COVID-19 and County Health

STAT 471 Final Project

Seokyoon Chang, Jessica Mixon, & Greta Maayan Waldman

May 2, 2021

Executive Summary

Despite the growing body of research focused on understanding the progression of COVID-19, there is still much to be learned about population-level health factors that make some populations more susceptible to COVID-19 spread, case severity, and mortality. Hence, for our final project, we decided to look into various measures of health across US counties and analyze which factors were most predictive of case fatality rate throughout 2020, the first year of the COVID-19 pandemic. While only analyzing 2020 COVID-19 data only inherently limits our results, we also note 2020 data may actually be most likely to reflect differences in healthcare quality across counties, as it was in the early months of the pandemic that many health systems faced unique issues of resource scarcity.

Our dataset pulled data from two sources: a New York Times time series dataset which includes cumulative counts of COVID-19 cases and deaths at the county level, and data from County Health Rankings & Roadmaps, a program focused on collecting county-level data on a variety of health determinants. For the latter data source, we pulled 2019 or older data from various datasets on their website, as many key variables we wanted to study are still missing for 2020 and we assume that most county-level health determinants stayed very similar from 2019 to 2020. Our explanatory variables span the four main categories of health factors that the program identifies: health behaviors (e.g., smoking, sexual activity), clinical care (e.g., flu vaccine rate), social and economic factors (e.g., unemployment rate), and physical environment (e.g., degree of air pollution). Our primary response variable of interest was deaths per cases, which we created by dividing county-level deaths by count-level cases.

Before exploring our data or running any analyses, we split our data into a training dataset and a test dataset, with the test dataset reserved for assessing and comparing model performance. Then, we explored our data to check for normality assumptions necessary for linear regression and to assess correlations between variables and between variables and the response. In order to build an optimal predictive model, we then built six different cross-validated models: ordinary least squares, ridge regression, LASSO regression, elastic net regression, random forest, and boosting. Of the regression models, elastic net had the lowest test error and of the tree-based models, the boosted model had the lowest test error (as well as the lowest test error overall).

Interestingly, we found that the boosted and elastic net regression both pointed to similar types of variables as the strongest predictors of deaths per cases. Specifically, our optimal boosted model revealed that variables related to residential segregation and unemployment emerged as the most significant predictors, revealing that structural economic and health access inequalities were more predictive of COVID-19 deaths per cases than other variables. We hope that this analysis can inform policies aimed at improving health outcome determinants, both in the context of COVID-19 and more generally going forward.

Introduction

Coronavirus disease (COVID-19) has had a devastating global impact, with a cumulative total of 149,987,772 confirmed cases and 3,157,594 deaths worldwide as of April 28, 2021.[1] About a fifth of these cases have been in the United States, with a recent count of 32,551,440 cases and 582,668 deaths.[2] With these staggering numbers still increasing despite recent large-scale vaccine rollouts, it is of vital importance to utilize various data sources to understand both the progression of COVID-19 thus far as well as the highest risk factors for contracting COVID-19. Furthermore, a thorough analysis of COVID-19 rates and predictive factors may help inform strategies to improve public health policies that could mitigate the negative impact of a future pandemic, which many scientists say is not a matter of if but of when.[3]

Past research has shown that infectious diseases are influenced by a variety of factors. Obesity, for instance, is associated with a higher likelihood of contracting influenza A, and seasonal temperature changes have shown to be predictive of the 2003 severe acute respiratory syndrome (SARS).[4] The CDC is currently in the process of identifying potential risk factors for severe COVID-19 illness,[5] and some that have already been identified include heart disease, diabetes, and pregnancy.[6] Yet despite these efforts, there is still much to be learned. Specifically, there is still insufficient research to explain the differences in COVID-19 susceptibility and mortality that exist not just on the individual level but also on broader population levels.

Given our knowledge of the capacity for a variety of factors to influence infectious disease spread as well as the fact that different counties in the US have differing levels of baseline health factors, we sought to investigate how rates of COVID-19 cases and deaths across the US are affected by various measures of community health. Specifically, we were interested in which kinds of factors (e.g., clinical, behavioral, health)—and which specific variables—are most predictive of deaths per cases (also known as case fatality rate). We hope that our analysis will contribute to the growing body of research on COVID-19 risk factors by expanding our understanding of COVID-19 and supporting efforts to mitigate the risk of future pandemics. Our results also shed light on the importance of analyzing social determinants of health in efforts to improve health outcomes.

---

[1] Coronavirus Cases: Worldometer. (n.d.). https://www.worldometers.info/coronavirus/.
[2] Ibid.
[3] Robbins, J. (2021, January 4). Heading Off the Next Pandemic. Kaiser Health News. https://khn.org/news/infectious-disease-scientists-preventing-next-pandemic/.
[4] Tian, T., Zhang, J., Hu, L., Jiang, Y., Duan, C., Li, Z., ... & Zhang, H. (2021). Risk factors associated with mortality of COVID-19 in 3125 counties of the United States. Infectious diseases of poverty, 10(1), 1-8.
[5] Centers for Disease Control and Prevention. (n.d.). Assessing Risk Factors for Severe COVID-19 Illness. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html.
[6] Centers for Disease Control and Prevention. (n.d.). Certain Medical Conditions and Risk for Severe COVID-19 Illness. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.

Data Description and Exploration

Data Sources

Our dataset merged data from two sources: a dataset on US county-level health and a COVID-19 dataset that includes average cases and deaths per county. Each data source includes data from January 2020 to December 2020.

The data regarding the health status of each county comes from County Health Rankings & Roadmaps, a program founded by the Robert Wood Johnson Foundation in collaboration with the University of Wisconsin Population Health Institute.[7] The program was designed to support community leaders in fostering equitable health outcomes by raising awareness about the variety of factors that influence length and quality of life, including economic and social factors. Specifically, the program considers all of the counties in the United States and includes measures of health behaviors, clinical care, social and economic factors, and the physical environment. In order to ensure a wide variety of explanatory variables in our dataset across the aforementioned categories, we pulled variables from the compiled dataset available on the program's website and merged these variables into our final dataset.[8]

The other dataset we drew from is the New York Times COVID-19 tracking dataset that includes cumulative COVID-19 cases and deaths in the United States at the county level.[9] This time series dataset, compiled from state and local governments and health departments, includes data beginning from the first reported coronavirus case in Washington on Jan. 21, 2020. Times notes that because of the widespread shortage of testing in the first few months of the pandemic, the true COVID-19 prevalence—for those few months especially—may not be accurately represented in the data.

Data Cleaning Process

Our central task in the data cleaning phase of the project was merging the data from the two sources described above. Both data sources provided their respective data on a county level (classified as a 5-digit FIPS code). For each county in the New York Times COVID-19 dataset, the numbers of deaths and cases of COVID-19 were aggregated across 2020. We merged the two datasets on the county level, and then calculated the explanatory variables in accordance with the County Health Rankings Data Documentation. For the most part, the latter step consisted of dividing raw counts of a variable by the population of interest for the variable to calculate a percentage.

Description of Variables

[7] County Health Rankings & Roadmaps. (n.d.) https://www.countyhealthrankings.org/
[8] County Health Rankings & Roadmaps. (n.d.) 2021 Measures. https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/2021-measures
[9] The New York Times Github. (n.d.) nytimes/covid-19-data. https://github.com/nytimes/covid-19-data

Observations

Our dataset has a total of 935 observations, corresponding to each of the counties included in our analysis.

Response Variable

Our response variable is the cumulative COVID-19 deaths per cases for each county in the US. We created this variable via a simple mutate operation of cumulative cases per county divided by cumulative deaths per county. We use deaths per cases as our response variable as a proxy for case severity in different counties across the United States, as it inherently controls for population size variability. We do acknowledge, however, that using a single COVID-19 metric may not account for underlying heterogeneities between subgroups in each county, a bias that results from varied distributions within and between populations.[10]

Explanatory Variables

Drawing on data from County Health Rankings & Roadmaps, we included 41 explanatory variables in our analysis, which fall into four main categories: health behaviors, clinical care, social and economic factors, and physical environment. For a detailed specification of these variables, refer to the Appendix.

Allocation of Data for Training and Testing

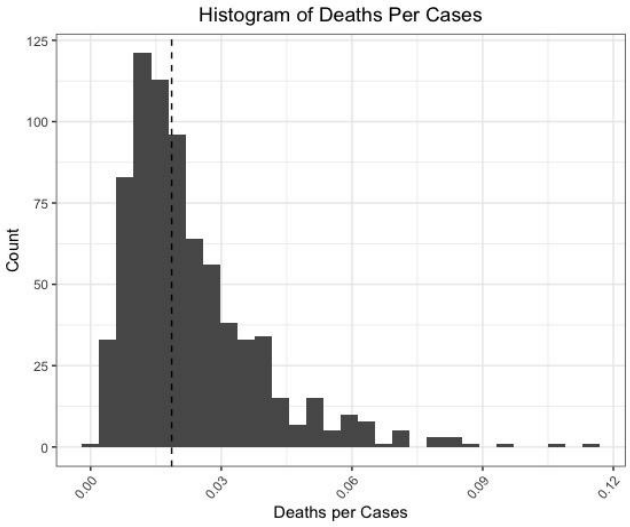
Before building our predictive models, we first removed observations from the dataset for which any variables had NA values. We decided to do this for consistency purposes, as some data analysis methods we employed require that no variables contain NA fields. We then split our dataset into two subsets: a training dataset used for building our predictive models and a test dataset used for evaluating our models. We used an 80-20 split, such that the training dataset consists of 80% of our observations and the dataset consists of 20% of our observations. Although this train-test split was performed separately for each class of methods, we utilized a random seed to ensure that each split led to the same results.

Exploratory Data Analysis of the Training Dataset

Response Variable

We first sought to understand the response variable's distribution. As seen in the histogram of deaths variable below, the data appears to be right-skewed, with some counties exceeding a death per cases rate of 0.1. The median deaths per cases is 0.018.

[10] U.S. News & World Report. (n.d.). COVID-19 Death Rate | Healthiest Communities. U.S. News & World Report. https://www.usnews.com/news/healthiest-communities/coronavirus-data/covid-death-rate?chart_type=line.



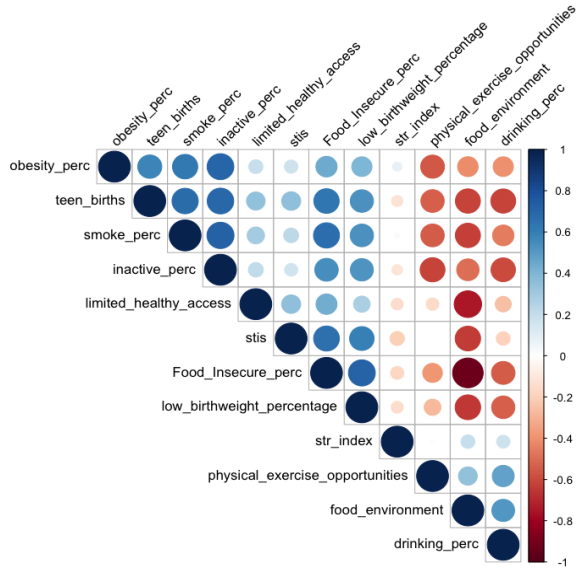
	County <chr>	State <chr>	deathspercases <dbl>
1	Orleans	New York	0.11600864
2	Sussex	New Jersey	0.10704202
3	Warren	New Jersey	0.09613882
4	Morris	New Jersey	0.08538473
5	Somerset	New Jersey	0.08490356
6	Hartford	Connecticut	0.08301789
7	Cape May	New Jersey	0.08271127
8	Bergen	New Jersey	0.07840372
9	Wayne	Michigan	0.07758423
10	Susquehanna	Pennsylvania	0.07749691

We proceeded to determine which counties had extreme response rates by looking at the sorted data. The sorted data above shows that aggregated across 2020, the highest rates of deaths per cases were primarily in northeastern states such as New York, New Jersey, and Connecticut.

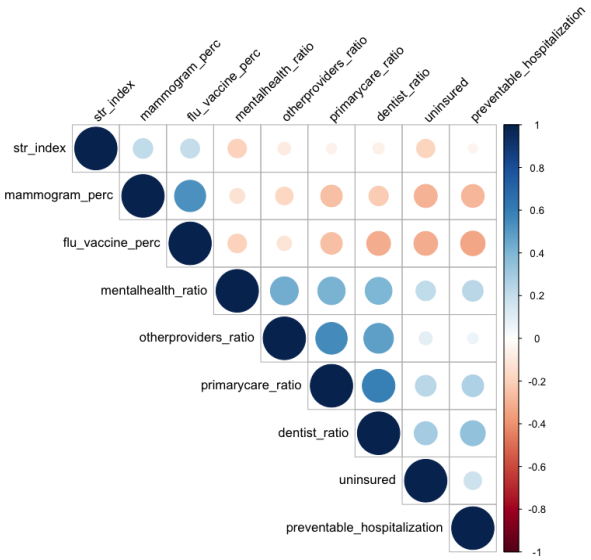
Explanatory Variables

Next, we sought to obtain a high-level view of the correlations of the predictor variables with each other and with the response variable. We first looked at correlations between explanatory variables within each category of health factors.

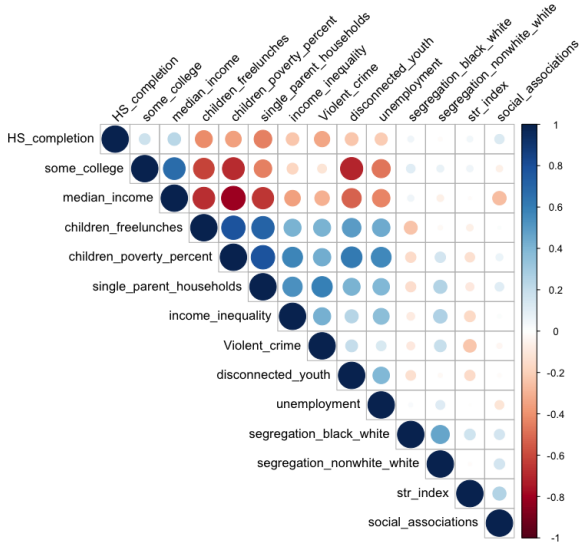
Health behaviors:



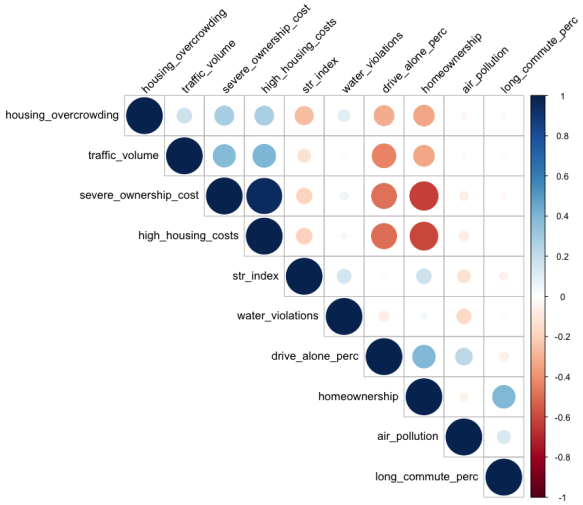
Clinical care:



Social and economic factors:



Physical environment:



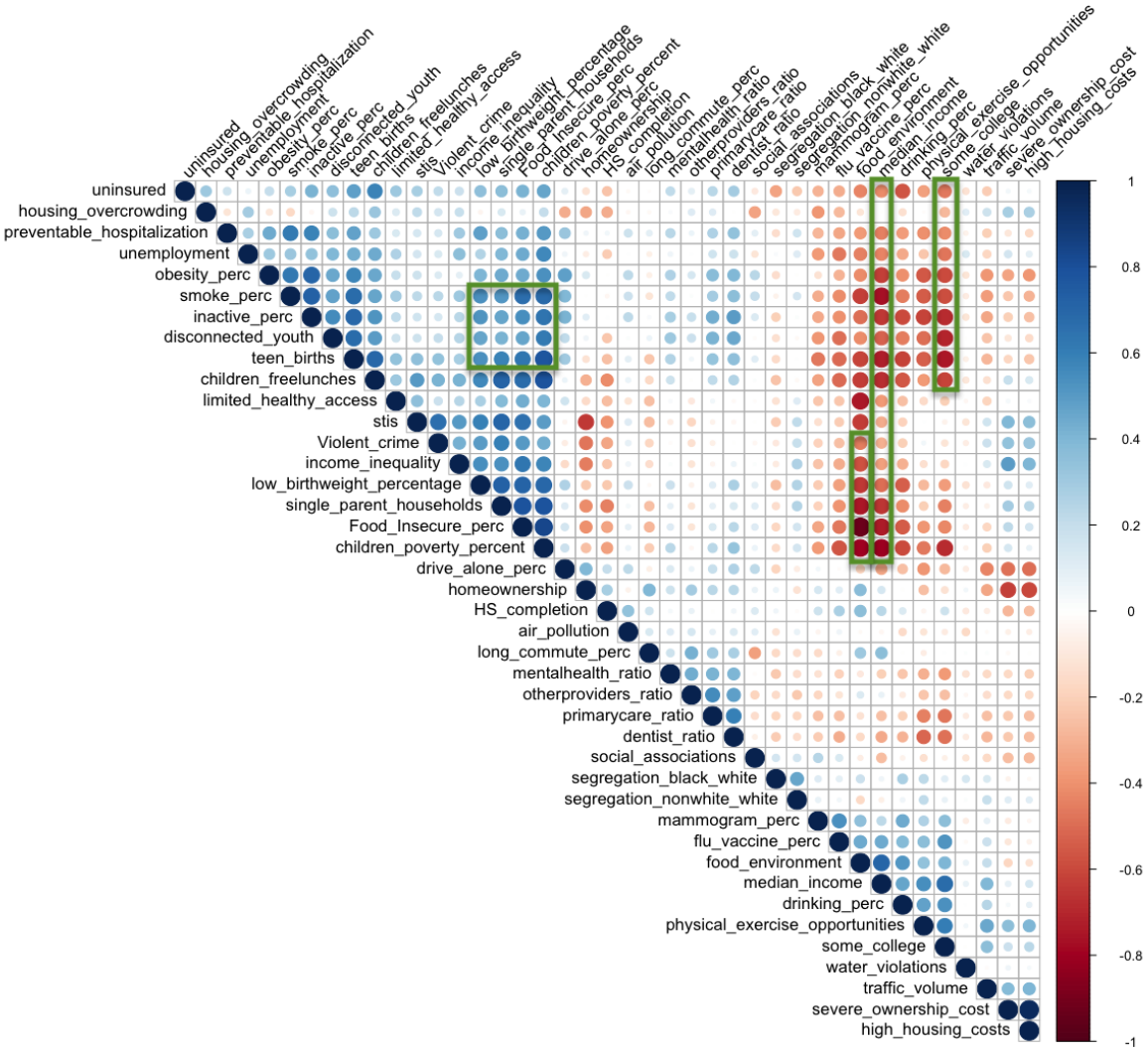
Health behaviors: We observe a negative correlation between food environment index and other variables that pertain to diet, such as food insecurity and limited access to healthy foods. In other words, the better the overall food environment, the less the population lacks adequate access to food, is low income, and lives far from a grocery store. We also notice a positive correlation between food insecurity and factors such as low birthweight, adult smoking, adult obesity, physical inactivity, and teen births.

Clinical care: We first note that there is a positive correlation between flu vaccinations and mammography screening. There is also a slightly negative correlation between uninsured and mammography screening and flu vaccinations. These observations make sense in that people who are uninsured may be less likely to receive a mammogram screening or flu vaccination. There are also positive correlations between access to different types of health providers (mental health providers, primary care providers, dentists, and other providers).

Social and economic factors: We observe a positive correlation between median household income (with higher values perhaps signalling higher income disparity) and other variables related to children, such as children in single-parent households, children in poverty, and children eligible for free or reduced price lunch. Median household income is also positively correlated with the percentage of adults ages 25-44 with some post-secondary education, and negatively correlated with variables that are themselves positively correlated such as disconnected youth, unemployment, and violent crime rate.

Physical environment: We observe a negative correlation between severe housing cost burden and home ownership, which is expected as higher housing costs make it difficult to purchase and own homes. Driving alone is slightly negatively correlated with housing problems or traffic volume.

Having analyzed correlations between variables in each category, we then assessed how the variables are correlated with each other overall.



This overall correlation plot allows comparison across different categories of variables. The top right of the plot notably shows several negative correlations. Specifically, median income and some college-related variables are negatively correlated with behavioral factors such as excessive drinking, physical inactivity, and adult obesity. Also, the food environment index is negatively correlated with median household income and other socioeconomic variables such as income inequality, children in poverty, and violent crime rate. The top left part of the plot shows several positive correlations. Notably, behavioral factors such as adult smoking, physical inactivity, and teen births are positively correlated with socioeconomic variables such as income inequality and children in poverty.

Model Building, Evaluation, and Interpretation

Ordinary Least Squares and Shrinkage Methods

Ordinary Least Squares

We began our analysis with an ordinary least squares regression of deaths per cases on all 41 explanatory variables. Notably, although our training dataset was shrunken due to our removal of any observations with NA fields, running the linear regression on the dataset in which NA values were present produced very similar results.[11] When checking for the normality assumptions, we noticed some nonlinearity in the residual plot and Normal quantile plot. Hence, we compared the normality when applying both a square root transformation and a log transformation to the response variable. Given that a log transformed response led to the closest semblance of normality, we report the corresponding results here.[12]

The OLS regression with a log-transformed response revealed that the following variables are significantly associated with the response at the 0.05 level: teen births, other primary care providers, high school completion, disconnected youth, unemployment, income inequality, children in single-parent households, housing overcrowding, residential segregation—nonwhite/white, homeownership, physical inactivity, flu vaccinations, and median household income. The multiple R-squared indicates that these features explain 32.5% of the variation in response.

Ridge Regression, LASSO Regression, and Elastic Net Regression

Despite the ordinary least squares method seeming to work well, we realized that fitting a linear model with so many explanatory variables might incur a large cost in variance and lead to suboptimal predictions. Hence, we decided to build and evaluate shrinkage models with the hopes of getting a more parsimonious and interpretable model. We ran three cross-validated regressions for which optimal values of lambda were chosen according to the one-standard-error rule: ridge, LASSO (Least Absolute Shrinkage and Selection Operator), and elastic net.[13] Since ridge regression does not have a selection feature, we report the variables selected by LASSO and elastic net. Notably, the elastic net regression selected all 13 variables by the LASSO plus an additional four variables.

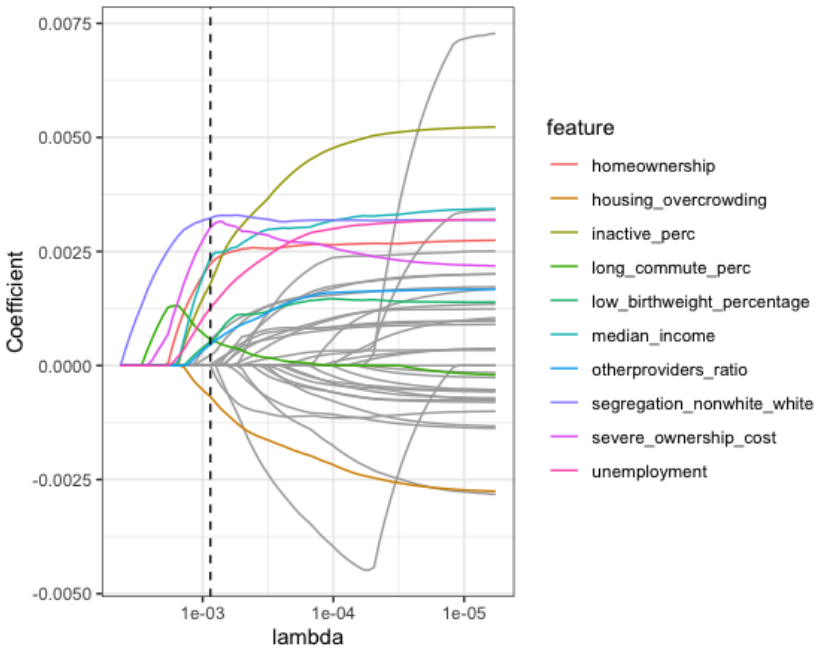
- **LASSO:** low birthweight percentage, other primary care providers, unemployment, income inequality, drinking water violations, housing overcrowding, residential segregation—non-White/White, homeownership, severe housing costs, physical inactivity
- **Elastic net:** low birthweight percentage, access to exercise opportunities, other primary

[11] For full summary statistics of linear regression using each dataset (with and without NAs), see Appendix.
[12] For more details on how we selected the log transformation, see Appendix.
[13] For cross-validation plots, see Appendix.

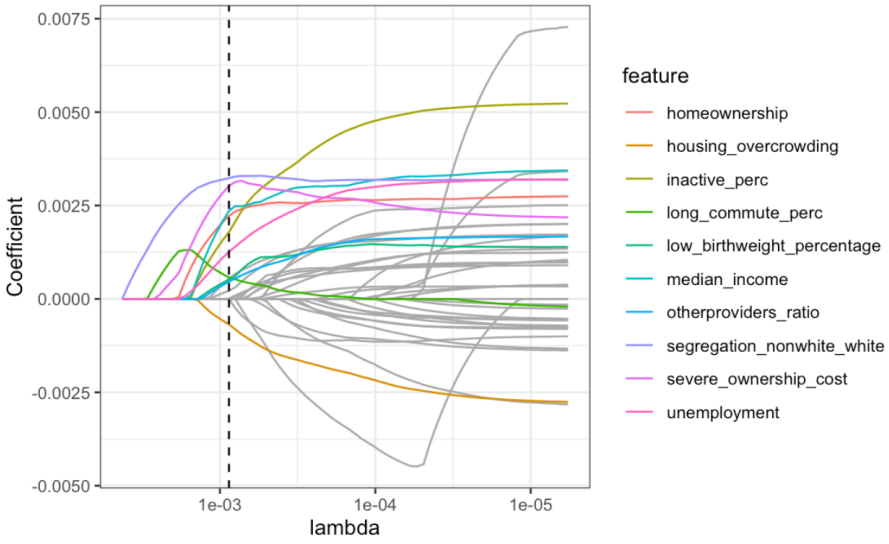
care providers, unemployment, income inequality, drinking water violations, severe housing costs, housing overcrowding, residential segregation—non-White/White, homeownership, severe housing cost burden, adult smoking, physical inactivity, food insecurity, flu vaccinations, median houosehold income, and long commute

Below are plots of the results of each regression, with the top 10 features identified by each method highlighted in color.

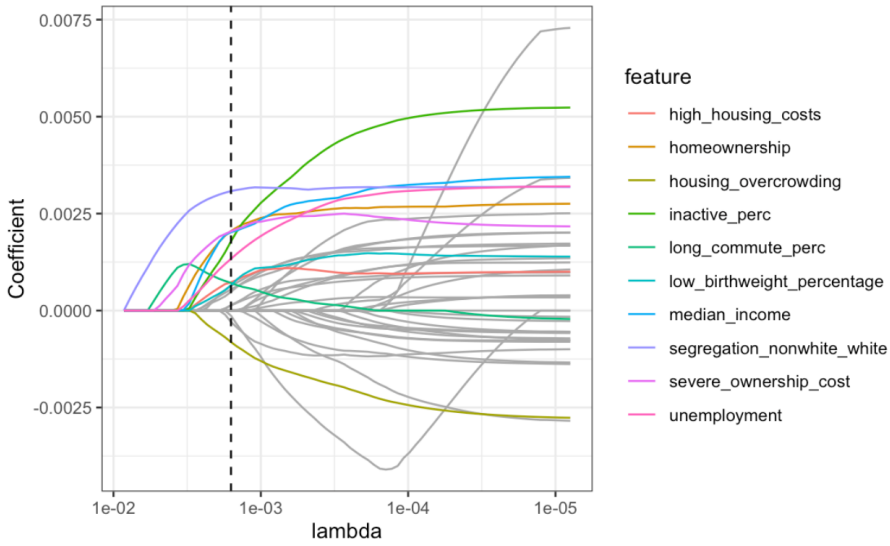
Ridge:



LASSO:



Elastic Net:

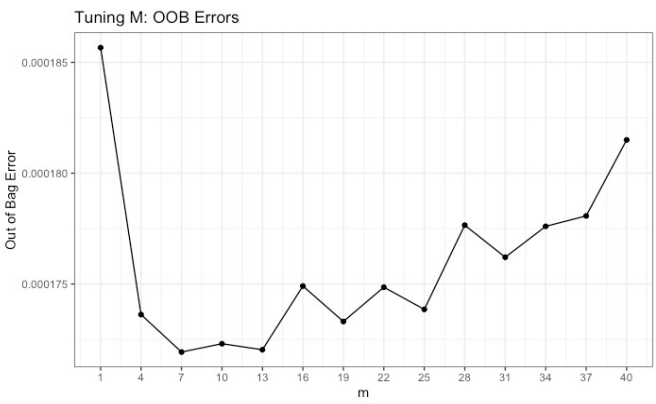


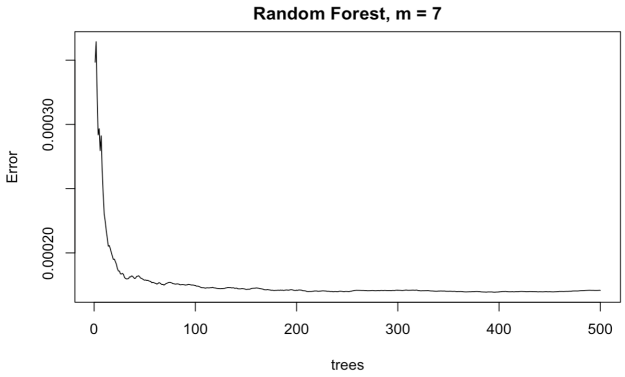
Tree-Based Methods

Random Forest

We then proceeded to fit a random forest to the data. While bagging is achieved when all 41 variables are considered at each tree split, leading to higher variance and suboptimal prediction performance, we tuned the random forest model for the optimal value of m, or the number of features to consider at each tree split, by training the model on different values of m, ranging from 1 to 41. The out of bag error for each value of m can be observed below, and we observe that the out of bag error is minimized at a value of $m = 7$, after which the error follows a loosely upward trend.

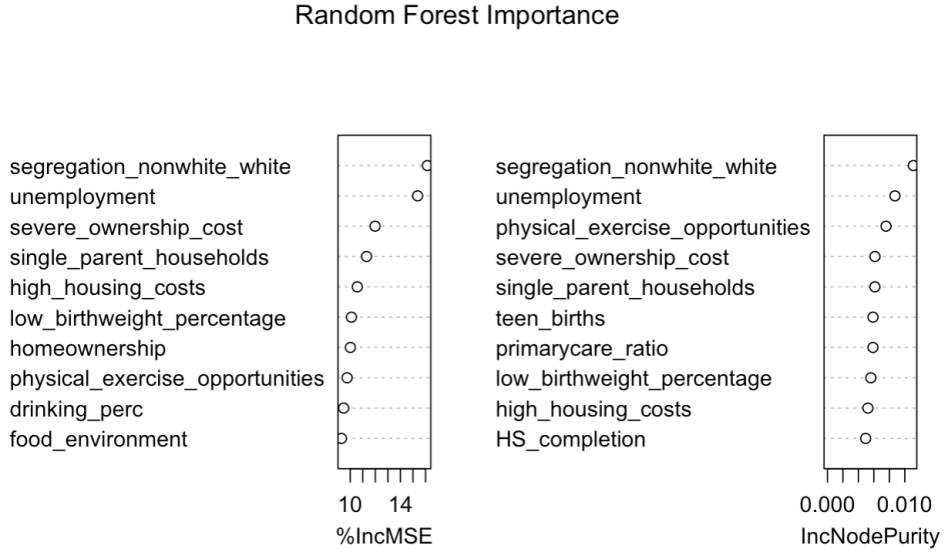
Next, we tuned the B parameter, which controls the number of bootstrap samples (corresponding to the number of fitted trees). As shown in the plot below, the cross-validated training error sharply decreases as the number of trees increases and then plateaus starting at around $B = 200$.





We then fit our tree using the *m* and *B* parameters specified above and assessed variable importance. In a random forest context, there are two notions of variable importance: purity based importance and OOB variable importance. OOB variable importance is a measure of the deterioration in prediction accuracy that results from scrambling a given feature out of bag. Purity based importance is a measure of the degree of improvement in node purity that results from splitting on a given feature. The results of both of these variable importance measures are summarized below.

OOB and Purity variable importance:



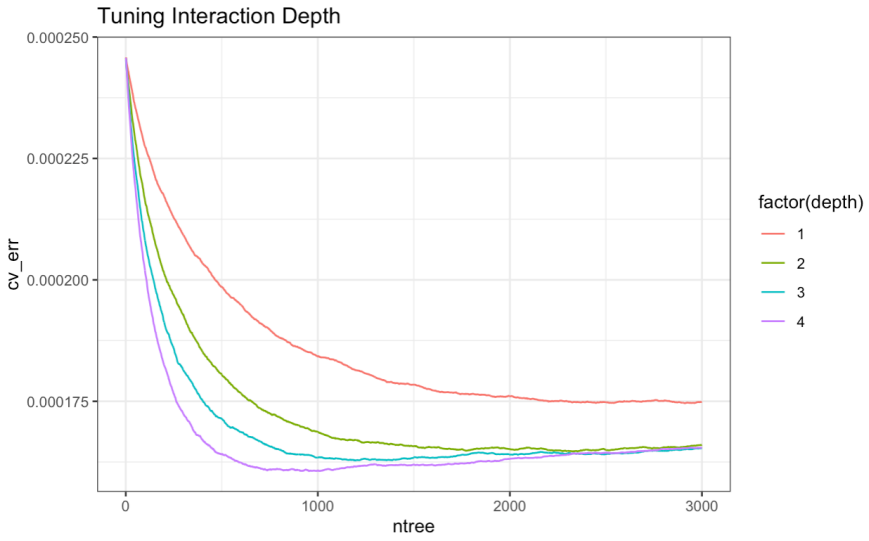
Looking at the purity based importance and OOB Error improvement plots above, we observe that unemployment and the segregation variable measuring the degree to which counties see segregation between White and non-White people have the highest importance as measured by both metrics. Physical exercise opportunities is a close third. This suggests that these variables are

most important in predicting COVID-19 deaths per cases across different counties in the US.

Boosting

Boosted Model

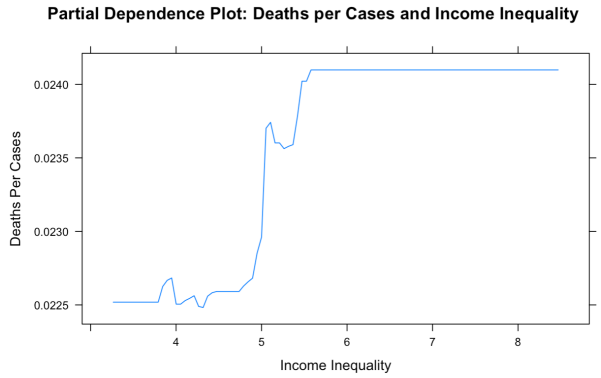
Boosting is another state-of-the-art prediction method that utilizes the aggregated outputs of multiple decision trees to achieve good prediction performance. Hence, we proceeded to fit our data using a boosted model. We began the model building process using the default parameters of 100 trees, a shrinkage factor of 0.1, an interaction depth of 1, and a subsampling fraction π of 0.5. In order to optimally tune our boosted model, we proceed to fit models with a larger number of trees and experiment with different shrinkage factors and interaction depths as well. Finding that the optimal number of trees is typically in the range of 1000 to 3000 across different model variations via experimentation, we then tuned the interaction depth of the model, testing interaction depths of 1, 2, 3, and 4 (all else equal). As per the cross validation plot shown below, we observe that the model with an interaction depth of 4 attains the minimum cross validated error, with 954 trees.

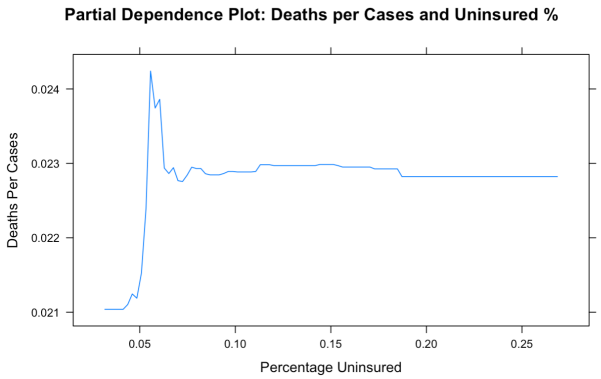


With the tuned boosted model as per the process outlined above, we then assessed variable importance via two measures of variable importance: purity-based importance (defined above in the description of random forest) and partial dependence plots. First, we report the results of purity-based importance, ranking the top 10 variables by their contributions to improvements in node purity.

Variable	Relative Importance
Segregation Non-white White	8.490
Unemployment	8.130

Physical Exercise Opportunities	5.874
Severe Ownership Cost	5.472
Teen Births	4.650
Homeownership	4.424
Ratio of Other Providers	4.423
STIs	3.420
Low Birth Weight Percentage	3.207
HS Completion	2.966





Next, we present partial dependence plots for the most important two variables: percentage uninsured and income inequality. Because our optimal tree has an interaction depth of 4, these plots estimate the approximate relationship of the two individual variables with our response variable. As the relative measure of income inequality increases, from a rating of 4 to 5, we see relatively stable deaths per cases, after which there is a significant spike. This suggests that income inequality could be a negative factor that significantly impacts a community's resilience to COVID-19 after a certain inequality threshold. Similarly, for the percentage of uninsured residents in a county, the percentage of uninsured residents increases dramatically with deaths per cases, before stabilizing. Taken together, these factors suggest that counties with high income inequality and higher percentages of uninsured individuals might face disproportionately higher burdens on their healthcare systems. Such factors might have reduced the resources available for individuals in these counties, leading to more COVID-19 fatalities.

Model Comparison

The mean squared errors (MSEs) of each of our models are presented below:

Model	Train MSE	Test MSE

Ordinary Least Squares (Log-Transformed Model)	0.288492	0.312119
Ridge Regression	0.000188	0.000158
Lasso Regression	0.000193	0.000164
Elastic Net Regression	0.000189	0.000161
Random Forest	0.00029	0.000141
Boosted Model	0.000058	**0.000139**

Conclusions

Comparison of Method Performance

As shown in the table above, the random forest and the boosted model have the lowest test errors. This is reasonable given these models' tendencies to have high predictive accuracy. Between the two, the boosted model has the lowest test error, with a mean squared error of 0.000139, but it is closely followed by random forest, which has a mean squared error of 0.00141. Notably, however, the ridge, LASSO, and elastic net regressions perform about as well, with test MSEs of 0.000158, 0.000164, and 0.000161, respectively. The unpenalized OLS model has the largest testing MSE of 0.312119, which is significantly higher than the train MSE, suggesting that the OLS model might be overfit.

Regardless of these differences in test MSE, the methods overlap significantly in their identification of important variables from the larger set. For instance, the elastic net regression selects the following variables, which are also selected by LASSO and deemed significant in the OLS model: other providers ratio, unemployment, income inequality, housing overcrowding, residential segregation—non-White/White, homeownership, and physical inactivity. The random forest and boosting models both include low birthweight percentage, median income, and unemployment percentage in the top 10 most important variables, as measured by their contributions to node purity.

Overall Conclusions, Recommendations, and Takeaways for Stakeholders

Our results point to a few key determinants of health that, given their impact on COVID-19 deaths per cases rates in 2020, policymakers should consider when aiming to improve factors that would improve health overall but also potentially mitigate the mortality risk of another pandemic. The boosted model, which had the strongest predictive performance, suggests that residential segregation between non-white and white residents is the most important variable in predicting a county's COVID-19 deaths per cases rate. The unemployment rate, availability of physical exercise opportunities, and measures of home ownership burden variables were also

highly important in this model. These variables are identified across all models suggesting these relationships are robust. Residential segregation, unemployment, and home ownership burden are socioeconomic factors that affect an individual's ability to access and pay for healthcare. In that regard, it is unsurprising that these factors would have a greater ability to predict differences in COVID-19 case fatalities across different counties in the United States. While some behavioral variables are also found to be significant (including STI incidence, high school completion, and others identified in the elastic net and lasso regressions), the variable importance ranking from the boosted model provides a highly interpretable hierarchy of the most influential factors from the greater set.

Given that socioeconomic factors were the strongest predictors of deaths per cases, it appears that on the county level, COVID-19 rates are most associated with community healthcare burdens. That is, COVID-19 outcomes appear to reflect the reality of healthcare accessibility across counties; those with high percentages of uninsured citizens, or with high degrees of segregation and inequality, might feature division of healthcare resources that reflects these disparities. It is thus reasonable that deaths per case rates would be higher when significant groups of a population have lesser access to healthcare resources and treatment. Notably, if our conclusions are indeed correct, this effect would likely be pronounced in situations of scarcity such as the early months of the pandemic studied here, when many hospitals faced shortages of ventilators and other resource shortages. Given the progression of COVID-19 since 2021 as well as the inherent complexity of fatality incidence, we are hesitant to make any assertive claims about the true predictive capacity of any of the top factors we identified. Nonetheless, these results can help inform policies directed toward improving various determinants of important health outcomes in counties across the US.

As the world shifts towards herd immunity as vaccines are made more widely available, it is important to reflect upon how the pandemic has asymmetrically impacted different counties across the country. Our results suggest that structural vulnerabilities can be captured by measures of inequality and poverty; the identification of counties high on these factors should serve as a warning for future vulnerability to health crises. These analyses can serve to protect already vulnerable communities from suffering disproportionately in the future.

Limitations and Future Directions

Dataset Limitations

As it is detailed in the Frequently Asked Questions page of County Health Rankings & Roadmaps program website,[14] all of the variables are from 2019 or earlier. Thus, it is possible that the values of the variables assessed were different in 2020, meaning that the interpretation of our

[14] County Health Rankings & Roadmaps. (n.d.) Frequently Asked Questions. https://www.countyhealthrankings.org/explore-health-rankings/faq-page

analysis may need to be taken with a grain of salt. Regardless, given that we analyzed data on the county level, it is unlikely that any county experienced enough drastic change over the course of the year to significantly affect our analysis. Furthermore, our dataset has a large number of observations to account for potential variability, although notably, many observations had to be removed as some of the R packages used to build some of our models required that no NA values were present in the data. In other words, since each observation represents a different US county, many counties were left out of our analysis. Another limitation is that, as described in the Exploratory Data Analysis section, there is evidence of correlation amongst some of our explanatory variables. This means that some variables can be confounding variables, which mask or distort the relationship between measured variables. Also, variables selected in the LASSO regression and elastic net regression as well as variables marked important in the tree methods might be misleading in that, given how variable selection works, it is possible that some selected variables are simply representative of a larger group of correlated variables.

Analysis Limitations

While splitting the data into training and testing datasets allows for a more unbiased test of the models, we recognize that our conclusions inherently contain some randomness due to the random split of the data. In other words, splitting the data again using a different random seed may have yielded different *p*-values in the OLS regression, different selected variables in the shrinkage methods, and different variables selected as important in the tree-based methods. Next, although we provide different methods for robust interpretation of the variables, our analysis incorporates only a specific subset of health related variables. The results of the analysis might change dramatically if we were to incorporate other variables. For example, as mentioned in the Exploratory Data Analysis, states in the northeastern part of the United States suffered from a high rate of COVID-19 cases and deaths in 2020. This may not have been because these states performed poorly in the health variables mentioned above. Rather, it may be due to the fact that these states are densely populated and hence were more susceptible to disease spread at the outset of the COVID-19 pandemic. In other words, other factors like geographic or demographic variables can hugely impact case fatality rate.

Recommended Follow-Up Analyses

To compensate for the limitations mentioned above, more extensive analysis can be done as we acquire more data from 2020 and 2021. Not only can we extend our analysis by utilizing the most up-to-date datasets, but we can also examine how COVID-19 cases and deaths have affected various health factors of each county. In other words, the explanatory and response variables can be reversed to conduct more dynamic data analyses. Next, given that many observations needed to be omitted in our dataset as they contained NA fields, we recommend that our analyses be reconducted once the missing data is collected. Finally, future work on the social determinants of health in the context of COVID-19 might also look at different population levels such as states, bigger geographical regions in America, or even different countries.

Appendix

Explanatory Variables

Below are the 41 explanatory variables we used for analysis. Words written in parentheses represent variable names used in R analysis. Unless noted otherwise, all variables are continuous.

Health behaviors:

- *Tobacco Use*
  - **Adult smoking (smoke_perc):** Percentage of adults who are current smokers.
- *Diet and Exercise*
  - **Adult obesity (obesity_perc):** Percentage of the adult population (age 20 and older) reporting a body mass index (BMI) greater than or equal to 30 kg/m^2.
  - **Food environment index (food_environment):** Index of factors that contribute to a healthy food environment, from 0 (worst) to 10 (best).
  - **Physical inactivity (inactive_perc):** Percentage of adults age 20 and over reporting no leisure-time physical activity.
  - **Access to exercise opportunities (physical_exercise_opportunities):** Percentage of population with adequate access to locations for physical activity
  - **Food insecurity (Food_Insecure_perc):** Percentage of population who lack adequate access to food.
  - **Limited access to healthy foods (limited_healthy_access):** Percentage of population who are low-income and do not live close to a grocery store.
- *Alcohol & Drug Use*
  - **Excessive Drinking (drinking_perc):** Percentage of adults reporting binge or heavy drinking.
- *Sexual Activity*
  - **Sexually transmitted infections (stis):** Number of newly diagnosed chlamydia cases per 100,000 population.
  - **Teen births (teen_births):** Number of births per 1,000 female population ages 15-19.
  - **Low Birth Weight Percentage(low_birthweight_percentage):** Percentage of live births with low birthweight (< 2,500 grams).

Clinical care:

- *Access to Care*
  - **Uninsured (uninsured):** Percentage of population under age 65 without health insurance.
  - **Primary care physicians (primarycare_ratio):** Ratio of population to primary care physicians.
  - **Dentists (dentist_ratio):** Ratio of population to dentists.
  - **Mental health providers (mentalhealth_ratio):** Ratio of population to mental health providers.
  - **Other primary care providers (otherproviders_ratio):** Ratio of population to primary care providers other than physicians.
- *Quality of Care*
  - **Preventable hospital stays (preventable_hopitalization):** Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees.
  - **Mammography screening (mammogram_perc):** Percentage of female Medicare enrollees ages 65-74 that received an annual mammography screening.
  - **Flu vaccinations (flu_vaccine_perc):** Percentage of fee-for-service (FFS) Medicare

enrollees that had an annual flu vaccination.
    - **Teen births (teen_births):** Number of births per 1,000 female population ages 15-19.

Social and economic factors:

- *Education*
    - **High school completion (HS_completion):** Percentage of adults ages 25 and over with a high school diploma or equivalent.
    - **Some college (some_college):** Percentage of adults ages 25-44 with some post-secondary education.
    - **Disconnected youth (disconnected_youth):** Percentage of teens and young adults ages 16-19 who are neither working nor in school.
- *Employment*
    - **Unemployment (unemployment):** Percentage of population ages 16 and older who are unemployed but seeking work.
- *Income*
    - **Children in poverty (children_poverty_percent):** Percentage of people under age 18 in poverty.
    - **Income inequality (income_inequality):** Ratio of household income at the 80th percentile to income at the 20th percentile.
    - **Median household income (median_income):** The income where half of households in a county earn more and half of households earn less.
    - **Children eligible for free or reduced price lunch (children_freelunches):** Percentage of children enrolled in public schools that are eligible for free or reduced price lunch.
- *Family & Social Support*
    - **Children in single-parent households (single_parent_households):** Percentage of children that live in a household headed by a single parent.
    - **Social associations (social_associations):** Number of membership associations per 10,000 residents.
    - **Residential segregation—Black/White (segregation_black_white):** Index of dissimilarity where higher values indicate greater residential segregation between Black and White county residents.
    - **Residential segregation—non-White/White (segregation_nonwhite_white):** Index of dissimilarity where higher values indicate greater residential segregation between non-White and White county residents.
- *Community Safety*
    - **Violent crime rate (Violent_crime)** Number of reported violent crime offenses per 100,000 residents.

Physical environment:

- *Air & Water Quality*
    - **Air pollution - particulate matter (air_pollution):** Average daily density of fine particulate matter in micrograms per cubic meter (PM2.5).
    - **Drinking water violations (water_violations):** Indicator of the presence of health-related drinking water violations. 1 indicates the presence of a violation, 0 indicates no violation.
- *Housing & Transit*
    - **Housing overcrowding (housing_overcrowding):** Percentage of households with overcrowding,
    - **Severe housing costs (high_housing_costs):** Percentage of households with high housing costs

- **Driving alone to work (driving_alone_perc):** Percentage of the workforce that drives alone to work.
- **Long commute—driving alone (long_commute_perc):** Among workers who commute in their car alone, the percentage that commute more than 30 minutes.
- **Traffic volume (traffic_volume):** Average traffic volume per meter of major roadways in the county.
- **Homeownership (homeownership):** Percentage of occupied housing units that are owned.
- **Severe housing cost burden (severe_ownership_cost):** Percentage of households that spend 50% or more of their household income on housing.

Summary Statistics for Ordinary Least Squares

Dataset without NA values:

```
Call:
lm(formula = log(deathspercases) ~ ., data = health_train)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2922 -0.3230  0.0091  0.3439  1.5493

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -8.722e+00  3.770e+00  -2.313  0.02100 *
low_birthweight_percentage      3.386e+00  2.426e+00   1.395  0.16332
food_environment                9.352e-02  3.419e-01   0.274  0.78453
physical_exercise_opportunities 1.876e-01  1.963e-01   0.955  0.33979
teen_births                    -7.772e+00  3.993e+00  -1.946  0.05200 .
limited_healthy_access          1.712e+00  3.076e+00   0.557  0.57792
stis                            1.616e+01  2.012e+01   0.803  0.42211
uninsured                      -7.578e-02  8.287e-01  -0.091  0.92717
primarycare_ratio              -1.629e-05  2.044e-05  -0.797  0.42553
dentist_ratio                  -5.582e-06  2.015e-05  -0.277  0.78181
mentalhealth_ratio              2.962e-06  1.975e-05   0.150  0.88083
otherproviders_ratio            4.800e-05  3.796e-05   1.264  0.20650
HS_completion                   8.085e-01  4.165e-01   1.941  0.05263 .
some_college                   -2.193e-01  4.639e-01  -0.473  0.63650
disconnected_youth              1.725e+00  7.652e-01   2.255  0.02445 *
unemployment                    9.189e+00  2.238e+00   4.105 4.51e-05 ***
income_inequality               4.713e-02  5.174e-02   0.911  0.36264
children_freelunches            1.995e-01  2.488e-01   0.802  0.42306
single_parent_households        1.638e+00  5.736e-01   2.856  0.00441 **
social_associations             1.375e+02  7.678e+01   1.791  0.07367 .
water_violations                8.006e-02  4.417e-02   1.813  0.07033 .
high_housing_costs             -5.516e-01  2.327e+00  -0.237  0.81265
housing_overcrowding           -5.828e+00  1.953e+00  -2.983  0.00295 **
segregation_black_white        -2.940e-03  2.137e-03  -1.375  0.16949
segregation_nonwhite_white      1.276e-02  2.621e-03   4.866 1.40e-06 ***
homeownership                   1.564e+00  5.395e-01   2.898  0.00387 **
severe_ownership_cost           2.990e+00  2.459e+00   1.216  0.22446
smoke_perc                     -1.743e+00  1.381e+00  -1.262  0.20725
obesity_perc                   -1.233e+00  6.916e-01  -1.783  0.07505 .
inactive_perc                   3.050e+00  6.975e-01   4.372 1.41e-05 ***
drinking_perc                  -2.068e+00  1.128e+00  -1.834  0.06713 .
Food_Insecure_perc             -3.930e+00  6.393e+00  -0.615  0.53892
preventable_hospitalization    -2.804e-05  1.838e-05  -1.525  0.12762
mammogram_perc                 -4.494e-01  5.172e-01  -0.869  0.38518
flu_vaccine_perc                1.040e+00  4.355e-01   2.388  0.01721 *
children_poverty_percent        7.877e-01  7.618e-01   1.034  0.30153
median_income                   6.273e-06  3.595e-06   1.745  0.08145 .
Violent_crime                   1.171e-04  1.456e-04   0.805  0.42129
air_pollution                   4.193e-03  1.565e-02   0.268  0.78888
drive_alone_perc               -2.744e-01  5.926e-01  -0.463  0.64348
long_commute_perc               4.377e-01  3.094e-01   1.415  0.15756
traffic_volume                  1.243e-04  7.821e-05   1.589  0.11249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5529 on 706 degrees of freedom
Multiple R-squared:  0.3249,    Adjusted R-squared:  0.2857
F-statistic: 8.287 on 41 and 706 DF,  p-value: < 2.2e-16
```

Dataset with NA values:

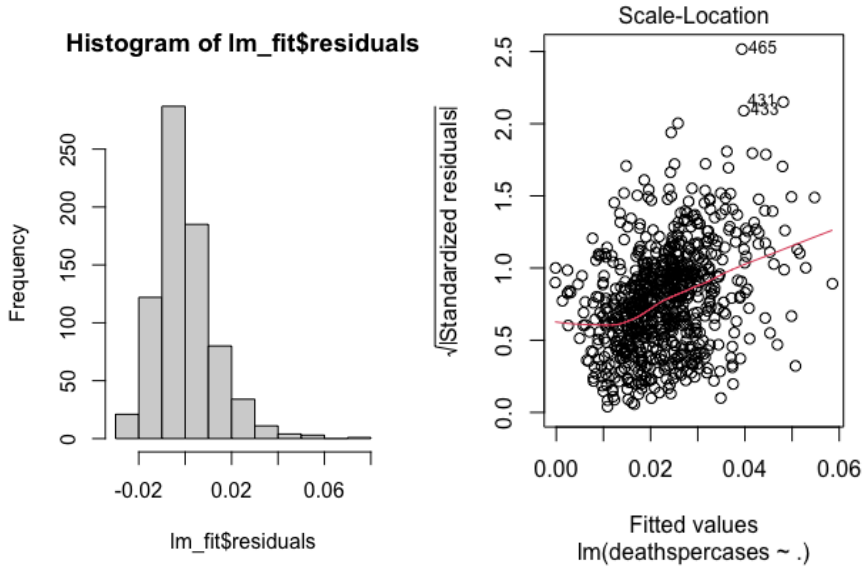
```
Call:
lm(formula = log(deathspercases) ~ ., data = health_yesna)

Residuals:
     Min       1Q   Median       3Q      Max
-2.31779 -0.33597  0.00392  0.34430  1.66765

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -1.161e+01  3.373e+00  -3.441 0.000606 ***
low_birthweight_percentage      2.792e+00  2.161e+00   1.292 0.196644
food_environment                3.346e-01  3.056e-01   1.095 0.273808
physical_exercise_opportunities 1.236e-01  1.736e-01   0.712 0.476575
teen_births                    -9.775e+00  3.488e+00  -2.802 0.005181 **
limited_healthy_access          3.964e+00  2.745e+00   1.444 0.149081
stis                            1.839e+01  1.709e+01   1.076 0.282244
uninsured                       2.129e-01  7.202e-01   0.296 0.767633
primarycare_ratio              -1.881e-05  1.863e-05  -1.009 0.313084
dentist_ratio                   4.020e-06  1.761e-05   0.228 0.819542
mentalhealth_ratio              1.468e-05  1.796e-05   0.818 0.413709
otherproviders_ratio           -1.376e-06  3.307e-05  -0.042 0.966806
HS_completion                   5.963e-01  3.726e-01   1.600 0.109928
some_college                   -6.373e-01  4.073e-01  -1.565 0.118010
disconnected_youth              2.238e+00  6.620e-01   3.381 0.000753 ***
unemployment                    1.043e+01  2.043e+00   5.108 3.98e-07 ***
income_inequality               5.147e-02  4.467e-02   1.152 0.249535
children_freelunches            1.314e-01  2.199e-01   0.597 0.550369
single_parent_households        1.891e+00  5.002e-01   3.781 0.000167 ***
social_associations             1.625e+02  6.785e+01   2.395 0.016832 *
water_violations                8.432e-02  3.896e-02   2.164 0.030725 *
high_housing_costs              6.290e-01  2.008e+00   0.313 0.754178
housing_overcrowding           -4.466e+00  1.706e+00  -2.617 0.009009 **
segregation_black_white        -3.520e-03  1.907e-03  -1.846 0.065213 .
segregation_nonwhite_white      1.248e-02  2.291e-03   5.447 6.61e-08 ***
homeownership                   1.530e+00  4.821e-01   3.173 0.001560 **
severe_ownership_cost           1.715e+00  2.122e+00   0.808 0.419030
smoke_perc                     -5.737e-01  1.196e+00  -0.480 0.631652
obesity_perc                   -1.237e+00  5.882e-01  -2.103 0.035733 *
inactive_perc                   2.262e+00  6.086e-01   3.717 0.000214 ***
drinking_perc                  -1.668e+00  9.874e-01  -1.690 0.091468 .
Food_Insecure_perc              1.327e+00  5.689e+00   0.233 0.815641
preventable_hospitalization    -2.429e-05  1.542e-05  -1.576 0.115370
mammogram_perc                 -2.780e-01  4.471e-01  -0.622 0.534240
flu_vaccine_perc                1.473e+00  3.821e-01   3.855 0.000124 ***
children_poverty_percent        3.459e-01  6.768e-01   0.511 0.609425
median_income                   8.935e-06  3.195e-06   2.796 0.005279 **
Violent_crime                   1.305e-04  1.303e-04   1.001 0.317015
air_pollution                   5.855e-03  1.357e-02   0.431 0.666344
drive_alone_perc               -1.107e-01  5.193e-01  -0.213 0.831203
long_commute_perc               5.533e-01  2.689e-01   2.058 0.039928 *
traffic_volume                  1.502e-04  6.869e-05   2.187 0.028994 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.549 on 893 degrees of freedom
  (2198 observations deleted due to missingness)
Multiple R-squared:  0.3224,    Adjusted R-squared:  0.2913
F-statistic: 10.36 on 41 and 893 DF,  p-value: < 2.2e-16
```

Assumptions for Ordinary Least Squares

To perform the Ordinary Least Squares regression, we checked 1) normality of residual variation

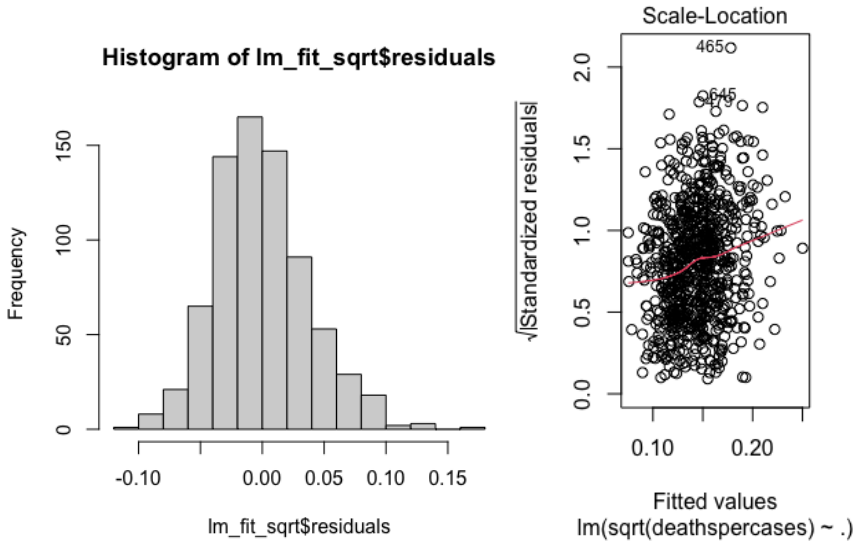
and 2) homogeneity of variance (homoscedasticity) of residuals. We check for the normality by looking at the histogram of the residuals, and for the homoscedasticity by looking at the scale-location plot.

Model Without Transformation:



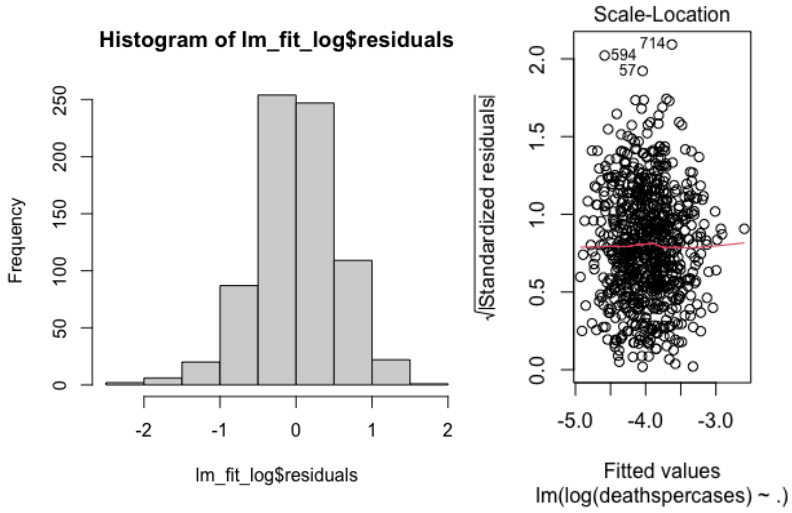
The histogram of the residuals is right-skewed, and the variance is heteroscedastic as it seems to increase with higher fitted values. Thus, this model violates both assumptions.

Square Root Transformation:



By applying the square root transformation, the histogram of the residuals is closer to normal, but the variance is still heteroscedastic as it seems to increase with higher fitted values. Thus, this model violates the homogeneity of variance assumption.

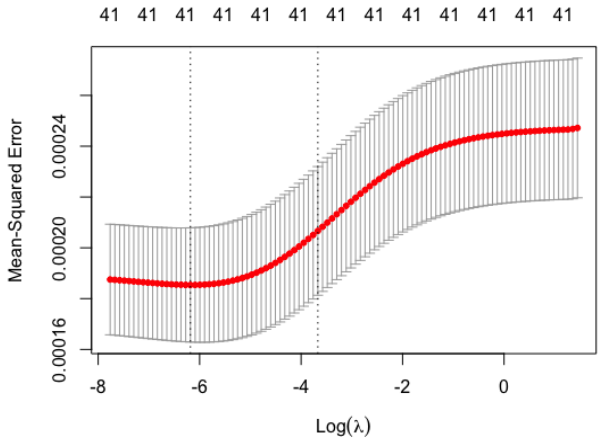
Log Transformation:



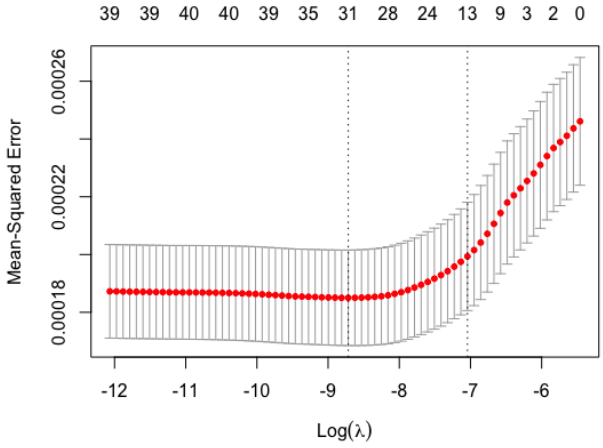
Finally, by applying the log transformation, the histogram of the residuals is approximately normal, and the variance is homogenous across fitted values. Thus, the log transformation does best, and we choose this model for analysis.

Shrinkage Methods Cross-Validation Plots

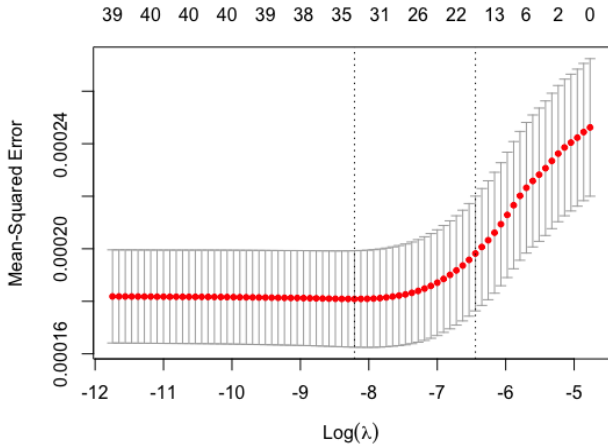
Ridge regression:



Lasso regression:



Elastic net regression:


39 40 40 40 39 38 35 31 26 22 13 6 2 0
Mean-Squared Error
0.00016
0.00020
0.00024
-12
-11
-10
-9
-8
-7
-6
-5
Log(λ)