonable to try and model the COVID death rate. It seems possible that food quantity could be indicative of a more equipped or less equipped country to handle the virus. Additionally, we want to see that if this model was successfully validated, if some variables were actually more impactful than others. For example, we would want to know if vegetable or animal product consumption would be more significant in the model. Finally, it would be interesting to see if these variables directly predicted the COVID death rate, or if we could hypothesize another intermediate variable that it predicts, like poverty or unemployment rate.

Section 2: Exploratory Data Analysis

2.1 Data Description

In the food supply data set, all the variables are as follows:

```
"Country"                       "Alcoholic.Beverages"
"Animal.Products"               "Animal.fats"
"Aquatic.Products..Other"       "Cereals...Excluding.Beer"
"Eggs"                          "Fish..Seafood"
"Fruits...Excluding.Wine"       "Meat"
"Miscellaneous"                 "Milk...Excluding.Butter"
"Offals"                        "Oilcrops"
"Pulses"                        "Spices"
"Starchy.Roots"                 "Stimulants"
"Sugar.Crops"                   "Sugar...Sweeteners"
"Treenuts"                      "Vegetal.Products"
"Vegetable.Oils"                "Vegetables"
"Obesity"                       "Undernourished"
"Confirmed"                     "Deaths"
"Recovered"                     "Active"
"Population"                    "Unit..all.except.Population."
```

There are 170 observations, with 31 variables excluding the last one that refers to the units of data. We treat `Country` as a categorical variable and all others to be numerical variables.

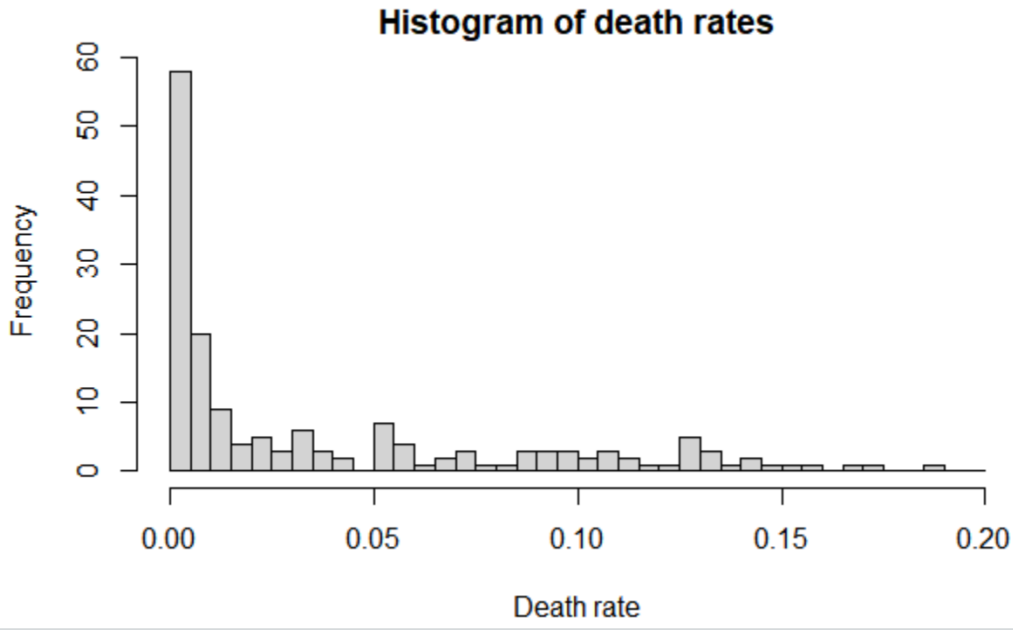
However, we may consider converting variables `Alcoholic Beverages`, `Aquatic product, Other`,`Sugar Crops` and `Sugar & Sweeteners` to be categorical since these columns have more than 90% values which are zeros, and we observe that there is generally no big difference between the rest of nonzero values, so it should be ok to convert and interpret these variables with two level: “Supplied” and “Not Supplied”.

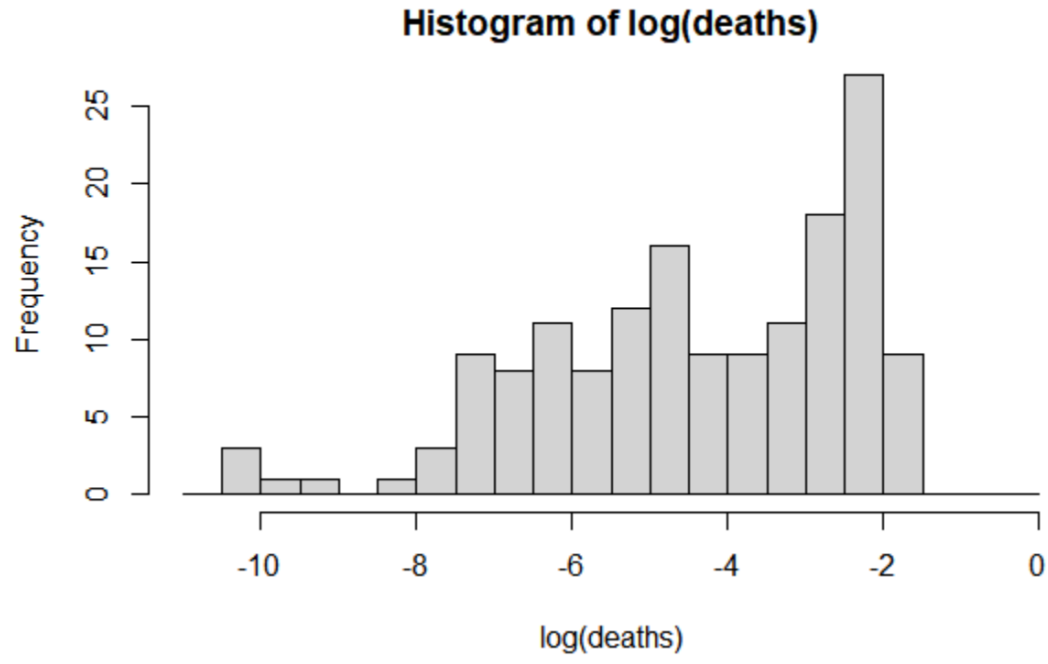
Next step we would consider removing the variable `Country`, since it’s not very useful to explain the response, and hard to interpret with too many levels as a categorical variable.

We would also remove observations with too many missing values, especially when important values, such as the death rate (the response), are missed; then we would replace the rest of NA values with zeros. We also notice that the variable `Undernourished` contains the values `<2.5`, which would be rewritten as 2.5 for type unification and analysis simplification.

2.2 Data Visualization

(1) Histogram

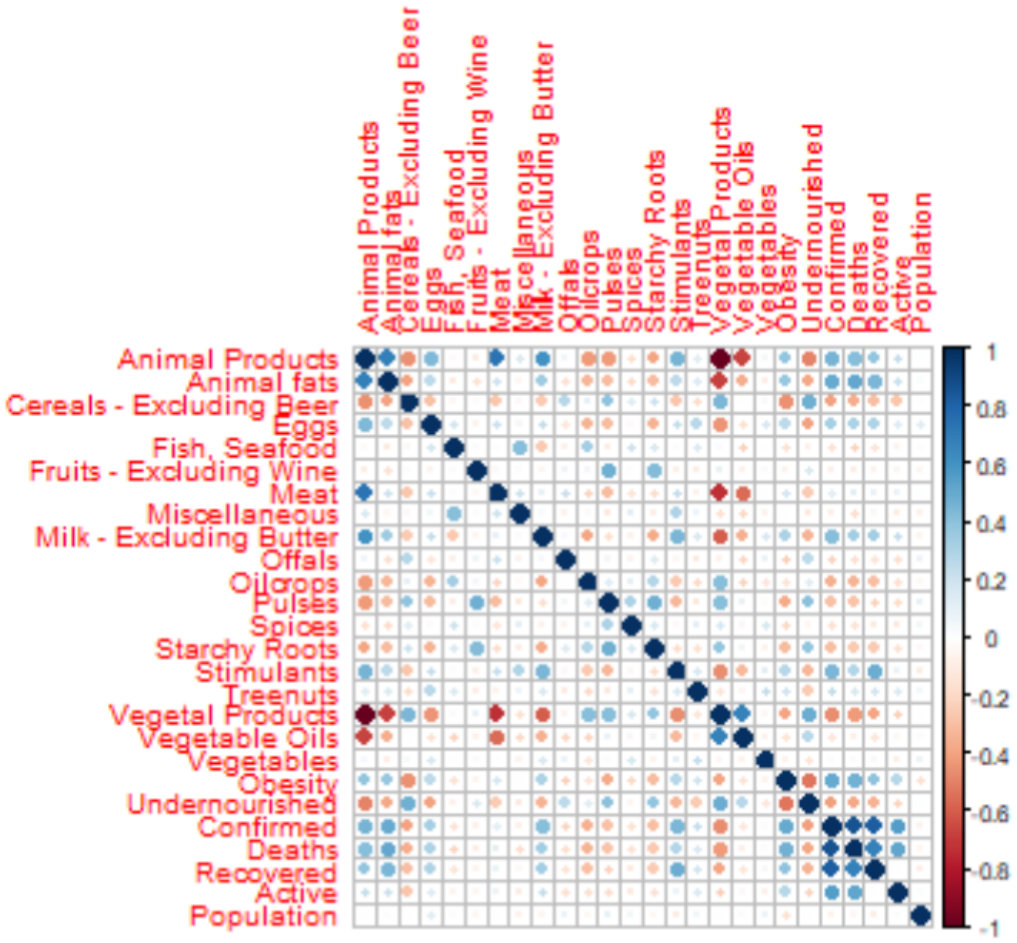




First we notice that our response variable `deaths` has a right skew, and a log transform can not normalize the distribution very well, another transformation will be considered.

(2) Correlation plots

We also check the correlation between the usable numeric predictors, where missing values are removed.



Note that some of the variables are highly correlated, such as the variable `Vegetal Product` and `Animal Product`, which we should pay attention to.

2.3 Outlier Detection

(1) Studentized Residual Test

To check for outliers, we calculated the studentized residuals for each data point fitted against a simple multivariate model, and compared it to the appropriate critical point in the t-distribution. After calculating the studentized residuals and sorting them by decreasing absolute value, our largest value was 2.930235.

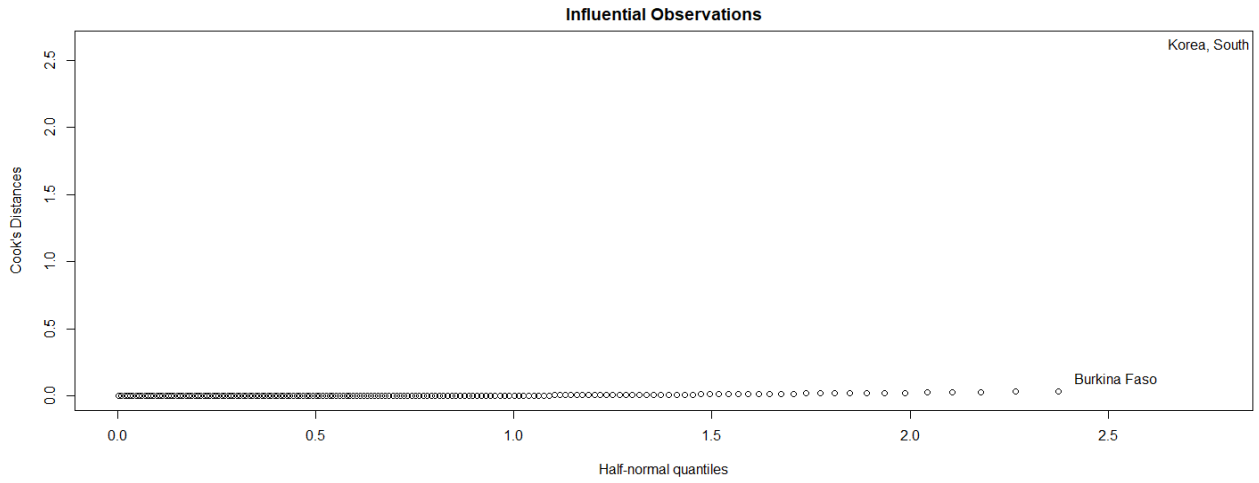
To obtain the critical point, we used a significance level of 0.05 and corrected the value with the Bonferroni Correction. By doing this, we obtained a critical point of 3.711081.

Since the largest studentized residual was still less than the critical point, we concluded that no outliers were present in the data and we could proceed as normal. Of course, this decision is still subject to change as we fit more complex models and do more in-depth analysis and validation.

(2) Cook's Distance

We also wanted to ensure that there were not too many influential points in the dataset. After calculating the Cook's Distance, we found that entry 83, corresponding to South Korea, had a high influence (Cook's Distance greater than 1). Hence, we will need to consider removing this entry when we start building our model.

We also built a half-normal plot to check the Cook's distance in comparison to every other point in the dataset



Clearly, `South Korea` has a much higher Cook's Distance, giving us more confidence that this point needs to be removed.

2.4 Interaction

Test if we should take interaction between categorical variable `Undernourished` and numerical variable `Cereals...Excluding.Beer` into consideration.

Our model to test this interaction has the following formula:

$$\text{deaths} = \beta_o + \beta_1 * \text{Cereals...Excluding.Beer} + \beta_2 * \text{Undernourished} + \beta_3 * \text{Cereals...Excluding.Beer} * \text{Undernourished}$$

We center the numerical predictor, then refit the model with interactions and attempt to test the following hypotheses at the significance level of 0.05:

H_o: The interaction between `Undernourished` and `Cereals...Excluding.Beer` is not significant and should not be included in the model

H_a: The interaction between `Undernourished` and `Cereals...Excluding.Beer` is significant and should be included in the model

We run a simple t-test to see if this coefficient in a linear model is significant, and we obtain a p-value of 0.046879. Since our p-value is less than the significance level, we can reject H_0 and claim that the interaction term is significant. Hence, the interaction between `Undernourished` and `Cereals...Excluding.Beer` should be included in the model

There are many possible combinations of interaction terms that we could consider. We know for a fact that this interaction was significant with `Cereals...Excluding.Beer`, so it's likely that it will be significant with other variables as well. Thus, when we focus more on model building, we will keep this in mind and test a variety of interaction combinations for our model.

2.5 Linearity

Assume we only have one predictor, then we fit the Polynomials Regression for the variable `Eggs` of quadratic and cubic terms as follows:

Linear Model: deaths = $\beta_0 + \beta_1 * \text{Eggs}$
Quadratic Model: deaths = $\beta_o + \beta_1 * \text{Eggs} + \beta_2 * \text{Eggs}^2$
Cubic Model: deaths = $\beta_o + \beta_1 * \text{Eggs} + \beta_2 * \text{Eggs}^2 + \beta_3 * \text{Eggs}^3$

We run the analysis linear model and obtain an overall p-value of 6.832e-06, meaning that the model is significant.

We do the same thing for the quadratic model and find an overall p-value of 2.762e-12. Additionally, both of our coefficients, β_1 and β_2, have respective t-values of 7.529 and -6.057, which both have p-values less than 0.05. It's safe to say that this quadratic model is valid.

Finally, we run the summary for the cubic model. We find an overall p-value of 1.693e-11. The coefficient β_1 has a t-value of 3.376, which is significant at our level of 0.05. However, β_2 and β_3 have t-values of 0.159 and 0.745, which means that both are insignificant. This gives us some doubts about the validity of this model.

By checking the significance of β_2 in the quadratic model, we are essentially performing an ANOVA test to compare the linear and quadratic models. Thus, our results tell us that the quadratic model is preferred, since β_2 was deemed to be significant.

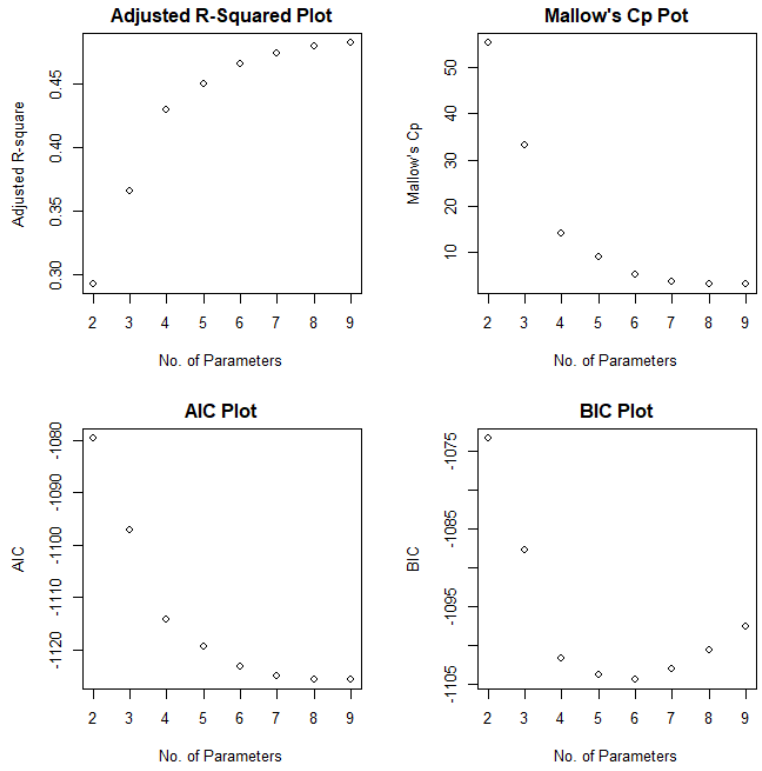
By checking the significance of β_3 in the cubic model, we are essentially performing an ANOVA test to compare the quadratic and cubic models. Thus, our results tell us that the quadratic model is preferred, since β_3 was deemed to not be significant.

In conclusion, after the Analysis of Variance Tests, we can know that the cubic term is not significant but the quadratic term is significant. Then for some numerical variables, like variable `Eggs`, there is evidence to support nonlinear trends. When building our model, we will keep this mind and frequently check for support for nonlinear trends for all of our numeric variables.

2.6 Variable Selection

As we further explore our dataset, we'll also do some basic variable selection techniques to select the optimal number of predictors. To do this, we'll generate a subset of models and select the number of predictors based on different metrics, namely Adjusted R-Square, Mallow's Cp, AIC, and BIC. We will look for a number that will give us low values for Mallow's Cp, AIC, and BIC, but a high value for Adjusted R-Squared. There are two primary searching algorithms recommended from the lecture, i.e., Leap and Bounds as well as Greedy algorithms. While Leap and Bounds can return the global optimal solution, Greedy algorithms can only return a local optimal solution. The former method can be realized with function *regsubsets* from library leaps to evaluate different scores for subsets of models up to size p (including the intercept). On the other hand, another function *step* from the stats library is to apply searching algorithms based on the AIC (default) or BIC criteria. Also, there are three direction choices (i.e., Backward, Forward and Stepwise) in the searching algorithms.

After performing these steps, we are given the following graphical outputs:



We notice that the parameter number 9 generates best values for Adjusted R-Squared, Mallow's Cp, and AIC. Interestingly, 6 parameters generate the best value for BIC. Hence, when we create our models in the next section, we will consider a very high number of parameters, 9 or more, as well as 6 parameters for our model building.

Section 3: Methodology

3.1 Simple Polynomial Model

We began our statistical modeling journey by creating a linear model that used all of the variables. After doing this, we observed from our correlation plot that “Confirmed”, “Active”, “Recovered”, and our response exhibited signs of collinearity, so they were removed from our dataset.

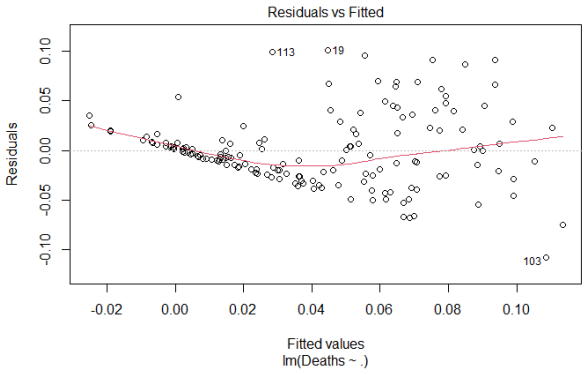
Now that our data has been fully cleaned, we can begin creating our initial model. To begin with, we needed to split our data into a training set and a testing set. We decided on a 90/10 train-test split because we believed that it would be sufficient to evaluate the effectiveness of our model.

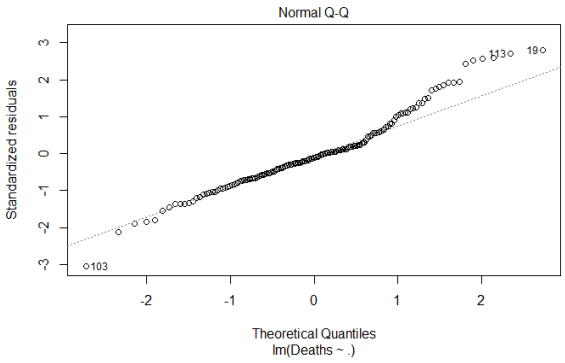
To avoid the problem of including too many parameters and overfitting the model, we looked back at the variable selection methods implemented in the previous section. We analyzed the AIC for our selection criterion and selected the parameters that minimized this value. After doing this, we were left with the following variables: "Animal Products", "Animal fats", "Miscellaneous", "Stimulants", "Vegetal Products", "Vegetable Oils", "Obesity", "Deaths"

We analyzed the summary output and obtained an R squared value of 0.4552 and an adjusted R squared value of 0.4294. The overall p-value for our model is 2.2e-16, which means that it is statistically significant in fitting our data. Now we can check the diagnostics of our model.

We began by checking for outliers. After including the Bonferroni correction, we calculated a cutoff value of -3.69398. However, none of our studentized residuals exceeded this absolute value, so we have no outliers in our data.

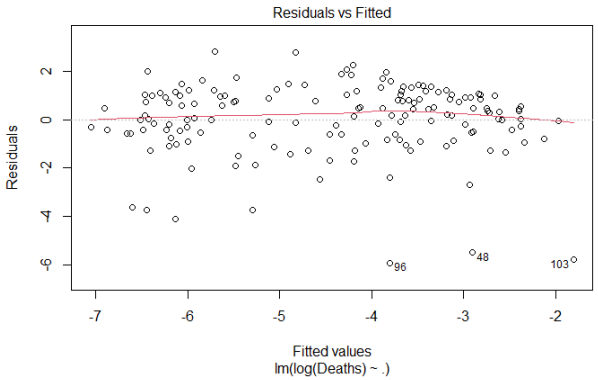
We created residuals vs fitted and Q-Q plots to assess our diagnostics, which are pasted below:

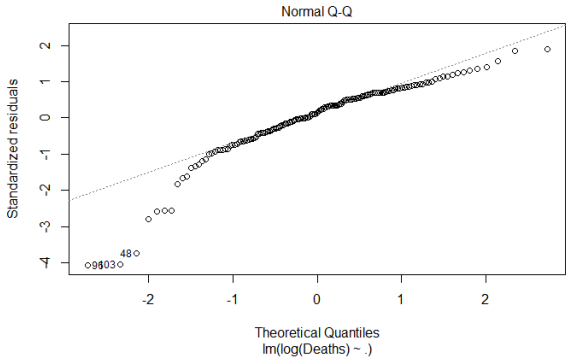




Along with this, we performed a Breusch-Pagan test to assess the constant variance assumption. This resulted in a p-value of 0.0001097, which means we are confident to reject the

null hypothesis and conclude that the assumption is violated. To remedy this, we used a variance stabilizing transformation on the response variable with the logarithm function. After fitting this model, we created the residuals vs fitted and Q-Q plots once more, which are below:



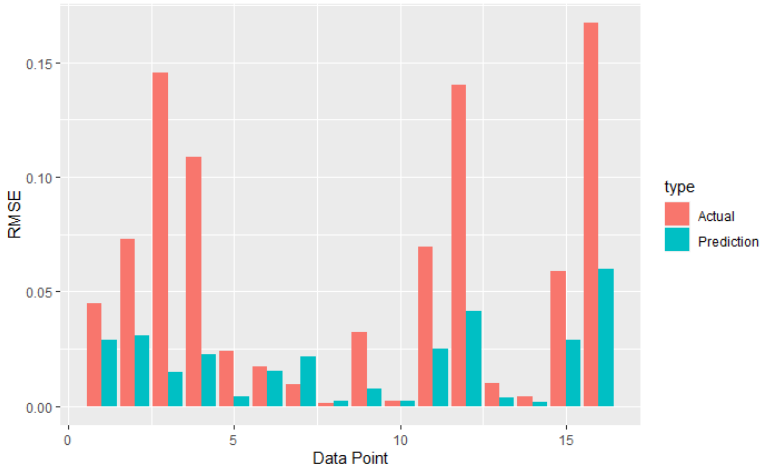


Our residuals vs fitted plot looks much better. The Breusch-Pagan test agrees, giving us a p-value of 0.498 which means that the constant variance assumption is not violated. However, the Q-Q plot looks slightly off and the Shapiro-Wilk normality test returned a p-value of 3.785e-09. This means that our normality of errors assumption is violated. We will look for a more robust model in the future that will be able to account for this flawl.

Additionally, the logarithmic transformation of the response variable slightly improved our other model diagnostics. We now have an R squared value of 0.4733 and an adjusted R squared value of 0.4483, which is slightly better than before.

3.2 Predictions

After fully fine-tuning our model, we can now make predictions on our test set. To evaluate how good our predictions were for the finalized transformed model, we used room mean square error. Our training set produced a value of 0.0424, while our testing set produced a value of 0.0569 This means that our model did an overall good job of predicting on our testing set. Below we have a visual representation of our predictions:



3.3 Mixed Effect Model

To address our issues of non-normality in the previous model, we decided to try to use a mixed effect model: a model that contains both fixed and random effects. For this model, we ended up using all of the variables that displayed no signs of collinearity (see section 2.2), and the following variables for our random effects: `Aquatic Products, Other`, `Alcoholic Beverages`, `Sugar Crops`, `Sugar & Sweeteners`. We used the same cleaned data and training set that was created in the previous sections to fit this model. The REML parameter was set to False in order to compute a maximum likelihood estimation.

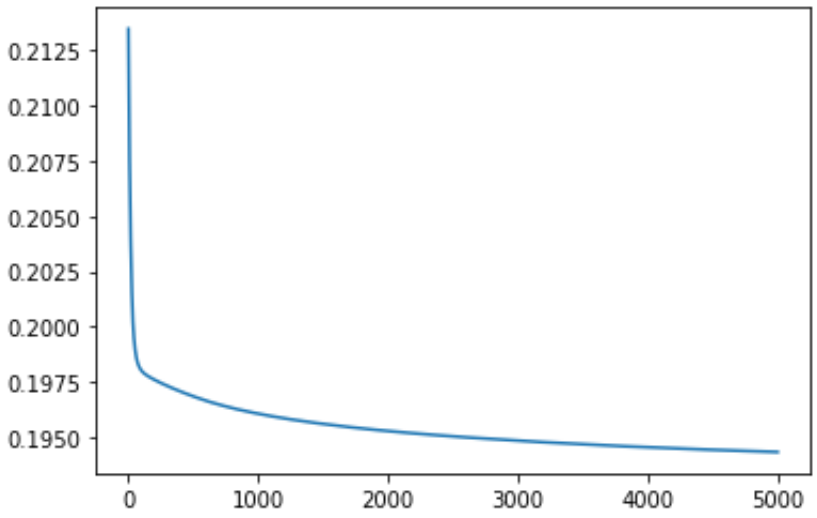
In order to evaluate the validity of this model, we used a Likelihood Ratio Test. After performing the necessary calculations, we obtained a test statistic of 98.79904, resulting in a p-value of 0. This gives us confidence that our model is valid.

We finally calculated the RMSE values of the training and testing sets. Our training RMSE was 0.0349 and the testing RMSE was 0.0418, meaning that this model does an excellent job at predicting new data as well as fitting the existing data we already gave it.

3.4 Deep Learning Model

Since machine learning has become omnipresent in the world of big data, we decided to additionally fit a deep neural network on our data. The data that we used for this model was the same in previous sections, so the only different task for us was to create our neural network. The model takes in our data and propagates it through a linear layer which converts it from a dimension of 29 to 100. Afterwards, it uses a ReLU activation function to bound its values. Finally, we pass it through a final linear layer which converts it to a dimension of 1, where it is activated using ReLU once more.

With the model created, we can now train it. It was trained over our data hundreds of times and the loss was recorded in each iteration. Since the other models used RMSE, we used RMSE to evaluate the accuracy of our model. The plot of our loss can be found below:



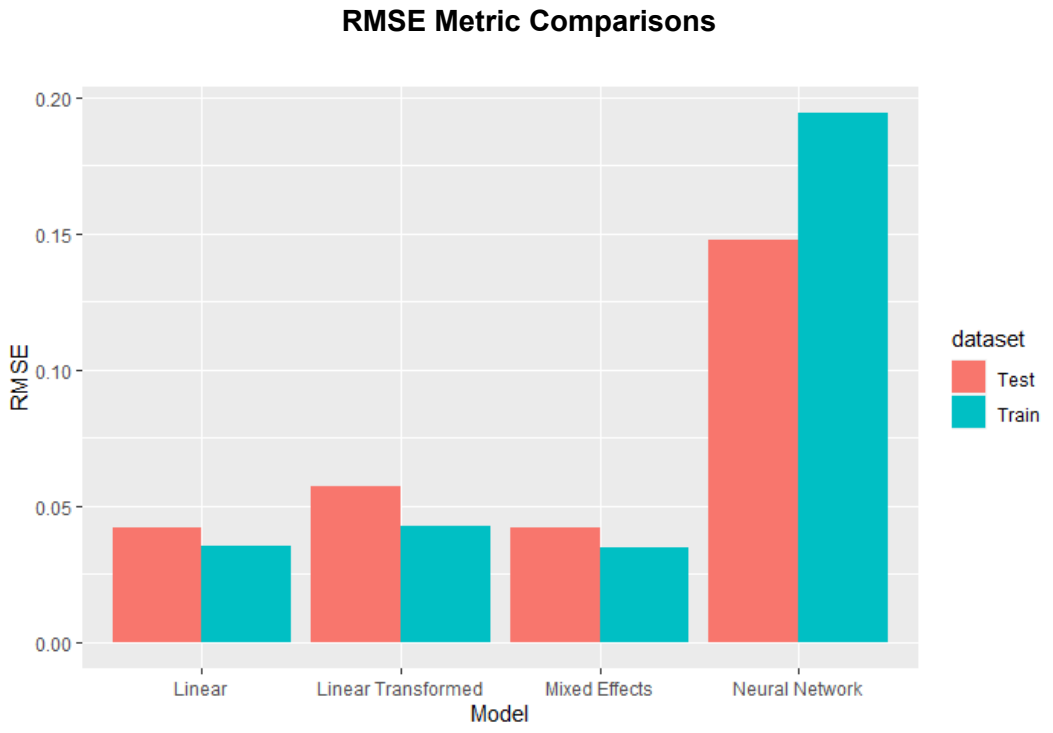
After the model was trained to its minimum loss value, we decided to evaluate the RMSE of the training and testing sets, obtaining values of 0.1943 and 0.1479, respectively.

3.5 Final Model Selection

Now that we have created, tested, and analyzed a variety of models, it is time for us to select one that we can use for predictions and more in-depth analysis. Our four models will be referred to as follows: linear, linear transformed, mixed effects, and machine learning.

The first metric we analyzed was the number of coefficients in the model, which directly relates to the complexity and overall difficulty of interpretation of the model. The linear and linear transformed models had 7 coefficients while the mixed effects model had 4 coefficients. The machine learning model is extremely complex and considers a multitude of interactions. Hence, the mixed effects model is less complex and easier to interpret, making it a better contender for our final selection.

We also analyzed the RMSE values of the training and testing sets for each model. Our linear model produced values of 0.0354 and 0.0419, the linear transformed model had 0.0424 and 0.0569, the mixed effects model had values of 0.0348 and 0.0418, and the machine learning model had values of 0.1943 and 0.1479. However, it's also important to note that the linear model had an adjusted R squared value of 0.4294, while the transformed model had an adjusted R squared value of 0.4483. A visual representation of the RMSE can be found below:



Due to its lower complexity and reduced RMSE value, we will select the Mixed Effects model as our final model

Section 4: Discussion and Conclusions

4.1 Data Summary

In this section we'll summarize our findings regarding our data and the data analysis process

When starting this project, we noticed that our data was extremely messy and required a lot of cleaning and transformations. Before we could even start fitting models, there was a lot to consider so that we could have a cleaner modeling process in the future.

Even after fitting our initial model, we found that there were more changes to consider so that they would be in line with our model assumptions. After removing outliers and transforming our variables so that they could fix the constant variance and normality of errors assumptions, we finally had a working initial model. Since we used sound variable selection techniques, these initial models did a sufficient job at modeling our data and produced low RMSE values on our testing set.

We then moved onto a mixed effects model to see if a different model could improve our predictions. After doing this, we found that the RMSE of the training and testing sets were slightly lower, indicating that we had a more robust model. Since the number of coefficients decreased with this model compared to the first two, we felt confident about this model as our final selection.

Finally, we wanted to see if the widely-discussed and flashy machine learning models that are currently being used would be able to solve this problem. However, despite including more interactions and having more complex calculus and mathematics behind the algorithm, it produced mediocre results. This was a little bit shocking to us, as it showed that just because a model is more complex does not necessarily mean that it will be better at making predictions.

4.2 Answering Initial Questions

In this section, we will make some conclusions surrounding our final model selection: the Mixed Effects model.

We'll start by answering the questions that we initially posed in the Introduction (see Section 1). We first wanted to see if it is even reasonable to try and model the COVID death rate using the variables that we were provided. After reflecting on this question, it becomes clear that this modeling is indeed possible. Almost all of our models were able to model this death rate. Each

model had overall significance and were able to fit the training data well. However, being able to fit the training data is not the tell-all, since the algorithm could be simply “memorizing” the data. Fortunately, we were able to see low RMSE values on our testing set predictions, which further confirms our hypothesis that we would be able to model this data.

Now that the model has been validated, we can move onto our second posed question, which was if some variables in the model could be considered more impactful than others. During our exploratory data analysis, we were able to answer this question by studying the collinearity and using variable selection techniques like AIC. Through the study of collinearity, we were able to find variables that go hand-in-hand, meaning that adding more than one of them would not be more impactful than just adding one of them. Through AIC selection, we found the specific variables that had the most impact on predicting the COVID death rate. These variables ended up being "Animal Products", "Animal fats", "Miscellaneous", "Stimulants", "Vegetal Products", "Vegetable Oils", and "Obesity", so we can say that these were more impactful than the rest in the dataset.

Finally, we wanted to know if these variables could directly predict the COVID death rate, or if there is some other intermediate variable that could explain this phenomenon. Unfortunately, it is hard to tell if our variables are the direct cause of differences in COVID death rates because as we know, correlation does not always imply causation. If we were to guess, we think that there is some intermediate variable that makes this correlation possible. Specifically, we hypothesize that lower supplies of food are indicative of a more impoverished and less-equipped in terms of resources to handle the pandemic, which could directly cause an increased COVID death rate. It would be interesting to do more research on this topic to try and uncover the true cause of the differences in death rate.

4.3 Main Conclusions

Taking all of this into consideration, we were able to make the following main conclusions:

- Just because a model is more complex does not mean that it will be better overall
- Even though country food supply and COVID deaths are objectively unrelated, we can still find a correlation between these two variables
- Certain variables, like *Animal Products* and *Vegetable Oils*, were indeed more impactful when making our predictions
- It is difficult to tell if food supply directly impacts COVID deaths, but did prompt further research and curiosity into this topic