Predicting COVID-19 Deaths: A Survey of Machine Learning Approaches

John Greer jcgreer@uchicago.edu

Richard Chang changrc@uchicago.edu

Harris School of Public Policy,
University of Chicago,
Illinois, USA

Abstract

State and local health agencies are increasingly reliant on statistical measures to direct COVID-19 resource allocation towards populations and geographic areas. However, the predictive value of these indicators is inherently limited by their construction, which occurred before COVID-19 and is used to predict or evaluate conditions that may only be peripherally related to COVID-19.

We attempt to construct a COVID-19 death predictor utilizing the most granular national data publicly available, at the county level, using measures that combine socio-economic and health indicators. Various machine learning models are applied and evaluated. Our results indicate that ridge regression generally produces the lowest error rates although error rates vary drastically based on the random test-train split.

1 Introduction

As of December 12, 2020, the U.S. has recorded 15,851,014 COVID-19 cases and 295,522 deaths.[i] The scope and speed at which COVID-19 has spread across the United States has forced state and local government agencies to consider how best to direct resources to assist particularly vulnerable populations while reducing the infection rate. However, the lack of a coordinated strategy or detailed guidance from the federal government has resulted in state and local health agencies relying on pre-existing indicators to try and predict populations and geographic areas disproportionately impacted by COVID-19. The potential mismatch between the indicator's original goal and its ability to identify priority populations and geographic areas could result in inefficient allocation of resources and exacerbation of COVID-19 related health and socioeconomic disparities.

This paper will briefly review some of the indicators currently being utilized by health agencies to direct resources and guide COVID-19 shutdown policies as well as an examine the potential for features utilized by those indicators to predict COVID-19 deaths in the U.S., at the county level.

2 Related Work

In October, 2020, the California Department of Public Health (CDPH) launched a "health equity metric" that would determine whether counties with a population greater than 106,000 could move to a less restrictive tier that allows for greater economic activity. The metric utilizes the Healthy Places Index[ii], which was designed to inform health policies by selecting predictors of life expectancy from birth, assigning weights developed in part through machine learning processes, and scoring census tracts with a percentile rank. The features included in the Healthy Places Index include health, education, and economic features. The Healthy Places Index was selected by CDPH to identify census tracts in the bottom quartile and require counties to ensure that the case rate and test positivity of those census tracts do not lag its county.

The Center for Disease Control and Prevention's Social Vulnerability Index[iii] (SVI) is another indicator being used to direct COVID-19 health policy. The National Academies published a Framework for Equitable Allocation of Vaccine for the Novel Coronavirus in October 2020, which utilized the SVI in its method for prioritizing geographic areas for vaccine allocation. The SVI rates the social vulnerability of geographic areas based on numerous economic, demographic, and residential metrics.

3 Approach

3.1 Data Preparation

The features utilized were sourced from the Census Bureau's 2014-2018 American Community Survey and Robert Wood Johnson Foundation's County Health Rankings. The target labels, COVID-19 related deaths, came from the Centers for Disease Control and Prevention's Provisional COVID-19 Death Counts in the United States by County.

The Census features were obtained using the Python CensusData 1.10 package[iv], which accessed the Census data through the Census Bureau's API, and the Census Bureau's Explore Census Data site.[v] The features included county population estimates, economic indicators, higher education attainment, health insurance, vehicle availability, and household density. Features were generated through linear operations, such as estimating minority population by subtracting the White Alone Non-Hispanic population from the total population, and household density by dividing the number of people in a household by the number of rooms in the household.

County Health Rankings data was downloaded from their website[vi]; this was the primary source of county-level health data. The health-related features included conditions that have been associated with COVID-19, including estimated percentages of residents that were smokers, obese, in poor health, and were physically inactive. It also included more generalized health statistics such as years of potential life lost, mental health, a food environment index, and alcohol use.

The Census and County Health Rankings data were prepared by eliminating features relating to margins of error and rank indices that were not expected to inform our models. The data sets were then joined on counties.

The COVID-19 death data was obtained from the CDC's National Center for Health Statistics website. This provided us with target labels for 1,571 counties. The counties included in the data set were limited to those that reported 10 or more COVID-19 deaths.

An inner join was then used to combine the features with COVID-19 deaths of respective counties while dropping counties that did not appear in the CDC data set. Features with greater than 5 missing entries were then dropped from the combined data set. Counties were then removed if they had missing features. This resulted in a reduction of 20 observations, leaving a total of 1,396 observations representing counties, 67 features, and the target label.

3.2 Models

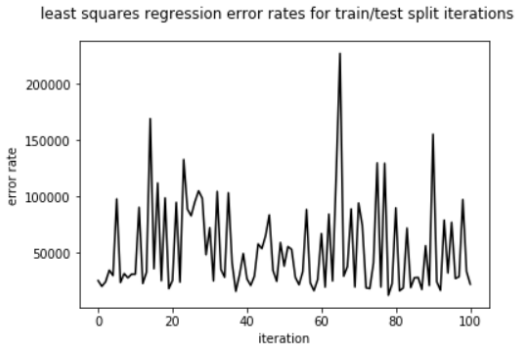
The models utilized included least squares linear regression and ridge regression. All the functions were coded from scratch, including the machine learning models and cross-validation implementation using NumPy and Pandas (code available upon request). The models were implemented by withholding a random selection of twenty percent of the observations for the test set. K-folds cross-validation was then utilized to train the model and determine the best weights based on the lowest mean squared error. The ridge regression implementation included a parameter search for a user-provided list of lambda values to search for the weights that resulted in the lowest mean squared error. Once the associated weights were calculated, they were applied to the holdout test data set that was initially separated from the training data set.

We also examined the potential for a model that utilized all the counties outside of California as a training data set and California's counties as the test data set. Cross-validation was then applied to the training data set, using each state's counties as the basis for splitting the training data. The weights with the lowest mean squared error were then applied to California's counties.

4 Results

Least Squares Linear Regression

The error rates were calculated by first iterating over a test-train split with a ratio of 0.75 of the data set used for training and the remainder as the test holdout. The k-fold cross-validation process was then applied to the training data set, iterating over k-folds from 3 to 10. The best error rate for each test-train split was then saved to a list. Over 100 iterations of test-train splits, the error rate varied significantly between 20,000 and 250,000. The median error rate was 33456.95 and the mean error rate was 53200.84.

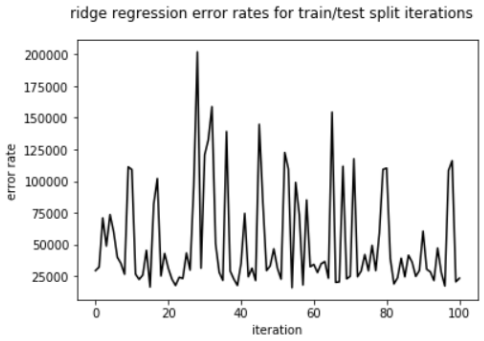


The weight vector calculated by the last train-test iteration indicates that the most significant predictor of COVID-19 deaths appears to be the county's total population.

Ridge Regression

A ridge regression model was applied with a parameter search function that examined the impact of lambda values between 0 and 1 on error rates. The result was a slight improvement in error rates over the least squares regression error rates, with the best ridge regression error rate being 22038.56.

However, this error rate may not be representative given the sensitivity of error rates to specific test-train sample selection. To determine this, a similar method to least squares linear regression was also applied in terms of iterating over test-train splits. K-fold = 5 was utilized and the parameter lambda value was set to 0.8. The best error rates were saved for each test-train iteration. There appeared to be similar variation to the least squares regression error rates. The median ridge regression error rate was 32612.77 and the mean ridge regression error rate was 52551.86.



Polynomial Expansion

To evaluate whether the relationship between the data might be better approximated by approaches other than strict linear regression, we also applied a polynomial feature expansion of the data before running least squares linear and ridge regression. The only degree tested was 2. The resulting error rates were significantly larger than those derived from the other approaches, so this approach was not pursued further.

Cross-Validation by State

Counties were also grouped by state and divided into a holdout test data set consisting of all California counties and a training data set consisting of remaining counties. Remaining counties would then be split by state with cross-validation being utilized with individual states as holdouts from cross-validation purposes. The weights associated with the lowest error rate were then applied to California county features to predict the COVID-19 death rate of California counties. The results were generally significantly high error rates, with the lowest error rate being 11395054708480.857. This strongly indicates that utilizing all states to predict COVID-19 death rates in California would not lead to accurate results and that a potential path for improvement would be to limit the training data set to states with similar demographics and distribution of features.

Limitations of Approach

This project was limited by use of COVID-19 county-level death counts as a predictor. This required the use of features that were also available at the same geographic level. The number of observations was further reduced by the CDC's suppression rule which did not report death counts for counties with fewer than 10 deaths. This resulted in observations for only 1,396 counties out of 3,007 in the U.S. being used. Race and ethnic population estimates for specific groups was also not used which may have increased the accuracy of the models, given that African American, Latino, and Native Hawaiian and Pacific Islander populations have experienced disproportionately high rates of age-adjusted COVID-deaths.

Avenues worth exploring in similar projects in the future include more geographically granular COVID-19 data, such as the ZCTA or census tract level, should it become available.

[i] COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), https://coronavirus.jhu.edu/map.html, last accessed 12/12/20.

[ii] The California Healthy Places Index, https://healthyplacesindex.org/, last accessed 12/12/20.

[iii] Center for Disease Control Social Vulnerability Index, https://www.atsdr.cdc.gov/placeandhealth/svi/index.html, last accessed 12/12/20.

[iv] CensusData 1.10, https://pypi.org/project/CensusData/, last accessed 12/12/20.

[v] Census Bureau, Explore Census Data, https://data.census.gov/cedsci/, last accessed 12/12/20.

[vi] https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation