Journal of Data Science 22 (4), 631–646
October 2024
DOI: 10.6339/23-JDS1111
Data Science in Action

The Effects of County-Level Socioeconomic and Healthcare Factors on Controlling COVID-19 in the Southern and Southeastern United States

Jackson Barth[1], Guanqing Cheng[1], Webb Williams[1], Ming Zhang[1], and Hon Keung Tony Ng[2,*]

[1]*Department of Statistical Science, Southern Methodist University, Dallas, Texas 75275, USA*
[2]*Department of Mathematical Sciences, Bentley University, Waltham, Massachusetts 02452, USA*

Abstract

This paper aims to determine the effects of socioeconomic and healthcare factors on the performance of controlling COVID-19 in both the Southern and Southeastern United States. This analysis will provide government agencies with information to determine what communities need additional COVID-19 assistance, to identify counties that effectively control COVID-19, and to apply effective strategies on a broader scale. The statistical analysis uses data from 328 counties with a population of more than 65,000 from 13 states. We define a new response variable by considering infection and mortality rates to capture how well each county controls COVID-19. We collect 14 factors from the 2019 American Community Survey Single-Year Estimates and obtain county-level infection and mortality rates from USAfacts.org. We use the least absolute shrinkage and selection operator (LASSO) regression to fit a multiple linear regression model and develop an interactive system programmed in R shiny to deliver all results. The interactive system at https://asa-competition-smu.shinyapps.io/COVID19/ provides many options for users to explore our data, models, and results.

Keywords *American Community Survey; interactive system; LASSO regression; R shiny*

1 Introduction

This paper is based on the entry of the 2021 Data Challenge Expo jointly sponsored by three American Statistical Association (ASA) Sections – Statistical Computing, Statistical Graphics, and Government Statistics. Beginning in late 2019 and early 2020, the novel coronavirus known as COVID-19 had an immediate and profound impact on the world, quickly becoming a leading cause of death in several countries, including the United States (Woolf et al., 2021). But compared to the rest of the world, COVID-19 seemed to substantially disrupt U.S. public health, with higher estimated deaths attributed to the virus than in similar countries (Bilinski and Emanuel, 2020). Some have attributed this to the inconsistent response of the U.S. government at federal and state levels (Haeder and Gollust, 2020), with some state and local governments ordering strict lockdown policies that endured much longer than others. This created a significant disparity in the way that the pandemic unfolded within different regions in the United States and served to expose existing social and economic imbalances. While in 2023, the threat of COVID-19 has ebbed significantly, understanding how the demographic, economic, and public

*Corresponding author. Email: tng@bentley.edu.

Received November 30, 2022; Accepted July 10, 2023

health factors present in different U.S. counties impacted incidence and death rates of the virus will help prepare for the next pandemic, both in terms of local public health policy and in the federal distribution of medical and financial aid.

In the months and years following the initial COVID-19 outbreak in the U.S., several studies sought to determine demographic factors associated with infections and death rates of the virus at the county level. Notable results included income inequality as a positive driver of incidence rates (Abedi et al., 2021) and social vulnerability factors as positive drivers of both infection and death rates (Karmakar et al., 2021; Clouston et al., 2021). Another commonly cited demographic factor is race, particularly in the early months of the pandemic. Cheng et al. (2020) shows that in March–July 2020, rural counties with higher rates of Black and Hispanic citizens tended to have higher levels of mortality due to COVID. In a study through December 2020, McLaren (2021) explored the disparity in death rates for racial minorities, identifying the underlying economic factors such as income, poverty levels, and health insurance. Studies have also reported partisanship as an underlying driver of infection and death rates, with "Trump-leaning" counties having lower death rates early in the pandemic but higher death rates in the fall of 2020 (Desmet and Wacziarg, 2022). While most studies looked at county-specific effects, Doti (2021) instead looked at state-specific effects, using the Oxford stringency index (a measure of government response to COVID-19) to study the impact of state-government mandates on COVID-19 death rates.

In order to fully understand the direct impact that demographic factors have on COVID-19 incidence and mortality rates, this study seeks to remove the "partisan effect" by normalizing the impact of state policies and focusing on states in the south and southeast, a region with a somewhat homogeneous political landscape. Novel death and infection scores are created to consider county performance compared to the overall region and the specific state to which it belongs. This was done to mitigate the effect of infection waves in specific regions and wildly differing state government policies. This analysis is distinct from those existing works in the literature in which we seek to isolate the demographic effects of COVID-19 at the county level while minimizing the impact of policies and regulations at the state level. To identify significant associated factors, we consider social and economic factors such as educational attainments, employment rates, the rate of high-risk jobs, poverty, and healthcare factors such as the percentage of the population with disabilities or insurance. Moreover, we include other variables include age, race, access to computers, and the Internet in the statistical analysis. A multiple linear regression model fitted by the Least Absolute Shrinkage and Selection Operator (LASSO) method (Tibshirani, 1996) is used to study the association between the infection and mortality scores and various covariates.

From a practical standpoint, the primary goal of this study is to provide actionable insights to the local governments about which demographic factors contribute to the pandemic's severity. These insights will help authorities to determine where to distribute additional COVID-19 aid in future pandemics. The results in this paper will also help the state and national leaders to identify counties that control COVID-19 more effectively so that policies and strategies associated with those counties can be applied on a broader scale in future public health emergencies.

This paper is organized as follows. In Section 2, we discuss the data management and cleaning, and the creation of a new response variable to reflect how well each county controls COVID-19. Section 3 describes the statistical modeling approach with the technical details in model selection, model validation, and the selection of the best hyper-parameter. In Section 4, we apply the proposed statistical model to analyze the COVID-19 data and present the results of two special cases. We also provide our suggestions for using the results from the statistical

analysis to help families, businesses, and communities respond to COVID-19. To make the results easily accessible to the public, we created an interactive system programmed in R shiny (Chang et al., 2022). Finally, some limitations of the current study and concluding remarks are provided in Section 5. The source files and other related materials for this paper to ensure reproducibility are available at the GitHub archive: https://github.com/chriszhangm/ASA-Data-Expo-2021.

2 Data Management and Cleaning

2.1 Data Sources

The target geographic area for this project includes counties in 13 Southern/Southeastern United States with a population of at least 65,000. These 13 states are chosen because of their geographical proximity while possessing a wide range of demographics. There are 328 counties in the 13 Southern/Southeastern United States with a population of at least 65,000 (21 in Alabama, 11 in Arkansas, 41 in Florida, 37 in Georgia, 13 in Kentucky, 17 in Louisiana, 10 in Mississippi, 41 in North Carolina, 11 in Oklahoma, 22 in South Carolina, 20 in Tennessee, 54 in Texas, and 30 in Virginia).

The data used for this paper come from two primary sources. The covariates in the model for each county came from the 2019 American Community Survey Single-Year Estimates, conducted by the United States Census Bureau, and the data are available at https://www.census.gov/newsroom/press-kits/2020/acs-1year.html. Although all data are entered as counts, the data were converted into proportions by adjusting for the weights (provided in the ACS tables) and scaling by the population size. The covariates included in the analysis are listed in Table 1. In two

Table 1: Descriptions of the covariates used in the analysis.

Covariate	Description
Under 5 Years	The proportion of the county population that is under the age of 5
15 to 44 Years	The proportion of the county population that is between ages 15 and 44
65 Years and Over	The proportion of the county population that is age 65 and older
75 Years and Over	The proportion of the county population that is age 75 and older
Bachelor	The proportion of the county population with a bachelor's degree's degree
Disability Rate	The proportion of the county population classified as having a disability
Employment	The proportion of the county population experiencing underemployment (as of 2019)
High Risk	The proportion of the county population with a "High Risk" job
Less than High School	The proportion of the county population with less than a high school education
No Computer	The proportion of the county population without computer access
No Insurance	The proportion of the county population without health insurance
No Internet	The proportion of the county population without internet access
Poverty	The proportion of the county population under the poverty line
Black	The proportion of the county population who racially identify as black
White	Proportion of the county population who racially identify as white

covariates (Employment and High-risk employment), eight counties had missing data (i.e., 16 values are missing in total). Given the small proportion of missing values in these two covariates ($8/328 = 2.44\%$), and the presence of a strong correlation between them, we perform regression imputation to handle the missing parts. Specifically, we regress both employment and high-risk employment on the remaining covariates, utilizing the non-missing observations to impute the missing records.

The COVID-19 infection and death data (by county) was collected from USAfacts.org. The data are listed as counts of new infections/deaths related to COVID-19 by day from 01/27/2020 to 03/27/2021. In the subsequent sections, we have utilized infection/death rates derived by dividing the raw counts by the population of each county.

2.2 Defining a New Response Variable

In this section, we define a new variable, which will be the response variable in our regression model, to reflect how well each county has performed against COVID-19.

While it might seem reasonable to just leave the response variable as infection rate and death rate, this method requires a fixed range of time in which to measure COVID-19 related infections and deaths for each county. While there are numerous interesting endpoints to select (i.e., directly before/after 2020 holiday surge or before/after the 2020 United States election), infections have peaked in different counties during different times. Additionally, a county's infection/death rate likely is affected by the policies and demographics of neighboring counties. Such effects likely cannot be explained by county demographic data. To mitigate these issues and to capture the monthly performance of each county relative to the entire Southeast region and the performance of nearby counties (i.e., the performance of the state), we first created two new score variables.

Let I_{ijk} be the infection rate and n_{ik} be the total population for county i (in state k) in month j. We define

$$I_{\bullet jk} = \frac{1}{\sum_i n_{ik}} \sum_i I_{ijk} \cdot n_{ik}$$

to be the infection rate for state k in month j and

$$I_{\bullet j\bullet} = \frac{1}{\sum_k \sum_i n_{ik}} \sum_k \sum_i I_{ijk} \cdot n_{ik}$$

be the infection rate of the 13 state regions in month j. Then, we define a total infection score for the ith county in state k as

$$\mathrm{SI}_{ik} = \sum_j 2I_{ijk} - (I_{\bullet j\bullet} + I_{\bullet jk}).$$

Similarly, let D_{ijk} be the death rate for county i (in state k) in month j and we define

$$D_{\bullet jk} = \frac{1}{\sum_i n_{ik}} \sum_i D_{ijk} \cdot n_{ik}$$

be the death rate for state k in month j and

$$D_{\bullet j\bullet} = \frac{1}{\sum_k \sum_i n_{ik}} \sum_k \sum_i D_{ijk} \cdot n_{ik}$$

be the death rate of the 13 state regions in month j. We define a total death score for the ith county as

$$\mathrm{SD}_{ik} = \sum_j 2D_{ijk} - (D_{\bullet j\bullet} + D_{\bullet jk}).$$

To create a combined score that considers the infection and death rates, we standardized the infection and death scores by subtracting the scores from their respective mean and dividing by their respective standard deviations. That is, the standardized infection and death scores are calculated as

$$\begin{aligned} \mathrm{SI}^*_{ik} &= \frac{\mathrm{SI}_{ik} - \overline{\mathrm{SI}}}{\sigma_{\mathrm{SI}}} \\ \text{and } \mathrm{SD}^*_{ik} &= \frac{\mathrm{SD}_{ik} - \overline{\mathrm{SD}}}{\sigma_{\mathrm{SD}}}, \end{aligned}$$

respectively, where $\overline{\mathrm{SI}} = \sum_i \sum_k \mathrm{SI}_{ik}/328$ and $\overline{\mathrm{SD}} = \sum_i \sum_k \mathrm{SD}_{ik}/328$. After the standardization, the scores SI^*_{ik} and SD^*_{ik} are centered at 0 and on the same scale.

In order to consider a cumulative effect on both the infection rate and death rate, we further created a variable that is a linear combination of the two standardized scores:

$$S_i = w\mathrm{SI}^*_i + (1-w)\mathrm{SD}^*_i \qquad w \in [0,1], \tag{1}$$

where w is the value to adjust the specific weights to put in the infection and death rates.

The purpose of using these scores was to synthesize two measures into a single score: (i) how well is a specific county performing with respect to the overall region, and (ii) how well has a specific county performed with respect to other counties in the state. The measure for (ii) is especially important in understanding how specific demographic/economic characteristics either overcome weak state-wide COVID-19 policies by outperforming other counties in the state or hinder strong policies by having worse infection/death rates. Additionally, higher infection/death rates in neighboring counties will indeed cause increases in the spread of COVID-19. Therefore, by considering the improvement in infection/death rate at the regional level (i.e., $I_{ijk} - I_{\bullet j\bullet}$) and at the state level (i.e., $I_{ijk} - I_{\bullet jk}$), we can mitigate the effects of state-wide policy while still taking into account the performance of a county with respect to the overall region. Hence, a county that performs better than average with respect to the overall region but worse than average compared to counties in its state will have a score of around 0. Alternatively, counties with lower rates on average at the overall region and state level have scores lower than 0 (and vice-versa).

In the R shiny app (Chang et al., 2022) we developed (will be discussed in Section 4), we allow the user to change the weight w. In this paper, we discuss three special cases with $w = 1$ (only consider the infection score), $w = 0$ (only consider the death score), and $w = 0.5$ (equal weight to both infection score and death score – referred to as total score). Note that a large value of S_i indicates that county i has higher infection/death rates on average compared to the state and the overall region.

2.3 Exploratory Data Analysis

In this subsection, we first look at a univariate geographical representation of the response variable S_i, $i = 1, 2, \ldots, 328$ for the 328 counties. Figures 1, 2, and 3 show the death scores ($w = 0$) infection scores ($w = 1$), and total scores ($w = 0.5$), respectively.

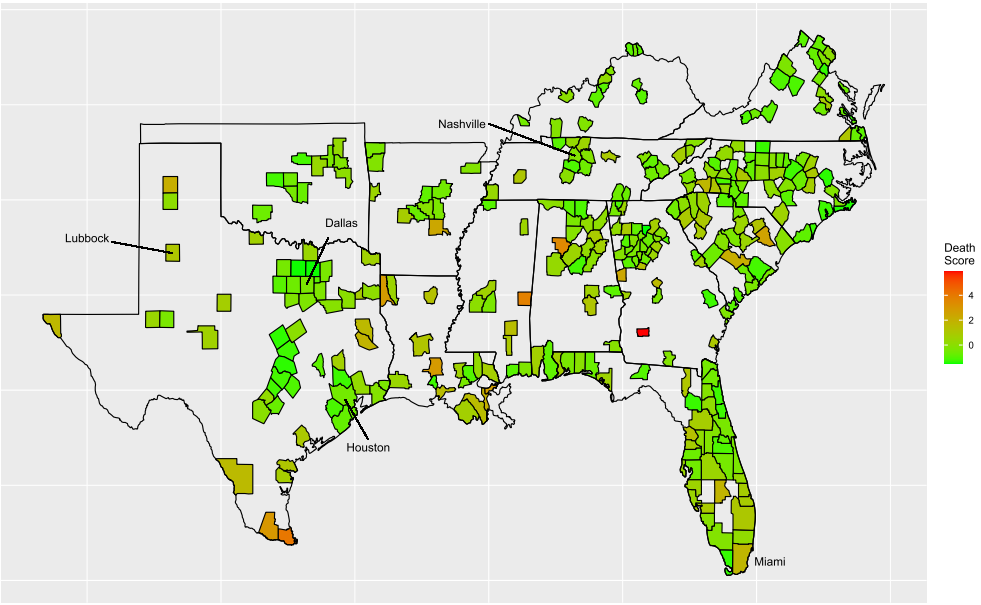

Figure 1: Death scores by county ($w = 0$), where w represents the relative weight of the Infection score (see Eq. (1)). Red counties are those with higher scores, while green counties have lower scores. Most larger cities tend to have lower death scores (except for Miami).

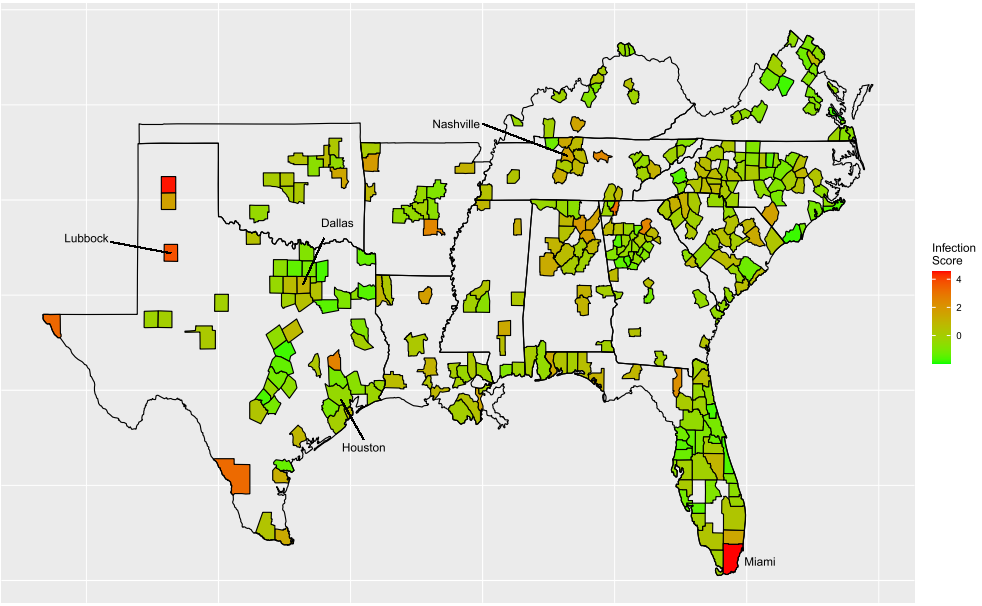

Figure 2: Infection scores by county ($w = 1$), where w represents the relative weight of the infection score (see Eq. (1)). Red counties have higher scores, while green counties have lower scores. Unlike death scores, larger cities tend to have higher infection scores. This is also true for college towns.

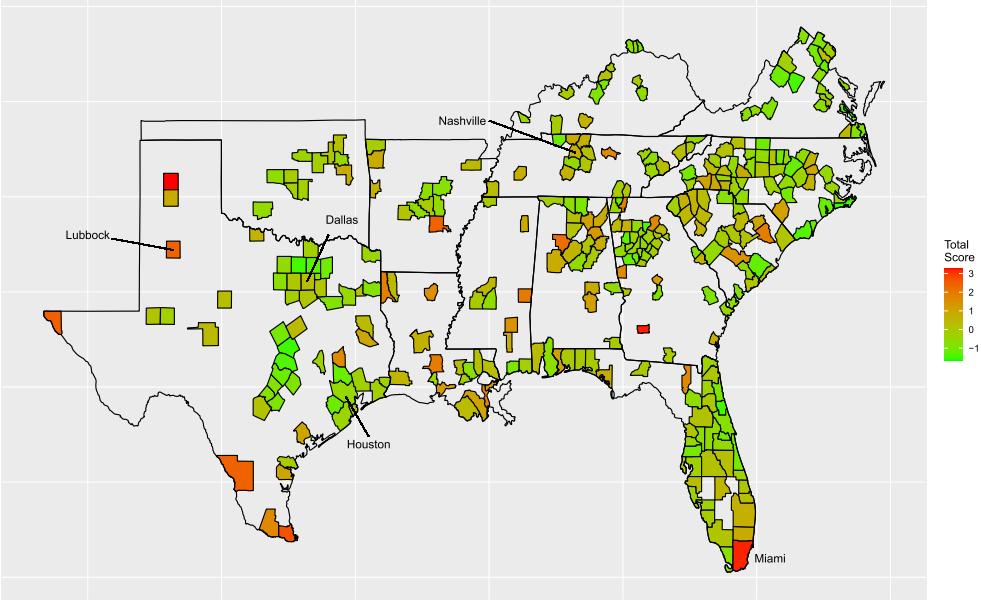

Figure 3: Total scores ($w = 0.5$) by county, where w represents the relative weight of the Infection score (see Eq. (1)). Red counties have higher scores, while green counties have lower scores. Virginia counties tend to perform best overall, even after considering state infection/death rates.

From Figure 1, we observe that large cities have relatively smaller death scores, indicating that large cities have lower death rates than the whole state and region. For example, metro areas such as Dallas-Fort Worth, Houston, and Nashville have very low death scores (Miami, FL is a notable exception to this). However, from Figure 3, these cities have higher infection scores, along with college towns (e.g., Lubbock, TX) and a few Texas counties near the U.S.-Mexico border. Despite being scaled by the state rates, counties in Virginia seem to perform quite well in both the death and infection scores. We also observe that several Florida counties have higher death scores, likely caused by higher populations of older citizens, who have an elevated mortality risk from COVID-19. Finally, from Figure 3, the total score map shows a similar pattern as the death score with bigger population centers performing reasonably well, with Miami again breaking the trend.

Secondly, we explore the relationships between the covariates considered here. Figure 4 presents the correlation matrix for the covariates and the individual death and infection scores. From Figure 4, we observe that the variables "No Computer" and "No Internet" have high correlations to the death score. We can also observe high multicollinearity within the covariates, which is one of the motivations for a LASSO regression (will be discussed in Section 3 for more details). Additionally, we see that the covariates "Less than Highschool" and "Poverty" positively correlate to death and infection scores. Therefore, we believe these variables can capture unique death and infection score information.

Thirdly, we study the relationship between the variable "65 and Over" and the death score, infection score, and total score. Figure 5 shows the scatterplots of the death score, infection score, and total score verse "65 and Over". From Figure 5, we observe that "65 and Over" has a positive relationship with the death score and a negative association with the infection score.

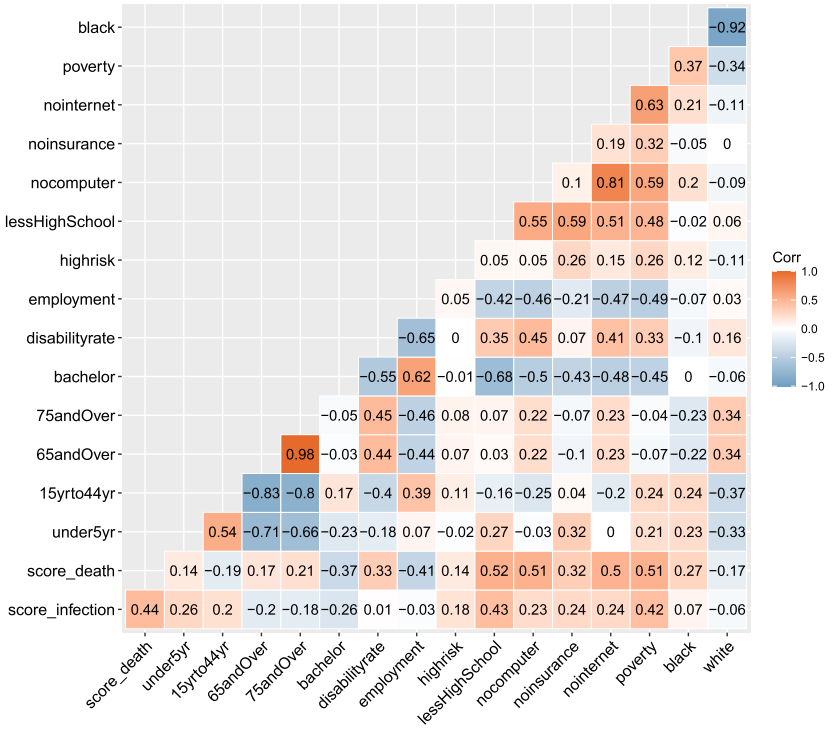

Figure 4: Correlation matrix between the covariates and the total score.

The positive association with death score is likely due to older people having a higher risk of dying from COVID-19, while the negative association with the infection score indicates that older people are going out less and being more careful in general, creating less exposure for older people and lower infection counts for the entire county. In contrast, we observe that the total score washes out these effects, showing the importance of individually observing the infection and death scores along with the combined score.

3 Statistical Analysis

In this section, we describe the statistical analysis of the data and modeling of the response variable "score." The method used for selecting the best hyper-parameter λ is described in the model validation section. R studio (R Core Team, 2022) and the R package "glmnet" (Friedman et al., 2010) are used for constructing and validating the multiple linear regression model.

We start with the multiple linear regression model

$$y_i = \alpha + \sum_{\ell=1}^{p} \beta_\ell x_{i\ell} + \varepsilon_i, \qquad i = 1, 2, \ldots, N,$$

where y_i is the ith response variable (i.e., S_i in Eq. (1) with weight w), β_ℓ is the coefficient for the ℓth covariate, $x_{i\ell}$ is the ℓth covariate for the ith observation, ε_i is the error term for ith observation, $p = 15$ is the number of covariates and $N = 328$ is the total number of counties. To estimate the model parameters α and $\beta = (\beta_1, \beta_2, \ldots, \beta_{15})$, we consider the LASSO method (Tibshirani, 1996) which performs both variable selection and regularization to improve the prediction accuracy of the regression model. Specifically, the LASSO estimates of the intercept

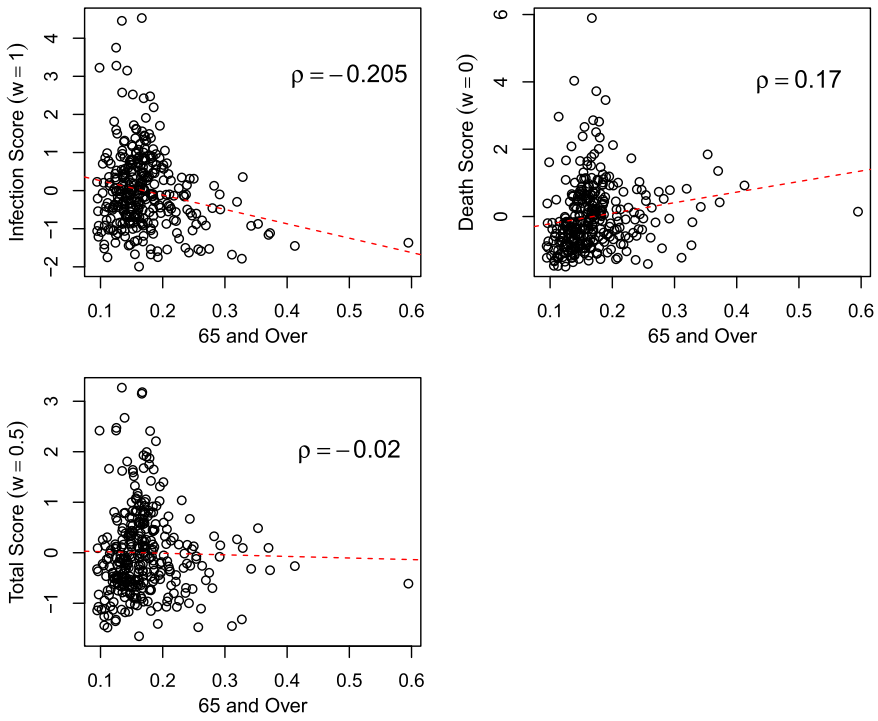

Figure 5: Scatter plots for infection score, death score, and total score versus the variable "65 and Over" (proportion of county citizens in this age range), where each marker represents a county. The relationship between this variable and score changes drastically depending on the specified value of w, with a significantly non-zero ($p < .001$) negative correlation when $w = 1$, a significantly non-zero ($p = .002$) positive correlation when $w = 0$ and virtually no linear relationship ($p = 0.72$) when $w = 0.5$.

α and the regression coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_{15})$, denoted as $\hat{\alpha}_{\text{lasso}}$, $\hat{\boldsymbol{\beta}}_{\text{lasso}}$, are obtained as

$$\left(\hat{\alpha}_{\text{lasso}}, \hat{\beta}_{\text{lasso}}\right) = \arg\min_{(\alpha,\beta)} \left\{ \sum_{i=1}^{328} \left(y_i - \alpha - \sum_{\ell=1}^{15} \beta_\ell x_{i\ell} \right)^2 + \lambda \sum_{\ell=1}^{15} |\beta_\ell| \right\},$$

where λ is a regularization term to control the model size. For instance, $\hat{\alpha}_{\text{lasso}}$ and $\hat{\boldsymbol{\beta}}_{\text{lasso}}$ are equivalent to the ordinary least squares (OLS) estimates of α and β, respectively, when $\lambda = 0$, while all coefficients are equal to 0 when $\lambda = 1$.

Compared to the conventional ordinary least-squares method for fitting a multiple linear regression, the LASSO method offers the advantage of allowing us to select the value of λ to regulate the model size. This capability not only helps in mitigating overfitting and multicollinearity issues but also makes the model more interpretable and parsimonious.

To determine the optimal value of λ in our model, we utilize the leave-one-out cross-validation method (Hastie et al., 2009). This technique is particularly recommended when working with small sample sizes, such as the data we have. We then select the λ such that the average mean squared error (AMSE) for the $N = 328$ folds is minimized. In this study, we divide the data into 328 parts, with each part containing exactly one observation. We then merge 327 of them to create the training set for model construction, and the remaining subject is reserved for

testing the model's performance. We utilize the MSE as the metric to measure the performance since our response variable S_i is continuous. Specifically, the MSE of the mth fold of the data is used as the test set is defined as:

$$\text{MSE}_m(\lambda) = \frac{1}{328} \sum_{i=1}^{328} \left(y_i - \hat{y}_i\right)^2,$$

where $m \in \{1, 2, \ldots, 328\}$, and y_i and $\hat{y}_i$ are the true value and the predicted value of the ith observation, respectively. Note that the MSE is a function of λ since different choices of λ would result in different MSEs. The AMSE for a particular value of λ is

$$\text{AMSE}(\lambda) = \sum_{m=1}^{328} \text{MSE}_m(\lambda)$$

and we choose the value of λ such that AMSE(λ) is minimized.

4 Major Results

To make our results in this paper easily accessible to the general public, we constructed an interactive system on a website at https://asa-competition-smu.shinyapps.io/COVID19/, which is programmed in R shiny (Chang et al., 2022). In Section 4.1, we describe the features of the interactive system and how to use it to get some meaningful results. Then, in Section 4.2, we show the details of the interactive system using two special cases when $w = 0$ and $w = 1$ to illustrate how the system can actually be used to identify some crucial significant factors that contribute to the county's performance in controlling the pandemic. As a result, in Section 4.3, we discuss how local governments can refer to the factors identified in our statistical analysis to adjust some strategies and policies in a different county to protect their residents better.

4.1 Delivery of the Results

We develop an interactive system for the public to access the results presented in this paper. Figure 6 presents the interface of the created R Shiny app (Chang et al., 2022). There are eight tabs available in the system: Data summary, Data variables, Correlation plot, Scores by county (Geographic graph), LOO-CV model selection result, Coefficients, Top 10 (Table), Top 10 (Geographic graph), and two interactive features: "Weight" and "log(λ)".

The brief descriptions of the eight tabs are provided as follows:

- The "Data summary" tab provides descriptive statistics of both independent and dependent variables.
- The "Data variables" tab shows each variable's definition, allowing users to understand the predictors and response variables without reviewing the information in this paper.
- The "Correlation plot" tab shows the Pearson correlation coefficient between the variables, which provides a big picture of all the variables and illustrates the high level of multicollinearity in the data.
- The "Scores by county (Geographic graph)" tab visualizes scores by county geographically.
- The tab "LOO-CV model selection" allows the users to obtain a figure of the average of mean squared error (AMSE) with its error bars for each value of λ using the leave-one-out cross-validation (LOO-CV) method. In the figure provided in the system (see Figure 7 as an

Figure 6: Interface of the R Shiny app at https://asa-competition-smu.shinyapps.io/COVID19/.

example), two vertical dashed lines are presented for different purposes of model selection, where the left dashed line selects the best value of λ, λ_{min}, with the highest predictive power. The right dashed line selects the value of λ that gives the most regularized model (denoted as λ_{lse}) such that the error is within one standard error of the minimum. Following the suggestions in Tibshirani (1996), we choose λ_{min} as the value of λ for the LASSO regression model, which mitigates the overfitting issue.

- The tab "Coefficients" shows all significant variables and their coefficients.
- The tab "Top 10 (Table) and (Geographic graph)" provides a table and a geographic graph containing the top 10 overperforming and underperforming counties, which have the smallest and largest residuals after fitting the model, respectively.

Two interactive features allow users to dive deeper into the data and further access more results from the statistical analysis, and users may use the sliders to adjust either "Weight" or "log(λ)":

- "Weight" allows users to build their response variable by adjusting the weight w in Eq. (1). In other words, users can put more weight on scores related to the death rates if they believe the performance of controlling COVID-19 is largely determined by how counties react to the mortality rates.
- "log(λ)" allows users to adjust the range of $\log(\lambda)$, which provides different cross-validation processes. For instance, the model size will be larger with the smaller $\log(\lambda)$, and users will have a parsimonious model by choosing larger values of $\log(\lambda)$. We recommend that the range of $\log(\lambda)$ is kept between -2 and 0 to get a model with a small sample size while keeping the predictive accuracy high.

4.2 Results of Two Special Cases

Since there is flexibility in choosing the weight w in Eq. (1), we illustrate the results of the statistical analysis by considering the cases where $w = 0$ and $w = 1$. We compare the model selection processes and selected variables for these two special cases. In Figure 7, we present the

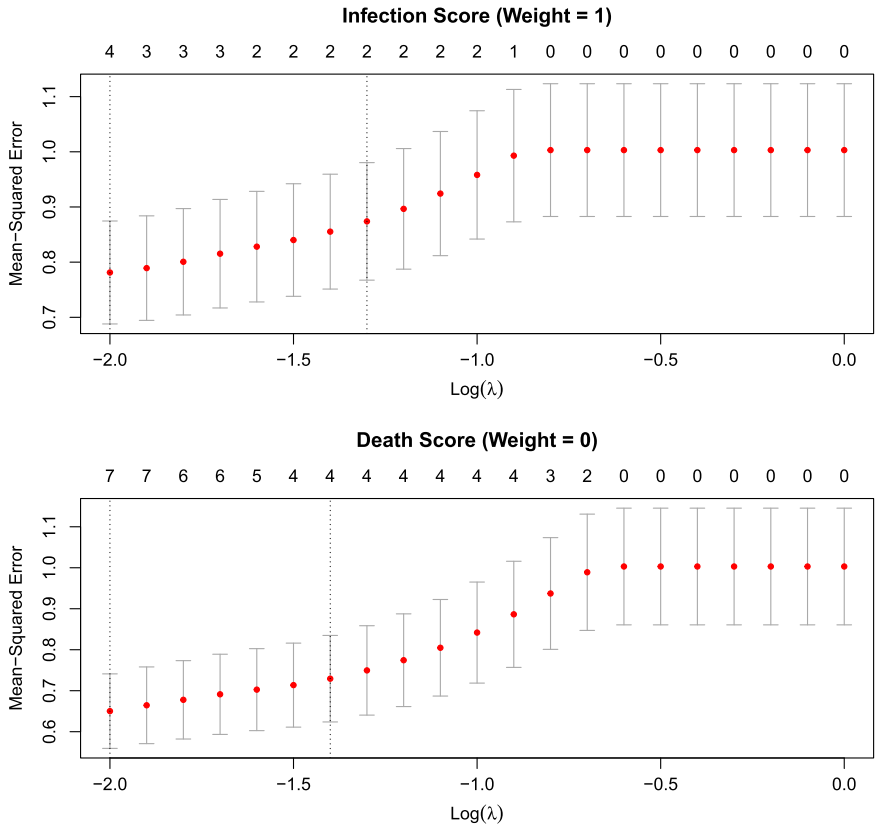

Figure 7: Leave-one-out cross-validation results for infection score (weight $w = 1$) and death score ($w = 0$).

results when $\log(\lambda) \in [-2, 0]$, we observe that the minimum average MSEs is achieved when $\log(\lambda) = -2$ or $\lambda_{\min} = e^{-2}$ for both special cases.

Given the selected hyper-parameter λ, Table 2 presents the selected variables based on the LASSO regression model with $w = 0$ and $w = 1$. We observe that counties are likely to have relatively poor control of the COVID-19 infection rates (i.e., $w = 1$) with a higher percentage of young people (the coefficient estimate for "15 years old to 44 years old" is 0.06), a lower education level (the coefficient estimate for "Less than High School" is 7.17) and a higher proportion of people under the poverty line (the coefficient estimate for "Poverty" is 3.26). However, when it comes to the performance of controlling the COVID-19 death rates (i.e., $w = 0$), we find that counties with a higher percentage of elder people (the coefficient estimate for "75 years old and over" is 1.41), a lower education level (coefficient estimate for "Less than High School" is 7.83), a higher proportion of people under the poverty line (the coefficient estimate for "Poverty" is 2.83), a higher percentage of people without computers (coefficient estimate of "No Computer" is 4.85) or the internet (coefficient estimate of "No Internet" is 3.02) are at a higher risk of COVID-19 death. Moreover, African Americans are more vulnerable to death due to COVID-19 than other ethnic groups (the coefficient estimate of "Black" is 0.46).

Table 2: Variables selected by the model for infection score (weight $w = 1$) and death score ($w = 0$).

Variables	Infection score ($w = 1$)	Death score ($w = 0$)
Intercept	−0.93	−1.36
15 to 44 Years	0.06	–
65 Years and Over	−1.14	–
75 Years and Over	–	1.41
Employment	–	−0.49
Less than High School	7.17	7.83
No Computer	–	4.85
No Internet	–	3.02
Poverty	3.26	2.83
Black	–	0.46

4.3 Helping Families, Businesses, and Communities Respond to COVID-19

From our results, the two most significant variables that showed up in the two special cases when $w = 0$ and $w = 1$ (i.e., the model purely based on the death score and the infection score) are the proportion of the population with less than a high school education and the proportion of the population under the poverty line. The following two most significant variables that only showed up in the death score model are the proportions of the people with no internet and no computer. The two variables on proportions of the population with no internet and no computer are highly correlated. Moreover, further analysis would require determining if these variables are only significant due to their correlation with old age or if they are significant enough by themselves. From the LASSO regression models we built, education is a significant factor separating counties that handled COVID-19 well versus those that did not. Local governments could use this information to target counties with high proportions of less than high school education and those under the poverty line and then provide better education regarding COVID-19 and the vaccines. While educating people about COVID-19 is not the most straightforward endeavor, this is the best course of action to take from our results, particularly at the local government level.

5 Conclusions and Discussion

In this paper, we present a statistical analysis based on a LASSO multiple linear regression model to determine the effects of socioeconomic and healthcare factors on the performance of controlling COVID-19 in the Southern and Southeastern United States. Our statistical analysis indicates that education and poverty levels emerge as two crucial factors in determining how well a county managed COVID-19 as these two factors were always selected by our model with different weights that control the proportion of infection and death rates. Specifically, our study discovered an association between higher infection and death rates in counties characterized by larger populations with low education and high poverty levels. Our results also show that a higher proportion of elderly people results in a lower infection rate and a higher death rate. The strength of this relationship changes from state to state, with Florida illustrating this concept the best. Moreover, counties bordering populous counties typically controlled infection rates

better. In contrast, counties on borders between states and counties that contained a "college town" typically underperformed in controlling the infection rates. These trends should be kept in mind for any future statistical analysis done on the counties in the United States.

There are both similarities and differences between our findings and those reported in the relevant literature. For instance, Mollalo et al. (2020) focused on COVID-19 incidence rates and identified income inequality as the most influential factor. Abedi et al. (2021) examined both incidence and mortality rates and concluded that higher education attainment was associated with higher infection rates but lower death rates. In contrast, higher poverty levels were linked to lower infection rates and higher death rates. Several plausible reasons contribute to the differences between their findings and ours. Firstly, they used data from different states, where people's behaviors and attitudes toward the pandemic may vary significantly. This regional variation could account for differences in the impact of various factors on COVID-19 outcomes. Secondly, their studies did not incorporate a standardized process for infection and death rates, nor did they consider spatial dependence. In contrast, we considered these factors, which may have influenced the observed associations between variables. Lastly, it is worth noting that there may be highly correlated variables in their analyses, which could lead to multicollinearity issues. Employing methods to address multicollinearity is crucial to ensure an accurate and meaningful interpretation of the coefficient estimates. Moreover, Karmakar et al. (2021) recently conducted analyses on COVID-19 incidence and death rates, finding that a 0.1 point increase in the Social Vulnerability Index (SVI) score was associated with a 14.3% increase in the incidence rate and a 13.7% increase in mortality rate. The SVI, developed by the Centers for Disease Control and Prevention (Flanagan et al., 2011), encompasses subindices including poverty and low education levels. Similar conclusions were also reported by Clouston et al. (2021), further supporting the association between social vulnerability factors and COVID-19 outcomes.

Our study had several limitations; we only had access to data from counties with a population greater than 65,000. Part of our results can be extrapolated to counties in the Southeastern United States with smaller populations. For instance, if a local government has good reason to believe a county with a population of about 25,000 has a large proportion of people with less than a high school education, the government officials should still launch education efforts for the county of interest. Moreover, we only considered the counties in the Southern and Southeastern United States. Expanding the statistical analysis presented in this paper to other regions of the United States may bring up several problems, with the major one being the differences in mask and lock-down policies between different regions. Lastly, while we took into account the possibility of spatial dependence by creating new score variables, further enhancing our model's flexibility and interpretability can be achieved by incorporating spatial factors through various spatial regression models (Ward and Gleditsch, 2018). By leveraging these approaches, we can better capture the spatial relationships and potential spatial autocorrelation in our data, leading to more comprehensive insights and a deeper understanding of the underlying mechanisms influencing the results.

Our R-shiny app and the proposed approach were specifically designed for COVID-19, which, as one of our reviewers pointed out, is no longer a global emergency. However, our tool can easily be adapted to address other emergencies, such as flu outbreaks. To achieve this, we need to gather infection and death data related to the flu and define a variable (e.g., Score) that combines infection and death rates, similar to what we have done for COVID-19 in this paper. By doing so, we not only have the ability to measure the impact of socioeconomic and healthcare factors quantitatively, but we can also analyze the spatial trends of different counties using the geographic graph provided in our app.

Supplementary Material

S1. Code

To ensure the reproducibility of the results presented in this manuscript, the following supplementary materials are provided at the GitHub archive https://github.com/chriszhangm/ASA-Data-Expo-2021:

- `Data_clean.R`: The R code for data cleaning;
- `modeling.R`: The R functions to show results in our paper and R shiny website.
- `app.R`: The R code to run the R shiny website;
- `full_data.csv`: full data set includes two response variables (score_infection, score_death) and socioeconomic and healthcare factors.
- `counties_prj.csv & states_SE.csv`: Two datasets for producing geographic graphs in the R shiny website.

Acknowledgement

The authors would like to thank the Editor and the referees for their valuable comments, which helped to improve the quality of this article.

References

Abedi V, Olulana O, Avula V, Chaudhary D, Khan A, Shahjouei S, et al. (2021). Racial, economic, and health inequality and COVID-19 infection in the United States. *Journal of Racial and Ethnic Health Disparities*, 8: 732–742. https://doi.org/10.1007/s40615-020-00833-4

Bilinski A, Emanuel EJ (2020). COVID-19 and excess all-cause mortality in the US and 18 comparison countries. *JAMA: The Journal of the American Medical Association*, 324(20): 2100–2102. https://doi.org/10.1001/jama.2020.20717

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. (2022). *shiny: Web Application Framework for R*. R package version 1.7.3.9001.

Cheng KJG, Sun Y, Monnat SM (2020). COVID-19 death rates are higher in rural counties with larger shares of blacks and hispanics. *The Journal of Rural Health*, 36(4): 602–608. https://doi.org/10.1111/jrh.12511

Clouston SA, Natale G, Link BG (2021). Socioeconomic inequalities in the spread of coronavirus-19 in the United States: A examination of the emergence of social inequalities. *Social Science & Medicine*, 268: 113554. https://doi.org/10.1016/j.socscimed.2020.113554

Desmet K, Wacziarg R (2022). JUE insight: Understanding spatial variation in COVID-19 across the United States. *Journal of Urban Economics*, 127: 103332. https://doi.org/10.1016/j.jue.2021.103332

Doti JL (2021). Examining the impact of socioeconomic variables on COVID-19 death rates at the state level. *Journal of Bioeconomics*, 23(1): 15–53. https://doi.org/10.1007/s10818-021-09309-9

Flanagan BE, Gregory EW, Hallisey EJ, Heitgerd JL, Lewis B (2011). A social vulnerability index for disaster management. *Journal of Homeland Security and Emergency Management*, 8(1), Article 3.

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22. https://doi.org/10.18637/jss.v033.i01

Haeder SF, Gollust SE (2020). From poor to worse: Health policy and politics scholars assessment of the U.S. COVID-19 response and its implications. *World Medical and Health Policy*, 12(4): 454–481. https://doi.org/10.1002/wmh3.371

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning*, chapter 7, 241–243. Springer, New York, twelve edition.

Karmakar M, Lantz PM, Tipirneni R (2021). Association of social and demographic factors with COVID-19 incidence and death rates in the us. *JAMA Network Open*, 4(1): e2036462. https://doi.org/10.1001/jamanetworkopen.2020.36462

McLaren J (2021). Racial disparity in COVID-19 deaths: Seeking economic roots with census data. *The B.E. Journal of Economic Analysis & Policy*, 21(3): 897–919. https://doi.org/10.1515/bejeap-2020-0371

Mollalo A, Vahedi B, Rivera KM (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the Total Environment*, 728: 138884. https://doi.org/10.1016/j.scitotenv.2020.138884

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Ward MD, Gleditsch KS (2018). *Spatial Regression Models*, volume 155. Sage Publications.

Woolf SH, Chapman DA, Lee JH (2021). COVID-19 as the leading cause of death in the United States. *JAMA: The Journal of the American Medical Association*, 325(2): 123–124. https://doi.org/10.1001/jama.2020.24865