# Does your lifestyle make you more vulnerable to COVID?

Authors: *Jose Repettoparede*; *Adegboyega During*; *Sakshi Joshi; Shikha Mohindra*; *Srinidhi Lakshminarayanan*; Business Analytics, *Eli Broad College of Business*, East Lansing, MI-48824, USA

## 1 Introduction

The current global pandemic and its effects on humanity has raised all kinds of speculations and conclusions from various quarters. Likewise, lifestyle choices and its effect on human health is a common topic that has been studied with respect to discussions on Health. While some are backed up with scientific evidence others are mere human thoughts. This project is largely based on lending a voice to the present situation and seeks to provide an overview of some evidence to further investigate how our lifestyle choices could dictate the impact of COVID-19. . The paper specifically investigates impact of various lifestyle factors on COVID-19 death rate on a county basis and uses data retrieved from various sources such as the Health Data and the Michigan County Health Rankings and Roadmaps. The variables of consideration are a combination of quantitative and qualitative variables to measure their impacts on the effects of contacting the virus.

The literature on the COVID-19 disease is very sparse due to its novelty and it started just five months ago. At this time, it is not yet known with certainty how to treat the virus and how to prevent people from continuing to infect themselves once everything reaches normality. Similarly, there have been speculations regarding which people are more likely to contract the disease and which factors most likely influence the person to aggravate once they are infected. Generally, it is perceived that people who smoke and those who do not take care of themselves in their diet are at higher risk. This study focuses on four major states in the United States to determine the relationship between the health indices at county level with the number of COVID-19 related deaths.

Data was retrieved for various health indicators from the County Health Ranking website while COVID-19 incidence datasets were retrieved from the worldometers website. This data was retrieved at county level for 147 counties making up four states in the United States. The datasets has exhaustive information on every state of the country and their counties ranked according to their happiness indexes.

To study this effect, we fit a multivariate mixed linear regression model to establish the relationship between the various variables and death at county level. We applied a stepwise regression using both forward and backward selections to select the best set of variables that predicts death outcome due to COVID-19 with the best accuracy.

The results suggested an increase in number of COVID-19 related deaths with variables such as HIV prevalence rate, population, percentage of people who cannot speak English while the average daily PM2.5 leads to a reduction in COVID-19 related death at a very high level of significance both individually and collectively. These results are statistically significant and robust to secondary and sensitivity analyses. Further, given the county data, we can predict the COVID-19 death rate for different counties with the best accuracy in some states.

## 2 Data Sources

There are many factors that influence how long and how well we live, from the quality of our home and the safety of our neighborhoods to the opportunities we have for good jobs and education. The coronavirus epidemic has widened existing social, geographic, and economic challenges

in communities across the country.
The data for this work consists of data from two different sources. We retrieved County Health Rankings that had a myriad of information from the website[1] which was used for this project. This has exhaustive information on every state of the country and their counties ranked according to their happiness index. With this data, we set to explore what all factors cause the most deaths in the counties. Rank data are helpful in providing local context on factors that impact health. Ranking factors such as housing, access to medical care, and unemployment provide a better understanding of community context and the places and people that could be most affected by the coronavirus epidemic. These rankings are a call to action about what can be done to improve community conditions so that an inclusive and equitable recovery for all is possible.
The County Health Rankings are based on a model of community health that emphasizes the many factors that influence how long and how well people live. The Rankings use more than 30 measures that help communities understand how healthy their residents are (health outcomes) and what will impact their health in the future. Health Outcomes are mainly a factor of Policies and Programs on one hand and Health Factors on the other. These indices are further summarized in the Figure 1 below:

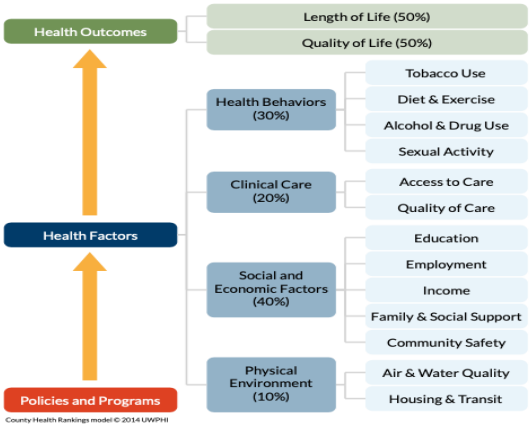



Figure 1: Measures of County Health Rankings.

The COVID data was retrieved from the worldometers website[2] which is the one-stop shop for COVID information throughout the world. This data set is comprehensive as it gives information that pertains to the number of tests conducted, number of incidences, number of deaths, number of recovery amongst others. For this study, we compare three states with Michigan that are very comparable in terms of deaths/population namely - Louisiana, District of Columbia, and Rhode Island. The COVID summary for this data is presented below in Table 1:

| State | Total Cases | Total Deaths | Active Cases | Cases/ 1M Pop. | Deaths/ 1M Pop | Total Tests | Tests/ 1M Pop. |
|---|---|---|---|---|---|---|---|
| Louisiana | 32,662 | 2,381 | 7,673 | 7,026 | 512 | 237,904 | 51,175 |
| D.C | 6,584 | 350 | 5,300 | 9,329 | 496 | 31,685 | 44,857 |
| Michigan | 48,391 | 4,714 | 20,991 | 4,845 | 472 | 329,639 | 33,007 |
| Rhode Island | 11,835 | 462 | 10,487 | 11,172 | 436 | 97,922 | 92,435 |

Table 1: Summary of COVID cases for States being considered

## 2.1 Identification Strategy

To estimate the COVID-19 related death in the county, we estimated a multivariate linear regression. This method has the capability of predicting the death variable by capturing the linear relationship

between various health choices and how they impact death. From the Corona dataset, we compute the standardized death which is computed as:

$$StandardizedDeaths = \frac{Deaths}{((positivecases/population) * 1000)} \quad (1)$$

For our estimation, let Dij represent death due to COVID-19 in county (i) and state (j) and the health choices be represented by a vector of X s with each X representing each health choice. The basic estimation equation is then

$$D_{ij} = \beta_0 + \sum_{n=1}^{\infty} \beta_n X_n + \epsilon_{ij} \quad (2)$$

where b0 is a constant term; bn captures the linear relationship between the health choice (Xn) and death; and eij is the error term. The main parameter of interest is the bn which captures the relationship between each health choice and death. Under the null hypothesis that each bn = 0, COVID-19 related death is not affected by any health choices. If any health choice impacts COVID-19 related death, it is expected that the coefficient bn would not be zero and will be statistically significant.

## 3 Exploratory Data Analysis

We selected states which have similar death rates as Michigan. These states were Louisiana, Rhode Island and District of Columbia. Out of these, District of Columbia is an outlier. Hence, we are left with 3 states. The distribution of deaths in these states are shown in the graph below.

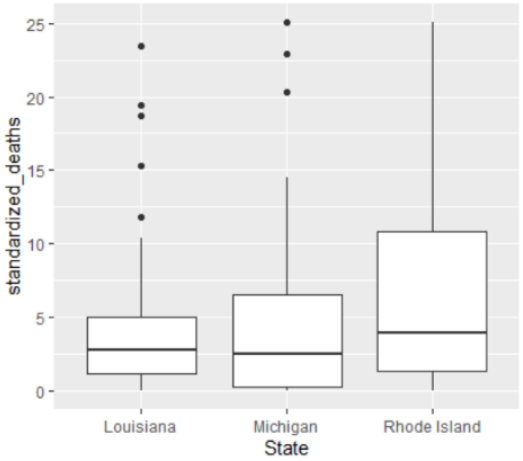



Figure 2: Death trend state wise

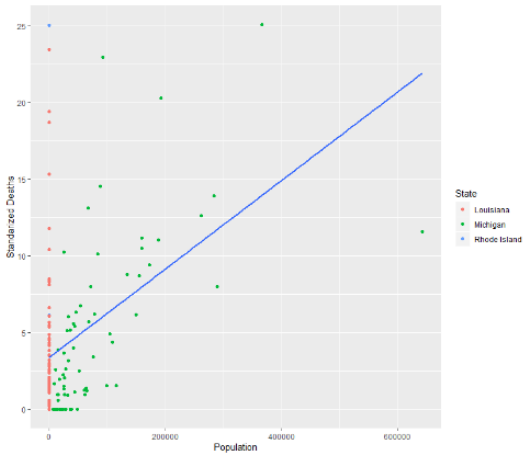



Figure 3: Population distribution

In the following figure the population is plotted against the output variable standardized deaths. The number of deaths in a county is directly proportional to the population of the county. This relationship can be explained with the social distancing norm. With high population, people will follow less social distancing and will be in contact with each other more compared to people in less populated counties. Hence, with increase in population, number of deaths due to COVID increases.
Next, we investigated the environmental factors that would contribute to the impact of COVID. Here we plotted percentage of rural population, average traffic on roadways and air quality.

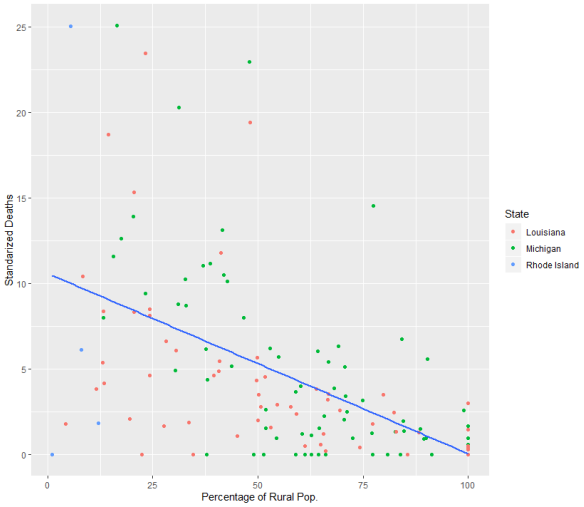



Figure 4: Rural population distribution

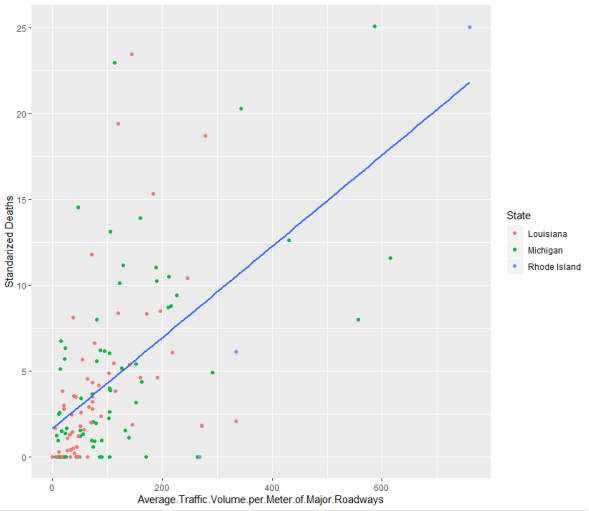



Figure 5: Average traffic distribution

For percentage of rural population in a county, deaths in a county is inversely dependent on the percentage of rural population in a county. This can be justified by multiple factors as rural population will have more physical activity, immunity, fresh air, less population, etc. which will make these counties less susceptible to severe impact of COVID. Next, the traffic on roadways in a county is highly impactful on number of covid deaths as more traffic means more population, less air quality and less physical inactivity.

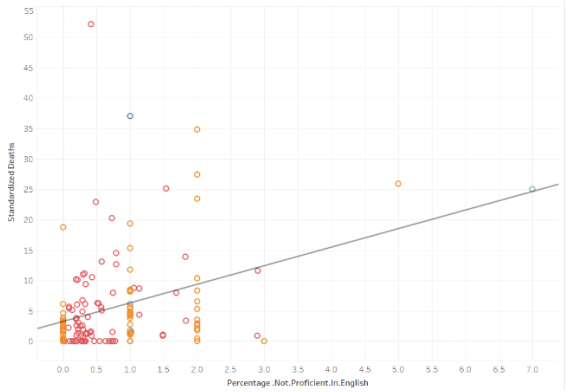



Figure 5: Language Proficiency

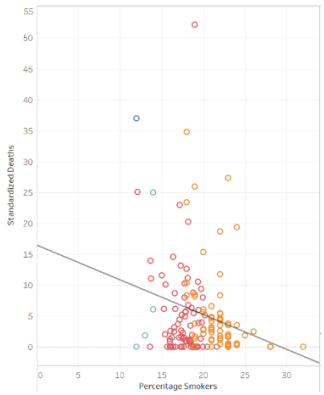



Figure 6: Precentage Smokers

Finally we looked into some individual characteristics that can impact the deaths due to COVID. Surprisingly, English language proficiency of adults has an impact on the number of deaths due to COVID. This can be explained as people with less proficiency will have hard time communicating with the doctors which can affect their medical care. Also, Percentage of smokers in a county is inversely related to the number of deaths. This is interesting as it is opposite of the normal belief that higher smoking rate will increase the chance of covid.

## 4 Methodology

### 4.1 Data Cleaning

Step 1: Data collection
We started our data with 66 health indicators from 147 countries through 4 states. These included data from policies and program, Health factors and health indicators.

Step 2: Multicollinearity and missing value check
We tried finding out the Variance Inflation Factor (VIF) of the all the variables and removed the ones that show multi-collinearity between them. We found percentage of adults with frequent physical distress and percentage of adults with food insufficiency to have high VIF values. Since, these two factors are dependent on a lot of other factors, we eliminate them. Surprisingly the "Percentage_.65.and.over" and "Percentage Smokers" was removed from the dataset due to multicollinearity.
This helped us reduce the number of indicators from 66 to 35 factors.

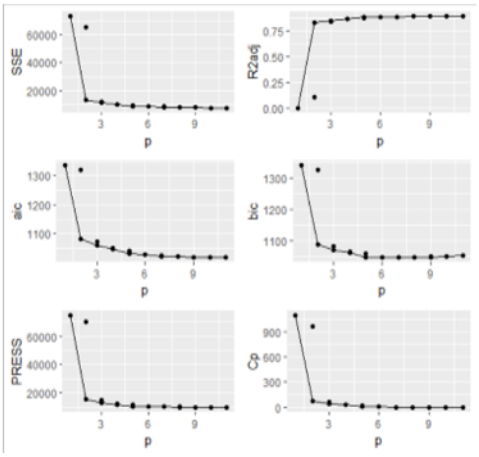
Step 3: Collinearity:
We used the correlation function to reduce the number of indicators from 35 to 32.
For example, the percentage of adults with obesity was highly correlated with adults with physical inactivity with 0.58. So, we went ahead to remove the percentage of adults with obesity.

## 4.2 Feature Selection Criterion

For this work, we estimated a linear regression to capture the average treatment effects of county wise lifestyle indices on COVID-19 related death. In addition, we predicted for our counties to see how our results differ from the actual data. To arrive at the final estimates of the regressions, this work conducted stepwise regression using the choice predictive variables to come up with the best model for prediction. In conducting the stepwise regression, this work considered both the forward and backward selection criterion. The selection criteria from each model are all summarized and presented in Table 2 below from which the best model is selected.
The stepwise regression and the various indices are summarized into Table 2. It is observed that as the number of variables increased, the precision power gets better up to the point where it gets saturated after which there is not much significant difference adding any extra variable.
In this case as witnessed in the summarized result, we settled for a six variable model inclusive of the intercept. This also helps reduced the level of complexity of our model. The variables that best predict death in our model include: Population, Average Daily PM2.5, HIV prevalence rate, Language, and Females.

| Sno. | p | R2 | SSE | R2adj | AIC | BIC | PRESS | Cp |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.000000 | 73544.794 | 0.000000 | 1334.805 | 1340.786 | 74555.707 | 1089.4928 |
| 2 | 2 | 0.825046 | 12866.971 | 0.823839 | 1080.550 | 1089.521 | 15412.207 | 72.9797 |
| 3 | 2 | 0.109908 | 65461.652 | 0.103769 | 1319.690 | 1328.661 | 69801.141 | 955.8125 |
| 4 | 3 | 0.851029 | 10956.059 | 0.848960 | 1058.917 | 1070.878 | 13180.933 | 42.9039 |
| 5 | 3 | 0.836904 | 11994.893 | 0.834638 | 1072.233 | 1084.195 | 14993.773 | 60.3414 |
| 6 | 4 | 0.864624 | 9956.201 | 0.861784 | 1046.849 | 1061.801 | 11852.330 | 28.1207 |
| 7 | 4 | 0.862455 | 10115.740 | 0.859569 | 1049.186 | 1064.138 | 12287.408 | 30.7987 |
| 8 | 5 | 0.880712 | 8773.041 | 0.877351 | 1030.252 | 1048.194 | 10497.532 | 10.2607 |
| 9 | 5 | 0.872442 | 9381.252 | 0.868849 | 1040.105 | 1058.048 | 11570.668 | 20.4699 |
| 10 | 6 | 0.885025 | 8455.849 | 0.880947 | 1026.839 | 1047.772 | 10483.365 | 6.9364 |
| 11 | 6 | 0.884304 | 8508.826 | 0.880202 | 1027.757 | 1048.690 | 10315.274 | 7.8257 |
| 12 | 7 | 0.889951 | 8093.535 | 0.885235 | 1022.401 | 1046.324 | 10206.066 | 2.8548 |
| 13 | 7 | 0.888275 | 8216.761 | 0.883487 | 1024.622 | 1048.546 | 10355.737 | 4.9232 |
| 14 | 8 | 0.892366 | 7915.909 | 0.886946 | 1021.139 | 1048.053 | 9998.290 | 1.8732 |
| 15 | 8 | 0.892149 | 7931.914 | 0.886717 | 1021.436 | 1048.350 | 10166.849 | 2.1419 |
| 16 | 9 | 0.895800 | 7663.403 | 0.889759 | 1018.373 | 1048.278 | 10010.808 | -0.3652 |
| 17 | 9 | 0.894864 | 7732.181 | 0.888770 | 1019.687 | 1049.591 | 9960.208 | 0.7892 |
| 18 | 10 | 0.897780 | 7517.755 | 0.891065 | 1017.553 | 1050.447 | 9702.139 | -0.8100 |
| 19 | 10 | 0.897589 | 7531.827 | 0.890861 | 1017.828 | 1050.722 | 9995.169 | -0.5738 |
| 20 | 11 | 0.899689 | 7377.368 | 0.892313 | 1016.782 | 1052.667 | 9686.962 | -1.1665 |
| 21 | 11 | 0.899097 | 7420.871 | 0.891678 | 1017.646 | 1053.531 | 9668.682 | -0.4363 |

Table 2: Selection Criterion Tables




| | **Variable** |
|---|---|
| $Y$ | Standardized Deaths |
| $X_1$ | Population |
| $X_2$ | Average Daily PM 2.5 |
| $X_3$ | HIV Prevalence Rate |
| $X_4$ | Female Population (%) |
| $X_5$ | Population Not Proficient in English (%) |
| $X_6$ | Drive Alone to Work (%) |

Figure 7: Output of regsubset and important features

## 4.3 Outlier Detection

Once we have obtained the final regression model, we analyze whether the assumptions of this statistical model are fulfilled. For this we can look at the residual and Q-Q Plot plots to verify that the assumptions of:

- Linearity
- Homoscedasticity (constant variance of the residuals)
- Independence
- Normality (outcome variable is normally distributed or any fixed value of the predictive variables)

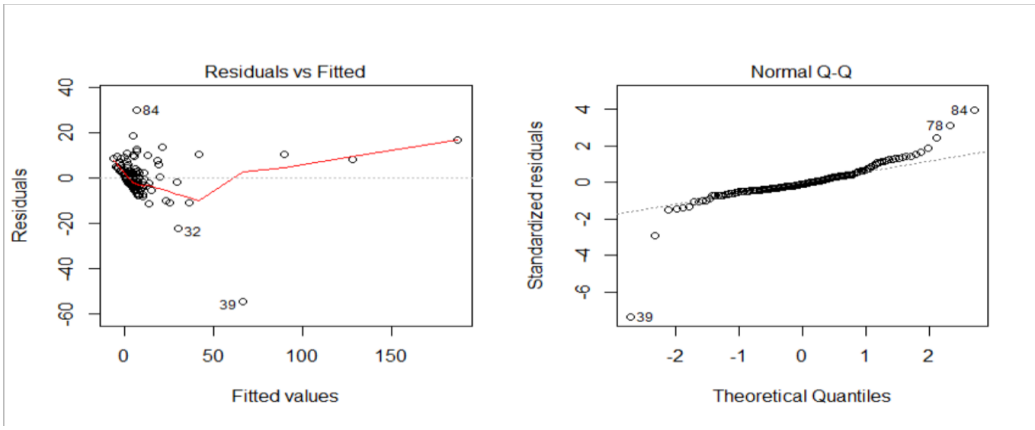



Figure 8: Residual vs Fitted and Normal QQ Plot.

Since we are fitting the standardized deaths per country information to a linear regression, we assume that the first assumption is true.
The red line is not completely horizontal but there is no obvious pattern of the residuals. No signs of Heteroscedasticity are observed because the residuals do not increase when the fitted values also increase. Also, we assume that there is no multicollinearity. This occurs when the predictors are highly correlated with each other. For this, prior to the analysis and implementation of the model, a multicollinearity analysis was performed to eliminate those variables that were related to each other and keep those that were the most significant to build the model. Finally, the normality assumption requires that the residuals should be normally distributed. By looking at the Q-Q-Plot, we can see that this assumption is not met, and some improvement must be done to get the final regression model. What we can identify from this analysis is that there are certain outliers that do not contribute to a good result and model precision. The graph identifies observations 84, 32 and 39 as outliers, so it has been decided to withdraw them to obtain a new model. These rows correspond to the counties of Kent, Washington, and Ingham.
From the elimination of the outliers, and using the same independent variables for the linear regression, the improvement of the results can be seen in the following table:

| Model | p | R2 | SSE | R2adj | AIC | BIC | PRESS |
|---|---|---|---|---|---|---|---|
| With Outliers | 6 | 0.885025 | 8455.849 | 0.880947 | 1026.839 | 1047.772 | 10483.37 |
| Without Outliers | 6 | 0.950043 | 3632.71 | 0.948233 | 887.4749 | 908.2636 | 4102.604 |

Table 3: Model with and without outlier

# 5 Model and Interpretation

Two graphical representations of the model and variables selected in this case the residuals plotted against the fitted values and the standardized residuals versus the theoretical quantiles in a normal plot indicated the presence of a few outliers. To address this challenge, these outliers are dropped and the model recomputed. The results from the initial model with the outliers and the ones without the outliers are presented in Tables 3 respectively.

## 5.1 Our Models

Table 3 below presents the results from our selected model. This result indicates that the population of a county influences the level of death due to COVID-19. Pre-existing conditions such as
HIV is also linked to increased death due to COVID-19. This could be because the presence of

pre-existing conditions in patients could have compromised their immune system as such, being infected with COVID-19 could further compromise their health therefore, leading to higher death. For example, an increase of a unit in the HIV prevalence rate would lead to about 0.018 death at county level.
More so, the population of a county is a major determinant of COVID-19 related death.
Our result suggested an increase in the population of a county by 10000 people would suggest an additional death due to COVID-19.

| Coefficients: | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | -40.114541 | 14.561947 | -2.755 | 0.006649** |
| Population | 0.000103 | 0.000003 | 29.683 | < 0.0000000000000002*** |
| Average.Daily.PM2.5 | -2.293218 | 0.484360 | -4.735 | 0.00000528786*** |
| HIV.Prevalence.Rate | 0.018363 | 0.002967 | 6.188 | 0.00000000624*** |
| %age_.Not.Proficient.in.English | 1.634994 | 0.710926 | 2.300 | 0.022927* |
| %age_Female | 1.148105 | 0.292271 | 3.928 | 0.000133*** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.744 on 141 degrees of freedom
Multiple R-squared: 0.885, Adjusted R-squared: 0.8809
F-statistic: 217.1 on 5 and 141 DF, p-value: < 0.00000000000000022

Table 4: Model with Outliers

In similar instance, the percentage of people living in a province who are not proficient in the English Language is seen to have a positive relationship with death due to COVID-19. This could be due to the inability to express their health condition to medical practitioners which makes the management difficult. A person who doesn't speak English has 63% chance of resulting to death compared to someone who speaks English. Likewise, it might be selecting the impact of ethnicity on COVID-19 related death. This would be further reviewed.
In addition, COVID-19 related death is more prominent in the Females when compared to the Males. This could suggest that men have some advantage in their immune system to fight the disease compared to females. A female has 15% chances of COVID-19 leading to death compared to males. This would need further studies to ascertain this claim.
Lastly, average daily PM2.5 suggested an inverse relationship between the particles inhaled and COVID 19 related death. Following the outbreak of COVID-19, there's been some level of fumigation going on at different levels which could suggest that people in provinces that do more outdoor fumigation could have been inhaling particles that fights the virus responsible for the disease. This is contrary to the result obtained by a similar research.
Following the presence of outliers in this model, we further dropped those variables and re-present our model below.

| Coefficients: | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | -38.666582 | 19.467413 | -1.986 | 0.04941* |
| Population | 0.000104 | 0.000004 | 27.326 | < 0.0000000000000002*** |
| Average.Daily.PM2.5 | -2.913660 | 0.601548 | -4.844 | 0.000004048*** |
| HIV.Prevalence.Rate | 0.017287 | 0.003340 | 5.175 | 0.000000988*** |
| %age_.Not.Proficient.in.English | 1.589823 | 0.806661 | 1.971 | 0.05116. |
| %age_Female | 1.250857 | 0.372357 | 3.359 | 0.00106** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.329 on 114 degrees of freedom
Multiple R-squared: 0.8889, Adjusted R-squared: 0.884
F-statistic: 182.4 on 5 and 114 DF, p-value: < 0.00000000000000022

Table 5: Model without Outliers

Table 5 below presents the results from the revised model. The results are very similar to that with the outliers in terms of magnitude and sizes. There are improvements in the overall performance of the model in terms of an improved R2 and adjusted R2. This implies a much a better predictability power of our model.

## 6 Assessing prediction accuracy of the models

In order to assess how well the model is predicting the standardized deaths of every county of the database, we calculated the MSE (Mean Squared Error) and the RMSE (Root Mean Squared Error) for the models described in Table 2 and 3 of the paper. Let us remember that the only difference between the two models is that the second one is modelled without the three observations identified as outliers. The RSME measures the model prediction error and communicates the average difference between the observed known values of the standardized deaths and the predicted value of each model. It is summarized in the following table the MSE and RMSE values for both models:

| Model | MSE | RMSE |
|---|---|---|
| With Outliers | 28.3 | 5.32 |
| Without Outliers | 25.2 | 5.02 |

Table 6: Model performance

As it can be seen, the second model has better prediction measures because it is not biased by the presence of outliers. The following table shows a sample of observations with their respective predictions. In addition, a prediction 95% confidence interval is calculated and the number of observations whose actual value falls within the confidence interval is verified.

| County | Actual standardized deaths | Predicted standardized deaths | 95% confidence interval of Standardized deaths | Lies within interval? |
|---|---|---|---|---|
| Avoyelles | 3.815653 | 4.1712825 | -6.1000858 to 14.44265 | Yes |
| St. Tammany | 23.434483 | 5.6338282 | -4.6295783 to 15.89723 | No |
| Pointe Coupee | 2.76975 | 5.0704806 | -5.150866 to 15.29183 | Yes |
| Washington | 3.531819 | 9.8335027 | -0.4501746 to 20.11718 | Yes |
| Jefferson | 25.905344 | 20.4209588 | 9.4692741 to 31.37264 | Yes |
| Oscoda | 1.6554 | 3.3123419 | -6.9243923 to 13.54908 | Yes |
| Menominee | 0 | 0.8362333 | -9.3954862 to 11.06795 | Yes |
| Barry | 1.251187 | 0.1576005 | -10.1511592 to 10.46636 | Yes |
| Lapeer | 14.533284 | 5.7495812 | -4.4886333 to 15.9878 | Yes |
| Ottawa | 13.908881 | 28.6412404 | 18.2945492 to 38.98793 | No |

Table 7: Prediction Result

After applying this procedure to all observations, it is verified how many observations fall within this range. Of the 144 observations, 135 falls inside the confidence interval.

## 7 Conclusion

This paper investigated the impact of lifestyle choices on COVID-19 related death at the county level and provides empirical evidence that suggest how these choices impact death. The results are quite interesting as it shows that contrary to the intuitive belief of some choices such as smoking and its relative impact on health outcome. Of the many lifestyle choices, our stepwise selection picked up the important variables that predicts COVID-19 related death. Identifying these effects rely on the independence between the variables of interest, which gave me the opportunity of studying their impact on average COVID-19 related death.

The findings revealed that the prevalence of HIV, population, inability to communicate in English language, and a bias against females are important variables in predicting directly COVID- 19 related death. An increase in any of these variables at the county level implies more death incidences. Individually and collectively, these variables were found to be very significant at various levels of significance up to 90%.

Conversely, the Average Daily PM2.5 variable has an inverse relationship with respect to death related to COVID-19. The conclusion from this result suggests that the impact of this variable leads to a reduced COVID-19 death. This effect may have been due to the sudden arrival of the disease which lead to a higher level of fumigation which could mean that inhaling these particles improved their health outcomes with respect to fighting the Corona Virus. This contrast with its effect in a related paper thus, the need to be further investigated.

We further used our results in predicting the likelihood of death in the various counties in our dataset and the outcome suggest that the model is able to predict correctly at least 80% of the times at a 95% confidence interval.

The results from this research cannot be easily extrapolated to other countries for some reasons such as the availability of data and differences in governmental structure among countries. Overall, while the existing literature on the impact of health choice on life outcome is vast, there are only a few empirical works have examined this impact at county levels and most importantly with respect to pandemics in general and COVID-19. These findings enrich the growing literature on the impact of health choices on life outcomes. The emphasis on COVID-19 and this studies in general at a county level amplifies the evidence on this topic.

Despite the inherent limitations of the ecological study design, our results underscore the importance of continuing to understand various factors including existing air pollution regulations and other lifestyle factors to protect human health both during and after the COVID-19 crisis.

The data and code are publicly available so our analyses can be updated routinely.

## 8 References


1. https://www.worldometers.info/coronavirus/
2. County Health Rankings and roadmaps, Building a culture of health county by county, a Robert Wood Jonson Foundation Program;https://www.countyhealthrankings.org
3. A national study on long term exposure to air pollution and COVID-19 mortality in the United States:https://projects.iq.harvard.edu/covid-pm