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.

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.

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
old_65	270	0	19.3	4.7	3.1	18.9	57.6
prop_higher_education	449	0	22.7	9.6	0.0	20.3	76.3

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

2.1 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.

2.2 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.

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

2.3 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.

To predict the COVID death rate across counties with various socio-economic features, I will merge the the five data sets from the above three data sources by the county and state. In the merged data set, each row represent a US county and provides the number of COVID cases and deaths. In addition, it also shows the demographic information such as the total population and voting pattern such as the number of votes for democratic and republican party. The merged data contains three main parts of the data. In order to show the data better, I will draw some visualizations to show the data.

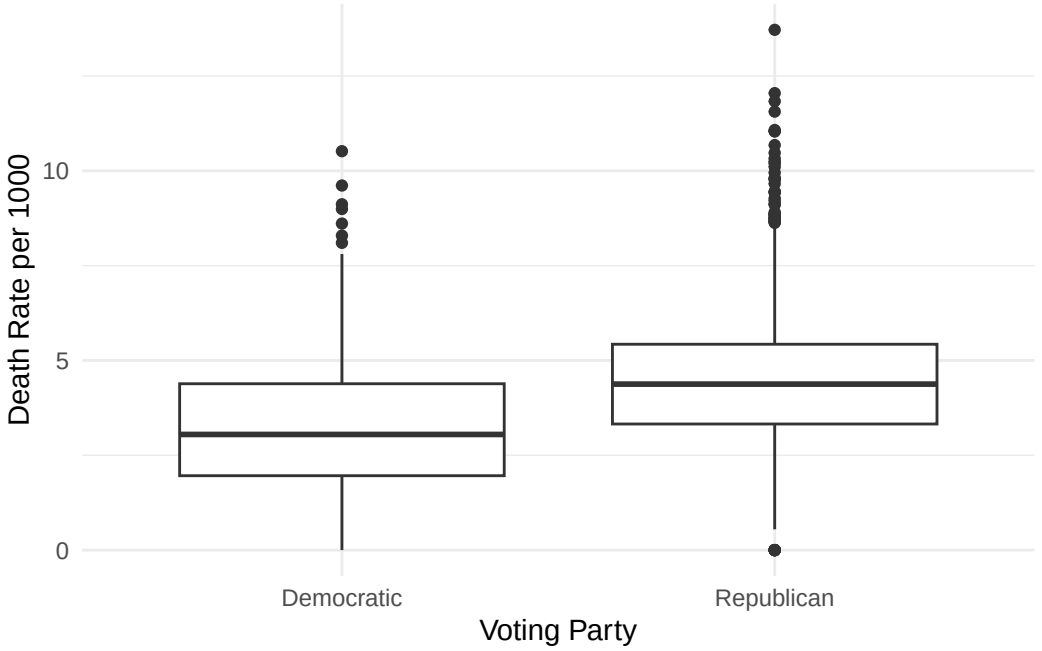

Figure 1: Summary of deaths rate for counties voting for Democratic and Republican

Sumamrise the above plot

Talk more about it.

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.

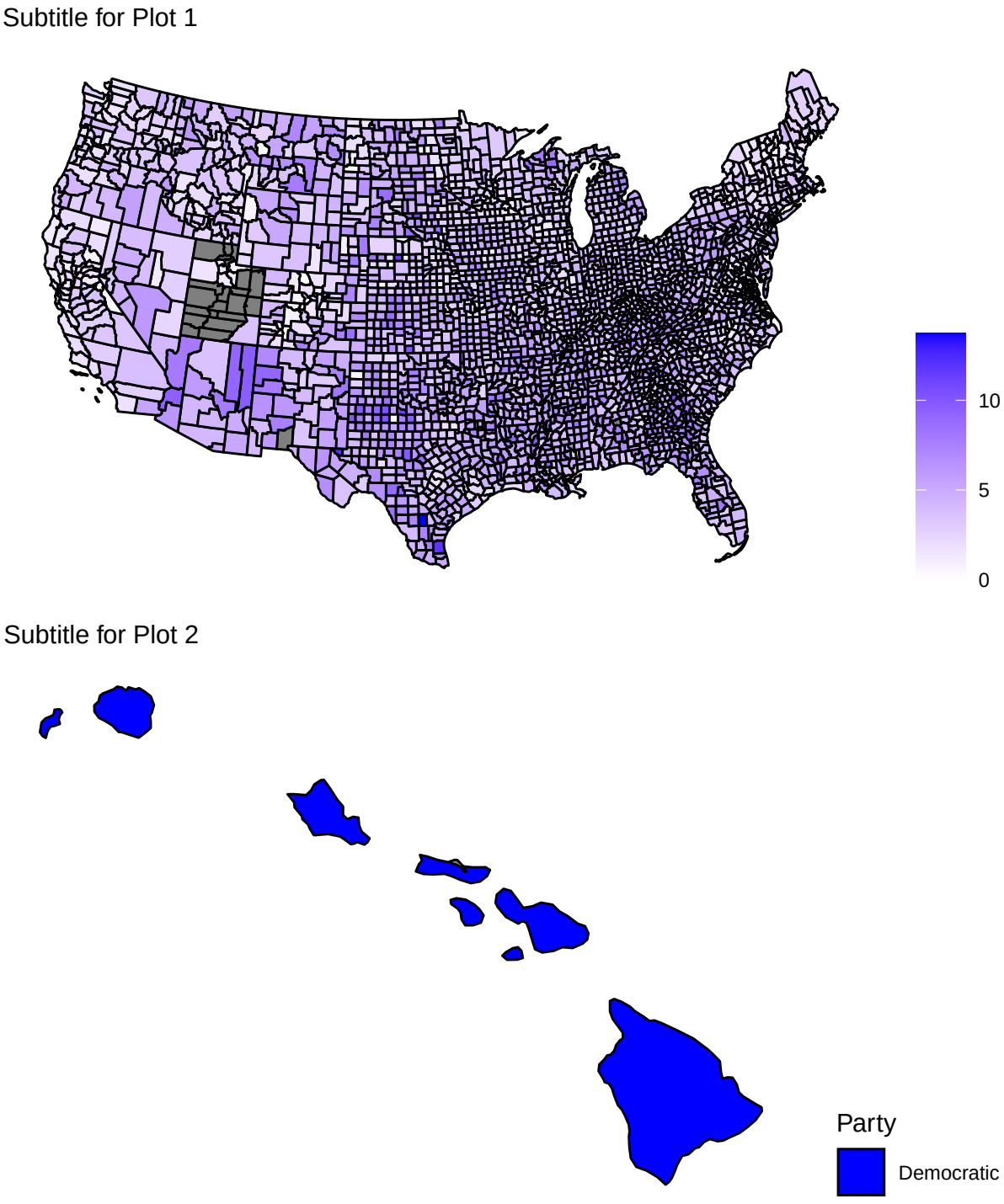

Figure 2: Number

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.

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, 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

(**first-model?**) shows the summary table for the model without political references. Using the above information, we can write the equation model:

```
model1 <- lm(death_rate ~ prop_higher_education + pctile + no_insurance + private_insuranc
summary(model1)
```

```
Call:
lm(formula = death_rate ~ prop_higher_education + pctile + no_insurance +
    private_insurance + males + old_85 + white_pct + black_pct,
    data = training_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1935 -0.7804 -0.0692  0.7095  8.1750
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.521125   0.793412   9.479  < 2e-16 ***
prop_higher_education -0.050927   0.004673 -10.899  < 2e-16 ***
pctile                -0.012395   0.001825  -6.792 1.45e-11 ***
no_insurance           0.065782   0.007981   8.242 3.02e-16 ***
private_insurance     -0.014423   0.005458  -2.642  0.00829 **
males                 -0.055192   0.012838  -4.299 1.80e-05 ***
old_85                 0.302689   0.033586   9.012  < 2e-16 ***
white_pct              0.010604   0.003833   2.766  0.00572 **
black_pct              0.012569   0.003991   3.149  0.00166 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.342 on 2004 degrees of freedom
Multiple R-squared:  0.3969,    Adjusted R-squared:  0.3945
F-statistic: 164.9 on 8 and 2004 DF,  p-value: < 2.2e-16
```

```r
model2 <- lm(death_rate ~ rep_rate + prop_higher_education + pctile +
                private_insurance  + no_insurance +
                males + old_85  + white_pct +
                black_pct, data = training_data)
summary(model2)
```

```
Call:
lm(formula = death_rate ~ rep_rate + prop_higher_education +
    pctile + private_insurance + no_insurance + males + old_85 +
    white_pct + black_pct, data = training_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5621 -0.7617 -0.0373  0.6445  7.9472

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.5957521  0.7738404   9.816  < 2e-16 ***
rep_rate              0.0033300  0.0003268  10.190  < 2e-16 ***
prop_higher_education -0.0238316  0.0052761  -4.517 6.64e-06 ***
pctile                -0.0103814  0.0017907  -5.797 7.80e-09 ***
private_insurance     -0.0281733  0.0054917  -5.130 3.17e-07 ***
```

```
no_insurance           0.0272949  0.0086520   3.155  0.00163 **
males                 -0.0591492  0.0125267  -4.722 2.50e-06 ***
old_85                 0.3076497  0.0327600   9.391  < 2e-16 ***
white_pct             -0.0083466  0.0041754  -1.999  0.04574 *
black_pct              0.0127542  0.0038926   3.277  0.00107 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.309 on 2003 degrees of freedom
Multiple R-squared:  0.4267,    Adjusted R-squared:  0.4241
F-statistic: 165.6 on 9 and 2003 DF,  p-value: < 2.2e-16
```

$$
\begin{aligned}
\text{Death Rate} = 7.23 & \\
& - 0.05 \times \text{prop\_higher\_education} \\
& - 0.02 \times \text{pctile} \\
& + 0.07 \times \text{no\_insurance} \\
& - 0.06 \times \text{males} \\
& + 0.29 \times \text{old\_85} \\
& + 0.01 \times \text{white\_pct} \\
& + 0.01 \times \text{black\_pct}
\end{aligned}
$$

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.

(**second-model?**) shows the summary of the best model without and with the political preferences. Based on them, we can write the regression question

$$
\begin{aligned}
\text{Death Rate} = 7.75 & \\
& + 0.0034980 \times \text{Rep\_Rate} \\
& - 0.4972847 \times \text{Rep\_Win} \\
& - 0.0120727 \times \text{Prop\_Higher\_Education} \\
& - 0.0099034 \times \text{Pctile} \\
& - 0.0284321 \times \text{Private\_Insurance} \\
& + 0.0281570 \times \text{No\_Insurance} \\
& - 0.0593644 \times \text{Males} \\
& + 0.3024402 \times \text{Old\_85} \\
& - 0.0169198 \times \text{White\_Pct} \\
& + 0.0130027 \times \text{Black\_Pct} \\
& - 0.0172526 \times \text{Rep\_Win:Prop\_Higher\_Education} \\
& + 0.0125389 \times \text{Rep\_Win:White\_Pct}
\end{aligned}
$$

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 for republican party, 0.3 more people are expected to dead. In addition, if the county votes for the Republican, then it is expected to have

5 Discussion

5.1 The impact of political preference on death rate is more significant than expected.

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 However, the socio-economic variables especially the private insurance, also play an important role.

5.3 The importantance matrix matches our models

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Table 3: Summary of the Best Model Without Politics Preference

Predictor	Estimate	Standard Error	Statistics	P-value
(Intercept)	7.52	0.79	9.48	0.000
prop_higher_education	-0.05	0.00	-10.90	0.000
pctile	-0.01	0.00	-6.79	0.000
no_insurance	0.07	0.01	8.24	0.000
private_insurance	-0.01	0.01	-2.64	0.008
males	-0.06	0.01	-4.30	0.000
old_85	0.30	0.03	9.01	0.000
white_pct	0.01	0.00	2.77	0.006
black_pct	0.01	0.00	3.15	0.002

Table 4: Summary of the Best Model With Politics Preference

Predictor	Estimate	Standard Error	Statistics	P-value
(Intercept)	7.60	0.77	9.82	0.000
rep_rate	0.00	0.00	10.19	0.000
prop_higher_education	-0.02	0.01	-4.52	0.000
pctile	-0.01	0.00	-5.80	0.000
private_insurance	-0.03	0.01	-5.13	0.000
no_insurance	0.03	0.01	3.15	0.002
males	-0.06	0.01	-4.72	0.000
old_85	0.31	0.03	9.39	0.000
white_pct	-0.01	0.00	-2.00	0.046
black_pct	0.01	0.00	3.28	0.001

Appendix

5.1 Model details

aaa

ojnoj

klmko

j jnjn

Table 5: Summary of the set of models with political preferences

	(1)	(2)	(3)	(4)	(5)
(Intercept)	6.524***	1.942+	−0.810	7.187***	7.521***
	(0.083)	(1.115)	(1.033)	(1.812)	(0.793)
prop_higher_education	−0.097***			−0.049***	−0.051***
	(0.003)			(0.005)	(0.005)
pctile		−0.024***		−0.009**	−0.012***
		(0.002)		(0.003)	(0.002)
unemployment		−0.078***		−0.035*	
		(0.015)		(0.016)	
no_insurance		0.100***		0.063***	0.066***
		(0.014)		(0.015)	(0.008)
private_insurance		0.017+		−0.017	−0.014**
		(0.010)		(0.012)	(0.005)
public_insurance		0.048***		0.005	
		(0.009)		(0.012)	
males			0.032*	−0.052***	−0.055***
			(0.016)	(0.014)	(0.013)
old_85			0.381***	0.293***	0.303***
			(0.041)	(0.038)	(0.034)
children			0.073***	0.013	
			(0.012)	(0.012)	
white_pct			0.009*	0.010*	0.011**
			(0.004)	(0.004)	(0.004)
black_pct			0.033***	0.014**	0.013**
			(0.005)	(0.004)	(0.004)
high_income				−0.212+	
				(0.120)	
Num.Obs.	2013	2013	2013	2013	2013
R2	0.290	0.333	0.074	0.400	0.397
R2 Adj.	0.289	0.331	0.071	0.396	0.395
AIC	7222.9	7105.9	7765.5	6905.9	6907.5
BIC	7239.7	7145.1	7804.7	6984.4	6963.6
Log.Lik.	−3608.428	−3545.940	−3875.735	−3438.973	−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 6: 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)
prop_higher_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)
old_85			0.300***	0.299***	0.308***
			(0.036)	(0.037)	(0.033)
children			−0.034**	−0.011	
			(0.011)	(0.012)	
white_pct			−0.041***	−0.011*	−0.008*
			(0.004)	(0.005)	(0.004)
black_pct			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