Application of Elastic Net Regression for Modeling COVID-19 Sociodemographic Risk Factors

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Statistical Science

in the

College of Graduate Studies

University of Idaho

by

Tristan A. Moxley

Major Professor: Benjamin Ridenhour, Ph.D.

Committee Members: Jennifer Johnson-Leung, Ph.D.; Christopher Williams, Ph.D.

Department Administrator: Hirotachi Abo, Ph.D.

May 2022

Abstract

COVID-19 has been at the forefront of global concern since its emergence in December of 2019. Determining the social factors that drive case incidence is paramount to mitigating disease spread. Simple predictive analysis in the form of multiple regression proves to be an inefficient method for predicting COVID-19 case rate using sociodemographic factors, as many of these factors are collinear; additionally, multiple regression is insufficient as this technique results in models that overfit the data, meaning the models cannot generalize when given new data and thus perform poorly. As such, biased estimation through elastic net regression was used to conduct a broad-based analysis across the ten HHS health regions for both the pre-Delta (March 22, 2020 to June 15, 2021) and Delta (June 15, 2021 to November 1, 2021) waves of the COVID-19 pandemic. Statistically, elastic net proved to be much more accurate in its prediction when compared to multiple regression, as almost every HHS model consistently had a lower root mean square error (RMSE); additionally, these models also succeeded in remedying overfitting through verification by way of training/testing R^2 evaluation. From an epidemiological standpoint, this research confirmed many of the known trends in terms of social factors that influence case incidence (such as group quarters percentage or mobile home percentage per county), while also discovering interesting trends occurring across different waves of the pandemic that give insight into the effect of measures such as vaccination. This research provides a novel approach to modeling sociodemographic risk factors against COVID-19 case rate which can easily be expanded upon in the future with a more robust set of sociodemographic factors.

Acknowledgements

I would like to thank Dr. Benjamin Ridenhour for his continuing support in my academic endeavors as my major professor. Taking on research of this magnitude would not have been possible without him, so I am forever grateful for that guidance, support, and expertise. I would also like to thank Dr. Erich Seamon, whom I worked under as a research assistant, and whom was the main inspiration for me conducting this research. A special thanks to my committee, Dr. Christopher Williams and Dr. Jennifer Johnson-Leung, for not only supporting me through the scope of my thesis, but for also being invaluable sources of knowledge. Having the opportunity to learn from such bright minds has been an absolute honor. I would like to also thank the Institute for Modeling Collaboration and Innovation for allowing me the opportunity to learn and contribute to COVID-19 discourse as a research assistant. Finally, I want to give thanks to Jana Joyce, Jaclyn Gotch, Melissa Gottschalk, and the rest of the staff within the Department of Mathematics and Statistical Science, for financially supporting me and keeping me on track through the course of my Master's degree. I couldn't have made it to this point without you; I am eternally grateful for your support.

Dedication

I dedicate this thesis to my wonderful girlfriend Jane, who has been an endless source of love and support for me as I've fought through these two years to obtain my Master's degree; to Ana, for being a wonderful friend and a great motivator in helping me finish my thesis; to my parents Michele and Sean, and to my brothers Ian and Michael, who have been supportive of me throughout my life and who encouraged me to get to where I am today; and to my wonderful cats, Hobbes, Gizzy, Dewey, and Mr. Baby for giving me affection to break up the stress and monotony of the long writing sessions.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

COVID-19, formally known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has made a significant impact on the world since its emergence in December of 2019, with a global case total of approximately 409 million cases and global death total of approximately 5.8 million deaths as of February of 2022 (The New York Times 2022). Coronavirus has seen various mutations throughout the course of its spread, with the most recent strain, Omicron, resulting in what appears to be vastly higher transmissibility and reinfection than for previous variants, though it is too early to tell with much scientific certainty as of the time of writing (Araf et al. 2022). Efforts to quell the pandemic to this point have included quarantining, mask mandates, social distancing measures, and vaccines. Developing vaccines, along with the other aforementioned preventative measures, has been the primary focus in eliminating COVID-19 from the general population since the onset of the pandemic in late 2019. However, in spite of the current abundance of vaccines and two years of ongoing quarantining procedures, the pandemic has spread globally in high numbers, especially in the United States, which, outside of having the highest total confirmed case count by a wide margin, just experienced a spike in Omicron cases with a single-day peak of just over 1,000,000 daily cases on January 24, 2022 (The New York Times 2022). These confirmed case numbers are significantly higher than other nations of comparable or larger population sizes. For instance, India and Brazil, the two nations which have the second and third most total confirmed COVID-19 cases, respectively, averaged 300,000 and 100,000 daily cases at their respective peaks from Omicron during that same time frame in late January 2022 (The New York Times 2022). What is important to note is that nations like India have a lower proportion of fully vaccinated individuals (55.16%) when compared to the United States (64.40%) (Ritchie et al. 2020). Even at an international level, the COVID-19 pandemic is an anomaly in an epidemiological sense, as wealthier countries with better health infrastructure have been hit much harder than their less wealthy counterparts (Bollyky et

al. 2022). Of course, disparities in case reporting in these less wealthy countries certainly may influence this trend, which is an important factor to acknowledge when viewing any epidemiological data. However, it is still reasonable to expect that nations with more wealth and better access to vaccines would have less disease burden overall, since studies have shown enhanced immune response when completing a full course of the COVID-19 vaccine (Bates et al. 2022). Thus, a search of which factors influence infection outside of reporting disparities may still be worthwhile.

Preventative measures show positive results in mitigating disease spread; social distancing measures effectively reduce the likelihood of contracting COVID-19 (Fazio et al. 2021), and vaccinations reduce disease severity, hospitalization risk, and deaths (Centers for Disease Control and Prevention 2021). In looking at potential risk factors for COVID-19 outside of preventative measures, social factors present themselves as a rational point of analysis, since environment is one of the pillars of epidemiological spread as mapped by the host-agent-environment model (Tsui, Deng, and Pan 2020). Two avenues present themselves when thinking about social factors: extrinsic social factors (i.e., qualities and behaviors that one acquires or changes throughout their daily lives, such as political ideology, metaphysical worldview, etc.), and intrinsic factors (i.e., factors that one cannot inherently change, or rather, cannot change with ease; these may include race, ethnicity, socioeconomic status, etc.).

One motivating example of an extrinsic social effect is the effect of political affiliation on COVID-19 response and behavior. COVID-19 rapidly became a highly-politicized issue in the U.S. as the initial spread of the disease progressed. The rapid spread and endorsement of misinformation regarding disease severity and origin (Motta, Stecula, and Farhart 2020), and the overall influence of the "anti-intellectual" ideology (Merkley and Loewen 2021) along partisan lines has resulted in a polarized divide in response to the pandemic in the United States, with the anti-intellectual side contributing to more lax, and often more risky, COVID-19 behaviors. Though not verifiable with absolute certainty, many of the individuals on the

anti-intellectualism side appear to be ideological conservatives and religious fundamentalists, as they "may see [scientific] experts as threatening to their social identities" (Merkley and Loewen 2021). This presents itself both at the national and state level. At a national level, for instance, the conservative individuals who cited former President Donald Trump and his task force as their primary news and information outlet were far less likely to receive a dose of the vaccine (Jurkowitz and Mitchell 2021), though these trends only serve to strengthen previously held concerns regarding vaccines, ones that traditionally do not break along partisan lines. A study conducted on Google trends data by Pullan and Dey reveal that the topic of COVID-19 vaccines accentuate previously established hoaxes on MMR vaccines containing mercury or causing autism (Pullan and Dey 2021). In other words, anti-intellectualism did not solely create vaccine hesitancy in the United States, but rather it exacerbated previously-held beliefs in individuals who had established concerns regarding vaccines, and contributed to individuals being more outspoken about vaccine misinformation along partisan lines. It still is a concern, though; anti-intellectualism has been a staple and a sort of "rallying cry" behind former President Donald Trump's rhetoric from the outset of his presidential campaign in 2016 (Motta, 2017), and these negative perspectives toward science that are common in anti-intellectualism can assist in spreading misinformation regarding treatments to the pandemic such as hydroxychloroquine and ivermectin (Mackey et al. 2021). However, this political divide for COVID-19 response goes beyond the national level. Differences in pandemic response split down the political divide at the state level, as well, with democratic states having more aggressive response to the disease (Grossman et al. 2020). Idaho, for instance, has had little to no state-wide regulation on prevention of COVID-19 through active precautions such as mask and vaccine mandates, and as such have suffered high spikes in cases and deaths due to the emergence of mutations such as Delta and Omicron (Baker 2022). It appears that these later variants, namely Delta and Omicron, have overall severity that is much more individually dependent when compared to earlier strains due to vaccine effectiveness on mitigating disease incidence (Bernal et al. 2021) and

the impact of political leaning on vaccination intent (Jurkowitz and Mitchell 2021).

Intrinsic social factors have broad-reaching effects on pandemic spread as well, although many of these factors are confounded and intertwined. Over the course of the pandemic, several studies have shown an adverse relationship between being a racial or ethnic minority (e.g., African American, Indigenous, Hispanic, etc.) and infection rate; this translates as well to higher incidences of severe cases and deaths (Gayle and Childress 2021). This increased disease burden on racial/ethnic minorities has several contributing factors. For example, some of the underlying trends in this increased risk for racial/ethnic minorities stems from factors such as higher rates of comorbidity, living in more crowded living conditions (Hooper, Nápoles, and Pérez-Stable 2020), and decreased ability to social distance due to working lower-paying interpersonal jobs throughout the pandemic (i.e., jobs in critical retail, transportation, agriculture, etc.) (Gayle and Childress 2021). Clearly, many of these intrinsic COVID-19 trends within the United States stem from other underlying social factors. In terms of racial disparity in the healthcare space, this result is not entirely unprecedented; for example, African Americans are eight times more likely to contract HIV compared to Caucasians on average, yet coverage of pre-exposure prophylaxis for treating HIV is seven times higher in Caucasians than in African Americans (Harris et al. 2019). This racial disparity in healthcare has existed prior to the COVID-19 pandemic, so clearly it is an issue to address in terms of modeling disease spread. Additionally, we can deduce that many underlying factors exist and are covarying, given that we know from Hooper, Nápoles, and Pérez-Stable (2020) that the disproportionate disease burden on racial/ethnic minorities stems from various underlying factors. Thus, it is within reason to believe that many other intrinsic social factors should play a similar role in COVID-19 spread and prophylaxis, and exploring what these specific effects are is of interest for this research.

Socioeconomic disparity in health crises is not a new issue when it comes to severity and impact. Setting a baseline at an international level to understand the global impact of socioeconomic disparity on COVID-19 is beneficial for understanding more localized effects.

According to Our World in Data, low-wealth nations have just above a 10% vaccination rate overall, which is critically low compared to nations like the United States which have relatively high vaccination rates and vaccine abundance (Ritchie et al. 2020). Despite the fact that higher-wealth countries have had a greater COVID-19 burden overall as mentioned previously, it is important to note the broad-reaching disparity as vaccines are proven to be effective. Furthermore, socioeconomic disparity negatively effects disease burden for individuals within these wealthier nations. At a national level, for instance, poverty and low income within the United States may have a damaging impact on disease burden and effectiveness in preventing infection, as many low-income individuals had to work in in-person jobs that resulted in them being vulnerable to the disease. Additionally, due to being lower-income, these individuals generally have higher rates of comorbidity and generally live in more crowded living spaces, both of which are confounding effects that result in hightened disease burden (Little et al. 2021). Given how these confounding factors increase disesase burden for COVID-19, it becomes a matter of great concern to map out the effect these intrinsic social factors have across the United States, so that ample resources are provided to everyone, regardless of inherent wealth, previous afflictions, and socioeconomic status. Interestingly, preliminary studies have been inconclusive in terms of direct analysis up to this point; the aforementioned retrospective study conducted by Little et al. (2021) determined that poverty was not a contributing factor to higher rates of hospitalization, in spite of the presence of many dangerous pre-existing conditions these more impoverished individuals have as a whole. Little et al. (2021) also identify that racial/ethnic minorities are unfavorably predisposed to negative COVID-19 burden due to confounding social factors. Though many of these results seem conflicting, this study aims to take a different approach to analyzing these sociodemographic factors with much more robust methods.

The contents of this thesis will look into the effect of sociodemographic factors on COVID-19 incidence. The analyses will first look into a meaningful and practical method to organize data to capture sufficient pandemic information across the entire United States. From there,

model development will be discussed and explored, looking into an appropriate method to analyze the data. Finally, the implications of the modeling will be discussed, along with the potential contribution to the broader body of COVID-19 research.

CHAPTER 2

Methods

2.1 Data and Analysis Overview

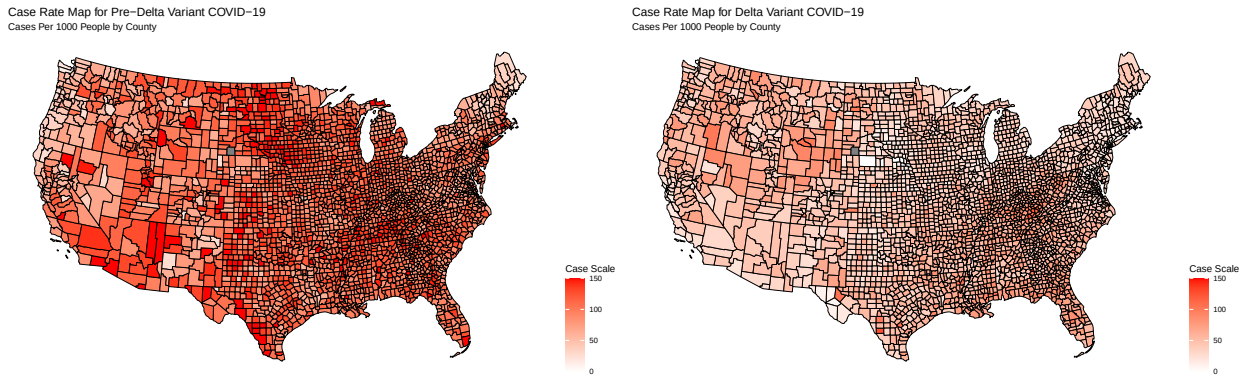
Data selection for this research involves choosing a source of sociodemographic factors that is comprehensive, reliable, and is measured at a sufficiently fine scale (i.e., county-level as opposed to state level) to capture sufficient variability by region. An analysis conducted by Karmakar, Lantz, and Tipirneni (2021) utilized the CDC's Social Vulnerability Index (SVI) in an exploratory sociodemographic analysis using negative binomial regression. In an effort to expand upon previous research, this paper will also use SVI as the primary source of data. The CDC's Social Vulnerability Index is a metric developed through synthesizing 15 census variables in order to map potential negative community effects from external forces such as natural disasters and public health crises (CDC/ATSDR 2018). The index is organized into four distinct categories each with an overarching "theme": socioeconomic status, household composition/disability, minority status/language, and housing type/transportation. Each of these larger categories have more focused metrics of vulnerability, such as poverty, minority status, education level, age demographics, and crowded living spaces. The Social Vulnerability Index is designed specifically with public health and safety in mind; for instance, the overall categorical scores calculated through SVI can help governmental agencies determine which communities need more resources and supplies during an emergency situation, as well as where to evacuate people in the case of a natural disaster. In essence, SVI is designed for determining the locations of need before, during, and after an emergency event.

This thesis will utilize the individual SVI categories to represent various intrinsic social factors, as well as per-county voting percentages from the 2020 U.S. presidential election as a proxy for the overall effect of political leaning. Each explanatory variable will be the percentage-by-county metric in order to keep standardization across varying county sizes. The full set of explanatory data for this thesis is given in Table 2.1.

Table 2.1: Explanatory variable abbreviations and their descriptions for SVI and voting percentage

Variable	Description
POV	Percentage of individuals below poverty estimate
UNEMP	Percentage of unemployed individuals
NOHSDP	Percentage of individuals with no high school diploma
AGE65	Percentage of individuals over the age of 65
AGE17	Percentage of individuals under the age of 17
DISABL	Percentage of individuals with a non-institutionalized disability
SNGPNT	Percentage of individuals who are single parents with a child below the age of 18
MINRTY	Percentage of minorities (all persons except white, non-hispanic)
LIMENG	Percentage of individuals over the age of 5 who speak English "less than well"
MUNIT	Percentage of housing in structures with 10 or more units
MOBILE	Percentage of mobile homes
CROWD	Percentage of occupied housing units with more people than rooms
NOVEH	Percentage of households with no vehicle available
GROUPQ	Percentage of persons in group quarters
pct	Percentage of democratic vote in 2020 Presidential Election

The response variable for all analyses will be cumulative case counts per 1000 individuals (calculated by dividing cumulative cases by county population, then multiplying the standardized result by 1000). The previous sociodemographic COVID-19 analysis by Karmakar, Lantz, and Tipirneni (2021) utilized negative binomial regression due to having limited county replicates (i.e., some counties, at the time of the aforementioned analysis, had little to no cases of COVID-19). Given that data are available for two complete "waves," this model restriction is not necessary. For overall model organization, creating one model with



(a) Pre-Delta (b) Delta

Figure 2.1: Heat maps for cumulative COVID-19 case rate for pre-Delta and Delta waves

each state using counties as replicates can limit the presentation of regional variance, since many areas of the country have different health infrastructures and disease dynamics. To remedy this, model subdivisions will be made using the Department of Health and Human Services (HHS) regions as a basis to attempt to capture some of that regional variability (note that all analyses will only be done for the continental United States). The use of HHS regions maintains national-level healthcare policy structure, making for an informed method of subdividing the United States for this thesis. The regions are generally known by number, but for the purposes of this thesis will be given names referring to their general geographic location. The plot of the HHS regions, along with their nicknames and abbreviations used in this thesis, are shown in Figure 2.2.

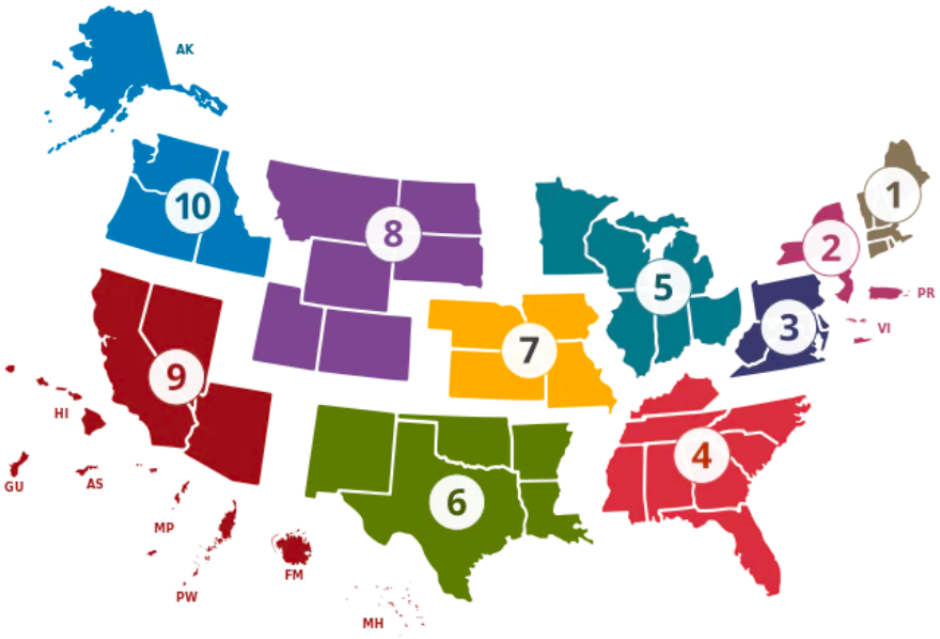


Figure 2.2: Map of the 10 HHS Regions in the United States (U.S. Department of Health and Human Services 2022)

Using this framework for organizing COVID-19 data in the U.S., the approach for analysis will be computing these ten sub-models for both the pre-Delta wave (incorporating both the original and Alpha variants, dating from March 22, 2020 to June 15, 2021) and the Delta

Table 2.2: HHS region names, abbreviations, and list of states used for all analyses

Number	Name	Abbreviation	States
1	Northeast	NE	CT, MA, ME, NH, RI, VT
2	Northeast Coast	NEC	NJ, NY
3	Mideast	MDE	DE, MD, PA, VA, WV
4	Southeast	SE	AL, FL, GA, KY, MS, NC, SC, TN
5	Midwest	MDW	IL, IN, MI, MN, OH, WI
6	Midsouth	MDS	AR, LA, NM, OK, TX
7	Middlewest	MDL	IA, KS, MO, NE
8	Midnorth	MDN	CO, MT, ND, SD, UT, WY
9	West	WST	AZ, CA, NV
10	Pacific Northwest	PNW	ID, OR, WA

wave (from June 15, 2021 to November 1, 2021), for a total of 20 models. These COVID-19 waves, of course, have some degree of overlap, so for the purposes of this research the cutoffs were determined by the local minima in the cumulative United States COVID-19 rolling case plot (The New York Times 2022). Figure 2.1 illustrates the distribution of cases per 1000 individuals across each county in the continental United States for both proposed pandemic waves. Changes in disease protocol over the course of the pandemic, such as the introduction of vaccines to the United States throughout the first quarter of 2021 (The New York Times 2022), may have an affect on the social response of individuals and for those at risk. As such, not only will individual sociodemographic effects be monitored in each model, but the comparison between significant factors from wave to wave will also be studied. The exact modeling technique will be elaborated on in the next two sections, and the metrics used for model comparison will be shown in Section 2.4.

2.2 Exploratory Analysis

The first modeling attempt for this study took a simple approach in using multiple regression conducted in R. This exploratory analysis revealed several issues in terms of model assumptions (and as such, most of the key results from the multiple regression will be ignored).

First, incorporating every explanatory variable in the model resulted in predictor multi-

Table 2.3: Variance Inflation Factor (VIF) for the Southeast HHS Region for cumulative data of pre-Delta variant COVID-19

Variable	VIF
POV	3.825
UNEMP	1.6217
NOHSDP	4.5653
AGE65	4.1719
AGE17	7.0804
DISABL	2.8828
SNGPNT	3.0779
MINRTY	8.7534
LIMENG	2.2571
MUNIT	2.9244
MOBILE	3.0128
CROWD	1.6046
NOVEH	3.5672
GROUPQ	2.5554
pct	6.6718

collinearity. Table 2.3 shows the Variance Inflation Factors (VIF) for the pre-Delta model based in the Southeast HHS region. A VIF of greater than five signals concerning collinearity, and a VIF of greater than 10 signals severe collinearity. As can be seen, many of these variables, such as percentage of minorities, individuals under 17, and democratic voting percentage, have a VIF of greater than five, and as such are correlated with other explanatory variables. This is no surprise, of course; the SVI metrics are grouped together based on overarching categories of social risk, so commonalities and shared information between these variables is to be expected. However, for statistical purpose, highly-correlated explanatory variables result in regression coefficient estimates that are generally unreliable, and as such we cannot make sound conclusions when high levels of collinearity are present. The reason why this occurs is due to how the coefficient estimates are computed. Regression assumes explanatory variables are independent, and will rely on this result for computing the minimum sum of squares. When they are not independent, a closed form of the regression coefficients cannot be found due to having to calculate a matrix inverse of a singular or nearly-singular matrix, and as such, reliable results cannot be determined. Some may argue that a VIF

of greater than 10 is when coefficient estimates delve into inaccuracy, though any amount is undesirable, and for the purposes of this thesis, any amount greater than five will be deemed detrimental given the dimensionality of the data set. The natural response to this issue is to remove variables that present redundant information (i.e., have VIF values over five). However, maintaining variables is desirable for this approach, as maintaining nuanced variable significance is at the core of the questions being asked. Low socioeconomic status and crowded housing may be correlated, for instance, but the individual influence of each variable on COVID-19 spread is desired. As well, some models, such as the Southeast HHS model for pre-Delta COVID-19 shown previously in Table 2.3, indicates a removal of the percentage minority variable, in spite of the knowledge that: a) minorities have been shown to be at higher risk for COVID-19 (Gayle and Childress 2021), and b) the Southeast HHS region has an average minority population of 29.05% across every county, which is the third highest among any HHS region behind the Midsouth and West regions. Clearly, we should expect to maintain a variable of importance in such a region, but under the current assumption we are unable to do so.

Table 2.4: R^2 coefficients for training and testing data set, along with testing set root mean square error (RMSE), for Pacific Northwest region pre-Delta multiple regression model

Metric	Value
Training R^2	0.7591
Testing $\mathbf{R^2}$	**0.5234**
RMSE	26.1974

Second, the multiple regression models overfit the data in most instances. Table 2.4 shows the result of model-fit assessments using a 70/30 training-testing set on the Pacific Northwest region for pre-Delta variant COVID-19. Under well-fit models, the R^2 coefficient should be similar for both models, as the training and testing sets should, in theory, be representative of the entire data set. However, given that the training set has a much higher R^2 value than the testing set, this indicates that the model suffers from overfitting. We

would expect a multiple regression model to be able to adapt well to new data. The primary reasoning for this overfitting is mainly due to the variance in the data. Given that the model has a wide gap in training and testing R^2 coefficients, we assume that the multiple regression model is fit too tightly to the training data set. Models generally behave this way when there is high variability in the data. Since multiple regression is shown here to fit models that have high variability, have collinear explanatory variables, and have difficulty generalizing to new data, multiple regression cannot provide accurate estimators for variable coefficients. As such, since the unbiased estimators (unbiased, in this case, referring to the sample regression coefficients having an expected value equal to the population coefficients) of multiple regression are unreliable for these data, we can utilize a biased estimation method that will not only reduce variability, but will also account for both collinear and redundant predictors. One such method is *elastic net regression*.

2.3 Elastic Net Regression

Elastic net regression is a regularization technique developed by Hui Zou and Trevor Hastie to overcome the limitations of l_1 (lasso) and l_2 (ridge) regularization (Zou and Hastie 2005). l_1 regularization is a penalty to the OLS estimators that results in feature selection, selecting against unimportant variables, whereas l_2 regularization is a penalty to the OLS (Ordinary Least Squares) estimators that shrinks correlated predictors towards each other to overcome multicollinearity (note: the specifications for the l_1 and l_2 penalties will be detailed below when presenting the elastic net model) (Zou and Hastie 2005). What elastic net achieves is a balance between these two penalties, resulting in a process by which models can be created that deal with multicollinearity among explanatory variables while simultaneously selecting the important features out of a large set of potential predictors. Zou and Hastie (paired with simplified derivations from Friedman, Hastie, and Tibshirani (2010)) describe the set up for the model as follows:

Suppose we have a vector $\mathbf{Y} = (y_1, ..., y_n)$, which is our observed data, and $\mathbf{X} = (x_1| \ ... \ |x_k)$

be our model matrix. Also suppose we have $\mathbf{x}_j = (x_{1j}, ..., x_{nj})^T$, $j = 1, ..., p$ are the predictor variables. We can construct a regression model such that $\mathbf{Y} = \beta_0 + x^T\beta + \epsilon$, where β_0 is the intercept of the regression equation, β are the coefficients for the p explanatory variables, and ϵ is our residual vector. For the elastic net penalty, we assume that the response is centered and the predictors are standardized. This means that the sum of our observed data should equal 0, the sum of our explanatory variables should equal 0, and the mean squared error of our explanatory variables should equal 1. This can be represented as:

$$\sum_{i=1}^{n} y_i' = 0, \qquad \sum_{i=1}^{n} x_{ij}' = 0, \qquad \frac{1}{n}\sum_{i=1}^{n}(x_{ij}')^2 = 1, \qquad j = 1, ..., p$$

Here, y_i' is the centered observed response for the i^{th} county, and x_{ij}' is the standardized measurement for the j^{th} explanatory variable in the i^{th} county. Once this pre-processing has been accomplished, we define the model. The proposition for this model is similar to that of a standard regression model. In standard OLS, we wish to minimize the sum of squared residuals in order to fit a linear model to the data. This is similar to that approach, but with the added l_1 and l_2 penalties. As such, we wish to find:

$$\hat{\beta} = \underset{(\beta_0,\beta)}{\operatorname{argmin}}\ R_\lambda(\beta_0, \beta) = \underset{(\beta_0,\beta)}{\operatorname{argmin}} \left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda P_\alpha(\beta)\right]$$

$$\begin{aligned} P_\alpha(\beta) &= (1-\alpha)\frac{1}{2}\|\beta\|_{l_2}^2 + \alpha\|\beta\|_{l_1} \\ &= \sum_{i=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right] \end{aligned} \tag{2.1}$$

This particular notation given for the derivation of the elastic net regression model comes from Friedman, Hastie, and Tibshirani (2010), which defines $P_\alpha(\beta)$ as the elastic net penalty first introduced by Zou and Hastie. α and λ are the hyperparameters for this model, and

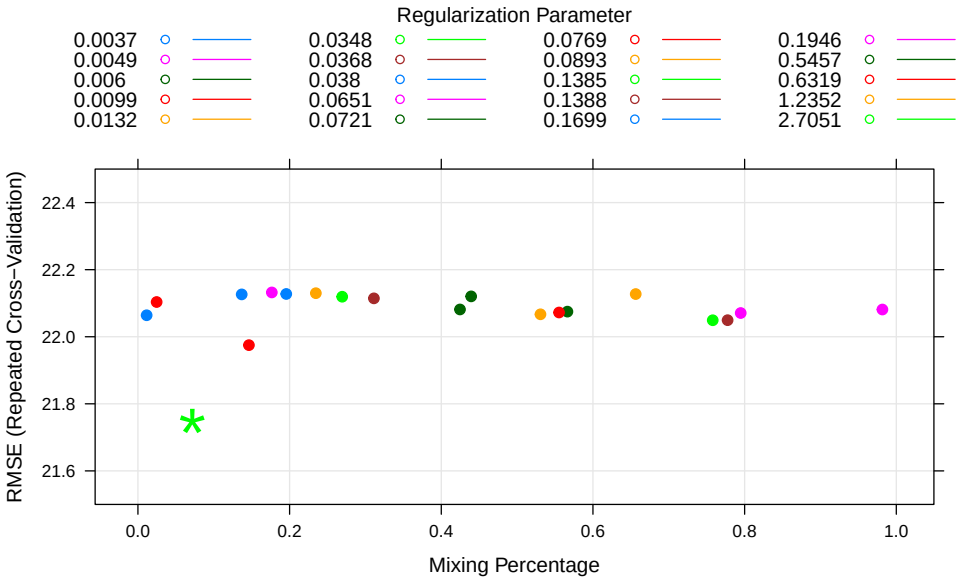
help to determine the mixture and amount of penalization, respectively. Intuitively, we can see why elastic net is a generalization of lasso and ridge. Setting $\alpha = 1$ results in the elastic net penalty only contributing a penalty equivalent to the magnitude of the β coefficients (also known as the l_1 norm), thus resulting in strictly l_1 regularization. Conversely, setting $\alpha = 0$ results in the elastic net penalty only contributing a penalty equivalent to the squared Euclidean distance of the β coefficients (also known as the l_2 norm), thus resulting in strictly l_2 regularization. These penalties together are scaled by λ; the higher the value for λ, the higher the penalty accrued to the regression. We saw earlier that poor prediction results in model overfitting. What λ does in this setting is scale the amount of bias we introduce into the model. This bias helps to remove redundant information from our data, thus improving prediction overall. This concept is known as a bias-variance tradeoff, and is a centerpiece for the concept of regularization techniques. The more bias we introduce, the less variance we have (note: this is shown through looking at the matrix form of the variance for the estimated elastic net coefficients $\hat{V}(\hat{\beta})$; for the scope of this thesis, the details of this inverse relationship between bias and variance will not be discussed in-depth). Of course, there is an optimal balance that minimizes the penalized regression shown earlier, and the metric to measure this will be specified in Section 2.4. To summarize the model, elastic net is a hybrid method that takes advantage of the bias-variance tradeoff that is beneficial from ridge while also utilizing the feature selection of lasso, making for a highly-effective method to combat overfitting.

This technique shows promise for modeling sociodemographic factors against COVID-19 cases given the previously described collinearity issues. Though Zou and Hastie describe a best-case scenario for elastic net as one where the number of predictors is much greater than the number of observations (i.e., p $\gg$ n), they also state that this technique is successful for modeling data with highly-correlated predictors, as we have in this case.

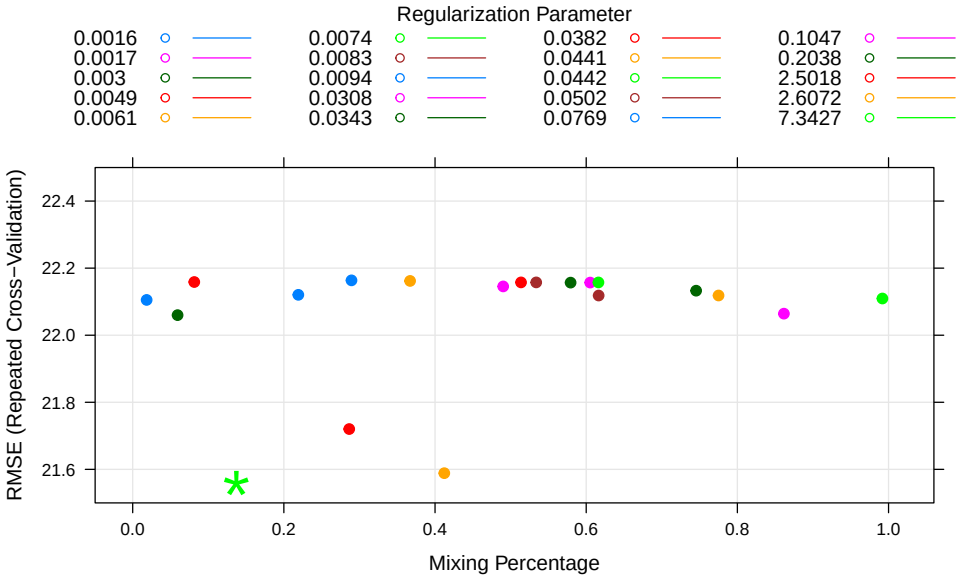
2.4 Metrics and Techniques

With the elastic net model in mind, we can organize and define the metrics that will be most useful in these analyses. Model creation will be exactly the same as that done in Section 2.2 during the exploratory analysis: ten HHS models for the pre-Delta wave, and ten HHS models for the Delta wave. Constructing these elastic net models will utilize the **caret** package in R, which is a package designed for machine learning of statistical parameters that can search for the optimal α and λ values to minimize $R_\lambda(\beta_0, \beta)$. In terms of theory, minimizing $R_\lambda(\beta_0, \beta)$ in a penalized regression is equivalent to minimizing the root mean squared error (RMSE) of the model. RMSE, in regression, is the standard deviation of the residuals and is an effective measure of how well a model fits the data. A 70/30 training/testing split will be used as previously, with the training data being utilized by the **caret** package to calculate these optimal values.

10-fold repeated cross validation will also be utilized for determining the hyperparameters with the lowest RMSE. Generally, repeated cross-validation is done five times in order to guarantee consistent results. However, cross validation can be a noisy process in some instances. For the data being used, each model has potentially high variance in cumulative case numbers for each replicate due to smaller counties simply having less COVID-19 cases; additionally, the **caret** package randomizes the selection of the mixing parameter α, so finding the optimal hyperparameters overall will yield different results from iteration to iteration. Figure 2.3 displays this concept; in Iteration 1, the α value with the lowest RMSE is just below 0.1 (as shown by the starred point on the respective parameter scatter plot), whereas Iteration 2 has an optimal α value of just under 0.2. It is important to note that the minimized RMSE between these two models is approximately the same, meaning both sets of tuning parameters yield similar model performance. Due to this stochasticity in selecting optimal mixing parameters that minimize the RMSE, the cross validation will be repeated 15 times. This is somewhat arbitrary, and is chosen simply to ensure the best model is



(a) Iteration 1



(b) Iteration 2

Figure 2.3: Display of varying optimal hyperparameters across two iterations of generating the elastic net model for the Pacific Northwest HHS region

selected. The **caret** package handles repeated cross-validation natively, and will repeat the procedure of randomly generating 20 potential α parameters for the alloted 15 iterations.

The R^2 coefficients for the training and testing data will be used to determine the degree of model overfitting. If they are fairly similar in value, we know that our model is correct in

its choice of hyperparameters. To further evaluate the effectiveness of the model, the testing set RMSE will be calculated to see how well the new data fits the pre-determined model. Finally, the same data set partition will be used for a multiple regression case, and the same metrics will be calculated. Since we know that multiple regression results in overfitting for this data set, we will compare the testing set R^2 and RMSE for the elastic net model and the multiple regression model to see which model performs better overall. To summarize, this thesis has three primary objectives:

1: Compare the features and coefficients present between models across the respective HHS regions for pre-Delta and Delta COVID-19 data to see which factors have a significant effect on COVID-19 cases,

2: Calculate the R^2 coefficients for the training and testing data in the elastic net model (and testing R^2 for the multiple regression model) in order to evaluate the elastic net's ability to correct the model overfitting present in the multiple regression models, and

3: Compare the root mean squared error (RMSE) between the elastic net models and standard multiple regression models for the testing data to see how each model performs when introduced to new data.

CHAPTER 3

Results

The results of these models will be grouped by pandemic wave (either pre-Delta or Delta) in order to easily view the metrics and model performance across each HHS region at a glance. The major modeling results and trends will be discussed both within and across each wave of the pandemic, while the implications and evaluation of this modeling technique will be discussed in Chapter 4. Despite needing to standardize our data to compute the elastic net coefficients as stated in Section 2.3, the **caret** package reports coefficient estimates in their original scale, allowing for ease in interpretability. Since each explanatory variable is represented as a percentage, the interpretation of these coefficients can be thought of in terms of the effects on COVID-19 cases per 1000 people by county while increasing the percentage of a given explanatory variable by 1%, leaving all other explanatory variables constant. Variable Importance Plots (VIP) will be used to help visualize the significance of these variables across all HHS regions within the two pandemic waves. Given the lack of a simple method to calculate p-values as with multiple regression coefficients (which use t-statistics and estimated standard errors), VIPs are the most effective method of viewing the significance of variables in regularized regression (note: some methods do exist for estimating lasso p-values, such as the method defined by Lockhart et al. (2014), but these methods will not be used in this thesis). Importance is calculated based on the absolute value of the coefficients, which are then standardized to a 0-100 scale. A score of 100 is given to the variable with the highest magnitude in the model. and 0 is given to the variable(s) with the lowest. The remaining variables lie within this range. These plots will include both the individual variable importances for each HHS region per wave, as well as box plots to view the average importance of each explanatory variable.

3.1 Pre-Delta

Table 3.1: Coefficients and metrics for the 10 HHS regions for the Pre-Delta COVID-19 wave, recorded from March 22, 2020 to June 15, 2021

Coefficients										
	PNW	MDW	MDL	MDS	SE	NE	NEC	MDN	MDE	WST
(Intercept)	71.515	106.127	106.377	107.568	109.204	65.139	87.203	113.583	85.057	88.214
In Poverty	3.159	−1.425	−5.51	0	0	−2.456	0	−1.804	−0.768	10.721
Unemployed	−1.071	−5.224	0	0	−1.186	−0.139	0.402	1.274	−1.279	1.541
No HS Diploma	0	−2.632	3.843	2.166	3.817	3.702	0	6.362	−1.656	2.346
Over 65	−6.144	0	2.918	−1.909	−1.892	−3.625	0	0	−0.97	8.175
Under 17	12.319	1.768	1.669	4.554	2.498	2.304	5.548	2.242	0.651	4.268
Disability	0	3.205	−0.837	0	0	−2.954	−0.001	−4.604	0	−6.107
Single Parent	0	3.333	6.684	4.191	0	2.168	−1.422	8.357	2.78	0.843
Minority	0	8.181	−1.134	0	5.067	3.708	4.586	0	0	24.012
Limited English	0	3.528	5.189	0	0.311	6.873	6.202	−4.757	4.595	0
Multi-Unit Home	2.589	2.019	1.734	0	0.357	0	0.556	3.964	−0.136	4.398
Mobile Housing	−1.025	−3.669	−6.612	−2.847	−6.157	−12.454	−7.949	−7.193	0	0
Crowded Housing	0	0	3.009	0.756	1.08	−2.001	2.471	1.798	1.054	−3.811
No Vehicle	−1.847	−3.554	−0.186	1.056	0	0	−0.888	0	3.584	−4.65
Group Quarters	4.108	5.288	6.214	6.433	6.663	−0.593	−3.612	17.063	11.004	13.588
Voting Percentage	−9.85	−10.495	−1.065	−0.865	−9.379	−5.676	−3.992	−1.773	−8.334	−11.205
Metrics										
α	0.417	0.363	0.731	0.544	0.911	0.104	0.194	0.809	0.5	0.266
λ	3.575	0.093	0.373	2.096	0.195	4.305	2.608	0.992	0.461	0.811
ENR Train R^2	0.659	0.405	0.353	0.282	0.196	0.781	0.784	0.40	0.444	0.706
ENR Test R^2	0.637	0.366	0.352	0.243	0.148	0.778	0.778	0.372	0.419	0.641
MR Testing R^2	0.601	0.322	0.297	0.203	0.107	0.764	0.631	0.247	0.339	0.456
ENR Test RMSE	19.265	15.683	21.052	23.945	19.997	15.479	11.52	28.114	12.88	26.789
MR Test RMSE	19.521	15.672	21.101	33.597	20.087	17.491	16.66	29.254	13.119	28.433

The complete list of coefficient estimates and metrics for the 10 pre-Delta elastic net models are shown in Table 3.1. Since this table is high-dimensional and is difficult to interpret upon first glance, we will break things down in the three steps stated at the end of Section 2.4 in order to aid in overall interpretation. As can be seen in the *sparse matrix* of coefficients (i.e., the matrix produced when variables are removed via the penalization parameter), the number of features removed across each model varied a fair amount, with regions such as the Pacific Northwest and Midsouth having a third of their explanatory variables removed from the final model, and regions such as the Middlewest having almost none removed. In assessing overall variable trends, it is important to see the overall variable effects across each region. Figure 3.1 displays the VIP plot for the pre-Delta COVID-19 wave. Here, we see that percentage of individuals living in group quarters (GROUPQ) has the highest overall variable importance across each region, with democratic voting percentage (pct) and

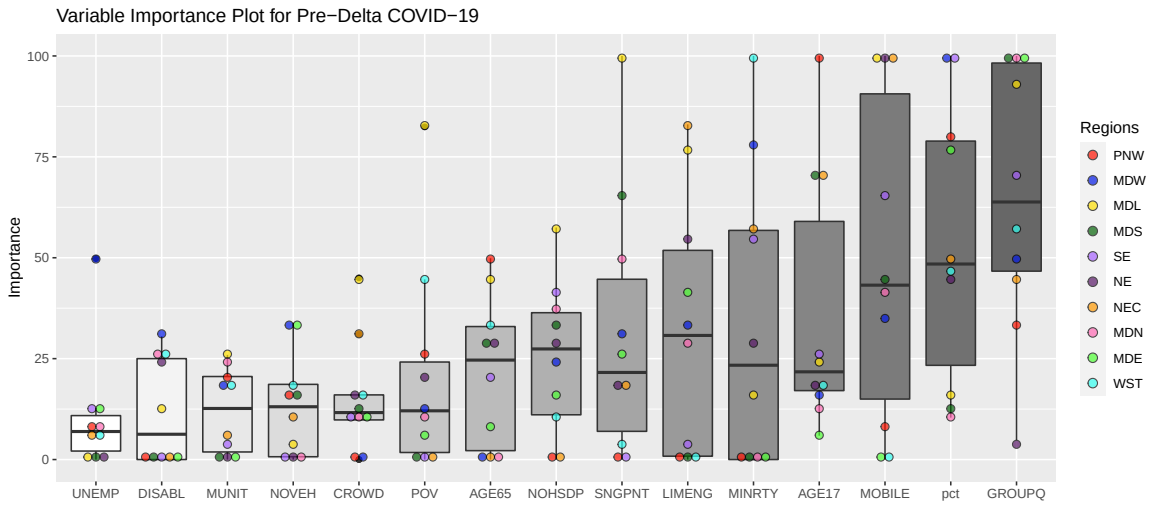


Figure 3.1: Pre-Delta Variable Importance Plot, organized from lowest to highest overall importance

percentage of individuals in mobile homes (MOBILE) being the next highest variables in terms of overall importance. As can be seen by the points in the dotplot portion of Figure 3.1, these three variables account for the highest variable importance across seven of the ten HHS regions for this wave of the pandemic. Note that percentage of single parents (SNGPNT) has the highest coefficient value for the Middlewest, yet has a nearly identical importance to percentage of mobile homes (MOBILE), thus resulting in the appearance of MOBILE having three models with which it has the highest importance, while it only has two. The other variables that hold the highest importance within an HHS region are percentage of individuals under 17 (AGE17), percentage of minorities (MINRTY), and percentage of single parents (SNGPNT). The variable with the lowest overall importance is percentage of unemployed individuals (UNEMP).

Turning now to the model metrics, not much can be assessed through analyzing the values for λ, as they are entirely dependent on the value for α. However, it may still be worthwhile to mention the trends for α in this pandemic wave, though these values can change quite a lot across model iteration, as shown in Figure 2.3. Overall, the mixing percentages across each HHS region do not show a consistent trend. Some models, such as the Southeast, display

an alpha value that trends toward the pure l_1 norm, whereas models such as the Northeast Coast sit towards the opposite extreme in the l_2 norm. However, most models display a fair amount of mixture between the two penalties, which is desirable as it shows the optimal RMSE for these models can be found in a more complex model than ridge or lasso alone. In terms of the R^2 results, the training set value, understandably, varies a significant amount across each region. More importantly, however, is that the testing set R^2 values for all HHS regions are within 5% of their respective training set R^2 values, showing that the models are generalizing well to new data. Additionally, every testing set R^2 for the elastic net models perform significantly better than the multiple regression training set R^2 counterpart, showing that in cases where the multiple regression models overfit the data, the elastic net models do not.

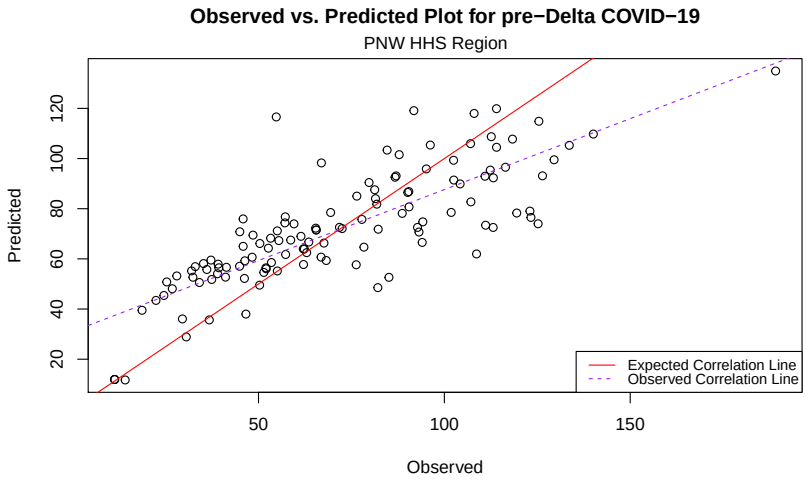


Figure 3.2: Observed versus predicted COVID-19 case rate plot for pre-Delta PNW HHS model, used to visualize the correlation between the observed data and the predicted data. The expected correlation line shows a perfect one-to-one correlation between the observed data and the fitted data ($R^2 = 0.6878$).

In terms of overall model performance, the RMSEs for the elastic net models are lower than their multiple regression counterparts in all but one model (Midwest, though the multiple regression model only slightly outperforms the elastic net model in terms of RMSE).

This result shows better model performance for the elastic net model overall, while simultaneously eliminating the overfitting issue for this data set. We will now analyze the Delta wave to evaluate how trends changed as the pandemic carried on.

3.2 Delta

Table 3.2: Coefficients and metrics for the 10 HHS regions for the Delta COVID-19 wave, recorded from June 15, 2021 to November 1, 2021

Coefficients										
	PNW	MDW	MDL	MDS	SE	NE	NEC	MDN	MDE	WST
(Intercept)	46.762	40.323	36.03	49.074	61.079	23.503	24.602	39.915	49.69	40.466
In Poverty	−0.297	1.167	0	0.311	2.816	1.377	1.252	−2.505	5.145	2.691
Unemployed	2.526	−2.088	1.011	0	1.01	0	1.258	2.719	0	−1.233
No HS Diploma	−0.203	−1.959	0	0.235	−1.193	0.225	0	−3.69	0	0.185
Over 65	−1.678	−0.523	−1.262	−2.493	0.756	0	0	−4.953	0	-7.405
Under 17	2.214	1.442	0	0	3.726	0	0.26	−0.435	0	3.039
Disability	4.159	3.771	3.755	0	2.202	1.308	0	6.638	0	2.571
Single Parent	4.696	1.87	1.475	0.898	1.12	0.522	0.995	−0.403	0	0.228
Minority	0	0.85	−1.338	−1.912	−2.34	−0.527	−0.685	−4.761	−2.302	−4.988
Limited English	−0.902	−0.438	0	−3.226	0	0	−2.031	0.02	0	−2.234
Multi-Unit Home	0	2.356	0	0	4.577	−1.034	0	0.95	0	−4.615
Mobile Housing	3.372	2.348	1.829	1.373	1.812	0.489	−0.718	4.738	5.206	−2.614
Crowded Housing	−2.334	−0.743	0	0.119	0.881	0	0	2.20	0	−1.562
No Vehicle	2.347	0.428	0.166	1.645	−0.086	0	0	2.171	0	3.605
Group Quarters	0.138	0.133	0	−0.976	0.192	0	−0.451	−1.729	−0.293	−1.459
Voting Percentage	−5.905	−8.087	0	0	−9.776	−1.367	−2.225	−1.448	−7.175	−7.635
Metrics										
α	0.201	0.67	0.954	0.903	0.655	0.298	0.841	0.115	0.709	0.144
λ	1.184	0.066	0.597	0.783	0.143	1.943	0.191	0.979	1.755	0.984
ENR Train R^2	0.603	0.553	0.244	0.163	0.461	0.471	0.616	0.284	0.722	0.731
ENR Test R^2	0.586	0.508	0.274	0.156	0.412	0.458	0.548	0.249	0.72	0.691
MR Testing R^2	0.463	0.449	0.215	0.071	0.37	0.148	0.286	0.175	0.567	0.615
ENR Test RMSE	14.094	9.146	9.957	15.213	13.086	7.044	4.491	15.047	12.33	9.301
MR Test RMSE	14.855	9.183	10.358	14.92	13.156	12.051	4.769	16.04	14.834	9.787

The complete list of coefficient estimates and metrics for the 10 Delta elastic net models are shown in Table 3.2. Compared to the pre-Delta wave, the range of removed features in the Delta elastic models is much larger overall. Models such as the Mideast region have up to two-thirds of their features removed through the penalization process. In contrast, the West model for the Delta wave has no features removed on average. In terms of individual features, the plot provided in Figure 3.3 gives some interesting insight. Percentage of group quarters individuals per county (GROUPQ) went from having the highest average variable importance in the pre-Delta wave to having the lowest average variable importance in the

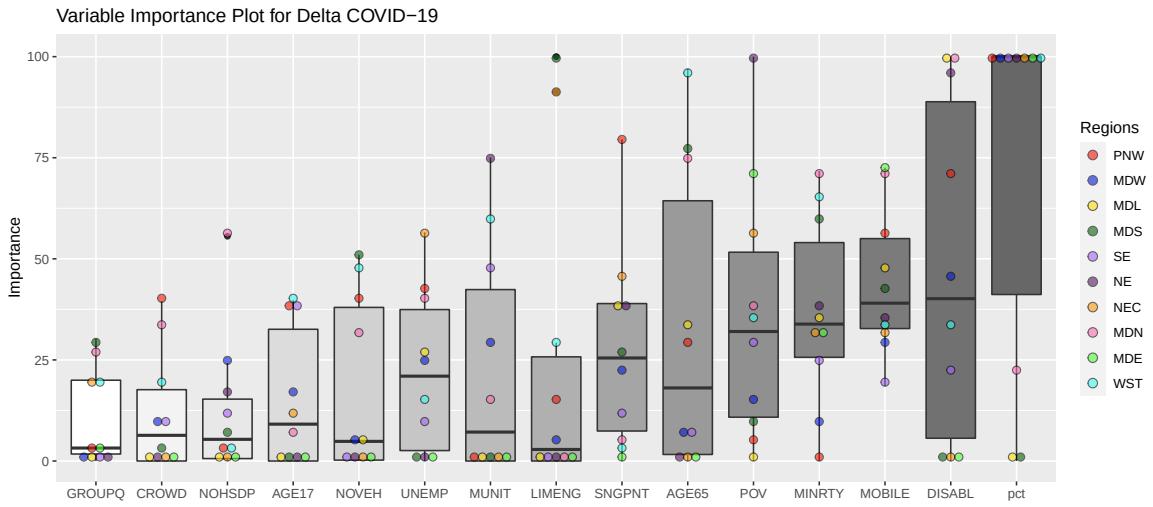


Figure 3.3: Delta Variable Importance Plot, organized from lowest to highest overall importance

Delta wave. Political voting percentage (pct) took its place as the variable with the highest average variable importance across all regions, followed by percentage of non-institutionalized disabled individuals (DISABL) and percentage of individuals in mobile homes (MOBILE). Whereas groups quarters percentage held highest significane overall in the pre-Delta wave with three individual regions of highest importance (i.e., group quarters percentage had a variable importane of 100 in three of the HHS region models), voting percentage had the highest individual variable importance in *seven* of the ten HHS regions for the Delta wave, as shown in the dotplot overlay in Figure 3.3. This matches the number of highest-importance variables for the first *three* explanatory variables combined in the pre-Delta wave. Overall, several variables increased in importance from pre-Delta to Delta such as percentage of individuals in poverty (POV) and percentage of individuals over 65 (AGE65), which can be seen by their positions in Figure 3.3. Other variables, such as percentage of individuals in crowded living spaces and percentage of individuals in group quarters housing, dropped dramatically from pre-Delta to Delta, with GROUPQ experiencing the most dynamic shift in variable importance overall.

The performance of the elastic net models in terms of RMSE was better in almost every

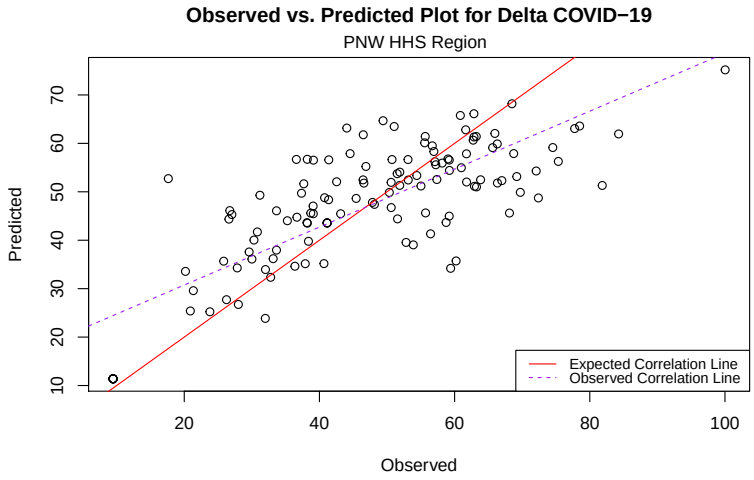


Figure 3.4: Observed versus predicted COVID-19 case rate plot for Delta PNW HHS model, used to visualize the correlation between the observed data and the predicted data. The expected correlation line shows a perfect one-to-one correlation between the observed data and the fitted data ($R^2 = 0.6077$).

instance. Only the Midsouth HHS region had a multiple regression RMSE that outperformed the elastic net RMSE. This reflects the same improvement in model performance that the elastic net models had on the pre-Delta data, showing consistency across the two pandemic waves.

For the metrics, the overall trends of the mixing percentages appear to be relatively similar to that of the pre-Delta models, generally trending towards a balanced mixture between l_1 and l_2 regularization. Some models in this wave had α values that appear to trend more toward pure lasso, such as the Middlewest, which had an α value of 0.954, and the Midsouth, which had an α value of 0.903. As with the previous wave, the training set R^2 values vary widely between HHS regions, with some explaining case variability well (i.e., Mideast), and some explaining case variability poorly (i.e., Midsouth). The testing set R^2 values are nearly identical to their respective training set R^2 values as in the previous wave, showing that the Delta COVID-19 behaves similarly with respect to cross validation. Further to this point, the multiple regression testing set R^2 values perform much worse than the elastic net testing set R^2 values, showing that, just as in the previous wave, elastic net

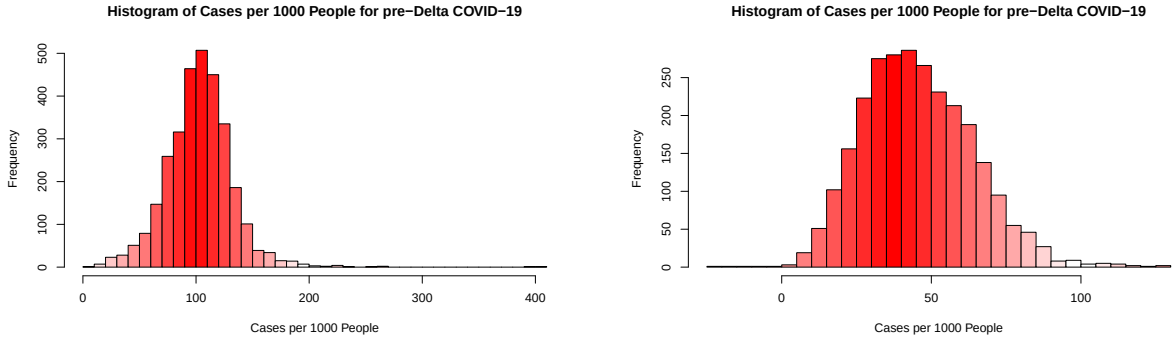
performs admirably in situations where multiple regression has difficulties. One model in this data set, interestingly, has an elastic net testing set R^2 value that is greater in value than its training set R^2 counterpart (Middlewest, with training/testing R^2 values of 0.244 and 0.274, respectively), though this is not necessarily a point of concern; rather, it simply indicates that the testing data set fits the model better, which could be due to either the model being able to generalize to new data well, or the model having an overall weak fit as shown in the low R^2 values.

CHAPTER 4

Discussion

4.1 Model Evaluation

Elastic net regression improved the prediction of COVID-19 cases when compared to multiple regression. All but two of the HHS regions across both the pre-Delta and Delta waves (Midwest, pre-Delta; Midsouth, Delta) had a lower testing set RMSE when compared to their multiple regression counterparts. Even the two elastic net models that performed slightly worse in terms of RMSE still corrected the overfitting of the multiple regression models when generalizing to new data, as the testing set R^2 values for the elastic net models always outperformed multiple regression. It is important to note that this method is beneficial when dealing with COVID-19 cases per 1000; in other words, this method is effective when standardizing cases across replicates. Standardizing the response variables in this method allows for use of standard multiple regression along with elastic net regression, as the case rates approximately follow a normal distribution (Figure 4.1 displays the distribution shape of the response variable for both pandemic waves, though it is important to note that the shape of our distribution could be approaching a normal distribution due to a large sample size). Previous predictive analyses for sociodemographic factors (many of which are used in this thesis through SVI), including Karmakar, Lantz, and Tipirneni (2021) and Millar et al. (2021), use similarly scaled COVID-19 cases/deaths (with Karmakar, Lantz, and Tipirneni (2021) using cases per 100,000 and Millar et al. (2021) using adjusted case-fatality rate). However, both of these models differ from the approach taken in this research by using negative binomial regression, though the use of generalized linear models such as negative binomial regression for COVID-19 cases is a natural path to take given that the response is still capturing count data at its core. This distinction is acknowledged simply for the purposes of identifying how this approach differs in its interpretation of the distribution for the response variable.



(a) Pre-Delta

(b) Delta

Figure 4.1: Distribution for pre-Delta and Delta cases per 1000 individuals per county

In looking at other analysis comparisons, Millar et al. (2021) also used the **caret** package to asses collinearity issues in their analysis, though their approach was in removing variables with correlation greater than 0.5, as opposed to the approach taken here where variables were removed based on the set of elastic net hyperparameters that minimize RMSE (Millar et al. 2021). Both approaches have sufficient validity based on the respective research questions; removing collinearity was done in the Millar et al. analysis to satisfy the assumptions of negative binomial regression (with a simple cutoff of 0.5 for correlation), while this thesis conducted elastic net for the purposes of outperforming standard multiple regression while potentially preserving variables that are correlated with one another, yet still contain some unique predictive information. All things considered, elastic net regression seems to provide a successful method for modeling COVID-19 case rate data in scenarios when high collinearity is present, and when maintaining important correlated features and removing unimportant features is a desired outcome.

4.2 COVID-19 Implications

Many of the results found in this thesis support findings found in previous research. For the pre-Delta wave, percentage of individuals in group quarters had the highest average significance across all regions, with most HHS regions having a strongly positive correlation

with cases (especially in regions where group quarters percentage was among the variables with the highest importance). It was known early on in the pandemic that close-quarters environments such as cruise ships were epicenters of COVID-19 spread, though nursing homes and prisons (i.e., examples of group quarters environments) were also shown to be highly dangerous environments for the spread of COVID-19 (Sloane 2020). Previous analyses from Karmakar, Lantz, and Tipirneni (2021) and Millar et al. (2021) also support these claims of group quarters being significant predictors for COVID-19 cases early on in the scope of the pandemic, with Karmakar, Lantz, and Tipirneni (2021), for instance, conducting their analysis through July of 2020. The results of this research not only confirm the dangers of group quarters for the first few months of the pandemic, but also show the risk of group quarters environments up through the end of the Alpha strain. It is interesting to observe that group quarters has the least average importance for the Delta variant data, however. This might be due to higher case rates in group quarters environments earlier on in the pandemic; another contributing factor could be the distribution of vaccines through the first half of 2021, as vaccines were shown to reduce the incidence of COVID-19 in nursing homes through the beginning of the Delta variant's emergence, which is an environment that is under the umbrella of group quarters (White et al. 2021).

Voting percentage was the most important predictor in seven of the ten Delta models, with each model containing negative coefficient values. The final model predicts that an increase in percentage of democratic voting according to the 2020 U.S. presidential election results in a decrease of case rate overall. Given that vaccines have shown high effectiveness in reducing case incidence up through the Delta variant (Bernal et al. 2021) and that republican individuals reportedly have 90% lower vaccination intent odds when compared to democrats (Dolman et al. 2022), this presents itself as a potential cause of this correlation. Of course, presidential voting is not the most foolproof proxy for the overall effect of political leaning on COVID-19; it fails to capture much of the extrinsic social effects caused by such factors as local legislature, personal risk evaluation, and media influence. However, this stands as a

topic for further research into the effects of nuanced political leaning factors on COVID-19 incidence.

Interestingly, percentage of mobile homes per county seems to have a significant effect on COVID-19 incidence according to this approach, being the third most important variable on average across both COVID-19 waves. Even more interestingly, the correlation of mobile home percentage per county shifts from negative in the pre-Delta wave across most regions, to *positive* in the Delta wave across most regions. These results are also backed by Karmakar, Lantz, and Tipirneni (2021) and Millar et al. (2021), which both reported mobile home percentage as a significant factor. To explain the negative trend in the pre-Delta model, Millar et al. (2021) elaborates on the potential cause of this effect being a "built-environment effect," since the ventilation and plumbing of mobile homes is separated from other housing units, unlike multi-unit housing such as apartment complexes; this is made even more significant when considering that fecal-aerosol transmission can occur for COVID-19 (Millar et al. 2021). To potentially explain the trend for the Delta model, we can explore data already gathered for this research, coupled with knowledge already presented regarding vaccination behavior. We know that political ideology has an effect on case incidence through overall behavior towards vaccination. As such, we can regress democratic voting percentage on mobile home percentage in a simple linear regression to make inferences on this relationship in order to possibly explain the positive trend of mobile home percentage on case rate in the Delta wave. The results of this exploratory analysis are shown in Table 4.1.

Table 4.1: Simple Linear Regression result for relationship between political leaning and mobile home percentage per county across the entire continental United States

Coefficients	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$
(Intercept)	39.244	0.509	77.09	$< 2e-16$
MOBILE	-0.456	0.032	-14.16	$< 2e-16$

Mobile home percentage has a significantly negative effect on democratic voting per-

centage. A 1% increase in mobile home percentage per-county decreases the percentage of democratic voting by 0.456%. Since we already know that a) vaccination mentality played a significant role in disease spread over the course of the Delta variant, and b) right-leaning individuals had much lower odds of intent for receiving the vaccine, this relationship presents itself as a possible explanation for the now-positive correlation between mobile home percentage and adjusted case rate. Of course, much like with this thesis as a whole, many factors could contribute to this inverse relationship. However, according to trends presented in previous literature when considering vaccination (e.g., Bernal et al. (2021)), this result holds a degree of significance. Overall, mobile home percentage seems like a strange predictor for COVID-19 on the surface, but it ultimately presents itself as an excellent case study as to how intrinsic and unchanging social factors can change in terms of predicting case incidence as the disease progresses.

Finally, these analyses found differences in variable importance and significance when compared to other sociodemographic analyses of COVID-19. Karmakar, Lantz, and Tipirneni (2021), for instance, found that every SVI entry was a significant predictor of COVID-19 in their negative binomial analysis. This result contrasts with many of the findings of this research, as the elastic net regularization procedure resulted in sparse matrices that removed features from most iterations of the included models. The VIP plots for both the pre-Delta and Delta waves of the pandemic also reveal this trend. Karmakar, Lantz, and Tipirneni (2021) presented percentage mobile homes as the least significant predictor among the SVI subcategories (though still significant, with a p-value of 0.01 as opposed to <0.001 for every other subcategory), contrasting directly with the high variable importance of mobile home percentage within this thesis for approximately (though not exactly) the same time period. Of course, their analysis was conducted somewhat earlier in the pandemic (through July 29, 2020), so this difference may be due in part to the cumulative gathering of data. Whatever the case may be, the cumulative wave of data used for the pre-Delta models in these analyses assists in establishing results that harbor higher accuracy in COVID-19 prediction due to

being more representative of initial pandemic trends.

4.3 Limitations

As with any data analysis, there are some limitations to this approach. For one, the analysis conducted on the pre-Delta wave had a much larger pool of data when compared to the Delta model, since the pre-Delta model contained 15 months worth of cumulative COVID-19 data, whereas the Delta model only contained five months worth of data. This effect is seen in Figure 2.1 especially, as the heat map for the Delta data is much weaker in terms of cases per 1000 people simply due to time. Further subdivision of COVID-19 waves may be an appropriate approach for refining this research, separating the models out into the original strain, Alpha, Delta, Omicron, etc. Additionally, many social factors were not considered in this thesis that have clear implications towards COVID-19. As alluded to in the discussion within Section 4.2, vaccine hesitancy is a highly-important extrinsic social factor that has clear and obvious implications towards case incidence, especially during the time frames used within these analyses. Inclusion of this factor (or, at minimum, a proxy for this factor), likely would have yielded significant results. As such, a more robust analysis using this modeling technique containing more social factors of perceived importance (such as a more nuanced subdivision of minorities within the United States, vaccine hesitancy, full age subdivisions, etc.) would be beneficial to our understanding of the spread and prevention of COVID-19. Additionally, given that elastic net regression is intended for data sets with high-dimensionality (Zou and Hastie 2005), inclusion of more sociodemographic factors would be a seamless process from a computational standpoint. In general, this approach would benefit from expansion on what features are included, as well as how the data are subdivided.

CHAPTER 5

Conclusion

Elastic net regression proved to be a highly successful modeling technique for predicting the effect of sociodemographic risk factors on COVID-19 case rate, performing significantly better than multiple regression in terms of prediction efficiency and accuracy. The results of this research reflect the findings of previous research with regard to COVID-19 risk factors, while simultaneously contesting some of the findings in analyses done with differing techniques earlier in the pandemic. This paper serves as both an introduction of the use of elastic net regression for the purpose of modeling sociodemographic risk factors, as well as a jumping-off point for expanding the analysis of risk factors on COVID-19 using a method that accounts for the inherent collinearity present in large feature sets of sociodemographic variables.

Bibliography

1. Araf, Yusha et al. (2022). "Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses to current COVID-19 vaccines". In: *Journal of medical virology*.

2. Baker, Mike (2022). "'Their Crisis' Is 'Our Problem': Washington Grapples With Idaho Covid Cases". In: *The New York Times*.

3. Bates, Timothy A et al. (2022). "Vaccination before or after SARS-CoV-2 infection leads to robust humoral response and antibodies that effectively neutralize variants". In: *Science Immunology*, eabn8014.

4. Bernal, Jamie Lopez et al. (2021). "Effectiveness of Covid-19 vaccines against the B. 1.617. 2 (Delta) variant". In: *New England Journal of Medicine*.

5. Bollyky, Thomas J et al. (2022). "Pandemic preparedness and COVID-19: an exploratory analysis of infection and fatality rates, and contextual factors associated with preparedness in 177 countries, from Jan 1, 2020, to Sept 30, 2021". In: *The Lancet*.

6. CDC/ATSDR (2018). "CDC/ATSDR Social Vulnerability Index". In: *Agency for Toxic Substances and Disease Registry*.

7. Centers for Disease Control and Prevention (2021). "COVID-19 Vaccines Work". In: *COVID-19*.

8. Dolman, Andrew J et al. (2022). "Opposing views: associations of political polarization, political party affiliation, and social trust with COVID-19 vaccination intent and receipt". In: *Journal of Public Health*.

9. Fazio, Russell H. et al. (2021). "Social distancing decreases an individual’s likelihood of contracting COVID-19". In: *Proceedings of the National Academy of Sciences* 118.8, e2023131118. DOI: 10.1073/pnas.2023131118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2023131118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2023131118.

10. Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1, p. 1.

11. Gayle, Helene D and James F Childress (2021). "Race, racism, and structural injustice: equitable allocation and distribution of vaccines for the COVID-19". In: *The American Journal of Bioethics* 21.3, pp. 4–7.

12. Grossman, Guy et al. (2020). "Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States". In: *Proceedings of the National Academy of Sciences* 117.39, pp. 24144–24153.

13. Harris, Norma S et al. (2019). "Vital signs: status of human immunodeficiency virus testing, viral suppression, and HIV preexposure prophylaxis - United States, 2013-2018". In: *Morbidity and Mortality Weekly Report* 68.48, p. 1117.

14. Hooper, Monica Webb, Anna María Nápoles, and Eliseo J Pérez-Stable (2020). "COVID-19 and racial/ethnic disparities". In: *Jama* 323.24, pp. 2466–2467.

15. Jurkowitz, Mark and Amy Mitchell (2021). "Americans who relied most on Trump for COVID-19 news among least likely to be vaccinated". In: *Pew Research Center*.

16. Karmakar, Monita, Paula M. Lantz, and Renuka Tipirneni (Jan. 2021). "Association of Social and Demographic Factors With COVID-19 Incidence and Death Rates in the US". In: *JAMA Network Open* 4.1, e2036462–e2036462. ISSN: 2574-3805.

17. Little, Christine et al. (2021). "The impact of socioeconomic status on the clinical outcomes of COVID-19; a retrospective cohort study". In: *Journal of community health* 46.4, pp. 794–802.

18. Lockhart, Richard et al. (Apr. 2014). "A SIGNIFICANCE TEST FOR THE LASSO". In: *Annals of statistics* 42.2, pp. 413–468. DOI: 10.1214/13-AOS1175.

19. Mackey, Tim K et al. (2021). "Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter". In: *The Lancet Digital Health* 3.2, e72–e75.

20. Merkley, Eric and Peter John Loewen (2021). "Anti-intellectualism and the mass public's response to the COVID-19 pandemic". In: *Nature Human Behaviour* 5.6, pp. 706–715.

21. Millar, Jess A et al. (2021). "Risk factors for increased COVID-19 case-fatality in the United States: A county-level analysis during the first wave". In: *PloS one* 16.10, e0258308.

22. Motta, Matt, Dominik Stecula, and Christina Farhart (2020). "How right-leaning media coverage of COVID-19 facilitated the spread of misinformation in the early stages of the pandemic in the US". In: *Canadian Journal of Political Science/Revue canadienne de science politique* 53.2, pp. 335–342.

23. Pullan, Samuel and Mrinalini Dey (2021). "Vaccine hesitancy and anti-vaccination in the time of COVID-19: A Google Trends analysis". In: *Vaccine* 39.14, pp. 1877–1881.

24. Ritchie, Hannah et al. (2020). "Coronavirus Pandemic (COVID-19)". In: *Our World in Data*. https://ourworldindata.org/coronavirus.

25. Sloane, Philip D (2020). "Cruise ships, nursing homes, and prisons as COVID-19 epicenters: a "wicked problem" with breakthrough solutions?" In: *Journal of the American Medical Directors Association* 21.7, pp. 958–961.

26. The New York Times (2022). "Coronavirus in the U.S.: Latest Map and Case Count". In: *The New York Times*.

27. Tsui, Ban CH, Aaron Deng, and Stephanie Pan (2020). "Coronavirus Disease 2019: Epidemiological Factors During Aerosol-Generating Medical Procedures". In: *Anesthesia and analgesia*.

28. U.S. Department of Health and Human Services (2022). "HHS Regional Offices - Regional Map". In: *U.S. Department of Health and Human Services*.

29. White, Elizabeth M et al. (2021). "Incident SARS-CoV-2 infection among mRNA-vaccinated and unvaccinated nursing home residents". In: *New England Journal of Medicine* 385.5, pp. 474–476.

30. Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.