Nick Friedel, Ryan Ferron, Henry Jones, Mario Silva, Matthew Schramm, Zijia Teng

Predictive Analytics Project Write-Up

Executive Summary

This report provides an analysis and evaluation of the factors that may affect the current mortality rate of the Covid-19 pandemic in each country across the globe. It also evaluates which factors determine how a country is classified as part of the “Developing World” or not. The methods of analysis include linear and logistic regression along with validation for each type of regression in order to determine the most effective predictive model. All calculations can be found in the R code uploaded to the group repository. The results of the data analysed show that the undernourished rate, the status whether a country is developing or not, and the percentage of sugars and cereal consumed (in kilocalories) to produce energy in the average adult’s daily diet seem to be the strongest predicting factors in determining the Covid-19 mortality rate within a country. On the other hand, for classification, it seemed that GDP per Capita had the strongest classifying effect on deciding if a country was or was not considered developing.

The data from the analysis was pulled from the following sources:

* Data for different food group supply quantities, nutrition values, obesity, and undernourished percentages were obtained from Food and Agriculture Organization of the United Nations [FAO website](http://www.fao.org/faostat/en/#home)
* Data for population count for each country came from Population Reference Bureau [PRB website](https://www.prb.org/)
* Data for COVID-19 confirmed, deaths, recovered and active cases were obtained from Johns Hopkins Center for Systems Science and Engineering [CSSE website](https://coronavirus.jhu.edu/map.html) and were chosen as of May 6, 2020.
* Data for determination of Developing Country status came from data prepared by the Development Policy and Analysis Division (DPAD) of the Department of Economic and Social Affairs of the United Nations Secretariat (UN/ DESA) as a [Statistical Annex](https://www.un.org/en/development/desa/policy/wesp/wesp_current/2012country_class.pdf)
* Data for national Gross Domestic Product and GDP per capita were gathered from The World Bank’s [National Accounts Data](https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)

Ultimately, the data selected was chosen prospectively through logic and research based on what seemed most reasonable to be the main factors of worse physical and economic health for each nation. In determining variables that may affect health, we were hoping to find patterns into the determination of the pandemic’s death and recovery rate and even developed nation status. We also felt this report would yield interesting results given the magnitude and eventual historical significance of the current dire circumstances of today’s reality.

In terms of the structure of the data, the variables and specifying units are as follows:

* Row Distinguishers are 170 countries worldwide acknowledged as sovereign by the UN.
* GDP variables:
  + Total GDP by a factor of 10 million US dollars
  + GDP per capita by 1,000 US dollars.
* Developing Countries variables:
  + First is classified by Y or N
  + Second is the binary version of the categorical variable (1 meaning it is a Developing Country, 0 meaning it isn’t)
* Fat Supply Quantity Data: Average percentages of fat consumed from each type of food by the average adult in that country including:
  + Alcoholic Beverages
  + Animal Products
  + Animal Fats
  + Aquatic Products
  + Cereals
  + Eggs
  + Fish and Seafood
  + Fruits excluding Wine
  + Meat
  + Miscellaneous
  + Milk
  + Offals
  + Oil Crops
  + Pulses
  + Spices
  + Starchy Roots
  + Stimulants
  + Sugar Crops
  + Sugar Sweeteners
  + Tree Nuts
  + Vegetal Products
  + Vegetable Oil
  + Vegetables
* Percent of population classified as Obese
* Percent of population classified as Undernourished
* Percent of confirmed Covid-19 cases amongst population
* Percent of deaths due to Covid-19 among population
* Percent of recovered cases from Covid-19 among population
* Percent of active cases of Covid-19 among population
* Total Population
* Mortality Rate of Covid-19 directly from confirmed cases in each country (death percentage / confirmed percentage)
  + **Critical to note that Mortality Rate was based only off the death rate by the infection rate as far as testing showed (Those not tested but potentially infected can skew the accuracy of this rate greatly)**
* Food Supply Kcal Data: Average percentages of energy (in kilocalories) used throughout the day by an average adult from each type of food
  + (Same types of food as seen for Fat Supply Quantity Data)
* Protein Supply Quantity Data: Average percentages of protein consumed from each type of food by the average adult in that country
  + (Same types of food as seen for Fat Supply Quantity Data)
* Food Supply Quantity kg Data: Average percentages of each type of food consumed daily (in kg) by the average adult in that country
  + (Same types of food as seen for Fat Supply Quantity Data)

It is also important to note that the analysis conducted has limitations.

Some of the limitations include:

* Hidden variables not gathered or considered may further or more effectively affect mortality rate and developing country status
* Multicollinearity amongst predictors within the model, specifically in regards to food group supply quantities
* Recording error in the raw data, especially for Covid-19 infection and mortality rate as testing is limited in many if not still all countries worldwide
* Mortality rate is also likely skewed given the lack of testing and the number of people not tested worldwide due to displaying no symptoms (hence why it is important to emphasize that **mortality rate was measured based only off of the death percentage by the infection percentage as far as testing showed**)

Cleaning the Data

After initially collecting and compiling all of the necessary data into one excel file, it was easy to first spot certain countries with invalid or incomplete data that could be deleted before uploading the file to R. These countries (rows) included those with numerous observations of “N/A” for Covid-19 data such as confirmed, recovered and active cases along with deaths due to inadequate reporting or public data available (North Korea) along with countries with no confirmed cases up to this point in time (Dominica, French Polynesia, Kiribati, Lesotho, Myanmar, New Caledonia, Saint Lucia, Saint Kitts and Nevis, Samoa, Solomon Islands, Turkmenistan, Vanuatu, Saint Lucia, Grenada, Saint Vincent and Grenadines). We also made the decision to still eliminate the bottom 10 percent of countries in population that were extremely small with very few Covid-19 cases (those remaining from the previous deletions were: (Malta, Belize, the Bahamas, Iceland, Barbados, Sao Tome and Principe, and Antigua & Barbuda).

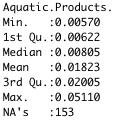
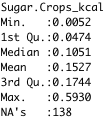
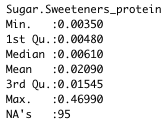
We did, however, decide to keep the countries remaining that had zero confirmed Covid-19 deaths as we felt this data could still prove important and measurable since there were still confirmed cases in all of these countries.

Upon uploading the pre-cleaned data into R, our initial exploratory analysis quickly revealed that many of the nutritional and food group variables had little-to-no weight on overall food supply quantities and needed to be removed to clean the dataset.

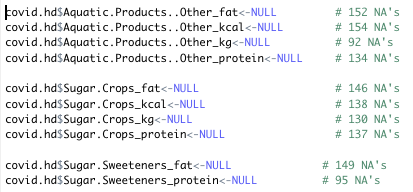
In order to do so, we set all quantities in the dataset initially listed at 0.0000 to NA in order to determine how many missing observations existed for each variable.

In doing so, we set all *logical* variables with the observation of 0.0000 back to 0.000 from NA, as to not overlook valid data. These variables included the mortality rate of Covid-19, the Undernourished and Obese % of populations in each country, the number of Covid-19 deaths and the binary variable for developed vs non-developed countries.

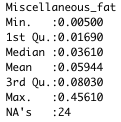
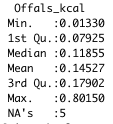
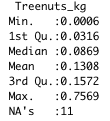
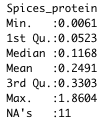
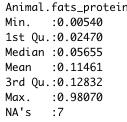
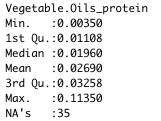
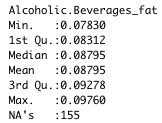
Following the observation of summary statistics for each of the variables, it was clear that variables like Aquatic Products had a significant amount of missing or inadequate data to where variables with over 50 NA observations could be removed. Included are a few examples of the summary statistics for the variables with large quantities of missing data:

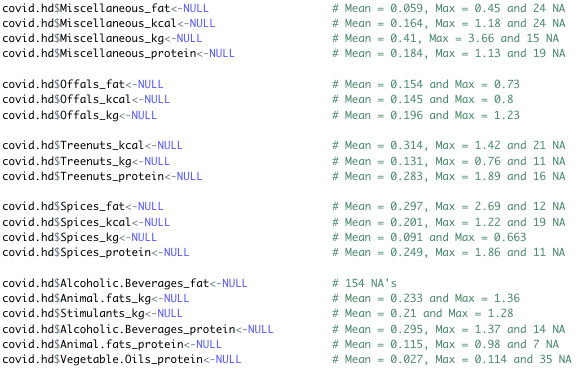
The following code shows these steps that we took to eliminate the variables of Aquatic Products, Sugar Crops and a few categories of the Sugar Sweeteners variable:



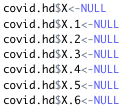
Then, we also decided that variables with minimal significance - Means of below 0.5 and a Max typically below 2 - could also be eliminated/nullified in the dataset. If a variable had a mean close to those constraints and still had over 10 NA values, we also decided to delete the variable. Some examples of the summary statistics of these variables are included again here:

From this cleaning, we ultimately also deleted the variables Miscellaneous Foods, and Spices, along with 3 of the 4 categories for Offals and Tree Nuts and various other low-yield variables, such as protein received from Animal Fats or Alcoholic Beverages as seen in our code here:



Finally, we also cleaned our separating columns that translated directly over from excel as X, X.1, X.2, etc:



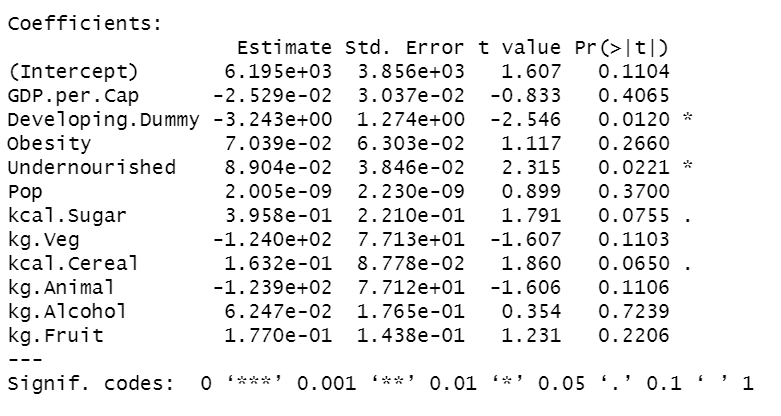
Linear Regression Task Model Proposals

Nick

For my regression model, I chose my variables based upon how much nutritional value each group would make. For example, I included variables like kcal of Sugars, how much protein each country gets from animal meats, as well as measures for grains, vegetables, fruit, and alcohol consumption. Intuitively, these are mostly what makes up many diets in almost all countries. I also included a variable to measure the GDP per Capita for each country. When running my model, the regression equation was as follows (coded in R):

lm(m.rate~GDP.per.Cap+Developing.Dummy+Obesity+Undernourished+Pop+kcal.Sugar+kg.Veg+kcal.Cereal+kg.Animal+kg.Alcohol+kg.Fruit, covid.v)

The fit of my model, measured by the Adjusted R-squared, was .077. The dummy variable denoting a developing country, as well as the variable for undernourishment rate were both significant at the 5% level. Kcal of Sugars and kcal of cereals were both significant at the 10% level. No other variables were significant in this model.



Matthew

First, I tested each category of variables as predictors of the variable Mortality.Rate to check for statistical significance.

Following my observations of each category of variables, I determined a final regression analysis which I felt might be the most appropriate which included all of the non-nutritional variables outside of Obesity rate, along with the percentages of energy consumed (in kcal) from fruits, sugar sweeteners, vegetable oils and vegetables. I removed obesity to try to reduce collinearity in the model given that some of the nutritional variables may closely correlate to or even predict obesity rate. Further, I chose to use the nutritional category of kilocalorie energy derived from each source of food in each country, specifically the food categories of sugar sweeteners and vegetable oils because typically those are unhealthy sources of food where one could derive high calorie intake. I also included fruits, vegetal products and animal products as predictors because I saw them as a healthier source of energy to potentially predict health and therefore mortality rate.

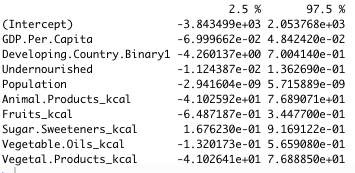
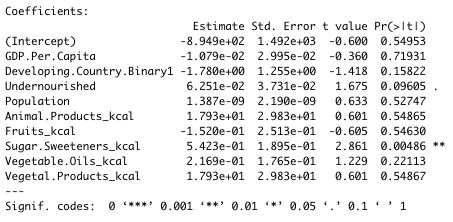
The model, as run in R, was:

lm(Mortality.Rate ~ GDP.Per.Capita+Developing.Country.Binary+Undernourished+Population

+Animal.Products\_kcal+Fruits\_kcal+Sugar.Sweeteners\_kcal

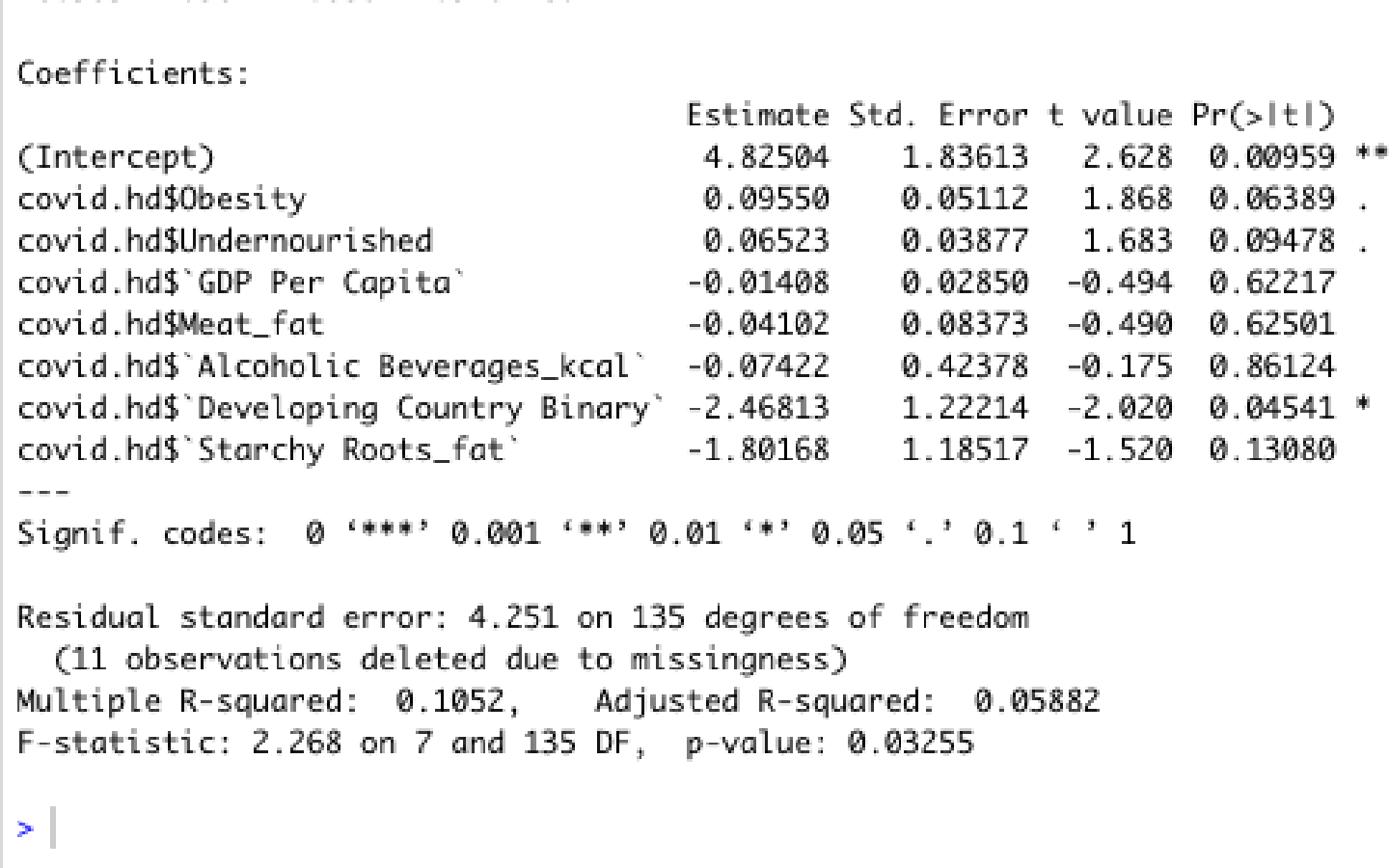
+Vegetable.Oils\_kcal+Vegetal.Products\_kcal, covid.hd)

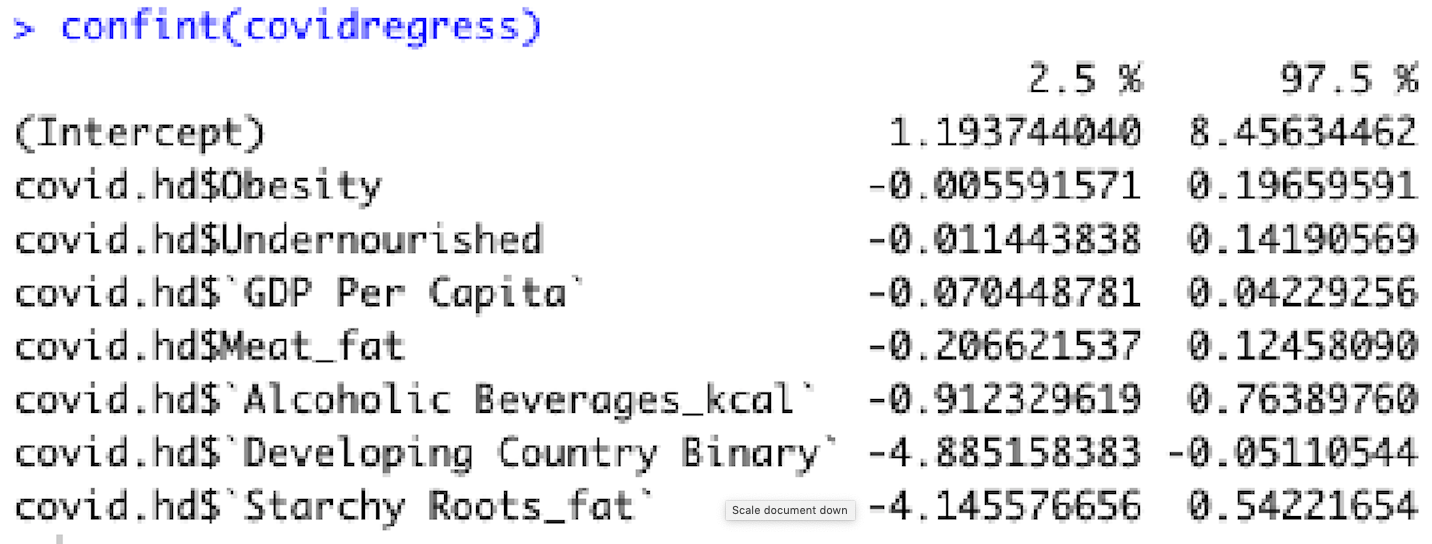
The fit of the model turned out to provide an R-squared of .1091 but an Adjusted R-squared of .05308, demonstrating a high level of complexity within the model to potentially increase out of sample error. No matter, I stuck with the model to only find 2 statistically significant predictors of Sugar Sweeteners at the .01 level of significance and the Undernourished rate at the .1 level of significance. The coefficients and their confidence intervals from the model can be seen here:



Henry

For my linear regression, I included variables I believed before doing the regression would cause certain countries to be more vulnerable to contracting the virus. Recent research on coronavirus has shown that those with weak immune systems tend to be more likely to test positive, and countries that consume an unhealthy diet, or are underfed, tend to have weaker immune systems. I also included variables that reflect a countries access to resources to help combat the virus, such as GDP per capita and the developing country binary. My left-hand side variable was Mortality Rate from COVID-19 and my righ-hand side variables included countries’ obesity rate, undernourished rate, GDP per capita, Meat Fat, Alcoholic Beverage kilocalories, Starchy Roots Fat, and the Developing Countries Binary. Here are my summary statistics and confidence intervals from my linear regression:





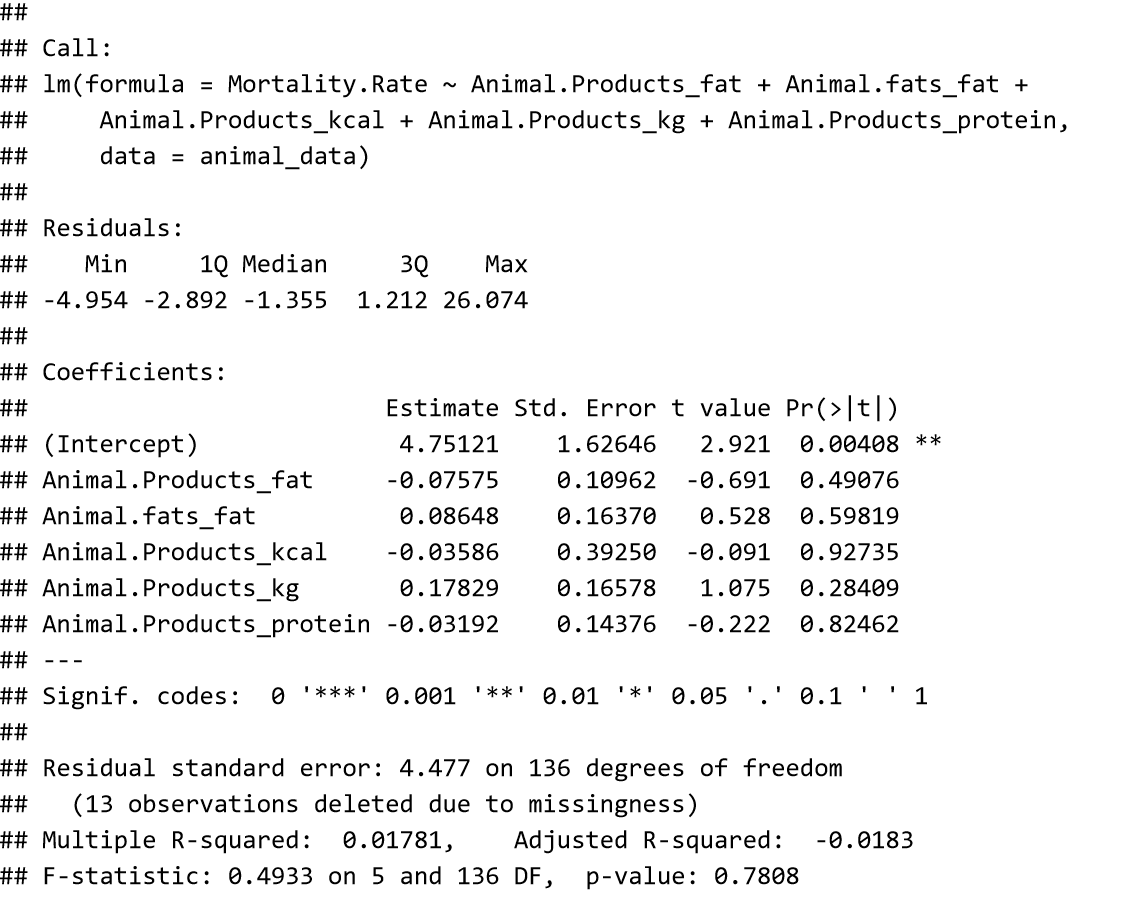
Zijia

The variables I chose for my regression model 1 is the animal products. There are 5 variables in my regression, for example, the animal fat, the weight of animal, animal product, the kcal of animal product. I use these variables to see if animal products have connections with the mortality rates.

The Code of my Regression Model 1:

animal\_reg = lm(Mortality.Rate ~ Animal.Products\_fat+Animal.fats\_fat+ Animal.Products\_kcal+Animal.Products\_kg+Animal.Products\_protein, animal\_data)

The result of Model 1:

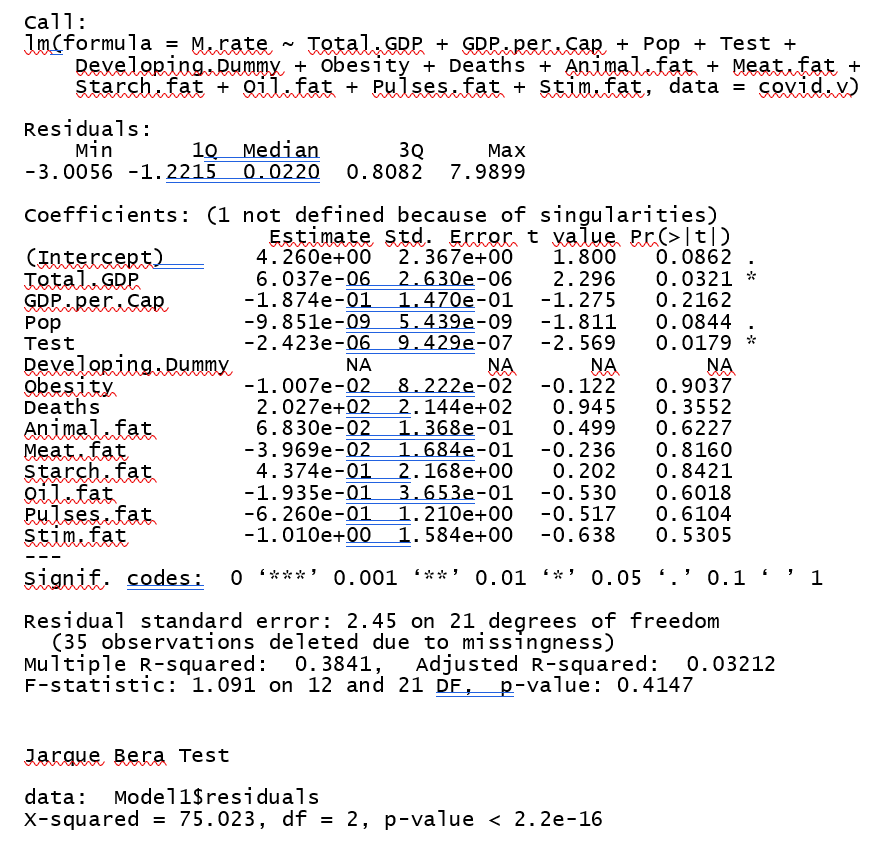


Mario

I chose my variables for my regression model , based on how many fats (Model 1) people consume. I included variables such as Oil Fats as bad fats, and Vegetable Fat as good fats. I also wanted to see what if wealth (GDP Per Capita Variable) played a role in the Mortality Rate. The code I ran in R was:

Model1 <- lm(M.rate~Total.GDP+GDP.per.Cap+Pop+Test+Developing.Dummy+Obesity+Deaths+Animal.fat+Meat.fat+Starch.fat+Oil.fat+Pulses.fat+Stim.fat, covid.v)

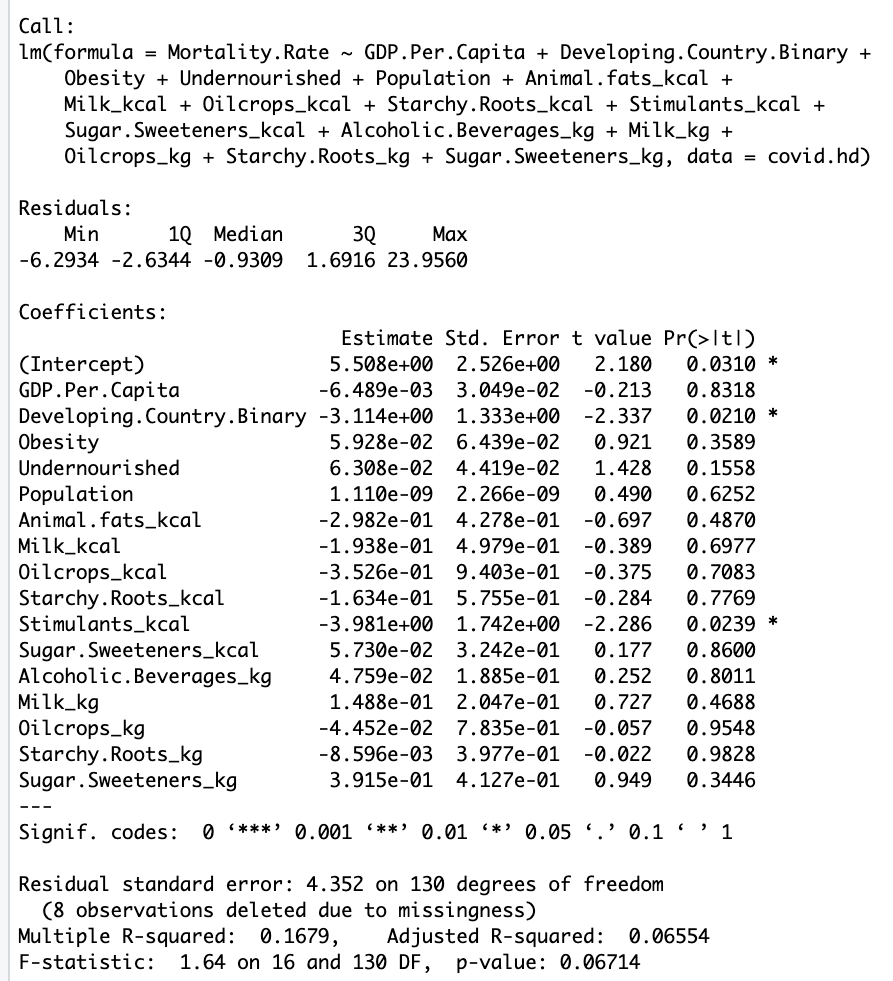
Model 1 Results:



Ryan

I wanted to mainly look at a few things in my linear regression to ultimately see what was the most correlating. I wanted to see the relationships between the quantities of food along with the calories that had, and then pair them up with obesity and GDP. I ended up choosing the following variables for my linear regression: Model1<-lm(Mortality.Rate~GDP.Per.Capita+Developing.Country.Binary+Obesity+Undernourished+Population+Animal.fats\_kcal+Milk\_kcal+Oilcrops\_kcal+Starchy.Roots\_kcal+Stimulants\_kcal+Sugar.Sweeteners\_kcal+Alcoholic.Beverages\_kg+Milk\_kg+Oilcrops\_kg+Starchy.Roots\_kg+Sugar.Sweeteners\_kg, covid.hd)

These were the results that I ended up with



Model Proposal Validation - Linear Regression

Nick

To validate my regression model, I partitioned the data set using seed = 1234, with 70% of the data in my training set, and 30% in my testing set. First, I constructed a new model using the same variables, except this time I fed the equation the training data. Then, I predicted the testing data using this model. After calculating the RMSE, the model was found to have an out-of-sample error of 3.99%.

Matthew

I used the same steps and values to partition and validate my data as Nick (above). However, after calculating my RMSE, the model was found to have an out-of-sample-error of 4.17.

Henry

Same steps as Nick with an of RMSE: out of sample error = 8.99.

Zijia

Same steps as Nick with an of RMSE: 5.407 out of sample error.

Mario

As with the rest of the team I used the same steps and values to partition and validate my data as Nick (above). However, after calculating my RMSE, the model was found to have an out-of-sample-error of 7.98.

Ryan

I ran the same steps to partition and validate my data which, after calculating the RMSE, produced an out of sample error of 4.55.

Best Model

As a group, we felt that Nick’s model best fit the data because he had one of the stronger Adjusted R-squared values and the lowest out of sample error percentage according to the RMSE. He also had the most statistically significant predictors in his model (four) even though we all found the data to be pretty difficult to navigate and find even one or two statistically significant variables in a strong model. This was likely because the variable Mortality Rate of Covid-19 is one that is very difficult to predict or find strong predictive variables for.

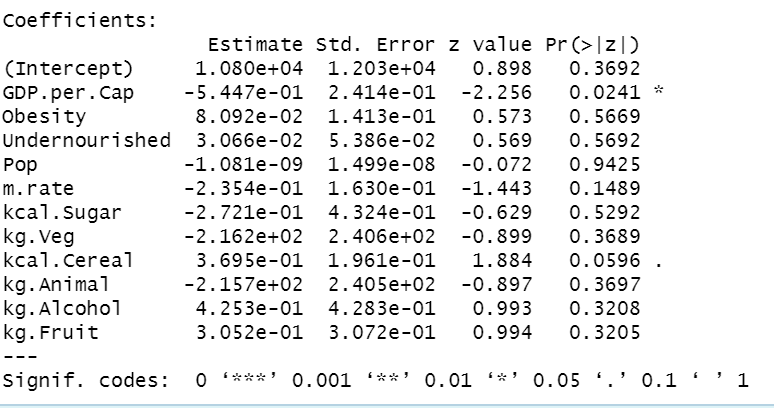
Classification Task Model Proposals

Nick

For the classification task, I chose to classify each country as either a developing country or not, using the same variables as my regression equation. However, since Developing.Dummy will be moving to the LHS as the classification, I moved mortality rate to be a predictor. Since the outcome will either be a ‘1’ or a ‘0’, I chose Logistic regression as the best method. The code in R looks like the following:

M\_LOG<-glm(Developing.Dummy~GDP.per.Cap+Obesity+Undernourished+Pop+m.rate+kcal.Sugar+kg.Veg+kcal.Cereal+kg.Animal+kg.Alcohol+kg.Fruit, data = Training, family = "binomial")

The fit of this model, given by the AIC, was 49.171. The only significant variables in this model were GDP per Capita (p=.024) and kcal of cereals (p=.059; significant at the 10% level).



Matthew

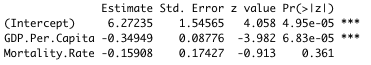
Using the Developing Country variable for classification, I used only 2 variables as predictors to the logistic regression model. I used GDP per capita, with the understanding that income and production levels are often thought of as the single-best measure of a country’s overall wealth, and Mortality Rate to re-check the previous findings of the linear regression in which the Developing Country variable was non statistically significant.

The code for the regression was:

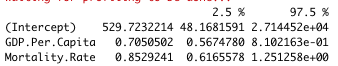
glm(Developing.Country.Binary ~ GDP.Per.Capita+Mortality.Rate, data = Training, family = "binomial")

The model turned out to produce a statistically significant intercept term and GDP Per Capita predictor each at the 0.001 level of significance. The coefficients were 6.272 and -0.349, respectively. Meaning, there exists a .349 decrease in the log of odds ratio for the predictor GDP per capita. The AIC fit for the model was 32.99.

The coefficients can be seen here:

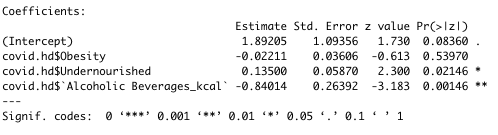


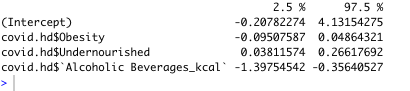
However, after improving the interpretive ability of the logistic regression by utilizing the exponential base function and raising E to the power of Beta, the confidence interval for the model came out to be:



Henry

After running the linear regression, I decided to focus on certain levels of what obesity, undernourished, and kilocalories consumed from alcoholic beverages and the effect on mortality rate from covid-19:



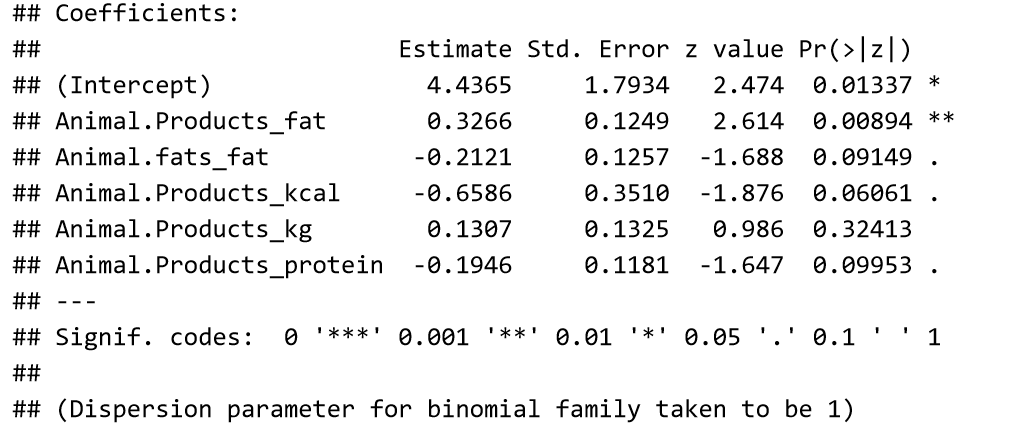


Zijia

I chose to classify if the country is a developing country or not, and use the model 1 animal data as the regression for the classification task. I measure the animal consuming. I changed the mortality rate as a predictor to re-check the data.

The code:

log\_reg<-glm(Developing.Country.Binary ~ Animal.Products\_fat+Animal.fats\_fat+ Animal.Products\_kcal+Animal.Products\_kg+Animal.Products\_protein, data = Training, family = "binomial")

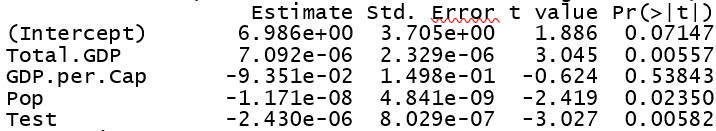


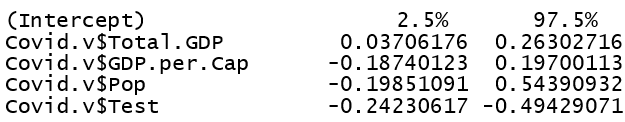
Mario

I have chosen to classify each country as either a developing country or not, using the same variables as my regression equation for the classification task. Being that the Developing. Dummy is becoming the Dependent Variable as part of the classification, however, I moved mortality rate to be a predictor or independent variable. Due to the fact that the result will be either a 1 or a 0, I chose Logistic regression as the best method and got an AIC Fit of 41.47. Total GDP with a p-value of 0.0321 and Test Per Population with a p-value of 0.0179 were the only two significant variables.

The code used is:

glm(Developing.Dummy ~ Total.GDP+GDP.per.Cap+Pop+Test+M.rate+Obesity+Deaths+Animal.fat+Meat.fat+Starch.fat+Oil.fat+Pulses.fat+Stim.fat, data = Training, family = "binomial")

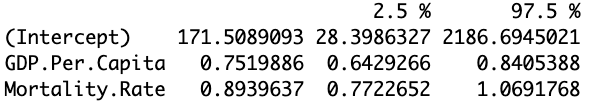
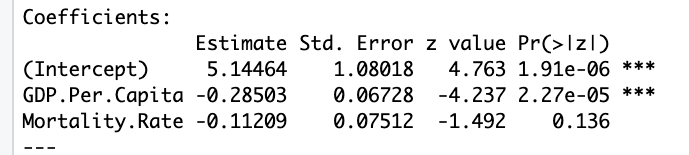




Ryan

I had the exact same ideas as Matthew, thinking that the most important important ways to determine the relationship between underdeveloped developed countries were by looking at the GDP and the mortality rate due to Covid-19. I, however, received slightly different results at 5.144 and -0.285, meaning, there exists a 0.285 decrease in the log of odds ratio for the predictor GDP per capita. The AIC Fit for me was 48.895.

glm(Developing.Country.Binary ~ GDP.Per.Capita+Mortality.Rate, data = Training, family = "binomial")



Model Proposal Validation - Logistic Regression

Nick

To validate my model, I used the same seed = 1234, as well as the same split of 70% training and 30% testing. For the accuracy of my model, I constructed a confusion matrix. The accuracy of the model was measured at 88.64%, **making the out-of-sample error 11.36%.**

Matthew

I followed the same steps as Nick (above), however I used a classification probability bound of 0.4 as opposed to 0.5 to classify the training and testing countries as Developing. Under these bounds, I ended with an accuracy of 95.33% for the training data and 93.48% for the testing data. Therefore meaning, the in-sample error for this classification model was 4.67% and **the out of sample error was 6.52%**.

Henry

In order to validate the data, I used dummy variables and made developing countries my training variables and Developed countries my testing variables, to see how much GDP per capita has an impact in predicting the spread of coronavirus.

Zijia

I used the same steps as Nick did, and came up with an **out of sample error of 19.6%** and an **accuracy of 80.4%.**

Mario

I used the same seed = 1234 to validate my model, as well as the same split of 70 per cent training and 30 per cent testing. For my model's accuracy I built a confusion matrix. The model's accuracy was calculated at 88.64%, **resulting in an out-of-sample error of 11.36%**

Ryan

I used the same steps as Nick and the same adjustment as Matthew and my accuracy was 93.33, which results in an **out-of-sample error of 6.67%.**

Best Model

The strongest model that we chose for the logistic regression was Matthew’s because of his low AIC at 32.99 and also the fact that his model produced the lowest out-of-sample error at 6.52%. His data also made it clear that the GDP per capita predictor seemed to be the most significant predictor of developing nation classification.

Conclusion and Lessons Learned

According to our strongest linear regression model, it can be predicted that the Undernourished rate, the status whether a country is developing or not, and the percentage of sugars and cereal consumed (in kilocalories) to produce energy in the average adult’s daily diet seem to be the strongest predicting factors in determining the Covid-19 mortality rate within a country. Of course, this is subject to a variety of testing, multicollinearity and missing variable error in the data. However, from the analysis, the health of a country does seem to play some part in the current mortality rate due to Covid-19.

According to the logistic regression analysis, we can understand that GDP per Capita seems to be the strongest predictive factor in determining whether a country is classified as Developing or not.

When all is said and done, we learned quite a bit about R, struggling through coding, specifically as it pertains to running regressions in R and partitioning the data, and also understanding how to analyze the differences between models.

Specifically, there was the issue of collinearity within our data, as we especially found this to be the case when comparing our regression models and seeing how for some of us, some variables would demonstrate little to no statistical significance, whereas others would find similar or the same variables significant based on the other gorupings of variables that were included in the regression. This is something to note going forward, especially in determining a good dataset to use, it can make it a lot easier on yourself if there is a larger variety and scope of variables to use as predictors in a model.

We also learned the importance of persistence with R, as many of us ran into problems dealing with NA variables that would prevent us from finding the RMSE or confusion matrix accuracy rate for either regressions. We had to help each other diagnose the issues and find the necessary code, like recoding NA to 0.000, in order to fix the error terms and troubleshoot.

Ultimately, we felt we were able to come to a decent conclusion into understanding and predicting the Covid-19 mortality rates, obviously with much skepticism due to a number of possible limitations. However, we were especially able to practice and begin to understand even just a piece of the backbone of computational methods in regression.