Brief Report: Influence of Nutrition on Global COVID-19 Confirmed Cases

David Moon

davidmoon@brown.edu

Brown University

Abstract

Undernourishment has been tied to coronavirus disease 2019 (COVID-19) rates globally. This study examined nutrition-related factors (obesity and percent energy intake of fruits and vegetables) and their relationship with confirmed COVID-19 rates at a country-level basis. Global data on 156 countries were included in a multiple linear regression (MLR) model regressing log-transformed undernourished counts, log-transformed population sizes, and percent energy intake of fruits and vegetables on log-transformed confirmed COVID-19 counts. This study demonstrated that log-transformed undernourished counts and log-transformed population sizes were significantly associated with log-transformed COVID-19 counts. An outstanding finding was that a 1% increase in undernourished count was associated with a 1.22% decrease in confirmed COVID-19 counts; a further study to control for additional confounding factors is warranted.

Keywords: COVID-19, coronavirus, nutrition, healthcare

Introduction

As coronavirus disease 2019 (COVID-19) continues to affect countries globally, methods to protect oneself (besides receiving a vaccine or receiving medical support) continue to be discovered.

Particularly, one such method is maintaining a balanced and nutritious diet. Gombart et al. (2020)[1] discusses how daily adequate amounts of micronutrients (especially vitamin C, vitamin D, and zinc) are necessary to ensure the proper functioning of immune. cells. Moveover, it has been shown that the effect of the virus is mostly on individuals with low immunity, as well as individuals affected with diseases like diabetes, and individuals using any immune-suppressed drug or having past history of major surgeries or severe medical conditions (Budwhar et al. 2020)[2].

Given these links between nutrition and the immune system, it is then important to define what constitutes a healthy diet. It is important to note that diet evolves and is heavily influenced by various social and economic factors (i.e. income, religious practices, geographical factors, etc). However, according to the World Health Organization, there are principles to what constitutes a healthy diet. One such principle is eating at least 400g (five portions) of fruits and vegetables daily. This is echoed by the American Heart Association, with recommendations to fill at least half of an individual's daily intake plate with fruits and vegetables. Slavin et al. (2012)[5] details specific benefits of fruits and vegetables: a supply of dietary fiber (which is linked to lower incidences of cardiovascular disease and obesity), phytochemicals (functioning as antioxidants, phytoestrogens, and anti-inflammatory agents) and various other vitamins and minerals.

The current study then sets out to investigate the relationship between nourishment (on the basis of fruits and vegetables) and COVID-19 confirmed cases. According to Johns Hopkins Center for Systems Science and Engineering (where COVID-19 data is obtained from), a confirmed case constitutes a positive viral test and presumptive positive cases (when a patient has tested positive but results are pending CDC confirmation). This study is a larger part of the current emphasis on nutrition research applied to the COVID-19 pandemic, such as the American Society for Parenteral and Enteral Nutrition (ASPEN) COVID-19 Nutrition Task Force, and beyond.

Methods

Dataset

The relevant dataset comes from Kaggle (dataset found here), which is a cleaned version of four datatsets derived from the following sources:

- Food and Agriculture Organization of the United Nations (FAO) data (specifically *Supply_Food_Data_Description.csv*).
- Population count data for each country from Population Reference Bureau (PRB) website
- Data for COVID-19 confirmed, deaths, recovered and active cases (as of February 2021) from Johns Hopkins Center for Systems Science and Engineering CSSE
- ChooseMyPlate.gov for the USDA Center for Nutrition Policy and Promotion diet intake guideline information

The R script used to clean the source datasets can be found here. Four measures of nutrition are used to generate 4 datasets in Kaggle (percentage of fat intake, percentage of food intake, percentage of energy intake, percentage of protein intake), of which the dataset measuring the percentage of energy intake is used. The basis of examining energy intake comes from the observation that the immune system demands energy (from exogenous sources, such as diet) to develop and mantain its effectiveness (Childs et al. 2019[3]).

The energy intake dataset compares percentages of energy intake from 23 different categories of foods from 170 countries' populations. These 23 categories of foods are obtained from the aforementioned FAO dataset. Each country's obesity rate and undernourished rate is included, as well as the percentage of confirmed COVID-19 cases.

Data Transformation

All rows (countries) with null values were removed, reducing the number of countries from 170 to 153. Undernourished rates with a string value of '<2.5%' were casted exactly to 2.5%. The percentages of confirmed COVID-19 cases, obesity, and undernourished individuals for each

country were mapped to the counts of confirmed COVID-19 cases, obesity, and undernourished individuals (using population sizes) in the regression model. The percentage of energy intake from fruits and vegetables were combined into one joint column. A log transform was applied to Obesity, Undernourished, Confirmed, and Population counts (to normalize their distributions).

Statistical Analysis and Methods

Prior to running the linear regression models, a Wilcoxon rank sum test was run to determine if the distribution of above-average obesity rates was significantly different from the distribution of below-average obesity rates (based on the median of the two distributions). This non-parametric test is used to bypass the assumption that the two populations being compared have a known, Gaussian distribution typical of a two-sample t-test. A one-way analysis of variance (ANOVA) test was run to compare the percent energy intake of fruits and vegetables between countries above and below the global undernourished rate (6.9%). All assumptions of this test were satisfied: independent variable consisted of two or more categorical and independent groups, independence of observations, no significant outliers, dependent variable (percent energy intake) normally distributed for all categories, and homogeneity of variances. Specifically, Levene's test was used to test for equality of variances.

Two different log-log multiple linear regression (MLR) model were run, both including the log-transformed count of confirmed COVID-19 cases as the response variable. The first model (denoted as model 1) was of the following form:

$$\log \text{Y} = \beta_1 log X_{ob} + \beta_2 log X_{und} + \beta_3 log X_{pop} + \beta_4 X_{frt} + \epsilon$$

The second model (denoted as model 2) was of the following form:

$$\log \text{Y} = \beta_1 log X_{und} + \beta_2 log X_{pop} + \beta_3 X_{frt} + \epsilon$$

Where X_{und} represents undernourished count, X_{ob} represents obesity count, X_{pop} represents population size, X_{frt} represents percent energy intake from fruits and vegetables, and Y represents confirmed COVID-19 cases. Each of these variables were measured at the country level. Because percentage variables were mapped to counts, population size must be included

in the model as a confounder (since it is associated with both confirmed counts and undernourished/obesity counts). This process is known as adjustment and can isolate relationships of interest (Pourhoseingholi et al. 2012[6]).

Prior to interpreting the model's results, multicollinearity between predictor variables was tested using variance inflation factor (VIF) values. A threshold of a VIF > 10 is chosen according to Vittinghoff et al. (2012)[9], with 1 being the lowest. All assumptions of the MLR model were validated: linearity, homoscedasticity, independent errors, normality of errors, and the mean of $\epsilon = 0$.

To assess model performance, the adjusted R^2 value is used to determine how well the model fits the response variable. The adjusted value is used to prevent overfitting and account for the number of predictors in the model. Root mean squared error (RMSE) represents the standard deviation of the residuals and and indicates how accurately the model predicts the response. F-tests (at the $\alpha = 0.05$ level) were also included to evaluate whether at least one of the regression coefficients $\neq 0$, which determined whether the proposed relationship was statistically reliable.

Results

The descriptive statistics for each of the five variables under investigation in this dataset is shown in table 1.

Further exploratory data analysis showed that Central African Republic had the highest undernourished rate (59.6%), with multiple countries having an undernourished rate of 2.5% or lower. Samoa had the highest obesity rate (45.5%), and Vietnam had the lowest obesity rate (2.1%). Montenegro had the highest confirmed COVID-19 rate (10.4082%), and Vanuata had the lowest confirmed COVID-19 rate (0.0003115%).

An initial one-way ANOVA test was run comparing percent energy intake of fruits and vegetables between countries below the global median undernourished rate and countries above the global median undernourished rate. The two groups were determined using the median of the undernourished rate present in the data (6.9%). The assumption for equal population variances was satisfied using Levene's test (F-statistic of 3.4994 and p-value of 0.06329). The ANOVA

	Overall (N=156)
FruitsAndVegetables	
Mean (SD)	3.07 (1.61)
Median [Min, Max]	2.87 [0.349, 9.28]
Obesity	
Mean (SD)	18.6 (9.55)
Median [Min, Max]	21.6 [2.10, 45.5]
Undernourished	
Mean (SD)	11.2 (11.7)
Median [Min, Max]	6.90 [2.50, 59.6]
Confirmed	
Mean (SD)	2.07 (2.39)
Median [Min, Max]	1.07 [0.000312, 10.4]
Population	
Mean (SD)	47700000 (163000000)
Median [Min, Max]	10700000 [72000, 1400000000]

Figure 1. Summary statistics of variables of interest.

test yielded significant results with an F-statistic of 6.061 and p-value of 0.0149 (< 0.05) (see table 2). Thus, we conclude there is a statistically significant difference in average percent energy intake of fruits and vegetables between below median undernourished rate countries and above median undernourished rate countries.

```
             Df Sum Sq Mean Sq F value Pr(>F)
level         1   15.1  15.137   6.061 0.0149 *
Residuals   154  384.6   2.497
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2. Summary table of one-way ANOVA test.

Given the heavily right-skewed distribution of undernourished rates and the use of percentage data, the Wilcoxon rank sum test was used to compare the distributions of undernourished rates for countries under/above the global obesity rate of 13% (as of 2016 from the World Health Organization). We conclude with a p-value of $4.42e-14$ that the two populations of undernourished rates have different distributions (with the associated CI for the difference between the two location parameters being [7.8, 14]).

Upon running model 1, multicollinearity was assessed using VIF values (table 3). The log-transformed obesity variable was 11.95 (> 10), and model results were not interpreted due to this violation. However, it is worth noting that the F-test yielded a test statistic value of 57.8 (p-value of $1.66e-29$). Though the log-transformed population variable also had a high VIF value (29.26), this variable must be kept in the model as a confounder. Running model 2 (without the log-transformed obesity variable), yielded the VIF values shown in table 3. All assumptions were satisfied and are shown in plot 4, and the mean of the residuals was $1.984e-17$. Table **??** denotes each explanatory variable's coefficient estimate, as well as p-values and confidence intervals. It should be noted that the percent energy intake of fruits and vegetables variable had an insignificant slope coefficient (p-value of 0.546). The F-test yielded a test statistic of 59.4 (p-value of $1.80e-25$). RMSE = 1.901 and adjusted $R^2 = 0.531$, which means that 53.1% of the variation in the log transformed COVID-19 confirmed count variable is explained by the model.

```
     Obesity      Undernourished          Population FruitsAndVegetables
   11.950529            8.869909           29.264800            1.054000

      Undernourished FruitsAndVegetables          Population
            5.099557            1.053105            4.983414
```

Figure 3. VIF values from model 1 (top) and model 2 (bottom).

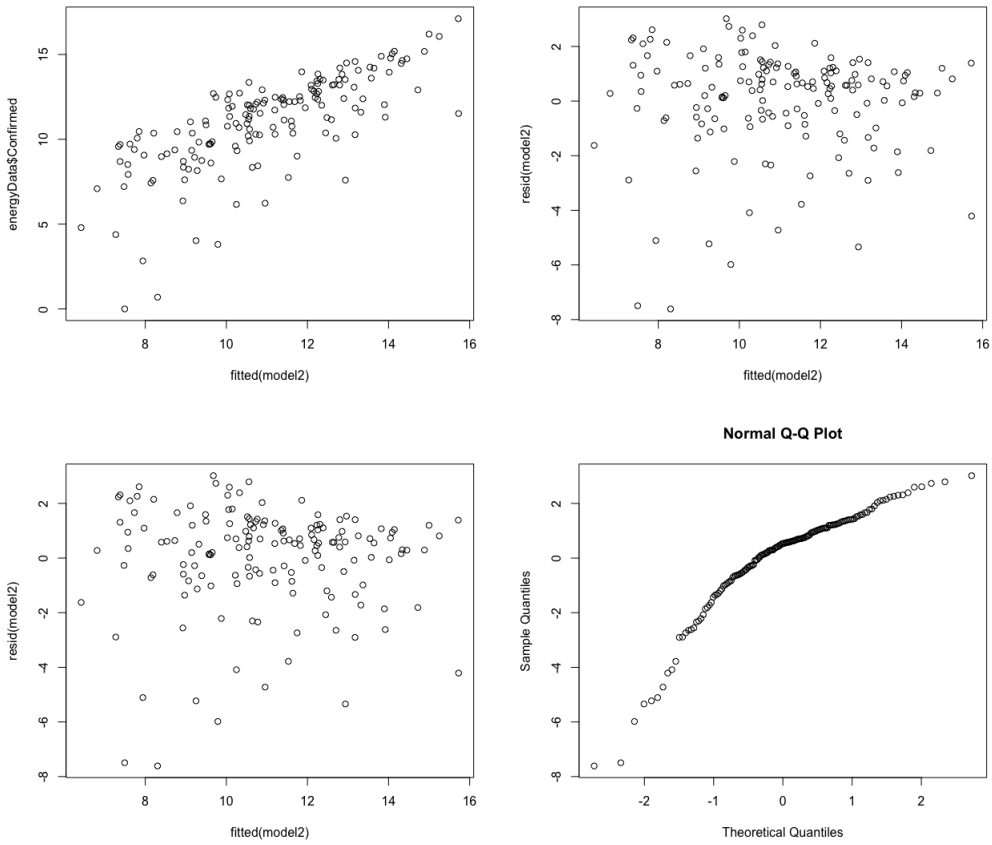

Figure 4. MLR model assumptions: scatterplot for linearity between response variable and fitted values, scatterplot for homoscedasticity between fitted values and model 2 residuals, scatterplot for independence of errors, QQ-norm plot for normality of errors (top-bottom, left-right).

	term	estimate	p.value	conf.low	conf.high
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	-8.59	1.29e- 7	-11.7	-5.53
2	Undernourished	-1.23	1.35e-11	-1.56	-0.896
3	FruitsAndVegetables	0.0599	5.46e- 1	-0.135	0.255
4	Population	2.23	1.33e-22	1.85	2.61

Figure 5. Model 2 summary statistics.

Discussion

The aim of this study was to investigate the relationship between nutrition-related factors and global confirmed COVID-19 cases. A 1% increase in undernourished count is associated with a 1.22% decrease in confirmed COVID-19 counts, and a 1% increase in population size is associated with a 2.24% change in confirmed COVID-19 counts (from model 2). The decrease in confirmed COVID-19 counts associated with an increase in undernourished rates is a potential limitation in this model, and the increase in confirmed COVID-19 counts associated with an increase in population size makes sense in this situation (more people means more potential for infection). Further analyses could involve modeling active or recovered COVID-19 rates as the response variable (both of which are included in the dataset).

The R-squared value (0.531) is not surprising as there are many other health-related, social, econonomic, and environmental factors that are relevant predictors in confirmed COVID-19 cases. While the proposed model may not be ideal for predictive analyses, this investigation (and the F-test) ascertains that the explanatory variables improve the model's fit, and future studies could explore other nutrition-related variables that could be included in the model. Butler et al. (2020)[7] specifically explores the impact of the western diet (which largely consists of saturated fats, sugars, refined carbohydrates) on the prevalence of obesity and type 2 diabetes, which in turn serve as two risk factors for COVID-19.

Existing literature and implications from this investigation suggest that there is a understudied link between diet and COVID-19 that could be worth further investigating. Namely, Belanger et al. (2020)[8] underscores the disproportionate impact of COVID-19 on minority groups (including Black, Latinx, and Native Americans) that correlate closely with nutrition disparities among these groups: while overall prevalence of obesity among U.S. adults is 42.4%, Black (49.6%), Native American (48.1%), and Latinx (44.8%) adults are disproportionately affected. This sets a larger context for systemic change involving increased awareness in public health policies and social support services. External, nutrition-related forces such as a lack of access to healthy foods and higher rates of food insecurity originate from historical injustices built by racism in the healthcare system. Moreover, Mehta (2020)[4] stresses the need to send

strong messaging on the importance nutrition at the government level and provide financial and practical aid to disadvantaged groups. Despite model results (specifically in regards to confirmed COVID-19 rates and undernourished rates), there is more that needs to be done in terms of putting nutritional care at the forefront of healthcare models, especially for vulnerable populations.

Acknowledgements

Special thanks to Professor Dunsiger for her advising regarding statistical techniques and general academic guidance.

References

[1] Gombart, A., et al. *A Review of Micronutrients and the Immune System-Working in Harmony to Reduce the Risk of Infection.* Nutrients, U.S. National Library of Medicine, 2020, pubmed.ncbi.nlm.nih.gov/31963293/.

[2] Budhwar, S., et al. *A Rapid Advice Guideline for the Prevention of Novel Coronavirus Through Nutritional Intervention.* Current Nutrition Reports, Springer US, 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7308604/.

[3] Childs, Caroline E, et al. *Diet and Immune Function.* Nutrients, MDPI, 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6723551//.

[4] Mehta, Shameer. *Nutritional Status and COVID-19: an Opportunity for Lasting Change?.* Clinical Medicine (London, England), Royal College of Physicians, 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7354054//.

[5] Slavin, Joanne L, and Beate Lloyd. *Health Benefits of Fruits and Vegetables.* Advances in Nutrition (Bethesda, Md.), American Society for Nutrition, 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3649719/.

[6] Pourhoseingholi, Mohamad Amin, et al. *How to Control Confounding Effects by Statistical Analysis.* Gastroenterology and Hepatology from Bed to Bench, Research Institute for Gastroenterology and Liver Diseases, 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC4017459/.

[7] Butler, Michael J, and Ruth M Barrientos. *The Impact of Nutrition on COVID-19 Susceptibility and Long-Term Consequences*. Brain, Behavior, and Immunity, Elsevier Inc., 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7165103/.

[8] Belanger, Matthew J., et al. *Covid-19 and Disparities in Nutrition and Obesity*. New England Journal of Medicine, 2021, www.nejm.org/doi/full/10.1056/NEJMp2021264.

[9] Vittinghoff E, et al. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. 2012 edition. Springer.