Associations Between Diet and COVID-19 Rates

Alex Shen

Introduction

When the novel coronavirus (COVID-19) pandemic spread through the world in late 2019 and 2020, it caused quarantines resulting in worldwide shutdowns of travel and commerce. Despite many limitations on travel, mask mandates, testing protocols, and social distancing requirements, the disease still caused shortages in hospital resources, including PPE, ventilators, and ICU beds. For example, the pandemic has caused up to 300,000 people in the US needing ICU beds simultaneously, while estimates for total ICU beds in the US are around 100,000.

The pervasiveness of the disease begs the question whether or not individual behaviors and life choices can decrease one's risk of either contracting or dying from the disease. Obesity, for example, is associated with higher rates of severe illness from COVID-19 (https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html). While obesity is influenced by genetic and environmental factors, it can be controlled by lifestyle choices such as exercise and healthy eating.

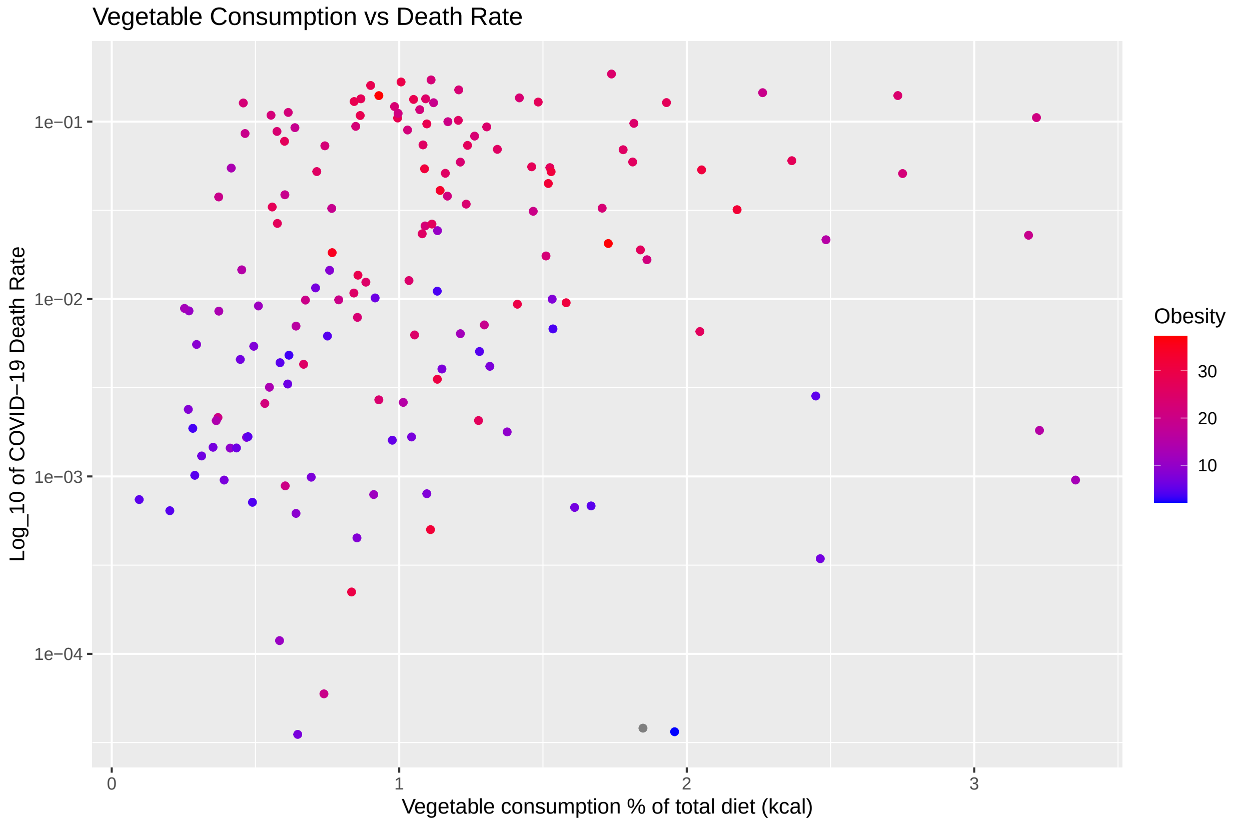
On the topic of healthy eating, Harvard T.H. Chan School of Public Health says that a "balanced diet consisting of a range of vitamins and minerals, combined with healthy lifestyle factors… most effectively primes the body to fight infection and disease." As a result, we are interested in whether or not dietary choices are associated with rates of contracting COVID-19 or rates of serious illness/death.

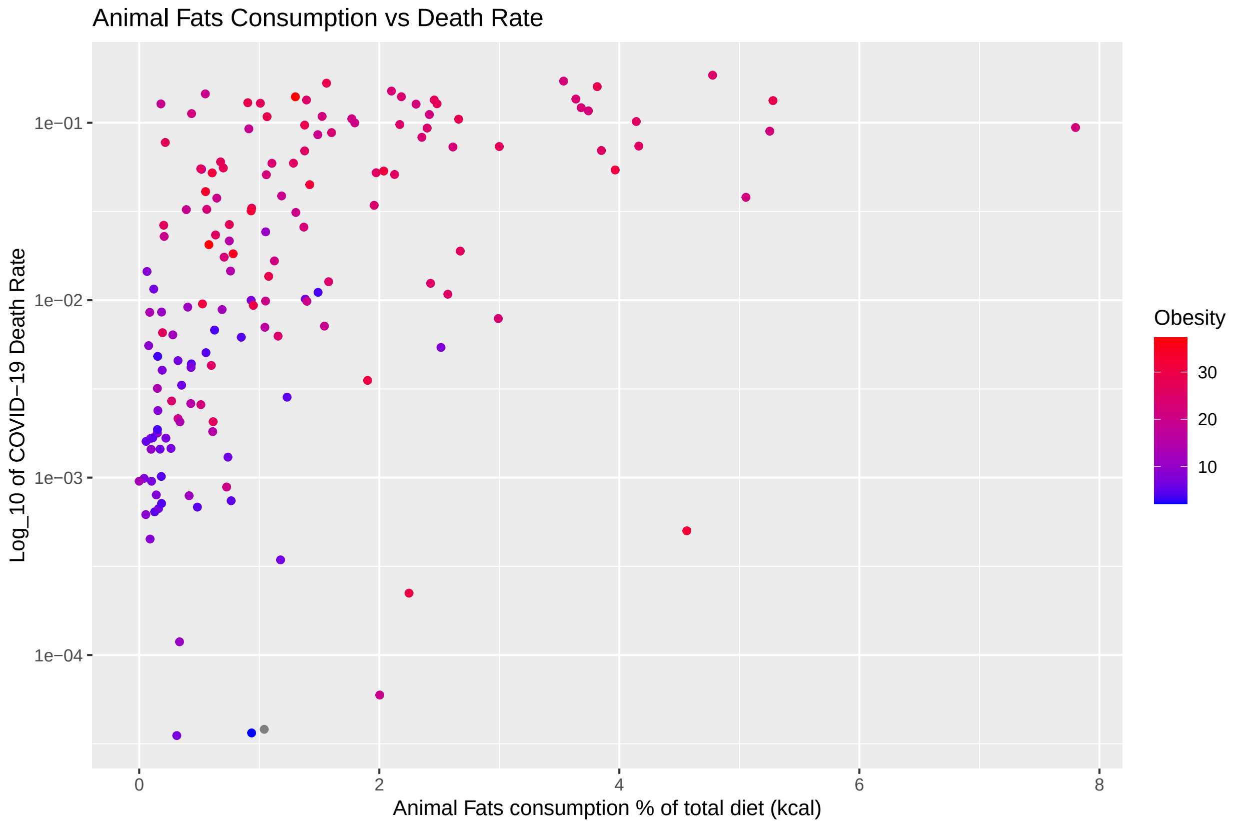
The data is called the COVID-19 Healthy Diet Dataset and comes from Kaggle courtesy of user Maria Ren, and was last updated the week of 02/06/2021, giving plenty of data from the COVID-19 pandemic to work with. Ren obtained all diet and obesity/undernourished rate information from the Food and Agriculture Organization of the United Nations website, all population measures from the Population Reference Bureau website, and all COVID-19 data from the Johns Hopkins Center for Systems Science and Engineering website. Ren combined data from these sources together, and did preliminary data cleaning.

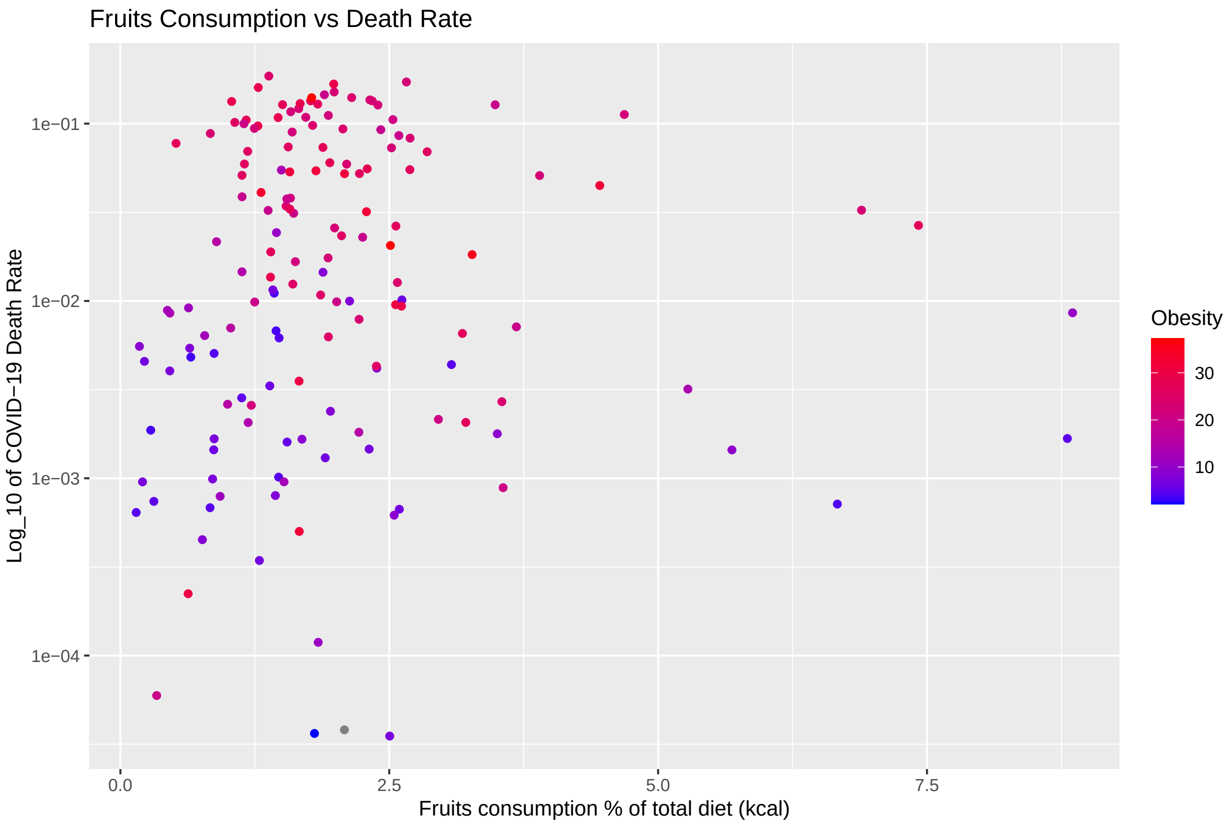
There are 4 datasets provided, each of which contains information about a part of people's diet. One contains information about people's consumption of fat, another for protein, another for overall food intake (in kg), and another for overall food intake (in kcal). Each dataset has 170 rows (countries) and the same 32 columns. There is a column for country name and many different columns representing food groups that contain information about what percentage of people's diet (in the category defined by the dataset it comes from) that each food group takes up. For example, the animal fats column in the Fats dataset represents the percentage of a countries total fat intake that is comprised of animal fats. Meanwhile, the column animal fats in the Overall Food Intake (kg) dataset will represent the percentage of countries total food intake in kg that is comprised of animal fats. In addition, there are columns (we will refer to these following columns as demographic/outcome information) for obesity rate, undernourished rate, confirmed COVID-19 case rate, COVID-19 death rate, active COVID-19 case rate, and total population. All of the rates are calculated as percentage of total population, for example, confirmed COVID-19 case rate in a country is the percent of people in that country that have COVID-19. The demographic/outcome information is identical for all 4 datasets.

The 4 datasets were combined by joining them together on the country field (all the countries in the 4 sets are the same), and deleting the duplicate demographic/outcome information. My overall research question will be to examine any associations between the different proportions certain food groups take up of people's diet and COVID-19 rates on a country level.

EDA


Vegetable Consumption vs Death Rate
Log_10 of COVID−19 Death Rate
1e−01
1e−02
1e−03
1e−04
0
1
2
3
Vegetable consumption % of total diet (kcal)
Obesity
30
20
10


Animal Fats Consumption vs Death Rate
Log_10 of COVID−19 Death Rate
1e−01
1e−02
1e−03
1e−04
0
2
4
6
8
Animal Fats consumption % of total diet (kcal)
Obesity
30
20
10



In these 3 plots, we have the percentage of each country's diet taken up by vegetables, animal fats, and fruit respectively on the x axes, and the rate of death from COVID-19 on the y axis. The darker the points, the lower the obesity rate is in the nation. The lighter and more yellow the points, the higher the obesity rate. From all 3 graphs, we can generally see that high death rates are associated with a higher rate of obesity, which makes sense.

For the vegetable graph, it is hard to make any discerning observations without stretching the imagination. From the animal fats graph, we can see that generally, countries with higher consumption of animal fats tend to be be ones with higher death rates. For the fruits graph, it seems that most countries with high death rates are clustered around 2.25% of their diet being fruits.

Methodology

Model formulation/selection

We decided to concentrate on two response variables in particular, rates of COVID-19 related deaths and confirmed cases. We will fit two models, one for each response variable. We are mostly interested in the relationship between each of the four predictor variables vegetable, fruit, meat, and animal fat consumption as a percentage of total food consumed (in kcal) and the response variables. These variables will go into my model. We also decided to control for the consumption of other dietary food groups as a percentage of total food consumed (in kcal). As a result, we also added variables for alcoholic beverages, cereals (wheat, rice, and other grain products), eggs, fish and seafood, milk, offals, oilcrops (including coconuts, peanuts, rapeseed, mustardseed, sunflower seeds, soybeans, etc.), pulses (legumes), spices, starchy root vegetables, spices, stimulants (including coffee, tea, and cacao products), tree nuts, vegetable oils, sugar and sweeteners, and finally miscellaneous, all representing percentage of total food consumed in kcal in order to control for these effects.

There was data available for the percentage of total protein and fats each of the food groups took up, but after careful consideration we decided against including them to eliminate high multicollinearity.

We also added in variables for obesity rate and population to control for these factors as well. We decided to not to include undernourished rate. This was because the data for undernourished rate was recorded in a strange way where anything under 2.5% was recorded in the same way. We did not see a principled way to create numerical values for this data or bin them into categories, and using another source of data was not feasible because we did not think we could find a measure of undernourished rate calculated in the same way as the original dataset. As a result, we decided to omit the undernourished rate from the model. Lastly, from a quick initial survey, it seems that how affluent a country is may effect COVID-19 case rates, deaths, etc. So, we downloaded the 2020 GDP per capita for all countries from the International Monetary Fund, and merged it into my dataset using the country name. We went through and cleaned it up and hand inputted a few data points due to certain countries being named slightly differently.

Because both of our response variables of deaths and confirmed case rates can be represented as proportions of the overall population of a country (bounded by 0 and 1), we decided to use a beta regression. The model formulation is shown below.

let $PropT\ X$ = proportion of a country's total food consumption in kcal coming from food X.

let $p_i = Death\ Rate_i | PropT\ Alcohol_i, PropT\ Animal\ Fats_i, PropT\ Cereals_i,$
$PropT\ Eggs_i, PropT\ Seafood_i, PropT\ Meat_i, PropT\ Fruits_i, PropT\ Milk_i,$
$PropT\ Miscellaneous_i, PropT\ Offals_i, PropT\ Oilcrops_i, PropT\ Pulses_i,$
$PropT\ Spices_i, PropT\ Starchy\ Roots_i, PropT\ Stimulants_i,$
$PropT\ Sugar\ and\ Sweeteners_i, PropT\ Treenuts_i, PropT\ Vegetable\ Oils_i,$
$PropT\ Vegetables_i, Obesity_i, GDP\ Per\ Capita_i, Population_i$

where $p_i \sim Beta(\alpha, \beta)$

$log(p_i/1 - p_i) = \beta_0 + \beta_1 PropT\ Alcohol_i + \beta_2 PropT\ Animal\ Fats_i + \beta_3 PropT\ Cereals_i + \beta_4 PropT\ Eggs_i +$
$\beta_5 PropT\ Seafood_i + \beta_6 PropT\ Meat_i + \beta_7 PropT\ Fruits_i + \beta_8 PropT\ Milk_i + \beta_9 PropT\ Miscellaneous_i +$
$\beta_{10} PropT\ Offals_i + \beta_{11} PropT\ Oilcrops_i + \beta_{12} PropT\ Pulses_i + \beta_{13} PropT\ Spices_i + \beta_{14} PropT\ Starchy\ Roots_i +$
$\beta_{15} PropT\ Stimulants_i + \beta_{16} PropT\ Sugar\ and\ Sweeteners_i + \beta_{17} PropT\ Treenuts_i + \beta_{18} PropT\ Vegetable\ Oils_i +$
$\beta_{19} PropT\ Vegetables_i + \beta_{20} Obesity_i + \beta_{21} GDP\ Per\ Capita_i + \beta_{22} Population_i$

The second model with confirmed COVID_19 case rate as the response has the same formulation, except replacing $Death\ Rate_i$ with $Confirmed\ Case\ Rate_i$

Missing Data

After subsetting the data for only predictors we are interested in, 7 of the 170 countries in the dataset contain missing values. All missing values occur in either the variables for rates of obesity, confirmed cases and deaths. Missing data for a country in the confirmed cases variable was always accompanied with a missing deaths variable. Generally speaking, most countries with missing values are either countries with very small populations or are special cases such as Taiwan or North Korea. We assume this data is MAR since it seems like the missing values are mostly explained by something observed in the data (population). Most of the missing data comes in the response variables of confirmed cases and deaths, and imputing response variables does not make much sense. As a result, we will drop these 7 observations with missing values and perform a complete case analysis.

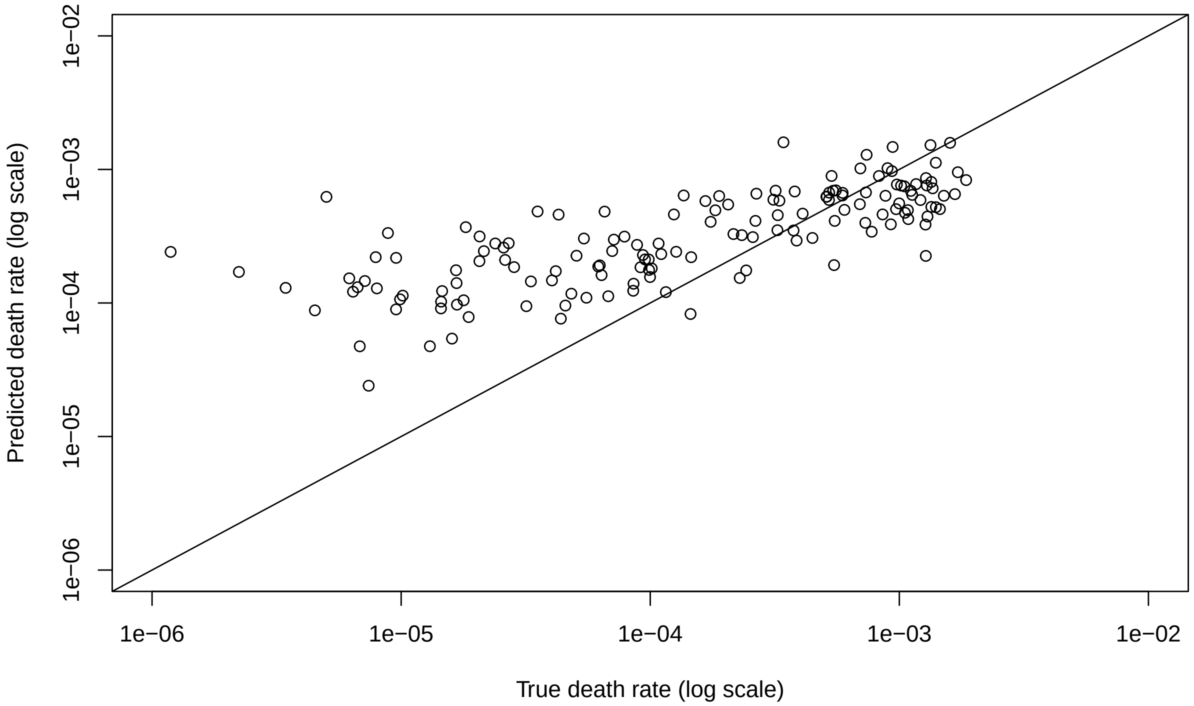
In addition, we will drop 8 additional observations that have values of 0 for deaths. For one or two of the 8 countries (such as Cambodia), we think that it is unfeasible to have 0 deaths from COVID-19 by February 2021. For the other countries, they are generally extremely small island nations, so it is feasible to have 0 deaths. We will drop these observations in either case because we don't want to be impute response variables, and beta regression models must have response variables between 0 and 1 non-inclusive. We will keep in mind that dropping these 15 total observations will bias the results.

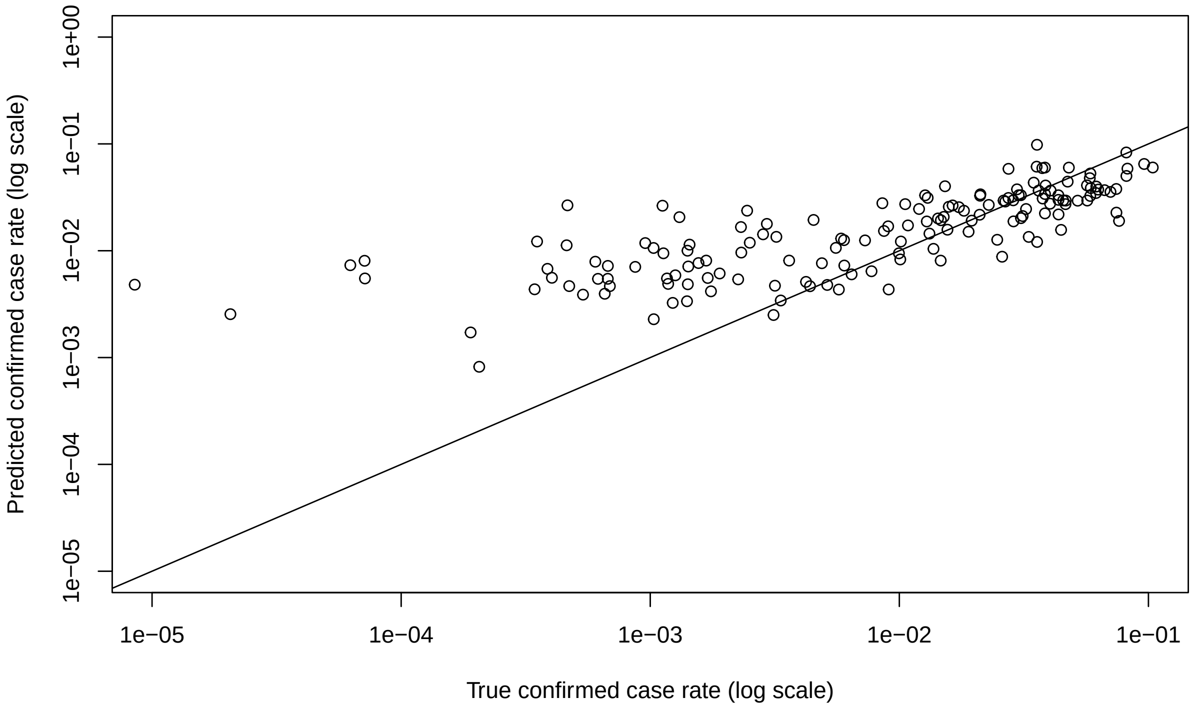
Model Diagnostics and Assumptions

One assumption necessary is the independence of observations. This assumption may not be completely met for both models. For example, bordering countries may often share similar diets due to similar cultures.

Having close geographical proximity may imply similar climate and habitablity for plant and species, leading to similar diets as well. Countries sharing borders may have more similar rates of COVID-19 as well due to more people (adjusted for population) moving between 2 bordering countries compared to 2 countries across the world. As a result, observations are not independent in this case, but we will proceed.

Next, we check the linearity condition. As seen in the below plot of predicted vs true death rates (both on log scales), this condition is not satisfied for the death rate model. If the condition were to be satisfied, the cloud of points would roughly fall around and along the plotted line of predicted death rates equaling true death rates. As we can see, this is not the case with patterning on the graph. On left of the graph, for countries with lower death rates, the model overpredicts death rates. On the very right of the graph, for countries with higher death rates the model tends to underpredict the death rates. For the confirmed case rate model, the linearity condition is also not satisfied. A very similar phenomenon occurs, where for countries with lower confirmed case rates the model overpredicts, on for countries with higher confirmed case rates the model starts to underpredict.





Results

The first model has death rates as the response variable, and the second model has confirmed case rates as the response. Otherwise, the formulations are the same. Some of the main parts of diet we were focused on was the percentage of total diet (in kcal) taken up by animal fats, fruits, and vegetables. For percentage of total diet taken up by animal fats, the coefficient from the model was 0.24. This means that holding other predictors constant, for every one unit increase in percentage animal fats take up of a country's diet, we expect to see a 0.24 increase in the logit of the deaths due to COVID-19 in that country as a proportion of total population with a 95% confidence interval of [-1.87, 2.35]. Holding all other predictors constant, for every one unit increase in percentage of a country's total diet taken up by fruits, we expect to see a 0.25 increase in the logit of death rate with a 95% confidence interval of [-1.86, 2.36]. Holding all other predictors constant, for every one unit increase in the percentage of a country's total diet taken up by vegetables, we expect to see a 0.12 increase in the logit of death rate with a 95% confidence interval from [-2.03, 2.27].

For the second model modeling confirmed cases as a proportion of population, holding all other predictors constant, for every one unit increase in the percentage of a country's diet taken up by animal fats (in kcal), we expect to see a 0.43 increase in the logit of confirmed cases as a proportion of total population, with a 95% confidence interval of [-1.66, 2.52]. For every one unit increase in the percentage of a country's diet taken up by fruit, we expect to see a 0.49 increase in the logit of rate of confirmed cases holding all other predictors constant, with a 95% confidence interval of [-1.60, 2.57]. Lastly, for every one unit increase in percentage of a country's diet taken up by vegetables, we expect to see a 0.35 increase in the logit of rate of confirmed cases holding all other predictors constant, with a 95% confidence interval of [-1.77, 2.47].

Curiously, for both models, at a significance level of 0.05, no predictors in either model are significant given the current model formulations. However, this does not mean we found that there is not an association between any of these predictors and the two response variables, but rather it means we fail to find significant evidence of an association given the current model formulations.

Discussion and Conclusions

The fact that no predictors in either model were significant was really surprising because we expected to see at least a few significant predictors, even if they did not come in the variables of interest. For example, in both models we control for obesity. As stated in the introduction, obesity is a known risk factor for serious complications or death in COVID-19 cases. However, in the first model, we found that obesity was not a significant predictor for death. The reason for this may be due to the data we have on diet sharing some of the same explanatory power that obesity has, since information about a country's diet can presumably be associated with obesity rates. Removing the diet related information, we can see that obesity is a significant predictor for death. However, this does not extend in the other direction; removing obesity in both models does not change the fact that all predictors in both are not significant.

One limitation is the missing data. Dropping missing values and unfeasible values to conduct a complete case analysis causes bias in the resulting estimates and results. Unfortunately, due to most of the missing data occurring in the response variables, imputation does not seem like a good solution to the problem. Further research through online databases could be used to find these values.

Another limitation is a the issue of countries not being independent observations. It might be possible to add random effects to the model to cope with the independence violations. With further research with groupings of countries can be formed based on a combination of geography, cultural boundaries, and other factors to be introduced as random effects.

Lastly, a limitation is the violation of the linearity assumption for both models. This may be due to not controlling for certain predictor variables that are not available in the original dataset. Perhaps an expanded dataset with more predictors allowing us to control for possible confounding variables (such as poverty levels, rates of other COVID-19 risk factors, etc.) could alleviate this issue.

One note about this analysis is that the data provides diet information and COVID-19 rates by country. Diet and COVID-19 cases happen on an individual level, and are aggregated at the country level for this dataset, and the analysis itself is therefore done on a country level. For examining potential associations between diet and COVID-19 rates, it may be better to have data on the individual level, where each observation is a person, with predictor variables pertaining to diet and response variables that are Bernoulli variables indicating whether or not the person got COVID-19, if they died or had severe illness from it, etc. It may be useful for further research to approach analysis on an individual level rather than aggregated on a country level.

Overall, no significant associations between diet and COVID-19 death rates were found given the current model formulation. It might be worth further investigating this topic, especially given the large impact that COVID-19 has on people's lives and the disruptive influence on public health systems worldwide. Further research may be conducted with different model formulations or acquiring more information to inform the model. However, digging deeper into this topic, a pivot might be in order to first examine the relationship between the breakdown of each country's diet and rates of obesity, in order to determine whether or not it is appropriate to control for obesity in the model.

Appendix

Table 1: Death Rate Model

Term	Estimate	Standard Error	Statistic	P-Value	95% CI Lower	95% CI Upper
Intercept	-16.55	53.72	-0.31	0.76	-121.84	88.74
Percentage of total consumption from Alcohol	0.34	1.07	0.32	0.75	-1.77	2.44
Percentage of total consumption from Animal Fats	0.24	1.08	0.22	0.82	-1.87	2.35

Term	Estimate	Standard Error	Statistic	P-Value	95% CI Lower	95% CI Upper
Percentage of total consumption from Cereals	0.15	1.08	0.14	0.89	-1.96	2.26
Percentage of total consumption from Eggs	0.61	1.09	0.56	0.57	-1.52	2.75
Percentage of total consumption from Seafood	-0.08	1.10	-0.07	0.94	-2.24	2.08
Percentage of total consumption from Meat	0.12	1.08	0.11	0.91	-2.00	2.23
Percentage of total consumption from Fruits	0.25	1.08	0.23	0.82	-1.86	2.36
Percentage of total consumption from Milk	0.21	1.07	0.20	0.84	-1.89	2.32
Percentage of total consumption from Miscellaneous	-0.39	1.12	-0.34	0.73	-2.59	1.81
Percentage of total consumption from Offals	-1.36	1.30	-1.05	0.29	-3.91	1.19
Percentage of total consumption from Oilcrops	0.02	1.07	0.02	0.99	-2.08	2.11
Percentage of total consumption from Pulses	0.05	1.08	0.05	0.96	-2.06	2.16
Percentage of total consumption from Spices	-0.17	1.14	-0.15	0.88	-2.41	2.06
Percentage of total consumption from Starchy Roots	0.13	1.07	0.12	0.91	-1.98	2.23
Percentage of total consumption from Stimulants	0.50	1.10	0.45	0.65	-1.66	2.65
Percentage of total consumption from Sugar/Sweeteners	0.18	1.08	0.17	0.87	-1.93	2.29
Percentage of total consumption from Treenuts	0.50	1.08	0.46	0.65	-1.63	2.62
Percentage of total consumption from Vegetable Oils	0.19	1.08	0.18	0.86	-1.92	2.31
Percentage of total consumption from Vegetables	0.12	1.10	0.11	0.91	-2.03	2.27
Obesity	0.02	0.01	1.78	0.08	0.00	0.05
GDP per capita (in thousands of dollars)	-0.01	0.00	-1.47	0.14	-0.01	0.00
Population (in hundreds of millions)	0.02	0.05	0.54	0.59	-0.06	0.11

Table 2: Confirmed Case Rate Model

Term	Estimate	Standard Error	Statistic	P-Value	95% CI Lower	95% CI Upper
Intercept	-23.84	53.07	-0.45	0.65	-127.86	80.18
Percentage of total consumption from Alcohol	0.53	1.06	0.50	0.62	-1.55	2.61
Percentage of total consumption from Animal Fats	0.43	1.06	0.40	0.69	-1.66	2.52
Percentage of total consumption from Cereals	0.37	1.06	0.35	0.73	-1.72	2.45

Term	Estimate	Standard Error	Statistic	P-Value	95% CI Lower	95% CI Upper
Percentage of total consumption from Eggs	1.01	1.08	0.94	0.35	-1.10	3.12
Percentage of total consumption from Seafood	0.24	1.08	0.22	0.83	-1.89	2.36
Percentage of total consumption from Meat	0.32	1.06	0.30	0.76	-1.76	2.41
Percentage of total consumption from Fruits	0.49	1.06	0.46	0.65	-1.60	2.57
Percentage of total consumption from Milk	0.46	1.06	0.43	0.67	-1.62	2.54
Percentage of total consumption from Miscellaneous	0.07	1.10	0.07	0.95	-2.09	2.24
Percentage of total consumption from Offals	-1.09	1.26	-0.87	0.39	-3.57	1.38
Percentage of total consumption from Oilcrops	0.21	1.06	0.20	0.84	-1.86	2.28
Percentage of total consumption from Pulses	0.28	1.06	0.26	0.79	-1.81	2.36
Percentage of total consumption from Spices	0.15	1.12	0.13	0.89	-2.04	2.34
Percentage of total consumption from Starchy Roots	0.36	1.06	0.34	0.74	-1.72	2.44
Percentage of total consumption from Stimulants	1.04	1.08	0.96	0.34	-1.08	3.15
Percentage of total consumption from Sugar/Sweeteners	0.41	1.06	0.38	0.7	-1.68	2.49
Percentage of total consumption from Treenuts	0.81	1.07	0.75	0.45	-1.30	2.91
Percentage of total consumption from Vegetable Oils	0.41	1.06	0.38	0.7	-1.68	2.49
Percentage of total consumption from Vegetables	0.35	1.08	0.32	0.75	-1.77	2.47
Obesity	0.02	0.01	1.90	0.06	0.00	0.05
GDP per capita (in thousands of dollars)	-0.01	0.00	-1.55	0.12	-0.01	0.00
Population (in hundreds of millions)	-0.01	0.05	-0.17	0.86	-0.10	0.08

References

https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset?resource=download&select=Food_Supply_Quantity_kg_Data.csv

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7306972/

https://www.hsph.harvard.edu/nutritionsource/nutrition-and-immunity/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7919160/

https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html

https://www.cdc.gov/healthyweight/index.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8954090/

https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD