Data 100 Undergrad Final Project PDF (due Wednesday!)

Jonathan Kung, Eric Wang, Athan Diep

TOTAL POINTS

100 / 100

QUESTION 1

1 Question Framing 10 / 10

✓ + 10 pts **Exemplary**

+ 6.66 pts Acceptable

+ 3.33 pts Inadequate

+ 0 pts Blank/Missing

+ 0 pts Dataset choice: Contraceptive

+ 0 pts Dataset choice: Basketball

✓ + 0 pts **Dataset choice: Covid-19**

QUESTION 2

2 Data Cleaning 15 / 15

✓ + 15 pts **Exemplary**

+ 10 pts Acceptable

+ 5 pts Inadequate

+ 0 pts Blank/Missing

QUESTION 3

3 Data Visualization 20 / 20

✓ + 20 pts **Exemplary**

+ 13.33 pts Acceptable

+ 6.66 pts Inadequate

+ 0 pts Blank/Missing

QUESTION 4

4 Method and Experiments 20 / 20

✓ + 20 pts **Exemplary**

+ 13.33 pts Acceptable

+ 6.66 pts Inadequate

+ 0 pts Blank/Missing

+ 0 pts Definition copied from wiki

QUESTION 5

5 Analysis and Conclusion 25 / 25

✓ + 25 pts **Exemplary**

+ 16.66 pts Acceptable

+ 8.33 pts Inadequate

+ 0 pts Blank/Missing

- 4.165 pts Code Unclear

QUESTION 6

6 Composition 10 / 10

✓ + 10 pts **Exemplary**

+ 6.66 pts Acceptable

+ 3.33 pts Inadequate

+ 0 pts Blank/Missing

- 5 pts Code is present in report

- 3.33 pts Code is split across multiple notebooks

+ 0 pts Missing citation(s) for additional dataset(s)

Introduction

In light of the current global pandemic, we wanted to ask the question: what factors are correlated with COVID-19 point prevalence at county levels? Understanding this question from a local perspective could unlock potential insights on how COVID-19 spreads or which communities might be the most vulnerable, and further application of this analysis could guide policy or medical decisions.

The datasets provided figures for confirmed cases and confirmed deaths, but we chose to focus on point prevalence as confirmed cases divided by county population rather than using fatality rates. We felt this value would be more representative of the effects of COVID-19 on a community as a whole for two reasons. First, we wanted to avoid potential misrepresentations of counties with death rates of 0% or 100% attributed to low figures in the data. Additionally, COVID-19 has historically held a relatively low mortality rate[2] and fatality rates are widely varied across age demographics[3].

To tackle our question, we examined the available data features on different counties within the dataset. We decided to bucket the available data into three groups: demographic factors, health risk factors, and social distancing factors.

Data Description

We conducted our analysis on three CSV datasets: an abridged dataset of county-level features, a time-series dataset on county-level US confirmed cases, and a time-series dataset on county-level US deaths. We appended the most current (April 18th, 2020) figures on confirmed cases and deaths with the abridged table and computed current death rates and confirmed prevalence rates based on population size.

The abridged county-level dataset included 3,244 county records with 87 feature columns per record. The confirmed cases and confirmed death tables included 3,255 records, with each record also corresponding to a specific county in the US. These tables provided data on geographic location and cases per day from January 1st, 2020 to April 18th, 2020.

We chose to focus on three types of features, as mentioned above: demographic factors, health risk factors, and social distancing factors.

For demographic factors, we isolated the following factors: `'FracMale2017'`, `'SVIPercentile','MedicarePercent',PopulationDensityperSqMile2010','MedianAge2010','65+Percent','dem_to_rep_ratio'`.

For health risk factors, we examined `'3-YrDiabetes2015-17','DiabetesPercentage', 'HDMortalityPercent', 'StrokeMortalityPercent','Smokers_Percentage', 'RespMortalityRate2014'`.

[2] https://news.berkeley.edu/2020/04/24/study-challenges-reports-of-low-fatality-rate-for-covid-19/
[3] https://www.businessinsider.com/coronavirus-compared-seasonal-flu-in-the-us-death-rates-2020-3

1 Question Framing 10 / 10

- ✓ **+ 10 pts** **Exemplary**
- **+ 6.66 pts** Acceptable
- **+ 3.33 pts** Inadequate
- **+ 0 pts** Blank/Missing
- **+ 0 pts** Dataset choice: Contraceptive
- **+ 0 pts** Dataset choice: Basketball
- ✓ **+ 0 pts** **Dataset choice: Covid-19**

For social distancing factors, we considered `'stay at home', '>50 gatherings','>500 gatherings', 'public schools', 'restaurant dine-in', 'entertainment/gym', 'federal guidelines', 'foreign travel ban'`.

All of the data we analyzed is quantitative. The primary keys are discrete integers, as are most raw demographic figures. Social distancing measures are categorical numeric values, while health risk percentages and rates are quantitative discrete values.

Methodology: EDA and Transformations

We used a variety of approaches for data cleaning and exploratory data analysis.

To make the dataset more manageable, we removed the columns that were not relevant to our analysis direction. These were primarily the columns on population breakdown by age and mortality breakdown by age.

Primary Key Cleaning

The primary key we used for merging tables and indexing the data was `countyFIPS` in the abridged county dataset, which is the Federal Information Processing Standards code for uniquely identifying county and county equivalents in the US. We ensured all primary keys were in numeric form prior to merging.

Upon examination of the counties in the dataframes, we also found various errors and inconsistencies in our selected primary key column, `FIPS`. Two of the records in abridged counties were filler records without corresponding information, so these records were dropped.

We found four records where the FIPS was NaN: Kansas City, Dukes and Nantucket, Michigan Department of Corrections, and the Federal Correctional Institution. To consider if we should remove these records, we looked at their number of confirmed cases and confirmed deaths. All of these counties had non trivial figures relative to US averages, so we decided to include them in our dataset. For Duke and Nantucket, we consulted a USDA directory of county FIPS codes and found that Duke and Nantucket are officially recognized as separate counties with unique FIPS. Since we had no way to determine the split of values between the two counties, we excluded this record from our analysis. Kansas City and the MDOC have similar issues of spanning multiple counties, so they were also excluded. The Federal Correctional Institute is officially considered part of Washtenaw county, so we amended the Washtenaw figures to include FCI values.

We also found 3 sets of duplicate FIPS in abridged counties: 60020, 66010, 69120. For 60020 and 69120, the only identical values were on healthcare shortage statistics: `HPSAShortage, HPSAServedPop, HPSAUnderservedPop`. All other fields were blank, so we removed these records. Upon examining the records for 66010, we found identical values in every non-NaN field. The two counties (Cocos Island and Guam) have the same FIPS, yet are technically part of different states[4]. Thus, we combined the two records into one with a joint name.

After these modifications, our data primary key had no duplicate or NaN values.

[4] https://en.wikipedia.org/wiki/Cocos_Island_(Guam)

plots demonstrate discernible patterns or linear relationships between the fields and the death rate. In addition, we realized that there's a large proportion of death rates at 0, which would cause potential issues with our intended linear regression model by biasing the predicted death rate towards 0. As such, we decided to change our response variable to the percentage of confirmed cases, which has a better spread illustrated below.

Death Rate

We found 7 counties with death rates = 1, and all have very low figures for deaths and confirmed (=1.0 or 2.0). Given that the figures are very low, the fact that their death rate is 100% is not as significant, and we exclude these rows to make our dataset more robust.

Confirmed Percent

We considered using the most current confirmed case values as a response variable in our model. Because the number of cases in a county is likely dependent on the population, we created the `ConfirmedPercent` variable as a normalized measure of confirmed case rates. This variable, along with different transformations of it, was plotted against the selected features. There seemed to be clearer relationships between the features and *Confirmed Percent* compared to *Death Rate* so the confirmed case rate was kept as our response variable in the final models.

Methodology: Models and Assumptions

$$(1)\ ConfirmedPercent = \alpha_0 + \sum_{i=1}^{7} \alpha_i \cdot demographic_i + \sum_{j=1}^{6} \beta_j \cdot health_j + \sum_{k=1}^{8} \gamma_k \cdot social_k + \varepsilon$$

$$(2)\ ConfirmedPercent = \alpha_0 + \sum_{i=1}^{7} \alpha_i \cdot demographic_i + \varepsilon$$

$$(3)\ ConfirmedPercent = \alpha_0 + \sum_{j=1}^{6} \beta_j \cdot health_j + \varepsilon$$

$$(4)\ ConfirmedPercent = \alpha_0 + \sum_{k=1}^{8} \gamma_k \cdot social_k + \varepsilon$$

We decided to initially try four different models based on our exploratory data analysis: one model for each of the three feature groups (demographics, health risks, social distancing) and one global model which included all features from all three categories. Equation (1) is the global model that fitted a total of 22 parameters to the data (including the intercept). Linear models were also fit to individual feature groups such as the models given in Equations (2), (3), and (4).

Cross Validation Procedure

For simplification purposes, we chose to use the built-in cross-validation functions included in scikit-learn. We split our existing data into a training set (85%) and test set (15%). After standardizing our features, we fitted the corresponding training data subsets to the four different linear regression models. We then calculated MSE on our training predictions and test predictions to obtain MSE values.

2 Data Cleaning 15 / 15

- ✓ **+ 15 pts** **Exemplary**
- **+ 10 pts** Acceptable
- **+ 5 pts** Inadequate
- **+ 0 pts** Blank/Missing

Geographic Features

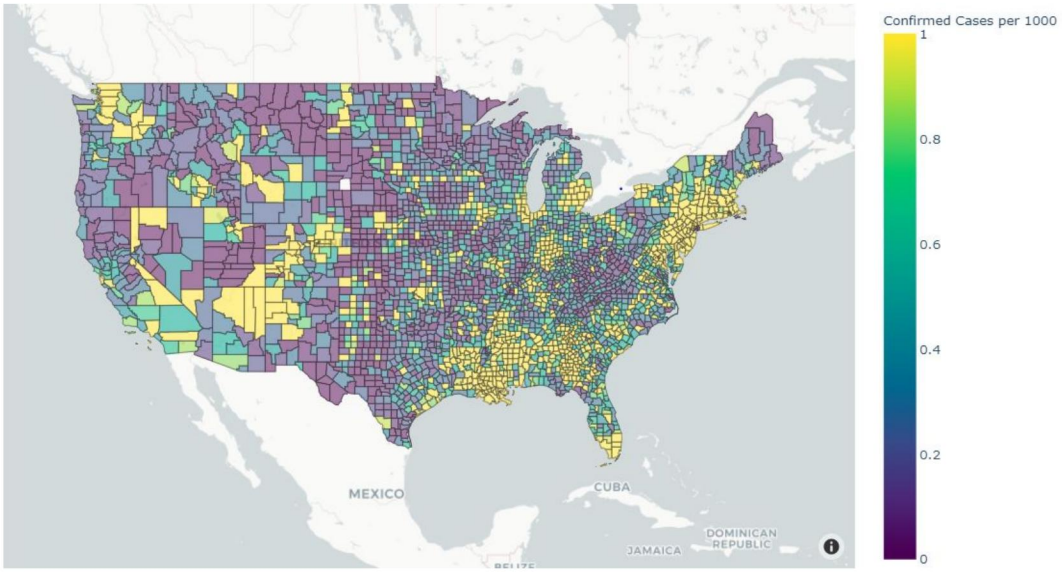

Figure 1a: Confirmed COVID-19 cases per 1000 people (separated by county)

In addition to the demographic, social distancing, and health risk factors, we wanted to check whether there were any geographical patterns that are present in the death rate and number of confirmed cases. Using the FIPS key in our dataset and a dataset found online mapping counties to geographical coordinates, we were able to create choropleth maps of our potential response variables and features. Figure 1a shows the confirmed case rate per 1000 people. There seems to be clustering of high confirmed case rates in particular regions such as in the northeast and southern United States.

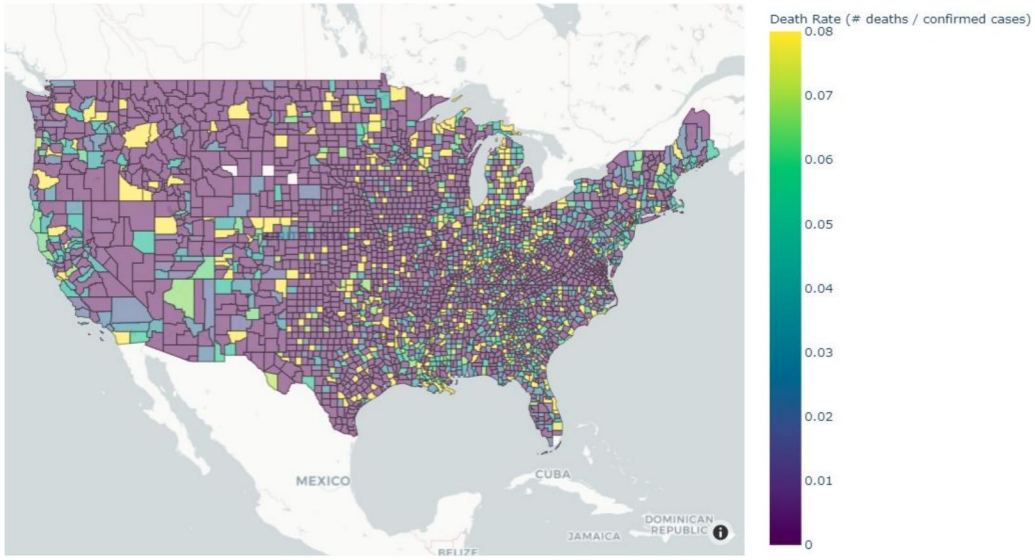

Figure 1b: Confirmed COVID-19 deaths per 1000 people separated by county

Exploring and Modifying Features

In order to make values comparable for counties of different sizes, we converted some raw figures into percentages. This applied to values on population estimates, Medicare Enrollment, and mortality rates.

In order to validate and begin exploring our selected feature buckets, we created correlation heat maps for each of the three categories:

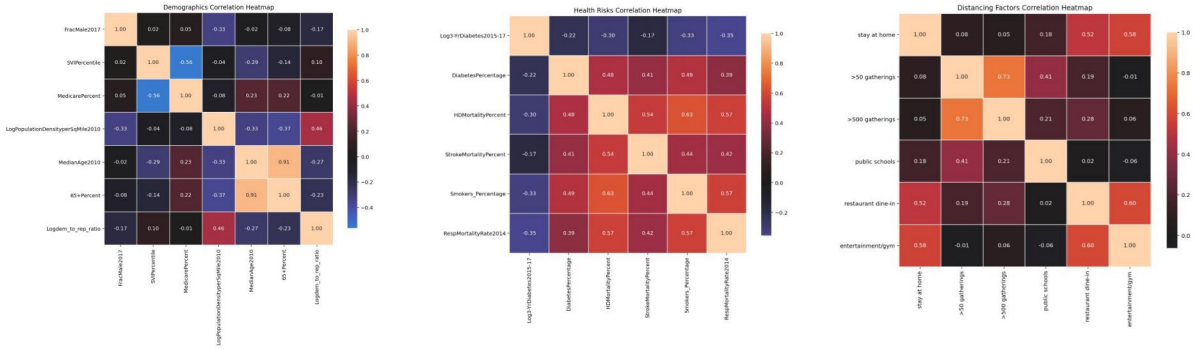

Figure 2: Correlation Heat Maps for Demographic, Health, and Social Distancing Factors

The social distancing correlation plot generally confirmed the intuition we had about how activities are restricted based on government social distancing mandates. For example, restaurant dine-in and entertainment have the highest non-trivial correlation, and we observed that the stay at home measure also correlates with both restaurant dine in and entertainment.

We found relatively significant correlations between most features in the health risk group. HDMortality presents the strongest correlations with the other features like diabetes percentage, smokers percentage, stroke mortality rate, and respiratory mortality rate. The correlations found in this data might suggest that correlations with other disease and health risks like COVID-19 might also be significant.

There were also a number of interesting correlations in the demographic data. Most notably, SVIPercentile and Medicare enrollment have a strong negative correlation. Political alignment is correlated with age distribution and population density, which also corroborates common knowledge.

We initially considered both death rate and confirmed rate as potential response variables to investigate in our analysis, so we furthered our exploratory analysis by generating pairplots of our feature groups against death rate and confirmed percentages.

Demographic Factors

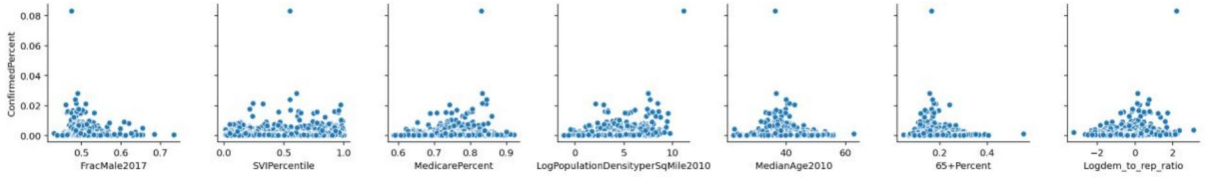

Figure 3a: Pairwise Plots for Demographic Factors vs. Confirmed Percent

We noted some insights on the variance of these plots. There appears to be relatively similar variance of `SVIPercentiles`, which supports the assumption of homoscedasticity in linear regression. We also see larger variance for a medium range of `MedicarePercent, MedianAge2010`, and `65+ Percent`, and high variance for a small range of `FracMales. PopulationDensity perSqMile2010` and `dem_to_rep_ratio` figures are all positive values, which is consistent with their semantics. There is only one zero value for population density, for which the death and confirmed counts are 0 and 1, respectively.

Since there are no zero or negative values and a right skewness for both fields, it seems reasonable to apply a log transformation. After doing so, there appears to be more of an equal spread.

Health Risk Factors

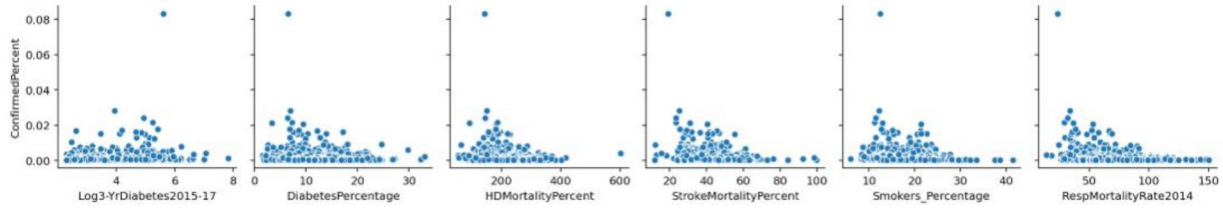

Figure 3b: Pairwise Plots for Health Risk Factors vs. Confirmed Percent

There are no negative or zero values for `3-YrDiabetes2015-17`. Since the field is a percentage and there is right skewness, it might be appropriate to apply a log transformation. After doing so, there appears to be more of an equal spread.

It looks like there's a few frequently occurring death rates across the five fields excluding `3-YrDiabetes2015-17`. Since there doesn't seem to be an obvious relationship between the fields and death rate, it would be useful to incorporate domain knowledge from an expert at this point.

Social Distancing Factors

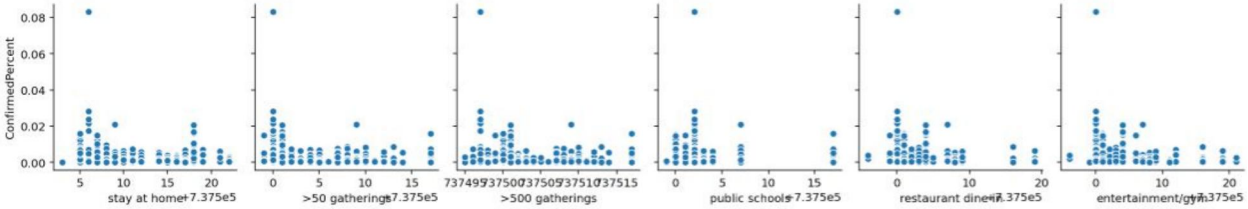

Figure 3c: Pairwise Plots for Social Distancing Factors vs. Confirmed Percent

We remove federal guidelines and foreign travel ban as both are uninformative features (only one value each)

We looked for any apparent relationships between the fields and the number of confirmed cases and deaths, the two fields used in constructing our death rate metric, as well as the metric itself. None of the scatter

3 Data Visualization 20 / 20

- ✓ **+ 20 pts** **Exemplary**
- **+ 13.33 pts** Acceptable
- **+ 6.66 pts** Inadequate
- **+ 0 pts** Blank/Missing

plots demonstrate discernible patterns or linear relationships between the fields and the death rate. In addition, we realized that there's a large proportion of death rates at 0, which would cause potential issues with our intended linear regression model by biasing the predicted death rate towards 0. As such, we decided to change our response variable to the percentage of confirmed cases, which has a better spread illustrated below.

Death Rate

We found 7 counties with death rates = 1, and all have very low figures for deaths and confirmed (=1.0 or 2.0). Given that the figures are very low, the fact that their death rate is 100% is not as significant, and we exclude these rows to make our dataset more robust.

Confirmed Percent

We considered using the most current confirmed case values as a response variable in our model. Because the number of cases in a county is likely dependent on the population, we created the `ConfirmedPercent` variable as a normalized measure of confirmed case rates. This variable, along with different transformations of it, was plotted against the selected features. There seemed to be clearer relationships between the features and *Confirmed Percent* compared to *Death Rate* so the confirmed case rate was kept as our response variable in the final models.

Methodology: Models and Assumptions

$$(1)\ ConfirmedPercent = \alpha_0 + \sum_{i=1}^{7} \alpha_i \cdot demographic_i + \sum_{j=1}^{6} \beta_j \cdot health_j + \sum_{k=1}^{8} \gamma_k \cdot social_k + \varepsilon$$

$$(2)\ ConfirmedPercent = \alpha_0 + \sum_{i=1}^{7} \alpha_i \cdot demographic_i + \varepsilon$$

$$(3)\ ConfirmedPercent = \alpha_0 + \sum_{j=1}^{6} \beta_j \cdot health_j + \varepsilon$$

$$(4)\ ConfirmedPercent = \alpha_0 + \sum_{k=1}^{8} \gamma_k \cdot social_k + \varepsilon$$

We decided to initially try four different models based on our exploratory data analysis: one model for each of the three feature groups (demographics, health risks, social distancing) and one global model which included all features from all three categories. Equation (1) is the global model that fitted a total of 22 parameters to the data (including the intercept). Linear models were also fit to individual feature groups such as the models given in Equations (2), (3), and (4).

Cross Validation Procedure

For simplification purposes, we chose to use the built-in cross-validation functions included in scikit-learn. We split our existing data into a training set (85%) and test set (15%). After standardizing our features, we fitted the corresponding training data subsets to the four different linear regression models. We then calculated MSE on our training predictions and test predictions to obtain MSE values.

Interestingly, we saw that the training MSE was significantly higher for global and demographic models. Test MSE was marginally higher for the social distancing and health factor models. Overall, the health risk factor model had the lowest MSE across both test and training sets. One note regarding error values is that our response variable, `ConfirmedPercent`, has an average value of 0.001017, so the magnitude of our MSE values should be viewed in the context of that scale.

The global model showed high error rates, so we were curious to see if we could create a more effective model that incorporated variables across different feature buckets. After experimenting with different subsets of features, we found that test and train set MSE were minimized for the following combination: `'Log3-YrDiabetes2015-17', 'HDMortalityPercent', 'Smokers_Percentage', 'MedianAge2010', '65+Percent'`.

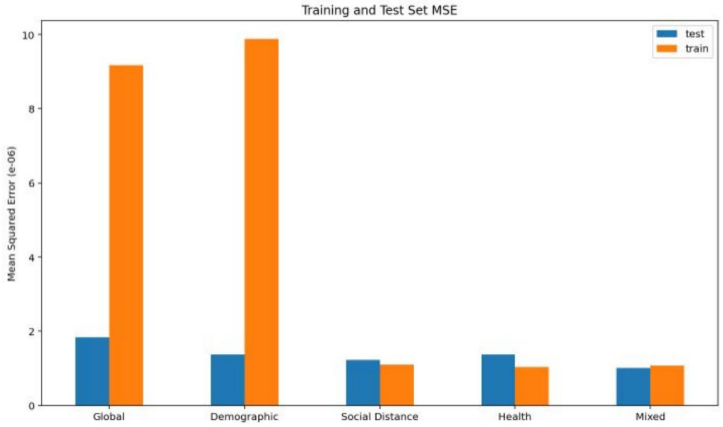

Figure 4: Training and Test Set MSE for Models

Regularization

Considering that models typically have higher performance on training sets than test sets, we wanted to further refine our model with cross-validation for regularization and hyperparameter tuning.

Our preliminary analysis using L1 regularization (lasso regression) indicated that the best alpha value is 0.001 and the cross validation error for optimal alpha is ~0.002995. The alpha value was always at the lower bound of the search range and the optimal (in terms of MSE) model under Lasso was that with all estimated coefficients equal to 0. This result is consistent with the characteristics of the Lasso model which tends to shrink coefficients to exactly 0. Even though the optimal Lasso model produced a lower test MSE than the unregularized model, it did not provide that much value as estimated coefficients of 0 failed to give us a better picture of the relationship between the response variable and features. Tuning the hyperparameter also yielded little difference in cross validation error which was largely constant as can be seen in Figure 5a below.

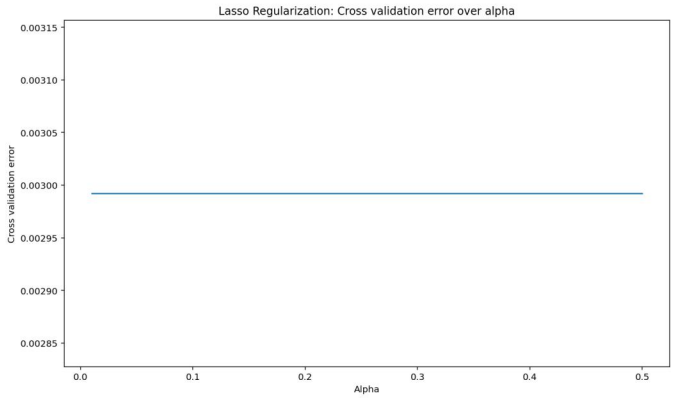

Figure 5a: Alpha vs. Cross Validation Error from Lasso Regularization

The L2 regularization (ridge regression) was also fit to the training data and its hyperparameter (alpha) was tuned using cross-validation. The tuning procedure resulted in an optimal alpha value that was always at the upper bound of the search range and estimated coefficients close to the unregularized full linear model for higher values of alpha.

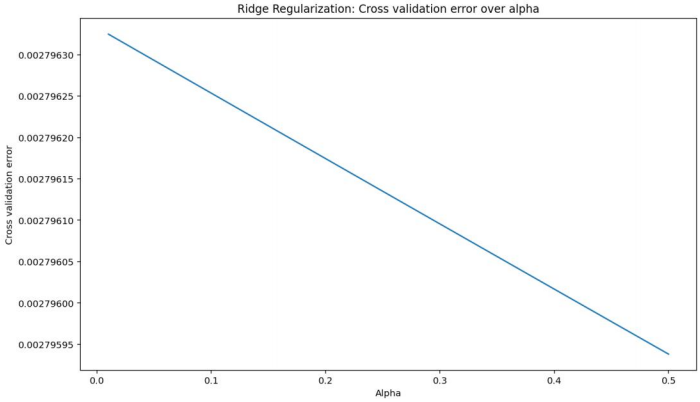

Figure 5b: Alpha vs. Cross Validation Error from Ridge Regularization

Regression Diagnostics

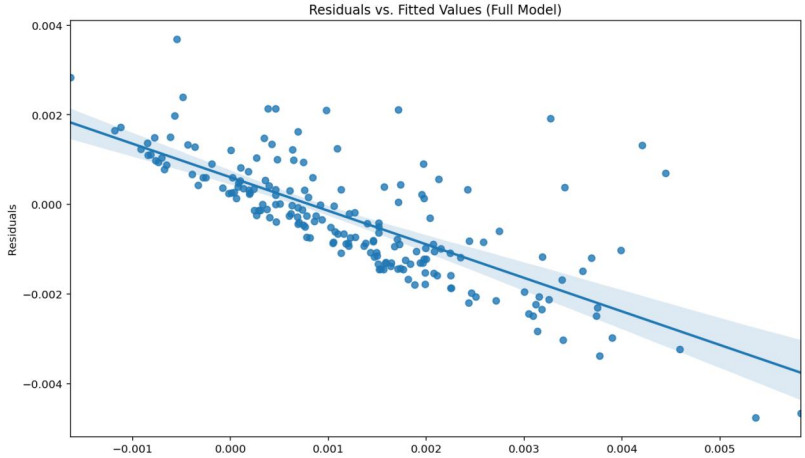

Figure 6: Residuals vs. Fitted values from full model

4 Method and Experiments 20 / 20

- ✓ **+ 20 pts** **Exemplary**
- **+ 13.33 pts** Acceptable
- **+ 6.66 pts** Inadequate
- **+ 0 pts** Blank/Missing
- **+ 0 pts** Definition copied from wiki

Our linear model relies on the assumption of homoscedasticity and errors with mean 0. Upon inspection of the residual plot shown in Figure 6, it seems that these assumptions are violated as the residuals are not centered around 0 and have non-constant variance across different fitted values. This indicates that a linear model may not adequately explain the relationship between the response variable and our features.

Summary of Results

We sought to explore the potential impact of certain demographic, health risk, and social distancing factors (county level) on the percent of confirmed COVID-19 cases. Through extensive EDA, feature engineering, and comparing models (OLS vs. Lasso vs. Ridge with different feature buckets and hyperparameters) via cross-validation errors, we came to several key conclusions:

- Correlation amongst feature buckets:
    - Within demographic features, SVIPercentile, a ranking demonstrating the county's social vulnerability, is negatively correlated with Medicare enrollment. Given that our health risk model obtained the lowest MSE across test and training sets, social vulnerability alongside Medicare or other health related features may be worth exploring further.
    - Many significant correlations amongst features in health risk groups suggest that these factors may also have a strong correlation with COVID-19 susceptibility, providing another avenue to explore.
- Non-significant relationships:
    - Our original pairwise plots between individual features, counts of confirmed cases, and death rates (our original response variable) demonstrate no distinguishable patterns or relationships despite many transformations.
- Model performance:
    - We determined that an Ordinary Least Squares model encompassing the following features was the most optimal in striking a balance between reducing MSE and being informative: `'Log3-YrDiabetes2015-17'`, `'HDMortalityPercent'`, `'Smokers_Percentage'`, `'MedianAge2010'`, `'65+Percent'`.
- Regularization:
    - Lasso Model: The optimal alpha values which minimized our MSE were consistently on the lower bound of our search range, resulting in extremely sparse coefficients and uninformative results.
    - Ridge Model: Resulting alpha values were always on the upper bound of our search range with estimated coefficients very similar to that of our unregularized, OLS model.
- Violated assumptions:
    - A close examination of our residual plots yields the conclusion that the assumption of homoscedasticity and errors with mean 0 have been violated, such that an alternative, non-linear model should probably be considered.
- Areas for further exploration:
    - Evaluating our models with further metrics including R^2 values would provide valuable insight on the significance of our conclusions.
    - Exploring more dates within the available time-series data may be beneficial in more accurately explaining growth patterns.
    - More granular data would not only improve the accuracy of our model but also make up for potential loss in information due to data aggregation.

Discussion

Notes on interesting and ineffective features

When experimenting with feature subsets in our mixed model, we found that the health factors `Log3-YrDiabetes2015-17`, `HDMortalityPercent` and `Smokers_Percentage` resulted in the lowest MSE. This suggests that comorbidities may be important to understanding COVID-19, and individuals with these health risk factors could be more susceptible. Additionally, we were surprised that certain demographic features were less effective than we initially hypothesized. For example, we intuitively believed that high population density would cause higher infection rates, but the data did not support this.

Data challenges

One challenge that we faced with the data was the sparsity of the data and the presence of 0 or missing values for our response variables and features. Many of these observations had to be excluded, providing us with a much smaller dataset than that which we started with. Additionally, it was difficult to identify any clear relationships between the feature variables and our response variable. As seen from the pairwise plots, the general shapes suggested fairly uniform distributions. Thus, it was hard to know if our features were truly relevant and meaningful variables in our model for COVID-19.

Limitations of analysis and assumptions

We used the most recent figures (April 18th, 2020) to calculate current confirmed cases for each county, but this may present some limitations. Some counties may be at different stages of contracting or containing the virus, and simply observing the point prevalence for one date does not capture overall growth patterns.

We are unable to fully assess data faithfulness, but it is likely that these figures do not capture the full reality of COVID-19 cases and deaths. This is primarily due to unavailability and incompleteness of testing across the US. Additionally, since the data is compiled from various sources, the time frames for the data are inconsistent. For example, median age statistics are from 2010 and Respiratory Mortality rates are from 2014. We assumed that these county-level metrics did not drastically change up to the present, but there may have been material differences that impact our model for COVID-19. Overall, we do believe our methods were reasonably sound in investigating general links between county-level attributes and COVID-19 risk.

Further analysis for expanded data

Having more complete data, particularly for our health related factors, would likely provide us with more accurate results on the relationship between county-level death rates and confirmed case rates.
More granular data, such as city-specific variables, might also help if some predictive information was lost when the variables were aggregated by county. Some factors such as the social distancing factors were noticed to only have state-level statistics which might lead our models' estimates of the impact of the factors on county-level rates to be inaccurate.

Ethical concerns

As COVID-19 is such an impactful and pressing issue during this time, policy decisions around the virus must be thought out thoroughly and hopefully backed by proper data analysis. This would require that our analysis is sound and the data is accurate. In regards to the data collection process, health related data inherently involve privacy issues as well as algorithmic biases which ultimately impact the outcome of our models and conclusions. Being unaware of these potential problems may lead us to not recognize what factors are actually affecting the spread of COVID-19 or wrongly underestimate the influence of other factors. We could address issues of privacy by encouraging data collectors to promote privacy impact assessments or establish simpler privacy policies. Algorithmic biases can be improved by establishing awareness amongst developers of the potential ethical implications of their work as well as involving domain experts during the construction process.

5 Analysis and Conclusion 25 / 25

✓ **+ 25 pts Exemplary**

+ 16.66 pts Acceptable

+ 8.33 pts Inadequate

+ 0 pts Blank/Missing

- 4.165 pts Code Unclear

Final Project: Exploring County-level Attribute Correlations to COVID-19 Point Prevalence

Athan Diep, Jonathan Kung, Eric Wang

Spring 2020 Data 100

May 13th, 2020

Abstract

COVID-19 has been the defining issue of 2020 across social, political, and economic landscapes. As Data Science 100 students, we wanted to examine available county-level data to understand how COVID-19 affects different communities across the United States. In this paper, we will present results from our analysis on 1288 counties based on datasets from US government and media sources compiled by UC Berkeley researchers[1]. Our study involved exploring features in these categories and their relationship to COVID-19: demographics, health risks, social distancing. We performed proper exploratory data analysis alongside different data science methods including feature engineering, cross validation, and model comparison. As a result of our project, we identified several interesting correlations, areas for further study, and the following key factors: percentage of those with diabetes, heart disease mortality, smokers, median age, and the count of individuals older than 65.

[1] https://github.com/Yu-Group/covid19-severity-prediction/tree/master/data

Introduction

In light of the current global pandemic, we wanted to ask the question: what factors are correlated with COVID-19 point prevalence at county levels? Understanding this question from a local perspective could unlock potential insights on how COVID-19 spreads or which communities might be the most vulnerable, and further application of this analysis could guide policy or medical decisions.

The datasets provided figures for confirmed cases and confirmed deaths, but we chose to focus on point prevalence as confirmed cases divided by county population rather than using fatality rates. We felt this value would be more representative of the effects of COVID-19 on a community as a whole for two reasons. First, we wanted to avoid potential misrepresentations of counties with death rates of 0% or 100% attributed to low figures in the data. Additionally, COVID-19 has historically held a relatively low mortality rate[2] and fatality rates are widely varied across age demographics[3].

To tackle our question, we examined the available data features on different counties within the dataset. We decided to bucket the available data into three groups: demographic factors, health risk factors, and social distancing factors.

Data Description

We conducted our analysis on three CSV datasets: an abridged dataset of county-level features, a time-series dataset on county-level US confirmed cases, and a time-series dataset on county-level US deaths. We appended the most current (April 18th, 2020) figures on confirmed cases and deaths with the abridged table and computed current death rates and confirmed prevalence rates based on population size.

The abridged county-level dataset included 3,244 county records with 87 feature columns per record. The confirmed cases and confirmed death tables included 3,255 records, with each record also corresponding to a specific county in the US. These tables provided data on geographic location and cases per day from January 1st, 2020 to April 18th, 2020.

We chose to focus on three types of features, as mentioned above: demographic factors, health risk factors, and social distancing factors.

For demographic factors, we isolated the following factors: `'FracMale2017'`, `'SVIPercentile','MedicarePercent',PopulationDensityperSqMile2010','MedianAge2010','65+Percent','dem_to_rep_ratio'`.

For health risk factors, we examined `'3-YrDiabetes2015-17','DiabetesPercentage', 'HDMortalityPercent', 'StrokeMortalityPercent','Smokers_Percentage', 'RespMortalityRate2014'`.

[2] https://news.berkeley.edu/2020/04/24/study-challenges-reports-of-low-fatality-rate-for-covid-19/
[3] https://www.businessinsider.com/coronavirus-compared-seasonal-flu-in-the-us-death-rates-2020-3

For social distancing factors, we considered `'stay at home', '>50 gatherings','>500 gatherings', 'public schools', 'restaurant dine-in', 'entertainment/gym', 'federal guidelines', 'foreign travel ban'`.

All of the data we analyzed is quantitative. The primary keys are discrete integers, as are most raw demographic figures. Social distancing measures are categorical numeric values, while health risk percentages and rates are quantitative discrete values.

Methodology: EDA and Transformations

We used a variety of approaches for data cleaning and exploratory data analysis.

To make the dataset more manageable, we removed the columns that were not relevant to our analysis direction. These were primarily the columns on population breakdown by age and mortality breakdown by age.

Primary Key Cleaning

The primary key we used for merging tables and indexing the data was `countyFIPS` in the abridged county dataset, which is the Federal Information Processing Standards code for uniquely identifying county and county equivalents in the US. We ensured all primary keys were in numeric form prior to merging.

Upon examination of the counties in the dataframes, we also found various errors and inconsistencies in our selected primary key column, `FIPS`. Two of the records in abridged counties were filler records without corresponding information, so these records were dropped.

We found four records where the FIPS was NaN: Kansas City, Dukes and Nantucket, Michigan Department of Corrections, and the Federal Correctional Institution. To consider if we should remove these records, we looked at their number of confirmed cases and confirmed deaths. All of these counties had non trivial figures relative to US averages, so we decided to include them in our dataset. For Duke and Nantucket, we consulted a USDA directory of county FIPS codes and found that Duke and Nantucket are officially recognized as separate counties with unique FIPS. Since we had no way to determine the split of values between the two counties, we excluded this record from our analysis. Kansas City and the MDOC have similar issues of spanning multiple counties, so they were also excluded. The Federal Correctional Institute is officially considered part of Washtenaw county, so we amended the Washtenaw figures to include FCI values.

We also found 3 sets of duplicate FIPS in abridged counties: 60020, 66010, 69120. For 60020 and 69120, the only identical values were on healthcare shortage statistics: `HPSAShortage, HPSAServedPop, HPSAUnderservedPop`. All other fields were blank, so we removed these records. Upon examining the records for 66010, we found identical values in every non-NaN field. The two counties (Cocos Island and Guam) have the same FIPS, yet are technically part of different states[4]. Thus, we combined the two records into one with a joint name.

After these modifications, our data primary key had no duplicate or NaN values.

[4] https://en.wikipedia.org/wiki/Cocos_Island_(Guam)

Geographic Features

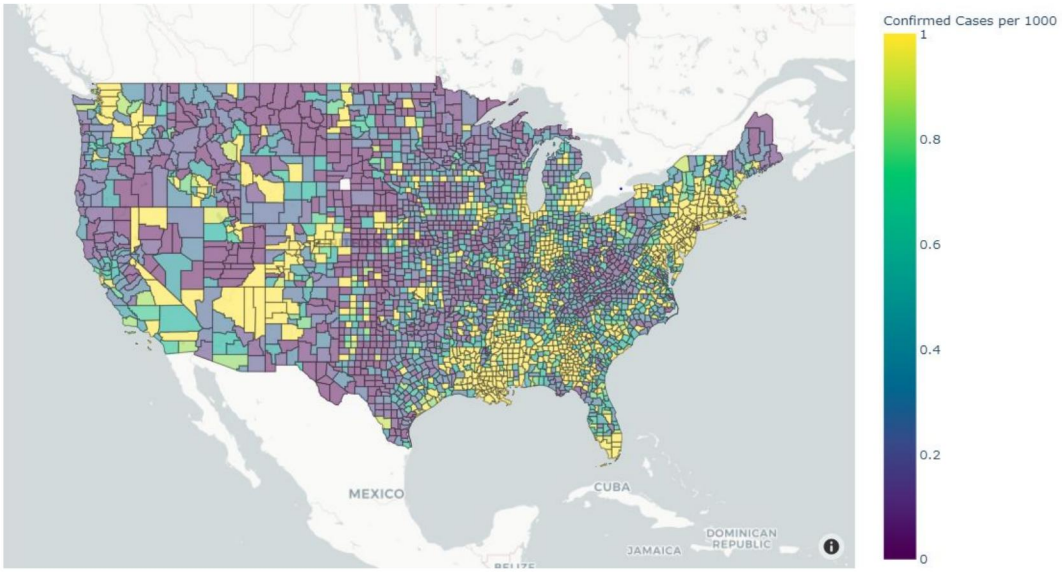

Figure 1a: Confirmed COVID-19 cases per 1000 people (separated by county)

In addition to the demographic, social distancing, and health risk factors, we wanted to check whether there were any geographical patterns that are present in the death rate and number of confirmed cases. Using the FIPS key in our dataset and a dataset found online mapping counties to geographical coordinates, we were able to create choropleth maps of our potential response variables and features. Figure 1a shows the confirmed case rate per 1000 people. There seems to be clustering of high confirmed case rates in particular regions such as in the northeast and southern United States.

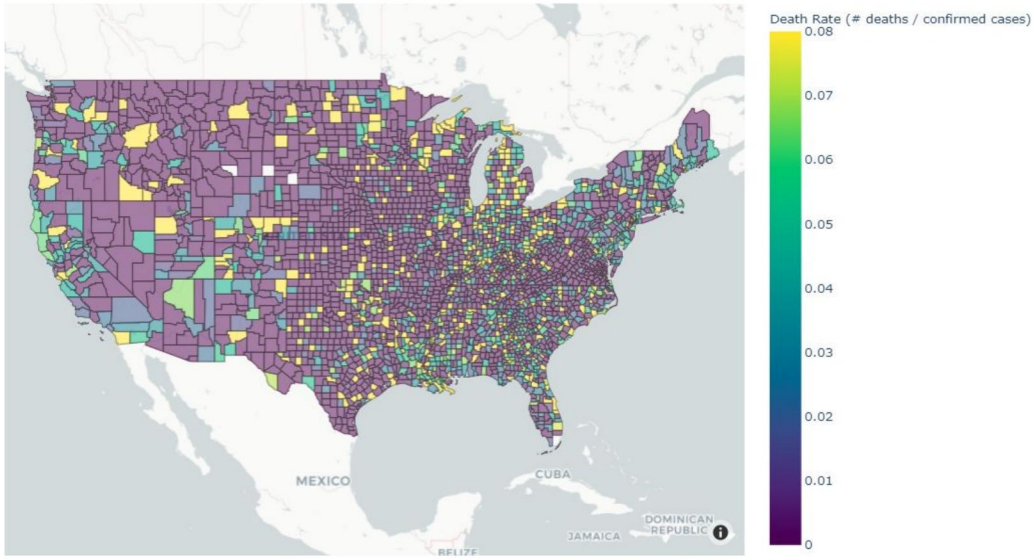

Figure 1b: Confirmed COVID-19 deaths per 1000 people separated by county

Exploring and Modifying Features

In order to make values comparable for counties of different sizes, we converted some raw figures into percentages. This applied to values on population estimates, Medicare Enrollment, and mortality rates.

In order to validate and begin exploring our selected feature buckets, we created correlation heat maps for each of the three categories:

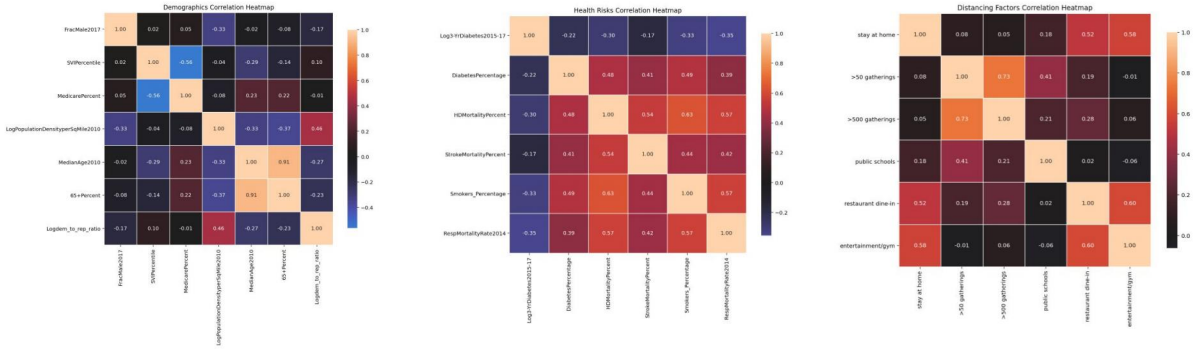

Figure 2: Correlation Heat Maps for Demographic, Health, and Social Distancing Factors

The social distancing correlation plot generally confirmed the intuition we had about how activities are restricted based on government social distancing mandates. For example, restaurant dine-in and entertainment have the highest non-trivial correlation, and we observed that the stay at home measure also correlates with both restaurant dine in and entertainment.

We found relatively significant correlations between most features in the health risk group. HDMortality presents the strongest correlations with the other features like diabetes percentage, smokers percentage, stroke mortality rate, and respiratory mortality rate. The correlations found in this data might suggest that correlations with other disease and health risks like COVID-19 might also be significant.

There were also a number of interesting correlations in the demographic data. Most notably, SVIPercentile and Medicare enrollment have a strong negative correlation. Political alignment is correlated with age distribution and population density, which also corroborates common knowledge.

We initially considered both death rate and confirmed rate as potential response variables to investigate in our analysis, so we furthered our exploratory analysis by generating pairplots of our feature groups against death rate and confirmed percentages.

Demographic Factors

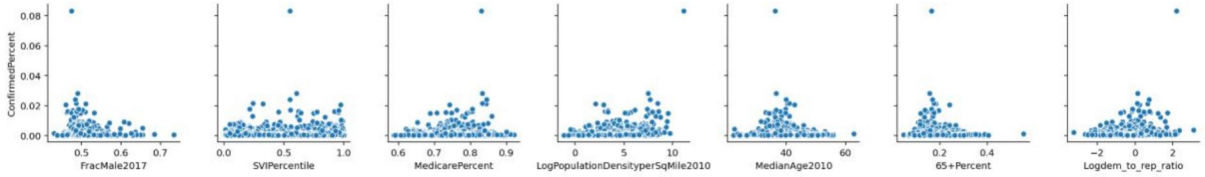

Figure 3a: Pairwise Plots for Demographic Factors vs. Confirmed Percent

We noted some insights on the variance of these plots. There appears to be relatively similar variance of `SVIPercentiles`, which supports the assumption of homoscedasticity in linear regression. We also see larger variance for a medium range of `MedicarePercent, MedianAge2010`, and `65+ Percent`, and high variance for a small range of `FracMales. PopulationDensity perSqMile2010` and `dem_to_rep_ratio` figures are all positive values, which is consistent with their semantics. There is only one zero value for population density, for which the death and confirmed counts are 0 and 1, respectively.

Since there are no zero or negative values and a right skewness for both fields, it seems reasonable to apply a log transformation. After doing so, there appears to be more of an equal spread.

Health Risk Factors

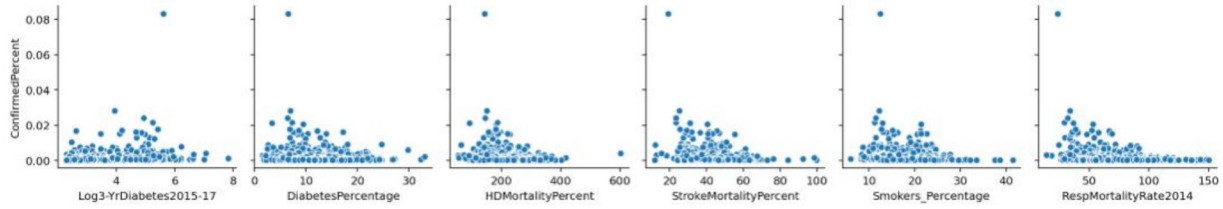

Figure 3b: Pairwise Plots for Health Risk Factors vs. Confirmed Percent

There are no negative or zero values for `3-YrDiabetes2015-17`. Since the field is a percentage and there is right skewness, it might be appropriate to apply a log transformation. After doing so, there appears to be more of an equal spread.

It looks like there's a few frequently occurring death rates across the five fields excluding `3-YrDiabetes2015-17`. Since there doesn't seem to be an obvious relationship between the fields and death rate, it would be useful to incorporate domain knowledge from an expert at this point.

Social Distancing Factors

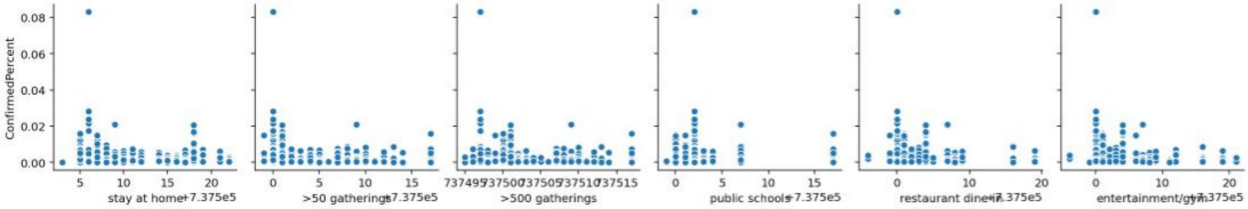

Figure 3c: Pairwise Plots for Social Distancing Factors vs. Confirmed Percent

We remove federal guidelines and foreign travel ban as both are uninformative features (only one value each)

We looked for any apparent relationships between the fields and the number of confirmed cases and deaths, the two fields used in constructing our death rate metric, as well as the metric itself. None of the scatter

plots demonstrate discernible patterns or linear relationships between the fields and the death rate. In addition, we realized that there's a large proportion of death rates at 0, which would cause potential issues with our intended linear regression model by biasing the predicted death rate towards 0. As such, we decided to change our response variable to the percentage of confirmed cases, which has a better spread illustrated below.

Death Rate

We found 7 counties with death rates = 1, and all have very low figures for deaths and confirmed (=1.0 or 2.0). Given that the figures are very low, the fact that their death rate is 100% is not as significant, and we exclude these rows to make our dataset more robust.

Confirmed Percent

We considered using the most current confirmed case values as a response variable in our model. Because the number of cases in a county is likely dependent on the population, we created the `ConfirmedPercent` variable as a normalized measure of confirmed case rates. This variable, along with different transformations of it, was plotted against the selected features. There seemed to be clearer relationships between the features and *Confirmed Percent* compared to *Death Rate* so the confirmed case rate was kept as our response variable in the final models.

Methodology: Models and Assumptions

$$(1)\ ConfirmedPercent = \alpha_0 + \sum_{i=1}^{7} \alpha_i \cdot demographic_i + \sum_{j=1}^{6} \beta_j \cdot health_j + \sum_{k=1}^{8} \gamma_k \cdot social_k + \varepsilon$$

$$(2)\ ConfirmedPercent = \alpha_0 + \sum_{i=1}^{7} \alpha_i \cdot demographic_i + \varepsilon$$

$$(3)\ ConfirmedPercent = \alpha_0 + \sum_{j=1}^{6} \beta_j \cdot health_j + \varepsilon$$

$$(4)\ ConfirmedPercent = \alpha_0 + \sum_{k=1}^{8} \gamma_k \cdot social_k + \varepsilon$$

We decided to initially try four different models based on our exploratory data analysis: one model for each of the three feature groups (demographics, health risks, social distancing) and one global model which included all features from all three categories. Equation (1) is the global model that fitted a total of 22 parameters to the data (including the intercept). Linear models were also fit to individual feature groups such as the models given in Equations (2), (3), and (4).

Cross Validation Procedure

For simplification purposes, we chose to use the built-in cross-validation functions included in scikit-learn. We split our existing data into a training set (85%) and test set (15%). After standardizing our features, we fitted the corresponding training data subsets to the four different linear regression models. We then calculated MSE on our training predictions and test predictions to obtain MSE values.

Interestingly, we saw that the training MSE was significantly higher for global and demographic models. Test MSE was marginally higher for the social distancing and health factor models. Overall, the health risk factor model had the lowest MSE across both test and training sets. One note regarding error values is that our response variable, `ConfirmedPercent`, has an average value of 0.001017, so the magnitude of our MSE values should be viewed in the context of that scale.

The global model showed high error rates, so we were curious to see if we could create a more effective model that incorporated variables across different feature buckets. After experimenting with different subsets of features, we found that test and train set MSE were minimized for the following combination: `'Log3-YrDiabetes2015-17', 'HDMortalityPercent', 'Smokers_Percentage', 'MedianAge2010', '65+Percent'`.

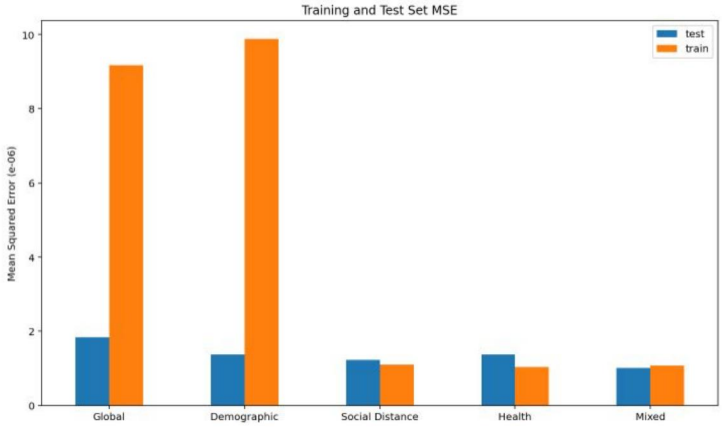

Figure 4: Training and Test Set MSE for Models

Regularization

Considering that models typically have higher performance on training sets than test sets, we wanted to further refine our model with cross-validation for regularization and hyperparameter tuning.

Our preliminary analysis using L1 regularization (lasso regression) indicated that the best alpha value is 0.001 and the cross validation error for optimal alpha is ~0.002995. The alpha value was always at the lower bound of the search range and the optimal (in terms of MSE) model under Lasso was that with all estimated coefficients equal to 0. This result is consistent with the characteristics of the Lasso model which tends to shrink coefficients to exactly 0. Even though the optimal Lasso model produced a lower test MSE than the unregularized model, it did not provide that much value as estimated coefficients of 0 failed to give us a better picture of the relationship between the response variable and features. Tuning the hyperparameter also yielded little difference in cross validation error which was largely constant as can be seen in Figure 5a below.

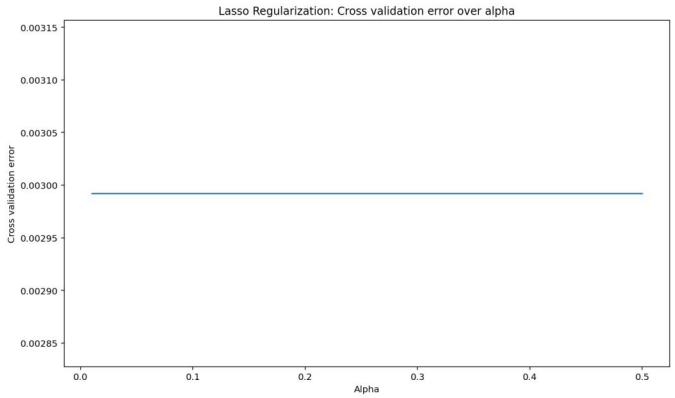

Figure 5a: Alpha vs. Cross Validation Error from Lasso Regularization

The L2 regularization (ridge regression) was also fit to the training data and its hyperparameter (alpha) was tuned using cross-validation. The tuning procedure resulted in an optimal alpha value that was always at the upper bound of the search range and estimated coefficients close to the unregularized full linear model for higher values of alpha.

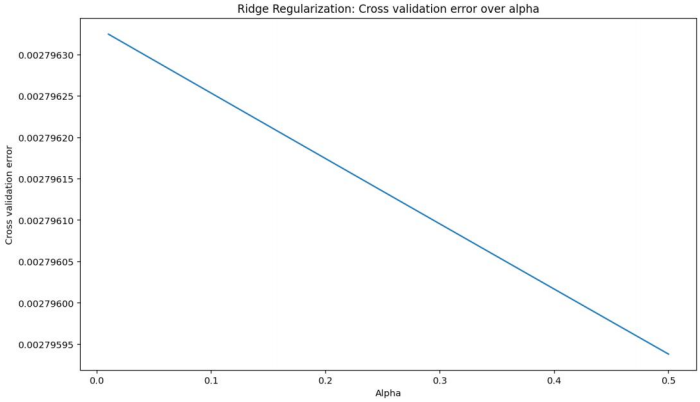

Figure 5b: Alpha vs. Cross Validation Error from Ridge Regularization

Regression Diagnostics

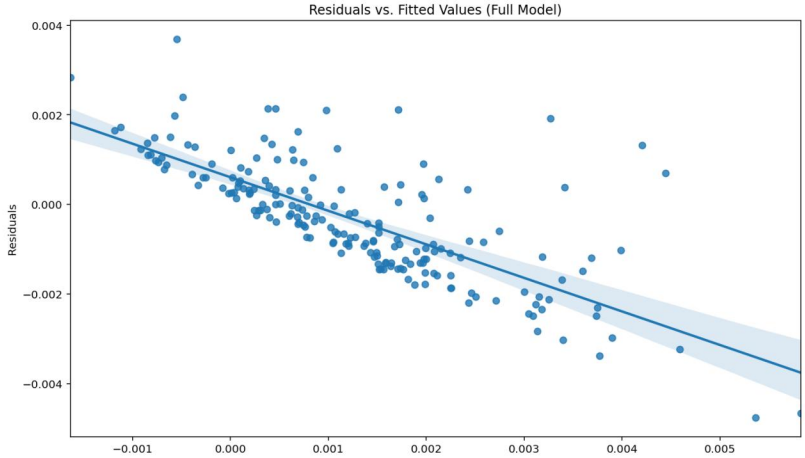

Figure 6: Residuals vs. Fitted values from full model

Our linear model relies on the assumption of homoscedasticity and errors with mean 0. Upon inspection of the residual plot shown in Figure 6, it seems that these assumptions are violated as the residuals are not centered around 0 and have non-constant variance across different fitted values. This indicates that a linear model may not adequately explain the relationship between the response variable and our features.

Summary of Results

We sought to explore the potential impact of certain demographic, health risk, and social distancing factors (county level) on the percent of confirmed COVID-19 cases. Through extensive EDA, feature engineering, and comparing models (OLS vs. Lasso vs. Ridge with different feature buckets and hyperparameters) via cross-validation errors, we came to several key conclusions:

- Correlation amongst feature buckets:
    - Within demographic features, SVIPercentile, a ranking demonstrating the county's social vulnerability, is negatively correlated with Medicare enrollment. Given that our health risk model obtained the lowest MSE across test and training sets, social vulnerability alongside Medicare or other health related features may be worth exploring further.
    - Many significant correlations amongst features in health risk groups suggest that these factors may also have a strong correlation with COVID-19 susceptibility, providing another avenue to explore.
- Non-significant relationships:
    - Our original pairwise plots between individual features, counts of confirmed cases, and death rates (our original response variable) demonstrate no distinguishable patterns or relationships despite many transformations.
- Model performance:
    - We determined that an Ordinary Least Squares model encompassing the following features was the most optimal in striking a balance between reducing MSE and being informative: `'Log3-YrDiabetes2015-17'`, `'HDMortalityPercent'`, `'Smokers_Percentage'`, `'MedianAge2010'`, `'65+Percent'`.
- Regularization:
    - Lasso Model: The optimal alpha values which minimized our MSE were consistently on the lower bound of our search range, resulting in extremely sparse coefficients and uninformative results.
    - Ridge Model: Resulting alpha values were always on the upper bound of our search range with estimated coefficients very similar to that of our unregularized, OLS model.
- Violated assumptions:
    - A close examination of our residual plots yields the conclusion that the assumption of homoscedasticity and errors with mean 0 have been violated, such that an alternative, non-linear model should probably be considered.
- Areas for further exploration:
    - Evaluating our models with further metrics including R^2 values would provide valuable insight on the significance of our conclusions.
    - Exploring more dates within the available time-series data may be beneficial in more accurately explaining growth patterns.
    - More granular data would not only improve the accuracy of our model but also make up for potential loss in information due to data aggregation.

Discussion

Notes on interesting and ineffective features

When experimenting with feature subsets in our mixed model, we found that the health factors `Log3-YrDiabetes2015-17`, `HDMortalityPercent` and `Smokers_Percentage` resulted in the lowest MSE. This suggests that comorbidities may be important to understanding COVID-19, and individuals with these health risk factors could be more susceptible. Additionally, we were surprised that certain demographic features were less effective than we initially hypothesized. For example, we intuitively believed that high population density would cause higher infection rates, but the data did not support this.

Data challenges

One challenge that we faced with the data was the sparsity of the data and the presence of 0 or missing values for our response variables and features. Many of these observations had to be excluded, providing us with a much smaller dataset than that which we started with. Additionally, it was difficult to identify any clear relationships between the feature variables and our response variable. As seen from the pairwise plots, the general shapes suggested fairly uniform distributions. Thus, it was hard to know if our features were truly relevant and meaningful variables in our model for COVID-19.

Limitations of analysis and assumptions

We used the most recent figures (April 18th, 2020) to calculate current confirmed cases for each county, but this may present some limitations. Some counties may be at different stages of contracting or containing the virus, and simply observing the point prevalence for one date does not capture overall growth patterns.

We are unable to fully assess data faithfulness, but it is likely that these figures do not capture the full reality of COVID-19 cases and deaths. This is primarily due to unavailability and incompleteness of testing across the US. Additionally, since the data is compiled from various sources, the time frames for the data are inconsistent. For example, median age statistics are from 2010 and Respiratory Mortality rates are from 2014. We assumed that these county-level metrics did not drastically change up to the present, but there may have been material differences that impact our model for COVID-19. Overall, we do believe our methods were reasonably sound in investigating general links between county-level attributes and COVID-19 risk.

Further analysis for expanded data

Having more complete data, particularly for our health related factors, would likely provide us with more accurate results on the relationship between county-level death rates and confirmed case rates.
More granular data, such as city-specific variables, might also help if some predictive information was lost when the variables were aggregated by county. Some factors such as the social distancing factors were noticed to only have state-level statistics which might lead our models' estimates of the impact of the factors on county-level rates to be inaccurate.

Ethical concerns

As COVID-19 is such an impactful and pressing issue during this time, policy decisions around the virus must be thought out thoroughly and hopefully backed by proper data analysis. This would require that our analysis is sound and the data is accurate. In regards to the data collection process, health related data inherently involve privacy issues as well as algorithmic biases which ultimately impact the outcome of our models and conclusions. Being unaware of these potential problems may lead us to not recognize what factors are actually affecting the spread of COVID-19 or wrongly underestimate the influence of other factors. We could address issues of privacy by encouraging data collectors to promote privacy impact assessments or establish simpler privacy policies. Algorithmic biases can be improved by establishing awareness amongst developers of the potential ethical implications of their work as well as involving domain experts during the construction process.

6 Composition 10 / 10

✓ **+ 10 pts Exemplary**

+ 6.66 pts Acceptable

+ 3.33 pts Inadequate

+ 0 pts Blank/Missing

- 5 pts Code is present in report

- 3.33 pts Code is split across multiple notebooks

+ 0 pts Missing citation(s) for additional dataset(s)