Final Project: Data Science Workflow
Exploration of Factors Influencing the Spread and Severity of COVID-19 in the Contiguous United States
Data 100/200A Spring 2020: Principles and Techniques of Data Science
Chrystal Chern, Ke Ma, Andrew Yeskoo

Abstract: In the ever-changing landscape of the COVID-19 pandemic, it can be difficult for the general public to make sense of why the virus seems to spread so aggressively in some areas while sparing others. At the same time, government leaders across the globe are weighing the benefits of instituting aggressive measures such as shelter-in-place orders against the economic hardship that it inevitably entails. As data scientists, we have the ability to look beyond the politics to explore the trends that the available data can reveal. In this report, we endeavor to explore the primary factors affecting the severity and rate of spread of COVID-19 within the contiguous United States, as well as the relative effectiveness of shelter-in-place orders instituted across the country. We also attempt to form a predictive model to predict the future spread of confirmed cases, as well as discuss several of the difficulties in creating a successful model.

The video overview of this project is linked here: https://youtu.be/QQWEcw7dU6E

Introduction

The current COVID-19 pandemic has touched all aspects of society, including the administration of the DATA100/200A course at UC Berkeley this Spring Semester. As aspiring data scientists, our research team has attempted to explore the trends and factors behind the disease's spread in the contiguous United States. In approaching this problem, we identified a series of three primary questions which we hoped to answer through our analysis:

1. Are a region's static factors (characteristics of each specific area relatively unchanged pre/post virus) correlated with the severity of COVID-19 spread in that region? If so, can we quantify what kinds of static factors are most highly correlated with measures of severity and thus speculate on a static factor's ability to predict severity?
2. Does the introduction of stay-at-home orders affect the rate of spread? Can we infer the effectiveness of the stay-at-home order based on the magnitude of the effect?
3. Can historical trends help inform future trends in severity? That is, can we create a model based on current data to predict future spread of the disease, and can we utilize both static factors and dynamic factors (interventions enacted to attempt to limit the spread of the virus) to build our model?

Through data analysis of county-specific metrics providing static descriptors (eg. population, level of urbanization, demographics, etc.) as well as a time series of confirmed cases divided by county, we systematically explore the above questions. This process included exploratory data analysis and visualization, data cleaning and merging of the relevant data sets, exploration of combinations of models and features to fit the data sets, and evaluation of the final model on its ability to fit data trends over time.

Exploratory Data Analysis and Data Cleaning

At the start of our analysis, we considered four primary data sets. The first set consisted of COVID-19 spread data at a statewide level, as well as for several foreign states and provinces. The second set consisted of metric data at a county level, listing both demographic information as well as the dates that various interventions were imposed to attempt to limit the virus spread. The third and fourth sets consisted of time-series data, divided by day, showing the number of confirmed cases and associated deaths, respectively. In researching the data sets, it was determined that all data was current through April 18, 2020, with the exception of the county metrics which were downloaded on April 26, 2020.

In the early data exploration, it was noted that several of the datasets contained records beyond the United States. As the chosen purpose of this project was to explore the trends related to the virus spread in the United States, the decision was made to drop these records from the considered data sets. This was done using trends identified in the UID and FIPS data series provided in varying form in all 4 data sets (Table 1). It was noted that these unique identifiers contained information about the country, state, and county of each associated record. By filtering based on the state FIPS code, we were able to limit our initial data analysis to the 50 states and Washington D.C.

Table 1: Examples of primary keys for each dataset

Original Provided Data	**Primary Key Example**
4.18states.csv	840~~000~~39, Ohio, USA
time_series_covid19_confirmed_US.csv	84039047, Fayette, Ohio, USA
time_series_covid19_deaths_US.csv	84039047, Fayette, Ohio, USA
abridged_counties.csv	39047, Fayette, Ohio

The dates in the various data sets were noted to be in different formats. Since our analysis would depend on comparing the cases and interventions on the same time period, all dates were converted to ordinal format to allow for direct comparison and analysis of specific periods in the virus spread. After converting the dates, the counties data and confirmed cases timeseries were joined, using the FIPS (at the correct level, whether county or state) as the joining key. Exploring the data set further, a set of static and dynamic factors of interest, described in Table 2, were selected as a subset of the complete list of factors in the counties data set. It was noted at this stage that a number of counties included null values for several of the static factors. Upon further exploration, it was determined that the majority of these were in Alaska or Hawaii. Since the spread of the virus in these states would likely not capture the trends in the majority of the United States, it was decided to further filter our data to only include the 48 contiguous states and Washington D.C.

Table 2: Static and dynamic factors of interest

Static	**Dynamic**
Population density per sq mi	Stay-at-home orders
Age and gender demographics	Limits on social gatherings
Mortality rates for various diseases	School closures
Medical insurance coverage	Restaurant closures
Hospital resources	Gym and entertainment closures
Level of urbanization	Foreign travel ban
Political party affiliation	Federal guideline implementation

In exploring our filtered and joined data, we first created a visualization (Figure 1) to explore the relationship between the population vs. the total confirmed cases on April 18, 2020.

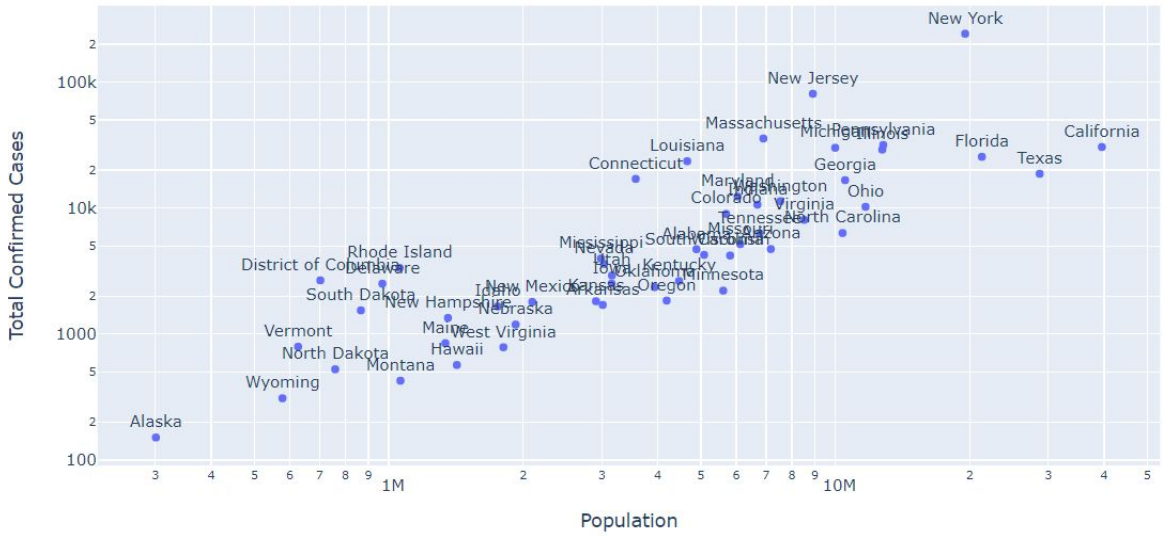

Figure 1: Relationship between population and total confirmed cases on a log-log scale

As can be seen, a clear linear trend is visible when the data is plotted on a log-log scale. This indicates that there may be a correlation between the population of a region and the total confirmed cases. This exploration was repeated for the total deaths as well, however it was decided to focus on the confirmed cases for the remainder of the analysis. Based upon the above observation, we proposed to create a new metric for the local severity of the outbreak, termed the case intensity. This was calculated by normalizing the total confirmed cases by the population of the specific region and multiplying by 100,000, representing the number of people in the population, per 100,000 people, confirmed with the virus. Plotting the local population versus the case intensity (Figure 2), we see that the log-linear trend has been removed.

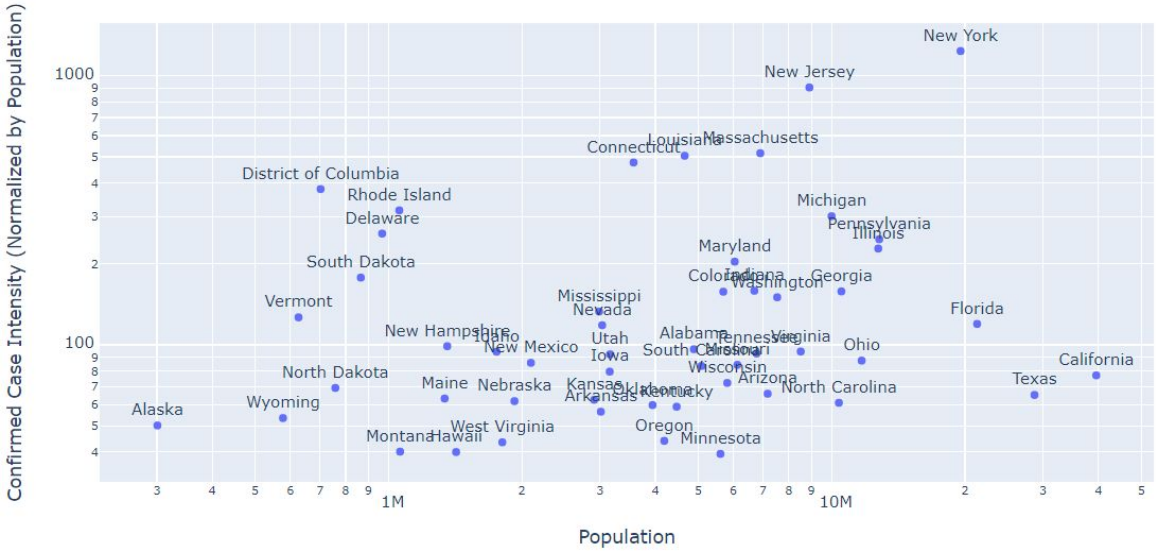

Figure 2: Relationship between population and confirmed cases per 100,000 people, on a log-log scale

We additionally explored the data to confirm our understanding of the virus trends based on media coverage of the virus. Looking at a heatmap for the geospatial density of total confirmed cases in each region of the contiguous US (Figure 3), it is confirmed that New York has the highest intensity within the United States, with California showing a small fraction of its intensity.

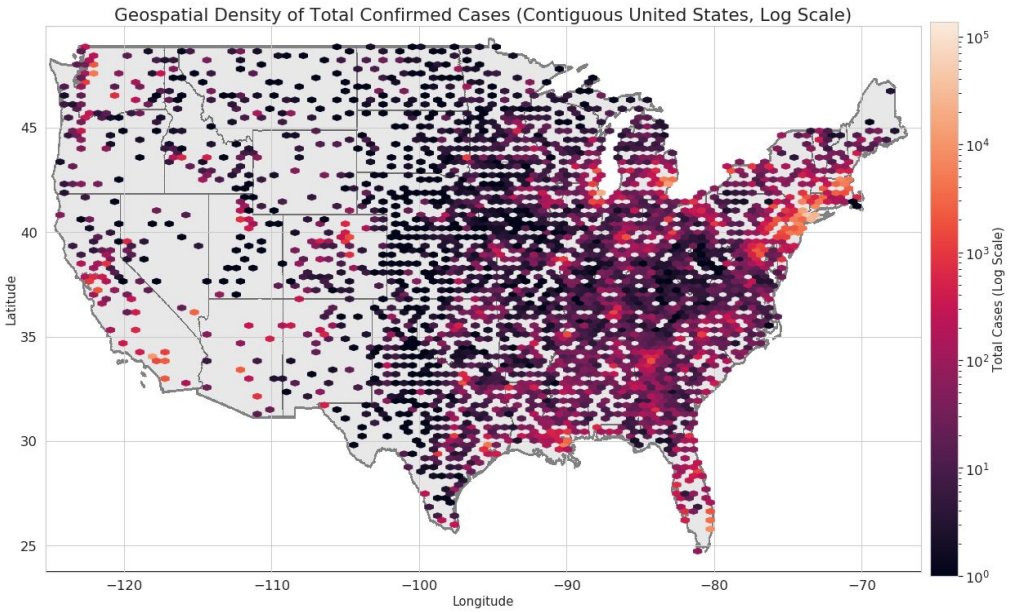

Figure 3: Geospatial density of total confirmed cases in the contiguous United State

During our analysis, we noticed that there were many counties with zero reported cases, even some in close geographic proximity to counties with high case numbers. Thus we decided to limit our analysis to counties with more than 10 confirmed cases per 100,000 people on April 18th.

Further data visualization was performed on the time series data sets using the case intensity metric, which showed that some states experienced decreases in rate of spread near the end of the time period studied, motivating us to look closely at different dynamic features as potential causes. The plots for this exploration can be found in the accompanying Jupyter Notebook.

Method and Experiments

The first and second questions of this analysis were to explore the impact of various static metrics on the initial spread of the virus in a given region, and the impact of stay-at-home orders enacted to limit the spread. To facilitate this analysis, the time series data sets for each county were divided into pre- and post- intervention time periods. In deciding which intervention to consider for the dividing date, several were explored but it was decided to focus on the stay-at-home orders. The reasons for this were two-fold; first, several of the dynamic interventions were imposed at the federal level, causing no variation in county data, or relatively early on in the virus spread in a region, making it difficult to determine any trends pre-intervention; and second, the stay-at-home orders have proved to be the most divisive of the interventions, with the national debate of their merit, impact, and continuation dominating much of the public discourse. For counties where no stay-at-home order was implemented prior to April 18, 2020, the entire time series was included in the pre-intervention analysis.

Inherent in these models are several assumptions that are important to recognize. The most significant assumption that our team identified is that the data provided is accurate. The problems and shortages associated with testing for the COVID-19 virus are extensive and well-reported, and this means that our time series are likely undercounting the actual values by an unknown degree. Estimates of the magnitude of this undercount vary widely and are highly region-dependent (e.g., studies in South Korea may not be applicable in the US). Due to the uncertainty pertaining to the undercount, the decision was made to fit the model to the reported data as-is, with the caveat that it represents a lower limit of the estimated responses as a result of this limitation. In the division of the data set to pre-/post-intervention, we begin with the assumption that the intervention becomes effective the date of implementation. In actual fact, the effect of the intervention will lag after it is implemented, with a lag likely between the

incubation period and the total infection period. We attempt to answer our first two questions of interest with this assumption. Then, for our third question of creating a predictive model, we explore the time lag between intervention implementation and effect.

We used linear least-squares regression to develop our models of the relationships between our factors of interest and response variables of interest. The governing equation and equation of error are:

$$Response\ Variable = \sum_{i=1}^{N} w_i\ Feature_i \quad ; \quad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Actual\ Response - Predicted\ Response)^2}$$

Prior to fitting the model, we standardized each feature by subtracting its mean and dividing by its standard deviation:

$z = \frac{x - mean(x)}{SD(x)}$, where z is the standardization of a feature x.

Further, ridge regression was used to regularize our models and prevent overfitting. The data was divided into a 90/10 train/test split, and the training data was further divided for k-fold cross validation with k = 5. Cross validation was used to tune the regularization hyperparameter, alpha, for each model by picking the alpha with the smallest cross validation error. An example of the cross validation vs. hyperparameter plot used to find the best alpha is shown in Figure 4.

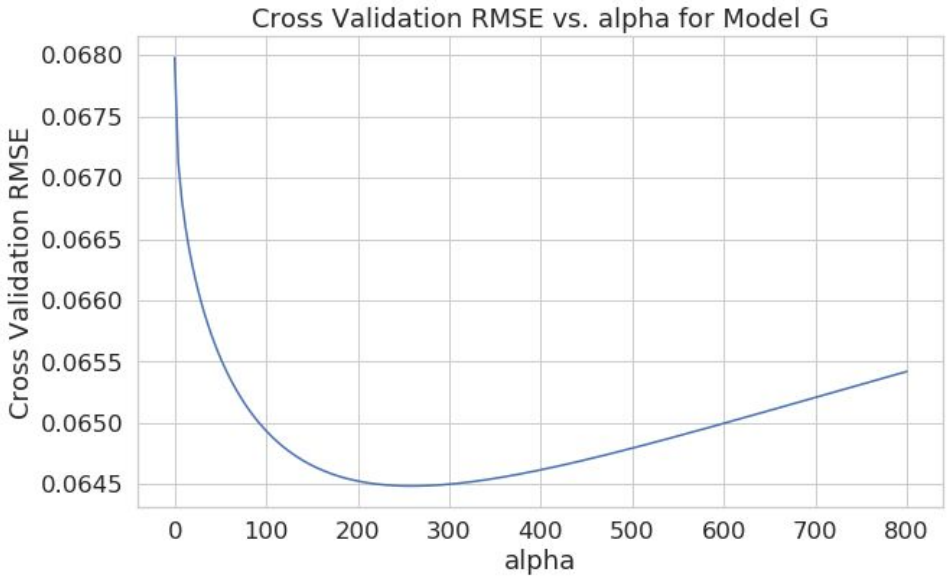

Figure 4: Cross validation RMSE vs. alpha for Model G. The best alpha in this case is $\alpha = 260$.

We developed 7 models, each targeting a specific one of our research questions, and developing on the lessons learned from the previous steps. We summarize the features and response variables for each model in Table 3 below:

Table 3: Features and Response Variables for Each Model

Model Name	**Features**	**Response Variable**
Model A	- All Static Factors (listed in Table 2)	Confirmed Case Intensity [cases per 100,000 people]
Model B	- All Static Factors - x^2 and $\sqrt{x}$ transform of each factor	Confirmed Case Intensity [cases per 100,000 people]
Model C	- All Static Factors - x^2 and $\sqrt{x}$ transform of each factor	Death Intensity [deaths per 100,000 people]
Model D	- All Static Factors - x^2 and $\sqrt{x}$ transform of each factor	Rate of Spread Before Interventions [increase in cases per 100,000 people, per day]
Model E	- All Static Factors - Time between first confirmed case and Stay-at-Home order - x^2 and $\sqrt{x}$ transform of each factor	Increase in Confirmed Case Intensity before Stay-at-Home [increase in cases per 100,000 people]
Model F	- All Static Factors - Time between Stay-at-Home order and April 18th - x^2 and $\sqrt{x}$ transform of each factor	Increase in Confirmed Case Intensity after Stay-at-Home [increase in cases per 100,000 people]
Model G	- All Static Factors - All Dynamic Factors (listed in Table 2)	Proportional Increase In Confirmed Cases, compared to previous week - Evaluated each week, and for a range of assumed delays in intervention effect, 3-24 days

As shown in Table 3, we began by studying static factors only, and described as purely linear the relationship between static factors and confirmed case intensity in Model A. Upon evaluation of Model A and reflection of the nonlinear trends identified in the exploratory data analysis, we employed feature engineering in Model B by using the square and square root transforms of the features. Similarly, Models C and D use transformations of the features, but study the relationship to the other two measures of severity: death intensity and rate of spread. To address our project goals related to evaluating the effect that the introduction of stay-at-home orders on the spread of the virus, we fit a second set of models in Models E and F to the pre- and post- stay-at-home order data sets, with the nonlinear transforms of both static features and time. Finally, to address our goal of creating a predictive model, we evaluated Model G, for which we used a weekly proportional increase in confirmed cases as the measure of spread rate. This was calculated by dividing the number of cases in a given week by the number of cases in the previous week. Model G was fit using daily data for a range of 3-24 days as assumed time lags, to iteratively determine the most likely period of time between the implementation and effect of interventions.

While each model provided valuable insight, Models B, F, and G performed best for fitting the response, and provided the best information to draw conclusions from regarding relative effects of the static and dynamic factors.

Results

From our experiments, we were able to perform some data-driven reasoning to address our project questions. First, in answering our project Question 1 regarding the relative effect of various static factors on severity of COVID-19 spread, we found that population density dominated by a wide margin

in Models A-C. Following population density, health-related features such as respiratory disease rate and stroke disease rate, and proportion of population that are elderly generally had the next highest weights. Next, in determining the relative effect of the stay-at-home order for project Question 2 using Models E-F, we found that the time between the first confirmed case and the stay-at-home order for a region ranked either first or second as the most important feature. This finding was particularly interesting because it indicates that implementing stay-at-home early on, before the virus is able to spread too quickly, may have an impact on spread rate in general. We also found that population density per square mile, disease mortality rates, and proportion of elderly in the population continued to be important factors in Models E-F.

Then, we moved on to our project Question 3 for creating a model which could model the weekly increase in spread rate in Model G. This model achieved a much better fit in response than any of the previous models (Figure 5), indicating that it is a good model to draw conclusions regarding effectiveness of each feature in predicting response. Comparing all of the static and dynamic factors as our features in this model, we found that the stay-at-home order had the highest assigned weight by far, then followed by population density. Limits to social gatherings and proportion of the population which are elderly followed with the next highest weights. We also determined through iterative analysis that an assumed delay of 20 days between intervention implementation and effect resulted in the best score for model fit (Figure 6). We note that there is a slight periodic pattern in the relationship between assumed delay times and model fit, which may be an artifact of taking per-week measures of confirmed cases.

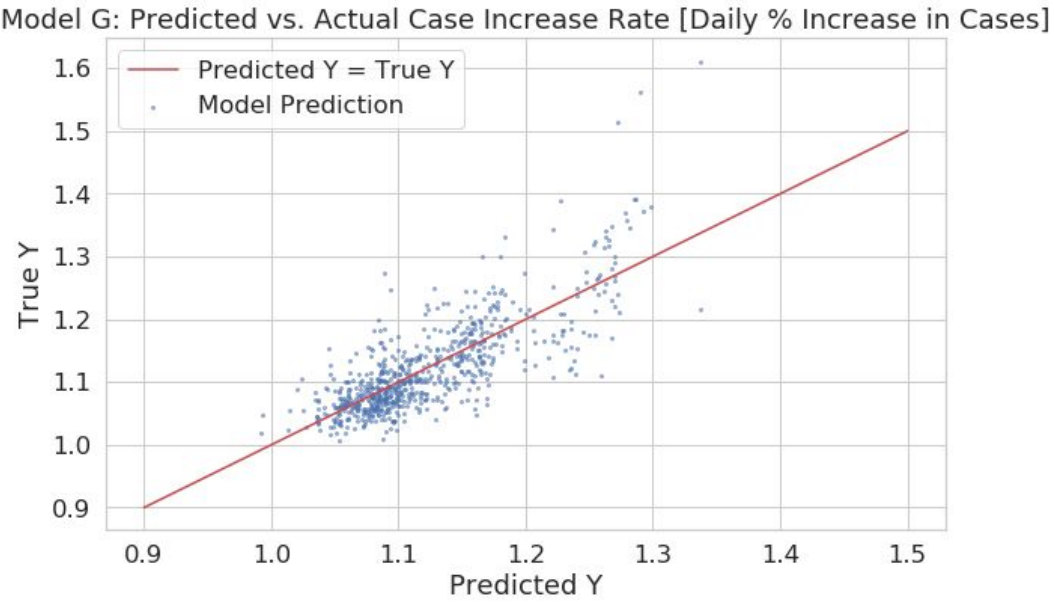

Figure 5: Predicted vs. actual rate of spread for Model G

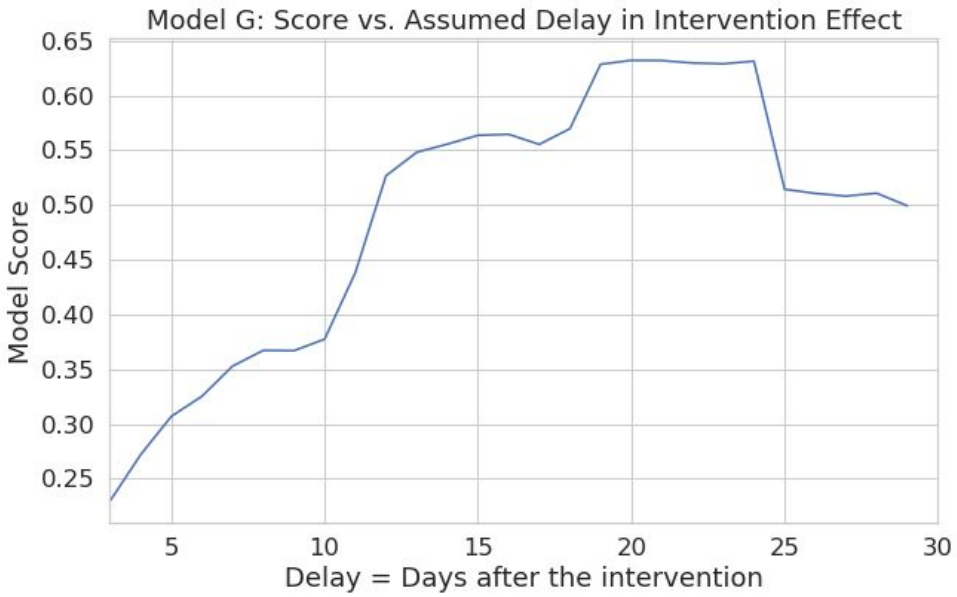

Figure 6: Model score vs. assumed delay in intervention effect for Model G

Following our modeling process and exploration of our key questions, we had the following observations.

1. What were two or three of the most interesting features you came across for your particular question?
    - As we initially suspected based on several of the outbreak hotspots, population density was consistently one of the most important factors across all models.
    - We found that the stay-at-home order was a strong predictor for rate of spread in comparison to other factors.
    - In several of our models, the population proportion by gender in several age ranges were in the top factors by weight, with older age groups having higher weights.
2. Describe one feature you thought would be useful, but turned out to be ineffective.
    - We expected the level of urbanization to be a strong predictor for the case intensity and spread rate, however it did not turn out to be dominant in our models.
    - One early hypothesis was that the democrat/republican ratio would be a strong predictor as a surrogate for a number of demographic factors, however it only had a high weight in one of our models with a poor fit and was not dominant in better models.
3. What challenges did you find with your data? Where did you get stuck?
    - The time dependence and time series nature of several of the variables were a challenge to deal with. We had to make decisions about granularity (daily, weekly, average?), as well as format the time data appropriately to superimpose the correct time periods in our analysis.
4. What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?
    - Our model relies on the assumption that the population will behave the same throughout the analyzed time period. For instance, adherence to the stay-at-home order will not wane with time, and testing availability remains constant (both not necessarily the case).
    - We made assumptions as to the validity of the data (particularly the case count) and incubation time, discussed previously in the report.
5. What ethical dilemmas did you face with this data?
    - It was often easy to lose sight of the human factor of the data - forgetting that each number in the case and death count represented a real person.

    - Several of the outbreaks have disproportionately affected minority populations; however there were concerns with the implications of bringing in racial demographic data.
6. What additional data, if available, would strengthen your analysis, or allow you test some other hypotheses?
    - A more accurate estimate of the testing undercount (the proportion of confirmed cases vs. the actual infected rate in the population) is critical to give an accurate estimate of the true spread and prevalence of the virus.
    - Demographic factors (affluence, education, etc.) may be correlated with higher access to testing, and therefore proportionally skew the confirmed case populations.
7. What ethical concerns might you encounter in studying this problem? How might you address those concerns?
    - We want to avoid confusing correlation with causation. It can be tempting at times to attribute blame to demographic patterns which are highly correlated with the response. Our earlier models did not show a strong fit between predicted and expected response; thus, we do not claim any causation between the factors and responses in these models, but only observe high correlations.

Conclusion

Our models were successful in identifying several potentially important factors that affect the severity of COVID-19 spread in a region. We found that population density, out of all static factors, is by far the most highly related to the intensity of confirmed cases and rate of spread of the virus. We also found that stay-at-home orders have the strongest relationship with the rate of spread out of all factors, indicating that they were having a significant impact on the rate of spread, once implemented. In attempting to predict weekly rates of spread, we were able to develop a good fit using a model that evaluates the relationship between all factors, both static and dynamic, on rate of spread, taking into consideration incubation time and total infection period of the virus. However, being limited to static linear regression models, even our best model is limited for evaluating this time-varying data, and we were forced to take average rates and make assumptions on accuracy of reported data and time-specific conditions in order to perform a prediction. This made it possible to fit the provided data but limits applicability in predicting future trends once additional unknown factors begin affecting the virus spread.

Tracking the spread of COVID-19 has been a well publicised challenge for our nation and its institutions. While some of the difficulties have been logistical in nature, others have stemmed from the nature of the virus itself in its long incubation period and asymptomatic carriers. Absent widespread testing and contact tracing, careful analysis and modeling of the available data can allow data scientists to estimate the trends behind the numbers. This will hopefully provide important information to the decision makers in our communities and government, allowing them to make informed decisions as they weigh the spread of the disease against the political and economic pressures counter to the available interventions. The main findings of this study are that, while there are a number of factors that contribute to the breadth and speed of the virus' spread in a community, our interventions, specifically staying at home, have had a real and measurable impact in slowing the spread of the virus. One key question which our study was unable to address was how best to determine when the stay at home orders have been effective to the point where they can be gradually lifted; a topic which our leaders are continuing to grapple with today. As with this report's analysis of the spread of COVID-19 in the US, our hope is that decisions made as it declines are informed by data science.