aaaa*

Yiliu Cao

October 31, 2023

SSSSS

Table of contents

*Code and data from this analysis are available at: https://github.com/yiliuc/Bicycle_Thefts_2023.git

1 Introduction

On May 23rd, the WHO chief Tedros announced that the COVID-19 ends and COVID-19 is no longer regarded as a global threat. At this point, the epidemic that lasted for three years has officially ended. During the three years, nearly 7.6 billion people got infected and about 6.9 million people lost their lives. With that said, the US seems to be one of the countries that had influenced by COVID-19 most, there was about 1.1 million deaths and 104 million infections. The mortality rate of COVID-19 in US is about 341 per 100,000, which is significantly higher than other western developed countries. Meanwhile, the COVID was entirely occured in the Biden's term and next year will be the Federal Election. It is worth to know how the COVID deaths is related to the politics and how to recover it, especially that we are now at the post-pandemic period.

In this paper, I aim to find the variations of death rate across different counties, to see how the counties with different socio-economic factors were affected by COVID differently. In addition, I will also include the political factors, to see whether the party that counties voted for will display any differences on COVID death. There are three primary data sources which are American Community Survey collecting the socio-economic information for each county, the John-Hopkins public data for COVID cases and deaths for each county and the 2020 US Federal Election from Fox News. With capturing this variations, it can help us to access the inequality of mortality caused by COVID across different counties. Besides, people living at each county will also know whether their political reference will affect the death rate, this can significantly judge them about how they will vote on the 2024 Federal Election.

There will be four main parts in this paper: Data, Models, Results, Discussion and Conclusion. In the Data part, I will introduce the data used in this paper and high light the key variables. The Models and Results session will introduce the model that will be used in this paper and the results from it. Moreover, the results will be interpreted and provide the insights about COVID deaths rate. Lastly, I will conclude the entire paper and discussion the limitations and drawbacks.

Table 1: The summary of important variabls from ACS

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
no_insurance	252	0	9.6	5.1	0.0	8.5	44.9
age_85	73	0	2.3	1.0	0.0	2.1	12.0
high_education	449	0	22.7	9.6	0.0	20.3	76.3

2 Data

In this paper, there are three main sources of data, which are American Community Survey (ACS), The Center for Systems Science and Engineering (CSSE) at John-Hopkins University, Fox News. These three sources contains different dimensional data in predicting the death rate of COVID in each county.

2.1 Source of Data

This paper will use five data sets from three different sources. They are shown below.

2.1.1 American Community Survey (ACS)

The data from ACS contains the socio-economic information for each county. While there are many tables from ACS, I will only take three of them which are DP02, DP03 and DP05. All the tables are the 2021 five-year ACS estimates

For DP02, it contains the data regarding the social characteristics. In this paper, I will only take the information regarding the educational attainment such as the percentage of residents in each county that have at a least bachelor degree. Compared to DP02, DP03 contains the information about the economic characteristics. To understand the deaths rate in terms of economy, I will grab the information about the proportion of residents with a private insurance, the mean household income and unemployment rate at each county. Finally, DP05 contains the demographic data and housing estimates. In this paper, I will take the total population, proportions of children and people above 85 and percentage of white and black residents.

2.1.2 CSSE at JHU

The CSSE at JHU collects the worldwide COVID data since 2020 to 2023. All the raw data are available on their CSSE GitHub which can easily access. In particular, the COVID daily report since 2020 for each county on US are listed. However, they stop to update the data after March 9th, 2023 as no significant changes. The COVID data for US used in the paper is data collected on March 9th.

Table 2: The summary of infection and death rate of COVID in US

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
infection_rate	2867	0	306.9	107.6	48.7	302.4	4855.4
death_rate	2847	0	4.3	1.7	0.0	4.2	13.7

In the data on March 9th, it contains the number of confirmed cases and deaths for each county in US, as well as the incident rate. To perform the data analysis later in this paper, I will only use the COVID cases and deaths in each county.

Descriptions here

2.1.3 Fox News

It contains the county-level election results during the 2020 Federal Election. For each county, it contains the number of votes for the demographic and republican party.

2.2 Data Cleaning

All data was collected and cleaned using the statistical programming language `R` [citation]. The packages used are `tidyverse` [citation], `dplyr` [citation], `stringr` [citation] and `janitor` [citation]. Besides, `here` [citation], `modelsummary` [citation], `girdExtra` [citation], `usmap` [citation], `knitr` [citation], `kableExtra` [citation] are used to analyse the data.

The procedure to clean the three sources of data are different. For all the ACS data downloaded using ACS API [citation], referring to the codebook published by ACS [citation], I only extract the necessary variables. Besides, I clean the name for each county and convert the state name to its abbreviation form. In terms of the COVID data from JHU, I only select the cases and deaths and filter only the counties in US. Lastly for the election data, I extract the county name and also transform the state name to its abbreviation form. I also calculate the number of people voting for Republican party per 1000 voters.

After cleaned each data set, I merge the the five data sets from the above three data sources by the county and state. Based on the merged data set, I calculate the infection and death number per 1000 residents in each county. Therefore, each row in the merged data represent a US county and provides the the demographic information such as the total population. In addition, the data also contains the voting pattern in each county such as the number of votes for democratic and republican party.

Table 3: Summary of important variables in the model

Variable	Encoded Name	Description
High education attainment	high_education	The proportion of local residents with at least Bachelor degree
Income Percentile	IncomePctile	The mean household income percentile representing the income level for each county
People with no insurance	no_insurance	The proportion of local residents without insurance
People with private insurance	private_insurance	The proportion of local residents with private insurance
Males	males	The proportion of male residents
People aged above 85	age_85	The proportion of people aged above 85
White people	white	The proportion of white people
Black people	black	The proportion of black people
People supporting Republican	rep_rate	Number of voters supporting Republican by 1000

The Table 3 summarizes all the important variables in this paper. It contains the social characteristics such as high education attainment and economic factors such as income percentile. Moreover, the demographic information and political preference for each county are also included

To understand the data better, Figure 1 shows the the distribution of death rate for the counties voting for the two parties. We can observe that difference in political preference results differences in death rate. Counties voting for republican has a relative higher death rate than the counties voting for the Democratic party. In addition, there have outliers for both groups.

To understand how the death rate correlates with the political preference, Figure 2 compares the support for Republican with corresponding death rate. From the graph, we can observe that the dark color corresponds to the dark blue, meaning that the counties voting for Republican seem to have a higher death rate than the counties voting for Democratic party. This pattern is especially significant for the central states.

2.3 Data limitations

Our data combines five data sets from three sources. However, it still have limitations which may affect the analyses, reducing the accuracy of our predictions.

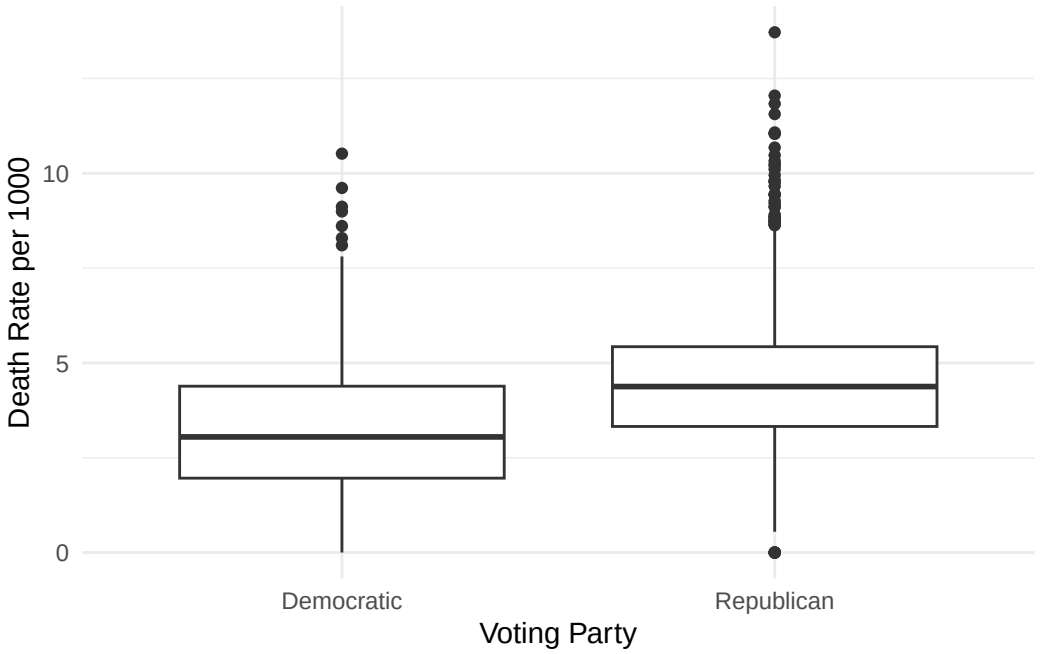


Figure 1: Summary of deaths rate for counties voting for Democratic and Republican

First of all, there is lack of COVID information in certain states, especially for the state "Utah." We can observe this from Figure 2 which some counties shows a color of grey. Due to this lackness, the merged data set will also lack the information on these counties even though other data sets have that information. The accuracy of the models may be also affected. In addition, the data we have may lack of spread, meaning that there could have other information which may also affect the death rate but are not included in this paper such as the hospital capacity. Our analyse may be less convincing.

3 Methods

As I introduced in the introduction, this paper aim to predict the death rate of COVID in each county of US and to detect whether this correlation related to their political preferences. Based on the merged data we have above, I will perform a series of models, from simple to complicated, to detect which predictors are significant in predicting the death rate.

3.1 Model Specifics

Before we fit the models, I will split the data into training and testing data. The training data is used to fit all the models and they are tested on the testing model, to see which one performs betters.

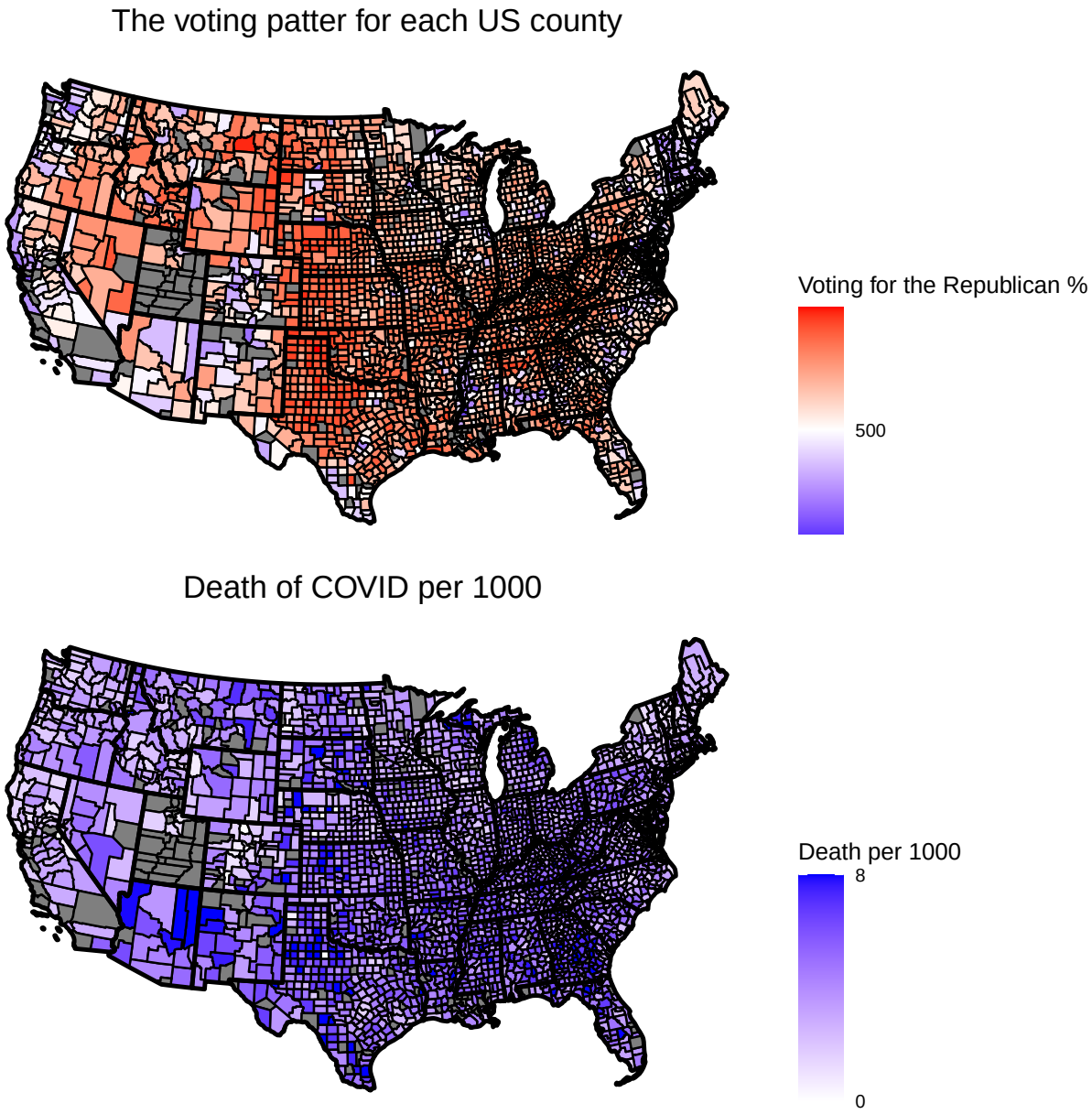


Figure 2: Election results of 2020 Federal Election and the corresponding death rate for each US county

For each OLS regression model, the response variable will always be the death rate of COVID in each county. However, since the deaths rate can be attributed by various perspectives such as demographic and economics, and we also want to see whether it depends on the party they voted during the last Federal Election. I will create two sets of models, which the first set does not consider the political preferences but the second does. In addition, for each sets of model, there will be three models with each one considering only one perspective of predictors, an overall model containing all predictors and one best model based on R^2_{adj} and $RMSE$. That said, I will fit 10 models in total along with two best models. To decide which model is the best one, I will use test them using the testing data, to see which one has a lower testing $RMSE$.

In addition to OLS regression models, I will also draw the regression tree and

The general form of the regression models is:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

where: - Y represents the response variables which is the death rate. - β_0 represents the intercept, which is the death rate holding all variables to zero - β_j represents the change of death rates with one unit increase in X_j - ϵ captures measurement errors and other discrepancies

3.2 Assumptions of the regression models

After we fit the linear regression models, it is essential to check the assumptions to ensure the accuracy of our predictions. There are four main assumptions which are linearity, uncorrelated error, constant variance and normality. For the first three assumptions, we can check it by plot the residuals with fitted values. However, we need to plot the Normal QQ plot to check the normality.

3.3 Regression Tree and Importance Matrix

A regression tree is a type of decision tree used specifically for regression problems, where the goal is to predict a continuous outcome variable based on one or more predictor variables. By using the regression tress, we can explain the decisions, identify possible events that might occur, and see potential outcomes

The objection function of regression tree is:

$$\min_{j,s} \left[\sum_{i:x_{i,j} \leq s, x_i \in R1} (y_i - \hat{y}_{R1})^2 + \sum_{i:x_{i,j} > s, x_i \in R2} (y_i - \hat{y}_{R2})^2 \right]$$

The objective function is to minimize the squared error. The j, s represents the index of the variable and threshold, respectively. In my regression tree, the two regions, R_1 and R_2,

Table 4: Summary of the Best Model Without Politics Preference

Predictor	Estimate	Standard Error	Statistics	P-value
(Intercept)	7.521	0.793	9.479	0.000
high_education	-0.051	0.005	-10.899	0.000
pctile	-0.012	0.002	-6.792	0.000
no_insurance	0.066	0.008	8.242	0.000
private_insurance	-0.014	0.005	-2.642	0.008
males	-0.055	0.013	-4.299	0.000
age_85	0.303	0.034	9.012	0.000
white	0.011	0.004	2.766	0.006
black	0.013	0.004	3.149	0.002

are split by the percentage of the White population in 2020, and the threshold here is 0.094. This means all the rows with the percentage of White people in 2020 less or equal to 0.094 are assigned to R1; otherwise, the rest are assigned to R2. In addition, the $\hat{y}_{R1}$ and $\hat{y}_{R1}$ represent the mean value of the y in each region. According to the regression tree, we can observe that the mean change of percentage vote for Trump is 0.157 and 0.015, which are the values of $\hat{y}_{R1}$ and $\hat{y}_{R2}$.

In addition to regression trees, we can also use the importance matrix which can tell us which predictors are important in predicting the death rate for each county. It can help us to verify our final model.

4 Results

4.1 Death rate without considering the political preferences

Table 4 shows the summary table for the model without political references. Using the above information, we can write the equation model:

$$
\begin{aligned}
\text{Death Rate} = 7.521 & \\
& - 0.051 \times \text{prop\_higher\_education} \\
& - 0.012 \times \text{IncomePctile} \\
& + 0.066 \times \text{no\_insurance} \\
& - 0.014 \times \text{private\_insurance} \\
& - 0.055 \times \text{males} \\
& + 0.303 \times \text{old\_85} \\
& + 0.011 \times \text{white\_pct} \\
& + 0.013 \times \text{black\_pct}
\end{aligned}
$$

Based on the above model, holding all variables to be zero, the expected deaths rate is about 7.2 per 1000. Holding other variables fixed, with one percent in the proportion of people who has at least bachelor degree, the death rate is expected to increase by 0.05. In addition, with income percentile increases by one, the death rate is expected to decrease by 0.02. Unsurprisingly, there is a negative relationship between the proportion of people with no insurance and death rate, 0.7 more people might die if the ratio of people living in a county without insurance increase by 10%. In terms of demographic factors, the number of death is inversely correlated with the number of males in a county. One percent increase in the males will result 0.06 less death. Moreover, it seems that old people are more likely to be died by COVID. That said, 0.29 more death if one percent increase in the proportion of people aged above 85. Lastly, one percent increase in the proportion of white and black people are both expected to have 0.01 more death.

4.2 Death rate with considering the plotical preference

Table 5 shows the summary of the best model without and with the political preferences. Based on them, we can write the regression question.

Table 5: Summary of the Best Model With Politics Preference

Predictor	Estimate	Standard Error	Statistics	P-value
(Intercept)	7.596	0.774	9.816	0.000
rep_rate	0.003	0.000	10.190	0.000
high_education	-0.024	0.005	-4.517	0.000
pctile	-0.010	0.002	-5.797	0.000
private_insurance	-0.028	0.005	-5.130	0.000
no_insurance	0.027	0.009	3.155	0.002
males	-0.059	0.013	-4.722	0.000
age_85	0.308	0.033	9.391	0.000
white	-0.008	0.004	-1.999	0.046
black	0.013	0.004	3.277	0.001

$$
\begin{aligned}
\text{Death Rate} = 7.596 & \\
& + 0.003 \times \text{Rep\_Rate} \\
& - 0.024 \times \text{High\_Education} \\
& - 0.010 \times \text{Pctile} \\
& + 0.027 \times \text{No\_Insurance} \\
& - 0.028 \times \text{Private\_Insurance} \\
& - 0.059 \times \text{Males} \\
& + 0.308 \times \text{Age\_85} \\
& - 0.008 \times \text{White} \\
& + 0.013 \times \text{Black}
\end{aligned}
$$

The difference between the second and the first model is that the second model includes the political preference for each county as one predictor to predict the death rate due to COVID. Holding other variables, we can see that every 100 more votes per 1000 for republican party, 0.3 more people are expected to dead.

4.3 Comparing the testing error

The above two models are fitted using the training data. Comparing the training RMSE from Table 7 and **?@tbl-all_models_2**, we can find that the model containing the political preference has a lower training RMSE. However, in order to test the prediction performance of the models we have varying data, I will use testing data to find the test error for the two models above.

Table 6: The testing error of the above two models

Model	Testinf Error
Model 1	1.769733
Model 2	1.675969

Table 6 shows the testing error of the above two models. We can observe that the errors for the two models are 1.769733 and 1.675969. Therefore, the model containing the political preferences has a lower expecter MSE, hence it has better predictions than another one.

5 Discussion

To sum up, the second model which contains the political reference is more appropriate than another one, representing that the party each county supported during the last election may change the attitude that people face with COVID and hence results differences in death rate. In addition to that, it is also important to consider the rest variables while predicting the death rate.

5.1 The impact of political preference on death rate is more significant than expected, it may overcome the change of income levels.

From the Table 5, we can observe that both income percentile and the support of Republican both are significant in predicting the death rate of COVID, but with an inverse direction of their coefficients. More people voting for Republican party is expected to increase the number of deaths due to COVID in this county and a wealthier counties usually have a less COVID deaths. However, if we compare their values of coefficients, we can find that the increase in death by every three more people in 1000 voting for Republican party is equivalent to decrease in death rate by one increase in the income percentile. In other words, counties with the lowest income levels might be the worst areas sacrificing from COVID if they support the Republican party.

This is a surprising but also horrible observation as Donald Trump is especially in favor of poor people. Eight out of the top 10 counties with highest poverty rate voted for Donald Trump. Referring to Figure 2, we can clearly see that the above 80% states in the poorer states, middle US, had voted for Trump. Conversely, above 80% wealthier coastal states, such as California and New York, had voted for The Democratic party. These both verify our observations from the model. What Trump's votes does not know is that their support may bring more deaths to their counties.

But why, why the counties with higher income levels are less likely to vote for the Republican party and hence results in a lower death rate? One thing that can explain this pattern from our model is the proportion of people that have higher education level. According to Table 5, we can find that there is an inverse relationship between the death rate and the number of residents with high education level. Its value of coefficient is much higher than the previous two, indicating that education level is an essential factor in predicting the death rate.

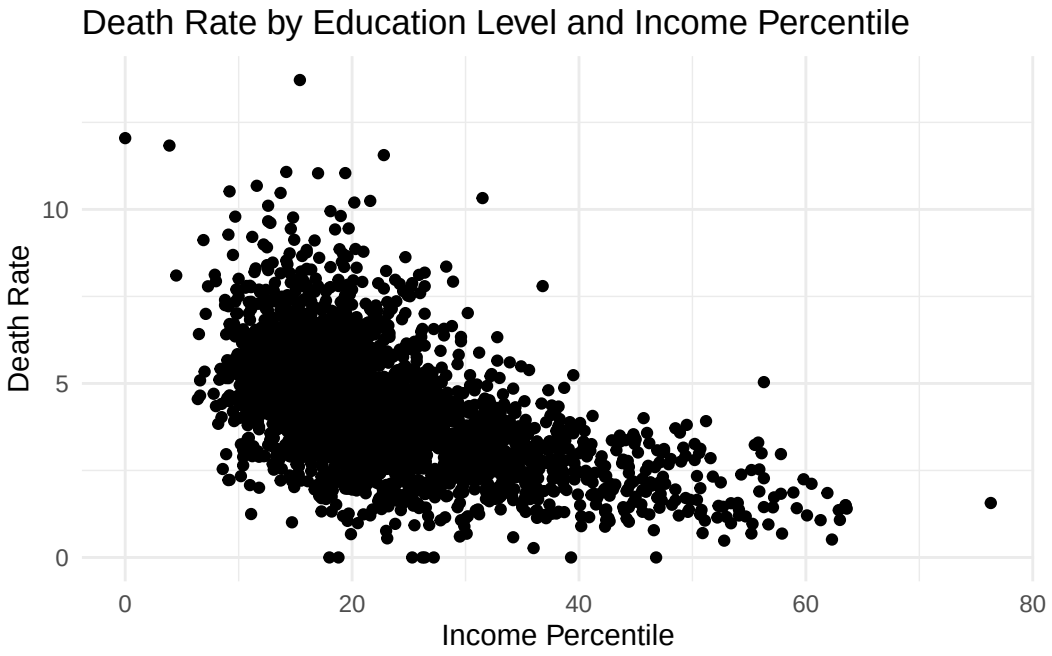
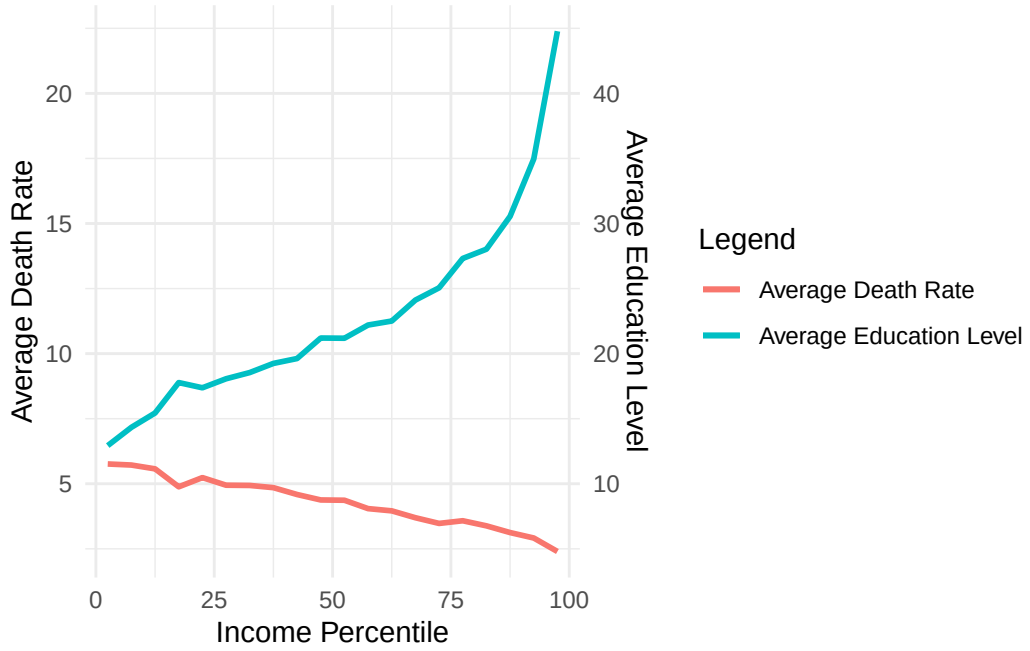


Figure 3: The testing error of the above two models

```r
summarized_data <- data %>%
  mutate(bin = cut(pctile, breaks = seq(0, 100, by = 5), include.lowest = TRUE, labels = s
  group_by(bin) %>%
  summarise(
    avg_death_rate = mean(death_rate, na.rm = TRUE),
    avg_education_level = mean(high_education, na.rm = TRUE)
  )

# Create the plot with dual y-axis
ggplot(summarized_data, aes(x = as.numeric(as.character(bin)))) +
  geom_line(aes(y = avg_death_rate, color = "Average Death Rate"), size = 1) +
  geom_line(aes(y = avg_education_level/2, color = "Average Education Level"), size = 1) +
  scale_y_continuous(sec.axis = sec_axis(~.*2, name = "Average Education Level")) + # Adju
  labs(x = "Income Percentile", y = "Average Death Rate", color = "Legend") +
```

```
  theme_minimal()
```



Education reflects this

5.2 Health insurance iss also an important factor in predicting the death rate of COVID.

In addition to the political preferences, the healthcare insurance is also significant in predicting the death rate for each county. Based on Table 5, it seems that both `no_insurance` and `private_insurance` are also significant in predicting the death rate for each county but they shows an inverse relationship on the death rate. In addition, the values of these two coefficients are almost same, meaning that the effects on death rate between them can be offset.

However, t

5.3 The importantance matrix matches our models

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

5.1 Model details

aaa

ojnoj

klmko

j jnjn

Table 7: Summary of the set of models with political preferences

	(1)	(2)	(3)	(4)	(5)
(Intercept)	6.524***	1.942+	−0.810	7.162***	7.521***
	(0.083)	(1.115)	(1.033)	(1.813)	(0.793)
high_education	−0.097***				−0.051***
	(0.003)				(0.005)
pctile		−0.024***		−0.013***	−0.012***
		(0.002)		(0.002)	(0.002)
unemployment		−0.078***		−0.035*	
		(0.015)		(0.016)	
no_insurance		0.100***		0.064***	0.066***
		(0.014)		(0.015)	(0.008)
private_insurance		0.017+		−0.016	−0.014**
		(0.010)		(0.012)	(0.005)
public_insurance		0.048***		0.005	
		(0.009)		(0.012)	
males			0.032*	−0.051***	−0.055***
			(0.016)	(0.014)	(0.013)
age_85			0.381***		0.303***
			(0.041)		(0.034)
children			0.073***	0.013	
			(0.012)	(0.012)	
white			0.009*		0.011**
			(0.004)		(0.004)
black			0.033***		0.013**
			(0.005)		(0.004)
prop_higher_education				−0.048***	
				(0.005)	
old_85				0.292***	
				(0.038)	
white_pct				0.010*	
				(0.004)	
black_pct				0.014**	
				(0.004)	
Num.Obs.	2013	2013	2013	2013	2013
R2	0.290	0.333	0.074	0.399	0.397
R2 Adj.	0.289	0.331	0.071	0.396	0.395
AIC	7222.9	7105.9	7765.5	6907.1	6907.5
BIC	7239.7	7145.1	7804.7	6980.0	6963.6
Log.Lik.	−3608.428	−3545.940	−3875.735	−3440.535	−3443.764
RMSE	1.45	1.41	1.66	1.34	1.34

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 8: Summary of the set of models with political preferences

	(1)	(2)	(3)	(4)	(5)
(Intercept)	5.336***	2.313*	6.182***	8.772***	7.596***
	(0.225)	(1.080)	(0.937)	(1.778)	(0.774)
rep_rate	0.001***	0.003***	0.007***	0.003***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
high_education	−0.085***			−0.026***	−0.024***
	(0.004)			(0.006)	(0.005)
pctile		−0.017***		−0.010***	−0.010***
		(0.002)		(0.002)	(0.002)
unemployment		−0.019		−0.019	
		(0.016)		(0.016)	
no_insurance		0.061***		0.024	0.027**
		(0.014)		(0.015)	(0.009)
private_insurance		−0.008		−0.032**	−0.028***
		(0.010)		(0.012)	(0.005)
public_insurance		0.028**		−0.002	
		(0.009)		(0.011)	
males			−0.062***	−0.064***	−0.059***
			(0.014)	(0.013)	(0.013)
age_85			0.300***	0.299***	0.308***
			(0.036)	(0.037)	(0.033)
children			−0.034**	−0.011	
			(0.011)	(0.012)	
white			−0.041***	−0.011*	−0.008*
			(0.004)	(0.005)	(0.004)
black			0.014***	0.011**	0.013**
			(0.004)	(0.004)	(0.004)
Num.Obs.	2013	2013	2013	2013	2013
R2	0.301	0.376	0.303	0.427	0.427
R2 Adj.	0.300	0.374	0.301	0.424	0.424
AIC	7192.9	6973.7	7195.5	6811.7	6807.8
BIC	7215.3	7018.6	7240.3	6890.2	6869.5
Log.Lik.	−3592.435	−3478.864	−3589.739	−3391.850	−3392.893
RMSE	1.44	1.36	1.44	1.30	1.31

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 References

https://news.un.org/en/story/2023/05/1136367 https://www.aljazeera.com/news/2023/5/11/three-years-1-1-million-deaths-covid-emergency-ending-in-us https://wisevoter.com/state-rankings/poorest-states/