STAT479
Final Project Report
Group member: Yuzhuo Kang, Mingke Wang, Zhiheng Shao

Summary:

Our group interest in the relationship between the age and death rate of COVID-19. The first section of this project introduces the background, motivation, the research question, and possible hypotheses. The second section, method, discusses the contents of our shiny app on the COVID-19 information visualization. Next, we talk about data sources and data cleaning process. In the last section, multiple linear regression model and a series of plots is applied to test whether the proportion of older people is associated with the death rate of COVID-19 in each state. We find that the association is not strong and not linear as we think. However, we do find that the share of elders aged 65+ among adults at risk have a strong and linear correlation with the death rate.

1.Introduction

Coronavirus disease affects everywhere and everyone. The death rate of COVID-19 varies across the states in the U.S. Death rate is the proportion of total death cases among all the positive diagnosis cases. The variations on death rates among states in the U.S associate with many factors, and we are interested in the relationship between age and death rates. The motivation is that this study between the age distribution and death rates could help the local government enact medical plans on some age groups which have a higher risk of being infected by COVID-19 and become dead. Another motivation is that the result could allow the public to know which age group has a higher death rate, so people could better protect themselves.

The research question is if the proportion of individuals 65 or older is associated with death rates in the U.S. There are several possible hypotheses. The first one is that immune functions decline with age, so older people are susceptible to diseases and have higher risks of dying. The second is that many older people do not have medical insurance, so they cannot afford to get treatments. While, there is another possibility that death rates correlate with the proportion of young adults. Compared to older people, they go out more often and have more chances to be infected.

By analyzing and visualizing the COVID-19 and age data, **we hypothesize that when controlling for other variables, the proportion of older people (65 years +) in each state correlates with the death rates of COVID-19**.

2.Methods

2.1 Shiny app for COVID-19 Visualization

We first create shiny app to visualize various information of Covid-19 in each state using COVID-19 tracking API and the American Community Survey (ACS) data. In this app, viewers can select any state or states that they wish to examine and compare for each topic. There are eleven topics that viewers can choose from: Cumulative and daily positive diagnosis rate, Cumulative and daily Death number, Total Death rate, Total recovered, Total positive and negative case, Current hospitalized, Current in ICU, and Current on ventilator. Viewers can choose the scale from standard to log to look more closely. Users can also choose the time ranges. In another tab, it will show the proportion of older people (65+) in the selected states. This tab also shows the U.S map that gives an intuitive comparison of the proportion of older people in each state.

2.2 Data sources and Data cleaning process

The URL link of COVID-19 tracking API is https://covidtracking.com/api/v1/states/daily.json. It collects the latest data for U.S states. It contains many relevant COVID-19 statistics such as the total positive, recovered and death cases. The data is from public health authorities and trusted news reporting.

state	date	positive	negative	death	total	hospitalizedCurrently	inIcuCurrently	onVentilatorCurrently	recovered
AK	20200505	371	22321	9	22692	13	*NA*	*NA*	277
AL	20200505	8285	98481	313	106766	*NA*	*NA*	*NA*	*NA*
AR	20200505	3496	51139	80	54635	89	*NA*	16	2041
AS	20200505	0	83	0	83	*NA*	*NA*	*NA*	*NA*
AZ	20200505	9305	78955	395	88260	728	303	185	1671
CA	20200505	56212	723690	2317	779902	4622	1388	*NA*	*NA*

The death rate and cumulative positive diagnosis rate of COVID-19 is calculated using this COVID-19 tracking API. Positive diagnosis rate is the total number of tests with positive results divided by the total number of tests. In order to test the relationship between the death rates and the proportion of older people, the observations of US Virgin Islands, Northern Mariana Islands, American Samoa, Guam, and Puerto Rico are dropped since the age data does not contain information for these regions. Also, the death rate and positive diagnosis rate on May 1st are used in the regression model because it does not contain missing values. The death rate and positive diagnosis rate are multiplied by 100%.

state_name	date	death_rate	cum_pos_rate
Alaska	2020-05-01	2.4725275	1.790898
Alabama	2020-05-01	3.8977368	7.786105
Arkansas	2020-05-01	1.9271304	6.685321
Arizona	2020-05-01	4.1446873	10.633155
California	2020-05-01	4.1096705	7.701245
Colorado	2020-05-01	5.0837477	21.113413

The data on the proportion of older people (65+), old, in every state uses ACS data https://usa.ipums.org/usa/index.shtml. ACS is an individual-level data, and the sample size is 3214539. The proportion of older people is calculated by grouping the individuals in the sample

by states. The advantage of ACS is that the size and scope are large, so the sample is representative.

STATEFIP	old
1	20.26079
2	12.60617
4	20.17431
5	19.58496
6	16.34404
8	16.17973

We also use other state-level data relevant to death rates. The data on the number of ICU beds per 10,000 population, hospbed, and the share (measured by 100%) of adults aged 65+ among all adults at risk for each state, risk, is from KFF, https://www.kff.org/health-costs/issue-brief/state-data-and-policy-actions-to-address-coronavirus/. It provides authorized data on national health statistics.

state_name	hospbed	risk
Alabama	3.9	51.0
Alaska	1.8	49.4
Arizona	2.5	59.1
Arkansas	2.9	50.3
California	2.1	56.0
Colorado	3.2	59.1

We also have data on state-level demographic characteristics, including the proportion of younger adults (aged 16-25), age25, the share of residents living in metropolitan areas, metrocity, and the percentage of residents without health insurance coverage, ins, in every state from the ACS https://usa.ipums.org/usa/index.shtml. These three variables are calculated by grouping the individuals in the sample by states. The proportion of residents in metropolitan areas in Wyoming is 0, so this observation is set as missing. Other two variables do not have missing or extreme values.

STATEFIP	metrocity	age25	ins
1	56.560688	29.06001	9.159219
2	20.995381	35.92609	13.738638
4	88.169739	30.88727	10.172882
5	39.874766	31.22316	7.609088
6	97.577986	30.92021	6.480438
8	83.745172	30.24424	6.474396

2.3 Statistics techniques and reason

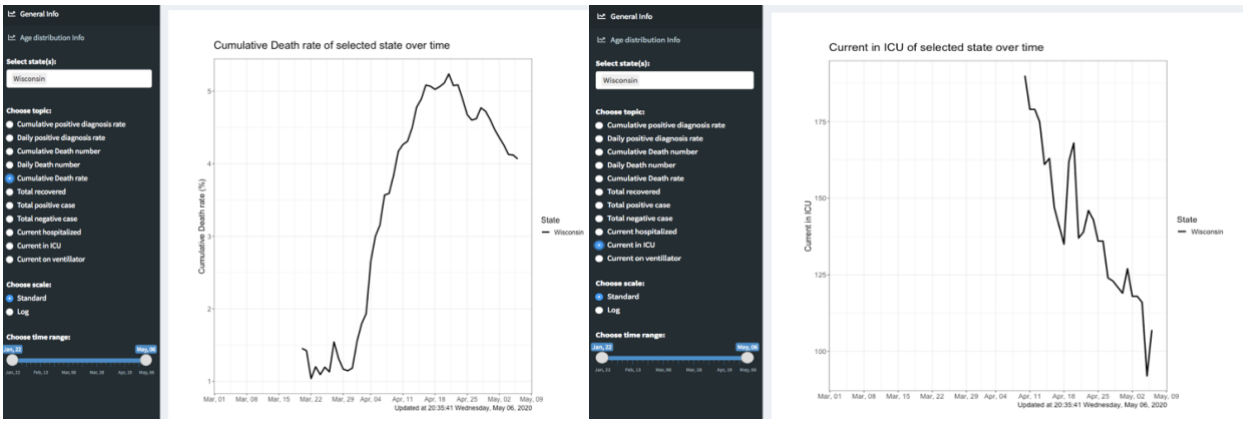
In order to show the relationship between the death rates and the proportion of older people age 65+, the multiple linear regression model is applied to test the significance of the age. The positive diagnosis rate, ICU beds per 10,000 population, share of older adults at risk, proportion of younger adults, proportion in metropolitan areas, and the percentage without health insurance are incorporated in the model as additional control variables. Share of adults aged 65+ among all the adults at risk may be a confounding variable because it correlates with the proportion of older people and relates with the death rate at the same time. The inclusion of control and confounding variables in analysis is an effective way to reduce the omitted variable bias by preventing spurious effects, so the estimation could reflect the true relationship between the age and death rates. When they are not included, the estimation is not reliable because the omitted variables have strong correlations with age. The model diagnostics is conducted to test the model and outliers.

In addition, a series of scatterplots with best fitted lines are adopted to visualize the relationships between the age distribution and the death rates in each state. The reason is to show whether there exist linear correlations between the variables.

3.Results

It displays some information on the COVID-19 shiny app. Take Wisconsin as an example. The proportion of older people is 19%. For the cumulative positive diagnosis rate, there is a peak at around March.15th. The positive rate keeps decreasing until around March.23th, and the rate increases again. For the cumulative death rate in Wisconsin, it has a peak in April.20th and is decreasing. For patients who are currently in ICU, the number of patients is decreasing.

Figure1. Shiny App of COVID-19



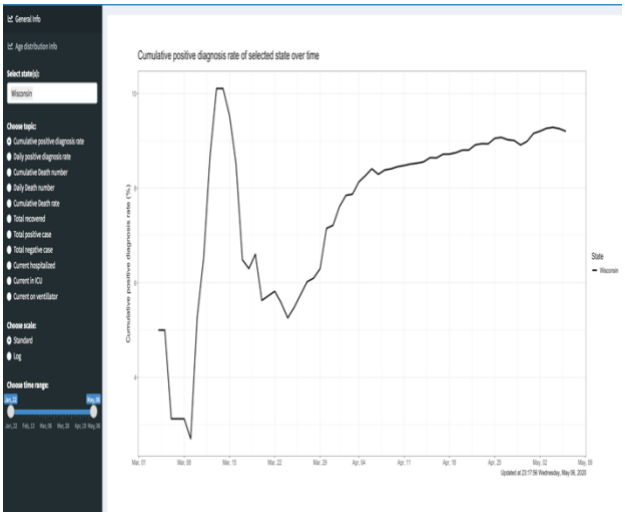
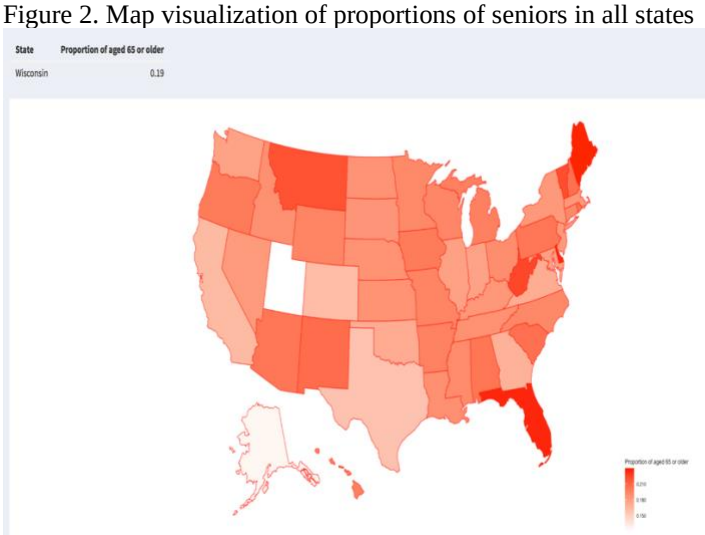


Figure 2. Map visualization of proportions of seniors in all states



The average death rates for all the 51 states until May 1st is 4.16%, and the average proportions of seniors is 18.88%. The standard deviation of death rates and proportions of seniors are 1.74% and 2.39%.

The regression result is shown in the Table 1. Positive diagnosis rate and the proportion of aged 65+ among all the adults at risk are significant predictors on the death rates because the p-values are less than 0.05. The proportion of older people aged 65+ and younger adults also have considerable relationships with the death rates since their p-values are less than 0.1. We find the death rate declines when the proportion of older people increases. The death rate also has negative association with the proportion of younger adults. Positive diagnosis rate and share of aged 65+ adults at risk have positive correlations with death rates.

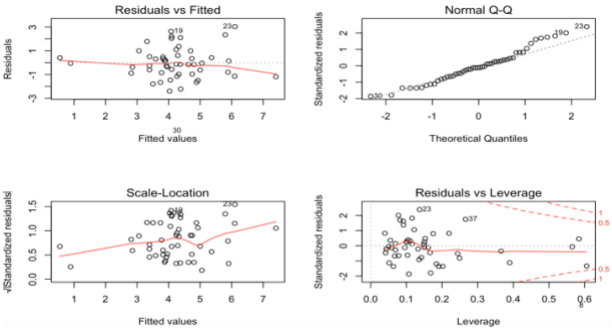
According to the diagnostic plot, the residuals have non-linear patterns, and the residuals are normally distributed with similar variance. There are influential points such as #23, Michigan. The death rate is 9.13%, and proportion of seniors is at average level. This is a case with

low proportion of seniors and high death rates. One plausible explanation is Michigan suffers its proximity to nation's COVID-19 hot spot. Another explanation is that Michigan does not have significant social distancing measures until April. While Michigan has an average population of seniors, it is a densely populated state.

Table 1: Regression Result of model

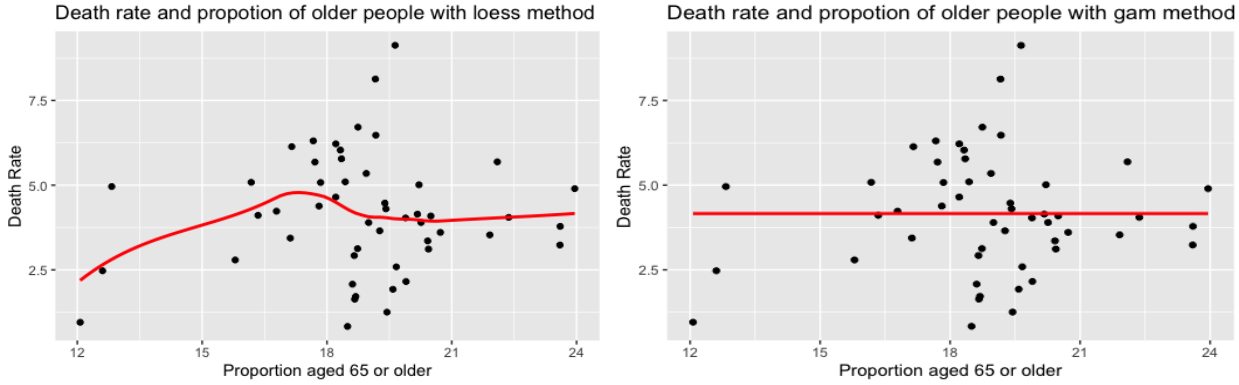
```
Call:
lm(formula = death_rate ~ old + cum_pos_rate + metrocity + age25 +
    hospbed + risk + ins, data = covid5)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3972 -0.8801 -0.1363  0.4958  3.0260

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.270970   8.708135   1.409   0.1662
old          -0.279934   0.157157  -1.781   0.0821 .
cum_pos_rate  0.067303   0.027289   2.466   0.0178 *
metrocity    -0.001708   0.009320  -0.183   0.8555
age25        -0.318693   0.161995  -1.967   0.0558 .
hospbed       0.129125   0.278947   0.463   0.6458
risk          0.101047   0.042107   2.400   0.0209 *
ins           0.018291   0.073818   0.248   0.8055
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.365 on 42 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4539,    Adjusted R-squared:  0.3629
F-statistic: 4.988 on 7 and 42 DF,  p-value: 0.0003589
```

Table 2: Diagnostic Plots



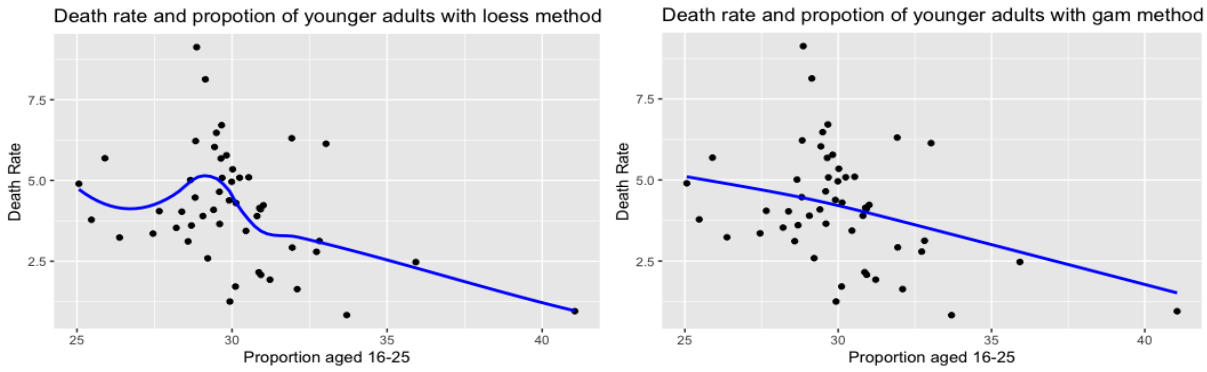
The scatterplots with best fitted lines using loess and generalized additive model methods show that a pretty flat curve between the proportion of seniors and the death rate, indicating that there may not be a very strong linear relationship between them.

Figure2: Scatterplots between death rate and proportion of older people



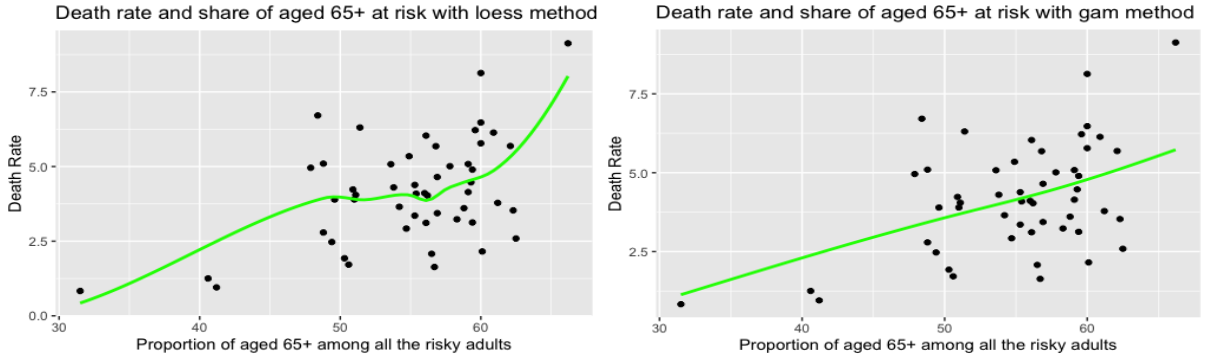
Following scatterplots with best fitted lines using loess and gam methods display the relationship between the proportion of younger adults and the death rate. It indicates that the death rate has a linear correlation with the proportion of younger adults. When the proportion of younger adults rises, the death rate diminishes.

Figure3: Scatterplots between death rate and proportion of younger adults



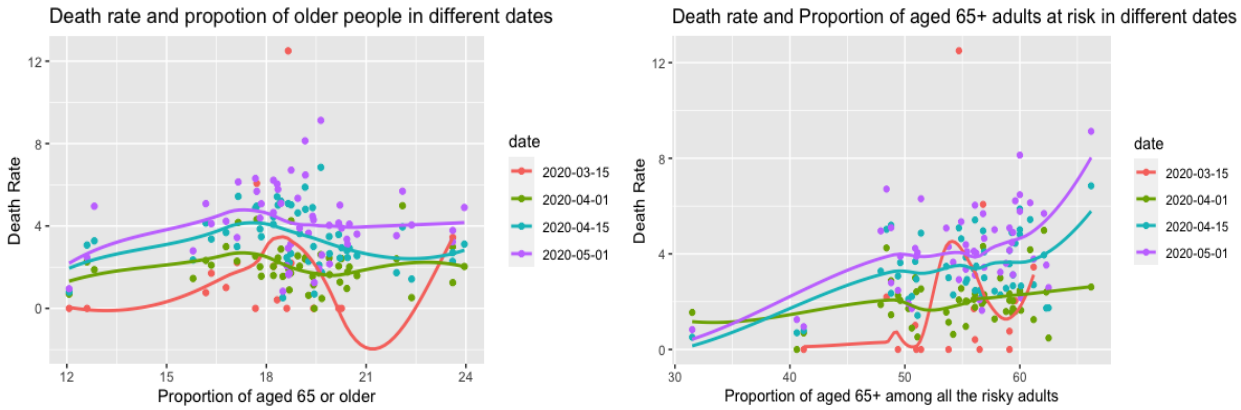
We also visualize the relationship between proportion of aged 65+ among adults at risk with scatterplots and best fitted lines using loess and gam methods. It shows that the death rate and the proportion of aged 65+ among adults at risk have positive and linear association.

Figure 4: Scatterplots between death rate and proportion of aged 65+ adults at risk



Following two plots show the associations between death rates and the proportion of older people and the share of age 65+ among adults at risk in each state on different dates from March 15th to May 1st. The first plot shows that death rates does not have linear correlations with the proportion of older people. However, the second plot shows that on April 15th and May 1st, the share of aged 65+ among all adults at risk has a positive and linear relationship with the death rate, but the association is lacking on March 15th and April 1st.

Figure 5: Scatterplots between death rate and two variables on different dates



In order to visualize the differences on the associations between the death rates and the proportion of older people in different regions in the U.S, the plot shows that in the west, south, northeast, and mid-west regions, the death rate does not have strong linear correlation with the proportion of older adults.

Figure 6: Scatterplot between death rate and share of older adults in different regions

4. Conclusion

Our study finds that the proportion of older people has a relationship with the death rate, but the association is not strong and not linear. Table 1 shows that the death rate varies by 0.28% for 1% growth in the proportion of older people, and the p-value is less than 0.1. We show the lack of linear correlation between age and death rates using Figure 2, 5, and 6 with various specifications.

Interestingly, we find the share of adults aged 65+ among adults at risk have a strong and linear correlation with the death rate. Table 1 shows that the share of older adults at risk is a significant predictor of the death rate since p-value is less than 0.05, and Figure 4 and 5 indicates its association with death rate is strong and positive.

The limitation of our data is that the age distribution and insurance coverage of each state is from 2018 data. It may deviate from the current age distribution, so the estimation result may be not accurate. Another issue is some variables such as current patients in ICU contains many missing values. The limitation of method is that the linear regression model may cannot reflect the true relationship between age and death rates because of the omitted variable bias. Also, we only select the death rate of one day as the response variable, so it may be not representative.

We hope to further improve this study by applying Cross-Validation and PCA to analyze the correlations between the proportion of older people and death rates. We also plan to use bootstrap to simulate the p-value.