Analysis Using Open Data

1.

The COVID-19 epidemic has affected people’s lives around the world. In addition to bringing inconvenience to people’s live, work, and study, it caused numerous deaths! With this concern, I want to explore the factors that contribute to the case fatality rate in the US. So, I started browsing the databases. I found that ACS provided most of the demographic data I needed. And then, I can get data of COVID-19 cases and deaths from the CDC website.

In order to get a good picture of the overall situation in the United States and cross-compare the data from different places, I decided to use the data at the state level. Due to a large amount of COVID-19 related data and the frequent changes in its statistical methods during the epidemic, I used the cross-sectional accumulated data of the epidemic on April 15, 2021.

I then reviewed relevant papers and reports and selected some demographic variables shown by some studies to be likely associated with COVID-19. The combined dataset includes the total population, the population over 85 (John Ng, 2020), the black population (Gould & Wilson, 2020), the population over 16 working, the population over 16 using public transportation to get to work, and the median household income (Takagi H, 2021). Through normalization, these variables are converted into ratios.

2.

Metadata is something that describes data, organizes information, and makes it easier to find and use. From libraries in the past to email, phone communications, social media accounts, web pages, and more, metadata is all around daily life.

In the ACS and CDC databases I used, they both generated statistical metadata through collection and processing. Then they use descriptive metadata to describe the basic information of these objects and publish it on the Internet for people to find and use.

This organized and publicly published statistical metadata is much more convenient and practical than the data sources I used in the past. In the past, I’ve used data sets that have titles and keywords. However, the ACS and CDC websites further set up interactive pages to select objects to generate tables and maps. It is helpful for users to see the data immediately and helps to organize the following information.

My work includes collecting information, combining and cleaning table, recording sources, describing information. Since it is difficult to find all the information at once, this process must be repeated many times. But organizing the data can help me record and find my own data and enable others to understand and trust my dataset.

3.

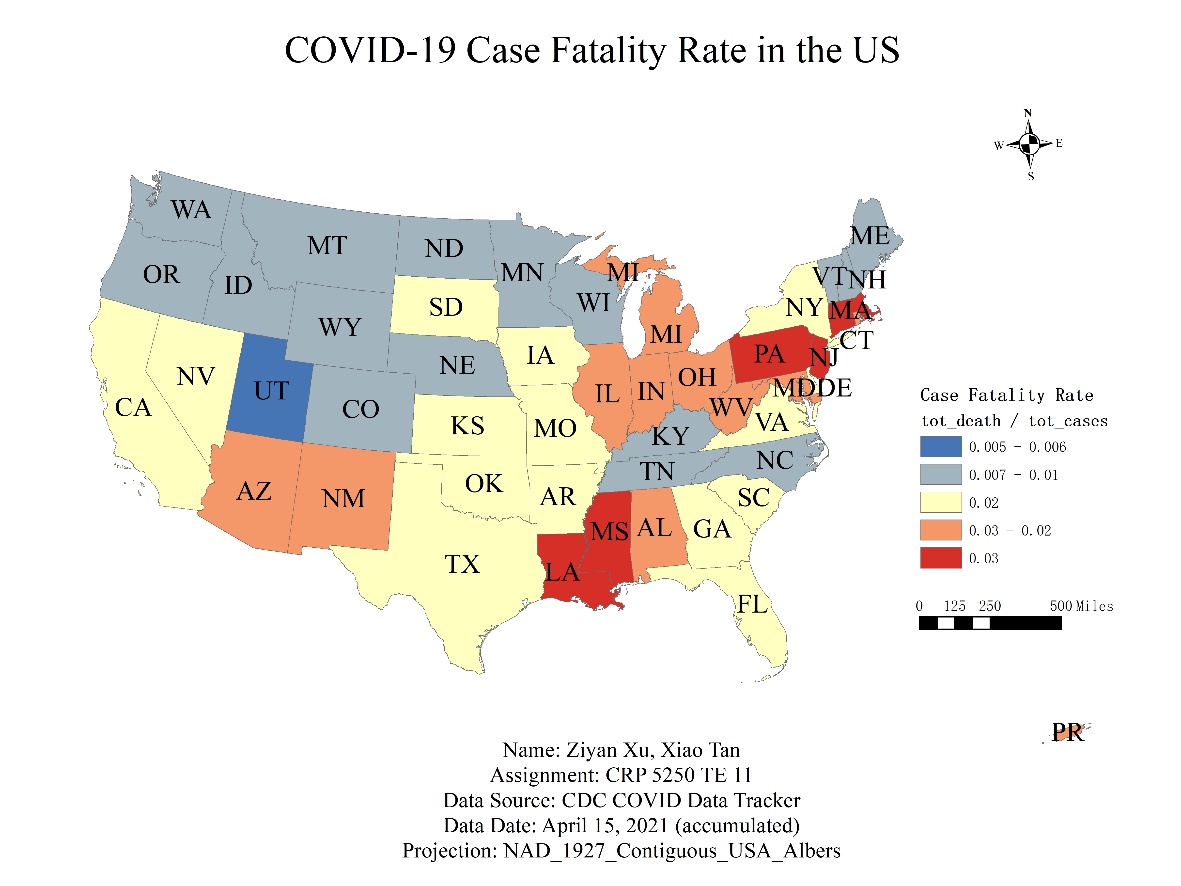


Figure 1 COVID-19 Case Fatality Rate in the US

According to the COVID-19 Case Fatality Rate Map, the mortality rates of people infected with COVID-19 are very different in different states. Overall, death rates were lowest in the northern parts of the West and highest in some states in the Northeast and South.

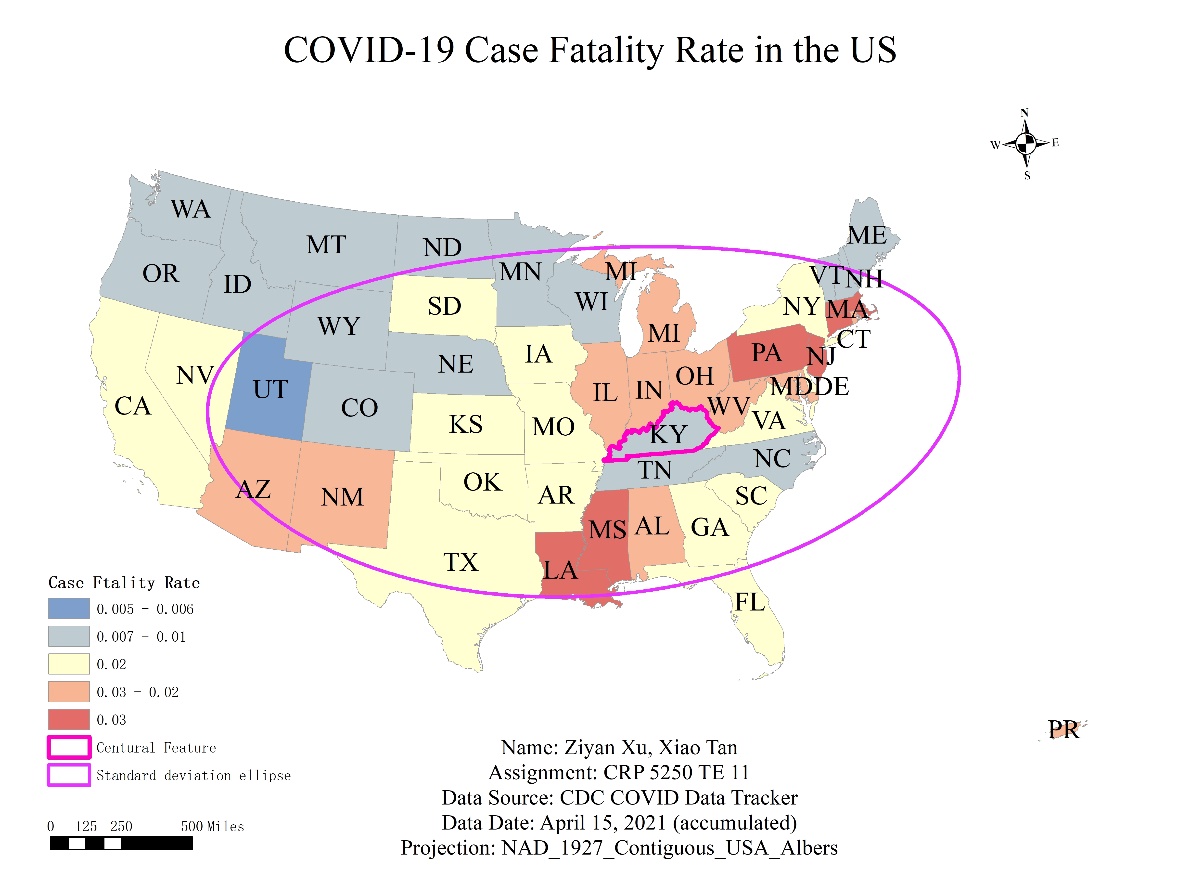
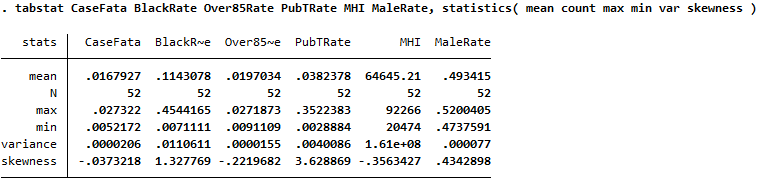


Figure 2 Spatial Feature of COVID-19 Case Fatality Rate in the US

Using spatial statistic methods, the standard deviation ellipse of the case mortality rate is long in the east-west direction, and the central feature is in Kentucky.

What are factors that contribute to different case fatality rates (CaseFata) in these states? With this question in mind, I read some articles and extract some possible variables: male population rate (MaleRate), aged over 85 years old population rate (Over85Rate), black or African American population rate (BlackRate), workers using public transportation go to work rate (PubTRate), median household income (MHI).

Table 1 Variable Statistics



Both BlackRate and PubTRate have a relatively significant right skewness.

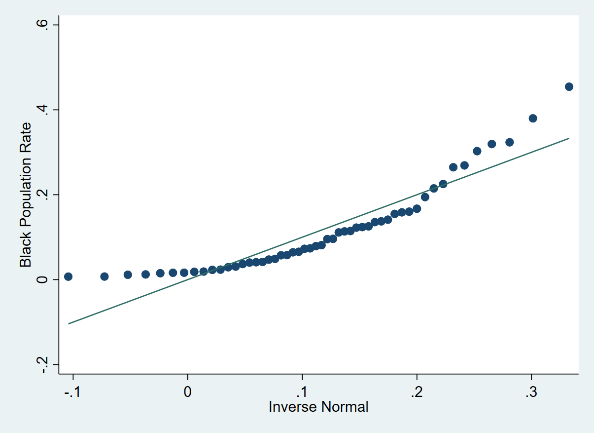
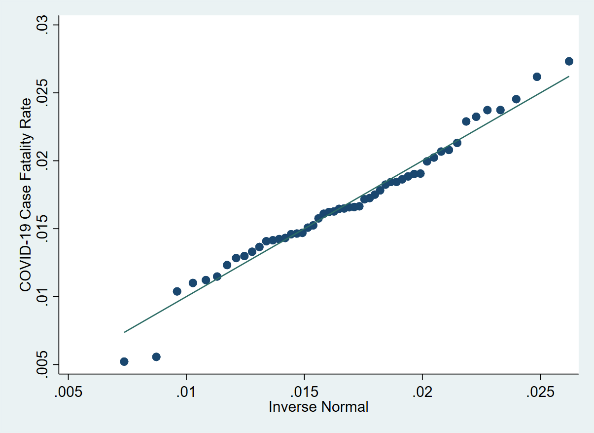


Figure 3 Inverse Normal of COVID-19 Case Fatality Rate. Figure 4 Inverse Normal of Black Population Rate

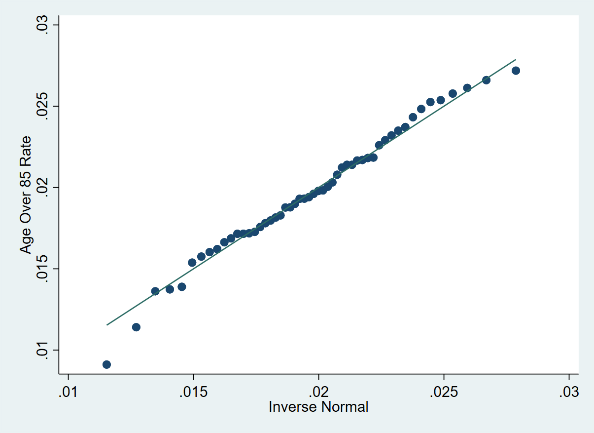
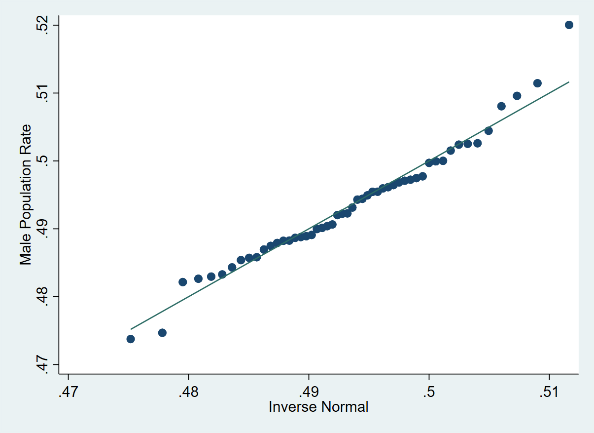


Figure 5 Inverse Normal of Male Population Rate. Figure 6 Inverse Normal of Age Over 85 Population Rate

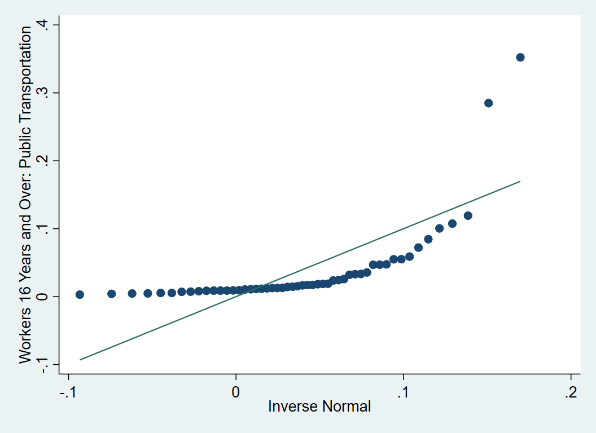
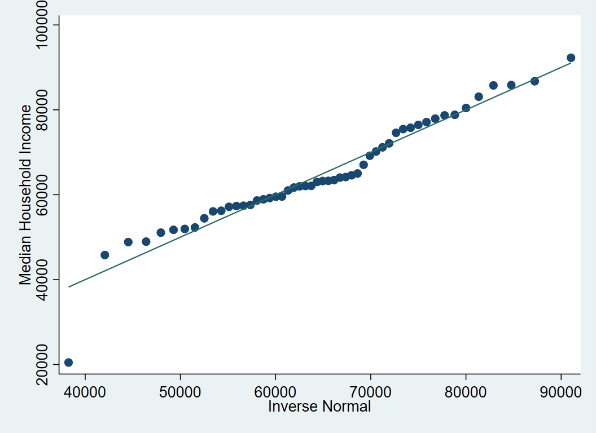
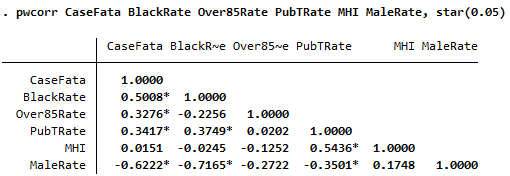


Figure 7 Inverse Normal of Median Household Income. Figure 8 Inverse Normal of Workers take Public Transportation to work

As can be seen from the normal distribution of these six variables, the CaseFata, MHI, and Over85Rate are close to the standard inverse line, while BlackRate and MaleRate have a few outliers. The PubTRate is not a normal distribution at all. Therefore, **PubTRate is not suitable to be an explanatory variable**.

Table 2 Correlation matrix



According to the correlation matrix, CaseFata has a significant relation to BlackRate, Over85Rate, PubTRate, MaleRate. However, it does not have a solid relation to MHI. Therefore, **MHI can be precluded from the model**.

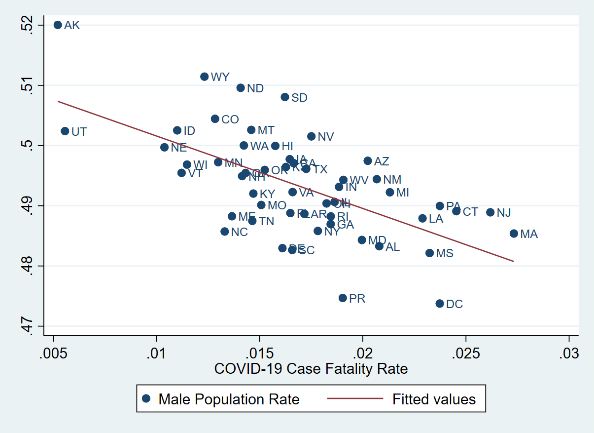
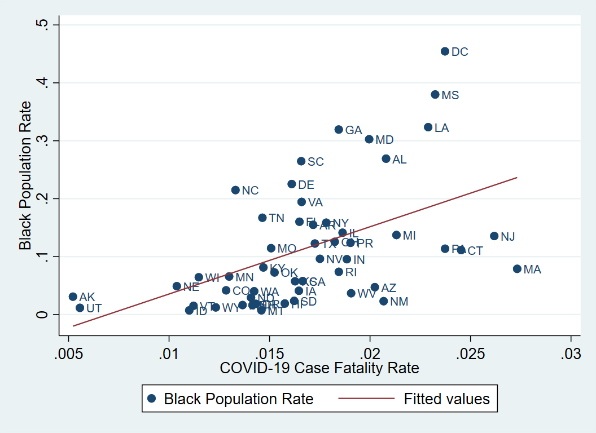


Figure 9 Black Population Rate/Case Fatality Rate. Figure 10 Male Population Rate/Case Fatality Rate

According to the scatter plot, CaseFata is positively related to BlackRate and negatively related to MaleRate. DC is the most significant outlier that has the highest black population rate.

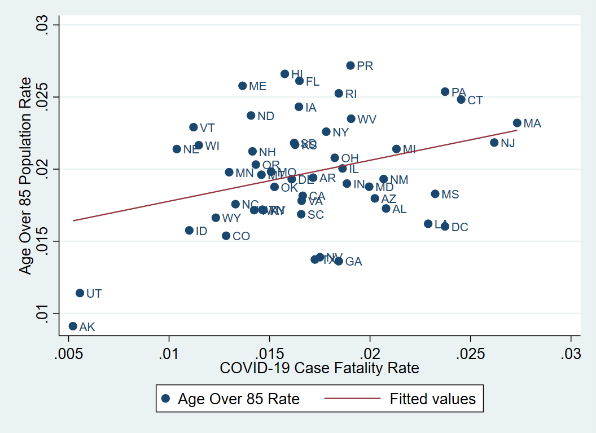
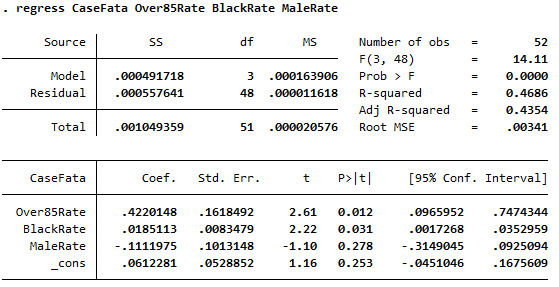


Figure 11 Age Over 85 Rate/Case Fatality Rate.

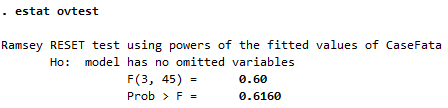
There may be a positive relation between CaseFata and Over85Rate. MA state has the biggest case fatality rate and PR state has the biggest percentage of age over 85 population.

Table 3 Regression Model



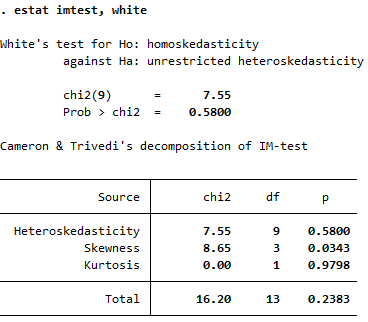
In this new model, the Prob>F is good enough, and the R square is 0.46. It means this model can explain 46% of the real data. The P>|t| value for Over85Rate and BlackRate are smaller than 0.05. So, they are relating to CaseFata. However, the P>|t| value for MaleRate is 0.278, bigger than 0.05. So, the MaleRate is not significantly relates to CaseFata.

Table 4 Ramsey Reset Test



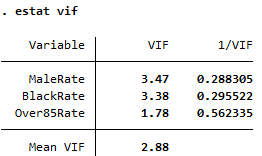
Check for the omitted variable. The Prob>F = 0.4 is too big to reject the Ho, so the model has no omitted variables.

Table 5 White’s Test



Check for homoskedasticity. The Prob>F = 0.16 is too big to reject the Ho, so there is no homoskedasticity.

Table 6 VIF Test



Check for VIF. All the VIF of variables are lower than 5, so there is no risk of multicollinearity.

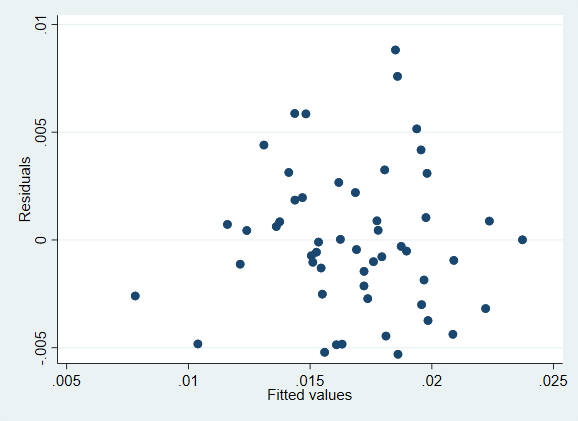


Figure 12 Residuals vs. Fitted Values

The residual vs. fitted values graph shows there is no relationship between them. So, the model is good.

In all, the Black population rate and age over 85 population rates respectively correlate with the COVID-19 case fatality in the United States. While the workers using public transportation rate, median household income, and male population rate are not relate to it.

# References

Gould, E., & Wilson, V. (2020). *Black workers face two of the most lethal preexisting conditions for coronavirus—racism and economic inequality.* Economic Policy Institute.

John Ng, K. B. (2020). *COVID-19 Mortality by Age, Gender, Ethnicity, Obesity, and Other Risk Factors: A Comparison Against All-Cause Mortality.* RGA.

Ladan Golestaneha, J. N. (2020). The association of race and COVID-19 mortality. *EClinicalMedicine*. doi:https://doi.org/10.1016/j.eclinm.2020.100455

Takagi H, K. T. (2021). Ethnicity/race and economics in COVID-19: meta-regression of data from counties in the New York metropolitan area. *Epidemiol Community Health*, 205-206.

Uma V. Mahajan, M. L.-P. (2020). Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties. *Journal of Public Health*, 445–447.

[Table 1 Variable Statistics 3](#_Toc70855837)

[Table 2 Correlation matrix 4](#_Toc70855838)

[Table 3 Regression Model 5](#_Toc70855839)

[Table 4 Ramsey Reset Test 5](#_Toc70855840)

[Table 5 White’s Test 6](#_Toc70855841)

[Table 6 VIF Test 6](#_Toc70855842)