S&DS 563/S&DS 363 Final Project

Evan Collins, Kelly Farley, Ken Stier

15 May 2021

Introduction

With the nadir of the SARS-CoV-2 pandemic finally subsiding, the task of understanding the epidemiological factors contributing to the propagation of COVID-19 has only begun. Understanding relationships between COVID-19 spread prevention and socioeconomic variables will prove vital to inform us of how to mitigate the propagation of the next pandemic. This report aims to understand how the economic, education, behavioral, and population data for the 3141 U.S. counties relate to COVID-19 case data.

We acknowledge that the impact of and response to COVID-19 has been very different from county to county. Looking at the current COVID-19 vaccination data in mid-May 2020, we note that the vaccination rate for ages 18+ ranges from 11% in some counties in Lousiana to 74% in some counties in New York. Our guiding question is: Who is the most vulnerable to COVID-19 infection and death? This knowledge will guide public health efforts as we continue to fight against the spread of COVID-19. Knowledge of what socioeconomic factors put people at risk will allow us to prioritize our vaccination and education efforts from those who need it the most and will also let us take a step back to acknowledge the systemic health inequalities in our country.

Design and Primary Questions

This report deploys three multivariate techniques to examine three questions in particular:

1. How do the 3141 counties differ from one another, i.e., how do the socioeconomic and COVID-19 data relate to one another when distinguishing U.S. counties? Principal components analysis (PCA) will help to reduce the dimensionality of our large dataset, increasing interpretability of underlying trends between clusters of variables. This metric technique works on the "columns" of our data set to reduce them into composite variables and make them more interpretable.
2. Which U.S. counties are similar to one another? Cluster analysis will enable the clustering of selected counties into a discrete number of groups based on similar socioeconomic and COVID-19 data. This metric technique works on the "rows" of our data set to find similar groups of observations.
3. Which U.S. county variable pairings are similar to one another? Correspondence analysis is similar to PCA but applies to categorical rather than continoous variables. This nonmetric technique works on both the "columns" and "rows" of our data set to visualize which rows and column points are similar in lower-dimensional space.

Using these techniques, we will be able to better understand our variables, our observations, and the interactions between our variables and observations. Who is most vulnerable to COVID-19 infection and death? This allows us to direct resources to protecting these vulnerable populations.

Data

The dataset referenced in this report includes COVID-19 infection and death statistics from U.S. counties (sourced from Johns Hopkins, as of 28 April 2021), combined with economic, education, and population data (sourced from various government agencies) and also survey responses about mask-wearing frequencies (sourced from NYT).

3141 complete observations on 10 continuous variables and 6 categorical variables. Continuous variables were rescaled as percentages of county population.

- **6 categorical variables**: FIPS, county name, state name, rural urban type, rural urban code, economic typology
- **9 continuous variables**: "Always" wear mask survey response percent, unemployment rate, median household income, percent poverty, percent of adults with less than a high school education, death rate, percent civilian labor force, percent of county population that has had confirmed COVID-19 cases, and percent of county population that has died from COVID-19.

[1] "FIPS" = State-County FIPS Code; Categorical (identifier)

[2] "County_Name" = US County Name; Categorical (identifier)

[3] "State_Name" = US State Name; Categorical

[4] "Rural_Urban_Type" = Regrouping of Rural-Urban Codes (2013) numbered 1-9 according to descriptions provided by the USDA. See variable [5]. Regroup codes 1 through 9 into three groups: (1) "Urban" for codes 1-3, (2) "Suburban" for codes 4-6, and (3) "Rural" for codes 7-9; Categorical (1-3)

[5] "Rural_Urban_Code_2013" = Rural-urban Continuum Code, 2013; (https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/); Categorical (1-9)

[6] "Economic_Typology_2015" = County economic types, 2015 edition (https://www.ers.usda.gov/data-products/county-typology-codes/); Non-overlapping economic-dependence county indicator. 0=Nonspecialized 1=Farm-dependent 2=Mining-dependent 3=Manufacturing-dependent 4=Federal/State government-dependent 5=Recreation; Categorical (0-5)

[7] "Always_Wear_Mask_Survey" = "Always" response. The New York Times administered a survey to 250,000 Americans from July 2 to July 14 asking the following question: How often do you wear a mask in public when you expect to be within six feet of another person?; Continous (%)

[8] "Unemployment_Rate_2019" = Unemployment rate, 2019; Continuous (%)

[9] "Median_Household_Income_2019" = Estimate of median household Income, 2019; Continous ($)

[10] "Percent_Poverty_2019" = Estimate of people of all ages in poverty 2019; Continuous (%)

[11] "Percent_Adults_Less_Than_HS" = Percent of adults with less than a high school diploma, 2014-18

[12] "Death_Rate_2019" = Death rate in period 7/1/2018 to 6/30/2019; Continuous (%)

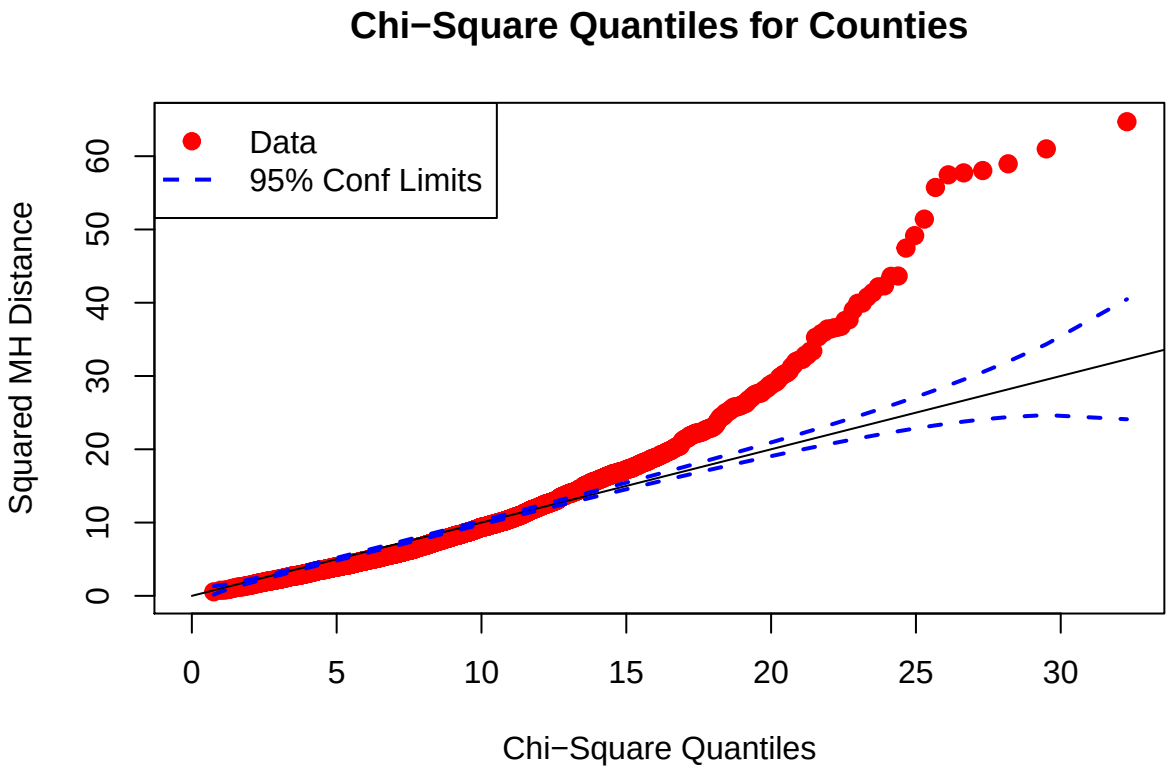
[13] "Civilian_Labor_Force_2019_as_pct" = Civilian labor force annual average, 2019, expressed as percent; Continuous (%)

[14] "Covid_Confirmed_Cases_as_pct" = Cumulative sum of COVID-19 cases expressed as percent. Reported from Johns Hopkins on 28 April 2021; Continuous (%)

[15] "Covid_Deaths_as_pct" = Cumulative sum of COVID-19 deaths expressed as percent. Reported from Johns Hopkins on 28 April 2021; Continuous (%)

Descriptive Plots

We made normal quantile plots for each of the 9 continuous variables in the dataset. This revealed that most variables initially did not have a univariate normal distribution. Taking the log-transform of the 10 continuous variables helped most variables have more linear quantile plots. Note that we also standardized the continuous variables since they were measured on different scales. Moreover, for death rate, percent COVID-19 cases, percent COVID-19 deaths, a 1.5 x IQR outlier exclusion method was applied to enable these variables to take on more normal univariate distributions. Note that the outlier exclusion method reduced the number of counties that we will analyze to 2,814 observation. Hence, this outlier exclusion method reduced the dataset by approximately 10%. This is somewhat high; however, we deemed that the benefits of having univariate and multivariate distributions to outweigh this disadvantage. We these changes made, the 9 continuous variables all had univariate normal distributions.



A chi-square quantile plot (shown above) reflects that our data does not have a multivariate normal distribution.

WE NEED TO ASK PROF ABOUT IF WE ABSOLUTELY NEED TO HAVE MULTIVARIATE NORMALITY.

Summary Statistics

Principle Components Analysis

Cluster Analysis

Correspondence Analysis

Conclusions and Discussion

Points for Further Analysis